ANA ROCÍO CÁRDENAS MAITA

Um estudo da aplicação de técnicas de
inteligência computacional e de aprendizado
de máquina em mineração de processos de
negócio

São Paulo

2016

ANA ROCÍO CÁRDENAS MAITA

**Um estudo da aplicação de técnicas de inteligência
computacional e de aprendizado de máquina em
mineração de processos de negócio**

Versão corrigida

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 04 de dezembro de 2015. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Marcelo Fantinato

Coorientador: Profa. Dra. Sarajane Marques Peres

São Paulo

2016

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Cárdenas Maita, Ana Rocío

Um estudo da aplicação de técnicas de inteligência computacional e de aprendizado de máquina em mineração de processos de negócio / Ana Rocío Cárdenas Maita ; orientador, Marcelo Fantinato. – São Paulo, 2016

129 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Versão corrigida

1. Negócios - Processos. 2. Negócios - Gerenciamento.
3. Mineração de dados. 4. Workflow. I. Fantinato, Marcelo,
orient. II. Título

CDD 22.ed.— 658.4063

Dissertação de autoria de Ana Rocío Cárdenas Maita, sob o título **“Um estudo da aplicação de técnicas de inteligência computacional e de aprendizado de máquina em mineração de processos de negócio”**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 04 de dezembro de 2015 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Adriana Backx Noronha Viana
Presidente
Faculdade de Economia e Administração da Universidade de São Paulo

Profa. Dra. Lucinéia Heloisa Thom
Instituto de Informática da Universidade Federal do Rio Grande do Sul

Prof. Dr. Luciano Antonio Digiampietri
Escola de Ciências Artes e Humanidades da Universidade de São Paulo

A minha mãe por ter me apoiado sempre na busca dos meus objetivos, e que com amor e carinho me impulsa cada dia a dar o melhor de mim.

Agradecimentos

Agradeço a minha família, pela força e apoio incondicional na realização desse sonho, desde antes de vir para o Brasil e em cada momento dessa etapa parte da minha vida acadêmica.

Meus agradecimentos sinceros aos meus orientadores Prof. Marcelo Fantinato e Prof^a. Sarajane Marques Peres pelo apoio, paciência e compreensão pessoal durante toda esta etapa de estudo. Além do direcionamento cuidadoso durante a realização do projeto e suas valiosas contribuições a minha formação acadêmica e profissional. Ficarei sempre grata a eles por terem me dado a oportunidade de ser parte desse programa de pós-graduação, conhecer o mundo da pesquisa e apresentado um caminho que pretendo continuar.

Aos professores do PPgSI, especialmente àqueles com quem tive a oportunidade de cursar as disciplinas e trabalhar diretamente: Prof. Marcos Chaim, Prof. João Bernardes, Prof. Edmir Prado, Prof. Luciano Araújo, Prof^a. Ariane Machado, Prof. Marcelo Eler, e Prof. Luciano Digiampietri, cujos ensinamentos foram muito valiosos para a culminação deste trabalho.

Agradeço também aos órgãos de fomento, especialmente à Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) pelo financiamento deste trabalho por meio de processo 2013/17520-7, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), e à Universidade de São Paulo.

“Não podemos resolver problemas pensando da mesma maneira que quando os criamos.”
(Albert Einstein, 1879-1955)

Resumo

MAITA, Ana Rocío Cárdenas. **Um estudo da aplicação de técnicas de inteligência computacional e de aprendizado de máquina em mineração de processos de negócio**. 2016. 129 f. Dissertação (Mestrado em Ciências) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

Mineração de processos é uma área de pesquisa relativamente recente que se situa entre mineração de dados e aprendizado de máquina, de um lado, e modelagem e análise de processos de negócio, de outro lado. Mineração de processos visa descobrir, monitorar e aprimorar processos de negócio reais por meio da extração de conhecimento a partir de *logs* de eventos disponíveis em sistemas de informação orientados a processos. O principal objetivo deste trabalho foi avaliar o contexto de aplicação de técnicas provenientes das áreas de inteligência computacional e de aprendizado de máquina, incluindo redes neurais artificiais¹ e máquinas de vetores de suporte, no contexto de mineração de processos. Considerando que essas técnicas são, atualmente, as mais aplicadas em tarefas de mineração de dados, seria esperado que elas também estivessem sendo majoritariamente aplicadas em mineração de processos, o que não tinha sido demonstrado na literatura recente e foi confirmado por este trabalho. Buscou-se compreender o amplo cenário envolvido na área de mineração de processos, incluindo as principais características que têm sido encontradas ao longo dos últimos dez anos em termos de: tipos de mineração de processos, tarefas de mineração de dados usadas, e técnicas usadas para resolver tais tarefas. O principal enfoque do trabalho foi identificar se as técnicas de inteligência computacional e de aprendizado de máquina realmente não estavam sendo amplamente usadas em mineração de processos, ao mesmo tempo que se buscou identificar os principais motivos para esse fenômeno. Isso foi realizado por meio de um estudo geral da área, que seguiu rigor científico e sistemático, seguido pela validação das lições aprendidas por meio de um exemplo de aplicação. Este estudo considera vários enfoques para delimitar a área: por um lado, as abordagens, técnicas, tarefas de mineração e ferramentas comumente mais usadas; e, por outro lado, veículos de publicação, universidades e pesquisadores interessados no desenvolvimento da área. Os resultados apresentam que 81% das publicações atuais seguem as abordagens tradicionais em mineração de dados. O tipo de mineração de processos com mais estudo é Descoberta 71% dos estudos primários. Os resultados deste trabalho são valiosos para profissionais e pesquisadores envolvidos no tema, e representam um grande aporte para a área.

Palavras-chaves: BPM. Gestão de Processos de Negócio. Mineração de dados. Mineração de processos. Mineração de *workflows*. Inteligência computacional. Aprendizado de máquina.

¹ Para fins de simplificação, denominadas no restante deste texto apenas como “redes neurais”.

Abstract

MAITA, Ana Rocío Cárdenas. **A study of the application of computational intelligence and machine learning techniques in business process mining**. 2016. 129 p. Dissertation (Master of Science) — School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2015.

Mining process is a relatively new research area that lies between data mining and machine learning, on one hand, and business process modeling and analysis, on the other hand. Mining process aims at discovering, monitoring and improving business processes by extracting real knowledge from event logs available in process-oriented information systems. The main objective of this master's project was to assess the application of computational intelligence and machine learning techniques, including, for example, neural networks and support vector machines, in process mining. Since these techniques are currently widely applied in data mining tasks, it would be expected that they were also widely applied to the process mining context, which has been not evidenced in recent literature and confirmed by this work. We sought to understand the broad scenario involved in the process mining area, including the main features that have been found over the last ten years in terms of: types of process mining, data mining tasks used, and techniques applied to solving such tasks. The main focus of the study was to identify whether the computational intelligence and machine learning techniques were indeed not being widely used in process mining whereas we sought to identify the main reasons for this phenomenon. This was accomplished through a general study area, which followed scientific and systematic rigor, followed by validation of the lessons learned through an application example. This study considers various approaches to delimit the area: on the one hand, approaches, techniques, mining tasks and more commonly used tools; and, on the other hand, the publication vehicles, universities and researchers interested in the development area. The results show that 81% of current publications follow traditional approaches to data mining. The type of mining processes more study is Discovery 71% of the primary studies. These results are valuable for practitioners and researchers involved in the issue, and represent a major contribution to the area.

Keywords: BPM. Business Process Management. Process Mining. Workflow Mining. Data Mining. Computational Intelligence. Machine Learning.

Lista de figuras

Figura 1 – Ciclo de vida de BPM – proposta 1	25
Figura 2 – Ciclo de vida de BPM – proposta 2	25
Figura 3 – Tipos básicos de mineração de processos em termos de entradas e saídas	46
Figura 4 – Problemas e desafios decorrentes da análise de <i>logs</i> de eventos	48
Figura 5 – Modelo de processo descoberto após a aplicação do algoritmo α no <i>log</i> de eventos da tabela 3	53
Figura 6 – Etapas seguidas durante a realização deste trabalho	56
Figura 7 – Passos seguidos na identificação de estudos primários	67
Figura 8 – Distribuição dos estudos primários por ano	68
Figura 9 – Classificação resumida dos estudos primários considerando tipos de mineração de processos contra tipos de tarefas de mineração de dados e técnicas usadas	73
Figura 10 – Distribuição de estudos primários por ano (Mapeamento sistemático) .	79
Figura 11 – Tipos de mineração de processos, quando tratados individualmente . .	84
Figura 12 – Distribuição dos estudos primários considerando os tipos de mineração de processos identificados	84
Figura 13 – Tipos de mineração de processos, quando tratados juntos no mesmo estudo primário	85
Figura 14 – Tarefas de mineração de dados identificadas nos estudos primários de mineração de processos	86
Figura 15 – Distribuição por ano, no período de 2004 a 2014, de tarefas de mineração de dados aplicadas em mineração de processos	87
Figura 16 – Análise cruzada entre tipos de mineração de processos e tarefas de mineração de dados identificadas nos estudos primários	87
Figura 17 – Tarefas de mineração de dados mais amplamente usadas em combinação com outras tarefas quando aplicadas em um mesmo estudo primário mineração de processos	88
Figura 18 – Técnicas em mineração de dados aplicadas identificadas nos estudos primários de mineração de processos	90
Figura 19 – Principais agrupamentos dos tipos de técnicas em mineração de dados aplicadas em mineração de processos identificadas nos estudos primários	90

Figura 20 – Técnicas de inteligência computacional e de aprendizado de máquina aplicadas em mineração de processos identificadas nos estudos primários	91
Figura 21 – Análise cruzada entre tipos de mineração de processos e técnicas de mineração de dados, identificadas nos estudos primários	92
Figura 22 – Domínios de aplicação em mineração de processos	93
Figura 23 – Ferramentas usadas em mineração de processos	94
Figura 24 – Adesão dos resultados às definições usadas neste estudo	95
Figura 25 – Exemplo de um trecho no <i>log</i> de eventos original do sistema EVC . . .	98
Figura 26 – Exemplo de um trecho do <i>log</i> de eventos do sistema EVC após pré-processamento	101
Figura 27 – Exemplo do conjunto A	103
Figura 28 – Exemplo do conjunto B	103
Figura 29 – Exemplo do modelo de processo gerado usando o <i>plug-in</i> “ <i>Heuristic Miner</i> ” da ferramenta ProM	105
Figura 30 – Processo seguido pelo grupo de alunos com desempenho “Excelente” .	107
Figura 31 – Processo seguido pelo grupo de alunos com desempenho “Regular” . .	108
Figura 32 – Processo seguido pelo grupo de alunos com desempenho “Insatisfatório”	109
Figura 33 – Perspectivas dos resultados alcançados neste trabalho	114

Lista de tabelas

Tabela 1 – Fragmento de um arquivo de <i>logs</i> de eventos	29
Tabela 2 – Taxonomia de técnicas de inteligência computacional e de aprendizado de máquina usada neste trabalho	37
Tabela 3 – Conjunto de atividades sequenciais gerado a partir do <i>log</i> de eventos da tabela 3	52
Tabela 4 – <i>String</i> de busca genérica para a revisão sistemática	59
Tabela 5 – Critérios de avaliação da qualidade	61
Tabela 6 – <i>String</i> de busca genérica para o mapeamento sistemático	64
Tabela 7 – Lista final de estudos primários para revisão sistemática	68
Tabela 8 – Lista final de estudos primários para revisão sistemática	69
Tabela 9 – Classificação dos estudos primários com base no tipo de arquitetura da redes neurais	70
Tabela 10 – Classificação dos estudos primários considerando as variações de má- quinas de vetores de suporte	70
Tabela 11 – Classificação dos estudos primários considerando os tipos de aprendizado de redes neurais e de máquinas de vetores de suporte	70
Tabela 12 – Classificação dos estudos primários considerando as tarefas de mineração de dados	71
Tabela 13 – Classificação dos estudos primários em relação ao tipo de mineração de processos	72
Tabela 14 – Classificação resumida dos estudos primários considerando todos os aspectos avaliados juntos	77
Tabela 15 – Dezesete conferências com maior número de publicações em mineração de processos nos últimos dez anos	80
Tabela 16 – Sete periódicos com maior número de publicações em mineração de processos nos últimos dez anos	81
Tabela 17 – Dezesete países com maior número de publicações em mineração de processos nos últimos dez anos	82
Tabela 18 – Doze instituições de pesquisa com maior número de publicações em mineração de processos nos últimos dez anos	82

Tabela 19 – Dezesesseis pesquisadores com maior número de publicações em mineração de processos nos últimos dez anos	83
Tabela 20 – Estrutura original de um registro de <i>log</i> de eventos do sistema EVC . .	97
Tabela 21 – Estrutura de um registro de <i>log</i> do sistema EVC após pré-processamento	101
Tabela 22 – Parâmetros usados na rede neural <i>Multilayer Perceptron</i>	111
Tabela 23 – Medidas de avaliação para o classificador (261 instâncias positivas e 30 instâncias negativas no conjunto A; 70 instâncias positivas e 30 instâncias negativas no conjunto A após amostragem)	112

Lista de abreviaturas e siglas

BPM	Gestão de Processos de Negócio – <i>Business Process Management</i>
BPMN	Notação e Modelo de Processos de Negócio – <i>Business Process Model and Notation</i>
EPCs	Cadeias de Processos Dirigida por EventosEvent-driven Process Chains <i>Event-driven Process Chains</i>
KDD	Descoberta de conhecimento a partir de dados – <i>Knowledge Discovery from Data</i>
LTL	Lógica Temporal Linear – <i>Linear Temporal Logic</i>
SOA	Arquitetura Orientada a Serviços – <i>Service-oriented Architecture</i>
SVM	Máquinas de vetores de suporte – <i>Support Vector Machine</i>
SOM	Mapas Auto-organizados – <i>Self Organized Map</i>
UML	Linguagem Unificada de Modelagem – <i>Unified Modeling Language</i>
WS-BPEL	Linguagem de Execução de Processos de Negócio para Serviços Web – <i>Web Services Business Process Execution Language</i>
WSDL	Linguagem de descrição de Serviços Web – <i>Web Service Description Language</i>

Lista de símbolos

α	Letra grega Alfa
β	Letra grega Beta
γ	Letra grega Gama
θ	Letra grega Teta

Sumário

1	Introdução	17
1.1	Motivação e justificativa	18
1.2	Objetivos	21
1.3	Estrutura do documento	22
2	Fundamentação teórica	23
2.1	Gestão de processos de negócio	23
2.1.1	Ciclo de vida de BPM	23
2.1.2	Modelagem e implementação de processos de negócio . .	26
2.1.3	<i>Logs</i> de eventos de processos de negócio	27
2.2	Mineração de dados, inteligência computacional e apren- dizado de máquina	28
2.2.1	Tarefas de mineração de dados	30
2.2.2	Técnicas de inteligência computacional e de aprendizado de máquina	34
2.2.3	Técnicas tradicionais para mineração de dados	39
2.3	Mineração de processos de negócio	42
2.3.1	Tipos de mineração de processos	44
2.3.2	Perspectivas e problemas de mineração de processos . . .	46
2.3.3	Algoritmos e ferramentas usados	49
2.3.4	Exemplo ilustrativo de mineração de processos	51
2.4	Considerações finais	53
3	Método de pesquisa	54
3.1	Protocolo resumido da revisão sistemática (específica)	57
3.1.1	Questões de pesquisa	58
3.1.2	Fontes de dados e estratégia de busca	59
3.1.3	Seleção de trabalhos	60
3.1.3.1	Critérios de inclusão	60
3.1.3.2	Critérios de exclusão	60
3.2	Protocolo resumido do mapeamento sistemático (geral)	61

3.2.1	Questões de pesquisa	62
3.2.2	Fontes de dados e estratégia de busca	64
3.2.3	Seleção de trabalhos	64
3.2.3.1	Critérios de inclusão	64
3.2.3.2	Critérios de exclusão	64
4	Resultados alcançados	66
4.1	Revisão sistemática (específica)	66
4.1.1	Resultados da revisão sistemática	67
4.1.2	Análise dos resultados da revisão sistemática	72
4.2	Mapeamento sistemático (geral)	78
4.2.1	Resultados do mapeamento sistemático	78
4.2.1.1	Tipos de mineração de processos atendidos por mineração de dados	83
4.2.1.2	Tarefas de mineração de dados em mineração de processos	85
4.2.1.3	Técnicas usadas para tratar as tarefas de mineração de dados no contexto de mineração de processos	89
4.2.1.4	Análises adicionais	92
4.3	Um exemplo de aplicação em mineração de processos	95
4.3.1	Exploração do domínio de aplicação	96
4.3.2	Definição das técnicas, tarefas e tipos de mineração	97
4.3.3	Pré-processamento e criação dos conjuntos de dados . . .	100
4.3.4	Resultados: ferramentas Disco Fluxicon e ProM	104
4.3.5	Resultados: predição categórica - <i>Multilayer Perceptron</i>	111
4.4	Lições aprendidas	113
4.4.1	Lições aprendidas da revisão sistemática	113
4.4.2	Lições aprendidas do mapeamento sistemático	114
4.4.3	Lições aprendidas do exemplo de aplicação	117
5	Conclusão	120
	Referências ²	125

² De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Gestão de Processos de Negócio (BPM – *Business Process Management*) inclui métodos, técnicas e ferramentas para apoiar o projeto, a realização, a gestão e a análise de processos de negócio; os quais são uma evolução dos anteriormente chamados *workflows* de negócio, ou apenas *workflows* (AALST; HOFSTEDE; WESKE, 2003). *Workflow* se refere à automação de processo de negócio, em que documentos, informações ou tarefas são passadas de um participante para o outro de acordo com um conjunto de regras de procedimentos para se atingir um objetivo de negócio (HOLLINGSWORTH, 1995). Já o ciclo de vida de BPM vai além e inclui as seguintes fases: (i) modelagem de processos; (ii) instanciação de modelos de processo; (iii) execução e administração de instâncias de processo; (iv) monitoramento e auditoria de instâncias de processo em execução; e (v) avaliação e melhoria de modelos de processo (WESKE, 2012). Nessa última fase, o histórico de execução das instâncias de processo pode ser avaliado visando a sua otimização, por meio dos dados de registro dos resultados das etapas de execução e de monitoramento de processos gerados para cada instância de processo colocado em execução. Tais dados são conhecidos como *log* de eventos, registro de eventos, ou simplesmente *log*¹.

De acordo com Han e Kamber (2006), mineração de dados se refere à extração ou mineração de conhecimento de grandes quantidades de dados. Além disso, mineração de dados pode ser vista como uma fase de um processo maior chamado Descoberta de Conhecimento (KDD — *Knowledge Discovery from Data*). Nesse processo maior, dados provenientes de um contexto específico são coletados, transformados e organizados antes de sofrer a mineração. Depois da mineração, o resultado deve ser organizado em uma estrutura acessível à interpretação direta por humanos. Existem diferentes tarefas a serem tratadas em mineração de dados e para cada uma delas existem diferentes técnicas que podem ser aplicadas. Dentre essas técnicas, aquelas provenientes de inteligência computacional são comumente usadas (WANG; FU, 2006; STAHL; JORDANOV, 2012; CÔRTEZ; PORCARO; LIFSCHITZ, 2002). Técnicas de inteligência computacional têm se apresentado eficientes na resolução de tarefas de mineração de dados, pois possuem boa capacidade de lidar com dados provenientes de contextos não definidos, já que a geração de modelos rígidos e altamente precisos e restritos são impraticáveis, o que é frequentemente encontrado em dados provenientes de problemas do mundo real (ABONYI; FEIL; ABRAHAM, 2005). Assim,

¹ Neste trabalho, o termo “*log* de eventos” é usado preferencialmente para se referir a esses dados

técnicas de inteligência computacional que lidam com raciocínio aproximado (como lógica *fuzzy* e raciocínio probabilístico), raciocínio indutivo (como redes neurais e máquinas de vetores de suporte) e raciocínio evolutivo (como técnicas de computação evolutiva) são altamente adequadas para mineração de dados, visto que seu contexto envolve conhecimento empírico e comportamental, em que a informação é imprecisa, incompleta e imperfeita (ABONYI; FEIL; ABRAHAM, 2005). Além disso, técnicas de aprendizado de máquina são amplamente usadas em diversas tarefas de mineração, conseguindo ótimos resultados na resolução de problemas complexos. De acordo com o paradigma de aprendizado, tais técnicas podem ser classificadas como: aprendizado supervisionado (tais como: *k-nearest neighbor classification*, *Naïve Bayes*, e árvores de decisão); e aprendizado não supervisionado tais como: *k-means* e EM (*expectation maximization*). Algumas dessas técnicas são classificadas entre as melhores para mineração de dados (WU et al., 2008).

Com a junção das duas áreas — BPM e mineração de dados — estabeleceu-se um novo campo de estudo, chamado mineração de processos de negócio, ou apenas mineração de processos (AALST, 2011). Trata-se de aplicar tarefas de mineração de dados sobre dados provenientes do ciclo de vida de BPM. O objetivo é extrair conhecimento dos *logs* de eventos provenientes do trabalho realizado nas diferentes fases de um processo de negócio, buscando melhorar tal processo, por meio da descoberta de associações entre variáveis, padrões de comportamento ou desvios de comportamento (AALST, 2011). Busca-se, assim, encontrar informações inerentes relacionadas ao processo real. Diversas abordagens, técnicas e algoritmos foram propostos para resolver esse tipo de situação. Especificamente em relação a técnicas de inteligência computacional e de aprendizado de máquina, pesquisadores com trabalhos importantes dentro da área já propuseram, por exemplo, abordagens baseadas em computação evolutiva (VERGIDIS et al., 2007; MEDEIROS; WEIJTERS, 2005) e lógica *fuzzy* (GÜNTHER; AALST, 2007), obtendo resultados satisfatórios na descoberta de processo; já para técnicas baseadas em redes neurais, poucos trabalhos já foram apresentados, como a proposta de um método baseado em redes neurais que poderia ter bons resultados, mas ainda há pouca maturidade para sua aplicação na prática (COOK; WOLF, 1998).

1.1 Motivação e justificativa

Atualmente, há um crescimento explosivo de dados de uma ampla variedade de fontes de dados provenientes de sistemas em muitos domínios e aplicações (BOSE; MANS;

AALST, 2013), facilitando a criação de históricos de processos gerados por sistemas de informação (GANTZ; REINSEL, 2010). Por outro lado, existe uma maior necessidade por parte das organizações de melhorar seus processos em ambientes competitivos e de rápida evolução (AALST, 2011). Mineração de processos foi introduzida há aproximadamente 20 anos, e recentemente tem sido consideravelmente aprimorada. De acordo com Aalst (2013), o incremento da disponibilidade e o crescimento do volume de dados de eventos sugere que o importante crescimento de mineração de processos será contínuo nos próximos anos.

De acordo com Tiwari e Turner (2008), o principal motivador para o desenvolvimento e o crescente uso de técnicas de mineração de processos é a necessidade das organizações aprender sobre seus próprios processos. Como mineração de processos deriva principalmente da mineração de dados, essa área foi aprimorada e adaptada para trabalhar sobre registros de dados que contêm dados de execução de processos. Esses registros são os *log* de eventos de execução, que são tipicamente armazenados em sistemas de BPM, embora eles também possam ser acessados por meio de outros sistemas relacionados ao processo. Além disso, técnicas e algoritmos personalizados foram desenvolvidos especificamente para atender às necessidades de especialistas em mineração de processos.

No trabalho de Tiwari e Turner (2008), alguns problemas característicos da área de mineração de processos são categorizados, e técnicas para suas resoluções são apresentadas, incluindo algumas técnicas de inteligência computacional e de aprendizado de máquina. O resultado desse trabalho indicou a seguinte proporção de aplicação das principais abordagens em mineração de processos, para pouco mais de 50 trabalhos analisados: técnicas e algoritmos personalizados (28%), técnicas baseadas em redes de Petri (21%), técnicas gerais de mineração de dados (7%), cadeias de processos dirigida por eventos (5%), cadeias de Markov (3%), análise de agrupamentos (3%), algoritmos genéticos (3%), e redes neurais (2%). Porém, não são apresentados detalhes sobre o uso dessas técnicas, como possíveis vantagens ou desvantagens para cada caso. Além disso, destaca-se o baixo uso de técnicas de inteligência computacional e de aprendizado de máquina, tais como redes neurais e algoritmos genéticos (totalizando apenas quatro trabalhos). Pela análise realizada nesse trabalho, as técnicas de mineração de processos mais empregadas são as baseadas em heurísticas, que são “regras de ouro”, baseadas no conhecimento prévio a aproximadamente padrões comuns de negócio, e normalmente usadas em mineração de dados para resolução de problemas mais triviais.

Já Yue et al. (2011) descrevem abordagens de mineração de processos considerando apenas as propostas por pesquisadores representativos na área. Os principais são: (i) Wil van der Aalst – algoritmo α (alpha) e β (beta) para descobrir processos por meio de redes de Petri, e ferramentas Little Thumb, EMIT e ProM, que estão entre as mais usadas na área; (ii) J. E. Cook e A. L. Wolf – uso de máquinas de estado, redes neurais, e cadeias de Markov; (iii) R. Agrawal, D. Gunopulos e F. Leymann – uso de expressões booleanas para tratamento de ruído e estruturas paralelas; (iv) J. Herbst – métodos MergeSeq, SplitSeq e SplitPar usando grafos de tarefas estocásticas, para tratar tarefas duplicadas; (v) G. Schimm – algoritmo para lidar com modelos de *workflow* de estrutura hierárquica, e ferramenta Process Miner que gera modelos de processo por meio da construção de blocos; e (vi) J. Wainer – métodos para reescrever modelos. De fato, os trabalhos relacionados no estudo de Yue et al. (2011) derivam, em geral, de abordagens inicialmente desenvolvidas pelo pesquisador Wil van der Aalst, atualmente o principal pesquisador desta área, em termos de citações em seus trabalhos. Uma desvantagem comum em muitos desses trabalhos é a presença de valores atípicos e excepcionais (conhecidos como ruído) no resultado.

Segundo a definição apresentada por Aalst (2011), mineração de processos situa-se entre “aprendizado de máquina” e “mineração de dados”, por um lado, e “modelagem e análise de processos”, por outro lado. Busca-se, então, aproveitar conceitos e técnicas já existentes na área de aprendizado de máquina, assim como técnicas afins provenientes da área de inteligência computacional, a fim de melhorar os resultados atualmente obtidos em mineração de processos. Além disso, a complexidade dos dados usados em mineração de processos demanda técnicas cada vez mais poderosas para realizar essa mineração. Técnicas atuais já são capazes de lidar com *log* de eventos massivos. Porém, apesar dos avanços alcançados em mineração de processos, de acordo com Aalst (2011), novos desafios surgiram, tais como: (i) integração e melhoria da qualidade dos dados de eventos; (ii) tratamento de *logs* de eventos complexos e com características específicas; (iii) criação de valores de referência para estabelecer pontos de comparação; (iv) tratamento de desvios do conceito de processo; (v) melhoria na representação de resultados para a descoberta de processos; (vi) melhoria no equilíbrio entre a qualidade dos critérios como robustez, simplicidade, precisão e generalização; e (v) melhoria da usabilidade e na compreensão para os que não são especialistas.

A avaliação exploratória inicial realizada no contexto deste trabalho encontrou que já havia algumas abordagens usando técnicas de inteligência computacional e de

aprendizado de máquina para tentar atingir alguns dos desafios mencionados por Aalst (2011), conforme apresentado por Tiwari e Turner (2008) e Yue et al. (2011). Embora ainda em pequeno número, esses trabalhos se apresentam importantes, principalmente os que usam técnicas de aprendizado de máquina, uma vez que a própria definição de mineração de processos, de acordo com Aalst (2011), situa essa área entre mineração de dados e “aprendizado de máquina”, de um lado, e modelagem e análise de processos de negócio, de outro lado. Além disso, técnicas de inteligência computacional e de aprendizado de máquina têm sido muito usadas para tratar problemas diversos de mineração de dados de forma geral, devido a suas numerosas aplicações bem sucedidas na resolução de tarefas de mineração de dados (WU et al., 2008) e a forma de lidar com problemas de grande volume de dados em diversas áreas (CHEN; ZHANG, 2014).

Considerando esse contexto, uma avaliação do estado atual da área de mineração de processos, incluindo suas características e limitações, por meio da análise qualitativa e quantitativa via revisão sistemática da literatura, podem contribuir para o entendimento das lacunas e dos problemas atuais existentes dentro do contexto de aplicação de técnicas de inteligência computacional e de aprendizado de máquina nesse contexto, assim como estabelecer as bases para trabalhos futuros. As lições aprendidas com esse estudo podem representar uma importante contribuição para os pesquisadores e profissionais da área.

1.2 Objetivos

Considerando que técnicas de inteligência computacional e de aprendizado de máquina são amplamente aplicadas de forma satisfatória em tarefas de mineração de dados, em contextos gerais, seria esperado que elas também estivessem sendo amplamente aplicadas em mineração de processos, o que não tem sido demonstrado na literatura recente. Assim, o principal objetivo deste trabalho foi investigar o cenário geral de mineração de processos, enfatizando o uso de técnicas de inteligência computacional e de aprendizado de máquina. Para alcançar esse objetivo geral, os seguintes objetivos específicos foram buscados:

- Realizar uma análise quantitativa e qualitativa da aplicação de técnicas de inteligência computacional e de aprendizado de máquina no contexto de mineração de processos, em comparação a outras técnicas tradicionais não consideradas “inteligentes”.

- Analisar os possíveis motivos pelos quais essas técnicas de inteligência computacional e de aprendizado de máquina não estariam sendo amplamente usadas no contexto de mineração de processos, em comparação a outras técnicas tradicionais não consideradas “inteligentes”.

Busca-se, com a realização deste trabalho, oferecer como resultado a pesquisadores e profissionais da área de mineração de processos uma série de benefícios para seus trabalhos futuros, incluindo uma compilação de lições aprendidas. Trata-se de um trabalho de pesquisa interdisciplinar com o objetivo de contribuir com a área de mineração de processos por meio do oferecimento de uma visão ampla de vantagens e desvantagens da aplicação das técnicas mais modernas de inteligência computacional e de aprendizado de máquina.

