UNIVERSIDADE DE SÃO PAULO

FFCLRP - DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

**Detecção de *outliers* baseada em caminhada determinística
do turista**

Rafael Delalibera Rodrigues

Dissertação apresentada à Faculdade de Filosofia,
Ciências e Letras de Ribeirão Preto da USP, como
parte das exigências para a obtenção do título de
Mestre em Ciências, Área: Computação Aplicada.

Ribeirão Preto - SP
2018

Detecção de *outliers* baseada em caminhada determinística do turista

Rafael Delalibera Rodrigues

ORIENTADOR: PROF. DR. ZHAO LIANG

CO-ORIENTADOR: PROF. DR. ANTONIO CARLOS ROQUE DA SILVA FILHO

Versão Revisada

(Versão original encontra-se na unidade que aloja
o Programa de Pós-graduação)

Ribeirão Preto - SP
2018

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.

Rodrigues, Rafael Delalibera

Detecção de *outliers* baseada em caminhada determinística do turista. Ribeirão Preto, 2018.

97 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/USP. Área de concentração: Computação Aplicada.

Orientador: Liang, Zhao.

1. Detecção de outlier. 2. Mineração de dados. 3. Descoberta de conhecimento. 4. Reconhecimento de padrões. 5. Computação bioinspirada.

Outlier detection based on deterministic tourist walk

Rafael Delalibera Rodrigues

SUPERVISOR: PROF. DR. ZHAO LIANG

CO-SUPERVISOR: PROF. DR. ANTONIO CARLOS ROQUE DA SILVA FILHO

Revised Version

(Original version is in the unit that hosts
the Graduate Program)

Ribeirão Preto - SP
2018

Rafael Delalibera Rodrigues

Detecção de *outliers* baseada em caminhada determinística do turista

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

Aprovado em: 03 de Abril de 2018

Banca Examinadora

Prof. Dr. Zhao Liang
Orientador

Prof. Dr. Renato Tinós

Prof. Dr. Jó Ueyama

Prof. Dr. João Roberto Bertini Junior

Ribeirão Preto - SP
2018

À minha filha Sophia.

AGRADECIMENTOS

Agradeço ao Prof. Dr. Zhao Liang, meu orientador pela oportunidade, amizade e ótima orientação, à minha família por todo apoio emocional e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

“If a machine is expected to be infallible, it cannot also be intelligent.”

Alan M. Turing; 1947

RESUMO

Detecção de *outliers* é uma tarefa fundamental para descoberta de conhecimento em mineração de dados. Cujo objetivo é identificar as amostras de dados que desviam acentuadamente dos padrões apresentados num conjunto de dados. Neste trabalho, apresentamos uma nova técnica de detecção de *outliers* baseada em caminhada determinística do turista. Especificamente um caminhante é iniciado para cada exemplar de dado, variando-se o tamanho da memória, assim, um exemplar recebe uma alta pontuação de *outlier* ao participar em poucos atratores, enquanto que receberá uma baixa pontuação no caso de participar numa grande quantidade de atratores. Os resultados experimentais em cenários artificiais e reais evidenciaram um bom desempenho do método proposto. Em comparação com os métodos clássicos, o método proposto apresenta as seguintes características salientes: 1) Identifica os *outliers* através da determinação de estruturas no espaço de dados ao invés de considerar apenas características físicas, como distância, similaridade e densidade. 2) É capaz de detectar *outliers internos*, situados em regiões entre dois ou mais agrupamentos. 3) Com a variação do valor de memória, os caminhantes conseguem extrair tanto características locais, quanto globais do conjunto de dados. 4) O método proposto é determinístico, não exigindo diversas execuções (em contraste às técnicas estocásticas). Além disso, neste trabalho caracterizamos, pela primeira vez, que as dinâmicas exibidas pela caminhada do turista podem gerar atratores complexos, com diversos cruzamentos. Sendo que estes podem revelar estruturas ainda mais detalhadas e consequentemente melhorar a detecção dos *outliers*.

Palavras-chave: outlier, outlier interno, caminhada do turista, valor de memória, valor crítico de memória, atrator, atrator com cruzamento.

ABSTRACT

Outlier detection is a fundamental task for knowledge discovery in data mining. It aims to detect data items that deviate from the general pattern of a given data set. In this work, we present a new outlier detection technique using tourist walks. Specifically, starting from each data sample and varying the memory size, a data sample gets a higher outlier score if it participates in few tourist walk attractors, while it gets a low score if it participates in a large number of attractors. Experimental results on artificial and real data sets show good performance of the proposed method. In comparison to classical methods, the proposed one shows the following salient features: 1) It finds out outliers by identifying the structure of the input data set instead of considering only physical features, such as distance, similarity or density. 2) It can detect not only external outliers as classical methods do, but also internal outliers staying among various normal data groups. 3) By varying the memory size, the tourist walks can characterize both local and global structures of the data set. 4) The proposed method is a deterministic technique. Therefore, only one run is sufficient, in contrast to stochastic techniques, which require many runs. Moreover, in this work, we find, for the first time, that tourist walks can generate complex attractors in various crossing shapes. Such complex attractors reveal data structures in more details. Consequently, it can improve the outlier detection.

Keywords: outlier, internal outlier, tourist walk, memory size, critical memory size, attractor, crossing-attractor.

LISTA DE FIGURAS

- Figura 1. Distinção entre *outlier* e ruído. (a) Um espaço de dados com um *outlier* denotado por A (em vermelho). (b) O mesmo espaço de dados, ainda com o *outlier* A, porém na presença de ruído. Fonte: Adaptado de (AGGARWAL, 2013).....36
- Figura 2. Tipos de *outliers*. (a) *Outlier* pontual. (b) *Outlier* coletivo. (c) *Outlier* contextual. Fonte: Adaptado de (AGGARWAL, 2013)37
- Figura 3. *Outlier interno* (asterisco vermelho) entre dois agrupamentos e três *outliers* externos (marcadores em cruz de cor preta).39
- Figura 4. Exemplo de detecção de *outlier* com abordagem baseada em profundidade. Fonte: Adaptado de (PREPARATA; SHAMOS, 1993)44
- Figura 5. Espaço de dados com agrupamentos de diferentes densidades e dois *outliers* em regiões distintas. Fonte: Adaptado de (AGGARWAL, 2013)46
- Figura 6. Exemplo de funcionamento do algoritmo DBSCAN. (a) Ilustração dos três tipos de pontos possíveis: central, borda e ruído; definidos com base na distância de acessibilidade e *MinPts*. (b) Um espaço de dados com dois agrupamentos e um *outlier*.....50
- Figura 7. Dois cenários bastante distintos, mas que determinam uma mesma pontuação de *outlier* para q com uso do método de detecção kNN. Fonte: Adaptado de (ANGIULLI; PIZZUTI, 2002)52
- Figura 8. Ilustração do conceito de distância de alcançabilidade. Fonte: Adaptado de (BREUNIG et al., 2000)55
- Figura 9. Determinação de *outliers* em espaços de dados com agrupamentos de densidade distinta. Fonte: Adaptado de (BREUNIG et al., 2000).....56
- Figura 10. Ilustração de uma caminhada do turista destacando a parte transiente (em amarelo) e o ciclo final (em verde).62
- Figura 11. Evolução dos atratores identificados para crescentes valores de μ num mesmo conjunto de dados de entrada. Cada imagem mostra todos os atratores identificados para o valor de memória fixado, diferentes cores nas linhas representam diferentes atratores.....64
- Figura 12. Comportamento dos atratores conforme incrementos no tamanho da memória. A linha azul mostra a evolução da quantidade de atratores, enquanto que as linhas laranja mostram as mudanças nos tamanhos médio, do maior e do menor atrator encontrado para cada valor de memória.65
- Figura 13. Três exemplos de atratores com cruzamento com diferentes quantidades de pontos de interseção (cor amarela) e semiciclos.66

- Figura 14. Dois atratores distintos detectados num conjunto de dados com dois agrupamentos em forma de banana. (a) Um atrator global é ilustrado, sem a presença de cruzamentos. (b) Um fragmento do conjunto de dados é destacado, onde ocorre um atrator com diversos cruzamentos. 67
- Figura 15. Comportamento dos atratores com os incrementos no tamanho da memória. (a) Conjunto de dados de entrada. (b) Comportamento dos atratores conforme incrementos no tamanho da memória. A linha azul mostra a evolução da quantidade de atratores, enquanto que as linhas laranja mostram as mudanças nos tamanhos médio, do maior e do menor atrator encontrado para cada valor de memória. 68
- Figura 16. Superposição de todos atratores identificados para todos os valores de memória até o antecessor do valor crítico, ilustrando a base para o mecanismo de detecção proposto. 72
- Figura 17. Exemplo de pontuação de *outlier* calculada para um conjunto de dados artificial. 74
- Figura 18. Cenário trivial com apenas um *outlier*. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. A linha de asteriscos vermelhos refere-se ao item de dado isolado. 80
- Figura 19. Cenário com agrupamentos de diferentes densidades. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados..... 81
- Figura 20. Cenário diverso com dois agrupamentos irregulares. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados..... 82
- Figura 21. Cenário com *outlier interno* e *externo*. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. A linha de asteriscos vermelhos e linha de asteriscos azuis (parcialmente sobrepostas) representam os itens de dado isolados interno e externo, respectivamente..... 83
- Figura 22. Cenário com *outlier interno* e *externo*, desconsiderando atratores com cruzamento. As cores representam a pontuação de *outlier* para cada ponto..... 84
- Figura 23. Experimento com a base *Wisconsin breast câncer*. Participação das amostras de dados nos atratores conforme crescimento do valor de memória. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados. 86

Figura 24. Experimento com a base <i>Ionosphere</i> . (a) Participação dos dados normais conforme o crescimento do valor de memória. (b) Participação dos <i>outliers</i> esperados conforme o crescimento do valor de memória.	87
--	----

LISTA DE TABELAS

Tabela 1. Conjunto de dados <i>Wisconsin breast cancer</i> : Comparação entre as técnicas de detecção de <i>outliers</i> mais conhecidas e a técnica proposta.	85
Tabela 2. Conjunto de dados <i>Ionosphere</i> : Comparação entre as técnicas de detecção de <i>outliers</i> mais conhecidas e a técnica proposta.	88

LISTA DE SÍMBOLOS

Notação	Descrição
O_i	<i>Outlier</i> isolado ou agrupamento de <i>outliers</i> de índice i .
A_i	Agrupamento de dados normais de índice i .
N	Número de elementos presentes no espaço de dados.
d	Número de dimensões (ou atributos) do conjunto de dados N .
O	Conjunto de todos os <i>outliers</i> de determinado espaço de dados; subconjunto de N .
q, r	Pontos quaisquer num espaço de dados; elementos de N .
ε	Raio da esfera d-dimensional que determinará uma ε -vizinhança.
$dist(q, r)$	Medida de distância entre os pontos q e r .
$N_\varepsilon(q)$	Conjunto dos pontos presentes na ε -vizinhança do ponto q .
$MinPts$	Quantidade mínima de pontos exigidos numa ε -vizinhança para determinar um ponto central.
N_C	Conjunto dos pontos centrais para o algoritmo DBSCAN.
N_B	Conjunto dos pontos de borda (fronteira) para o algoritmo DBSCAN.
N_R	Conjunto dos pontos de ruído para o algoritmo DBSCAN.
π	Percentual de <i>outliers</i> esperado; pelo método DB-Outlier.
$O(\varepsilon, \pi)$	Conjunto de <i>outliers</i> detectados pelo método DB-Outlier.
k	Quantidade de vizinhos mais próximos a serem considerados.
$D^k(q)$	Distância até o k -ésimo vizinho mais próximo de q .
$N_k(q)$	Vizinhança do ponto q com relação à quantidade de vizinhos k .

Notação	Descrição
$lrd_k(q)$	Densidade local da vizinhança do ponto q , utilizada pelo método LOF.
$LOF_k(q)$	Fator de <i>outlier</i> local estimada para o ponto q .
δ	Limiar para votação do algoritmo NDoT.
μ	Valor de memória da caminhada do turista; tamanho da janela de restrição.
t	Comprimento da trajetória transiente determinada por um caminhante turista.
p	Período do ciclo final (atrator) determinado por um caminhante turista.
μ_c	Valor crítico de memória do processo geral de caminhada do turista.
$\mu_{\text{máx}}$	Valor máximo de memória no processo de caminhada do turista; equivale a $N-I$.
S_i	Pontuação (ou <i>score</i>) de <i>outlier</i> atribuída ao item de dado com índice i .
$\Phi_i(\mu)$	Participação do item de dado i nos atratores das caminhadas com memória μ .