1.3 Estrutura do documento

Esta dissertação está dividida em cinco capítulos, considerando esta introdução. Os demais capítulos estão organizados da seguinte forma:

- o capítulo 2 contextua o tema em estudo, apresentando os principais conceitos relacionados a gestão de processos de negócio, mineração de dados incluindo inteligência computacional e de aprendizado de máquina, e mineração de processos de negócio;
- o capítulo 3 expõe de forma resumida o método de pesquisa, incluindo o protocolo de revisão sistemática e protocolo de mapeamento sistemático, que foram seguidos durante a realização deste trabalho;
- o capítulo 4 apresenta uma análise qualitativa e quantitativa dos resultados alcançados com as revisões aprofundadas da literatura e com um exemplo de aplicação.
- o capítulo 5 apresenta a conclusão deste trabalho, com destaque às principais contribuições do estudo e as possibilidades de extensão do mesmo.

2 Fundamentação teórica

Este capítulo apresenta conceitos básicos envolvidos na realização deste trabalho, incluindo: gestão de processos de negócio, mineração de dados, técnicas de inteligência computacional e de aprendizado de máquina, e mineração de processos .

2.1 Gestão de processos de negócio

Gestão de Processos de Negócio (BPM) visa organizar a realização do trabalho nas organizações visando assegurar resultados consistentes e a aproveitar as oportunidades de melhoria. BPM não se preocupa com a realização de atividades individuais, mas apenas com a gestão de cadeias inteiras de eventos, atividades e decisões — chamadas de processos de negócio, buscando acrescentar valor às organizações e aos clientes (DUMAS et al., 2013).

Segundo Weske (2012), um processo de negócio consiste em um conjunto de tarefas realizadas em uma sequência específica para alcançar um objetivo comum. BPM apresenta-se como um conjunto de conceitos, abordagens e ferramentas que, com o apoio da Tecnologia da Informação (TI), podem transformar os processos das organizações. Com base em BPM, é possível atingir mais facilmente o alinhamento estratégico entre áreas de negócio e a área de TI, gerenciando e aprimorando as soluções tecnológicas a partir de processos que agregam valor significativo para as organizações. Assim, BPM pode ser vista como uma vantagem competitiva para as organizações (FANTINATO; SOUZA; TOLEDO, 2010).

De acordo com Aalst (2011), BPM pode ser considerada uma extensão de sistemas e abordagens clássicas de gestão de processos, incluindo *workflows*. Porém, BPM tem um alcance ainda maior, desde a automação e a análise de processos até a gestão de processos e o trabalho organizacional. Assim, BPM inclui métodos, técnicas e ferramentas para apoiar a concepção, a representação, a gestão e a análise de processos.

2.1.1 Ciclo de vida de BPM

O ciclo de vida de processos de negócio consiste em fases organizadas em uma estrutura sequencial que apresenta a dependência lógica entre elas. Segundo a visão de Weske (2012), o ciclo de vida de BPM é composto pelas seguintes fases (conforme

representado na figura 1). Busca-se realizar esse ciclo de vida de forma rápida devido aos ambientes dinâmicos das organizações em que BPM é usada.

- **Projeto e análise:** primeiramente, ocorre o projeto do processo de negócio, que consiste no levantamento de como o trabalho é realizado na organização e na modelagem do processo. Durante a modelagem do processo de negócio, as práticas organizacionais são definidas, estipuladas e formalizadas em modelos de processos, isto é, artefatos gráficos que, por meio de alguma notação (conforme apresentadas na seção 2.1.2), definem o fluxo das tarefas envolvidas no processo. Por fim, durante a análise do processo, o modelo de processo deve ser validado, simulado e verificado para garantir que ele represente realmente o processo correto desejado.
- **Configuração:** o modelo de processo de negócio definido é usado para realizar a implementação do processo. Seleciona-se o sistema a ser usado para apoiar a implementação do processo, que podem incluir diferentes plataformas e ferramentas, ou até mesmo uma solução individual e específica para cada caso. Por fim, o processo de negócio implementado deve ser testado para posteriormente ser implantado no ambiente de produção.
- **Execução:** a realização de processos é acompanhada e monitorizada em tempo de execução. Além disso, deve ocorrer também a manutenção dos processos em execução. Os processos em execução podem ser auxiliados por meio de previsões e recomendações baseadas em modelos originários de dados históricos.
- **Avaliação:** os dados gerados durante a execução de processos são avaliados para verificar se os objetivos organizacionais estão sendo atingidos de maneira satisfatória. Assim, eventuais ineficiências no processo podem ser identificadas possibilitando que aprimoramentos nos processos possam ocorrer. É aqui quando o histórico de execução é analisado por meio das técnicas de mineração de processos. Isso pode levar à remodelagem de processos, de forma a reiniciar o ciclo de vida novamente.

Há outras propostas para definir e representar o ciclo de vida de BPM com algumas variações, embora todas sejam bastante similares. Aalst (2011), por exemplo, descreve o ciclo de vida de BPM em relação ao apoio à fase de operação, como ilustrado na figura 2. Nesse ciclo de vida, as fases de (re)Projeto, Configuração/Implementação, Execução/Monitoramento, são similares às descritas anteriormente. No entanto, de acordo com Aalst (2011), uma nova fase é considerada — a fase de Ajuste. Não se faz nenhum

redesenho nem criação de software, apenas os critérios já predefinidos são usados para realizar algumas adaptações necessárias. Já na fase Diagnóstico/Requisitos, os processos operacionais são analisados e monitorados para identificar problemas, o que pode ser usado nas novas iterações para tomar decisões sobre o redesenho de processo.

Figura 1 – Ciclo de vida de BPM – proposta 1

Fonte: Traduzido de Weske (2012)

Figura 2 – Ciclo de vida de BPM – proposta 2

Fonte: Traduzido de Aalst (2011)

2.1.2 Modelagem e implementação de processos de negócio

Um modelo de processo de negócio é uma representação gráfica do processo que descreve as dependências entre atividades que precisam ser executadas coletivamente para a realização de um objetivo específico (WESKE, 2012). Modelagem de processos de negócio tem, por sua vez, o objetivo de oferecer um instrumento formal para uma descrição precisa de processos de negócio usando uma linguagem com sintaxe e semântica bem definidas (DUMAS; AALST; HOFSTEDE, 2005). Diversas linguagens de modelagem e especificação foram propostas para ser usadas em BPM, sendo *Business Process Model and Notation* (BPMN) atualmente a mais usada (GROUP, 2011). BPMN é destinada ao uso comum para ambas as áreas de negócio e de TI de uma organização, visto que suas propriedades visuais facilitam sua compreensão. Os modelos de processo são representados via um fluxograma, com flexibilidade e legibilidade. No entanto, outras linguagens também têm sido usadas para modelar processos de negócio, tais como: diagramas de atividades da Linguagem de Modelagem Unificada (UML – *Unified Modeling Language*), que permitem representar um sistema de forma padronizada (RUSSELL et al., 2006; ZHANG; DUAN, 2008); e Cadeias de Processos Dirigidos a Eventos (EPC – *Event-driven Process Chains*), que oferece uma notação gráfica dedicada a processos (DUMAS; AALST; HOFSTEDE, 2005).

No contexto de mineração de processos, redes de Petri é a linguagem de modelagem mais usada (TIWARI; TURNER, 2008). De uma forma geral, uma rede de Petri é um grafo bipartido que representa tarefas por meio de transições (retângulos) e estados distribuídos por um conjunto de lugares (círculos) (WESKE, 2012). Segundo Dumas, Aalst e Hofstede (2005), redes de Petri podem ser vistas de várias formas: como uma linguagem e um formalismo de modelagem; como um método apoiado por um conjunto de ferramentas desenvolvidas por um fabricante comercial e por instituições acadêmicas; como uma teoria de grande tradição e centenas de teoremas; como uma disciplina científica; e, até mesmo, como uma filosofia. A diferença entre redes de Petri do ponto de vista de linguagem visual, ou de teoria matemática ou ainda de linguagem formal é importante para compará-la e combiná-la a outras abordagens. Quando usadas como linguagem de modelagem, redes de Petri são uma linguagem visual; quando usadas sob semânticas de alguma outra linguagem, elas são consideradas uma teoria matemática; quando usadas para representar uma simulação, de forma precisa e ilustrativa, e para a análise de informação de controle do fluxo de componentes recorrentes e de atividades sincronizadas, elas são consideradas

uma linguagem formal (DUMAS; AALST; HOFSTEDE, 2005). Esta última é a definição normalmente usada para mineração de processos (AALST, 2011).

Outras classes de redes de Petri também podem ser usadas em mineração de processos, incluindo aquelas definidas por alguma restrição estrutural, com alguma interpretação semântica, tais como: máquinas de estado; *marked graph*; *free-choice nets*; redes de Petri coloridas; e redes de *workflow*, conhecidas como WF-nets (DUMAS; AALST; HOFSTEDE, 2005). WF-nets, por exemplo, bastante usadas em mineração de processos, têm uma sintaxe simples e uma notação dedicada ao início e ao fim de um processo (DUMAS et al., 2013). WF-nets também incorporam locais que representam condições e *tokens* para representar o ponto na execução de processo. Tais *tokens* mudam de posição segundo os dados da aplicação. WF-nets podem ser estruturas hierárquicas; assim, a estrutura interna de uma atividade complexa é realizada por meio de outra WF-net dedicada (WESKE, 2012).

A fim de realizar a gestão e a integração de possíveis processos de negócio, considerando o ponto de vista técnico, diferentes tecnologias têm sido propostas. Tecnologias de *middleware* foram propostos e usados, tais como CORBA, DCOM e Java-RMI (ALONSO et al., 2004); os quais puderam ser adequadamente usadas em contextos intra-organizacionais. À medida que a necessidade de interoperabilidade evolui para as cooperações interorganizacionais, as soluções existentes deixaram de atender seus objetivos. Esse problema foi minimizado com o surgimento da Arquitetura Orientada a Serviços (SOA — *Service-oriented Architecture*) e da tecnologia de serviços web, que oferecem novas perspectivas para a necessidade de interoperabilidade de BPM. SOA e serviços web proporcionam, por exemplo, a composição de serviços web por meio da linguagem WS-BPEL (*Web Services Business Process Execution Language*) (JORDAN; EVDEMON, 2007), que integrada à linguagem WSDL (*Web Service Description Language*) (JORDAN; EVDEMON, 2007), permite a execução de processos de negócio. WS-BPEL tem sido considerada a linguagem padrão para a implementação de processos, via composição de serviços web descritos usando WSDL.

2.1.3 Logs de eventos de processos de negócio

Um *log* de eventos é basicamente uma tabela contendo os registros dos eventos que relatam a execução das atividades do processo de negócio (DUMAS; AALST; HOFSTEDE, 2005). Por exemplo, um *log* pode registrar informações sobre quando um usuário faz *login*

ou quando um erro é encontrado. Dependendo do tipo de processo a ser executado, um *log* podem registrar informações referentes a: eventos de aplicativo ou de programa (ex.: um erro ou um aviso); eventos relacionados a segurança (ex.: auditorias a usuários ou registros de acesso); eventos de instalação; eventos do sistema, entre outros. Especificamente para este trabalho, são considerados os *logs* gerados durante a fase de execução e monitoramento no ciclo de vida de BPM, como dados para conduzir as atividades de mineração de processos. É assumida aqui a estrutura de *log* de eventos apresentada por Dumas, Aalst e Hofstede (2005) e Aalst (2011), que consideram possível registrar os eventos de forma que: (i) cada evento refere-se a uma *atividade* (ex.: uma etapa bem definida dentro de processo); (ii) cada evento refere-se a um *caso* (ex.: uma instância de processo); (iii) cada evento pode ter um *usuário*, também referido como *originador* (ou seja, pessoa que executa ou inicializa a atividade); (iv) cada evento tem um *timestamp* (ou seja, o horário em que foi executado); e (v) todos os eventos dentro do *log* estão totalmente ordenados.

Na tabela 1, ilustra-se um fragmento curto de um *log* de eventos gerado por um sistema de pago de compensações. Tal exemplo é usado posteriormente na seção 2.2.3, como parte de um exemplo de mineração de processos. Cada linha dentro do arquivo que armazena o *log* de eventos representa informações sobre um evento: identificador do caso; data e hora de ocorrência; atividade associada ao evento; e usuário da atividade. O identificador de caso agrupa os eventos executados em uma mesma instância de processo. Por exemplo, o caso 1 possui cinco eventos associados. O primeiro evento do caso 1 representa a execução da atividade “Registrar requerimento”, executada no dia “30-12-2010” às “11 h e 2 min”, por “Pete”. Cada evento é único, ou seja, uma mesma pessoa não poderia executar duas atividades pertencentes ao mesmo caso, no exatamente mesmo *timestamp*. Outros tipos de informações podem ser registrados para cada evento, a depender do sistema e do interesse para o negócio. Por exemplo: hora de início e fim da atividade, estado da atividade, recursos usados, custo de execução, etc.

2.2 Mineração de dados, inteligência computacional e aprendizado de máquina

Organizações em geral têm se apresentado eficientes em capturar, organizar e armazenar dados. Contudo, ainda há dificuldade em transformar essa grande quantidade de dados em conhecimento útil. Nesse contexto, mineração de dados é usada para encontrar

Tabela 1 – Fragmento de um arquivo de *logs* de eventos

IdCaso	<i>Timestamp</i>	Atividade	Usuário
1	30-12-2010:11.02	Registrar requerimento	Pete
1	31-12-2010:10.06	Analisar o detalhe	Sue
1	05-01-2011:15.12	Verificar bilhete	Mike
1	06-01-2011:11.18	Decidir	Sara
1	06-01-2011:11.18	Rejeitar requerimento	Pete
2	30-12-2010:11.32	Registrar requerimento	Mike
2	30-12-2010:12.12	Verificar bilhete	Mike
2	30-12-2010:14.16	Analisar informalmente	Pete
2	05-01-2011:11.22	Verificar bilhete	Sara
2	08-01-2011:12.05	Pagar compensação	Ellen
3	30-12-2010:14.32	Registrar requerimento	Pete
3	30-12-2010:15.06	Analisar informalmente	Mike
3	30-12-2010:16.34	Verificar bilhete	Ellen
3	06-01-2011:09.18	Decidir	Sara
3	06-01-2011:12.18	Reiniciar requerimento	Sara
3	06-01-2011:13.06	Analisar o detalhe	Sean
3	08-01-2011:11.43	Verificar bilhete	Pete
3	09-01-2011:09.55	Decidir	Sara
3	09-01-2011:09.55	Pagar compensação	Ellen
4	06-01-2011:15.02	Registrar requerimento	Pete
4	07-01-2011:12.06	Verificar bilhete	Mike
4	08-01-2011:14.43	Analisar a profundidade	Sean
4	09-01-2011:12.02	Decidir	Sara
4	12-01-2011:15.44	Rejeitar requerimento	Ellen
5	06-01-2011:09.02	Registrar requerimento	Ellen
5	07-01-2011:10.16	Analisar informalmente	Mike
5	08-01-2011:11.22	Verificar bilhete	Pete
5	10-01-2011:13.28	Decidir	Sara
5	11-01-2011:16.18	Reiniciar requerimento	Sara
5	14-01-2011:14.33	Verificar bilhete	Ellen
5	16-01-2011:15.50	Analisar informalmente	Mike
5	19-01-2011:11.18	Decidir	Sara
5	20-01-2011:12.48	Reiniciar requerimento	Sara
5	21-01-2011:09.06	Analisar informalmente	Sue
5	21-01-2011:11.34	Verificar bilhete	Pete
5	23-01-2011:13.12	Decidir	Sara
5	24-01-2011:14.56	Rejeitar requerimento	Mike
6	06-01-2011:15.02	Registrar requerimento	Mike
6	06-01-2011:16.06	Analisar informalmente	Ellen
6	07-01-2011:16.22	Verificar bilhete	Mike
6	07-01-2011:16.52	Decidir	Sara
6	16-01-2011:11.47	Pagar compensação	Mike

Fonte: Adaptação de Aalst (2011)

conhecimento útil, e muitas vezes inesperado, a partir da análise dos dados produzidos nos mais diversos tipos de sistemas e processos, por meio da revelação de padrões de comportamento que podem servir como base para a tomada de decisão com alto nível de valor estratégico (NAZEM; SHIN, 2002). Dessa forma, mineração de dados é uma atividade para explorar e analisar dados, de forma automática ou semiautomática, visando descobrir padrões relevantes, ocultos à base de dados.

Existem várias aplicações para mineração de dados, em diversas áreas. Alguns exemplos são: *(i)* agrupamento de clientes por perfis para gestão do relacionamento com os clientes; *(ii)* predição de séries temporais hidrológicas a partir de redes neurais (TAMPELINI et al., 2009); e *(iii)* classificação (início de um módulo de reconhecimento de padrões) na Língua Brasileira de Sinais (LIBRAS) (PERES et al., 2010). Além disso, há diversos exemplos de aplicação de mineração de dados em áreas tais como *marketing*, vendas, manufatura, finanças, saúde e energia.

Um problema associado à mineração de dados, e a KDD de uma forma geral, é que além da possibilidade de garimpar relacionamentos inúteis, o número de correlações possíveis de serem obtidas tende a ser muito grande, o que impede a análise exaustiva de cada uma (THOMÉ, 2002). Apesar de existirem diversas ferramentas para auxiliar na execução das técnicas de mineração, os resultados ainda precisam de análise humana. Porém, ainda assim, a mineração de dados contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos especialistas concentrarem esforços apenas em partes mais significativa dos dados (CAMILO; SILVA, 2009).

Mineração de dados é comumente classificada considerando sua capacidade em realizar determinadas tarefas (GRZYMALA-BUSSE, 2008). Para cada tarefa de mineração de dados, diferentes técnicas podem ser aplicadas (HAN; KAMBER, 2006; WU et al., 2008), incluindo aquelas das áreas de inteligência computacional e de aprendizado de máquina (WANG; FU, 2006). A seguir são apresentadas as principais e mais comuns tarefas dentro do contexto deste trabalho. Enquanto que, as técnicas mais usadas para cada uma dessas tarefas são descritas nas seções subsequentes.

2.2.1 Tarefas de mineração de dados

As tarefas a serem executadas em mineração de dados visam a construção de modelos matemáticos ou heurísticos que permitem a definição de mapeamentos e regras

capazes de revelar conhecimento novo sobre um domínio. Essas tarefas podem prever tendências e comportamentos, permitindo tomar decisões com base em conhecimento acumulado, e frequentemente desprezado, contido nos próprios bancos de dados. Assim, durante a resolução das tarefas de mineração de dados, padrões e relacionamentos existentes, porém desconhecidos, podem ser identificados, levando à geração de hipóteses úteis ao usuário. A seguir, são apresentadas descrições das seguintes principais tarefas existentes para mineração de dados segundo as definições de (HAN; KAMBER, 2006): *(i)* predição categórica (ou classificação), predição numérica (tipo “regressão”), e predição numérica (tipo “identificação de tendências”); *(ii)* análise de agrupamentos; *(iii)* descoberta de padrões frequentes, regras de associação e correlação; *(iv)* análise de *outliers* e ruído; *(v)* análise de evolução; e *(v)* análise de dados estruturados.

- **Predição:** trata-se da construção de um modelo de predição com base em exemplos (contidos em uma base histórica). A construção do modelo, segundo essa tarefa, pressupõe o conhecimento prévio das possíveis classes (ou rótulos) e a correta classificação dos exemplos usados na modelagem. Existem várias aplicações para esse tipo de abordagem; em análise de risco, por exemplo, o objetivo pode ser o de classificar um potencial cliente entre as classes de “excelente”, “bom” ou “mau” pagador. Segundo Han e Kamber (2006), há três tipos de predição:
 - **Predição categórica (ou classificação):** consiste em buscar uma função que classifique uma determinada ocorrência em uma categoria dentre um conjunto finito e predefinido de classes. As principais técnicas que podem ser usadas para esse tipo de predição são árvores de decisão e redes neurais.
 - **Predição numérica (do tipo “regressão”):** consiste em buscar uma função que represente, de forma aproximada, o comportamento apresentado pelo fenômeno em estudo. Esse modelo estatístico relaciona uma ou mais variáveis independentes (ou variáveis de predição) e variáveis dependentes (ou variáveis de resposta), as quais têm valores contínuos. A variável de resposta representa o que se quer prever. A mais conhecida é a regressão linear; por exemplo, uma reta que minimiza o erro médio entre todos os valores considerados. Entretanto, a regressão não linear é mais útil em problemas reais. Uma das principais técnicas que podem ser usadas para esse tipo de predição são as redes neurais.

- **Predição numérica (do tipo “identificação de tendências”)**: envolve uma componente temporal, isto é, representa problemas em que o interesse é prever o valor futuro de uma determinada variável (mono-variável) ou em valores anteriores da variável de interesse e de outras variáveis (multi-variável). As principais técnicas que podem ser usadas nesse tipo de predição são métodos estatísticos, máquinas de vetores de suporte e redes neurais.
- **Análise de agrupamentos**: consiste em buscar similaridades entre dados para definir um conjunto finito de grupos de dados, organizando os dados em agrupamentos de forma a maximizar a similaridade intra-agrupamentos e minimizar a similaridade interagrupamentos. Consiste também em determinar um conjunto não conhecido de classes para os dados de forma também a descrevê-los. Um exemplo de aplicação para essa tarefa é descobrir grupos homogêneos de clientes, que pode ajudar na definição de uma estratégia de *marketing* a ser adotada. As principais técnicas que podem ser usadas para essa tarefa são os métodos estatísticos e as redes neurais.
- **Descoberta de padrões frequentes, regras de associação e regras de correlação**: consiste em identificar ocorrências frequentes que possam estar direta ou indiretamente associados (ou correlacionados) na ocorrência de eventos ou transações. Essa estratégia é geralmente usada em aplicações em que se busca identificar itens que possam ser colocados juntos em um mesmo pacote de negociação. Por exemplo, a constatação de que vendas de cerveja e de fraldas descartáveis apresentam um comportamento ascendente às sextas-feiras pode levar à construção de uma hipótese que associe ambos os itens a um tipo especial de cliente. Regras de associação também podem ser usadas para avaliar a existência de algum tipo de relação temporal entre os itens constantes em uma base de dados. Podem ser usadas para essa tarefa, por exemplo, métodos estatísticos.
- **Análise de *outliers* e de ruído**: uma base de dados pode conter objetos que não obedecem ao comportamento geral do modelo dos dados, os quais são chamados *outliers*. Em muitos casos, a mineração de dados descarta os *outliers* como ruído ou exceções. No entanto, para aplicações como detecção de fraude, esses eventos raros podem ser mais importantes do que as ocorrências regulares. Os *outliers* podem ser detectados usando, por exemplo: (i) métodos estatísticos que assumem um modelo de distribuição de probabilidades para os dados; (ii) modelos de distância em que os

objetos que são substancialmente distantes de qualquer outro grupo são considerados *outliers*; ou métodos baseados em desvio que examina deferências nas principais características dos objetos no grupo.

- **Análise de evolução:** descreve modelos regulares e tendências de objetos, os quais mudam de comportamento no tempo. Apesar desse tipo de análise poder ser incluído em tarefas como “Descoberta de padrões frequentes, regras de associação e regras de correlação”, “Predição (classificação, regressão ou identificação de tendências)” ou “Análise de agrupamentos” de dados quando relacionados ao tempo, existem características próprias que os diferenciam, incluindo: análise de dados de séries temporais; união de padrões em sequências ou periodicidades; e análise de dados baseada em similaridade. As principais técnicas que podem ser usadas para essa tarefa são: métodos para lidar com *sliding windows model*, mineração de padrões frequentes em fluxos de dados, mineração de padrões sequenciais em bases de dados transacionais, e algoritmo GSP (baseado em geração-e-teste para um novo candidato).
- **Análise de dados estruturados:** mineração de dados foi inicialmente concebida para ser usada com base em repositórios estruturados de dados, tais como arquivos, banco de dados, e data *warehouse*. Porém, atualmente, os dados são representados por diversos formatos não estruturados, tais como: espacial e temporal, multimídia, web, entre outros. Assim, existe uma necessidade crescente de mineração de dados nesses tipos de dados não estruturados. As principais técnicas aplicáveis a essa tarefa são: mineração de sub-grafos frequentes e mineração baseada em restrições. Alguns exemplos específicos são apresentados a seguir:
 - **Grafos:** usados na modelagem e na mineração de estruturas complexas tais como: circuitos, imagens, proteínas, redes biológicas, redes sociais, processos, documentos XML, etc. Um tipo de subestrutura frequente em grafos são os padrões básicos, que podem ser descobertos em uma coleção de grafos, como cliques ou componentes conexas. Tais padrões são úteis para caracterizar conjuntos de grafos, discriminar diferentes grupos de grafos, classificar e agrupar grafos, e buscar similaridades em bases de grafos.
 - **Relacionamentos:** baseados na análise de redes sociais, em que é representado o relacionamento entre as entidades envolvidas (que é similar a uma estrutura de grafos). As redes sociais são conjuntos de dados heterogêneos e relacionais

representados por grafos. Elas não estão necessariamente em um contexto social, podendo estar relacionadas especificamente ao contexto de tecnologia, negócios, economia, entre outros.

2.2.2 Técnicas de inteligência computacional e de aprendizado de máquina

Atualmente, não existe consenso na literatura para definir tanto a área de inteligência computacional quanto a área de aprendizado de máquina. Para este trabalho, foram adotadas algumas das premissas dadas pelos principais autores e instituições relacionados a esse contexto, conforme apresentadas a seguir.

Inteligência computacional é um ramo do estudo da inteligência artificial. A comunidade internacional de líderes profissionais e pesquisadores em inteligência computacional “IEEE Computational Intelligence Society” sinala que a inteligência computacional visa o uso de aprendizado, adaptativo ou de computação evolutiva para criar software que é, em algum sentido, inteligente (IEEE, 2008). Uma coleção de paradigmas computacionais com inspiração biológica e linguística, em que a teoria, o modelo, a implementação e o desenvolvimento de redes neurais (incluindo também máquinas de vetores de suporte, por semelhança), algoritmos evolutivos, sistemas baseados em lógica *fuzzy* e sistemas inteligentes híbridos, são considerados dentro esta área (ONG; LIM; CHEN, 2010). Abordagens de inteligência computacional têm sido usadas com sucesso para resolver problemas em diferentes áreas, tais como: negócios, gestão, diagnóstico médico, engenharia, entre outras. Recentemente, inteligência computacional vem sendo usada em tarefas de auxílio em sistemas de tomada de decisão para selecionar ações em tempo real, reduzindo a sobrecarga de informação, permitindo a atualização dos dados e lidando com incerteza em problemas de decisão (PHILLIPS-WREN; ICHALKARANJE; JAIN, 2008).

Em relação à área de aprendizado de máquina, uma definição é apresentada por Tom Mitchell em seu livro “Machine Learning” (MITCHELL, 1997). Mitchell define que um programa de computador aprende se ele é capaz de melhorar seu desempenho em uma determinada tarefa, sob alguma medida de avaliação, a partir de experiências passadas. Alternativamente, Faceli et al. (2011) definem aprendizado de máquina como um processo de indução de uma hipótese (ou aproximação de uma função) a partir da experiência passada. Ambas as definições tem como base um tipo de raciocínio indutivo. No raciocínio

indutivo, entende-se que as explicações para os fenômenos surgem a partir da observação de fatos e, assim, como definido por Russell e Norving (2004), uma tarefa de indução pode ser definida da seguinte forma: “Dada uma coleção de exemplos de uma função F , deve-se retornar uma função H que se aproxime de F .” Nessa definição, a função H é uma hipótese e a função F é desconhecida. Nesse sentido, pode-se afirmar que um processo algorítmico capaz de processar um conjunto de observações e formular uma regra geral que as explique pode ser considerado uma técnica de aprendizado de máquina. Isso torna muito complexa a tarefa de enumerar as técnicas que implementam o paradigma de aprendizado de máquina.

De forma mais geral, as técnicas de aprendizado de máquina podem ser classificadas em dois tipos de estratégia de aprendizado: aprendizado supervisionado e aprendizado não supervisionado (HAYKIN, 2009), tal como definido a seguir:

- **Aprendizado de máquina supervisionado:** estratégia para inferir uma função a partir de dados de treinamento rotulados. Os dados de treinamento consistem de um conjunto de exemplos de treinamento. Cada exemplo é um par constituído por um objeto de entrada (tipicamente um vetor) e um valor de saída desejado (também chamado de sinal de controle). O algoritmo analisa os dados de treinamento e produz uma função inferida, que pode ser usada para o mapeamento de novos exemplos. Um cenário ideal permitiria que o algoritmo determinasse corretamente os rótulos de classe para instâncias ocultas. Isso exige que o algoritmo de aprendizado possa generalizar a partir dos dados de treinamento para situações ocultas em uma forma “razoável”. Exemplos de técnicas usadas para tratar o aprendizado de máquina supervisionado são: redes neurais; máquinas de vetores de suporte; árvores de decisão; classificador Naïve Bayes; e algoritmo *K-Nearest Neighbour* (KNN).
- **Aprendizado de máquina não supervisionado:** estratégia para encontrar uma estrutura oculta em dados não rotulados. Considerando que os exemplos dados para o algoritmo não são rotulados, não há nenhum sinal de erro ou recompensa para avaliar uma solução em potencial. Isso distingue o aprendizado não supervisionado do aprendizado supervisionado. Exemplos de técnicas usadas para tratar o aprendizado de máquina não supervisionado são: *Self-Organizing Map* (SOM) – um tipo de redes neurais; *k-means*; *Partitioning Around Medoids* (PAM); Agrupamento hierárquico; e algoritmo *Expectation-Maximization* (EM).