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO.....	27
1.1 Contextualização	27
1.2 Motivação e Objetivos	31
1.3 Organização do Trabalho.....	32
CAPÍTULO 2 - DETECÇÃO DE OUTLIERS.....	33
2.1 Considerações Iniciais.....	33
2.2 Definição de <i>Outlier</i> e Taxionomia.....	34
2.3 Tipos de <i>Outliers</i>	36
2.4 Aplicações	39
2.5 Diferentes Cenários de Dados	41
2.6 Principais Abordagens na Detecção de <i>Outliers</i>	42
2.6.1 Abordagem Estatística.....	43
2.6.2 Abordagem Baseada em Profundidade	44
2.6.3 Abordagem Baseada em Distância	45
2.6.4 Abordagem Baseada em Densidade	46
2.7 Principais Métodos de Detecção de <i>Outliers</i>	47
2.7.1 DBSCAN.....	47
2.7.2 DB-Outlier.....	51
2.7.3 kNN	52
2.7.4 LOF	53
2.7.5 Outros	57
2.8 Considerações Finais	59
CAPÍTULO 3 - CAMINHADA DETERMINÍSTICA DO TURISTA.....	61
3.1 Contextualização	61
3.2 Atratores Com Cruzamento	66
3.3 Detecção de Atratores	69
3.4 Considerações Finais	70
CAPÍTULO 4 - MÉTODO DE DETECÇÃO PROPOSTO	71

4.1 Nova Abordagem de Detecção	71
4.2 Cálculo da Pontuação de <i>Outlier</i>	73
4.3 Critérios de Controle da Caminhada do Turista	75
4.4 Algoritmo de Detecção de <i>Outlier</i>	76
4.5 Considerações Finais	77
CAPÍTULO 5 - EXPERIMENTOS COMPUTACIONAIS	79
5.1 Experimentos em Cenários de Dados Artificiais	80
5.1.1 <i>Outliers Internos</i>	83
5.2 Experimentos com Dados Reais	85
5.3 Considerações Finais	88
CAPÍTULO 6 - CONCLUSÕES	89
6.1 Publicações	90
6.2 Trabalhos Futuros	90
REFERÊNCIAS	93

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Outlier é uma observação que desvia tanto das outras observações de modo a levantar suspeita de que foi gerada por um mecanismo diferenciado (HAWKINS, 1980). Sendo, também definido, tipicamente, como uma observação inconsistente quando comparada à maioria dos dados; divergindo dos padrões esperados. E geralmente, são objetos ou eventos de ocorrência rara em um processo ou conjunto de dados, muitas vezes com características bastante variadas que dificultam a estipulação de um modelo capaz de distingui-los do padrão apresentado pelos demais dados considerados normais.

Assim, é difícil estabelecer uma definição conceitual rígida e precisa para os *outliers*, pois a anormalidade das amostras de dados é contextual, sendo fortemente dependente da aplicação e da natureza dos dados; que se mostram extremamente variadas. Vale mencionar inclusive, que uma definição rígida, tende a tornar-se muito restritiva para essa variabilidade de cenários existente. A literatura estatística emprega definições conceituais mais precisas (GRUBBS, 1969; TUKEY, 1977), porém essas definições não são apropriadas para as aplicações mais recentes do estado da arte que exigem maior robustez.

A detecção de *outliers* apresenta-se, então, como uma tarefa de extrema e crescente importância para descoberta de conhecimento e mineração de dados, especialmente ao lidarmos com problemas em ciências aplicadas. E embora seja um tema abordado de longa data através de métodos estatísticos, se renova como um tópico de pesquisa de grande relevância nos dias atuais, devido ao grande crescimento na disponibilidade de dados para os pesquisadores e indústria. Em diversos cenários, os dados são tantos que realizar o

processamento de todo o conjunto disponível é impraticável ou até mesmo indesejável. Assim, métodos capazes de selecionar aqueles dados com alto grau de distinção em meio a todo esse volume despertam grande interesse.

Identificar *outliers* é importante em diversas aplicações práticas, como: detecção de fraudes em cartões de crédito, detecção de erro humano, descoberta de eventos raros (por exemplo, na análise de dados astronômicos), detecção de intrusão em *cyber* segurança, detecção de falhas em linhas de produção, distúrbios em ecossistemas, análise médica de novas terapias para identificação de reações anormais, identificação de consumidores com grande potencial para ações de marketing e segmentação de mercado (AGGARWAL, 2013; HODGE; AUSTIN, 2004).

Existem aplicações onde o tratamento desejado consiste em apenas excluir os *outliers* do conjunto de dados, como um passo intermediário de filtragem desses dados rotulados como indesejáveis; o que consiste num tratamento mais clássico dos *outliers*. Para os casos em que esse tipo de tratamento é aplicado, normalmente, é adotada uma definição conceitual mais precisa de *outliers* (possivelmente em conformidade com as definições clássicas da estatística); isso ocorre (e deve ocorrer apenas) quando o significado desses valores anômalos é substancialmente conhecido e/ou de pouca (ou nenhuma) relevância semântica para o processamento e análise dos dados. Devemos ressaltar que nestes cenários apesar de pouca relevância semântica, a grande importância na filtragem dos *outliers* decorre do fato de que estes normalmente comportam-se como valores extremos e mesmo em pequeno número podem impactar significativamente nas medidas estatísticas da distribuição dos dados, prejudicando bastante a análise e possível tomada de decisão posterior.

Porém, outras aplicações são centradas nos *outliers*. Onde estes possuem grande valor semântico para antecipar e evitar situações críticas e de risco ou, até mesmo, para lucrar com eventos e situações raras ou de grande valor agregado (ANGIULLI et al., 2016; ANGIULLI; PIZZUTI, 2005; FRÉNEY; VERLEYSEN, 2016; KANG et al., 2016; SALEHI et al., 2016; SU; TSAI, 2011). Nessas aplicações o tratamento desejado consiste em evidenciar estes *outliers* para analisar com maior profundidade as características de cada amostra anômala e, em certos casos, até mesmo avaliar o mecanismo de geração das mesmas.

Para ilustrar possíveis consequências associadas ao tratamento incorreto (e interpretação) dos *outliers*, é interessante mencionarmos um caso de análise de dados com bastante impacto ecológico e relevância científica. O qual ocorreu no processamento dos dados obtidos pelo satélite Nimbus 7 da NASA (*National Aeronautics and Space Administration*), responsável por prover informações atmosféricas e, dentre elas, monitorar

dados relativos à camada de ozônio. Apesar de este satélite ter iniciado suas operações no ano de 1978, em 1985 a *British Antarctic Survey* divulgou uma pesquisa (FARMAN; GARDINER; SHANKLIN, 1985) reportando um fenômeno crítico de queda brusca e crescente nos níveis de ozônio da região da Antártida (durante a primavera) a partir de dados coletados diretamente em suas bases de operação; tendo sido essa pesquisa a responsável por identificar pela primeira vez o buraco presente na camada de ozônio sobre a Antártida. O curioso foi que os autores, inicialmente, hesitaram em publicar seus resultados, visto que os dados obtidos pelo Nimbus 7 (mais precisos) não apontavam esta mesma queda. Porém, após esse relato os pesquisadores da NASA realizaram um exame mais detalhado de seus dados e em 1986 publicaram (STOLARSKI et al., 1986) resultados evidenciando a observação do mesmo fenômeno. Então, foi identificado que a discrepância se devia ao programa de análise da captura dos dados que descartava observações abruptas e extremamente baixas (assumindo a presença de algum erro); o que acabou atrasando a descoberta do buraco na camada de ozônio em muitos anos, simplesmente por uma decisão inapropriada no tratamento dos *outliers*. Poucos anos depois, o mecanismo gerador dessa anomalia foi descoberto e relacionado com a ação de clorofluorocarbonetos (CFC) produzidos pelo Homem sobre o ozônio presente na estratosfera terrestre (CHRISTIE, 2001).

São muitos os algoritmos de detecção de *outliers* já propostos e, igualmente, diversificadas as abordagens conceituais que tomam por base (CHANDOLA; BANERJEE; KUMAR, 2009; HODGE; AUSTIN, 2004; SU; TSAI, 2011). Para exemplificar, podemos citar algumas: abordagem estatística, baseada em profundidade, baseada no desvio, baseada em proximidade (ou distância), baseada em densidade, baseada em ângulo, baseada em agrupamento, baseada em rede, baseada na teoria da informação e abordagem de alta dimensionalidade.

Cada uma dessas abordagens apresenta pressupostos teóricos próprios que visam resolver certos tipos de dificuldades encontradas nos diversos cenários possíveis. Assim, a detecção de *outlier* se manifesta fortemente dependente do contexto do problema e da natureza do conjunto de dados, onde determinadas abordagens podem apresentar melhor ou pior acoplamento a determinado problema ou conjunto de problemas.

Porém, ao analisarmos de maneira mais geral, o princípio central da detecção de *outliers* parece se sustentar nas suposições acerca dos padrões formados pelo conjunto de dados (AGGARWAL, 2013), os quais normalmente não são conhecidos e são complexos de serem modelados matematicamente. Por esta razão, vários métodos de detecção de *outliers* com base nos princípios de aprendizado de máquina foram desenvolvidos e propostos nos

últimos anos, alcançando excelentes resultados em espaços de dados bastante irregulares, volumosos e de diferentes dimensionalidades.

Um algoritmo com propriedades interessantes para extração de padrões num espaço de dados é a caminhada do turista (do inglês *tourist walk*), uma estratégia de exploração determinística abstraída como um “turista” caminhando sempre em direção à cidade mais próxima, desde que essa ainda não tenha sido visitada nos últimos passos (definido por um parâmetro de memória do modelo). Apesar de ser um mecanismo de exploração aparentemente simples, a dinâmica gerada por este pode extrair informações importantes como conjuntos de pontos atratores; locais de onde o caminhante não mais escapa, estando preso num ciclo (LIMA; MARTINEZ; KINOUCI, 2001).

Esse modelo já foi utilizado, por exemplo, para explicar o movimento migratório de pássaros que seguem uma rota determinada, revisitando certos sítios de alimentação numa mesma sequência em busca de maior disponibilidade de recursos alimentares durante o ano e permitindo que esses se renovem em tempo do ciclo se repetir. Inclusive, o modelo permite que esse evento seja descrito com base apenas no conhecimento das características locais do espaço (pelos indivíduos), sem a necessidade de conhecimento das características globais do sistema (impraticável). Essa abstração pode, inclusive, ser expandida, filosoficamente, para tentar explicar alguns padrões comportamentais que interpretamos como livre arbítrio, mas seriam apenas reflexo da aplicação de regras determinísticas num ambiente estocástico (STANLEY; BULDYREV, 2001).

Os atratores identificados pelos caminhantes para diferentes pontos de partida e valores de memória parecem revelar propriedades estruturais do espaço de dados que exploram. Propriedades essas que servem, inclusive, ao propósito de reconhecimento de padrões invariante (SILVA; ZHAO, 2015) com resolução local para global (conforme incremento no parâmetro de memória). Portanto, a caminhada do turista é essencialmente uma caminhada parcialmente autorrepulsiva controlada por um único parâmetro de memória, sendo capaz de reconhecer padrões através da identificação de ciclos (atratores) onde seus caminhantes ficam presos (CAMPITELI et al., 2006)

Neste trabalho é proposto, então, um método de detecção de *outliers* que faz uso do algoritmo de caminhada do turista num dado conjunto de dados (BACKES; MARTINEZ; BRUNO, 2011; CAMPITELI et al., 2006; LIMA; MARTINEZ; KINOUCI, 2001), utilizando os atratores identificados como representantes dos mecanismos geradores dos dados normais. E, por consequência, determinando os *outliers* como sendo aqueles exemplares de dados que possuem pouca participação nesses atratores; numa varredura local

para global pelo espaço de dados. A capacidade de identificar estruturas presentes no espaço de dados ao invés de considerar apenas características físicas (como distância, similaridade e densidade), pretende tornar este método, naturalmente, mais robusto e capaz de detectar padrões de alto nível.

1.2 Motivação e Objetivos

Dentre as motivações que nos direcionaram para concepção deste trabalho, devemos ressaltar a crescente importância da detecção de *outliers* para inúmeras aplicações, apresentando grande relevância para diversos campos das ciências aplicadas. E sendo por consequência um tópico importante para ser investigado.

Entre as diversas propriedades do algoritmo de caminhada determinística do turista que nos atraíram a atenção, podemos citar:

- Os caminhantes ao determinar atratores em seu percurso revelam propriedades estruturais do espaço de dados; possibilitando o reconhecimento de padrões.
- Os atratores extraídos apresentam certas propriedades típicas de reconhecimento de padrões invariante e de alto nível, pois não se baseiam apenas nas características físicas do espaço, como distância, similaridade e densidade.
- O processo de caminhada, com incrementos constantes nos valores de memória, dá-se de uma maneira local para global no conjunto de dados.
- O algoritmo de caminhada possui apenas um parâmetro, para atualização do valor de memória.

Todas essas características apresentadas acima são muito relevantes para o problema de detecção de anomalias e acabaram inspirando e motivando a proposição do novo método. Além disso, a abordagem apresentada neste trabalho é diferenciada daquelas tradicionais, seguindo então numa direção certamente inovadora e de aspecto mais geral.

Inclusive, durante o desenvolvimento da pesquisa foi caracterizado um novo fenômeno, decorrente das caminhadas do turista: a extração de atratores com cruzamento (com dinâmicas mais complexas que os atratores convencionais). Esse achado também motivou posteriormente o estudo do conceito de *outlier interno*; levando a uma melhora de precisão da técnica.

Por fim, o objetivo geral da pesquisa é a proposição e avaliação de um método de detecção de *outliers* inédito que se utiliza do algoritmo de caminhada do turista para extrair dos conjuntos de dados padrões de dados de alto nível, possibilitando a identificação de anomalias de maneira mais robusta e automatizada.

1.3 Organização do Trabalho

O Capítulo 2, intitulado “Detecção de *Outliers*”, apresenta uma revisão sobre *outliers* e as diferentes abordagens de detecção já propostas na literatura, reforçando os aspectos teóricos e aplicados que serão importantes para posteriormente melhor caracterizar o método que propomos.

No Capítulo 3, intitulado “Caminhada Determinística do Turista”, dissertamos a respeito do algoritmo que nos inspiramos para propor a nova técnica, esclarecendo como se dá o processo de caminhada e descrevendo as dinâmicas apresentadas por este e as estruturas envolvidas nesse processo; ressaltando aquelas propriedades que posteriormente serão aproveitadas em nossa proposta.

Já no Capítulo 4, intitulado “Método de Detecção Proposto”, será apresentado e descrito o novo método de detecção de *outliers* que é fruto deste trabalho. Serão discutidos os aspectos conceituais da nova abordagem de detecção, o mecanismo de diferenciação dos dados anormais e os critérios de controle da caminhada do turista.