Para fins de delimitação neste trabalho, foram consideradas técnicas de aprendizado de máquina apenas aquelas que, por princípio, possuem estratégias de exploração do espaço de hipóteses baseadas no uso da informação presente diretamente no conjunto de observações (ou seja, no conjunto de dados). Por exemplo, a técnica de aprendizado não supervisionado *k-means* implementa análise das relações existentes entre os dados de um conjunto de dados (vetores) para inferir sua hipótese, que é representada na mesma forma dos dados — um vetor protótipo. Já um algoritmo genético (técnica estudada em inteligência computacional), embora possa ser usado para resolver uma tarefa de aprendizado de máquina, tem suas estratégias baseadas na otimização de uma função (projetada para avaliar uma hipótese sob um conjunto de observações no caso da resolução de uma tarefa de aprendizado de máquina), por meio de operadores probabilísticos independentes das relações existentes entre as observações. Assim, entende-se que a primeira, *k-means*, implementa aprendizado de máquina de forma direta, enquanto a segunda técnica, algoritmos genéticos, o implementa de forma indireta, não sendo, portanto, considerada uma técnica de aprendizado de máquina no escopo deste trabalho.

Sendo assim, considerando as definições apresentadas para as áreas de inteligência computacional e de aprendizado de máquina, a tabela 2 resume a taxonomia que foi estabelecida para as técnicas consideradas neste trabalho. Além de considerar as definições apresentadas para cada área, apenas as técnicas que foram identificadas em pelo menos um dos trabalhos analisados, presentes na literatura, foram incluídas neste trabalho. A seguir, cada uma das onze técnicas de inteligência computacional e/ou aprendizado de máquina consideradas são descritas.

- **Computação evolutiva:** propõe um paradigma computacional inspirado na seleção natural de Darwin, de 1859. Compreende um conjunto de técnicas de busca e otimização, nas quais cria-se uma população de indivíduos que vão reproduzir e competir pela sobrevivência. Tais indivíduos representam a solução para o problema sob estudo. Os melhores sobrevivem e transferem suas características a novas gerações. As técnicas atualmente incluem: programação evolutiva, estratégias evolutivas, algoritmos genéticos e programação genética (POZO et al., 2015).
- **Lógica *fuzzy*:** uma forma de lógica multivalorada em que os valores “verdade” de variáveis podem ser qualquer número real entre 0 e 1, em contraposição à lógica booleana em que os valores “verdade” de variáveis só podem ser 0 ou 1. A lógica *fuzzy*

Tabela 2 – Taxonomia de técnicas de inteligência computacional e de aprendizado de máquina usada neste trabalho

Técnica / Abordagem	Inteligência computacional	Aprendizado de máquina	
		supervisionado	não supervisionado
Computação evolutiva	X		
Lógica <i>fuzzy</i>	X		
Máquinas de vetores de su- porte	X	X	
Redes neurais	X	X	
SOM	X		X
Árvores de decisão		X	
Classificador Naïve Bayes		X	
k-Vizinho mais próximo		X	
Agrupamento hierárquico			X
Algoritmo EM			X
<i>K-means</i>			X
PAM			X

Fonte: Ana Rocío Cárdenas Maita, 2015

lida com o conceito de verdade parcial, em que o valor “verdade” pode variar entre completamente verdadeira a completamente falsa. Além disso, quando as variáveis linguísticas são usadas, esses graus podem ser geridos por funções especiais.

- **Redes neurais e máquinas de vetores de suporte (SVM — *Support-Vector Machines*):** redes neurais e máquinas de vetores de suporte representam juntas um poderoso conjunto de técnicas amplamente usado na resolução de tarefas de predição e análise de agrupamentos para mineração de dados. Ambas usam raciocínio indutivo, e são similares tanto em relação à forma de explorar um problema de mineração de dados quanto em relação à forma de fornecer a solução, sendo consideradas equivalentes em suas formas mais básicas (CRISTIANINI; SHAW-TAYLOR, 2000). Máquinas de vetores de suporte, no entanto, apresentam características adicionais no modelo de otimização da solução, já que elas consideram também a minimização do risco estrutural do modelo, além da minimização do risco empírico (VAPNIK, 2013). Ambas as técnicas apresentam boa capacidade para realizar mapeamentos não lineares complexos, comuns em contextos de mineração de dados, e por isso possuem potencial para alcançar predições com alto grau de acuidade. Elas se constituem como modelos a partir dos quais é possível extrair conhecimento, o que é útil em ambientes de apoio à tomada de decisão (CÔRTEZ; PORCARO; LIFSCHITZ, 2002). Ambas se destacam dentre as técnicas de inteligência computacional (STAHL; JORDANOV, 2012;

CÔRTEZ; PORCARO; LIFSCHITZ, 2002). Máquinas de vetores de suporte, por exemplo, foram incluídas entre os “principais 10” algoritmos em mineração de dados de acordo com a “6th IEEE International Conferência on Data Mining – ICDM 2006” (WU et al., 2008).

- **Mapas Auto-organizados (SOM – *Self Organized Map*):** é uma rede neural não supervisionada, frequentemente usada em tarefas de agrupamento e visualização de dados. O objetivo é encontrar um conjunto de valores de referência e associar cada objeto do conjunto de dados ao vetor de referência mais próximo. O algoritmo depende da inicialização dos vetores de referência. O resultado consiste em um conjunto de vetores de referência que definem implicitamente os grupos (KOHONEN, 2001).
- **Árvores de decisão:** funcionam como um fluxograma em forma de árvore, em que cada nó (não folha) indica um teste feito sobre um valor. As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence.
- **Classificador Naïve Bayes:** esse algoritmo computa a probabilidade de um documento pertencer a uma determinada classe a partir da probabilidade *a priori* de um documento ser dessa classe e das probabilidades condicionais de cada termo ocorrer em um documento da mesma classe. Seu objetivo é encontrar a melhor classe para um documento, maximizando a probabilidade a posteriori conforme a uma equação definida.
- **K-Vizinho mais próximo (kNN – *k-Nearest Neighbor*):** esse algoritmo armazena os dados de treinamento e quando um novo objeto é submetido para classificação, o algoritmo procura os k registros mais próximos (medida de distância) desse novo registro. O novo registro é classificado na classe mais comum entre todos os k registros mais próximos.
- **Abordagens de agrupamento hierárquico:** estes métodos constroem recursivamente uma organização hierárquica de grupos, onde a cada nível de abstração maior, são também maiores as diferenças entre elementos contidos em cada grupo. Fraley e Raftery (1998) sugerem dividir estes métodos em dois grupos principais: agrupamento hierárquico aglomerativo, onde grupos aninhados são organizados como uma árvore hierárquica, e agrupamento hierárquico divisivo, busca dividir o conjunto

de dados em um vários grupos distintos entre si, maximizando as dissimilaridades dos diferentes grupos.

- **Algoritmo EM (*Expectation-Maximization*)**: pode ser aplicado em situações em que se deseja estimar um conjunto de parâmetros que descreve uma distribuição de probabilidade. O algoritmo se baseia em estimar parâmetros de máxima verossimilhança para problemas em que os estados das variáveis não foram observados.
- **K-means**: método de agrupamento que objetiva particionar n observações dentre k grupos em que cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um diagrama de Voronoi. (HARTIGAN; WONG, 1979).
- **Particionamento ao Redor de Medóides (*Partitioning Around Medoids*)**: também chamado de algoritmo K-Medóides, foi desenvolvido como uma variação do K-Means. A diferença básica entre esses dois métodos está na forma como essas técnicas encontram o ponto de referência, K-Means realiza o cálculo de um centroide como ponto de referência para a formação dos agrupamentos, enquanto a técnica de K-Medoids busca um objeto, na posição mais cêntrica possível como ponto de referência.

2.2.3 Técnicas tradicionais para mineração de dados

Especificamente no contexto deste trabalho, as técnicas usadas em mineração de dados, em geral, ou em mineração de processos, de forma específica, mas que não são diretamente relacionadas a inteligência computacional ou a aprendizado de máquina, tal como definido na seção 2.2.2 (ou seja, técnicas que não são consideradas “inteligentes”), são chamadas de técnicas “tradicionais”. A seguir, as técnicas tradicionais que foram consideradas neste trabalhos são apresentadas e descritas brevemente. Apenas as técnicas tradicionais que foram identificadas em pelo menos um dos trabalhos analisados, presentes na literatura, foram incluídas.

- **Algoritmo α (alpha)**: o algoritmo α é um dos primeiros algoritmos desenvolvidos especificamente para o contexto de mineração de processos. Proposto por Aalst (2011), esse algoritmo visa descobrir um modelo de *workflow* baseado em redes de Petri de dependência do local entre os eventos e no uso de regras heurísticas para a

definição do modelo. A última versão do algoritmo considera elementos temporais. Existem muitas variações desse algoritmo, as quais seguem o mesmo princípio.

- **Baseadas em agentes:** conforme Bonabeau (2002) na modelagem baseada em agentes, o sistema é modelado como uma coleção de entidades autônomas de tomadas de decisão chamadas agentes. Em relação aos outras técnicas de modelagem, esta técnica permite capturar fenômenos emergentes, fornece uma descrição natural de certos tipos de sistemas, e principalmente é flexível.
- **Baseadas em casos:** técnicas que buscam resolver novos problemas adaptando soluções usadas para resolver problemas anteriores. São caracterizadas por: extração do conhecimento a partir de casos (ou experiências) com que o próprio sistema se depara (ou solução); identificação das características mais significantes dos casos apresentados a fim de devolver uma melhor solução (resposta); e, armazenamento e indexação do caso e sua respectiva solução.
- **Baseadas em dados temporais:** trata-se da análise de estruturas complexas que se baseiam no tempo ou na sequência cronológica de um evento, com o objetivo de modelar e prever séries temporais. Muitas técnicas são usadas nesse contexto, e em mineração de processos destacam-se: fragmentação de processos com restrições de tempo; técnicas para mineração de sequências temporais; e Lógica Temporal Linear (LTL — *Linear Temporal Logic*).
- **Baseadas em distância:** usualmente, técnicas baseadas em distância são usadas na tarefa de classificação, mas ao contrário das outras abordagens, não ocorre a construção de um modelo de classificação explícito. Novos exemplos são classificados com base na comparação direta e similaridade aos exemplos de treinamento. Diferentes medidas de similaridade, baseadas em funções de distância, podem ser especificadas para diferentes contextos de aplicação. Algumas das medidas de distância tradicionais são Euclidiana e Manhattan. Um exemplo de método baseado em distância é o “método baseado em distância para detecção de outliers” (CAMILO; SILVA, 2009).
- **Baseadas em estruturas de grafos:** as estruturas chamadas de grafos são muito importantes na modelagem de estruturas complexas, como circuitos, imagens, proteínas, redes biológicas, etc. No contexto de mineração de processos, as estruturas usadas como técnica de modelagem mais usadas são as redes de Petri, que permitem representar sistemas usando como alicerce uma forte base matemática. A estrutura das redes de Petri é um grafo dirigido com dois tipos de vértices (lugares e transições),

os lugares são representados graficamente por círculos (ou por elipses) e as transições por retângulos. Outra forma de usar os grafos em mineração de processos é para representar o modelo organizacional de processo como redes sociais, esse tipo de grafo representa relações (laços) entre agentes independentes (usuários).

- **Baseadas em heurísticas:** de maneira geral, os métodos heurísticos estão constituídos por: princípios (sugestões para a solução: analogia e redução); regras (ajuda, ou pistas, para encontrar maneiras de resolver problemas); e, estratégias (recursos inerentes ao processo de resolução, que podem ser: construtivo, de melhoria, trabalho para a frente ou para trás). No contexto de mineração de processos, referem-se a “regras de ouro” que são suposições a respeito de padrões de processo (AALST, 2011).
- **Baseadas em lógica:** abordagens cujo princípio de funcionamento se baseia em lógica matemática, em que essa lógica proporciona uma maneira de demonstrar se uma questão é verdadeira ou falsa.
- **Baseadas em regras:** em geral, estas técnicas usam as regras como uma forma de representação do conhecimento. Essas regras têm uma estrutura na forma “Se (antecedente ou premissa ou condição) Então (consequente ou conclusão)” que relaciona informações ou fatos (no antecedente) a alguma ação ou resultado (no consequente). São fáceis de criar e de interpretar. Em muitos casos, aparecem na classificação de regras de negócio.
- **Baseadas em semântica:** a semântica é o estudo do significado, incide sobre a relação entre significantes. Dependendo da concepção de significado que se tenha, têm-se diferentes semânticas. As abordagens baseadas em semântica tratam de extrair modelos semânticos, nos quais o conhecimento é expresso em ontologias e permite realizar a descrição semântica de processos a serem minerados. Algumas dessas abordagens são: descrição de processos semânticos, análise e mineração de processos semânticos.
- **Operações OLAP (*Online Analytical Processing*):** Processamento Analítico On-line é uma classe de sistemas que suportam armazenamento, gestão e análise de dados históricos, geralmente derivados de diferentes fontes de dados, incluindo os sistemas OLTP (*Online Transaction Processing*). Esses sistemas operam sobre bases de dados históricas e multidimensionais a partir do uso de algoritmos específicos para carga e análise de dados multidimensionais (por exemplo, análise de cubos de dados). Tais algoritmos são aqui denominados como “operações OLAP”.

- **Probabilidade e estatística:** se caracterizam por terem um modelo de probabilidade subjacente explícito, o qual proporciona uma probabilidade de um exemplo de dado estar, por exemplo, em cada classe em vez de simplesmente uma classificação. Por exemplo, são considerados dentro desse grupo: abordagens baseadas em teorias de decisão, abordagens baseadas no modelo de Markov, e algoritmo Apriori.

Mesmo que uma técnica de inteligência computacional ou de aprendizado de máquina (tal como apresentadas na seção 2.2.2) também seja baseada em princípios de alguma dessas técnicas classificadas aqui como tradicional, ela é classificada, no contexto deste trabalho, apenas como técnica de inteligência computacional ou de aprendizado de máquina. Por exemplo, embora *k-means* seja uma técnica baseada em distância, ela é classificada aqui apenas como uma técnica de aprendizado de máquina, uma vez que ela implementa aprendizado a partir de experiência; da mesma forma, embora lógica *fuzzy* seja uma técnica baseada em lógica, ela é classificada aqui apenas como uma técnica de inteligência computacional, uma vez que implementa raciocínio aproximado.

2.3 Mineração de processos de negócio

Mineração de processos, ou mineração de *workflows*, é uma área de pesquisa relativamente recente que se situa entre mineração de dados e aprendizado de máquina, de um lado, e modelagem e análise de processo, de outro lado (AALST, 2011). Mineração de processos visa descobrir, monitorizar e melhorar processos reais, ou seja, um modelo de processo é construído, avaliado ou melhorado de forma automática por meio da extração de conhecimento a partir dos *logs* de eventos usualmente disponíveis nos sistemas de informação que apoiam a execução dos processos de negócio (AALST, 2011). Apesar de serem usados dados referentes a acontecimentos passados, os resultados podem ser aplicados a casos em execução e futuros. Por exemplo, o tempo de conclusão de uma ordem de compra em execução pode ser previsto usando um modelo de processo.

Modelos de processo de negócio elaborados por humanos não necessariamente representam o processo real. Atualmente, cada vez mais grandes volumes de dados são registrados pelos sistemas de informação acerca da execução de seus processos de negócio, na forma de *log* de eventos. Assim, é possível combinar os pontos fortes de ambas as áreas – mineração de dados (na qual modelos são frequentemente treinados para realizar previsões

sobre casos futuros semelhantes em um mesmo contexto) e modelagem de processos (para a construção de modelos mais próximos ao mundo real). Esses modelos de processo, por serem mais próximos da realidade, constituem-se em uma grande fonte de discernimento para a compreensão e a melhoria dos processos de negócio.

Mineração de processos pode ser vista como a ligação que faltava entre as áreas de mineração de dados e BPM tradicionais. Assim, mineração de processos não deveria ser vista como simplesmente um tipo específico de mineração de dados, uma vez que as tarefas e técnicas de mineração de dados não seriam aplicáveis diretamente ao contexto de BPM. Nesse caso, novas abordagens específicas para mineração de processos seriam necessárias. Por exemplo, a maioria das técnicas comumente usadas em mineração de dados, de uma forma geral, não são orientadas a processos; havendo, portanto, dificuldade em descrever características tais como concorrência existente no contexto de processos. De acordo com Aalst e Dustdar (2012), por exemplo, modelos de processo de negócio não são comparáveis às estruturas simples usadas em mineração de dados, tais como árvores de decisão e regras de associação. Mineração de processos deve tratar modelos de processo de forma completa, levando em consideração, por exemplo, aspectos temporais e considerando a execução do processo como uma sequência de atividades que têm sido executadas.

Porém, a área geral de mineração de dados é bastante ampla, busca tratar dados de todas as naturezas, de diferentes complexidades, e tem apresentado importantes evoluções nos últimos anos. Enquanto as tarefas de mineração de dados mais básicas e tradicionais poderiam ser consideradas muito simples para o contexto de mineração de processos (tais como classificação, regressão, análise de agrupamentos, e Descoberta de padrões frequentes, regras de associação e regras de correlação, conforme apresentadas na seção 2.2.1); por outro lado, mais recentemente, outras tarefas têm sido tratadas em mineração de dados, que visam justamente a lidar com tipos de dados mais complexos, em que os processos de negócio parecem se encaixar de forma satisfatória (tais como análise de *outliers* e de ruído, análise de evolução e análise de dados estruturados, conforme apresentadas na seção 2.2.1).

Uma comparação entre mineração de dados e mineração de processos pode ser feita em relação a análise de *outliers* e de ruído. Em mineração de dados, a generalização é importante para evitar o que é chamado de “*overfitting* dos dados”. Isso significa que, normalmente, busca-se livrar-se de todos os dados que não combinam com a regra geral, considerados ruídos. Por outro lado, em alguns casos, é importante observar que tais

dados destoantes se referem a *outliers*, e não ruídos, ou seja, dados que não obedecem ao comportamento geral do modelo dos dados mas que para o contexto em questão são mais importante do que o próprio comportamento geral. Em mineração de dados, já houve importantes avanços na proposta de abordagens, técnicas e algoritmos para tratar essa tarefa, incluindo o caso de dados de natureza complexa, tais como dados temporais (HAN; KAMBER, 2006). O uso de tais técnicas facilita a construção de modelos de processos de qualidade, robustos, simples, precisos e genéricos. De forma similar, em mineração de processos, a generalização também é necessária para lidar com processos complexos e entender os principais fluxos de processo. Apesar disso, também é importante compreender as exceções ou casos particulares, para descobrir ineficiências e pontos de melhoria. É importante investigar o comportamento de variantes menos executadas, contrastando com a verificação de conformidade. Ruído e dados incompletos transformam a descoberta do modelo de processo em um desafio (AALST et al., 2012); para os quais técnicas já usadas na mineração de dados tradicional podem ser úteis.

2.3.1 Tipos de mineração de processos

Aalst (2011) define três tipos básicos de mineração de processos: descoberta de processos, conformidade de processos, e aprimoramento de processos. Cada um desses tipos de mineração de processos está associado a diferentes objetivos de mineração e pode estar relacionado a diferentes fases do ciclo de vida de BPM. Os três principais tipos de mineração de processos são descritos a seguir:

- **Descoberta de processos:** usa um *log* de eventos para produzir um modelo de processo, sem basear-se em qualquer informação prévia. Diferentes algoritmos podem ser aplicados para, com base em um número suficiente de execuções de um processo, analisar *logs* de eventos e construir o modelo que represente o comportamento desse processo automaticamente. Além do modelo de processo principal propriamente dito, que se refere à perspectiva de controle de fluxo, modelos adicionais que descrevem outras perspectivas também podem ser produzidos. Por exemplo, se o *log* de eventos contiver informações a respeito de recursos, é possível descobrir modelos relacionados a recursos, tal como uma rede social, mostrando como pessoas trabalham em conjunto em uma organização.

- **Conformidade de processos:** compara um modelo de processo existente com o modelo resultante de um *log* de eventos do mesmo processo em execução. Pode ser usada para validar se a realidade, refletida no *log*, está em conformidade com o modelo de processo e vice-versa. Por exemplo, pode ser usada para mostrar se alguma regra existente no modelo de processo foi ou não implementada e se está ou não realmente sendo seguida pelos usuários. Auditorias podem ser beneficiadas por esse tipo de mineração, já que fraudes podem ser buscadas ao se verificar, por exemplo, se há ações obrigatórias de um processo que não estão sendo executadas.
- **Aprimoramento de processos:** busca consertar ou melhorar um modelo de processo já existente com base em informações registradas em um *log* de eventos. O aprimoramento pode ser aplicado para “consertar” um modelo de processo; ou seja, alterar o modelo para que ele reflita melhor a realidade. Por exemplo, duas atividades que estão modeladas em paralelo mas que na verdade podem ser executadas em qualquer ordem podem levar a um remodelo de processo. Já o aprimoramento aplicado para “melhorar” adiciona uma nova perspectiva ao modelo de processo. Por exemplo, um modelo de processo pode ser estendido com dados de desempenho, recursos, métricas de qualidade, etc. Assim, usando referências temporais dos *logs*, é possível complementar o modelo de modo que informações sobre níveis de serviço, tempos de execução e frequências sejam também consideradas para análise.

Independentemente do tipo básico de mineração, existe uma relação importante entre os tipos de mineração de processos tanto com “modelos de processo” quanto com “*log* de eventos”, conforme ilustrado com maior destaque na figura 3, que descreve os três principais tipos de mineração de processos em termos de entradas e saídas (AALST, 2011). A descoberta de processos usa como entrada um conjunto de *logs* de eventos e produz como saída um modelo de processo. A verificação de conformidade de processos usa como entrada tanto um conjunto de *logs* de eventos quanto um modelo de processo exemplo e produz como saída informações de diagnóstico que identificam diferenças ou semelhanças entre as duas entradas. O aprimoramento de processos também usa como entrada um conjunto de *logs* de eventos e um modelo de processo, mas produz como saída um novo modelo de processo – estendido; ou seja, uma nova versão do modelo de processo usado como entrada com base nas informações existentes na outra entrada. Enquanto a conformidade de processos determina o posicionamento entre o modelo e a realidade, o

aprimoramento de processos complementa ou melhora o modelo fornecido à priori. Já a descoberta apenas cria um modelo que representa a realidade.

Figura 3 – Tipos básicos de mineração de processos em termos de entradas e saídas

Fonte: Adaptação de Aalst (2011)

2.3.2 Perspectivas e problemas de mineração de processos

Mineração de processos não está limitada aos tipos básicos descoberta, conformidade e aprimoramento de processos. Considerando que um *log* de eventos fornece dados sobre a realidade de uma organização, tais dados podem fornecer diferentes pontos de vista sobre essa realidade e, conseqüentemente, proporcionar diferentes abstrações intencionais de determinados comportamentos capturados em tais organizações. Nesse sentido, para um determinado *log* de eventos, pode haver múltiplos e diferentes pontos de vista úteis. Além disso, diferentes partes interessadas podem pretender analisar diferentes pontos de vista. Assim, diferentes modelos de processos podem ser orientados a determinadas perspectivas, sendo descritos de formas diferentes, e enfatizando características relevantes para determinados tipos de usuários.

Aalst (2011), por exemplo, apresenta, como um possível conjunto de diferentes perspectivas, as quatro seguintes:

- **Perspectiva de controle de fluxo**: focada na ordenação de atividades, visando encontrar uma boa caracterização de todas as sequências possíveis. O resultado é

expresso tipicamente em termos de uma rede de Petri ou alguma outra notação de processo, tal como EPC, BPMN, ou diagramas de atividade da UML.

- **Perspectiva organizacional:** focada em informações implícitas no *log* de eventos relacionadas a recursos, ou seja, a atores envolvidos (ex.: pessoas, sistemas, papéis, ou departamentos) e a como eles se relacionam entre si. Busca-se estruturar a organização via classificação de atores avaliando diferentes aspectos.
- **Perspectiva de caso:** focada em aspectos específicos de casos individuais gerados na execução de um processo. Uma instância pode ser caracterizada pelo seu percurso no processo, pelos atores que a executam, ou por valores descritos nos dados. Por exemplo, se uma instância representa uma ordem de reposição de estoque, pode ser importante saber qual é o fornecedor ou o número de produtos requisitados.
- **Perspectiva temporal:** focada na temporização e na frequência de eventos. Nesse caso, referências temporais são anotadas nos eventos, possibilitando: a descoberta de gargalos; a medição de níveis de serviço; o monitoramento do uso de recursos; e a predição de tempo de processamento necessário para concluir instâncias existentes.

A descrição de processos considerando sua “perspectiva de controle de fluxo” é a espinha dorsal de qualquer modelo de processo. Algumas construções básicas de fluxo apoiadas pela generalidade das linguagens tradicionais são: sequência, encaminhamento paralelo (AND-split/AND-join), escolha (XOR-split/XOR-join), e ciclos. Técnicas de mineração de processos devem ser capazes de apoiar esses tipos de construções (AALST et al., 2012). Porém, nem todas as técnicas conseguem lidar com todos os casos, tal como “concorrência” em processos.

Devido à natureza do processo e características de qualidade dos *logs* de eventos, abordagens atuais em mineração de processos apresentam problemas para lidar com dados reais, visto que esses dados tendem a ser incompletos, ruidosos e imprecisos. Além disso, processos atuais tendem a ser complexos e sujeitos a um grande número de variações. Isto tem como resultado *logs* de eventos com uma granularidade muito fina, heterogêneos e voluminosos. Bose, Mans e Aalst (2013) classificam os problemas e desafios decorrentes na análise de *logs* de eventos em duas categorias, descritas a seguir e apresentadas de forma gráfica na figura 4:

- **Características do processo:** dado um conjunto de *logs* de eventos, algumas métricas básicas podem ser extraídas, tais como: “número de casos”, “número de

Figura 4 – Problemas e desafios decorrentes da análise de *logs* de eventos

Fonte: Adaptação de Bose, Mans e Aalst (2013)

atividades únicas”, e “número de instâncias de processo únicas”. Dependendo como essas métricas apareçam nos *logs* de eventos, alguns desafios podem ser encarados, tais como: *(i)* números voluminosos de casos ou eventos, por exemplo, sistemas de alta tecnologia produzem grandes quantidades de dados, eventos são capturados em um baixo nível de execução, como eventos no nível de aplicação ou redes de comunicação; *(ii)* números grandes e heterogêneos de casos e instâncias diferentes, processos atuais estão desenhados para ser “flexíveis”, como resultado um *log* de eventos contém uma mistura heterogênea entre diversos cenários de uso e comportamentos não estruturados; *(iii)* números altos de granularidade de eventos para atividades diferentes, de um lado, processos são concebidos com muito de detalhe comprometem a compreensão do seus modelos, em tanto que, analistas e usuários finais preferem níveis de abstração bem altos, encontrar um equilíbrio entre esses dois níveis de granularidade é um dos maiores desafios em mineração de processos; *(iii)* processos flexíveis e sujeitos a mudanças, usualmente processos de negócio são executados em entornos dinâmicos, eles estão sujeitos a um amplo número de variações, mudanças nos processos podem se manifestar de forma latente no *log* de eventos, tais mudanças podem ser evolucionárias ou momentânea dependendo sua duração.

- **Qualidade do *log* de eventos:** estes problemas podem ser agrupados em quatro categorias principais: *(i)* dados incompletos, corresponde ao cenário onde diferentes tipos de informações são perdidas mesmo sendo obrigatórias; *(ii)* dados incorretos, cenário onde apesar de que os dados estão presentes no *log* de eventos, eles estão baseados em contexto diferentes; *(iii)* dados imprecisos, as entradas registradas são demasiado gerais e conduz a uma perda de precisão; e *(iv)* dados irrelevantes, cenários onde é necessária uma filtragem e/ou transformação dos registros irrelevantes. Sendo que, não necessariamente todos esses problemas aparecem juntos em um mesmo *log* de evento.

2.3.3 Algoritmos e ferramentas usados

O resultado da mineração de processos depende principalmente da técnica de mineração usada. Segundo Aalst (2011), um grande desafio para técnicas que tratam mineração de processos com qualidade é obter um equilíbrio entre os critérios abaixo, sendo que ainda não há técnicas que os satisfazem de forma adequadamente equilibrada:

- **Robustez:** um modelo de processo é robusto se diversas variações descritas nos *logs* de eventos puderem ser reproduzidas no modelo, do início ao fim.
- **Simplicidade:** um modelo de processo é simples se ele for capaz de explicar o comportamento descrito no *log* de eventos.
- **Precisão:** um modelo de processo é preciso se ele não permitir variantes em excesso.
- **Generalização:** um modelo de processo deve generalizar os comportamentos descritos no *log* de eventos, sem restringir o comportamento apenas aos dados registrados.

Um dos primeiros algoritmos propostos especificamente para mineração de processos, para ser usado mais especificamente na descoberta de processos, e um dos ainda mais importantes, é o algoritmo α , proposto por Aalst (2011). Ele foi desenvolvido visando lidar de forma adequada com a concorrência entre as atividades de processo. O algoritmo α recebe como entrada um *log* de eventos simples, procura padrões particulares, baseando-se em um conhecimento *a priori* sobre regras e estruturas usuais dentro do negócio. Produz, como saída, um modelo de processo representado por uma rede de Petri que, espera-se, possa refletir a realidade do processo (AALST, 2011). Por exemplo, se uma atividade a é seguida por b , e a atividade b nunca é seguida por a , então se presume que existe

uma dependência causal entre a e b . Para refletir essa dependência, a rede de Petri correspondente deveria ter uma ligação de a até b .

A principal limitação do algoritmo α é que os processos geralmente podem conter vários tipos de construções que esse algoritmo não pode descobrir, como pequenos ciclos de atividades e tarefas duplicadas (MEDEIROS et al., 2004). Para lidar com esse problema, o algoritmo β (beta) foi proposto, com capacidade de lidar com o problema de ciclos curtos e concentrado mais em tarefas ocultas e duplicadas, e construções avançadas de roteamento. Versões mais recentes do algoritmo α já consideram também as variações temporais nas sequências de atividades.

Yue et al. (2011) descrevem outras abordagens de mineração de processos, embora a maioria delas sejam baseadas nas abordagens propostas inicialmente pelo pesquisador Prof. Dr. Wil van der Aaslt, principal pesquisador em mineração de processos. Alguns dos algoritmos descritos são: algoritmo para lidar com modelos de processos de estruturas hierárquicas; métodos de agrupamento para lidar com tarefas duplicadas; variações do algoritmo α , como o algoritmo γ (gama) e o algoritmo θ (teta); método de derivação e estatística; etc. Além disso, algumas ferramentas adicionais são também mencionadas: *InWoLvE* e *Process Miner tool*.