O Capítulo 5, intitulado “Experimentos Computacionais”, apresenta e analisa os resultados obtidos nos diversos experimentos computacionais realizados com o novo método, tanto em cenários artificiais quanto em bases de dados reais.

Por fim, as conclusões, contribuições e trabalhos futuros estão descritas no Capítulo 6, intitulado “Conclusões”.

Capítulo 2

DETECÇÃO DE OUTLIERS

O presente capítulo apresentará a definição de outlier, sua relevância, aplicações, tipos e outros aspectos relevantes para a detecção e identificação desses nos conjuntos de dados. Também apresentará as dificuldades oriundas do campo de aplicação e as principais técnicas de detecção no estado da arte. Com isso, pretendemos revisar os conceitos que concernem ao mecanismo de detecção que propomos no presente trabalho.

2.1 Considerações Iniciais

Detecção de *outlier* é uma tarefa de extrema importância para descoberta de conhecimento e mineração de dados ao lidarmos com problemas aplicados. E embora seja um tema já estudado de longa data através de métodos estatísticos, ainda permanece como um tópico de pesquisa de grande interesse na atualidade, especialmente com o grande crescimento na disponibilidade e dimensão dos dados para os pesquisadores e a indústria.

O cenário atual de aplicação das técnicas de detecção de *outliers* apresenta-se cada vez mais amplo e generalizado. Espera-se que estas técnicas se adaptem aos cenários de dados mais variados e desafiadores com: quantidades massivas de dados, alta dimensionalidade, aquisição de dados incremental e dinâmica, tipos de dados mistos, indisponibilidade de dados anotados, entre outros (SU; TSAI, 2011). Portanto, esse cenário deu origem e destaque às abordagens de detecção que fazem uso dos princípios da área de aprendizado de máquinas, possibilitando o surgimento de diversas técnicas generalistas e de grande robustez. Isso se dá, pois a grande variabilidade desses cenários de aplicação torna muito difícil ou extremamente custoso estabelecer modelos estatísticos adequados, que representem bem os dados normais, e, conseqüentemente, dificultam a determinação de um mecanismo para identificação dos

outliers com base naquelas amostras que não se adequam à função geradora dos dados normais; ou seja, em muitos desses cenários as técnicas de detecção de *outliers* trabalham sem nenhum conhecimento prévio acerca dos dados ou apenas com alguma noção da natureza desses dados.

2.2 Definição de *Outlier* e Taxionomia

Uma definição conceitual de *outlier* não é direta, mas fortemente relacionada ao contexto específico do problema ao qual se aplica. Porém, uma definição bem aceita é dada em (HAWKINS, 1980), onde *outlier* é caracterizado como “uma observação que desvia tanto das outras observações que levanta suspeita de que foi gerada por um mecanismo diferenciado”. Enquanto em (GRUBBS, 1969) é definido como uma observação periférica que pode ter duas origens distintas: ser “meramente uma manifestação extrema da variedade aleatória inerente aos dados”, ou ainda “resultado de um desvio grosseiro do procedimento experimental”. Portanto, de maneira mais ampla, *outliers* podem ser entendidos, intuitivamente, como observações inconsistentes quando comparados à maioria dos dados (também denominados por *inliers*) (BARNETT; LEWIS, 1994) e que divergem dos padrões esperados. Sendo, normalmente, objetos ou eventos de ocorrência rara em um conjunto de dados ou processo; pela própria definição.

É comum que autores de áreas específicas do conhecimento (ou de áreas específicas de aplicação) estabeleçam definições mais precisas para o termo *outlier*, porém ao fazê-lo tornam também o escopo do conceito mais restritivo, que acaba por não se adequar às diversas interpretações nos diferentes campos de aplicação existentes. Assim, entendemos que as definições intuitivas mais amplas são suficientemente adequadas para definição conceitual e aquelas mais precisas devem ser estabelecidas apenas no tratamento dos contextos diferenciados das aplicações e nas proposições dos métodos que deverão lidar com estes dados de características bem determinadas.

Para exemplificar a complexidade na determinação de uma definição conceitual precisa e o amplo espectro de campos de aplicação, basta mencionar a vasta terminologia apresentada na literatura: anormalidades, anomalias, contaminantes, desviantes, discordantes, exceções, novidades (AGGARWAL, 2013; AGGARWAL; SATHE, 2017; HAWKINS, 1980; SU; TSAI, 2011); são todos termos diferentes que podem fazer referência a *outliers*. Vale

observar, que certos autores podem entender que estas terminologias se diferenciam umas das outras e estabelecem definições diferenciadas que se adequam aos seus problemas de pesquisa, porém não há um consenso na comunidade científica e novamente uma compreensão mais ampla é mais adequada.

Apesar dessa amplitude conceitual mencionada, é importante fortalecer e esclarecer a distinção entre ruído e *outlier*. Pois assim como ocorre com a definição de *outlier*, também não há consenso na literatura sobre esta distinção; talvez até por consequência da ampla definição e subjetividade do termo *outlier* (como esclarecemos acima). Porém, há extrema importância na distinção destes conceitos e consideramos a seguinte diferenciação apropriada: o ruído está associado à imprecisão (relacionada a limitações na medida ou até mesmo à natureza estocástica dos dados normais, se referindo a quaisquer flutuações aleatórias de dados que impeçam a percepção de um sinal esperado) (STEIN; GOSSEN; JONES, 2005), enquanto que *outliers* são exemplos de dados que podem carregar informação valiosa, relevante ou até mesmo crítica que se destaca do restante (os quais normalmente possuem um mecanismo gerador específico e não aleatório, mas difícil de ser encontrado dada a raridade de ocorrência) (KANG et al., 2016; CHANDOLA; BANERJEE; KUMAR, 2009; TAN; STEINBACH; KUMAR, 2005).

Essa distinção está ilustrada com maior clareza na Figura 1, onde são apresentados dois cenários com o mesmo espaço de dados na presença de um *outlier* (denotado por A, na cor vermelha), porém na Figura 1(a) não há presença de ruído, enquanto que na Figura 1(b) este se faz presente. Obviamente, é possível deduzir que a identificação de *outliers* em espaços onde haja forte presença de ruídos se apresentará como uma tarefa ainda mais complexa, ou, em alguns casos, até mesmo impraticável pela indissociação dos elementos *outlier* com relação ao ruído; por permear todo o espaço de dados, acaba por dificultar a determinação do mecanismo gerador dos dados normais e com maior intensidade a detecção dos *outliers*, que se manifestam como eventos raros e normalmente dispersos. Por consequência disso, alguns autores preferem considerar que o termo *outlier* engloba o (é um superconjunto do) conceito de ruído, possuindo esse último uma distribuição ligeiramente mais comportada que os *outliers* (AGGARWAL, 2013). Assim, em vários cenários a distinção destes dois conceitos é bastante sutil ao visualizarmos a manifestação no espaço de dados, porém muito evidente quando conhecida a natureza de geração dos dados; o que normalmente não é conhecimento prévio.

Figura 1. Distinção entre *outlier* e ruído. (a) Um espaço de dados com um *outlier* denotado por A (em vermelho). (b) O mesmo espaço de dados, ainda com o *outlier* A, porém na presença de ruído.

Fonte: Adaptado de (AGGARWAL, 2013)

2.3 Tipos de *Outliers*

Os diversos contextos em que a análise de *outliers* pode se manifestar permitem a diferenciação e classificação desses *outliers* em diversos tipos. Estes são determinados e nomeados conforme os distintos comportamentos que podem apresentar quando inseridos em diferentes cenários de dados e contextos de aplicação. É comum encontrar na literatura a distinção em três categorias de *outliers*: *outliers* pontuais, *outliers* contextuais e *outliers* coletivos (CHANDOLA; BANERJEE; KUMAR, 2009).

- ***Outlier* pontual:** é o tipo mais comum e intuitivo de *outlier*. Consiste naquela observação considerada anormal quando comparada aos outros objetos como um todo. Cenários de dados onde os dados normais tendem a formar algum tipo de agrupamento, situando-se próximos entre si, e caracterizando o *outlier* pontual como aquele que está evidentemente distante de todos os demais pontos. Deve-se observar que pequenos agrupamentos de *outliers* (que satisfaçam a definição acima) ainda são considerados *outliers* pontuais. É preciso haver cuidado para não interpretá-los erroneamente como *outliers* coletivos, bastando avaliar se cada um dos *outliers* do agrupamento ainda se manifesta como uma anomalia mesmo no

caso de apresentarem-se isoladamente. Um exemplo de *outlier* pontual está ilustrado na Figura 2(a), destacado com a coloração vermelha.

- **Outlier contextual:** cenário onde um objeto é considerado anormal num certo contexto específico. Assim, esse tipo de *outlier* pode assumir um valor contido no mesmo intervalo de valores dos dados normais, porém possui características anômalas pelo contexto (ou momento) em que se manifesta; Figura 2(c).
- **Outlier coletivo:** ocorrem em cenários onde a anomalia consiste não na localização e valor assumido pela observação, mas na periodicidade anômala quando comparada com o todo. Assim, o *outlier* coletivo exige a análise de uma faixa ou grupo de observações que se manifestam coletivamente como uma anomalia, pois ao serem analisadas de forma desassociada, como observações individuais, são similares aos dados normais. Um exemplo desse tipo de *outlier* pode ser observado na Figura 2(b), onde um *outlier* coletivo está destacado na cor vermelha.

Figura 2. Tipos de outliers. (a) Outlier pontual. (b) Outlier coletivo. (c) Outlier contextual.
Fonte: Adaptado de (AGGARWAL, 2013)

Além dos tipos apresentados acima, neste trabalho, nós fazemos distinção entre dois tipos de *outliers*:

- a) ***outliers externos***: aqueles geometricamente localizados na região periférica do agrupamentos de dados, sendo aquelas amostras anômalas mais comumente tratadas nos diversos trabalhos publicados;
- b) ***outliers internos***: *outliers* localizados na região central entre dois ou mais agrupamentos de dados.

Por conveniência, neste trabalho quando mencionarmos apenas “*outlier*”, estaremos nos restringindo aos *outliers externos*. Ao passo que ao lidarmos com cenários de *outliers internos*, a nomenclatura estará diferenciada de forma explícita. Vale observar que as definições apresentadas podem ser interpretadas como subclassificações dos *outliers* pontuais.

Certamente, *outliers internos* são mais difíceis de detectar do que *outliers externos*, dado que os agrupamentos de dados estão normalmente espalhados pelo espaço de dados, mas não evidentemente separados (HUBERTY; OLEJNIK, 2006), dificultando bastante a diferenciação dos *outliers internos* daqueles elementos de dados regulares. Se os agrupamentos de dados estão bem separados, os *outliers* acima mencionados tendem a enquadrar-se no comportamento esperado para os *outliers externos* (mais usuais) e conformariam com os paradigmas de detecção empregados pelas diversas técnicas de detecção de *outliers* convencionais.

A Figura 3 ilustra a presença de um *outlier interno* (marcado com um asterisco vermelho) situado entre dois agrupamentos regulares de dados, cada um deles seguindo uma distribuição normal. Quanto mais os grupos se aproximam do *outlier interno*, mais difícil será a sua detecção, pois será necessário distingui-lo dos dois padrões distintos (o que na maioria dos casos não é um conhecimento prévio e é de difícil reconhecimento; por ser um padrão múltiplo). Também estão presentes na Figura 3, outros três *outliers externos* (marcador preto de forma cruzada). Se estes *outliers externos* se aproximarem demasiadamente de um dos grupos, a dificuldade de detecção também será aumentada. Mas, nessa situação, será suficiente verificar que tais exemplos de dados não conformam com o padrão apresentado pelo agrupamento de dados mais próximo. Portanto, esta tarefa de detecção tem um envolvimento mais local quando comparado à detecção de *outlier interno*; onde deverão ser cheçadas as conformações com diferentes padrões apresentados pelos diversos agrupamentos de dados (e seus mecanismos geradores distintos) presentes num reconhecimento com características local para global.

Figura 3. *Outlier interno* (asterisco vermelho) entre dois agrupamentos e três *outliers externos* (marcadores em cruz de cor preta).

2.4 Aplicações

Como descrito na subseção de taxionomia do presente capítulo, para cada tipo de aplicação, o termo “*outlier*” apresenta uma conotação específica e dependente do contexto. Portanto, o mesmo também ocorre com a tarefa que se dará após a detecção, ou seja, devemos ressaltar que para cada aplicação é destinado um tratamento específico para estas amostras anômalas. Em certas aplicações a intenção pode ser apenas excluir ou ignorar os *outliers* do conjunto de dados, como um passo intermediário do processamento de dados, porém em outras os *outliers* são exemplos de dados desejáveis; seja para antecipar e evitar situações críticas e de risco ou, até mesmo, para lucrar com eventos e situações raras e ou de grande valor, onde sua identificação consiste no grande propósito de todo o processamento dos dados (ANGIULLI et al., 2016; ANGIULLI; PIZZUTI, 2005; FRÉNAY; VERLEYSSEN, 2016; KANG et al., 2016; SALEHI et al., 2016; SU; TSAI, 2011).

Assim, existem diversas aplicações onde a tarefa de detecção de *outliers* possui grande relevância, presentes numa grande variedade de domínios, como:

- **Financeiro (detecção de fraudes):** são aplicações onde é desejável diferenciar e identificar possíveis transações fraudulentas e criminosas que ocorram em sistemas comerciais. Aplicações específicas desse domínio são largamente adotadas na

detecção de fraudes em: transações com cartões de crédito, chamadas de telefonia celular, reivindicação de benefícios de seguradoras ou benefícios estatais, lances em leilões e operações no mercado de ações (CHANDOLA; BANERJEE; KUMAR, 2009). Grande parte dessas aplicações estabelece como estratégia o monitoramento de atividades em tempo real, onde é caracterizado e mantido um perfil de atividade de cada cliente (ou empresa) para detectar comportamentos incomuns que desviem substancialmente deste (KOU et al., 2004). No caso de uma anomalia, diferentes intensidades no desvio podem direcionar para fluxos de tratamento específicos como: impedimento da transação, separação e direcionamento para avaliação de um especialista de domínio ou até mesmo o bloqueio geral da atividade daquele cliente.

- **Medicina e saúde pública:** são diversas as aplicações possíveis na área médica e de saúde pública. Desde aquelas centradas no paciente, onde é realizado o monitoramento de suas condições médicas para diagnosticar doenças. O que pode ocorrer a partir de seus sinais vitais, registros médicos ou a partir de exames de imagem. Podendo também decorrer da análise de séries temporais obtidas a partir de sensores nos eletrocardiogramas ou eletroencefalogramas. Outro enfoque aplicado possível é o epidemiológico, onde o objetivo é identificar precocemente os surtos de doenças em regiões ou condições específicas a partir de grandes bases de dados de pacientes (WONG et al., 2003).
- **Descoberta científica (eventos raros):** muitas pesquisas científicas podem se beneficiar da descoberta de eventos raros (uma das possíveis interpretações para *outliers*), principalmente nos cenários onde há a aquisição constante de enormes quantidades de dados; implicando num custo computacional que inviabiliza o processamento e análise de todos esses dados obtidos. Assim, métodos de detecção de *outliers* são aplicados para auxiliar na distinção daquelas situações inusitadas, que podem significar uma nova descoberta científica ou eventos de maior interesse científico para execução de uma análise mais detalhada (seja computacional ou investigativa). Como exemplo, um dos campos de pesquisa que faz uso intenso desse contexto de aplicação é a astronomia (e astrofísica), pois diariamente são coletadas quantidades exorbitantes de dados que impõem a necessidade de tratamento automatizado (HILBE, 2014). Uma aplicação específica que podemos

menção é o SETI (busca por inteligência extraterrestre, do inglês *Search for Extraterrestrial Intelligence*) (PATHAK, 2014). Esse contexto também engloba pesquisas de perturbações dos ecossistemas, bancos de dados genômicos e proteômicos; entre diversas outras aplicações de descoberta de conhecimento científico.

Além dos domínios de aplicação mencionados acima, podemos citar também as aplicações seguintes: detecção de intrusão em *cyber* segurança, detecção de falhas em linhas de produção, detecção e filtragem de erro humano, identificação de consumidores com grande potencial para ações de marketing e segmentação de mercado, detecção de danos e defeitos estruturais a partir de imagens (AGGARWAL, 2013; CHANDOLA; BANERJEE; KUMAR, 2009; HODGE; AUSTIN, 2004). Assim, os vastos e diferenciados campos de aplicação da detecção de *outliers* deixam evidente sua relevância e o potencial impacto e contribuição da evolução do estado da arte das técnicas de detecção.

2.5 Diferentes Cenários de Dados

Devido aos diversos contextos de aplicação existentes para a análise de *outliers* (mencionados na seção anterior), são variados e bastante distintos os cenários de dados onde os algoritmos de detecção poderão ser executados. Portanto, as técnicas de detecção de *outliers* podem ser classificadas de acordo com as características apresentadas pelo cenário de dados onde deverá ser aplicada.

Ao considerarmos os atributos dos conjuntos de dados de entrada a partir da visão do aprendizado de máquina, podemos dividir os algoritmos em três classes distintas: supervisionados, não supervisionados e semissupervisionados (AGGARWAL, 2013; SILVA; ZHAO, 2016). Métodos supervisionados são aqueles utilizados em cenários onde é possível aplicar procedimentos de treinamento, por haver disponibilidade de um conjunto de dados rotulados de tamanho suficientemente grande, enquanto que os métodos não supervisionados são mandatórios quando estão disponíveis apenas dados não rotulados; ou seja, não há informação prévia sobre a classificação dos dados. Já os semissupervisionados são aqueles que se adequam aos cenários híbridos, onde apenas uma pequena proporção dos dados

possuem rótulos, não o suficiente para prosseguir com uma fase de treinamento, mas de valia para melhorar a precisão do método de detecção.

Ao analisarmos os cenários com relação ao tipo de saída esperado, os algoritmos de detecção de *outlier* podem tanto retornar um rótulo (também denominado por *hard-label*) para cada item de dado (podendo ser “normal” ou “*outlier*”), quanto uma pontuação (do inglês *soft-label*). Porém, vale ressaltar que este último pode ser facilmente convertido para um rótulo binário num momento tardio, através da imposição de um limiar.

A escolha de um algoritmo de detecção de *outliers* apropriado está diretamente relacionada com o contexto do problema e com a natureza do conjunto de dados (cenário de dados). Porém, a maioria dos métodos disponíveis na literatura são propostos para atuar em cenários não supervisionados, com a devolução de uma lista de pontuações (como saída) (ZIMEK; SCHUBERT; KRIEGEL, 2012). Isso se deve às principais dificuldades enfrentadas no campo aplicado, já que devido os *outliers* serem tipicamente eventos raros, os exemplos rotulados não costumam estar disponíveis (ou seja, comumente não há nenhum conhecimento a priori sobre o conjunto de dados de entrada). Assim, nesta situação fortemente desbalanceada, torna-se impraticável o uso de técnicas supervisionadas. Exibir os rótulos na saída ao invés da pontuação também não é desejável na maioria das aplicações, pois os métodos que exibem como saída um *soft-label*, se beneficiam do fato de não ser necessário fixar, inicialmente, o número esperado de *outliers* (informação que comumente também é desconhecida); tornando-os mais robustos e auxiliando no aumento da precisão média destes métodos. Observamos que algumas técnicas de detecção podem se utilizar de uma noção da intensidade de desvio para cada item de dados, permitindo distinguir quais exemplos de dados possuem características mais fortes de *outlier* dentre aqueles identificados (CHANDOLA; BANERJEE; KUMAR, 2009).

2.6 Principais Abordagens na Detecção de Outliers

É comum categorizar as técnicas de detecção de *outliers* em classes relacionadas às diferentes abordagens de detecção empregadas na identificação dos *outliers*, das quais podemos citar as seguintes: abordagem estatística, baseada em profundidade, baseada no desvio, baseada em proximidade (ou distância), baseada em densidade, baseada no ângulo, baseada em agrupamento, baseada em rede, baseada na teoria da informação e abordagem de

alta dimensionalidade (CHANDOLA; BANERJEE; KUMAR, 2009; HODGE; AUSTIN, 2004; SU; TSAI, 2011). Cada uma dessas abordagens prevê diferentes suposições e restrições acerca do comportamento dos *outliers* e utilizam-se de diferentes referenciais teóricos para iterar sobre os conjuntos de dados de maneira a reconhecer corretamente os conjuntos de referência (aquele que contém os dados ditos normais) e identificar esses *outliers*.

Um dos aspectos que está de certa forma embutido nas diversas abordagens existentes consiste na granularidade do conjunto de referência a ser considerado para determinar quais amostras são anômalas, neste quesito dois tratamentos são possíveis: global e local. Na resolução global presume-se que o conjunto de referência é global e, portanto, contém todos os dados, exceto os *outliers*. Assim, assume-se que há apenas um mecanismo gerador dos dados normais. Já o tratamento local possibilita a coexistência de vários conjuntos de referência no espaço de dados e, portanto, não estabelece previamente uma quantidade fixa de mecanismos geradores para os dados (AGGARWAL, 2013). No tratamento global é comum que certos *outliers* sejam forçados a participar do conjunto de referência para que outros sejam identificados, enquanto que nas abordagens locais o grande problema é a complexidade em se definir uma boa estratégia para identificar os conjuntos de referência apropriados e a granularidade a ser adotada. Algumas abordagens de detecção podem, inclusive, se utilizar de um tratamento misto, vasculhando diferentes resoluções automaticamente ou através de algum parâmetro definido pelo usuário.

2.6.1 Abordagem Estatística

As primeiras técnicas de detecção de *outlier* propostas foram baseadas na abordagem estatística, sendo importante ratificar que o próprio termo “*outlier*” e as primeiras definições conceituais acerca deste surgiram no campo da estatística (GRUBBS, 1969). Simplificadamente os métodos de detecção dessa classe assumem determinada distribuição de probabilidade ou modelo estatístico para o conjunto de dados, sendo, então, possível aplicar diversos testes estatísticos para diferenciar e identificar as amostras que mais desviam da distribuição assumida. Assim, os dados normais serão aqueles que se situam nas regiões de alta probabilidade do modelo e *outliers* aquelas observações que desviam fortemente da distribuição assumida no modelo adotado.

Os aspectos negativos dessa abordagem se devem à falta de robustez das técnicas, pois apenas um *outlier* (com valor extremo; e muitas vezes caracterizado exatamente por esse comportamento) pode impactar significativamente nas medidas estatísticas (por exemplo,

média e desvio padrão) da distribuição em uso. Normalmente a distribuição é acoplada (vinculada) aos dados levando-se em conta todo o conjunto, ou seja, acaba incluindo *outliers* em potencial; tratamento global. E uma vez estabelecida uma distribuição para os dados, essa permanece fixada; porém em cenários incrementais essa distribuição provavelmente se mostrará inapropriada com o tempo.

Por fim, no caso de análise estatística não paramétrica, o problema consiste na grande dificuldade em determinar um modelo estatístico apropriado sem informação prévia sobre o comportamento dos dados e dos *outliers*; o que é ainda mais crítico nos conjuntos de alta-dimensionalidade (HAWKINS, 1980).

2.6.2 Abordagem Baseada em Profundidade

Neste conjunto de técnicas, a premissa é que os *outliers* frequentemente situam-se nas bordas do espaço de dados; independentemente da distribuição estatística que venham a seguir. E conseqüentemente, as amostras de dados normais ocupam o centro do espaço.

A estratégia comumente adotada é determinar camadas geométricas convexas numa busca hierárquica em profundidade, seguindo das camadas mais externas até a definição das mais internas (numeradas nesse mesmo sentido, de maneira crescente). Após isso, determina-se um valor de corte e as camadas abaixo desse valor (portanto mais externas) serão consideradas *outliers*; ou seja, as camadas serão sempre descartadas no sentido externo-interno (PREPARATA; SHAMOS, 1993).

Figura 4. Exemplo de detecção de *outlier* com abordagem baseada em profundidade.
Fonte: Adaptado de (PREPARATA; SHAMOS, 1993)

Um exemplo desse mecanismo está ilustrado na Figura 4, onde (a) ilustra o conjunto de dados a ser analisado e (b) mostra as camadas convexas já demarcadas para este mesmo conjunto. O valor de corte deverá, então, ser definido para determinar quantas camadas serão excluídas da mais externa para mais interna de modo a eliminar os *outliers*.

Dentre os principais problemas apresentados por essa abordagem, podemos destacar que a complexidade computacional da tarefa de demarcação das camadas convexas cresce exponencialmente com a dimensionalidade do espaço de dados (AGGARWAL, 2013). E, além disso, por uma consequência geométrica, quão maior a dimensionalidade, maior será a proporção dos itens de dados situados nas camadas mais externas (PREPARATA; SHAMOS, 1993); um problema muito grave naqueles cenários onde se supõe que os *outliers* se manifestam como elementos de ocorrência rara. Por fim, é importante observar, ainda, que esta abordagem se caracteriza como um tratamento de resolução global.

2.6.3 Abordagem Baseada em Distância

Esta é uma das abordagens mais produtivas, que deu origem a diversas variantes e inúmeras técnicas de detecção de *outliers*. A ideia central desta abordagem é que cada amostra de dado seja julgada (avaliada) com base na distância para seus vizinhos; estabelecendo como premissa que os dados normais devem ter uma vizinhança densa, enquanto que os *outliers* seriam aqueles pontos localizados em vizinhanças com densidade mais baixa quando comparada ao restante dos dados (KNORR; NG; TUCAKOV, 2000). Essa abordagem é considerada local, porém boa parte das técnicas dessa família possui um parâmetro de ajuste de resolução para possibilitar aplicação em cenários mais diversificados, porém algumas controlam a variação da resolução automaticamente.

Porém, um dos problemas apresentados por estas técnicas ocorre quando o espaço de dados a ser analisado possui regiões com diferentes densidades (como no cenário da Figura 5), fazendo com que para uma resolução fixada certas amostras de dados não se manifestem como *outliers*, por não haver uma premissa estabelecida a respeito da comparação de áreas com diferentes densidades. Por conta disso, os métodos baseados em distância podem ser interpretados como métodos de densidade global.

Dentre as várias técnicas de detecção de *outliers* que se enquadram nessa abordagem, podemos destacar as seguintes: DBSCAN (ESTER et al., 1996), DB-Outlier (KNORR; NG, 1997, 1998), kNN (RAMASWAMY; RASTOGI; SHIM, 2000) e kNNW (ANGIULLI; PIZZUTI, 2002); estas técnicas possuem uma descrição mais detalhada na seção 2.7.

2.6.4 Abordagem Baseada em Densidade

A abordagem baseada na densidade é uma evolução da abordagem baseada em distância, que surgiu com o propósito de resolver o problema de detecção de *outliers* em espaços de dados com agrupamentos de diferentes densidades (AGGARWAL, 2013).