O algoritmo “GeneticMiner” foi proposto por Medeiros, Weijters e Aalst (2007) para a descoberta de processos e é baseado em técnicas de computação evolutiva. Seu método de descoberta parte de uma população inicial de indivíduos (nesse caso, os processos que são candidatos à solução); várias iterações são realizadas até estar mais próximo da solução apropriada, de forma que as soluções encontradas vão sendo validadas por meio de medições da precisão. Quando comparado aos demais algoritmos de mineração, o “GeneticMiner” apresenta melhor precisão do modelo de processo extraído especialmente para processos que apresentam comportamentos complexos. No entanto, ele normalmente demanda alto tempo de processamento para chegar ao resultado.

Dentre as ferramentas específicas para mineração de processos, a mais conhecida é ProM (TU/E, 2015). Trata-se de uma plataforma extensível e de código aberto, que apoia uma ampla variedade de técnicas de mineração de processos sob a forma de *plug-ins*. Ela foi desenvolvida pelo grupo da mineração de processos da “Eindhoven University of Technology” (grupo liderado pelo Prof. Dr. Wil van der Aaslt) (TU/E, 2015). Ela possui uma ampla variedade de tipos de modelos, como redes de Petri, redes heurísticas, EPCs, redes sociais, e YAWL. Tais modelos podem ser usados tanto como entradas quanto

como saídas para os *plug-ins* disponíveis. E esses *plug-ins* podem ser do tipo: importação, exportação, conversão, filtro de *log* de eventos, mineração e análise.

Disco Fluxicon é outra ferramenta específica para mineração de processos. Trata-se de uma ferramenta comercial, cujos desenvolvedores trabalham em estreita colaboração com a equipe da ProM. A Disco Fluxicon realiza a descoberta automática do modelo de processo com base em um arquivo de dados importado em formato “*Comma-separated values*”. O algoritmo usa os fundamentos da abordagem “Fuzzy Miner”, e segundo seus criadores implementa otimização significativa do desempenho de processamento. A Disco Fluxicon foi desenvolvida com foco em seu uso comercial, e apresenta um claro foco em questões como usabilidade, mas não em facilidades para uso de diferentes algoritmos e técnicas na forma de *plug-ins*, como ocorre com a ProM. Na verdade, a Disco Fluxicon implementa apenas uma versão de seu próprio algoritmo, sem divulgar detalhes de implementação, e busca oferecer bons resultados nesse único algoritmo, incluindo bons resultados de desempenho, qualidade e usabilidade. Do ponto de vista de visualização gráfica dos resultados, por exemplo, suas funcionalidades são muito melhores do que da ProM. Além disso, ela apresenta um conjunto completo de métricas de processo para as atividades e caminhos no processo, incluindo: frequência absoluta, frequência do caso, número máximo de repetições, duração total, duração média, duração máxima. Ela permite ainda explorar cada “Caso” ou conjunto de “Casos” relevantes, e inspecionar rapidamente o seu histórico, por meio de uma tabela detalhada de atividades, recursos, tempos de início, duração e um conjunto completo de atributos para cada evento do “Caso”, fluxos de processo comuns, e *outliers* (LABORATORIES, 2015).

2.3.4 Exemplo ilustrativo de mineração de processos

Esta seção apresenta um exemplo ilustrativo de mineração de processos, do tipo descoberta de processo, visando esclarecer esse conceito, com base na proposta de Aalst (2011). Considera-se, como ponto de partida para mineração de processos, o *log* de eventos apresentado na tabela 1. Nesse exemplo, em particular, as atividades são consideradas como atômicas, a duração das atividades não é informada, e cada evento é associado a uma atividade. As colunas “IdCaso”, “Atividade” e “*Timestamp*”, representam os atributos mínimos do *log* de eventos, necessários para a mineração de processos. Inicialmente, na tabela 1, apresenta-se uma visão do *log* de eventos focada nos eventos. Uma forma de

modelar esses dados a fim de aplicar uma técnica de mineração de processos é convertendo-os em uma estrutura focada em instâncias de execução, de um ou vários processos. Assim, para facilitar a visualização, é possível adotar a seguinte notação: $a = Realizar\ requerimento$; $b = Analisar\ informalmente$; $c = Analisar\ a\ detalhe$; $d = Verificar\ bilhete$; $e = Decidir$; $f = Reiniciar\ requerimento$; $g = Pagar\ compensação$; e, $h = Rejeitar\ requerimento$. Na sequência, seguindo a ordem de ocorrência das informações contidas nas colunas “IdCaso” e “Timestamp”, obtém-se uma representação mais compacta, que é apresentada na tabela 3. Diversos algoritmos podem ser aplicados sobre tal estrutura. O resultado da aplicação do algoritmo α , por exemplo, é a rede de Petri apresentada na figura 5.

Verifica-se que as seis instâncias de processos contidos na tabela 3 estão contidos também no modelo da figura 5. Como este é um exemplo com poucos dados, é possível encontrar de forma visual as seis instâncias de processos dentro do modelo. Contudo, em situações reais quando trata-se com grandes volumes de registros nos *log* de evento, realizar tal verificação não é possível; logo, algoritmos de mineração de processos devem dar a confiabilidade de que o comportamento contido no *log* de eventos realmente representa os principais processos de negócio. E não, um modelo tão específico que permite apenas aqueles “comportamentos casuais” observados (este caso é chamado de *overfitting*), nem um modelo muito genérico e impreciso (este caso é chamado de *underfitting*) (AALST, 2011).

Esse exemplo apresenta apenas um tipo de representação de modelo do processo. Além disso, em situações reais, podem existir dezenas de atividades potencialmente concorrentes. Assim, os modelos resultantes podem ser muito mais abstratos (ou seja, podem refletir diversos tipos de comportamentos na execução de processo); ou também, eles podem ser extremamente complexos (por exemplo, comportamentos de atividades concorrentes não são representados dentro do modelo).

Tabela 3 – Conjunto de atividades sequenciais gerado a partir do *log* de eventos da tabela 3

IdCaso	Sequencia de atividades
1	(a, b, d, e, h)
2	(a, d, c, e, g)
3	$(a, c, d, e, f, b, d, e, g)$
4	(a, d, b, e, h)
5	$(a, c, d, e, f, d, c, e, f, c, d, e, h)$
6	(a, c, d, e, g)

Fonte: Adaptação de Aalst (2011)

Figura 5 – Modelo de processo descoberto após a aplicação do algoritmo α no *log* de eventos da tabela 3

Fonte: Adaptação de Aalst (2011)

2.4 Considerações finais

Neste capítulo foram apresentados aspectos gerais das áreas estritamente relacionadas à mineração de processos, gestão de processos de negócio por um lado, e mineração de dados, inteligência computacional e de aprendizado de máquina do outro lado. A junção dessas duas áreas cria a área de mineração de processos de negócio. O principal foco de pesquisa da área de mineração de processos é a extração de conhecimento a partir de um *log* de eventos registrados durante as etapas do ciclo de vida de BPM, a fim de diminuir a brecha existente entre o modelo proposto e a realidade do negócio.

Também foram apresentados conceitos relacionados aos algoritmos e técnicas que atualmente são os mais usados na área de mineração de dados, principalmente aqueles provenientes das áreas de inteligência computacional e de aprendizado de máquina, por serem bem sucedidas para a resolução de problemas complexos usando dados reais. Esse tipo de problemas geralmente requerem alguma forma de aquisição automática de conhecimento a partir do domínio de aplicação e, portanto, não podem ser solucionados por técnicas computacionais convencionais. Por outro lado, uma das maiores dificuldades na área de mineração de processos é lidar com tal complexidade em dados provenientes de execuções de processos em situações reais. Além do mais, a maioria dos algoritmos e técnicas tradicionalmente usados em mineração de processos seguem as abordagens propostas por Aalst (2011), deixando de lado o potencial que tem as técnicas de inteligência computacional e de aprendizado de máquina para solucionar os problemas similares aos encontrados em mineração de processos.

3 Método de pesquisa

Este capítulo visa descrever, em linhas gerais, o método de pesquisa usado neste trabalho a fim de atingir os objetivos.

Este trabalho visa estudar de forma ampla o contexto atual da área de mineração de processos, sendo necessários o uso de métodos de pesquisa formais e sistemáticos. De forma mais ampla, o método de pesquisa usado enquadra-se como:

- **Natureza “básica”**: uma vez que objetiva gerar conhecimentos novos, úteis para o avanço da ciência – neste caso específico, as áreas de BPM, mineração de dados, e mineração de processos; sem a previsão imediata de uma aplicação prática.
- **Gênero “empírico”**: já que esse trabalho foi direcionado principalmente pela busca de dados relevantes e convenientes obtidos por meio da experiência, da vivência de outros pesquisadores, publicadas por meio de trabalhos de pesquisa existentes na literatura científica; tendo como objetivo chegar a novas conclusões a partir da maturidade experimental desses trabalhos avaliados.
- **Objetivos “exploratório” e “descritivo”**: primeiramente com caráter “exploratório” por proporcionar maior familiaridade com o problema, que pode ainda ser considerado novo, principalmente considerando o uso de técnicas de inteligência computacional e de aprendizado de máquina, com vistas a torná-lo mais explícito ou construir hipóteses; mas também com caráter “descritivo” por já possuir o objetivo de descrever os fatos e/ou fenômenos da realidade estudada.
- **Abordagem mista “qualitativa” e “quantitativa”**: em relação à análise e apresentação dos resultados, sendo que, em alguns casos, os resultados podem ser mensurado em números, e então classificados e analisados com estatística descritiva básica, e, em outros, uma opinião crítica das lições aprendidas no processo é sustentada com bases nas teorias encontradas.

Como procedimento técnico, esse trabalho foi baseado principalmente em uma pesquisa do tipo “bibliográfica”, em que os principais trabalhos já publicados na área de mineração de processos foram estudados; tendo como fonte de informação artigos científicos publicados nos principais veículos científicos internacionais. A técnica para coleta e análise de dados foi “revisão sistemática da literatura”, conforme diretrizes estabelecidas por Kitchenham (2007). Duas revisões sistemáticas foram conduzidas: a primeira focada em

duas técnicas bem específicas de inteligência computacional e de aprendizado de máquina – redes neurais e máquinas de vetores de suporte; e a segunda, mais geral, abordando todo o contexto de mineração de processos independentemente do tipo de técnica aplicada. De fato, a segunda revisão sistemática realizada, de acordo com a teoria seguida, se refere a um “mapeamento sistemático da literatura” dada a amplitude em que ele foi realizado.

De forma geral, o desenvolvimento deste trabalho compreendeu quatro etapas importantes, sendo que em cada uma delas atividades para alcançar os objetivos específicos foram realizadas. Tais etapas são apresentadas resumidamente na figura 6 e descritas a seguir:

1. Revisão bibliográfica estendida, por meio da revisão sistemática e do mapeamento sistemático.
2. Delimitação da área de mineração de processos, em relação às técnicas e as tarefas de mineração de dados, quando aplicados em tal contexto, buscando evidências para a justificativa ou não de uma maior ou menor aplicação de técnicas provenientes especificamente das áreas de inteligência computacional e de aprendizado de máquina.
3. Realização de um exemplo de aplicação, como um meio para compreender e sustentar argumentos sobre aplicações reais na área de mineração de processos, considerando o ponto de vista tanto de técnicas tradicionais quanto de técnicas de inteligência computacional e de aprendizado de máquina.
4. Análise dos resultados obtidos com um exemplo de aplicação visando generalizar as conclusões obtidas para cenários diversos de mineração de processos, para contribuir com o levantamento das lições aprendidas com este trabalho.

Inicialmente, previa-se realizar uma única revisão sistemática da literatura envolvendo todos os estudos existentes na área de mineração de processos. Porém, com as primeiras buscas exploratórias, verificou-se que o escopo dos trabalhos era consideravelmente maior que o planejado inicialmente. Assim, percebeu-se a necessidade de realizar um mapeamento sistemático da literatura de toda a área de mineração de processos, e uma revisão sistemática detalhada focada apenas em algumas das técnicas de importância para as áreas de inteligência computacional e de aprendizado de máquina, para as quais foram escolhidas redes neurais e máquinas de vetores de suporte. Essas duas técnicas foram escolhidas por sugestão da especialista na área de mineração de dados com uso de técnicas inteligentes – a Profa. Dra. Sarajane Marques Peres, coorientadora deste trabalho,

Figura 6 – Etapas seguidas durante a realização deste trabalho

Fonte: Ana Rocío Cárdenas Maita, 2015

considerando os bons resultados que elas apresentam quando aplicadas em problemas complexos da área de mineração de dados (WU et al., 2008; STAHL; JORDANOV, 2012).

O método de revisão sistemática proposto por Kitchenham (2007) facilita o levantamento bibliográfico que consiste em identificar, avaliar e interpretar as pesquisas relevantes dentro de uma determinada área de pesquisa. O objetivo desse método é criar uma organização sistemática para conduzir a revisão de forma que seja possível documentar todo o processo de levantamento bibliográfico, possibilitando aos leitores a oportunidade de avaliar a qualidade da revisão ou até mesmo reproduzi-la. Os conceitos apresentados pela autora diferenciam esses dois tipos de revisão da literatura. Por um lado, o mapeamento sistemático é centrado na classificação e na categorização geral dos resultados, a fim de ter uma visão mais ampla dos estudos primários, aplicável – como é caso deste trabalho – quando a área em estudo possui muitos estudos primários. Já a revisão sistemática visa identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para questões de pesquisa definidas, de forma mais focada em um ponto específico da área de estudo.

Os dois tipos de revisão são necessários como parte dos resultados para alcançar os objetivos de pesquisa deste trabalho, não sendo tratadas como simples revisões bibliográficas, mas sim como estudos documentários que permitem a obtenção de uma visão mais geral dos direcionamentos atuais da área. A revisão sistemática realizada permite verificar que, embora haja interesse científico na área de mineração de processos, pouco tem sido investido especificamente em técnicas como redes neurais e máquinas de vetores de suporte.

Por outro lado, o mapeamento sistemático permitiu confirmar que as técnicas usuais provenientes das áreas de inteligência computacional e de aprendizado de máquina, realmente vem sendo pouco aplicadas em mineração de processos, embora haja um crescimento de tais técnicas nos últimos anos.

Nas secções seguintes, os protocolos resumidos, tanto para a revisão sistemática quanto para o mapeamento sistemático, são apresentados, os quais se referem às principais contribuições deste trabalho. Esses protocolos foram definidos seguindo as diretrizes estabelecidas por Kitchenham (2007), e compreendem basicamente três fases: *(i)* planejamento, que inclui o desenvolvimento de um protocolo de revisão, o qual especifica os métodos para executar a revisão, incluindo uma análise exploratória para definição de conceitos, a definição de *strings* de buscas, e o estabelecimento de critérios de inclusão e de exclusão de estudos encontrados; *(ii)* condução, que inclui a seleção dos estudos, a avaliação da qualidade dos estudos selecionados, a extração de dados, e a organização da informação extraída; e *(iii)* relatório, que inclui a documentação dos resultados em relatórios técnicos, trabalhos de dissertação ou teses, e em artigos científicos.

Uma descrição detalhada do protocolo definido e usado neste trabalho para a realização da revisão sistemática está sendo submetida como parte de um artigo científico. Já para a realização da revisão sistemática, tal descrição encontra-se em um artigo já publicado em um periódico internacional (MAITA et al., 2015).

A fim de ajudar no levantamento de possíveis hipóteses ou então de comprovar hipóteses já levantadas durante o mapeamento sistemático, um exemplo de aplicação foi realizado. Um exemplo de aplicação foi realizado com dados provenientes de um *log* de eventos registrados durante uso de um sistema de ensino a distância de cursos de especialização oferecidos pela Universidade de São Paulo em conjunto com a Universidade Virtual do Estado de São Paulo.

3.1 Protocolo resumido da revisão sistemática (específica)

O objetivo da revisão sistemática foi avaliar a aplicação das técnicas redes neurais e máquinas de vetores de suporte em tarefas de mineração de dados no contexto de mineração de processos, com o objetivo de entender como tais técnicas de inteligência computacional são aplicadas no contexto de mineração de processos.

3.1.1 Questões de pesquisa

As seguintes questões de pesquisa nortearam essa revisão sistemática:

QP1. Que tipos de redes neurais ou de máquinas de vetores de suporte são usados para implementar as tarefas de mineração de dados quando aplicados ao contexto de mineração de processos?

Para responder essa pergunta, os seguintes tipos de classificação foram considerados. Primeiro, em relação às redes neurais, elas podem ser classificadas em relação ao tipo de arquitetura ou em relação ao tipo de aprendizado. Em relação ao tipo de arquitetura, que se refere a como suas unidades de processamento (neurônios) são estruturadas, as redes neurais podem ser classificadas como: *feedforward* com camada única; recorrente com camada única; *feedforward* multicamada; recorrente com multicamada; e de camada competitiva (HAYKIN, 2009). Em relação ao tipo de aprendizado, que se refere à estratégia usada pela técnica para corrigir ou aproximar o modelo mapeado ou a superfície de decisão que está sendo procurada. As redes neurais podem ser classificadas como de aprendizado supervisionado, aprendizado não supervisionado, aprendizado de reforço, e redes neurais de pesos fixos (FAUSETT, 1994).

Já em relação às máquinas de vetores de suporte, elas podem ser classificadas em relação às suas variações, juntamente com o tipo de aprendizado. O modelo básico de máquinas de vetores de suporte foi considerado, categorizada como “aprendizado supervisionado”, adicionado de suas duas variações: Regressão por Vetores Suporte (SVR – *Support Vector Regression*), categorizada como “aprendizado supervisionado”; e Agrupamento por Vetores Suporte (SVC – *Support Vector Clustering*), categorizada como aprendizado não supervisionado (HAYKIN, 2009; CRISTIANINI; SHAW-TAYLOR, 2000).

QP2. Quais tarefas de mineração de dados são tratadas nos trabalhos de mineração de processos por meio da aplicação de redes neurais ou de máquinas de vetores de suporte?

Para essa questão de pesquisa, as principais tarefas de mineração de dados segundo Han e Kamber (2006), foram consideradas: (i) predição categórica (ou classificação); (ii) predição numérica (do tipo “regressão”); (iii) predição numérica (do tipo “identificação de tendências”); (iv) análise de agrupamentos; e (v) Descoberta de padrões frequentes, regras

de associação e regras de correlação; conforme apresentados na seção 2.2.1.

QP3. Quais tipos de mineração de processos são atendidos por redes neurais ou máquinas de vetores de suporte, quando tais técnicas são usadas para resolver tarefas de mineração de dados?

Para essa questão de pesquisa, foram considerados os três principais tipos de mineração de processos: (i) descoberta de processos; (ii) conformidade de processos; e (iii) aprimoramento de processos (AALST, 2011); conforme definidos na seção 2.3.1.

3.1.2 Fontes de dados e estratégia de busca

A um fim de maximizar o número de trabalhos candidatos encontrados, foram escolhidas como fontes de dados de pesquisa as bases: Scopus e ISI Web of Science. Essas duas bases juntas indexam a maioria dos artigos existentes nas mais importantes bibliotecas digitais, tais como: IEEEExplore, Biblioteca Digital ACM e SpringerLink. A *string* de busca definida (Tabela 4) visa encontrar trabalhos que sejam relacionados a mineração de processos usando apenas as técnicas de redes neurais e de máquinas de vetores de suporte. Essa *string* foi aplicada pela primeira vez em 2013, posteriormente reaplicada ao final de 2014.

Tabela 4 – *String* de busca genérica para a revisão sistemática

((“process mining” OR “processes mining” OR “workflow mining” OR “workflows mining” OR “mining process” OR “mining processes” OR “mining workflow” OR “mining workflows” OR (“business process” OR “business processes” OR workflow OR workflows) AND “data mining”)) AND ((neural OR SOM OR “Self Organizing” OR “Self-Organizing” OR “Organizing Map” OR “Organizing Maps” OR MLP OR “Multilayer Perceptron” OR “Backpropagation” OR “Back-Propagation” OR RBF OR “Radial Basis Function” OR artmap OR “Adaptive Resonance Theory” OR hopfield OR LVQ OR “Learning Vector Quantization”) OR (SVM OR SVR OR SVC OR “Support Vector” OR “Support-Vector”))

Fonte: Maita et al. (2015)

3.1.3 Seleção de trabalhos

Foram especificados critérios de inclusão e de exclusão baseados no escopo e na qualidade dos artigos encontrados para garantir que, ao final, fosse selecionados apenas trabalhos realmente relacionados ao contexto desejado.

3.1.3.1 Critérios de inclusão

- **CI-1:** o artigo trata essencialmente mineração de dados, ou seja, mineração de dados está diretamente relacionada ao escopo principal do trabalho em vez de ser meramente mencionada de forma genérica.
- **CI-2:** o artigo trata essencialmente processos de negócio (incluindo *workflows* de negócio), ou seja, processos/ *workflows* estão diretamente relacionados ao escopo principal do trabalho em vez de serem meramente mencionados de forma genérica.
- **CI-3:** ambos, mineração de dados e processos de negócio, são tratadas em conjunto no artigo para apresentar uma abordagem de mineração de processos em vez de cada abordagem ser tratada de forma independente no artigo.
- **CI-4:** a abordagem apresenta como objetivo principal o uso de redes neurais ou máquinas de vetores de suporte como uma técnica para implementar tarefas de mineração de dados no contexto de mineração de processos.

3.1.3.2 Critérios de exclusão

- **CE-1:** o artigo não está disponível eletronicamente na web;
- **CE-2:** o artigo não é apresentado inteiramente no idioma inglês;
- **CE-3:** o artigo não está relacionado, principalmente, às áreas de ciência da computação ou de sistemas de informação. Por exemplo, o artigo está relacionado principalmente às áreas de medicina ou engenharia industrial;
- **CE-4:** o registro de dados identificado na verdade não se refere a um artigo científico, mas sim a alguma publicação sem revisão por pares, tais como: relatórios técnicos; livros e capítulos de livros; prefácios de anais de conferência; ou editoriais de periódicos;

- **CE-5:** o artigo apresenta algum tipo de avaliação de outros trabalhos, como uma pesquisa ou alguma revisão sistemática (ou seja, um estudo secundário), e não os resultados de algum trabalho de pesquisa específica (ou seja, um estudo primário);
- **CE-6:** o artigo foi publicado há mais de 11 anos.

Além, critérios de avaliação da qualidade são definidos a seguir na tabela 5, adaptados da proposta de Dybå e Dingsøyr (2008) para ser aplicados nessa revisão.

Tabela 5 – Critérios de avaliação da qualidade

Critério	Descrição
<i>Relatório</i>	<i>Interessada com a qualidade da informação que está sendo relatada em termos de uma lógica de estudo, objetivos e contexto</i>
Rela.1	O estudo esta baseado em pesquisas (ou meramente um “lições aprendidas” relatório com base na opinião de especialistas)?
Rela.2	Existe uma declaração clara dos objetivos da pesquisa?
Rela.3	Existe uma descrição adequada do contexto em que a pesquisa foi realizada?
<i>Rigor</i>	<i>Interessada com a precisão dos métodos de investigação usados para estabelecer a validade de instrumentos de recolha de dados e os métodos de análise e, portanto, a confiabilidade dos resultados</i>
Rig.1	O projeto de pesquisa foi adequado para tratar os objetivos da pesquisa?
Rig.2	A estratégia de coleção era apropriada para os objetivos da pesquisa?
Rig.3	Houve um grupo de controle para comparar os tratamentos?
Rig.4	Os dados foram recolhidos, de tal maneira que eles tratam a questão de pesquisa?
Rig.5	A análise dos dados foi suficientemente rigorosa?
<i>Credibilidade</i>	<i>Interessada com a confiabilidade dos métodos de estudo usados para assegurar que os resultados sejam válidos e significativos</i>
Cre.1	A relação entre pesquisador e participantes foi considerada de forma adequada?
Cre.2	Existe uma declaração clara dos resultados?
<i>Relevância</i>	<i>Interessada com a avaliação da importância do estudo para a indústria de software em geral e da comunidade de investigação</i>
Rele.1	O estudo é de valor para a investigação ou para aplicação?

Fonte: Adaptação de Dybå e Dingsøyr (2008)

3.2 Protocolo resumido do mapeamento sistemático (geral)

De acordo com Kitchenham (2007), mapeamento sistemático é um tipo de revisão sistemática. Enquanto que uma revisão sistemática tenta se aprofundar na análise de cada estudo primário identificado, em um baixo nível de granularidade; o mapeamento sistemático é um estudo de escopo, ou seja, uma ampla revisão da estudos primários em

uma área de tópico específico que tem o objetivo de identificar quais evidências estão disponíveis sobre o tema. Um mapeamento sistemático é preferível em vez de uma revisão sistemática quando o tema é muito amplo, como a mineração de processos se mostrou. Um mapeamento sistemático permite coletar evidências em um domínio para ser tratado em um alto nível de granularidade.

O principal objetivo do mapeamento sistemático realizado neste trabalho foi identificar e avaliar todos os trabalhos que propõem abordagens para mineração de processos de negócio, independentemente do tipo de técnica usada, mesmo que o foco de interesse aqui seja as técnicas de inteligência computacional e de aprendizado de máquina. Para compreender qual a importância desse tipo de técnica no contexto geral de mineração de processos, é necessário conhecer o contexto completo, incluindo o uso das técnicas tradicionais.

3.2.1 Questões de pesquisa

As seguintes questões de pesquisa nortearam essa revisão sistemática:

QP1. Quais tipos de mineração de processos são atendidos pelas técnicas e tarefas de mineração de dados?

Para essa questão de pesquisa, foram considerados os três principais tipos de mineração de processos: *(i)* descoberta de processos; *(ii)* conformidade de processos; e *(iii)* aprimoramento de processos (AALST, 2011); conforme definidos na seção 2.3.1. O objetivo desta pergunta foi identificar quais tipos de mineração de processos têm sido mais amplamente aplicados no período avaliado, ou seja, nos últimos dez anos. Com tal análise, é possível identificar, por exemplo, se houve alguma tendência crescente ou decrescente em termos de quais tipos de mineração têm sido tratados durante esse tempo. Além disso, é possível realizar análises transversais considerando como as tarefas de mineração de dados e as técnicas usadas são tratadas pelos diferentes pesquisadores em relação aos diferentes tipos de mineração de processos.

QP2. Quais tarefas de mineração de dados são aplicadas em mineração de processos?

A fim de direcionar essa questão de pesquisa, foram consideradas as tarefas de mineração de dados de acordo com a taxonomia apresentada por Han e Kamber (2006): *(i)*

predição categórica (ou classificação), predição numérica (tipo “regressão”), e predição numérica (tipo “identificação de tendências”); (ii) análise de agrupamentos; (iii) Descoberta de padrões frequentes, regras de associação e regras de correlação; (iv) análise de ruído e de *outliers*; (v) análise de evolução; e (vi) análise de dados estruturados. Em relação a primeira revisão sistemática, houve o acréscimo de algumas tarefas, para permitir uma análise mais completa. De forma ortogonal à questão de pesquisa QP1, esta questão busca identificar quais destas tarefas de mineração de dados têm sido aplicadas especificamente para o contexto de mineração de processos.

QP3. Quais técnicas são usualmente aplicadas para implementar as tarefas de mineração de dados, no contexto de mineração de processos? Quais destas técnicas são de inteligência computacional ou de aprendizado de máquina?

A fim de tratar essa pesquisa pergunta, foi considerado um vasto conjunto de técnicas normalmente usadas para resolver diferentes tarefas de mineração de dados. O conjunto de técnicas inclui: (i) técnicas de inteligência computacional, incluindo redes neurais, máquinas de vetores de suporte, lógica *fuzzy*, e computação evolutiva; (ii) técnicas de aprendizado de máquina, incluindo aprendizado supervisionado e aprendizado não supervisionado; e (iii) técnicas tradicionais, incluindo: técnicas baseadas em agentes; derivadas do algoritmo α ; técnicas baseadas em casos; técnicas baseadas em distância; técnicas baseadas em estruturas de grafos; técnicas baseadas em heurística; técnicas baseadas em lógica; operações OLAP; probabilidade e estatística; técnicas baseadas em regras; técnicas baseadas em semântica; e técnicas baseadas em dados temporais. Complementarmente à questão de pesquisa QP2, o propósito desta questão é identificar quais são as técnicas aplicadas especificamente ao contexto de mineração de processos. Na mesma linha de pensamento da questão de pesquisa anterior, poderia se esperar que todas as essas técnicas comumente aplicadas em mineração de dados também estivessem sido aplicadas à mineração de processos, principalmente aquelas de campos de inteligência computacional e de aprendizado de máquina, uma vez que elas têm apresentado bons resultados para uma série de problemas de mineração de dados em geral.

3.2.2 Fontes de dados e estratégia de busca

Similarmente ao protocolo seguido na revisão sistemática, as fontes de dados do “Scopus” e “ISI Web of Science” foram usadas. A *string* de busca definida (Tabela 6) visa encontrar trabalhos que sejam relacionados a mineração de processos de forma geral, independentemente do tipo de técnica usada. Essa *string* foi aplicada pela primeira vez em 2013 e reaplicada no início de 2015.

Tabela 6 – *String* de busca genérica para o mapeamento sistemático

(“mining process” OR “process mining” OR “workflow mining” OR “workflows mining” OR “mining process” OR “mining processes” OR “mining workflow” OR “mining workflows” OR (“business process” OR “business processes” OR workflow OR workflows) AND “data mining”))

Fonte: Ana Rocío Cárdenas Maita, 2015

3.2.3 Seleção de trabalhos

Foram especificados critérios de inclusão e de exclusão baseados no escopo e na qualidade dos artigos encontrados para garantir que, ao final, fosse selecionados apenas trabalhos realmente relacionados ao contexto desejado. Tais critérios são muito similares aos usados para a revisão sistemática, conforme apresentados na seção 3.1.3. A seguir apenas as diferenças específicas para esse mapeamento sistemático são apresentados.