Figura 5. Espaço de dados com agrupamentos de diferentes densidades e dois *outliers* em regiões distintas.
Fonte: Adaptado de (AGGARWAL, 2013)

O cenário da Figura 5 ilustra a motivação para o surgimento dos métodos baseados em densidade. Pois, neste espaço de dados os métodos baseados em distância conseguiriam identificar apenas o *outlier* O_1 , não obtendo sucesso no caso de O_2 . Pois, apesar de O_2 situar-se afastado do agrupamento de alta densidade A_2 , essa distância é similar ou inferior àquela que separa os elementos normais do agrupamento de baixa densidade A_1 ; dando condições para que o *outlier* O_2 ainda respeitasse a densidade global do cenário e não fosse corretamente identificado como dado anômalo.

Os métodos de detecção deste grupo consideram a densidade relativa durante o processo de iteração sobre o espaço de dados. Sendo a premissa básica a seguintes: a densidade ao redor de um dado normal é similar à densidade dos seus vizinhos, enquanto que as amostras anômalas possuem densidade consideravelmente diferente dos seus vizinhos.

Assim, o mecanismo de detecção consiste em comparar a densidade ao redor de um ponto com a densidade dos seus vizinhos mais próximos, quando, então, será possível estabelecer uma pontuação de *outlier* com base na intensidade dessa disparidade. Os diferentes métodos existentes diferem, basicamente, quanto à estratégia utilizada para estimar estas densidades.

O LOF (BREUNIG et al., 2000) foi o método de detecção de *outliers* responsável por originar esta nova abordagem baseada em densidade. Sua importância foi tamanha que diversos outros métodos foram propostos com base no seu núcleo de funcionamento, dando origem a uma família de métodos LOF (do inglês, *LOF family methods*); uma descrição mais detalhada sobre o funcionamento do método LOF pode ser encontrada no item 2.7.4.

2.7 Principais Métodos de Detecção de Outliers

A quantidade de métodos de detecção existente é muito vasta, refletindo a grande e crescente importância da tarefa de detecção e análise de *outliers*, pelos diversos fatores que já mencionamos nas seções anteriores. Portanto, não pretendemos nesta seção realizar uma revisão extensa e nem mesmo descrição histórica rigorosa dos muitos métodos existentes, suas variações e evolução. Por outro lado, damos destaque àqueles que serviram como ponto de inflexão para o avanço da área ou revelaram mecanismos valiosos para proposta de abordagens diferenciadas na detecção dos *outliers* e que auxiliam na posterior percepção e compreensão das características apresentadas pela nova abordagem proposta nesse trabalho.

Assim, os métodos selecionados são apresentados conforme a ordem cronológica de sua proposição e publicação, visando facilitar a compreensão da evolução para o estado da arte e o aperfeiçoamento nos cenários de detecção possíveis. Observe que a seção dá ênfase apenas a métodos das abordagens baseada em distância e em densidade, pois são aqueles orientados a mesma categoria de problemas que o método proposto.

Todos os métodos descritos a seguir compartilham da mesma definição base para o problema de detecção de *outlier*, onde é fornecido como entrada um conjunto de dados composto por N elementos com d dimensões (ou atributos) e a saída desejada é que o método determine quais destes elementos pertencem ao conjunto O dos *outliers* (com maior nível de acerto possível). Assim, $O \subset N$ e é esperado que $|O| \ll |N|$, pela própria definição de *outlier*.

2.7.1 DBSCAN

O algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise*; em português, Clusterização Espacial Baseada em Densidade de Aplicações com Ruído) (ESTER et al., 1996) não é propriamente um método de detecção de *outliers*, mas de

agrupamento de dados (ou clusterização) por densidade. Porém, ainda assim merece destaque no contexto de detecção de *outliers*, pois seu desenho conceitual prevê a presença de “ruídos” (que neste contexto são amostras de dados que interfiram na correta definição dos agrupamentos de dados) e por consequência os trata de forma inovadora e elegante (para o período em que foi proposto), acabando, então, por servir de inspiração para a proposta de diversos métodos focados na detecção de *outliers* que surgiram posteriormente.

Sua abordagem conceitual consiste numa noção intuitiva de que agrupamentos de dados tendem a manifestar certa densidade típica, consideravelmente maior que a região externa (não pertencente ao grupo; e que pode conter ruídos prejudiciais à correta separação dos agrupamentos). E estendendo essa abstração, pode-se interpretar que ruídos normalmente situam-se em regiões do espaço com densidade muito menor.

Para efetuar essa abordagem, o algoritmo exige apenas dois parâmetros:

- **ε (*eps*):** raio da esfera d-dimensional que determina a distância máxima entre dois pontos, de modo que possam ser considerados pertencentes à mesma vizinhança. Assim, diz-se que ε forma a ε -vizinhança de um dado ponto q , denotada por $N_\varepsilon(q)$ e expressa pela Equação (1), que inclui todos aqueles vizinhos que são alcançados por este raio a partir do ponto q ; onde $N_\varepsilon(q) \subset N$ qualquer que seja q . Estabelecendo o conceito de distância de alcançabilidade (*reachability distance*);

$$N_\varepsilon(q) = \{r \in N \mid \text{dist}(q, r) \leq \varepsilon\} \quad (1)$$

- ***MinPts*:** trata-se da quantidade mínima de pontos exigidos na ε -vizinhança de um dado ponto (incluindo ele mesmo) para que ele seja considerado um ponto central pertencente ao agrupamento. Assim, esse parâmetro controla a densidade mínima da vizinhança na região central do agrupamento e estabelece o conceito de distância central (*core distance*).

Algumas definições sobre os tipos de pontos presentes no espaço de dados também são empregadas da seguinte forma:

- **Ponto Central:** aqueles pontos que possuem ao menos *MinPts* pontos em sua ε -vizinhança (incluindo o próprio ponto); respeitando a Equação (2). Portanto, para que um ponto q seja um ponto central, $N_\varepsilon(q)$ deve conter ao menos *MinPts* elementos. São assim denominados, pois se situam nas regiões centrais dos agrupamentos de dados e constituem o conjunto N_C ;

$$N_C = \{q \in N \mid |N_\varepsilon(q)| \geq \text{MinPts}\} \quad (2)$$

- **Ponto de Borda:** são os pontos que não satisfazem a condição de ponto central, por possuírem menos que *MinPts* pontos em sua ε -vizinhança. Porém, possuem ao menos um ponto central contido nessa. Ou seja, são aqueles que se situam nas regiões de borda (fronteira) dos agrupamentos de dados, mas ainda são considerados pertencentes ao grupo desses dados. Sendo denotados pelo conjunto N_B e descritos pela Equação (3);

$$N_B = \{q \in N \mid q \notin N_C \wedge (\exists r \in N_C \mid r \in N_\varepsilon(q))\} \quad (3)$$

- **Ruído:** são os pontos que não satisfazem as condições prévias para serem denotados por ponto central ou de borda. Assim, são aqueles que não possuem ao menos *MinPts* pontos em sua ε -vizinhança e também não alcançam diretamente nenhum ponto central. A Equação (4) descreve o conjunto N_R que contem todos os pontos de ruído pertencentes ao conjunto de dados N ;

$$N_R = \{q \in N \mid q \notin N_C \wedge q \notin N_B\} \quad (4)$$

Vale observar, ainda, que N_C , N_B e N_R são conjuntos disjuntos entre si e sua união é equivalente ao conjunto N .

Dados os conceitos e definições acima, o DBSCAN delimita um agrupamento de dados a partir da conexão por densidade existente entre os diversos pontos centrais de forma transitiva; interconexão entre pontos centrais. Já as fronteiras do agrupamento serão constituídas apenas pelos pontos de borda que forem alcançados diretamente por pelo menos um destes pontos centrais. Um exemplo desse mecanismo está ilustrado na Figura 6(b), onde estão presentes dois agrupamentos identificados pelas máscaras azul e laranja.

Figura 6. Exemplo de funcionamento do algoritmo DBSCAN. (a) Ilustração dos três tipos de pontos possíveis: central, borda e ruído; definidos com base na distância de acessibilidade e *MinPts*. (b) Um espaço de dados com dois agrupamentos e um *outlier*.

O preceito central para que o DBSCAN possa funcionar corretamente consiste na determinação dos parâmetros ϵ e *MinPts*. Da maneira que foi inicialmente proposto (ESTER et al., 1996), este algoritmo prevê valores globais para esses parâmetros e seus autores propuseram uma heurística eficaz para estimá-los com base num ponto central do agrupamento menos denso presente no espaço de dados. Assim, com grande probabilidade os outros grupos seriam corretamente identificados, pois seus pontos iriam satisfazer o *MinPts* por situarem-se numa região de maior densidade. Porém, essa abordagem prevê que *outliers* situem-se apenas em regiões esparsas num contexto global.

Portanto, um problema específico para detecção de *outliers* surge naquele cenário relatado na seção 2.6.4 e que foi ilustrado pela Figura 5; ou seja, quando da presença de um *outlier* próximo a um agrupamento de maior densidade e tendo sido definida uma densidade global com base no agrupamento de menor densidade. Assim, no contexto de detecção de *outliers*, esse método se enquadra na abordagem baseada em distância (há também autores que citam um subgrupo dessa para comportar abordagens com base em agrupamento), por tratar da densidade como aspecto global.

Enfim, além de servir como inspiração para diversos dos algoritmos de detecção de *outlier* que foram propostos na sequência (alguns dos quais serão abordados no prosseguir da seção), o DBSCAN ainda é considerado um ótimo algoritmo de agrupamento de dados em conjuntos com grupos de formatos arbitrários e diferentes tamanhos (SCHUBERT et al., 2017).

2.7.2 DB-Outlier

O método DB-Outlier (*Distance-Based Outlier*; em português, *Outlier Baseado em Distância*) (KNORR; NG, 1997, 1998) foi proposto com base nas noções de acessibilidade e conectividade propostas pelo DBSCAN, porém com foco na detecção de *outliers*; sem relação com a tarefa de agrupamento de dados.

O algoritmo requisita apenas dois parâmetros: um raio ε e um percentual de *outliers* esperados π . Os autores usam outros símbolos para esta notação, porém mantemos o ε , pois se trata do mesmo conceito utilizado pelo DBSCAN; como informado pelos próprios autores (KNORR; NG, 1997, 1998). Então, a definição de *outlier* utilizada é a seguinte: Um ponto q é considerado um *outlier* se ao menos π por cento de todos os outros pontos tem uma distância até q menor que ε . Essa definição está formalizada em notação matemática na Equação (5), onde $O(\varepsilon, \pi)$ corresponde ao conjunto de *outliers* detectados em N para os valores de parâmetros informados. Ou seja, os parâmetros ε e π são definidos pelo usuário; restringindo a quantidade de *outliers* encontrados com base num determinado percentual esperado ou desejado.

$$O(\varepsilon, \pi) = \left\{ q \in N \mid \frac{|\{r \in N \mid \text{dist}(q, r) < \varepsilon\}|}{|N|} \leq \pi \right\} \quad (5)$$

Quanto à semelhança com o DBSCAN, vale mencionar inclusive, que no caso da escolha de um mesmo valor de ε para ambos os algoritmos, o conjunto de *outliers* obtidos pelo DB-Outlier será equivalente ao conjunto dos pontos de borda e *outliers* definidos pelo DBSCAN. Porém, na prática o valor de ε utilizado para detecção de *outlier* deverá ser maior do que aquele antes usado com a finalidade de agrupamento.

Assim, apesar de não haver uma mudança conceitual considerável com relação ao DBSCAN, o DB-Outlier já se apresenta como um método, de fato, orientado à detecção de *outliers* e possibilita uma significativa redução na complexidade computacional envolvida; pelo simples fato de não ter de considerar os aspectos relacionados a agrupamento de dados. O DB-Outlier também se enquadra na abordagem de detecção baseada em distância; não sendo, também, capaz de resolver os problemas que envolvem regiões com densidades distintas em proximidade com os *outliers*.

2.7.3 kNN

Outro método de detecção de *outliers* bastante simples, porém eficiente, é o kNN (*k-Nearest Neighbors*; em português, *k*-ésimo Vizinho Mais Próximo) (RAMASWAMY; RASTOGI; SHIM, 2000). Nessa técnica a definição de *outlier* consiste em determinar a pontuação de *outlier* de certo ponto, a partir do valor de distância do *k*-ésimo vizinho mais próximo desse ponto; essa distância é denotada por $D^k(q)$, onde q é o ponto cuja pontuação pretendemos determinar e k é a quantidade de vizinhos mais próximos que deverão ser considerados.

Assim, aqueles pontos que possuem maior distância até o seu *k*-ésimo vizinho, terão também maior pontuação de *outlier*. Podemos verificar que a definição não se mostra conceitualmente divergente daquela proposta para o método DB-Outlier, já que ambas utilizam-se de alguma estimativa de densidade para encontrar pontos de baixa densidade (interpretados como *outliers*).

Dentre algumas vantagens do kNN com relação ao DB-Outlier, podemos citar que a definição do parâmetro k (com a quantidade de vizinhos mais próximos) é mais intuitiva e menos propensa a erros do que determinar o valor de um raio ε e, também, que o kNN fornece uma pontuação de *outlier*, o que permite diferenciar, por exemplo, os *outliers* mais característicos daqueles relativamente mais próximos aos agrupamentos de dados; que podem possuir significado distinto nos cenário de aplicação.

Figura 7. Dois cenários bastante distintos, mas que determinam uma mesma pontuação de *outlier* para q com uso do método de detecção kNN.

Fonte: Adaptado de (ANGIULLI; PIZZUTI, 2002)

Dado que este método ignora as informações das observações mais próximas (vizinhos mais próximos que antecedem o k -ésimo), há certamente uma perda de informação que pode ser relevante em determinados casos; a Figura 7 ilustra dois casos cenários bastante distintos que acabariam determinando a mesma pontuação de *outlier* para o ponto q com o uso do kNN; já que em ambos os casos o valor de $D^k(q)$, com $k = 8$, seria o mesmo.