3.2.3.1 Critérios de inclusão

Os seguintes critérios de inclusão, incluídos na seção 3.1.3.1, também foram incluídos aqui de forma idêntica: **CI-1**, **CI-2** e **CI-3**. O seguinte critério de inclusão, incluídos na seção 3.1.3.1, não foi incluído aqui: **CI-4**.

3.2.3.2 Critérios de exclusão

Os seguintes critérios de exclusão, incluídos na seção 3.1.3.2, também foram incluídos aqui de forma idêntica: **CE-1**, **CE-2**, **CE-3**, **CE-4**, **CE-5** e **CE-6**. O seguinte critério de exclusão, não incluído na seção 3.1.3.2, foi incluído para este caso:

- **CE-7:** o artigo foi classificado pelo veículo de publicação dentro da categoria “artigo curto” (*short paper*); ou o artigo foi classificado pelo veículo de publicação, explicitamente, como: *work in progress*, *position paper*, *doctoral consortium*, *doctoral workshop*, *keynote*, ou *poster*.

4 Resultados alcançados

A partir do estudo realizado para a área mineração de processos, considerando uma análise direcional considerado vários pontos de vista para uma melhor delimitação e entendimento das pesquisas atuais na área, quatro pontos de vista são discutidos neste trabalho: *(i)* quais técnicas e tarefas originárias da área de mineração de dados são as mais usadas, quando aplicadas para resolver problemas de mineração de processos; *(ii)* quais tipos de mineração de processos atualmente são os mais usados; *(iii)* o que representa a quantidade de técnicas provenientes das áreas de inteligência computacional e de aprendizado de máquina, especificamente redes neurais e máquinas de vetores de suporte, em proporção com outras técnicas tradicionais, em mineração de processos; *(iv)* quais veículos para publicação, países, pesquisadores e universidades estão apresentando maior interesse em estudar aspectos relacionados em essa área. Considerando que os conceitos tratados nos itens *(i)*, *(ii)* e *(iii)* estão estreitamente relacionados, uma análise cruzada entre eles foi necessária para uma melhor apresentação dos resultados.

Este capítulo está organizado da seguinte forma: na seção 4.1 e 4.2 os resultados da revisão sistemática (revisão específica) e do mapeamento sistemático (revisão expandida) são apresentados; a seção 4.3 apresenta um exemplo de aplicação em mineração de processos; por fim, as lições aprendidas com esses três tipos de resultados são apresentadas na seção 4.4.

4.1 Revisão sistemática (específica)

A revisão foi conduzida em duas etapas principais: *(i)* identificação e seleção dos estudos primários bem como a avaliação da qualidade dos estudos primários; e *(ii)* extração e síntese dos dados dos estudos primários. Esta seção apresenta um resumo dos resultados alcançados na execução de tais etapas. Os resultados completos dessa revisão sistemática, descritos com maior detalhe, encontram-se no artigo que já foi publicado Maita et al. (2015).

4.1.1 Resultados da revisão sistemática

A primeira etapa, de identificação e seleção dos estudos primários, foi realizada seguindo o método sugerido por Chen e Babar (2011), em três passos *(i)* identificação dos registros relevantes nas bases de dados com a aplicação da *string* de busca, tendo sido obtidos 188 resultados; *(ii)* aplicação dos critérios de inclusão e de exclusão propostos no protocolo da revisão, resultando na seleção de 11 estudos primários; *(iii)* aplicação dos critérios de avaliação da qualidade: relatório, rigor, credibilidade e relevância para a área, (cf. 5). As etapas desta aplicação são apresentadas na figura 7.

Figura 7 – Passos seguidos na identificação de estudos primários

Fonte: Traduzido de Maita et al. (2015)

A tabela 7 apresenta a lista dos 11 estudos primários, resultantes da aplicação dos critérios de inclusão e exclusão, descritos no protocolo da revisão. Os dados apresentam que sete dos 11 estudos primários foram publicados em periódicos, e quatro em conferências.

Os resultados da avaliação de qualidade para esses 11 estudos primários estão apresentados na tabela 8. Os dados apresentam a avaliação dos estudos primários selecionados tendo em conta os atributos definidos na tabela 5. Cada sub-atributo foi avaliado e pontuado como “Completo – 1,0 ponto”, “Parcial – 0,5 ponto” ou “Nada – 0 ponto”. Os estudos foram melhor avaliados em termos dos atributos “relatórios” e “relevância”, com taxas relativas de 91 e 82 por cento, respectivamente. Em contraste, foram pior avaliada

Tabela 7 – Lista final de estudos primários para revisão sistemática

Ano	ID	Título	Tipo de publicação
2004	J-1	<i>An intelligent information infrastructure to support the streamlining of integrated logistics workflow</i>	Periódico
2006	J-2	<i>An intelligent production workflow mining system for continual quality enhancement</i>	Periódico
	C-1	<i>Development of a distributed process mining system for reactive ion etching enhancement</i>	Conferência
	C-2	<i>A fuzzy paradigm approach for business process intelligence</i>	Conferência
2011	J-3	<i>KNNI-SVM: A hybrid algorithm integrating imputation and support vector machine for real-time business process monitoring</i>	Periódico
	C-3	<i>Process optimization of candy production based on data mining</i>	Conferência
2012	J-4	<i>Mining event logs to support workflow resource allocation</i>	Periódico
2013	J-5	<i>A comparative study of dimensionality reduction techniques to enhance trace clustering performances</i>	Periódico
	C-4	<i>Data mining for “big archives” analysis: A case study</i>	Conferência
2014	J-6	<i>Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the CloudFlows platform</i>	Periódico
	J-7	<i>A dynamic understanding of customer behavior processes based on clustering and sequence mining</i>	Periódico

Fonte: Traduzido de Maita et al. (2015)

em termos dos atributos “credibilidade” e “rigor”, com taxas relativas de 61 e 56 por cento, respectivamente.

A figura 8 apresenta a distribuição temporal dos 11 estudos primários (no período de 2004 até 2014). Verifica-se uma concentração dos estudos publicados (64%) nos últimos quatro anos do período avaliado.

Figura 8 – Distribuição dos estudos primários por ano

Fonte: Traduzido de Maita et al. (2015)

A análise dos 11 estudos primários selecionados resultou nas classificações apresentadas a seguir, especificamente relacionadas à avaliação das questões de pesquisa.

Tabela 8 – Lista final de estudos primários para revisão sistemática

ID	Relatório			Rigor					Credi- bilidade		Rele- vância	Índice de qualidade
	Rel.1	Rel.2	Rel.3	Rig.1	Rig.2	Rig.3	Rig.4	Rig.5	Cre.1	Cre.2	Rele.1	
J-1	C	C	C	C	P	N	P	P	P	C	C	8,0
J-2	C	C	P	C	C	N	C	P	P	C	C	8,5
C-1	C	P	P	C	N	N	C	P	N	P	P	5,5
C-2	C	C	P	P	N	N	P	N	N	P	P	4,5
J-3	C	C	C	C	N	N	P	N	N	P	P	5,5
C-3	C	P	P	P	N	N	P	N	N	P	P	4,4
J-4	C	C	C	C	C	N	C	C	P	C	C	9,5
J-5	C	C	C	C	C	N	C	C	C	C	C	10,0
C-4	C	C	C	C	C	N	C	C	C	C	C	10,0
J-6	C	C	C	C	P	N	C	P	P	C	C	8,5
J-7	C	C	C	C	C	N	C	C	P	C	C	9,5

Notas: C – completo (1,0 ponto); P – parcial (0,5 ponto); N – nenhum (0 pontos)

Fonte: Traduzido de Maita et al. (2015)

QP1. Que tipos de redes neurais ou de máquinas de vetores de suporte são usados para implementar as tarefas de mineração de dados quando aplicados ao contexto de mineração de processos?

As tabelas 9, 10 e 11 apresentam um resumo do resultado da extração e da classificação dos dados dos 11 estudos primários, escopo da revisão sistemática, com o objetivo de responder essa primeira questão de pesquisa. Em relação às abordagens identificadas nos estudos primários que usam redes neurais, os dados da tabela 9 e da tabela 11 apresentam que a maioria delas apresentam arquitetura multicamada com aprendizado supervisionado. Já em relação às abordagens identificadas nos estudos primários que usam máquinas de vetores de suporte, os dados da tabela 10 e da tabela 11 apresentam que todos os trabalhos usam o modelo básico de máquinas de vetores de suporte para classificação com aprendizado supervisionado.

QP2. Quais tarefas de mineração de dados são direcionadas em mineração de processos quando aplicadas a redes neurais e máquinas de vetores de suporte?

Em relação às tarefas de mineração de dados, a tabela 12 mostra que para as abordagens identificadas nos estudos primários selecionados para esta revisão sistemática, três das cinco categorias são cobertas quando se considera a sua resolução com redes neurais e máquinas de vetores de suporte. Assim, foram encontrados representantes para as seguintes tarefas: previsão categórica (ou classificação), previsão numérica (do tipo “regressão”), e análise de agrupamentos. Nenhum estudo foi identificado para as outras

Tabela 9 – Classificação dos estudos primários com base no tipo de arquitetura da redes neurais

ID	<i>Feedforward</i> com camada única	Recorrente com camada única	<i>Feedforward</i> multicamada	Recorrente multicamada	Camada competitiva
J-1			X		
J-2			X		
C-1			X		
C-2			X		
C-3			X		
J-5					X
J-7					X

Fonte: Traduzido de Maita et al. (2015)

Tabela 10 – Classificação dos estudos primários considerando as variações de máquinas de vetores de suporte

ID	Vetores suporte para classificação	Regressão por vetores suporte	Agrupamento por vetores suporte
J-3	X		
J-4	X		
C-4	X		
J-6	X		

Fonte: Traduzido de Maita et al. (2015)

Tabela 11 – Classificação dos estudos primários considerando os tipos de aprendizado de redes neurais e de máquinas de vetores de suporte

ID	Aprendizado supervisionado	Aprendizado não supervisionado	Aprendizado de reforço	Redes neurais de pesos fixos
J-1	X			
J-2	X			
C-1	X			
C-2	X			
C-3	X			
J-3	X			
J-4	X			
J-5		X		
C-4	X			
J-6	X			
J-7		X		

Fonte: Traduzido de Maita et al. (2015)

duas categorias: previsão numérica (do tipo “identificação de tendências”), e descoberta de padrões frequentes, regras associação e correlação. De fato, dois estudos envolvem a tarefa de descoberta de padrões frequentes, mas são resolvidos com outras técnicas, assim, não foram levadas em conta na tabela 12.

Tabela 12 – Classificação dos estudos primários considerando as tarefas de mineração de dados

ID	Predição categórica (classificação)	Predição numérica Regressão	Identificação de tendências	Análise de agrupamentos	Descoberta de padrões frequentes regras de associação e correlação
J-1	X				
J-2		X			
C-1		X			
C-2	X				
C-3	X				
J-3		X			
J-4	X				
J-5			X		
C-4	X				
J-6	X				
J-7			X		

Fonte: Traduzido de Maita et al. (2015)

QP3. Quais tipos de mineração de processos são atendidos por redes neurais ou máquinas de vetores de suporte quando tais técnicas são usadas para tratar tarefas de mineração de dados neste contexto?

Em relação ao tipo de mineração de processos, cujo objetivo é verificar as aplicações de redes neurais e máquinas de vetores de suporte na área de interesse, os dados da tabela 13 apresentam que dois estudos se referem ao tipo descoberta, três ao tipo aprimoramento, e seis ao tipo conformidade. Embora essa classificação tenha sido realizada considerando as definições apresentadas por Aalst (2011), de fato, a maioria dos trabalhos não seguem estritamente as definições propostas por esse especialista para a área de mineração de processos. Dessa forma, a classificação apresentada na tabela 13 se refere a uma aproximação de essas propostas por Aalst (2011), levando em consideração a análise realizada pelos autores deste estudo.

Tabela 13 – Classificação dos estudos primários em relação ao tipo de mineração de processos

ID	Descoberta	Conformidade	Aprimoramento
J-1		X	
J-2			X
C-1			X
C-2			X
C-3		X	
J-3			X
J-4			X
J-5	X		
C-4			X
J-6		X	
J-7	X		

Fonte: Ana Rocío Cárdenas Maita, 2015

4.1.2 Análise dos resultados da revisão sistemática

Com base nas informações apresentadas na seção anterior, algumas análises foram realizadas, cujos resultados podem ser úteis para a comunidade de pesquisadores interessados no uso de redes neurais ou máquinas de vetores de suporte para realizar mineração de processos.

Primeiro, em relação ao número de estudos primários identificados, ainda há um número limitado de estudos sendo conduzidos com o objetivo de aplicar redes neurais e máquinas de vetores de suporte em mineração de processos. Considerando que aproximadamente 620 trabalhos foram inicialmente identificados na área de mineração de processos, os 11 estudos finais selecionados representam apenas 2% de todo o universo de estudo. Com respeito a possíveis em relação à distribuição durante o período de 11 anos avaliado como se mostra na figura 8. Verifica-se que dois terços dos estudos foram publicados nos últimos quatro anos, depois de um período de quatro anos sem qualquer publicação, o que pode indicar uma aparente retomada de pesquisa e interesse nesta área. Além disso, nos estudos avaliados, os autores relatam geralmente impressões otimistas em relação ao uso destas técnicas no contexto de mineração de processos, incluindo os salientes resultados numéricos, apoiando a produção de conhecimento útil para as áreas de negócio envolvidos nos estudos. No entanto, considerando o ponto de vista da área de inteligência computacional, uma análise dos testes e métricas de estratégias para avaliar os resultados quantitativos apresentados nestes estudos revela um grau de imaturidade

Figura 9 – Classificação resumida dos estudos primários considerando tipos de mineração de processos contra tipos de tarefas de mineração de dados e técnicas usadas

Fonte: Ana Rocío Cárdenas Maita, 2015

dessas obras na padronização dos métodos de avaliação e da disponibilidade de referências de conjuntos de dados que permitam a reprodutibilidade dos estudos.

A fim de apoiar as análises e discussão além das apresentadas aqui, a tabela 14 apresenta um resumo dos dados anteriormente apresentados nas tabelas 8 e 13. Ela permite visualizar de forma integrada todos os 11 estudos primários considerando todos os aspectos avaliados juntos, como definidas por meio das três questões de investigação, ou seja: tipos de mineração de processos, tipos de tarefas de mineração de dados, e os detalhes das técnicas usadas. Por uma questão de simplicidade a tabela 14 apresenta apenas os atributos e as combinações de atributos existentes apresentadas nas tabelas anteriores. Também com o objetivo de resumir todos os resultados em conjunto, a figura 9 apresenta uma análise cruzada considerando os tipos de mineração de processos contra os tipos de tarefas de mineração de dados e algumas informações sobre a técnica usada em cada estudo preliminar. As obras identificadas nesta revisão sistemática compartilham um padrão em relação aos itens que nós pesquisamos por meio das perguntas de pesquisa específicas. A maioria dos trabalhos são estudos que realizam alguma forma de aprimoramento de mineração de processos, principalmente usando redes neurais. As análises a seguir, apresentam para cada uma das questões de pesquisa investigadas nesta revisão sistemática, bem como alguns análises cruzados da revisão sistemática.

Em relação ao tipo de mineração de processos tratados nos estudos primários avaliados, de acordo com a tabela 14 e à figura 9, aproximadamente 55% dos artigos referem-se ao tipo “Aprimoramento”, seguido por cerca de 27% que se referem ao tipo “Conformidade”, e cerca de 18% referem-se ao tipo “Descoberta”. Embora esta classificação leva em consideração as definições propostas por Aalst (2011), a maioria dos estudos primários identificados nesta revisão sistemática não se aderiram totalmente aos conceitos apresentados por este especialista em mineração de processos. Por exemplo, apenas quatro dos 11 estudos primários referem explicitamente ao termo mineração de processos (J-4, J-5, C-1, C-2) e um deles refere-se explicitamente ao mineração de *workflow* (J-2). Assim, pouco mais da metade dos estudos primários identificados foram avaliados por referir-se à aplicação de mineração de dados no contexto de processos de negócio (ou *workflow*), o que foi entendido indiretamente como mineração de processos. Além disso, apenas cinco dos 11 estudos mencionam explicitamente o uso de *log* de eventos (com diferentes termos) na sua abordagem (J-2, J-3, J-4, J-5, C-1). Isso significa que pouco mais da metade dos estudos primários não são identificados com base na análise de *log* de eventos para realizar tarefas de mineração de dados e podem não representar exatamente o que é definido como mineração de processos. Finalmente, apenas cinco estudos explicitamente lidam com o modelo de processos ou modelo de *workflows*, que é considerado base para os três tipos de mineração de processos de acordo com Aalst (2011) (J-4 J-5, J-6, C-2, C-3). No geral, apenas dois estudos atendem a essas três características avaliadas em conjunto (J-4 e J-5) (ou seja, 22%), enquanto três estudos não cumprem quaisquer de tais características (J-1, J-7, C-4) (ou seja, 33%).

Em relação aos tipos de redes neurais e máquinas de vetores de suporte usados para implementar tarefas de mineração de dados, quando especificamente aplicadas no contexto de mineração de processos, os dados apresentados mostram que:

- Enquanto 64% dos estudos usam redes neurais, apenas 36% deles usam máquinas de vetores de suporte. Redes neurais são mais populares do que máquinas de vetores de suporte em diversas áreas de aplicação, dado que as máquinas de vetores de suporte começaram a ser estudadas, de forma mais intensa, somente em meados de 1990. Redes neurais, no entanto, têm sido estudada desde a década de 1940, embora na maior parte da década de 1950 elas receberam muito pouca atenção do meio acadêmico. Dado que as máquinas de vetores de suporte tornaram-se populares

muito recentemente, é perfeitamente normal que exista uma lacuna em seu uso em comparação com redes neurais nas áreas de aplicação, incluindo mineração de processos, embora as duas técnicas apresentem resultados semelhantes.

- Em relação aos estudos que usam redes neurais, 71% usam a arquitetura *feedforward* multicamadas com aprendizado supervisionado e apenas 29% usam a arquitetura de camada competitiva com o aprendizado não supervisionado. O uso desses tipos de redes neurais é esperado, visto que eles são os mais populares na mineração de dados, principalmente devido a seu potencial para a resolução de problemas, desse modo, uma maior disponibilidade de ferramentas computacionais de suporte a aplicações. No entanto, as arquiteturas de *feedforward* de camada única, recorrentes de camada única, recorrentes multicamada, e também tipos de redes de aprendizado por reforço e de pesos fixos não são usadas em nenhum dos estudos primários selecionados. Isso também era esperado, uma vez que, mesmo na área de mineração de dados em geral, esses tipos de redes neurais aparecem com menor frequência, ou porque eles são muito simples e limitados – como a arquitetura de camada única, ou porque eles são só adequados para o tratamento tipos de problemas variados – como o aprendizado por reforço, altamente aplicáveis em simuladores e robótica.
- Em relação aos estudos que usam máquinas de vetores de suporte, todos eles usam a classificação básica do modelo de máquinas de vetores de suporte para classificação (com aprendizado supervisionado); isto é, as variações de “classificação por vetores suporte” e “regressão por vetores suporte” não são usadas em qualquer estudo. Isto é provavelmente porque redes neurais são mais populares, desde que “classificação por vetores suporte” e “redes neurais” são igualmente aplicáveis para contextos de agrupamento, como máquinas de vetores de suporte e para contextos de regressão. De qualquer forma, considerando um total de apenas 11 estudos primários, a falta de estudos que usam “classificação por vetores suporte” e “regressão por vetores suporte” não representa necessariamente um padrão para área de mineração no processo, mas sim apenas um indicativo.

Com relação às tarefas de mineração de dados que estão sendo tratados em estudos de mineração de processos que usam redes neurais ou máquinas de vetores de suporte, os dados apresentados mostram que: 55% dos estudos identificados tratam a tarefa predição categórica (ou classificação); 27% tratam a tarefa previsão numérica (considerando o tipo

de “regressão”); e 18% tratam a tarefa de análise de agrupamentos. A tarefa predição numérica (do tipo “identificação tendências”) não é referenciado em nenhum dos estudos. Em relação às tarefas de Descoberta de padrões frequentes, regras de associação e regras de correlação, dois estudos comentam sobre a resolução de tais tarefas, no âmbito de uma abordagem mais global, combinada com a previsão categórica (ou classificação), em um dos estudos; e com a análise de agrupamentos em outro estudo. No entanto, nesses casos, essa tarefa não está diretamente ligada às atividades que as redes neurais ou máquinas de vetores de suporte resolvem, portanto, não se inserem no âmbito da análise desta revisão sistemática.

A análise cruzada apresentada na tabela 14 e figura 9 mostrou que dos seis estudos que lidam com a tarefa predição categórica (ou classificação), quatro usam máquinas de vetores de suporte, enquanto apenas dois usam redes neurais. Considerando a tarefa de predição numérica (do tipo “regressão”) todos os estudos usam redes neurais. Dado que as máquinas de vetores de suporte estão se tornando cada vez mais populares e, tradicionalmente, o conhecimento das variações “classificação por vetores suporte” e “regressão por vetores suporte” ocorre somente mais tarde, é provável que os estudos na área de mineração o processo também começassem a usar “classificação por vetores suporte” e “regressão por vetores suporte”, em vez de redes neurais, para estas duas tarefas que até agora têm sido tratados apenas com redes neurais. Em geral, máquinas de vetores de suporte muitas vezes provou ser superior a redes neurais, considerando o tratamento do risco estrutural, além do risco empírico que gera potencialmente resultados mais precisos do que tratar apenas o risco empírico, como realizada pelas redes neurais.

Em geral, cerca de 80% dos estudos primários avaliados nesta revisão sistemática apresentaram resultados positivos, com benefícios de mineração de processos, como o aumento da produtividade e eficiência nos processos de negócio. No entanto, nem todos estes resultados positivos podem ser atribuídos ao uso de redes neurais e máquinas de vetores de suporte, uma vez que a maioria dos estudos sugerem abordagens híbridas que combinam uma série de diferentes técnicas e tecnologias.

Tabela 14 – Classificação resumida dos estudos primários considerando todos os aspectos avaliados juntos

(Tipo de mineração de processos)	Descoberta	Conformidade		Aprimoramento			Índice de qualidade
				Predição Categórica (ou classificação)	Predição numérica tipo “Regressão”	Predição Categórica (ou classificação)	
		(Tipo de mineração de dados)	Análise de agrupamentos	Predição categórica (ou classificação)	Aprendizado supervisionado		
(Tipo de aprendizado)	Aprendizado não supervisionado	Aprendizado supervisionado		Aprendizado supervisionado			
(Tipo de técnica)	Redes neurais artificiais	Redes neurais artificiais	Máquinas de vetores suporte	Redes neurais artificiais	Máquinas de vetores suporte	Máquinas de vetores suporte para classificação	
(Arquitetura de redes neurais /variação de máquinas de vetores suporte)	Camada competitiva	<i>Feedforward</i> multicamada	Máquinas de vetores suporte para classificação	<i>Feedforward</i> multicamada	Máquinas de vetores suporte para classificação	Máquinas de vetores suporte para classificação	
J-1		X			X		8.0
J-2					X		8.5
C-1					X		5.5
C-2						X	4.5
C-3			X			X	5.5
J-3						X	4.0
J-4							9.5
J-5	X						10.0
C-4			X			X	10.0
J-6			X				8.5
J-7	X						9.5

Fonte: Ana Rocío Cárdenas Maíta, 2015

De acordo com a avaliação dos estudos primários em relação à sua qualidade, como mostra a tabela 8, estes estudos foram geralmente bem avaliados, especialmente em relação aos critérios de “relatório” e “relevância”. No entanto, apenas artigos publicados nos últimos três anos foram completamente bem avaliados em relação à maioria dos critérios, incluindo também “rigor” e “credibilidade”, com exceção de um estudo publicado em 2006, que em sete dos critérios foi bem avaliado. Os documentos menos recentes referem-se a estudos mais básicos, e que não apresentam claramente a aplicação, testes e verificação dos resultados obtidos com as abordagens propostas. Estes concentram a maior parte dos estudos publicados em anais de congressos, o que poderia explicar um pouco a avaliação desfavorável em termos de qualidade. Além disso, os documentos mais antigos referem-se ao período em que a mineração de processos estava ainda sendo divulgada, com o trabalho de pesquisa focado em mostrar o seu potencial. A boa avaliação sobre a qualidade dos trabalhos recém-publicados mostra que os pesquisadores estão mais preocupados com a contribuição real das redes neurais e máquinas de vetores de suporte no contexto de mineração de processos.

4.2 Mapeamento sistemático (geral)

O mapeamento sistemático foi conduzido de forma similar à revisão sistemática apresentada na seção anterior, mas em um nível de granularidade maior por ser um estudo mais amplo e menos detalhado.

4.2.1 Resultados do mapeamento sistemático

Estudos primários foram identificados também com base na estratégia estabelecida por Chen e Babar (2011). Tal estratégia consiste em três etapas principais: *(i)* identificação dos estudos relevantes em bancos de dados de pesquisa (resultado = 3.962 registros encontrados); *(ii)* aplicação dos critérios de inclusão e de exclusão (resultado = 705 estudos primários selecionados); e *(iii)* avaliação da qualidade dos estudos primários (resultado = 705 estudos primários avaliados).

Em termos gerais, a figura 10 apresenta uma distribuição temporal dos 705 estudos primários relacionados a mineração de processos (de 2005 a 2014), classificados por tipo de

publicação (periódico ou conferência). Destaca-se um aumento considerável e progressivo no número de publicações tanto em periódicos quanto em conferências.

Figura 10 – Distribuição de estudos primários por ano (Mapeamento sistemático)

Fonte: Ana Rocío Cárdenas Maita, 2015

Primeiramente, uma análise foi realizada especificamente para a conferência ou periódico em que cada um dos 705 estudos primários selecionados, procurando a existência de veículos que, devido a sua natureza, poderiam ter a publicação de um maior número de trabalhos relacionados a mineração de processos. Considerando conferências, 243 diferentes veículos foram responsáveis por publicar a totalidade de 473 trabalhos publicados em anais de eventos. Por outro lado, considerando periódicos, 117 diferentes veículos foram responsáveis por publicar a totalidade de 232 estudos primários publicados em periódicos. Os dados detalhados relacionados com a conferência ou periódico específico em que foram publicados esses estudos primários são os seguintes:

- Dezesete conferências publicaram pelo menos cinco estudos primários cada um, conforme apresentado na tabela 15. Além disso, estas 17 conferências mais populares foram em conjunto responsáveis pela publicação de 183 documentos (que representam 39% dos 473 trabalhos publicados em anais de congressos, isto é, quase a metade deles). Os restantes 290 estudos foram espalhados entre 226 conferências diferentes.
- Sete revistas publicaram pelo menos cinco estudos primários cada uma, conforme apresentado na tabela 16. Além disso, essas sete principais revistas mais populares em conjunto foram responsáveis pela publicação de 72 artigos (que representam 31%

dos 232 artigos publicados em revistas, isto é, cerca de um terço deles). O resto, ou seja, 160 estudos foram espalhados entre 110 revistas diferentes.

Tabela 15 – Dezesete conferências com maior número de publicações em mineração de processos nos últimos dez anos

Pos	Iniciais	Título da Conferência	# de estudos primários
1	BPM	<i>International Conference on Business Process Management</i>	61
2	CAiSE	<i>International Conference on Advanced Information Systems Engineering</i>	16
3	SIMPDA	<i>International Symposium on Data-Driven Process Discovery and Analysis</i>	11
4	Petri Nets	<i>International Conference on Application and Theory of Petri Nets and Concurrency</i>	10
5	CIDM	<i>IEEE Symposium on Computational Intelligence and Data Mining</i>	9
6	SAC	<i>ACM Symposium on Applied Computing</i>	8
6	BPMDs	<i>International Workshop on Business Process Modeling, Development and Support</i>	8
8	CEC	<i>IEEE Congress on Evolutionary Computation</i>	7
8	EDOC	<i>IEEE International Enterprise Distributed Object Computing Conference</i>	7
8	ICT&KE	<i>International Conference on ICT and Knowledge Engineering</i>	7
11	BIS	<i>International Conference on Business Information Systems</i>	6
11	BPI	<i>International Workshop on Business Process Intelligence</i>	6
11	ICDM	<i>IEEE International Conference on Data Mining</i>	6
11	ICEIS	<i>International Conference on Enterprise Information Systems</i>	6
15	CSCWD	<i>IEEE International Conference on Computer Supported Cooperative Work in Design</i>	5
15	ICEBE	<i>IEEE International Conference on e-Business Engineering</i>	5
15	PAKDD	<i>Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining</i>	5

Fonte: Ana Rocío Cárdenas Maita, 2015

Ainda em termos gerais, por meio de uma verificação manual nos dados de autoria de cada estudo, encontrou-se que os 705 estudos primários selecionados nesse mapeamento sistemático foram produzidos por 1.275 autores diferentes, de 469 instituições de pesquisa diferentes, localizadas em 61 países diferentes. Esses números representam o quanto generalizada é a pesquisa relacionada a mineração de processos. Uma análise foi realizando levando-se em conta apenas os mais produtivos, para permitir uma análise mais

Tabela 16 – Sete periódicos com maior número de publicações em mineração de processos nos últimos dez anos

Pos	Iniciais	Título do periódico	# de estudos primários
1	ESA	<i>Expert Systems with Applications</i>	19
2	JIS	<i>Information Systems</i>	17
3	DKE	<i>Data and Knowledge Engineering</i>	9
4	DSS	<i>Decision Support Systems</i>	8
4	IJBPM	<i>International Journal of Business Process Integration and Management</i>	8
6	TKDE	<i>IEEE Transactions on Knowledge and Data Engineering</i>	6
7	SMC	<i>IEEE Transactions on Systems, Man, and Cybernetics</i>	5

Fonte: Ana Rocío Cárdenas Maita, 2015

aprofundada de onde o maior número de trabalhos nas áreas de mineração de processos têm sido desenvolvidos. Os seguintes dados foram levantamos:

- Os principais países na publicação de estudos primários neste contexto, contribuíram com pelo menos 50 estudos cada um, e em conjunto contribuíram com a produção de 62% do total de 705 estudos primários selecionados (ou seja, quase dois terços do total), conforme apresentado na tabela 17.
- As doze principais instituições de pesquisa, que contribuíram com a produção de pelo menos dez estudos primários cada uma, como apresentado na tabela 18, contribuíram em conjunto de 34% do total de 705, ou seja, aproximadamente um terço do total.
- Os dezesseis principais pesquisadores, que contribuíram com a produção de, pelo menos, dez estudos primários cada um, como apresentado na tabela 19, contribuíram em conjunto com a produção de 29% do total de 705 estudos primários selecionados, ou seja, aproximadamente um quarto do total.