Para possibilitar a diferenciação dos *outliers* nesses casos, foi proposto o kNNW (*kNN-weight*; em português, kNN ponderado) (ANGIULLI; PIZZUTI, 2002). Esse método utiliza a mesma abordagem conceitual que o kNN, porém para realizar o cálculo da pontuação de *outlier* ao invés de utilizar apenas a distância até o k -ésimo vizinho mais próximo, ele agrega todas as distâncias do ponto q até todos os seus vizinhos mais próximos.

Ambos kNN e kNNW, assim como todos os métodos até então mencionados, se enquadram na abordagem de detecção baseada em distância; apresentando os mesmos problemas na detecção de *outliers* em regiões com razoável variação de densidade.

2.7.4 LOF

Finalmente iremos descrever o método de detecção de *outliers* que resolveu o problema de detecção de em cenários com densidade variada, dando origem à abordagem baseada em densidade. Esse algoritmo é denominado LOF (*Local Outlier Factor*; em português, Fator de Anomalia Local) (BREUNIG et al., 2000).

Assim como ocorre com o DB-Outlier, o LOF também compartilha algumas das noções conceituais introduzidas pelo DBSCAN, como distância central (*core distance*) e distância de alcançabilidade (*reachability distance*).

A ideia geral deste algoritmo é comparar a densidade ao redor de certo ponto, com a densidade ao redor de seus vizinhos; com isso a densidade relativa do ponto é calculada, resultando na pontuação de *outlier*. Essa abordagem pressupõe que a densidade ao redor de um dado normal seja similar à densidade apresentada por seus vizinhos e que para o caso dos *outliers* seja consideravelmente diferente dos seus vizinhos.

O algoritmo LOF tem como base apenas um parâmetro denominado por *MinPts* (quando introduzido pelos autores) ou por k em grande parte da literatura. Esse parâmetro representa o número de vizinhos mais próximos, assim como nos métodos kNN e kNNW, porém aqui é usado para definir a vizinhança local de um objeto e não diretamente associado à pontuação de *outlier*. É intuitivo verificar que o uso dos k -vizinhos mais próximos permite uma boa estimativa da densidade local, dado que tende a gerar medições distintas em

agrupamentos de diferentes densidades; observe que a densidade se manifesta inversamente proporcional à média das distâncias dos k -vizinhos mais próximos. Ressaltamos que identificar essas densidades distintas não seria possível, por exemplo, com a utilização de uma hipersfera com raio fixo.

Esse algoritmo estabelece um conjunto de definições que dão suporte ao seu funcionamento e fundamentam sua abordagem diferenciada, sendo as seguintes:

- **k -distance(q):** é definida como a distância até o k -ésimo vizinho mais próximo do ponto q . Mesmo conceito empregado pelo kNN e análogo à ϵ -vizinhança do DBSCAN; porém aqui trata-se de uma medida local e não mais global.
- **Vizinhança k -distance de q :** é formada pelo conjunto dos objetos cuja distância até q não supera k -distance(q). É importante observar que esta vizinhança pode conter mais que k objetos, devido à possibilidade de empates. A formalização dessa definição é dada pela Equação (6).

$$N_{k\text{-distance}(q)}(q) = \{r \in N - \{q\} \mid \text{dist}(q, r) \leq k\text{-distance}(q)\} \quad (6)$$

- **Distância de alcançabilidade:** a distância de alcançabilidade entre os objetos r e q é, de fato, a distância entre eles, mas que seja ao menos a k -distance(q); a Equação (7), mostra a definição formal dessa distância. Intuitivamente, se um objeto q está longe de r , então a distância de alcançabilidade entre estes é simplesmente a distância que os separa, porém se estiverem suficientemente perto a distância de alcançabilidade será representada pela k -distance de r . Esta medida atua como um fator de alisamento (*smoothing factor*). Algumas implementações utilizam-se apenas da k -distance, descartando a distância de alcançabilidade; o que deu origem ao método Simplified-LOF (SCHUBERT; ZIMEK; KRIEGEL, 2014b), que se comporta de maneira ligeiramente diferente. Esta definição está ilustrada na Figura 8.

$$\text{reachability-distance}_k(r, q) = \max\{k\text{-distance}(q), \text{dist}(r, q)\} \quad (7)$$

Figura 8. Ilustração do conceito de distância de alcançabilidade.
 Fonte: Adaptado de (BREUNIG et al., 2000)

- **Densidade de alcançabilidade local:** é comumente denotada por lrd (*local reachability density*), sendo responsável por estimar a densidade (local) da vizinhança de um dado objeto. É definida como o inverso da média das distâncias de alcançabilidade de q a partir de seus vizinhos r ; formalmente definido na Equação (8). Ou seja, quão maior for o valor de $lrd_k(q)$, maior é a densidade da região onde q se encontra.

$$lrd_k(q) = 1 / \left(\frac{\sum_{r \in N_k(q)} reachability\text{-}distance_k(q, r)}{|N_k(q)|} \right) \quad (8)$$

- **Fator de outlier local:** ou *local outlier fator* (LOF) define a pontuação de *outlier* para determinado ponto q ; ou seja, é a definição final desejada que irá diferenciar os dados normais dos *outliers*. Esse fator é dado pela média da razão entre a densidade dos vizinhos de q e a densidade de q . A definição formal é dada na Equação (9).

$$LOF_k(q) = \frac{\sum_{r \in N_k(q)} \frac{lrd(r)}{lrd(q)}}{|N_k(q)|} \quad (9)$$

Assim, estando o fator de *outlier* local definido é possível determinar as pontuações (*scores*) de *outlier*. Para isso, este fator, basicamente, irá verificar se as regiões da vizinhança próxima de q são relativamente (proporcionalmente) mais densas que o próprio q . São esperados valores próximos de 1 quando essas densidades são, aproximadamente, a mesma (região com densidade homogênea); caracterizando os dados normais (*inliers*). Já para os *outliers* os valores serão tão maiores quanto mais extremos forem esses.

Com isso, o problema identificado nos conjuntos de dados com presença de *outliers* e agrupamentos com diferentes densidades é resolvido. Pois diferente dos outros métodos o tratamento da densidade aqui não é global, mas local; sendo, de fato, determinadas as regiões de densidade local, possibilitando a comparação das alterações na vizinhança. A Figura 9, ilustra simplificada como o algoritmo LOF se comporta num espaço de dados com agrupamentos de diferentes densidades com *outliers* próximos às regiões de maior densidade; problema similar àquele descrito na Figura 5 (vide página 46).

Figura 9. Determinação de *outliers* em espaços de dados com agrupamentos de densidade distinta.
Fonte: Adaptado de (BREUNIG et al., 2000)

O método LOF teve tamanha importância para o problema de detecção de *outliers* que acabou por gerar uma grande quantidade de variantes, sendo utilizada, inclusive, a expressão *LOF family methods* (métodos da família LOF). Dentre as principais variantes, podemos mencionar as seguintes: KDEOS (SCHUBERT; ZIMEK; KRIEGEL, 2014a); Simplified LOF (SCHUBERT; ZIMEK; KRIEGEL, 2014b); LoOP (KRIEGEL et al., 2009); LDOF (ZHANG; HUTTER; JIN, 2009); LDF (LATECKI; LAZAREVIC; POKRAJAC, 2007); INFLO (JIN et al., 2006); COF (TANG et al., 2002).

Cada variante, basicamente, assume formulações distintas para o conjunto de definições proposto pelo LOF original, muitas são focadas na alteração da função que calcula

a densidade local, com a finalidade de reduzir na complexidade computacional; sempre com alguma relação de troca (*tradeoff*) com a acurácia obtida, mas com alguma vantagem para cenários específicos. Porém, para cenários diversificados ou sem conhecimento prévio sobre a organização dos dados o LOF original ainda tende a obter melhores resultados que sua família de métodos derivados (CAMPOS et al., 2016), principalmente, no quesito de acurácia.

2.7.5 Outros

Os métodos destacados nas subseções prévias foram selecionados com enfoque na importância para o aperfeiçoamento das classes de abordagens mencionadas na seção 2.6 - Principais Abordagens na Detecção de *Outliers*, respeitando uma ordem cronológica característica dessa evolução. Como já mencionado, foi dada ênfase às abordagens baseada em distância e em densidade, pois estas são mais robustas e generalistas permitindo aplicação em cenários de dados variados; que são o foco, também, da técnica proposta.

Para finalizar a subseção dos métodos de detecção de *outliers*, foram selecionados outros métodos de grande importância na literatura. Estes não tiveram subseções específicas por não pertencerem à evolução das abordagens baseadas em distância ou densidade, mas, certamente, merecem ao menos uma descrição breve, pois também apresentam resultados compatíveis com o estado da arte; em cenários mais restritos. Essas técnicas foram comparadas com o mecanismo de detecção proposto pelo trabalho; vide Capítulo 5 - Experimentos Computacionais; com exceção da KDEOS.

COF (*Connectivity-based Outlier Factor*, em português, Fator de *Outlier* Baseado em Conectividade) (TANG et al., 2002) é uma das variantes do LOF, inclusive já mencionado na subseção anterior. Sua modificação com relação ao LOF consiste, resumidamente, em substituir o termo de “densidade de alcançabilidade total” original por uma densidade estimada (mais simples) que considera a conectividade de uma vizinhança via uma árvore geradora mínima (do inglês, *minimum spanning tree*).

FastABOD (*Fast Angle-Based Outlier Detection*, em português, Detecção Rápida de *Outlier* Baseada em Ângulo) (KRIEGEL; S HUBERT; ZIMEK, 2008) consiste em um método de detecção mais especializado e pertencente à abordagem de detecção baseada em ângulo. Foi projetada como uma variante mais rápida (de menor complexidade computacional) que a ABOD (*Angle-Based Outlier Detection*) (KRIEGEL; S HUBERT; ZIMEK, 2008). Enquanto a ABOD calcula, para cada ponto tomado como origem, os ângulos para todos os outros pares de pontos possíveis, utilizando-se de uma variação ponderada

desses ângulos como um fator de *outlier*, a FastABOD considera apenas os pares formados pelos k -vizinhos mais próximos; permitindo uma redução considerável na complexidade computacional.

KDEOS (*Kernel Density Estimation Outlier Score*) (SCHUBERT; ZIMEK; KRIEGEL, 2014a) foi desenvolvido como um método de *ensemble*, incorporando à estrutura do LOF uma estimativa de densidade de *Kernel* (do inglês, *Kernel Density Estimation*; KDE) de largura variável. Para comparação com as densidades dos vizinhos, as densidades KDE são padronizadas por ponto utilizando os z -scores relativos às KDE do conjunto dos k -vizinhos mais próximos. Sua média, então, é calculada sobre diferentes tamanhos de vizinhança, para um intervalo $[k_{min}, k_{max}]$ e a pontuação de *outlier* é obtida com a função normal de densidade acumulada (caso seja assumida uma distribuição normal para esta pontuação).

NDoT (*Nearest Neighbor Distance Based outlier Detection Technique*, em português, Técnica de Detecção de *outlier* Baseada em Distância do Vizinho Mais Próximo) (HUBBALLI; PATRA; NANDI, 2011) é uma técnica que classifica os dados como *outlier* (*hard-label*) com base num mecanismo de votação. Um elemento será classificado como *outlier* caso não esteja em conformidade com sua vizinhança (definida pelos k -vizinhos mais próximos, sendo k definido pelo usuário). Simplificadamente é calculado um fator de vizinho mais próximo (do inglês, *Nearest Neighbor Factor*; NNF) para um ponto p , calculando a distância deste (p) para cada um dos seus k -vizinhos mais próximos e comparando com a média dessas distâncias (apenas para a k -vizinhança). A votação ocorre da seguinte maneira: quando o fator NNF de p com relação a um de seus k -vizinhos é superior ao limiar δ (*threshold*), é contabilizado um voto para p . Ao final da análise de p com relação a sua k -vizinhança, p será um *outlier* se possuir quantidade de votos superior a $2/3$ da quantidade de elementos em sua vizinhança.

SSOD (*Semi-Supervised Outlier Detection*, em português, Detecção de Outlier Semissupervisionada) (GAO; CHENG; TAN, 2006) é uma técnica de detecção direcionada à aplicação em cenários semissupervisionados. Esta se utiliza do algoritmo de agrupamento K-Médias (*K-Mean*) para estabelecer centroides no espaço de dados, podendo guiar-se tanto pelos dados rotulados como normais quanto como *outliers*.

É interessante notar que até o momento presente, nenhum dos métodos mais recentes de detecção de *outliers* supera, consistentemente, os métodos seminais; como kNN, kNNW e LOF que compõem o estado da arte na detecção de *outliers*; com ênfase para a abrangência e robustez do LOF. Porém, os métodos FastABOD e KDEOS também se distinguem, mas em cenários mais especializados que os primeiros (CAMPOS et al., 2016).

2.8 Considerações Finais

O capítulo apresentou diversos aspectos que envolvem desde a definição conceitual de dados anômalos até os princípios e mecanismos para sua detecção. Ficando evidente que todos esses conceitos se entrelaçam fortemente e são bastante variados conforme as diferentes abstrações e aplicações necessárias.

Foi possível observar que o princípio central da detecção de *outliers* se sustenta nas suposições acerca dos padrões apresentados pelo conjunto de dados (AGGARWAL, 2013), os quais normalmente são, a priori, desconhecidos e complexos de serem modelados matematicamente. Por esta razão, os métodos de detecção de *outliers* que tomam como base princípios de aprendizado de máquina estão em forte evidência e sendo continuamente desenvolvidos e aprimorados nos últimos anos; proporcionando a possibilidade de aplicação em cenários antes impraticáveis, seja pela ineficácia ou pelo custo computacional proibitivo dos métodos predecessores.