Nos três casos anteriormente mencionados, um mesmo trabalho pode ser produzido por autores de diferentes instituições ou mesmo diferentes países. Assim, mais de uma instituição ou mais de um país pode ter contribuído para sua produção.

As seções a seguir apresentam os resultados relacionados às respostas das questões de pesquisa estruturados nessa ordem.

Tabela 17 – Dezesete países com maior número de publicações em mineração de processos nos últimos dez anos

Pos	País	# de estudos primários
1	Holanda	147
2	China	125
3	Alemanha	68
4	Estados Unidos de América	67
5	Itália	61
6	Reino Unido	39
7	Austrália	36
8	Coreia do Sul	35
9	Espanha	34
10	Bélgica	33
11	Áustria	22
11	Brasil	22
11	França	22
14	Hong Kong	16
14	Portugal	16
16	Índia	15
17	Tailândia	10

Fonte: Ana Rocío Cárdenas Maita, 2015

Tabela 18 – Doze instituições de pesquisa com maior número de publicações em mineração de processos nos últimos dez anos

Pos	Instituição de pesquisa	País	# de estudos primários
1	Univ. Tecnológica de Eindhoven	Holanda	129
2	Univ. Católica de Leuven	Bélgica	26
3	Univ. Politécnica da Catalunha	Espanha	20
4	Univ. de Tsinghua	China	18
5	Univ. Tecnológica de Queensland	Austrália	16
5	Univ. da Calábria	Itália	16
7	Inst. de Computação de Alto Desempenho e Networking	Itália	15
8	Univ. Técnica de Lisboa	Portugal	14
8	Univ. de Southampton	Reino Unido	14
10	Univ. de Twente	Holanda	12
11	Univ. Politécnica de Hong Kong	China	10
12	Univ. de Ulm	Alemanha	10

Fonte: Ana Rocío Cárdenas Maita, 2015

Tabela 19 – Dezesseis pesquisadores com maior número de publicações em mineração de processos nos últimos dez anos

Pos	Pesquisador	Instituição de pesquisa	País	# de estudos primários
1	W. M. P. Aalst	Univ. Tecnológica de Eindhoven	Holanda	105
2	B. Baesens	Univ. de Southampton	Reino Unido	20
2	J. Vanthienen	Univ. Católica de Leuven	Bélgica	20
4	B. F. van Dongen	Univ. Tecnológica de Eindhoven	Holanda	19
5	J. Carmona	Univ. Politécnica da Catalunha	Espanha	18
6	M. Song	Ulsan Inst. Nacional de Ciência e Tecnologia	Coréia do Sul	15
7	G. Greco	Univ. de Calábria	Itália	14
7	A. Guzzo	Inst. de Computação de Alto Desempenho e Networking	Itália	14
7	L. Pontieri	Inst. de Computação de Alto Desempenho e Networking	Itália	14
7	J. Wang	Univ. de Tsinghua	China	14
11	D. R. Ferreira	Univ. Técnica de Lisboa	Portugal	13
12	R. P. J. C. Bose	Univ. Tecnológica de Eindhoven	Holanda	12
12	L. Wen	Univ. de Tsinghua	China	12
14	A. Rozinat	Univ. Tecnológica de Eindhoven	Holanda	11
15	W. Gaaloul	Inst. Nacional de Pesquisa em Informática e Automação	França	10
15	A. K. A. Medeiros	Univ. Tecnológica de Eindhoven	Holanda	10

Fonte: Ana Rocío Cárdenas Maita, 2015

4.2.1.1 Tipos de mineração de processos atendidos por mineração de dados

De acordo com os dados apresentados na figura 11, evidências da aplicação dos três tipos de mineração de processos consideradas neste estudo foram coletadas. O tipo “Descoberta de processos” foi o mais comum, encontrado em 71% dos estudos primários; seguido pelo tipo “Conformidade de processos” encontrado em 37% deles; e, o tipo “Aprimoramento de processos” foi o menos comum, encontrado em apenas 17% deles.

A figura 12 apresenta o número de estudos publicados focados nos diferentes tipos de mineração de processos distribuídos por ano, considerando os últimos dez anos. Essa figura exhibe que, apesar de que a aplicação dos três diferentes tipos de mineração de processos incrementou-se durante tal período, o incremento ocorreu em diferentes momentos com maior intensidade para diferentes tipos de mineração. Por exemplo, o tipo “Descoberta de processos”, além de ser o tipo mais amplamente usado, tem aumentado consistentemente

Figura 11 – Tipos de mineração de processos, quando tratados individualmente

Fonte: Ana Rocío Cárdenas Maita, 2015

desde 2008, exceto para o último ano, em que os dados podem não estar completos ainda. Seguindo do tipo “Conformidade de processos”, o qual foi quase constante durante seis anos (2005-2010) e há apenas dois anos teve um incremento (2011-2012). Por fim, um ano mais tarde, isto é, apenas em 2012, o tipo “Aprimoramento de processos”, a menos aplicada, começou a apresentar um incremento depois de um período de sete anos (2005-2011) de praticamente um uso constante.

Figura 12 – Distribuição dos estudos primários considerando os tipos de mineração de processos identificados

Fonte: Ana Rocío Cárdenas Maita, 2015

O número total de estudos primários tanto na figura 12 quanto na figura 11 é maior que 705, pois em muitos casos o mesmo trabalho pode tratar dois ou até mesmo três diferentes tipos de mineração de processos. A figura 13 apresenta os números detalhados de 155 estudos primários nos quais aplicações de mais de um tipo de mineração de processos foram identificados, este número representa 22% dos 705 estudos em total, que aplicam

mais de um tipo de mineração de processos. A combinação mais comum é tratar os tipos “Descoberta de processos” e “Conformidade de processos” juntos, que foi encontrado em 90 estudos (ou seja, em 13% do total); as combinações menos comuns foram “Descoberta de processos” e “Aprimoramento de processos” juntas, encontradas em 24 estudos (ou seja, 3% do total), seguidas por “Conformidade de processos” e “Aprimoramento de processos”, encontrado em apenas 16 estudos (ou seja, 2% do total). A combinação dos três tipos de mineração de processos tratados em conjunto no mesmo estudo foi encontrada em apenas 24 estudos (ou seja, 3% do total).

Figura 13 – Tipos de mineração de processos, quando tratados juntos no mesmo estudo primário

Fonte: Ana Rocío Cárdenas Maita, 2015

4.2.1.2 Tarefas de mineração de dados em mineração de processos

Os dados apresentados na figura 14 representam evidências da aplicação das oito tarefas de mineração de dados consideradas neste estudo, e definidas na seção 2.2.1. Verifica-se que “predição categórica (ou classificação)” foi a tarefa mais usada, identificada em 25% dos estudos primários; “análise de dados estruturados” foi a segunda mais usada, identificada em 23% dos estudos primários; “análise de agrupamentos” foi a terceira mais usada, identificada em 21% dos estudos primários; “descoberta de padrões frequentes”, “regras de associação”, e “regras de correlação”, com 10%, 8%, e 3% dos estudos, respectivamente; seguidas pela tarefa mais comum é “análise de evolução”, identificadas em 18% dos estudos primários; seguidas por “análise de ruído e de *outliers*”, identificada em 10% dos estudos primários; finalmente, as tarefas menos comuns foram “predição numérica (do tipo

regressão)” e “predição numérica (do tipo identificação de tendências)”, identificados em 6% e 3% dos estudos primários, respetivamente. Finalmente, “Outras ou não identificadas”, contendo aqueles estudos onde não foi possível uma definição da tarefa usada com os dados informados pelo autor, isso representa apenas um 4% do total de estudos primários.

Figura 14 – Tarefas de mineração de dados identificadas nos estudos primários de mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

A figura 15 mostra o número de estudos primários considerando as diferentes tarefas de mineração de dados distribuídas por ano, durante a última década (com exceção apenas da categoria “Outras ou não identificadas”). Cada barra vertical, de cada tarefa de mineração de dados, representa um ano; isto é, dez anos em total para cada tarefa. Esta figura mostra um incremento nas aplicações de todas as diferentes tarefas de mineração de dados para a área específica de mineração de processos. Alguns pontos excepcionais podem ser observados tais como, por exemplo, um grande incremento na aplicação da tarefa “análise de agrupamentos”. De igual forma, observa-se para a tarefa “predição categórica (ou classificação)”, em 2010, e para “análise de dados estruturados”, em 2009.

A figura 16 representa a análise cruzada entre as tarefas de mineração de dados e os tipos de mineração de processos tratados, a fim de ver quais tarefas são mais aplicadas em determinados tipos de mineração de processos. Verifica-se que o maior número de ocorrências se refere ao tratamento da tarefa de predição categórica (ou classificação) para

Figura 15 – Distribuição por ano, no período de 2004 a 2014, de tarefas de mineração de dados aplicadas em mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

Figura 16 – Análise cruzada entre tipos de mineração de processos e tarefas de mineração de dados identificadas nos estudos primários

Fonte: Ana Rocío Cárdenas Maita, 2015

resolver a descoberta de processos de negócio, representando 15% do total de combinações. Esse número é bastante alto em comparação às combinações de menor ocorrência, ou seja, a tarefa análise de ruído e de *outliers* para resolver o aprimoramento de processos, ou a tarefa predição numérica (do tipo “regressão”) para resolver a conformidade de processos, que representam apenas 0,1% do total de combinações cada uma delas. Destacam-se outras combinações mais representativas, principalmente para o tipo de descoberta de processo, que é o tipo mais usado, em conjunto com as cinco primeiras tarefas de mineração mais usadas, na ordem do gráfico (segundo a mesma ordem da figura 14). Essas cinco maiores combinações em conjunto representam 60% do total de combinações identificadas neste gráfico; ou seja, representam mais da maioria.

Ainda em relação às combinações de tarefas de mineração de dados quando aplicadas na mineração de processos, a figura 17 mostra que as tarefas mais amplamente aplicadas puramente ou combinados com outras. Embora o uso de tarefas de mineração de dados combinados com outras tarefas tem uma média de cerca de 42%, essa taxa varia entre 55% (considerando a tarefa “análise de *Outliers*”) para 16% (considerando a tarefa “previsão numérica”).

Figura 17 – Tarefas de mineração de dados mais amplamente usadas em combinação com outras tarefas quando aplicadas em um mesmo estudo primário mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

4.2.1.3 Técnicas usadas para tratar as tarefas de mineração de dados no contexto de mineração de processos

Para avaliar as técnicas usadas nos estudos primários selecionados, elas foram agrupadas em função do princípio fundamental usado na solução da tarefa de mineração. Devido à grande variedade de abordagens usadas, assim como de variações dos algoritmos propostos pelos autores, para essa classificação, houve uma forte necessidade do auxílio dos conhecimentos da Profa. Dra. Sarajane Peres, quem é a coorientadora deste trabalho e especialista na área de mineração de dados.

A figura 18 representa as evidências de aplicação para 13 diferentes técnicas de inteligência computacional e de aprendizado de máquina; assim como, 12 técnicas tradicionais em mineração de dados consideradas em estudo e definidas na seção 2.2.3. O gráfico está dividido em duas partes: no lado esquerdo, as ocorrências das técnicas específicas das áreas de inteligência computacional ou aprendizado de máquina; e, no lado direito, as ocorrências das técnicas consideradas neste trabalho como tradicionais para este contexto. Note-se que um mesmo trabalho pode apresentar a ocorrência de mais do que uma técnica sendo usada para resolver uma ou mais tarefas de mineração de dados; assim, a quantidade total de técnicas é muito maior ao total de estudos primários.

No total, são 1.253 ocorrências de 25 classes de técnicas em 705 estudos primários, 234 de tais técnicas são relacionadas às áreas de inteligência computacional e de aprendizado de máquina, as quais representam um 19% do total encontrado; em contraste, 1019 técnicas são consideradas como tradicionais em mineração de dados, as quais representam um 71% do total encontrado. Além do mais, considerando as técnicas das áreas puramente de inteligência computacional e de aprendizado de máquina, apenas três delas – isto é, computação evolutiva, lógica *fuzzy* e árvores de decisão – conseguiram mais do que 5% do total encontrado. Entretanto, considerando as técnicas tradicionais de mineração de dados, a maioria delas, exceto por apenas três – isto é, “técnicas baseados em agentes”, “operações OLAP” e “técnicas baseadas em caso” - conseguiram mais de 5% do total encontrado.

Ainda sobre os diferentes tipos de técnicas a ser aplicadas, de acordo com os dados apresentados na figura 19, do número total de estudos primários selecionados no mapeamento realizado 705, apenas 12% deles usam puramente (uma ou mais) técnicas de inteligência computacional de aprendizado de máquina, exclusivamente. Por outro lado, 72% deles usam puramente (uma ou mais) as técnicas tradicionais de mineração de

Figura 18 – Técnicas em mineração de dados aplicadas identificadas nos estudos primários de mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

Figura 19 – Principais agrupamentos dos tipos de técnicas em mineração de dados aplicadas em mineração de processos identificadas nos estudos primários

Fonte: Ana Rocío Cárdenas Maita, 2015

Figura 20 – Técnicas de inteligência computacional e de aprendizado de máquina aplicadas em mineração de processos identificadas nos estudos primários

Fonte: Ana Rocío Cárdenas Maita, 2015

dados, exclusivamente. Por último, 16% dos estudos primários usam ambos os tipos de técnicas em um mesmo trabalho: (uma ou mais) técnicas de inteligência computacional ou aprendizado de máquina e (uma ou mais) técnicas de mineração de dados tradicionais. Ao todo, apenas 28% dos estudos primários usam pelo menos uma técnica de inteligência computacional ou o aprendizado da máquina; ou seja, 198 dos 705 estudos.

Especificamente em termos das 234 ocorrências de técnicas de inteligência computacional e de aprendizado de máquina, conforme apresentado na figura 18, a figura 20 apresenta uma visão detalhada segundo o tipo classificação principal usado; ou seja, inteligência computacional ou aprendizado de máquina, e neste último caso, também em termos de aprendizado, seja Aprendizado supervisionado ou Aprendizado não supervisionado. No total, existem mais de 234 ocorrências na figura 20 uma vez que três das técnicas são classificadas como técnicas de inteligência computacional e de aprendizado de máquina, ou seja, redes neurais, máquinas de vetores de suporte e SOM. Como resultado, os estudos que usam essas técnicas são contados duas vezes. Em linhas gerais, 140 estudos usam alguma técnica de inteligência computacional, ou seja, 55% do montante total, ao passo que 116 usam alguma técnica de aprendizado de máquina, ou seja, 45% do montante total (dos quais, 89 usam abordagens de aprendizado supervisionado e 27 usam aprendizado não supervisionado, isto é 35% e 11% do total, respectivamente).

A figura 21 mostra uma análise cruzada entre as técnicas de mineração de dados (aquelas de inteligência computacional e de aprendizado de máquina do lado esquerdo e as técnicas tradicionais do lado direito da imagem), quando aplicadas em cada um dos três tipos de mineração de processos. Em termos das combinações técnica usada com tipo de mineração de processos, as quatro combinações mais frequentes são: primeiro, “Técnicas

baseadas em estrutura de grafos” e o tipo “Descoberta de processos”, identificada em 195 estudos; segundo, “Técnicas baseadas em estrutura de grafos” e o tipo “Conformidade de processos”, identificada em 116 estudos; terceiro, “Probabilidade e estatística” e o tipo “Descoberta de processos”, identificada em 90 estudos; quarto, “Técnicas baseadas em dados temporais” e o tipo “Descoberta de processos”, identificada em 86 estudos; e quinto, “Técnicas baseadas em heurísticas” e o tipo “Descoberta de processos”, uma vez que juntas representam o 41% do total. Em contraste, as combinações menos usadas são “K-means” e “classificador Naive Bayes” quando aplicada no tipo “aprimoramento de processos”, apenas um artigo para cada caso, tratou essas combinações. Destaca-se que redes neurais é a uma técnica usada de forma quase similar tanto para Descoberta de processos como para Aprimoramento de processos, embora estudos da aplicação dela em Conformidade não foram encontrados.

Verifica-se que a proporção da figura 16 para tarefas e técnicas de mineração de processos, ainda se mantém, visto que o existe maior concentração de estudos de técnicas de mineração de dados aplicadas no tipo Descoberta de processos – cerca de 68% do total, frente ao restante dividido entre a Conformidade e o Aprimoramento do processo.

Figura 21 – Análise cruzada entre tipos de mineração de processos e técnicas de mineração de dados, identificadas nos estudos primários

Fonte: Ana Rocío Cárdenas Maita, 2015

4.2.1.4 Análises adicionais

A fim de prover uma extensa perspectiva da área de pesquisa de mineração de processos, as seguintes três análises são apresentadas: domínio de aplicação, ferramentas usadas, aderências às definições de mineração de processos.

Em relação aos domínios de aplicação, a figura 22 mostra as diversas áreas em que foram aplicados estudos de caso ou experimentos, quando identificados nos estudos primários analisados no mapeamento realizado. No total, 48% dos estudos primários relataram algum estudo de caso ou experimento. Os dados mostram que as três áreas em que mineração de processos têm sido mais frequentemente aplicada são: “Processos clínicos”, “Manufatura/Produção” e “Auditoria” com 34, 28 e 22 estudos primários, respectivamente, sendo que as três juntas representam 12% do total de estudos. Outros domínios de aplicação como Redes sociais, Sistemas empresariais, Componentes de serviços web, Suporte de TI, Medicina, Municipalidades, Telecomunicações, Economia, Sistemas acadêmicos, Fluxos de trabalho transacional, Sistemas de e-mail, Comercio eletrônico, Logística, Milícia, polícia e seguridade social, e Sistemas bancários aparecem em 11 estudos primários em média cada um. Para outros 85 estudos primários, não foi possível identificar a área de aplicação do estudo de caso ou experimento.

Figura 22 – Domínios de aplicação em mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

Em termos de ferramentas usadas para aplicar as técnicas de mineração de dados, considerando os estudos que mencionaram o uso de alguma ferramenta de apoio 84% usou a ProM, apenas 15% usaram outras ferramentas, entre elas Weka, identificada em 14 estudos; e a Disco Fluxicom, identificada em 12 estudos. Em relação ao total de estudos analisados, 401 (dos 705) não apresentaram detalhes da ferramenta usada em sua proposta (Figura 23).

Figura 23 – Ferramentas usadas em mineração de processos

Fonte: Ana Rocío Cárdenas Maita, 2015

Com relação à adesão às definições de mineração de processos proposto por Dumas et al. (2013), a maioria dos estudos primários identificados no mapeamento sistemático realizado não se aderem plenamente aos conceitos apresentados por este especialista em mineração de processos como apresentado na figura 24. Por exemplo, a partir dos 705 estudos primários, 518 estudos referem explicitamente ao termo “mineração de processos” ou de “mineração de *workflow*”. Assim, cerca de 27% dos estudos primários identificados foram avaliados por referir-se à aplicação de mineração de dados no contexto de processos de negócio (ou *workflow*), o que foi entendido como mineração de processos indiretamente. Além disso, 400 estudos mencionaram explicitamente o uso de *log* em suas abordagens, chamados “log de eventos”, “log de processo”, “registros de instância” ou outros termos similares. Isto significa que cerca de 43% dos estudos primários identificados não são explicitamente com base na análise de *log* de eventos para realizar tarefas de mineração de dados e podem não representar exatamente o que é definido como a mineração de processos. Finalmente, 449 estudos explicitamente lidam com modelos de processos ou modelos de fluxo de trabalho, que é considerado a base para os três tipos de mineração de processos de acordo com Aalst (2011). Portanto, 36% dos estudos primários não trata explicitamente modelos de processos ou modelos de fluxo de trabalho na sua abordagem em qualquer tipo de mineração do processo – descoberta, conformidade ou aprimoramento. Em geral, apenas 259 estudos satisfazem estas três características avaliadas em conjunto (isto é, 37%), que podia ser considerada totalmente aderente à definição de mineração

de processos dada por Dumas et al. (2013). Por outro lado, 63 estudos não satisfazem nenhuma de tais características (isto é, 9%), e, portanto, pode ser considerada totalmente não aderente a tal definição, embora eles satisfazem todos os critérios de inclusão.

Figura 24 – Adesão dos resultados às definições usadas neste estudo

Fonte: Ana Rocío Cárdenas Maita, 2015

4.3 Um exemplo de aplicação em mineração de processos

Esta seção apresenta um exemplo da execução de mineração de processos, usando diferentes tipos, técnicas e tarefas, conforme apresentado nas seções anteriores deste trabalho. Esse exemplo tem como propósito caracterizar com mais detalhes o cenário de aplicação de mineração de processos em um ambiente real visando facilitar tanto a compreensão dos problemas envolvidos em tal cenário quanto a geração de hipóteses que justifiquem de forma empírica, pelo menos parcialmente, o contexto atual de mineração de processos.

Tal facilitação é proporcionada pela apresentação de lições aprendidas obtidas a partir de duas estratégias: *(i)* aplicação de duas ferramentas (Disco Fluxicon e ProM), nas quais técnicas elaboradas, ou adaptadas, especialmente para resolver problemas de mineração de processos são implementadas; e *(ii)* definição de um problema de predição categórica (ou classificação) que foi concebido em dois diferentes níveis de complexidade, de forma a proporcionar algumas reflexões referentes à construção de modelos de classificação,

usando uma ferramenta em que técnicas de inteligência computacional e de aprendizado de máquina são implementadas (neste caso, especificamente, a rede neural “Multilayer Perceptron” disponível na ferramenta “Weka”). Esses casos de mineração de processos foram executados nos contextos de descoberta de processos e de conformidade de processos.

O ambiente real usado como domínio de aplicação se refere ao processo de negócio apoiado pelo uso de um sistema de ensino a distância por alunos de cursos de especialização oferecidos pela Universidade de São Paulo em conjunto com a Universidade Virtual do Estado de São Paulo. Trata-se de um processo não estruturado, em que os alunos não precisam seguir um modelo pré-definido rígido e, portanto, tal modelo nunca foi formalizado junto a qualquer entidade ligada ao curso, incluindo: alunos, coordenadores pedagógicos, administradores de sistema, desenvolvedores de sistema, etc. Assim, o processo sendo minerado está apenas implícito no *log* de eventos obtido durante o uso do sistema. Embora o processo seja não estruturado, pode ser útil para as entidades envolvidas no curso o entendimento do processo sendo seguido pelos alunos a fim de verificarem diversas informações relativas a eles, incluindo: perfil de alunos versus perfil de processo sendo realizado; ou se o processo esperado pelos coordenadores estaria sendo executado ou não. Assim, o objetivo da mineração de processos executada nesse domínio é, em termos gerais, descobrir e verificar qual é o processo seguido pelos alunos durante o uso da plataforma virtual.

Como procedimento para construção do exemplo, as seguintes etapas foram definidas: (i) exploração do domínio de aplicação; (ii) definição dos tipos, técnicas e tarefas de mineração a serem explorados; (iii) pré-processamento dos registros do *log* de eventos e criação de conjuntos de dados para mineração; (iv) aplicação de técnicas para mineração de processos disponíveis nas ferramentas Disco Fluxicon e ProM; (v) aplicação de uma técnica de inteligência computacional e/ou aprendizado de máquina em mineração de processos (considerando aquelas disponíveis da ferramenta “Weka”); (vi) reflexão sobre as lições aprendidas durante a execução do exemplo.

4.3.1 Exploração do domínio de aplicação

O exemplo de aplicação de mineração de processos descrito aqui foi executado a partir de registros de *log* de eventos referentes à navegação realizada por alunos de cursos de especialização em um sistema de ensino a distância especialmente desenvolvido para apoiar

o processo de ensino e aprendizagem dos cursos “Ética, Valores e Cidadania na Escola – EVC”. O sistema de apoio aos cursos (chamado aqui de sistema EVC) foi desenvolvido sobre a plataforma Moodle, a qual, de forma nativa, realiza o armazenamento das interações dos usuários com os objetos de aprendizagem disponíveis no sistema. Informações mais específicas sobre o sistema EVC e o processo de seu desenvolvimento são descritas por Junior et al. (2013), PERES et al. (2012).

Aproximadamente 1.500 usuários (entre alunos, tutores e coordenadores pedagógicos) interagiram com o sistema e com a plataforma Moodle durante a realização desses cursos. Para fins de desenvolvimento deste exemplo prático de mineração de processos, foram considerados 291 alunos matriculados no terceiro oferecimento do curso EVC, realizado entre as datas de 23/08/2013 a 09/12/2014. As interações dos usuários com o sistema durante esse período resultaram em cerca de 300.000 registros em *log* de eventos. A estrutura dos registros do *log* de eventos armazenados pela plataforma Moodle para o sistema EVC é descrita na tabela 20, e um trecho do arquivo de *log* de eventos é apresentado na figura 25.

Tabela 20 – Estrutura original de um registro de *log* de eventos do sistema EVC

Rótulo do campo	Descrição	Tipo de dado
Curso	Nomenclatura técnica usada pelos desenvolvedores do sistema	Texto
Hora	Momento em que o usuário executa cada ação no sistema	Data/hora
Endereço IP	Endereço do <i>Internet Protocol</i> da máquina de acesso do usuário	Máscara para endereço IP
Nome completo	Nome completo do usuário logado no sistema EVC	Texto
Ação	Identificação da ação do usuário na navegação no sistema EVC. Trata-se do acesso a um objeto de aprendizagem no sistema	Texto
Informação	Informação adicional sobre a ação realizada pelo usuário	Texto

Fonte: Ana Rocío Cárdenas Maita, 2015

4.3.2 Definição das técnicas, tarefas e tipos de mineração

O exemplo de aplicação de mineração de processos foi dividido em duas grandes partes: (i) primeiramente, por meio da aplicação de técnicas de mineração disponíveis em ferramentas desenvolvidas especificamente para o contexto de mineração de processos, o

Figura 25 – Exemplo de um trecho no *log* de eventos original do sistema EVC

Curso	Hora	endereço IP	Nome completo	Ação	Informação
PP	2014 maio 6 9:05	111.111.111.11	Aluno X	url view	M3-M4 LORENA - TURMA E
PP	2014 maio 6 9:05	111.111.111.11	Aluno X	course view	Projeto de pesquisa
EVC	2014 maio 6 9:05	111.111.111.11	Aluno X	course view	Ética, Valores e Cidadania na Escola
TCC	2014 maio 6 9:01	111.111.111.11	Aluno X	course view	Sistema de Acompanhamento de TCC
EVC	2014 maio 6 9:00	111.111.111.11	Aluno X	course view	Ética, Valores e Cidadania na Escola
EVC	2014 maio 6 9:00	111.111.111.11	Aluno X	user login	42
Módulo III	2014 março 25 20:37	222.22.22.222	Aluno X	course view	Direitos humanos e convivência democrática
EVC	2014 março 25 20:37	222.22.22.222	Aluno X	course view	Ética, Valores e Cidadania na Escola
EVC	2014 março 25 20:36	222.22.22.222	Aluno X	course view	Ética, Valores e Cidadania na Escola
EVC	2014 março 25 20:36	222.22.22.222	Aluno X	course view	Ética, Valores e Cidadania na Escola
EVC	2014 março 25 20:36	222.22.22.222	Aluno X	user change password	42
EVC	2014 março 25 20:36	222.22.22.222	Aluno X	user login	42

Fonte: Ana Rocío Cárdenas Maita, 2015

que poderia levar à aplicação de técnicas tradicionais de acordo com o contexto tratado neste trabalho; e (ii) depois, por meio da aplicação de técnicas especificamente da área de inteligência computacional e/ou aprendizado de máquina, o que demandaria o uso de ferramentas mais gerais da área de mineração de dados.

Considerando, a primeira parte do exemplo, os resultados do mapeamento sistemático realizado neste trabalho sinalizam as ferramentas Disco Fluxicon e ProM como as mais usadas em mineração de processos. A primeira, uma plataforma de código aberto, possui uma grande variedade de *plug-ins* disponíveis para os três tipos de mineração de processos. Já a Disco Fluxicon é uma ferramenta comercial (porém com licenças especiais para uso acadêmico) exclusiva para o tipo de mineração de processos “Descoberta de processos”. Essas duas ferramentas foram usadas no domínio de aplicação do sistema EVC a fim de exemplificar a mineração de processos. Assim, essas duas ferramentas foram selecionadas para uso neste exemplo.

Em relação aos tipos de mineração a serem usados, foram selecionados descoberta de processos e conformidade de processos. mineração de processos se encaixa no domínio de aplicação do sistema EVC visto que não existe um modelo de processo explicitamente definido em tal contexto, embora pudesse ser útil para os diversos atores envolvidos. Assim, para esse caso específico, o exemplo construído consistiu em apresentar os *log* de eventos, pré-processados de forma a atender aos requisitos de entrada de dados de ambas ferramentas Disco Fluxicon e ProM, para resolver o tipo de mineração descoberta de processos.

Além da descoberta de processos, o tipo de mineração conformidade também foi explorado nesse exemplo, uma vez que, além de estar disponível na ferramenta ProM, é o segundo tipo mais usado de acordo com o mapeamento realizado e tem potencial de trazer

informações importantes para os atores do sistema EVC. Para isso, o conjunto completo *log* de eventos pré-processado foi dividido em grupos (três, nesse caso específico) para explorar a conformidade do processo entre diferentes grupos de usuários do sistema: assim, primeiramente, um dos grupos seria usado para se descobrir o processo, e outros grupos seriam usados para verificar a conformidade dos processos que eles estavam executando com o processo descoberto inicialmente para o primeiro grupo. As características usadas para a separação dos usuários em grupos está apresentada na seção 4.3.3.