Ressaltamos, então, que o método de detecção que propomos neste trabalho, tem como intuito acrescentar uma nova e distinta abordagem ao estado da arte em detecção de *outliers*. E vários dos conceitos apresentados nesse capítulo são de grande importância para avaliar a proposta diferenciada que estudamos. Dentre as propriedades de grande relevância que caracterizam nosso método, podemos mencionar: o reconhecimento de estruturas de alto nível para determinar o conjunto de referência (ao invés de considerar apenas características físicas como distância, similaridade ou densidade), uma dinâmica de busca local para global no espaço de dados (sem necessidade de um parâmetro controlado pelo usuário, que decorre da dinâmica natural do método) e uma nova definição conceitual para as amostras anômalas.

Com isso, prosseguiremos no próximo capítulo para a descrição do algoritmo de caminhada do turista, que também faz parte do referencial teórico central da técnica de detecção desenvolvida e, então, posteriormente passaremos para a descrição do novo método.

Capítulo 3

CAMINHADA DETERMINÍSTICA DO TURISTA

Neste capítulo o algoritmo de caminhada do turista é revisado e algumas de suas propriedades são destacadas, as quais serão de essencial importância para a posterior compreensão do desenvolvimento conceitual do método de detecção de outliers proposto. É ressaltada também a dinâmica de atratores com cruzamento, que terá grande importância na determinação daquelas amostras de dados caracterizadas como outliers internos.

3.1 Contextualização

A caminhada do turista pode ser intuitivamente descrita como um “turista” (caminhante) que segue uma regra determinística elementar para traçar sua rota: ele caminha em direção à cidade mais próxima, que ainda não tenha visitado nos últimos μ passos. No contexto da análise de dados, cada cidade corresponde a um exemplo em um determinado conjunto de dados. Por esta característica, a caminhada do turista é classificada como uma caminhada determinística (BUNIMOVICH, 2004; FREUND; GRASSBERGER, 1992), em contraste às caminhadas aleatórias (outro tópico de pesquisa muito estudado e produtivo) (CAMPITELI et al., 2006; LIMA; MARTINEZ; KINOCHI, 2001). De maneira mais específica, este modelo pode ser caracterizado como uma caminhada determinística parcialmente autorrepulsiva, devido ao envolvimento do parâmetro de restrição μ no processo de caminhada. Este parâmetro age como uma janela de memória (ou janela de restrição), evitando que o caminhante visite os últimos μ elementos contidos nesta memória num dado momento. Assim que o caminhante executa um novo passo na caminhada, o elemento mais

antigo (ainda presente) nesta janela é liberado da restrição e o atual elemento (recém-visitado) é adicionado.

Embora a caminhada do turista siga uma regra bem simples, a dinâmica gerada pelo caminhante é complexa e exibe diversas características interessantes ao variarmos o parâmetro de memória μ . Para $\mu \geq 1$, o caminhante já pode apresentar um comportamento não-trivial (LIMA; MARTINEZ; KINOCHI, 2001) durante seu deslocamento. Uma característica muito interessante a ser observada é que após certo número de passos, o caminhante ficará obrigatoriamente preso num ciclo (também denominado por atrator) em sua trajetória. Portanto, a trajetória do turista se distingue, tipicamente, em duas partes: transiente (com comprimento t) e ciclo final (com período p); onde $\mu + 1 \leq p \leq N$, sendo N o número de elementos no espaço de dados (CAMPITELI et al., 2006). O transiente é composto por todos os pontos do espaço que foram visitados pelo caminhante antes deste adentrar o ciclo final, portanto o ciclo final será determinado pelo conjunto de pontos visitados ciclicamente após sair do transiente (determinando um atrator).

Figura 10. Ilustração de uma caminhada do turista destacando a parte transiente (em amarelo) e o ciclo final (em verde).

O termo ciclo final é utilizado, pois certas trajetórias podem formar ciclos intermediários (temporários) dos quais o caminhante é capaz de escapar. Portanto, estes ciclos não configuram atratores reais (BACKES et al., 2010). Ao seguir com a caminhada, em algum momento o caminhante ficará preso num ciclo final, do qual ele não poderá escapar mesmo num tempo de simulação infinito.

A Figura 10 ilustra um processo de caminhada do turista que se inicia num exemplar rotulado por 1, onde após a fase de transiente (elementos de cor amarela) composta por três elementos rotulados por 1, 2 e 3 ($t = 3$), o processo de caminhada converge para um atrator (elementos de cor verde) constituído pelos elementos rotulados por 4, 5, 6, 7, 8, 9 e 10 que identificam um ciclo de período $p = 7$, indicando que o valor do parâmetro de memória deve ser $\mu \leq 6$ na ilustração apresentada.

Podemos observar que tendo múltiplos caminhantes, cada um iniciando a partir de um dos N elementos, geralmente, para simulações com pequenos valores de memória μ , é obtido um grande número de pequenos atratores distintos (ciclos com poucos elementos), mas ao passo que aumentamos o tamanho da memória, o número de atratores distintos (únicos) tende a decrescer e o período p destes tende a aumentar (BACKES et al., 2010); essa característica pode ser observada na Figura 11 e na Figura 12. Os extremos podem exemplificar esse comportamento de forma mais clara e intuitiva: quando $\mu = 0$, cada elemento inicial da caminhada se constitui num atrator único (*singleton*), onde iremos obter N atratores de período $p = 1$, após uma simulação completa pelo espaço de dados. Entretanto, para $\mu = N - 1$ (valor máximo possível para o parâmetro de restrição), independentemente do ponto de início (saída), todos os atratores obtidos terão período $p = N$. Portanto, com a variação do tamanho da memória μ os diferentes atratores identificados acabam por serem capazes de capturar características locais-para-globais presentes no espaço de dados (CAMPITELI et al., 2006; SILVA; ZHAO, 2015). Em geral, o intervalo de interesse para os valores de memória das caminhadas reside no intervalo $1 \leq \mu \leq N - 1$.

Outro fenômeno interessante, que pode reduzir efetivamente o valor máximo do intervalo de interesse dos valores de memória das caminhadas do turista, é a existência de um valor de memória crítico (tipicamente denotado por μ_c). Este valor crítico é caracterizado pelo valor de memória com o qual os caminhantes são capazes de atingir o estado de “saturação de complexidade” no espaço de dados (SILVA; ZHAO, 2015), situação na qual um atrator global (que contém em seu ciclo final todos os N elementos do conjunto de dados) é identificado. Geralmente, quando a caminhada ocorre num conjunto de dados bem arranjado, o valor crítico apresenta-se muito menor que $N - 1$ (CAMPITELI et al., 2006). As caminhadas que ocorrem além deste valor crítico acabam presas no atrator global; por esta razão o termo saturação de complexidade é introduzido na literatura. A Figura 11 ilustra esse fenômeno, que ocorre quando $\mu=35$, mesmo sendo o valor máximo de memória $\mu_{\text{máx}} = 45$.

Figura 11. Evolução dos atratores identificados para crescentes valores de μ num mesmo conjunto de dados de entrada. Cada imagem mostra todos os atratores identificados para o valor de memória fixado, diferentes cores nas linhas representam diferentes atratores.

A Figura 11 ilustra as características da dinâmica de identificação dos atratores na caminhada do turista, evidenciando a evolução desses atratores para incrementos no valor de memória μ num mesmo conjunto de dados de entrada (composto por 46 elementos); para cada valor de memória é iniciado um caminhante em cada ponto (de saída) do conjunto de dados. É possível observar que para valores pequenos de μ são identificados vários atratores, que seguem crescendo em tamanho (incorporando mais elementos) até atingir $\mu=5$, quando apenas um grande atrator (com 35 elementos) se torna dominante pelo intervalo de $\mu \in [5, 34]$; ou seja, em todo esse intervalo, nenhum atrator distinto é identificado e todos os pontos de saída são atraídos até ele. Isso ocorre por tratar-se de uma forma muito regular (que foi identificada logo para $\mu=2$), da qual o caminhante só escapará quando o valor de $\mu=35$ for atingido, o obrigando a deixar o atrator; o qual possui exatamente 35 elementos que preencherão toda a memória, impondo a busca por um atrator maior.

Para esta configuração de dados, o valor de memória $\mu=35$ já possibilita a identificação de um atrator global (aquele que engloba todos os pontos), sendo, portanto o valor crítico de memória (μ_c) n este cenário de dados. Deste valor até o valor máximo de μ que será 45 (ou seja, $N-1$) alguns outros atratores distintos se manifestam, mas há sempre a presença de um atrator global. A Figura 12 ilustra como ocorre essa evolução na quantidade e no tamanho dos atratores para o mesmo conjunto de entrada utilizado na Figura 11; sendo possível identificar facilmente os comportamentos que foram descritos logo acima.

Figura 12. Comportamento dos atratores conforme incrementos no tamanho da memória. A linha azul mostra a evolução da quantidade de atratores, enquanto que as linhas laranja mostram as mudanças nos tamanhos médio, do maior e do menor atrator encontrado para cada valor de memória.

3.2 Atratores Com Cruzamento

Embora já tenhamos descrito a essência do processo de caminhada do turista, devemos enfatizar que a dinâmica cíclica exibida pela caminhada é muito rica e se mostra como o mecanismo principal do processo. O atrator periódico ilustrado pela Figura 10 é um atrator típico (normal) e pode levar a interpretação incorreta de que o processo de caminhada pode gerar apenas atratores geometricamente simples. Aqui, mostramos que o caminhante pode gerar atratores complexos com várias formas com cruzamentos. Ao aumentar o valor de memória é possível verificar o emergir de uma dinâmica complexa e elegante, que dá origem a vários atratores não triviais. Alguns ciclos são bastante complexos, exibindo um ou mais pontos de interseção em sua trajetória (BACKES et al., 2010) e , então, constituindo o que denominamos por atratores complexos (ou *crossing-attractors*). Estes atratores são compostos por duas ou mais seções (semicíclós).

Figura 13. Três exemplos de atratores com cruzamento com diferentes quantidades de pontos de interseção (cor amarela) e semicíclós.

A Figura 13 ilustra três possíveis exemplos de atratores com cruzamento, onde a dinâmica exibida para conformação do ciclo final é obviamente mais complexa do que aquela ilustrada pela Figura 10. Na Figura 13(a), o atrator com cruzamento tem apenas um ponto de interseção e, por consequência, dois semicíclós. O atrator apresentado na Figura 13(b) é

similar ao anterior, porém conta com a presença de dois pontos de interseção consecutivos; separando os dois semiciclos identificados. Já na Figura 13(c) é ilustrado um atrator que contém quatro semiciclos com apenas um ponto de interseção na trajetória do ciclo final.

Figura 14. Dois atratores distintos detectados num conjunto de dados com dois agrupamentos em forma de banana. (a) Um atrator global é ilustrado, sem a presença de cruzamentos. (b) Um fragmento do conjunto de dados é destacado, onde ocorre um atrator com diversos cruzamentos.

Embora os atratores ilustrados na Figura 13 apresentem maior complexidade quando comparados ao atrator da Figura 10, eles ainda são simples, pois ao explorarmos grandes conjuntos de dados visualizamos o emergir de uma grande quantidade de atratores geometricamente muito mais complexos. Um exemplo é mostrado na Figura 14(b), onde um atrator com cruzamento foi detectado num conjunto de dados com 200 pontos divididos em dois agrupamentos com formato de banana. Em contraste, a Figura 14(a) ilustra todo o conjunto de dados com o atrator global (aquele que determina um ciclo final contendo todos os pontos do conjunto); observe que a área destacada corresponde ao local onde o atrator com cruzamento foi identificado. Este exemplo é bastante interessante, pois temos um atrator gigante obtido com um alto valor de memória ($\mu = 178 = \mu_c$) que não apresenta a dinâmica de cruzamento que estamos ressaltando (não há sequer um ponto de interseção), enquanto que o

atrator com cruzamento foi obtido num processo de caminhada com pequeno valor de memória ($\mu = 8$) e contém 19 pontos de interseção com apenas 34 elementos de dados distintos. Esse exemplo é suficiente para esclarecer que a relação de complexidade dos atratores não se dá exclusivamente pelo aumento nos valores de memória e tamanho dos ciclos finais identificados.

Figura 15. Comportamento dos atratores com os incrementos no tamanho da memória. (a) Conjunto de dados de entrada. (b) Comportamento dos atratores conforme incrementos no tamanho da memória. A linha azul mostra a evolução da quantidade de atratores, enquanto que as linhas laranja mostram as mudanças nos tamanhos médio, do maior e do menor atrator encontrado para cada valor de memória.

Na Figura 15(b) analisamos o comportamento dos atratores conforme incrementos sucessivos no tamanho de memória para o conjunto de entrada ilustrado na Figura 15(a); que possui 500 elementos de dados. A simulação determinou o valor crítico de memória como sendo $\mu_c=309$, porém é interessante notar que a Figura 15(b) mostra diversos atratores com tamanhos superiores à quantidade total de elementos antes do valor crítico ser atingido; observe as linhas laranja que superam a linha horizontal pontilhada (preta) traçada na altura do valor de N . Todos esses casos evidenciados pela Figura 15(b) são, certamente, manifestações de atratores com cruzamento; porém observe que atratores com cruzamento também podem se manifestar com cardinalidade inferior a N , como já mostramos no exemplo da Figura 14, porém estes casos não são visíveis nesse tipo de gráfico.

Outros aspectos da dinâmica exibida pelos atratores podem ser analisados, como a degeneração e superposição de atratores (LIMA; MARTINEZ; KINOCHI, 2001) que ocorrem com incrementos consecutivos nos valores de memória; reforçando, também, a complexidade na dinâmica de identificação destes atratores. Ao prosseguirmos para grandes

valores de memória, grandes atratores podem surgir, enquanto outros atratores presentes nos valores de memória das etapas anteriores podem se degenerar. Assim sendo, a evolução dos atratores de uma janela de memória para outra não é trivial; eles podem crescer, degenerar, unir, interseccionar e até mesmo desaparecer.