Para essa primeira parte do exemplo, não foi possível haver uma definição prévia da técnica a ser usada e da tarefa de mineração de dados a ser resolvida assim como da tarefa de mineração de dados a ser tratada. Isso ocorreu por limitação da abordagem seguida que se baseava no uso das ferramentas Disco Fluxicon e ProM. Para o caso da ferramenta Disco Fluxicon, por ser uma ferramenta comercial, e até onde foi possível averiguar em um trabalho de busca de informação em artigos científicos, técnicos e em *websites* correlatos ao tema, não há informações sobre o algoritmo ou a técnica usada para resolver os problemas de mineração de processos; mas apenas que essa ferramenta realiza uma análise baseada nas frequências de ocorrência de cada atividade. Em relação à ferramenta ProM, ainda seria necessário executar uma exploração prática dos diversos “plug-ins” nela disponíveis, pois são inúmeros, com diferentes características e limitações.

Na segunda parte do exemplo, para exercitar a aplicação de uma técnica de inteligência computacional e/ou de aprendizado de máquina no contexto de mineração de processos, considerando o domínio de aplicação do sistema EVC, optou-se por, primeiramente, escolher uma tarefa de mineração de dados a ser usada no exemplo, e então definir a técnica a ser aplicada. Seguindo os resultados do mapeamento sistemático realizado, verificou-se que a tarefa de predição categórica (ou classificação) é a mais frequentemente usada; e, considerando que no domínio de aplicação do sistema EVC a resolução da tarefa de classificação é potencialmente útil, essa tarefa foi escolhida para ser tratada nesse exemplo de aplicação.

No contexto do sistema EVC, o problema de predição categórica foi definido como um problema binário expresso em termos do resultado final do aluno no curso (aprovado; reprovado) em relação ao seu comportamento em termos do processo seguido no sistema EVC durante o curso. A ideia por trás dessa modelagem é que seja possível obter um modelo capaz de prever o resultado de final de aprovação ou reprovação do aluno com base no processo que ele está seguindo no sistema EVC, assumindo que o comportamento

de alunos anteriores pode ser usado como uma generalização do comportamento de novos alunos em novos oferecimentos do curso.

A técnica escolhida para resolução dessa tarefa de classificação foi redes neurais, mais especificamente, a arquitetura “Multilayer Perceptron” - uma arquitetura *feedforward* multicamada presente na maioria dos trabalhos de mineração de processos que usam redes neurais artificiais, conforme identificado na revisão sistemática realizada neste trabalho (MAITA et al., 2015). Como apoio para a aplicação da rede neural no problema em questão, foi escolhida a ferramenta Weka (*Waikato Environment for Knowledge Analysis*), uma ferramenta para processamento, mineração e análise de dados escrita em Java (distribuído sob *GNU Public License*).

O problema de classificação como definido aqui exigiu que o conjunto de dados gerado a partir do pré-processamento do *log* de eventos passasse por mais alguns procedimentos de adequação, o que gerou dois diferentes conjuntos que representaram dois níveis diferentes de complexidade para o problema, como descrito na seção 4.3.3.

4.3.3 Pré-processamento e criação dos conjuntos de dados

Os dados originais que compõem um *log* de eventos apresentavam problemas que poderiam interferir no processo de mineração de dados. Durante a análise dos registros, os seguintes problemas foram encontrados:

- **inconsistência:** nomes de alunos e de atividades não padronizados;
- **incompletude:** registros de *log* sem valores para alguns atributos;
- **ruído:** registros de *log* referentes a navegação feita pelos desenvolvedores do sistema durante os testes de desenvolvimento;
- **erro de usuário:** erros de autenticação de sessão.

Todas essas imperfeições foram tratadas manualmente, ou por meio da correção de valores ou por meio da exclusão de registros. Após essa limpeza, o conjunto de dados foi reduzido a 254.161 registros de *log* de eventos e, a partir deles, os registros foram transformados em nova organização que mais bem se adequou a realização da mineração de processos. A estrutura do conjunto de dados obtido está apresentada na tabela 21, e um trecho do conjunto de dados é apresentado na figura 26.

Tabela 21 – Estrutura de um registro de *log* do sistema EVC após pré-processamento

Rótulo do campo	Descrição	Tipo de dado
IdCaso	Identifica uma sequência de eventos relacionados a uma sessão de uso do sistema EVC, representando uma instância de processo	Autocontador
Hora	Data e hora de uma execução da ação	Data/hora
Usuário	Identificação do usuário que executa a ação	Texto
Atividade	Descrição da ação executada pelo aluno no sistema EVC	Enumerado

Fonte: Ana Rocío Cárdenas Maita, 2015

Figura 26 – Exemplo de um trecho do *log* de eventos do sistema EVC após pré-processamento

IdCaso	Hora	Usuario	Atividade
8826	18/06/2014 18:17	Aluno Y	Projeto de pesquisa
8826	18/06/2014 18:17	Aluno Y	Portifolio
8827	23/06/2014 10:32	Aluno Y	Profissao docente e educacao especial/inclusiva
8827	23/06/2014 10:32	Aluno Y	Profissao docente e educacao especial/inclusiva
8827	23/06/2014 10:50	Aluno Y	Profissao docente e educacao especial/inclusiva
8827	23/06/2014 10:50	Aluno Y	Portifolio
8827	23/06/2014 10:50	Aluno Y	Profissao docente e educacao especial/inclusiva
8827	23/06/2014 10:51	Aluno Y	Legenda video-aula
8827	23/06/2014 10:51	Aluno Y	Legenda video-aula
8827	23/06/2014 10:51	Aluno Y	Profissao docente
8827	23/06/2014 10:51	Aluno Y	Profissao docente e educacao especial/inclusiva
8827	23/06/2014 11:15	Aluno Y	Educacao especial/inclusiva
8827	23/06/2014 11:16	Aluno Y	Legenda video-aula
8827	23/06/2014 11:57	Aluno Y	Legenda video-aula
8827	23/06/2014 12:34	Aluno Y	Profissao docente

Fonte: Ana Rocío Cárdenas Maita, 2015

Os valores para o campo “IdCaso” foram criados baseados no seguinte procedimento: um identificado único foi criado, na forma de um autocontador, com base na informação do campo “Hora”, de forma que ações realizadas em um intervalo de tempo com variações de até duas horas (entre o final de uma ação e o início da próxima ação) fossem consideradas como parte de uma mesma instância de processo; ou seja, dentro de um mesmo “IdCaso”. Por causa desse procedimento, o número de instâncias de processo (valores para o atributo “IdCaso”) totalizou 25.013.

Para o campo “Atividade”, foi realizado um procedimento de uniformização e sumarização baseado no campo original dos registros de *log* “Informação”. Esse procedimento

reduziu o conjunto de 493 valores diferentes para o campo “Informação” para um conjunto de 12 valores diferentes para o campo “Atividade”.

A fim de viabilizar a realização de alguns teste com as ferramentas Disco Fluxicon e ProM, o conjunto de dados foi dividido em três subconjuntos, usando como atributo de divisão uma informação externa aos registros de *log*: a nota final obtida por cada aluno em uma das disciplinas. Apenas uma das disciplinas foi usada como exemplo de aplicação. Foram criadas três categorias de alunos com base nesse critério de divisão:

- excelente: alunos que obtiveram notas finais iguais ou maiores que 9,0, com 225.439 registros;
- regular: alunos que obtiveram notas finais iguais ou maiores que 7,0 e menores que 9,0, com 20.489 registros;
- insuficiente: alunos que obtiveram notas finais menores 7,0, com 6.233 registros.

Em relação à resolução da tarefa de predição categórica, os dados provenientes dos registros de *log* de eventos não estão organizados na forma como é esperado para processamento por uma rede neural *Multilayer Perceptron*. Além disso, para possibilitar a supervisão no processo de indução do modelo classificador, a informação referente à categoria à qual cada dado pertence precisa estar presente. Assim, dois novos conjuntos de dados foram gerados a partir das informações presentes no conjunto de dados descrito na tabela 21.

Os novos conjuntos de dados, embora gerados a partir de informações do *log* de eventos, são representações centradas em alunos e não centradas em ações; ou seja, enquanto nos *logs* de eventos cada registro diz respeito a uma ação de um aluno, esses novos conjuntos de dados agrupam todas as informações das instâncias de processos referentes às ações de cada aluno em um único registro.

Os novos conjuntos de dados possuem diferentes níveis de complexidade em relação à descrição das instâncias de processos. Para ambos, cada instância de processo (ou seja, um “IdCaso”) executada no sistema constitui um atributo descritivo de aluno. Os valores assumidos para cada atributo em cada registro (cada aluno) é o número de vezes que aquele aluno executou aquela instância do processo (zero ou mais vezes). A diferença entre os dois conjuntos se refere na forma como as atividades em um instância do processo são agrupadas, conforme descrito a seguir:

- **conjunto A:** as atividades executadas dentro da instância de processo são listadas desconsiderando a ordem de execução, excluindo repetições (conforme exemplo na figura 27). Esse conjunto possui 576 atributos e 94% de taxa de esparsidade.
- **conjunto B:** as atividades executadas dentro da instância de processo são listadas na ordem de execução, excluindo primeiramente repetições sequenciais e depois os laços (conforme exemplo na figura 28). Esse conjunto possui 5.839 atributos e 99% de taxa de esparsidade.

Figura 27 – Exemplo do conjunto A

Usuário	Categoria	1-2-3-4-5-6-7-9-10	2-3-4-6-7-9-10-11	2-3-4-6-10-12	2-3-5-6-9-10	2-3	2-3-5-6-9-10-11	2-3-6-10-11	2-4-5-6-7-9-10	2-3-4 (...)
Aluno 1	Aprovado	1	0	0	0	12	0	0	0	13
Aluno 2	Aprovado	2	0	0	0	9	0	0	0	18
Aluno 3	Aprovado	2	0	0	0	35	0	0	0	0
Aluno 4	Aprovado	0	1	0	0	23	0	0	0	12
Aluno 5	Aprovado	0	0	0	0	6	0	1	0	0
Aluno 6	Aprovado	3	0	0	0	21	0	0	0	0
Aluno 7	Aprovado	3	0	0	0	82	0	0	0	0
Aluno 8	Reprovado	1	0	0	0	1	0	0	1	0
Aluno 9	Aprovado	1	0	0	0	32	1	0	1	0
Aluno 10	Aprovado	0	0	0	0	20	0	0	0	11
Aluno 11	Aprovado	0	0	0	0	19	0	0	0	23
Aluno 12	Aprovado	0	0	0	0	25	0	0	0	0
Aluno 13	Aprovado	1	0	0	0	2	0	0	0	2
Aluno 14	Aprovado	1	0	0	0	5	0	0	0	1
Aluno 15	Aprovado	0	0	0	0	22	0	0	0	0

Fonte: Ana Rocío Cárdenas Maita, 2015

Figura 28 – Exemplo do conjunto B

Usuário	Categoria	04-08-04-05-10-05-04-08-04-05	04-08-04-05-10-07-04-08	04-08-04-05-10-07-11-04-08-04-07-11-07	04-08-04-05-10-08-04	04-08-04-05-10-08-04-06-09-07-04 (...)
Aluno 1	Aprovado	1	0	0	0	0
Aluno 2	Aprovado	2	0	0	0	0
Aluno 3	Aprovado	2	0	0	0	0
Aluno 4	Aprovado	0	0	0	0	0
Aluno 5	Aprovado	0	0	0	0	0
Aluno 6	Aprovado	3	0	0	0	0
Aluno 7	Aprovado	3	0	0	0	0
Aluno 8	Reprovado	1	0	0	0	0
Aluno 9	Aprovado	1	0	0	0	0
Aluno 10	Aprovado	0	0	0	0	0
Aluno 11	Aprovado	0	0	0	0	0
Aluno 12	Aprovado	0	0	0	0	0
Aluno 13	Aprovado	1	0	0	0	0
Aluno 14	Aprovado	1	0	0	0	0
Aluno 15	Aprovado	0	0	0	0	0

Fonte: Ana Rocío Cárdenas Maita, 2015

Por fim, para a categoria, uma distribuição binária foi obtida, conforme descrição a seguir:

- aprovado: alunos que obtiveram notas finais maiores ou iguais que 7,0, com 261 registros;
- reprovado: alunos que obtiveram notas finais menores que 7,0, com 30 registros.

4.3.4 Resultados: ferramentas Disco Fluxicon e ProM

A ferramenta ProM disponibiliza uma série de *plug-ins*, produzido por diferentes pesquisadores autores, que permitem a realização dos três tipos de mineração de processos. Apesar do grande número de *plug-ins* disponíveis, há problemas de usabilidade e complexidade no uso em grande parte deles, de forma que seu uso não é trivial. Na maioria dos casos, existe pouca documentação sobre características específicas do funcionamento desses *plug-ins*, incluindo sobre a descrição dos tipos de dados de entrada e de saída, embora na maioria dos casos os dados de entrada devam estar estruturados seguindo estruturas e formatos particulares pré-definidos. A maioria desses *plug-ins* são apenas citados e brevemente explicados em alguns artigos científicos, mas não são encontrados facilmente manuais de uso ou relatórios técnicos de demonstração ou outros tipos de documentos desse tipo.

Vários *plug-ins* da ferramenta ProM foram testados, mas as tentativas realizadas com os *plug-ins* de descoberta do modelo “*Mine for a Petri Net using Alpha-algorithm*” e “*Heuristic Miner*” foram os que permitiram a geração de resultados gráficos que permitissem a melhor observação natural do modelo de processo descoberto. No entanto, a interpretação de tais resultados se mostrou bastante complexa, o que levaria à necessidade de um estudo dos resultados ainda mais detalhado do que foi inicialmente definido como objetivo para este exemplo. Contudo, para efeitos dos objetivos básicos para este exemplo, a descoberta do modelo de processo foi realizada usando o conjunto de dados “excelente” a partir da aplicação do *plug-in* “*Heuristic Miner*”. O “*Heuristic Miner*” foi usado em preferência ao “*Mine for a Petri Net using Alpha-algorithm*” porque mostrou melhor visualização dos resultados.

A figura 29 apresenta o resultado do modelo de processo gerado para essa parte do exemplo, considerando o uso do *plug-in* “*Heuristic Miner*” da ferramenta ProM, cuja saída é apresentada em termos de uma rede de Petri. Tal modelo representa as execuções dos processos mais frequentes seguidas pelos alunos da categoria “excelente”, os retângulos de cor amarela representam atividades, enquanto que os laços entre elas representam os caminhos nas execuções do modelo. Embora tal detalhe na representação significa um problema na compreensão visual do modelo, assim como a análise do processo de negócio em si.

Figura 29 – Exemplo do modelo de processo gerado usando o *plug-in* “*Heuristic Miner*” da ferramenta ProM

Fonte: Ana Rocío Cárdenas Maita, 2015

Já para os testes de conformidade do modelo usando a ferramenta ProM, o *plug-in* “*ET Conformance*” foi o único que permitiu a execução com os dados existentes. Todos os demais não eram possíveis de serem usados com os dados disponíveis, ou apresentaram erros que não puderam ser tratados por falta de documentação apropriada. O *plug-in* “*ET Conformance*” permite medir a conformidade entre o arquivo de *logs* de eventos e sua rede de Petri correspondente (o processo explícito (existente ou descoberto) para o sistema sobre o qual os *logs* foram gerados). Os resultados, em função de métricas podem apresentar, por exemplo, a qualidade de uma rede de Petri para representar o comportamento descrito em um *logs* de eventos; ou quantificar a qualidade de um modelo descoberto em função da precisão dele com o respectivo arquivo de *logs* de eventos. Há pouca documentação existente sobre os detalhes técnicos deste *plug-in*. Para a interpretação das medidas oferecidas por tais métricas, é necessária uma análise ainda mais aprofundada sobre o método usado na obtenção de tais métricas, além do seu significado, o que não faz parte do escopo deste projeto.

A ferramenta Disco Fluxicon realiza a descoberta do modelo do processo por meio da interpretação das sequências de atividades existentes em um arquivo de *logs* de eventos. Seu maior benefício, frente à ferramenta ProM, é em termos de usabilidade e de legibilidade dos resultados apresentados. No entanto, poucas informações estão disponíveis sobre a técnica usada para a descoberta. O pouco de informação disponível apenas leva à conclusão de ser uma análise baseada em frequências de ocorrência para cada atividade.

Para descoberta de processo com a ferramenta Disco Fluxicon, foram realizados testes com os três subconjuntos de dados: “excelente”, “regular” e “insuficiente”. A ferramenta gera uma representação gráfica do modelo que descreve o processo descoberto. As figuras 30, 31, e 32 apresentam os três modelos de processo gerados respectivamente para os três subconjuntos de dados – “excelente”, “regular” e “insuficiente”. Nos gráficos apresentados, os retângulos representam atividades, a intensidade da cor dos retângulos representam a frequência com a qual as atividades foram acessadas (as mais escuras foram as mais acessadas), as ligações direcionadas entre os retângulos expressam a ordem de execução das atividades, e os números associados às ligações indicam o número de vezes com o qual as ligações foram seguidas.

108

Uma completa e correta interpretação do processo descoberto deveria ser realizada junto a especialistas no domínio do processo. No entanto, por se tratar de apenas um exemplo de aplicação de mineração, os especialistas de domínio não foram consultados, dada a ainda imaturidade dos resultados obtidos. Assim, nesse contexto, as análises realizadas apenas no escopo de exemplificação deste trabalho, foi possível observar que:

- atividades como “Legenda vídeo-aula” foram as mais acessadas nos três grupos de alunos, ou seja, independentemente de seu desempenho final;
- alunos com desempenho “excelente” e “regular” acessam mais frequentemente as atividades relacionadas a desenvolvimento de “Projeto de pesquisa”, “Documentos tutoriais” e “Avaliação do projeto”, quando comparados aos alunos com desempenho “insuficiente”;
- Provavelmente alunos com desempenho “insuficiente” apresentam baixa frequência em relação ao acesso a essas atividades por não terem cumprido adequadamente tais tarefas e por isso tiveram um desempenho abaixo do esperado.
- Alunos com desempenho “excelente”, por exemplo, aparentemente mostraram ser mais organizados em visitar as atividades relacionadas tanto às matérias, tais como: “Video de apoio”, “Convivência democrática”, “Educação especial/inclusiva”, “Profissão docente”, “Direitos humanos” e “Saúde na escola”, como àquelas atividades relacionadas ao acompanhamento do curso, tais como: “Manual”, “Documentos e tutoriais”, “Avaliação de projetos”, “Nota Portfólio”, etc. Por outro lado, alunos com desempenho “regular” mostram menos conexões entre o fluxo de execução de atividades relacionadas às matérias ou acompanhamento do curso.

Um entendimento mais refinado sobre tais resultados poderia ainda considerar outras perspectivas, tais como: quais outros erros cometeram os alunos de desempenho insuficiente; ou se houve falhas de projeto de interface gráfica no sistema EVC que dificultaram o acesso a outros recursos. Essas questões indicam que é necessária a realização de trabalhos futuros mais centrados em resolvê-las. Embora, para o objetivo deste trabalho, os resultados apresentam o potencial de mineração de processos para gerar processos alinhados com a realidade que permitam a melhoria de processo de negócio.

4.3.5 Resultados: predição categórica - *Multilayer Perceptron*

A construção de modelos de classificação com a rede neural artificial *Multilayer Perceptron*, usando a ferramenta Weka, foi realizada sob três estratégias diferentes, considerando o **conjunto de dados A**: resubstituição, *holdout* e validação cruzadas. Os parâmetros *default* sugeridos na ferramenta Weka (Tabela 22) para a rede neural em questão foram mantidos, uma vez que variá-los não levou a melhorias significativas nos resultados obtidos (Tabela 23).

Tabela 22 – Parâmetros usados na rede neural *Multilayer Perceptron*

Parâmetros	Descrição	Valor <i>default</i>
Número de épocas	Controla do tempo de treinamento	500
Tamanho do conjunto de validação	Se é diferente de 0, permite adiantar o término do treinamento, se necessário	0
Normalização dos atributos	Se aplicada (verdadeiro) coloca os valores dos atributos no intervalo [-1,1]	Verdadeiro
Número de neurônios na camada escondida	Influencia o tipo de mapeamento realizado na camada escondida da rede neural, determinando a dimensionalidade do dado após o mapeamento	(# de atributos + # de classes)/2
Taxa de aprendizado inicial	Controla a intensidade com que as alterações de pesos são aplicadas	0.3
Decaimento da taxa de aprendizado	Escolha sobre usar (verdadeiro) ou não (falso) uma taxa de aprendizado que muda de valor durante o treinamento da rede neural	Falso
Momentum	Se usado influencia a alteração de pesos de forma que mínimos locais possam ser evitados	0.2
Reset	Se ativado (verdadeiro) reinicia o treinamento da rede com uma taxa de aprendizado menor caso a rede entre em um processo de divergência no treinamento	Verdadeiro

Fonte: Ana Rocío Cárdenas Maita, 2015

O primeiro classificador, obtido sob a estratégia de resubstituição (treino e teste sobre a mesma amostra de dados) teve o objetivo de verificar a viabilidade de construção de um modelo de decisão para o problema em questão. Resubstituição é uma estratégia de indução de classificadores otimista, e não fornece informações sobre a capacidade de generalização do modelo. O resultado obtido mostra que a obtenção da superfície de decisão

é factível, mas que há uma dificuldade com a classificação da classe "reprovado" (classe negativa), já que mesmo no teste otimista, falsos positivos são encontrados.

O segundo classificador, obtido sob a estratégia de validação cruzada com 10 pastas, é um modelo realista e sua capacidade de generalização está aferida. Observando os resultados percebe-se a grande dificuldade do modelo em aprender a classe negativa. Essa dificuldade se dá principalmente por conta do desbalanceamento das classes no conjunto de dados – há muito mais alunos aprovados do que alunos reprovados no contexto sob análise.

O terceiro classificador tem o objetivo de ilustrar uma possibilidade no contexto onde o balanceamento das classes é presente. Para realização deste teste, o conjunto de dados foi amostrado de forma a diminuir o desbalanceamento. A estratégia *holdout* foi aplicada considerando 70% dos dados para treinamento do modelo e 30% dos dados para teste do modelo. Os resultados mostram o sucesso do modelo, no entanto, se trata de um teste em um conjunto de dados bastante reduzido.

Tabela 23 – Medidas de avaliação para o classificador (261 instâncias positivas e 30 instâncias negativas no conjunto A; 70 instâncias positivas e 30 instâncias negativas no conjunto A após amostragem)

Medida	Ressubstituição	Validação cruzada	<i>Holdout</i>
Número de instâncias classificadas corretamente	274	263	30
Número de instâncias classificadas incorretamente	17	28	0
Número de verdadeiros positivos (acertos na classe “aprovado”)	261	248	26
Número de verdadeiros negativos (acertos na classe “reprovado”)	13	15	4
Número de falsos negativos (erros na classe “aprovado”)	0	13	0
Número de falsos positivos (erros na classe “reprovado”)	17	15	0
Precisão (ou preditividade positiva)	0,94	0,94	1,00
Preditividade negativa	1,00	0,54	1,00
Revocação (ou taxa de verdadeiros positivos)	1,00	0,95	1,00
Especificidade (taxa de verdadeiros negativos)	0,43	0,50	1,00

Não foi possível obter um classificador para o conjunto de dados B no ambiente de desenvolvimento usado, pois trata-se de um conjunto de dados bastante grande em termos de atributos (alta dimensionalidade) que torna a complexidade de treinamento também bastante alta. Em algumas tentativas a ferramenta Weka acusou problema de falta de memória computacional e em outras tentativas, embora a ferramenta tenha acusado o término do treinamento, ela não informou os resultados. Um trabalho de exploração direta sobre a implementação provida pela ferramenta, manipulando diretamente as variáveis da implementação, e o uso de um ambiente de teste mais robusto, poderiam levar a possibilidade de obtenção de resultados. Essa é uma das oportunidades de trabalhos futuros que este estudo possibilita, uma vez que a modelagem do problema e o conjunto de dados já estão devidamente estruturados.

4.4 Lições aprendidas

Este trabalho alcançou resultados direcionados a quatro perspectivas a fim de permitir uma análise mais extensa e global da área de mineração de processos. Tais perspectivas são (cf. figura 33): *(i)* tipos de mineração de processos; *(ii)* técnicas e tarefas originárias de mineração de dados; *(iii)* quantidade de técnicas especificamente de inteligência computacional (IC) e de aprendizado de máquina (AM); *(iv)* detalhamento sobre aplicação de redes neurais e máquinas de vetores de suporte (SVM – Support Vector Machines); e *(v)* publicações em relação a veículos, países, pesquisadores e instituições.

As seções seguintes apresentam as lições aprendidas de cada etapa deste trabalho.

4.4.1 Lições aprendidas da revisão sistemática

Os resultados da revisão sistemática focada em redes neurais e máquinas de vetores de suporte mostraram que essas técnicas aparecem em apenas 2% de todos os estudos publicados entre 2003 e 2013 na área de mineração de processos. Esse cenário identificado para mineração de processos não reflete o contexto geral em mineração de dados, uma vez que, neste segundo caso, redes neurais e máquinas de vetores de suporte são amplamente usadas para resolver vários tipos de problemas específicos. Esse baixo uso de redes neurais e máquinas de vetores de suporte pode ser causado devido a uma falta de conhecimento sobre sua capacidade potencial para este tipo de problema especificamente. Dada que a

Figura 33 – Perspectivas dos resultados alcançados neste trabalho

Fonte: Ana Rocío Cárdenas Maita, 2015

área de mineração de processos envolve conhecimentos multidisciplinares, isso poderia eventualmente resultar da falta de especialização em determinadas áreas de conhecimento necessárias. Além disso, os estudos encontrados não cobrem as várias combinações possíveis nesse contexto, ou seja, eles não exploram as diferentes possibilidades existentes em relação aos tipos de mineração de processos ou aos tipos de arquitetura e de aprendizado de redes neurais e às variações de máquinas de vetores de suporte.

Além disso, existe uma concentração de estudos publicados nos últimos quatro anos, provavelmente porque neste último período houve uma maior divulgação em estudos relacionados a essa área. Este resultado mostra também um maior interesse por parte dos pesquisadores envolvidos, embora ainda haja imaturidade no desenvolvimento de trabalhos que apliquem redes neurais e máquinas de vetores de suporte.

4.4.2 Lições aprendidas do mapeamento sistemático

A distribuição temporal dos estudos primários selecionados destaca um aumento progressivo no número de publicações de 2005 a 2014, com maior incremento desde 2008. Tal interesse pela área de mineração de processos é corroborado pela diversidade de autores

e veículos para publicação (705 estudos primários, produzidos por 1.275 autores, em 469 instituições de pesquisa, localizadas em 61 países). A grande diversidade de autores, instituições de pesquisa, e países onde foram publicados esses artigos representam a ampla abrangência deste estudo e verifica-se, assim, um crescente interesse por parte dos autores em realizar pesquisas na área de mineração de processos. Apesar disso, conforme apresentado nas tabelas 17, 18 e 19, ainda existe uma grande concentração das pesquisas em países (17 países contribuem com 62% do total), em instituições (12 instituições contribuem com 34% do total), e em pesquisadores (16 pesquisadores contribuem com 29% do total). Especificamente em termos de pesquisador individual, o mais produtivo é Wil van der Aaslt, quem é responsável por cerca de 15% dos estudos, enquanto que os próximos três pesquisadores mais populares foram responsáveis por cerca de apenas 3% cada um. Tais concentrações podem ser uma das responsáveis por algumas das conclusões deste mapeamento em termos de, por exemplo, os tipos de técnicas que estão sendo aplicadas visto que, embora exista uma vasta gama de técnicas aplicáveis, é natural que um grupo específico de pesquisadores se torne especializado em algum conjunto determinado de técnicas.

Em relação aos tipos de mineração de processos, dos três tipos usados como referência neste projeto, descoberta de processos é o mais usado (identificado em 70% dos estudos). Embora a principal preocupação das organizações pode ser o aprimoramento de processos, um maior número de estudos que trata a descoberta de processos, isto pode ser entendido como um fenômeno natural, uma vez que o tratamento de tais melhorias depende da existência prévia dos modelos de processos para ser melhorados. Além disso, conformidade de processos pode ser vista como um passo intermediário necessário para encontrar problemas no processo que precisa de ser melhorado, tendo como uma entrada o modelo de processo, o que poderia explicar que seja o tipo de mineração de processos intermediário em termos de ocorrência, como indicado nos resultados deste mapeamento sistemático.

Em relação às tarefas de mineração de dados, verifica-se que há cinco tarefas bastante usadas neste contexto, nesta ordem: “predição categórica (ou classificação)”, “análise de dados estruturados”; “análise de agrupamentos”, “Descoberta de padrões frequentes, regras de associação e regras de correlação” e “Análise de evolução”. As outras três tarefas são bem menos usadas neste contexto. Esse resultado não é inesperado, já que essas tarefas também não são uniformemente distribuídas para a área de mineração

de dados em geral. No entanto, as tarefas que têm um maior número de ocorrências em mineração de processos diferem daquelas em mineração de dados em geral. Por exemplo, no caso específico de mineração de processos, as tarefas de “Análise de evolução” e “Análise de dados estruturados”, que são aplicadas em cenários com dados imprevistos, são tão amplamente usadas quanto as tarefas de “Predição categórica”, “Análise de agrupamentos”, e “Descoberta de padrões frequentes, regras de associação e regras de correlação”, que são comumente aplicadas em problemas clássicos de mineração de dados. Esse resultado era também esperado, visto que as primeiras tarefas são necessárias para lidar com cenários não clássicos de mineração de dados tais como BPM. “Análise de dados estruturados”, por exemplo, uma das tarefas aplicadas em problemas não clássicos de mineração de dados, foi classificada como a segunda mais amplamente aplicada em mineração de processos, e pode ser vista como uma tarefa perfeitamente adequada para este contexto, uma vez que ele trata de estruturas complexas como redes, grafos, etc. Complementarmente, “Análise de evolução”, outra tarefa aplicada em problemas não clássicos de mineração de dados (e classificada em quinto lugar no mapeamento realizado), também é bem adequada em mineração de processos, uma vez que o comportamento de processos de negócio permite alterações ao longo do tempo.

As outras três tarefas aplicadas em menor frequência em mineração de processos são: “análise de *outliers* e análise de ruído”, “previsão numérica – do tipo identificação de tendências” e “previsão numérica – do tipo regressão”. “Análise de *outliers* e análise de ruído” tem uma aplicabilidade significativa em mineração de processos e, portanto, poderia ser ainda mais explorada. Em termos de “*outliers*”, exceções ou casos particulares devem ser cuidadosamente observados, a fim de encontrar ineficiências e oportunidades de melhorias em processos de negócio, por exemplo, por meio da análise do comportamento e de exceções nas execuções dos processos ou de instâncias de processos menos executadas. Em termos de “ruído”, os dados brutos a serem minerados (ou seja, os *logs* de eventos) são frequentemente compostos por dados irrelevantes, incompletos, inconsistentes e imprecisos, gerados pelo uso indevido dos sistemas de informação. Por outro lado, “previsão numérica”, tem uma aplicabilidade menor em mineração de processos visto que, como qualquer previsão numérica, ela se refere a problemas de mineração que podem ser descritos por uma função (isto é, uma fórmula matemática) capaz de prever um valor numérico. Tais problemas numéricos atualmente são menos comuns em mineração de processos, embora o tipo de predição “identificação de tendências”, que envolve uma componente de tempo, é

mais comum em mineração de processos quando comparado com o tipo “regressão” de predição, o que diverge de mineração de dados em geral.

Em relação às técnicas de mineração de dados aplicadas em mineração de processos, 1.253 ocorrências de 25 classes de técnicas foram encontradas nos 705 estudos primários selecionados; 19% de tais técnicas são relacionadas às áreas de inteligência computacional e de aprendizado de máquina, e 71% são consideradas como tradicionais em mineração de dados. Em termos de técnicas de inteligência computacional e de aprendizado de máquina, pode-se observar por meio dos dados apresentados na figura 21 que elas estão basicamente concentradas em uma técnica de aprendizado de máquina supervisionado (ou seja, árvores de decisão) e duas técnicas de inteligência computacional (ou seja, computação evolutiva e lógica *fuzzy*). Árvores de decisão podem ser vistas como mais fáceis de ser interpretadas dentro das técnicas de aprendizado de máquina. Computação evolutiva e lógica *fuzzy* já foram implementadas em alguns *plug-ins* nas ferramentas desenvolvidas especificamente para a mineração de processos, conforme apresentado na Seção 2.3.3.

Essas três técnicas têm sido usadas para resolver as oito tarefas de mineração de dados em mineração de processos, embora o uso de cada uma tenha sido predominante em tarefas específicas considerando suas particularidades. Por exemplo, árvores de decisão são usadas principalmente para resolver a tarefa “Previsão categórica” visto que esta técnica foi originalmente proposta para este fim assim como a maioria de seus algoritmos é destinada a resolver essa tarefa. Lógica *fuzzy* é usada principalmente para resolver tarefas de “Descoberta de padrões frequentes, regras de associação e regras de correlação” provavelmente porque esta tarefa é implementada pela ferramenta Disco (a terceira ferramenta mais usada em mineração de processos), que usa uma técnica baseada em lógica *fuzzy*. Computação evolutiva é usada principalmente para resolver, por exemplo, a análise de evolução. Embora, nenhuma razão forte tenha sido identificada para essa correlação, a flexibilidade de modelos de computação evolutiva poderia ser visto como um facilitador para lidar com características de problemas temporais.

4.4.3 Lições aprendidas do exemplo de aplicação

Finalmente, o exemplo de aplicação realizado permitiu mais bem compreender os resultados obtidos com a revisão sistemática e com o mapeamento sistemático apresentados em seções anteriores, assim como, oferecer subsídios para a análise dos resultados de tais

estudos sistemáticos, de forma retroalimentada. A avaliação inicial do estado da arte em relação às ferramentas atualmente mais usadas (ou seja, Disco Fluxicon e ProM) assim como suas aplicações possibilitou a verificação de variedades de *plug-ins* e algoritmos disponíveis para serem aplicados em diversos contextos desta área. Disco Fluxicon, por um lado, permite realizar a descoberta do processo, embora não tenha se encontrado informações detalhadas acerca do algoritmo usado por ela. Além disso, ela permite realizar uma análise baseada nas frequências de ocorrência de cada atividade. Já a ferramenta ProM contém inúmeros *plug-ins* com características particulares em relação ao tipo de mineração de processos a ser usada e ao tipo de dados de entrada e de saída desejados em cada análise em particular.

Especificamente em relação à segunda parte do exemplo de aplicação, foi tratada a resolução da tarefa de predição categórica (ou classificação) por meio de redes neurais com a função “Multilayer Perceptron” da ferramenta Weka. Durante a etapa de pré-processamento dos dados, algumas dificuldades foram encontradas em relação ao tipo de dados representado em um *log* de eventos, visto que trata-se de um processo de negócio com características particulares (expostas na seção 2.3.2) semanticamente importantes do ponto de vista de BPM. No entanto, tais características podem representar uma limitação do ponto de vista da área de mineração de dados, pois não é possível aplicar o tratamento usual aos dados. Assim, a análise e o pré-processamento dos *logs* de eventos foram realizados levando sempre em consideração o valor que eles representam neste contexto.

Com a aplicação da técnica de classificação *Multilayer Perceptron*, os conceitos apresentados na seção 2.3.2 sobre os problemas em mineração de processos foram verificados, especificamente as dificuldades existentes devido à natureza do processo, tais como: heterogeneidade de instâncias de eventos, numerosas instâncias de eventos, e alta granularidade. De fato, a modelagem realizada tentou reduzir esses problemas, embora os resultados alcançados mostraram que tal modelagem não foi suficiente para conseguir resultados mais precisos. Esse tipo de dados deve ser tratado de forma mais aprofundada, visto que modelar um problema em mineração de processos implica gerar um conjunto de dados adequado para aplicar uma determinada técnica de mineração de dados – nesse caso, uma técnica de aprendizado de máquina ou inteligência computacional. Porém, os dados, ou seja, os *logs* de eventos, não devem perder a semântica do processo que eles representam. Tal dificuldade pode ser uma das razões pelas quais essas técnicas vem sendo pouco usadas em mineração de processos – como foi exposto com os resultados do

mapeamento sistemático. Desta forma, verifica-se uma lacuna de pesquisa para trabalhos futuros.

De forma geral, o exemplo de aplicação permitiu encontrar duas limitações importantes ao aplicar técnicas de inteligência computacional e de aprendizado de máquina. Por um lado, encontrou-se uma limitação em relação à modelagem do problema em si; e, por outro lado, a adaptação de tais técnicas antes de serem aplicadas com esse tipo de dados. Assim, provavelmente novas metodologias de pré-processamento dos *logs* de eventos devem ser propostas assim como adaptações nas abordagens de inteligência computacional e de aprendizado de máquina devem ser tratadas para serem aplicadas no contexto específico de mineração de processos. Ambas as lacunas são propostas de trabalhos futuros.

5 Conclusão

Mineração de processos é uma área de pesquisa relativamente recente, de forma que poucos trabalhos de pesquisa já realizaram estudos de revisão sistemática a seu respeito. Além disso, os poucos trabalhos de revisão sistemática existentes não trataram esse assunto de forma tão abrangente como foi realizada neste trabalho. O estudo cujos resultados são apresentados aqui permitem visualizar a área de mineração de processos de forma global mostrando que em geral a maioria das publicações entre 2005 e 2014 estão direcionadas a técnicas e algoritmos inicialmente formulados pelos primeiros pesquisadores da área, deixando de lado o conceito fundamental da mineração de processos que é vir da junção das principais potencialidades de duas áreas – BPM e mineração de dados. Dentro dessa perspectiva interdisciplinar, seria esperado mais abordagens de mineração de processos que usassem mais o potencial de técnicas existentes da área de mineração de dados, incluindo, principalmente, as técnicas de inteligência computacional e de aprendizado de máquina. No entanto, os resultados deste trabalho mostraram que apenas 19% dos 705 estudos primários analisados usam tais técnicas.

Este estudo apresentou uma abordagem interdisciplinar, já que análises tanto da perspectiva de BPM quanto de mineração de dados foram apresentadas, seguindo uma metodologia sistemática com rigor científico. Assim, as principais contribuições para a área são: *(i)* o método sistemático seguido, incluindo os critérios de inclusão e de exclusão aplicados, permite sua replicabilidade, de modo que as informações apresentadas podem ser reproduzidas e contrastáveis por outros pesquisadores ou especialistas e *(ii)* as informações apresentadas na forma de quadros, tabelas e figuras sintetizam os principais resultados, de modo a facilitar a descoberta de novas lacunas de pesquisa para projetos de pesquisa futuros assim como a discussão dos resultados por meio de novos estudos aprofundados em temas específicos.

Este trabalho teve como objetivo principal delimitar a área de mineração de processos por meio de um estudo bibliográfico amplo, que mostrasse a área de estudo em questão considerando vários enfoques: por um lado, as abordagens, técnicas, tarefas de mineração e ferramentas comumente mais usadas; e, por outro lado, veículos de publicação, universidades e pesquisados interessados no desenvolvimento da área. Assim, as contribuições deste trabalho formam as bases para novos trabalhos de pesquisa na área de mineração de processos, principalmente aqueles que possuem o objetivo de aplicar técnicas

de inteligência computacional ou de aprendizado de máquina, servindo de referência para profissionais e pesquisadores melhorarem seu entendimento da área.

Por meio da revisão sistemática realizada para identificar e avaliar os trabalhos que propõem o uso de redes neurais ou máquinas de vetores de suporte no contexto da mineração de processos verifica-se que, embora haja interesse científico na área de mineração de processos, pouco tem sido investido especificamente em redes neurais e máquinas de vetores de suporte. Além do número absoluto de estudos primários identificados ter sido relativamente baixo, eles não cobrem as diversas combinações possíveis nesse contexto, ou seja, sem explorar as diferentes possibilidades existentes em relação aos tipos de mineração de processos ou aos tipos de arquitetura e de aprendizado relacionados às redes neurais ou às variações e aos tipos de aprendizado de máquinas de vetores de suporte.

Por outro lado, considerando que essas técnicas tem sido amplamente usadas para resolver uma série de diferentes problemas em diferentes áreas em termos de mineração de processos, levanta-se aqui a hipótese, ainda a ser investigada com mais profundidade, de que tais técnicas também podem oferecer bons resultados para tarefas específicas em mineração de processos. Um cenário similar é esperado para técnicas similares de inteligência computacional e de aprendizado de máquina.

Embora poucos estudos primários tenham sido identificados na revisão sistemática, as avaliações realizadas para esses estudos indicaram que aproximadamente 80% desses estudos que tiveram algum tipo de avaliação apresentaram resultados positivos. Isso ressalta os benefícios de mineração de processos tais como aumento tanto da produtividade, como da eficiência nos processos de negócio, quando usadas técnicas de inteligência computacional e de aprendizado de máquina. Porém, de fato, nem todos esses resultados positivos podem ser atribuídos exclusivamente ao uso de redes neurais ou máquinas de vetores de suporte, uma vez que a maioria dos trabalhos propõe abordagens híbridas que mesclam uma série de diferentes técnicas e tecnologias.

Essa parte das contribuições deste trabalho relacionada à revisão sistemática realizada aparece no artigo já publicado *“Process mining through artificial neural networks and support vector machines: a systematic literature review”* (MAITA et al., 2015).

Já o mapeamento sistemático, foi um estudo bem mais abrangente – tratando 705 estudos primários, enquanto que trabalhos realizados nos últimos anos com algum motivo similar tratou um escopo bem menor – entre 7 a cerca de 50 trabalhos. Além disso, a grande maioria dos trabalhos de revisão existente na área de mineração de processos

não teve o objetivo de ser um trabalho de revisão de abordagem sistemática. Por fim, nenhum trabalho tinha tido ainda o propósito de avaliar a contribuição de técnicas de inteligência computacional e de aprendizado de máquina no contexto de mineração de processos, embora essa área seja comumente definida como uma junção de: *(i)* mineração de dados e “aprendizado de máquina”, de um lado; e *ii* modelagem e análise de processos de negócio, de outro lado.

Assim, o trabalho de análise da literatura realizado aqui apresenta um direcionamento maior em relação às técnicas de inteligência computacional tais como lógica *fuzzy*, computação evolutiva, redes neurais, e máquinas de vetores de suporte – e técnicas de aprendizado de máquina incluindo tanto aprendizado de máquina supervisionado como aprendizado de máquina não supervisionado. Para apresentar uma visão mais clara do uso de tais técnicas, seu uso foi comparado em relação às técnicas tradicionais de mineração de dados sendo aplicadas em mineração de processos. Tal avaliação mostrou que aproximadamente 81% das abordagens encontradas seguem técnicas tradicionais em mineração de dados e apenas aproximadamente 19% usam técnicas de inteligência computacional ou de aprendizado de máquina.

Além disso, como complemento, em relação às tarefas de mineração sendo usadas em mineração de processos, as cinco mais usadas são: predição categórica (ou classificação) (25%); análise de dados estruturados (23%); análise de agrupamentos (21%); descoberta de padrões frequentes, regras de associação, e regras de correlação (21%); e análise de evolução (18%). Já em relação aos tipos de mineração de processos, os mais usados de acordo com os 705 estudos primários avaliados são: descoberta de processos (71%); conformidade de processos (37%); e aprimoramento de processos (17%).

Em termos de resultados gerais obtidos com o mapeamento sistemático realizado, verifica-se que o cenário identificado para o contexto de mineração de processos não reflete o contexto geral de mineração de dados. No contexto mais amplo de mineração de dados, técnicas de inteligência computacional e de aprendizado de máquina são amplamente aplicadas, considerando os bons resultados normalmente obtidos na resolução de problemas complexos que envolvem situações reais. Apesar de a área de mineração de processos também lidar com problemas altamente complexos, com características muitas vezes similares a outros problemas já tratados no contexto geral de mineração de dados, parece haver uma falta de intercâmbio entre as diferentes áreas de forma que experiências bem sucedidas em mineração de dados pudessem ser incorporadas pela mineração de processos.

O baixo uso de técnicas de inteligência computacional e de aprendizado de máquina pode ser decorrência de um possível baixo conhecimento de suas potencialidades para esse tipo de problema, uma vez que muitos pesquisadores de mineração de dados são advindo da área de BPM e não de mineração de dados. Considerando que a área de mineração de processos envolve uma série de conhecimentos multidisciplinares, isso pode provocar, em muitos casos, a falta de experiência em determinadas áreas de conhecimento. Entretanto, essa hipótese precisaria ser explorada de forma mais profunda como uma possível causa do cenário identificado por esse trabalho.

Em termos de veículos de publicação, universidades e cientistas interessados no desenvolvimento da área, verifica-se que a maior parte dos trabalhos encontram-se concentrados em universidades de um país específico (na Holanda) com 126 publicações, do total de 705; sendo a seguinte universidade colocada na lista (na Bélgica), com 26 publicações. A grande diferença entre a primeira e segunda pode ser explicada pelo fato de ter como parte o grupo de pesquisa em mineração de processos dirigido pelo principal pesquisador desta área, Prof. Dr. W. van der Aalst; quem também é co-autor do maior número de publicações, 105 do total de 705; seguido por um pesquisador do Reino Unido com 20 publicações.

O número de 705 estudos primários identificados nos últimos dez anos (período de 2005 a 2014) não significa ser o número exato de trabalhos publicados nessa área em toda a literatura. De fato, o número real de trabalhos publicados nessa área, para esse período, tende a ser maior, principalmente por dois motivos: *i* dado o grande número de trabalhos já publicados na área, houve critérios de exclusão que teve o objetivo de excluir trabalhos publicados que indicassem dúvida na qualidade do trabalho, incluindo trabalhos curtos (*short papers*) e trabalhos que tivessem sido revisado por pares (ex: capítulos de livros convidados); e *ii* embora as duas bases de dados mais abrangentes disponíveis atualmente tenham sido usadas – Scopus e ISI Web of Science, é possível que haja outros trabalhos de qualidade que não tenham sido indexados por essas bases, ou que tenham sido indexados porém que não tenham sido identificados por meio da estratégia de busca usada, por alguma limitação existente, por exemplo, na *string* de busca.

Este trabalho permitiu verificar que o principal grupo de trabalhos da área de mineração de processos encontra-se centralizada a um número limitado de abordagens. Um dos motivos para isso poderia ser a complexidade que exige a modelagem do problema para a aplicação de técnicas mais específicas para mineração de dados relacionadas a

inteligência computacional e aprendizado de máquina, tais como redes neurais. Isso abre uma série de novas possibilidades de pesquisa para serem tratadas em mais profundidade em futuros projetos, incluindo, por exemplo:

- Analisar a detalhe as técnicas de inteligência computacional e de aprendizado de máquina, buscando aplicá-las ou adaptá-las para este contexto. Estudar técnicas de modelagem de problemas complexos com dados estruturados em mineração de dados, tentando aplicá-las na modelagem deste tipo de problemas.
- Pesquisar em profundidade sobre a natureza dos processos, para lidar situações tais como concorrência, ciclos, execuções de processos excepcionais, relação atividade-recursos, entre outras que dificultam a aplicação de técnicas já existentes para outros contextos em mineração de processos. Plantear possíveis variações dos algoritmos existentes em mineração de dados para eles conseguir melhores resultados no contexto de mineração de processos.
- Desenvolver técnicas ou ferramentas que facilitem a visualização do modelo de processo de negócio gerado após a aplicação de técnicas de inteligência computacional e de aprendizado de máquina, já que essas técnicas apenas produzem dados brutos que podem ser usados na abstração de modelos que descrevam processos de negócio, mas não diretamente os modelos de processo propriamente ditos.

Referências¹

- AALST, W. M. P. van der. *Process Mining – Discovery, Conformance and Enhancement of Business Processes*. 1. ed. [S.l.]: Springer, 2011. Citado 25 vezes nas páginas 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, 39, 41, 42, 44, 45, 46, 49, 51, 52, 53, 59, 62, 71, 74 e 94.
- AALST, W. M. P. van der. Mediating between modeled and observed behavior: the quest for the “right” process. In: *IEEE International Conference on Research Challenges in Information Science*. [S.l.: s.n.], 2013. p. 31–43. Citado na página 19.
- AALST, W. M. P. van der; DUSTDAR, S. Process mining put into context. *Web-Scale Workflow*, IEEE Internet Computing, v. 16, n. 1, p. 82–86, 2012. Citado na página 43.
- AALST, W. M. P. van der; HOFSTEDE, A. T.; WESKE, M. Business process management: A survey. *International Conference on Business Process Management*, Springer, p. 1–12, 2003. Citado na página 17.
- AALST, W. V. D. et al. Process mining manifesto. In: SPRINGER. *Business process management workshops*. [S.l.], 2012. p. 169–194. Citado 2 vezes nas páginas 44 e 47.
- ABONYI, J.; FEIL, B.; ABRAHAM, A. Computational intelligence in data mining. *Informatica*, v. 29, n. 1, p. 3–12, 2005. Citado 2 vezes nas páginas 17 e 18.
- ALONSO, G. et al. *Web Services – Concepts, Architecture and Applications*. [S.l.]: Springer-Verlag, 2004. v. 1. 360 p. Citado na página 27.
- BONABEAU, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 99, n. suppl 3, p. 7280–7287, 2002. Citado na página 40.
- BOSE, R. P. J. C.; MANS, R. S.; AALST, W. M. P. van der. Wanna improve process mining results? In: *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*. [S.l.]: IEEE, 2013. p. 127–134. Citado 3 vezes nas páginas 19, 47 e 48.
- CAMILO, C. O.; SILVA, J. C. da. *Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas*. [S.l.], 2009. Citado 2 vezes nas páginas 30 e 40.
- CHEN, C. L. P.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, Elsevier, v. 275, p. 314–347, 2014. Citado na página 21.
- CHEN, L.; BABAR, M. A. A systematic review of evaluation of variability management approaches in software product lines. *Information and Software Technology*, Elsevier, v. 53, n. 4, p. 344–362, 2011. Citado 2 vezes nas páginas 67 e 78.
- COOK, J. E.; WOLF, A. L. Discovering models of software processes from event-based data. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, ACM, v. 7, n. 3, p. 215–249, 1998. Citado na página 18.
- CRISTIANINI, N.; SHAW-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. [S.l.]: Cambridge University Press, 2000. Citado 2 vezes nas páginas 37 e 58.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- CÔRTEZ, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados – funcionalidades, técnicas e abordagens. *PUC-Rio Inf. MCC10/02*, 2002. Citado 3 vezes nas páginas 17, 37 e 38.
- DUMAS, M.; AALST, W. M. P. van der; HOFSTEDE, A. ter. *Process-aware information systems: bridging people and software through process technology*. [S.l.]: John Wiley & Sons, Hoboken, New Jersey, 2005. Citado 3 vezes nas páginas 26, 27 e 28.
- DUMAS, M. et al. *Fundamentals of Business Process Management*. [S.l.]: Springer-Verlag Berlin, 2013. Citado 4 vezes nas páginas 23, 27, 94 e 95.
- DYBÅ, T.; DINGSØYR, T. Strength of evidence in systematic reviews in software engineering. In: ACM. [S.l.], 2008. p. 178–187. Citado na página 61.
- FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. 1. ed. [S.l.]: Editora LTC, 2011. Citado na página 34.
- FANTINATO, M.; SOUZA, I. M. G. de; TOLEDO, M. B. F. Product line in the business process management domain. *Applied Software Product Line Engineering*, Auerbach Publications, p. 497–530, 2010. Citado na página 23.
- FAUSETT, L. *Fundamentals of Neural Networks Architectures, Algorithms and Applications*. [S.l.]: Prentice Hall, Upper Saddle River, NJ, 1994. 461 p. Citado na página 58.
- FRALEY, C.; RAFTERY, A. E. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, Br Computer Soc, v. 41, n. 8, p. 578–588, 1998. Citado na página 38.
- GANTZ, J.; REINSEL, D. The digital universe decade – are you ready? *IDC White Paper*, EMC Corporation, 2010. Citado na página 19.
- GROUP, O. O. M. *Business Process Model and Notation (BPMN) Version 2.0*. 2011. [Http://www.omg.org/spec/BPMN/2.0/PDF/](http://www.omg.org/spec/BPMN/2.0/PDF/). Disponível em: <<http://www.omg.org/spec/BPMN/2.0/PDF/>>. Citado na página 26.
- GRZYMALA-BUSSE, J. W. Three approaches to missing attribute values: A rough set perspective. In: *Data Mining: Foundations and Practice*. [S.l.]: Springer, 2008. p. 139–152. Citado na página 30.
- GÜNTHER, C. W.; AALST, W. M. P. van der. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In: SPRINGER. *Business Process Management*. [S.l.], 2007. Citado na página 18.
- HAN, J.; KAMBER, M. *Data mining: Concepts and Techniques*. 2. ed. [S.l.]: Morgan Kaufmann San Francisco, Calif, USA, 2006. Citado 6 vezes nas páginas 17, 30, 31, 44, 58 e 62.
- HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, JSTOR, p. 100–108, 1979. Citado na página 39.
- HAYKIN, S. *Neural networks and learning machine*. 3. ed. [S.l.]: Prentice Hall, 2009. Citado 2 vezes nas páginas 35 e 58.

HOLLINGSWORTH, D. Workflow management coalition. *The Workflow Reference Model*, Document Number WFMC-TC00-1003, Brussels, p. 22–23, 1995. Citado na página 17.

IEEE. *IEEE Computational Intelligence Society*. 2008.

[Http://ewh.ieee.org/r7/montreal/cis/](http://ewh.ieee.org/r7/montreal/cis/). Disponível em: <<http://ewh.ieee.org/r7/montreal/cis/>>. Citado na página 34.

JORDAN, D.; EVDEMON, J. *Web Services Business Process Execution Language Version 2.0*. 2007. [Http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf](http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf). Disponível em: <<http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf>>. Citado na página 27.

JUNIOR, M. P. S. et al. Experimenting with design thinking in requirements refinement for a learning management system. *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, p. 182–193, 2013. Citado na página 97.

KITCHENHAM, B. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. United Kingdom, 2007. Citado 4 vezes nas páginas 54, 56, 57 e 61.

KOHONEN, T. *Self-organizing Maps*. [S.l.]: Springer Berlin, 2001. Citado na página 38.

LABORATORIES, F. P. *Disco Fluxicon*. 2015. Disponível em: <<<http://fluxicon.com/>>>. Citado na página 51.

MAITA, A. R. C. et al. Process mining through artificial neural networks and support vector machines: a systematic literature review. *Business Process Management Journal*, Emerald Group Publishing Limited, v. 21, n. 6, 2015. Citado 10 vezes nas páginas 57, 59, 66, 67, 68, 69, 70, 71, 100 e 121.

MEDEIROS, A. K. A. de et al. Process mining: Extending the α -algorithm to mine short loops. *Eindhoven University of Technology, Eindhoven*, Department of Technology Management, Eindhoven University of Technology, v. 19, p. 24, 2004. Citado na página 50.

MEDEIROS, A. K. A. de; WEIJTERS, A. J. M. M. Genetic process mining. In: CITESEER. *Applications and Theory of Petri Nets 2005, volume 3536 of Lecture Notes in Computer Science*. [S.l.], 2005. Citado na página 18.

MEDEIROS, A. K. A. de; WEIJTERS, A. J. M. M.; AALST, W. M. P. van der. Genetic process mining: an experimental evaluation. *Data mining and Knowledge Discovery Journal*, Springer, p. 245–304, 2007. Citado na página 50.

MITCHELL, T. M. *Machine Learning*. 1. ed. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. Citado na página 34.

NAZEM, S. M.; SHIN, B. Data mining: New arsenal for strategic decision-making. In: IRM PRESS. *Data warehousing and web engineering*. [S.l.], 2002. p. 103–112. Citado na página 30.

ONG, Y.-S.; LIM, M. H.; CHEN, X. Research frontier-memetic computation—past, present & future. *IEEE Computational Intelligence Magazine*, v. 5, n. 2, p. 24, 2010. Citado na página 34.

- PERES, S. M. et al. Apoio ao aprendizado semipresencial: Uma aplicação no curso de especialização Ética, valores e saúde na escola. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2012. p. 8–1–8–10. Citado na página 97.
- PERES, S. M. et al. A committee machine implementing the pattern recognition module for fingerspelling applications. In: 25TH ACM SYMPOSIUM ON APPLIED COMPUTING – SAC 2010. [S.l.], 2010. Citado na página 30.
- PHILLIPS-WREN, G.; ICHALKARANJE, N.; JAIN, L. C. *Intelligent decision making: an AI-based approach*. [S.l.]: Springer Science & Business Media, 2008. v. 97. Citado na página 34.
- POZO, A. et al. Computação evolutiva – apostila. *Material Didático* <<http://www.inf.ufpr.br/aurora/tutoriais/Ceapostila.pdf>>, Departamento de Informática - Universidade Federal do Paraná, 2015. Citado na página 36.
- RUSSELL, N. et al. On the suitability of UML 2.0 activity diagrams for business process modelling. In: *Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modelling*. Hobart, Tasmania, Australia: Australian Computer Society, 2006. (APCCM 2005), p. 95–104. Citado na página 26.
- RUSSELL, S.; NORVING, P. *Artificial intelligence*. 2. ed. [S.l.]: Elsevier, 2004. Citado na página 35.
- STAHL, F.; JORDANOV, I. An overview of the use of neural networks for data mining tasks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Periodicals, Inc. New York, USA, v. 2, n. 3, p. 193–208, 2012. Citado 4 vezes nas páginas 17, 37, 38 e 56.
- TAMPELINI, L. G. et al. Uma aplicação de redes de elman no tratamento e predição de séries temporais hidrológicas. In: IX CBRN – CONGRESSO BRASILEIRO DE REDES NEURAIAS. [S.l.], 2009. Citado na página 30.
- THOMÉ, A. C. G. Redes neurais: uma ferramenta para kdd e data mining. *Material Didático* <http://equipe.nce.ufrj.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf>, Outubro, 2002. Citado na página 30.
- TIWARI, A.; TURNER, C. J. A review of business process mining: State-o-the-art and future trends. *Business Process Management Journal*, Emerald Group Publishing Limited, v. 14, n. 1, p. 5–22, 2008. Citado 3 vezes nas páginas 19, 21 e 26.
- TU/E, E. U. of T. *Prom Tools*. 2015. Disponível em: <<<http://www.promtools.org/>>>. Citado na página 50.
- VAPNIK, V. *The Nature of Statistical Learning Theory*. [S.l.]: Springer Science & Business Media, 2013. 314 p. Citado na página 37.
- VERGIDIS, K. et al. Optimisation of business process designs: An algorithmic approach with multiple objectives. *International Journal of Production Economics*, Elsevier, v. 109, n. 1, p. 105–121, 2007. Citado na página 18.
- WANG, L.; FU, X. *Data mining with computational intelligence*. Berlin Heidelberg: Springer, 2006. Citado 2 vezes nas páginas 17 e 30.

WESKE, M. *Business Process Management*. 2. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2012. Citado 5 vezes nas páginas 17, 23, 25, 26 e 27.

WU, X. et al. Top 10 algorithms in data mining, knowledge information systems. *Know Inf Syst*, Springer-Verlag, v. 1, n. 14, p. 1–37, 2008. Citado 5 vezes nas páginas 18, 21, 30, 38 e 56.

YUE, D. et al. A review of process mining algorithms. In: *Business Management and Electronic Information (BMEI), 2011 International Conference on*. [S.l.]: IEEE, 2011. v. 5, p. 181–185. Citado 3 vezes nas páginas 20, 21 e 50.

ZHANG, M.; DUAN, Z. From business process models to web services orchestration: The case of UML 2.0 activity diagram to BPEL. In: *Proceedings of the 6th International Conference on Service-Oriented Computing*. Sydney, Australia: Springer Verlag, 2008. (ICSOC 2008), p. 505–510. Citado na página 26.