3.3 Detecção de Atratores

Dado que a caminhada do turista pode gerar atratores complexos, a detecção destes atratores também se torna uma tarefa desafiadora. Neste trabalho usamos a seguinte estratégia para encontrar os atratores: inicialmente, implementamos um contador de passos para registrar os instantes em que um dado caminhante visita os mesmos elementos de dados repetidas vezes. Quanto mais vezes este visitar um mesmo elemento, maior será a chance de que esteja preso num ciclo composto por tal elemento. É importante observar que apesar de intuitivamente assumirmos que bastaria visitar um mesmo elemento por duas vezes para considerarmos que o ciclo já estaria determinado, esta intuição não é sempre uma verdade (BACKES et al., 2010), pois o caminhante poderia ainda não estar sobre efeito da janela de restrição de memória durante a primeira passagem (poderia não haver tempo desta ter sido preenchida no momento de tal passagem). Com base nesse comportamento, é necessário checar se o elemento atual da trajetória em andamento já foi visitado ao menos três vezes pelo caminhante antes de analisarmos a trajetória, já realizada, em busca da presença de um ciclo final; evitando assim desperdício de recursos computacionais.

Na ocorrência de três ou mais visitas (caso em que numa etapa anterior do algoritmo o ciclo não foi identificado) ao elemento atual da caminhada, será necessário considerar a lista de pontos formada por aqueles visitados entre a última visita ao elemento atual (momento atual da trajetória) e a visita imediatamente antecessora no mesmo elemento (podemos chamar essa lista de última seção) e, após isso, compará-la com a seção de pontos situados entre esta visita imediatamente antecessora com a sua precedente (duas atrás da contagem atual). Caso as duas seções sejam compostas pelos mesmos elementos, consideramos ter encontrado o ciclo final.

Porém, na ocorrência de um atrator com cruzamentos se faz necessário estender esse procedimento para comparar a última seção determinada pelas interseções não só com sua seção imediatamente anterior, mas também com as seções prévias recorrentemente até que

seja encontrada uma seção composta pelos mesmos elementos ou que não haja mais seções para comparação (caso em que o ciclo final ainda não pôde ser determinado).

Projetamos esta tarefa de comparação com a utilização de uma estrutura de dados que simplesmente mantém duas medidas para cada seção: a quantidade de elementos e a soma dos identificadores de seus elementos. Esta suposição funciona corretamente, pois os identificadores dos elementos dos conjuntos de dados são números inteiros únicos, o que também torna esse conjunto de medidas único para cada um dos semiciclos possíveis numa trajetória.

Ainda no nível de implementação, uma característica importante que deve ser mencionada é que o algoritmo de caminhada do turista pode ser completamente executado com base, exclusivamente, na utilização de uma tabela de classificação da vizinhança (do inglês *neighborhood rank table*), não sendo mais necessário manter, calcular ou utilizar as medidas de distâncias entre os pontos do espaço de dados assim que esta tabela tiver sido construída. Inclusive, no caso de sabermos qual o valor máximo de memória μ desejado para a caminhada, bastará que a tabela de classificação da vizinhança seja calculada para $\mu+1$ vizinhos de cada elemento.

3.4 Considerações Finais

O capítulo apresentou diversas das características dinâmicas da caminhada do turista que, por sua vez, atraíram nossa atenção para proposição do novo método de detecção, que se utiliza de uma abordagem conceitual diferenciada.

Assim, com os conceitos peculiares acima descritos, finalizamos o presente capítulo e podemos prosseguir para o próximo onde serão descritas as características específicas do método de detecção de *outlier* proposto.

Capítulo 4

MÉTODO DE DETECÇÃO PROPOSTO

Aqui é descrito o método de detecção de outliers proposto. São ressaltadas as suas peculiaridades, quanto à abordagem conceitual diferenciada na identificação dos outliers e os detalhes de controle da dinâmica dos caminhantes de modo a extrair dados relevantes para a tarefa de identificação dos outliers.

4.1 Nova Abordagem de Detecção

Muitas das características da caminhada do turista mencionadas no capítulo anterior atraíram nossa atenção para aplicá-la na tarefa de detecção de *outliers*. Mais especificamente, vale reforçar que os elementos de dados que formam os atratores identificados nas diversas caminhadas parecem relacionar-se a agrupamentos de dados, padrões e características organizacionais do espaço de dados (CAMPITELI et al., 2006; SILVA; ZHAO, 2015). E posto que *outliers* são eventos raros provavelmente gerados por mecanismos que diferem daquele responsável pela maioria dos dados, hipotetizamos que os *outliers* não conformariam com esses padrões identificados pelo caminhante turista e, portanto, deveriam ter uma participação mínima nos atratores identificados.

O mesmo raciocínio foi estendido para *outliers internos* que também deveriam apresentar uma participação mínima nos atratores (por não conformar com os padrões exibidos), porém nestes casos mais específicos o comportamento deveria diferir ligeiramente, havendo apenas pouco mais de participação quando comparado aos *outliers externos*. Dadas tais considerações, prosseguimos para a investigação científica que resultou no método de detecção de *outliers* aqui proposto.

Figura 16. Superposição de todos atratores identificados para todos os valores de memória até o antecessor do valor crítico, ilustrando a base para o mecanismo de detecção proposto.

A Figura 16 possui 100 pontos divididos em dois agrupamentos (pontos de coloração preta) e dois *outliers* (cor vermelha). As diferentes linhas coloridas indicam a superposição de todos os 612 atratores distintos identificados para todas as caminhadas possíveis com o parâmetro de memória variando no intervalo $[1, (\mu_c - 2)]$; o valor crítico foi removido pois define-se como aquele em que surge o atrator global (que contém todos os pontos do espaço de dados) e seu antecessor por ser um atrator com características de união dos agrupamentos. Podemos, então, ver que nenhum dos atratores identificados neste intervalo englobou os *outliers* destacados, o que aconteceria apenas para o valor crítico de memória e seu antecessor imediato. Assim, simplificada, a Figura 16 ilustra e reforça as bases que idealizamos ao propor o mecanismo de detecção de *outlier* com base no caminhante do turista.

Formalmente, definimos o problema de detecção de *outlier* que será abordado pelo método proposto da seguinte maneira: em um dado conjunto de dados, composto por N elementos com d dimensões (ou atributos), o objetivo é descobrir os O elementos que não conformam com o padrão constituído pela maioria dos exemplos de dados, onde $O \subset N$ e espera-se que $|O| \ll |N|$, satisfazendo a interpretação de *outliers* como eventos de ocorrência rara. Vale observar que algumas definições de outliers não se baseiam nessa restrição, porém reforçamos que o escopo do nosso estudo está fortemente relacionado a esta característica.

Posto isso, o processo de detecção de *outlier* pode ser descrito. Dado um conjunto de dados onde é desejável identificar *outliers*, primeiramente deve ser construída a matriz de distâncias (ou matriz de similaridade). Prosseguimos, então, para a construção da tabela de

classificação da vizinhança (do inglês *neighborhood rank table*) a partir da matriz de distâncias. Após isso, a matriz de distâncias poderá ser descartada, pois o algoritmo da caminhada do turista poderá ser todo executado com base na tabela de classificação. Além do que, não haverá necessidade de atualizarmos novamente esta tabela até o final do algoritmo. Vale observar que utilizamos a distância Euclidiana para construção da matriz de distâncias, porém a dinâmica do método não é fortemente impactada pela escolha desta métrica já que este terá como base a tabela de classificação resultante (CAMPITELI et al., 2006). Após essa etapa de pré-processamento, iniciamos a caminhada determinística do turista com N caminhadas para um dado valor de μ (que posteriormente será incrementado); cada caminhante inicia seu trajeto a partir de um item distinto do conjunto de dados. Para cada caminhada, o caminhante procura sempre visitar o item de dado (vizinho) mais próximo, desde que este não tenha sido visitado nos últimos μ passos. Após algumas iterações, a caminhada acabará presa num atrator; definido pela presença de um ciclo final. Com isso, devemos manter uma estrutura que armazene a participação de cada exemplar do conjunto de dados nos atratores identificados. Em resumo, cada vez que certo item de dado faz parte de um atrator identificado este terá seu contador incrementado em 1. Ao término de uma rodada completa das N caminhadas mencionadas, o tamanho da memória μ é incrementado e o procedimento repetido para este novo valor de memória, o que deve ocorrer sucessivas vezes até que o valor crítico de memória μ_c seja alcançado com a identificação de um atrator global.

4.2 Cálculo da Pontuação de Outlier

A pontuação dos *outliers* para a técnica proposta é formalmente definida com a seguinte equação:

$$S_i = 1 - \frac{\sum_{\mu=1}^{\mu_c} \Phi_i(\mu)}{\sum_{\mu=1}^{\mu_c} \sum_{j=1}^N \Phi_j(\mu)}, \quad (10)$$

onde S_i é a pontuação de *outlier* do item de dado i ; $\Phi_i(\mu)$ é o número de participação do item de dado i nos atratores identificados para as caminhadas com tamanho de memória μ ; μ_c é o valor crítico de memória e N é a quantidade de dados presentes num dado conjunto de dados.

Simplificadamente, esta pontuação compara a participação do elemento i (a ser pontuado) nos atratores em que esteve presente, com o total da participação de todos os outros elementos em todos atratores identificados durante a caminhada. Assim, da equação (10), podemos observar que um item de dado recebe uma alta pontuação de *outlier* se este participa em poucos atratores nas caminhadas do turista, enquanto que deverá receber uma pontuação relativamente baixa ao participar de uma grande quantidade de atratores. Logo, os elementos anômalos (*ouliers*) deverão possuir sempre pontuação próxima de 1, enquanto os dados normais (*inliers*) terão pontuações inferiores; relativas a disparidade observada nessas quantidades de participação.

É importante observar, então, que a saída do método de detecção proposto é uma pontuação (*soft-label*), como descrito na seção 2.5, onde cada item do conjunto de dados terá um valor atribuído com a finalidade de estimar quão *outlier* aquele item é em relação ao seu conjunto. Uma das formas de analisar os dados resultantes é ordená-los de forma decrescente, com relação a esta pontuação, e estabelecer um ponto de corte com relação base no número de *outliers* esperados, tornando os n elementos, do topo da lista, *outliers*. Outra maneira é estabelecer um limiar de pontuação, considerando qualquer valor acima de tal limiar como sendo *outlier*, mas sem fixar a quantidade esperada.

A Figura 17 ilustra as pontuações obtidas para um cenário de dados artificial, onde é possível ver que o elemento de dado mais distante recebeu uma pontuação muito alta de *outlier* (como esperado) e o elemento central que não conforma bem com os padrões da espiral recebeu a segunda pontuação mais alta.

Figura 17. Exemplo de pontuação de *outlier* calculada para um conjunto de dados artificial.

4.3 Critérios de Controle da Caminhada do Turista

Resumidamente, a base para o mecanismo proposto consiste em executar um processo de caminhada do turista sobre um conjunto de dados e extrair, deste, informações úteis para distinguir os *outliers* dos exemplos de dados normais. Para este propósito, precisamos estabelecer alguns critérios específicos para o controle da caminhada do turista, como: condição inicial, intervalo de memória, elementos equidistantes, aceitação dos atratores e critério de parada. Cada um deles está descrito em maiores detalhes abaixo:

- **Condição inicial:** Exerce forte influência sobre o processo de caminhada. Cada caminhada determinará apenas um atrator, mas diferentes pontos de partida podem levar a diferentes atratores. Portanto, para nosso objetivo executamos uma busca completa no espaço de dados, iniciando uma caminhada para cada um dos N elementos para um valor de memória μ .
- **Intervalo de memória:** O intervalo de interesse é limitado para valores de μ entre 1 e μ_c . Para valores superiores ao crítico há uma intensa presença de atratores globais, o que não gera ganho de informação relevante para a tarefa de detecção de *outliers*; condição denominada saturação de complexidade (SILVA; ZHAO, 2015).
- **Elementos equidistantes:** Na ocorrência de empates ao escolher o vizinho mais próximo (presença de elementos equidistantes do ponto atual), esses serão resolvidos aleatoriamente durante o processo de caminhada. Vale observar que esse fenômeno se torna crítico em configurações de dados que apresentam estrutura muito regular (como grades), onde a ocorrência de uma direção preferencial impactará negativamente na dinâmica e precisão da detecção de *outliers*; as pontuações dos *outliers* tendem a se concentrar erroneamente nos pontos situados na direção preferencial estabelecida.
- **Aceitação dos atratores:** Dada a condição inicial acima descrita, cada caminhada irá extrair apenas um atrator, assim que o ciclo final for identificado. Porém, muitos dos atratores extraídos não serão únicos, já que algumas caminhadas (mesmo iniciando de pontos de partida distintos) podem ficar presas num mesmo atrator. Para a tarefa de detecção de *outliers*, estamos mais interessados nos atratores distintos e, portanto, cada atrator é contado apenas uma vez durante um

mesmo valor de memória μ . Ou seja, para cada rodada de certo valor de μ , são contabilizados apenas os atratores distintos.

- **Critério de parada:** No contexto de cada caminhada, o critério de parada é atingido quando um atrator (ciclo final) é determinado pelo caminhante. Mas o critério de parada global é definido pela iteração na qual um caminhante qualquer encontra um atrator global; momento no qual o μ_c é definido como sendo o μ atual.

4.4 Algoritmo de Detecção de *Outlier*

O Algoritmo 1, abaixo, descreve os passos principais da técnica de detecção de *outliers* baseada em caminhada do turista. A entrada será um conjunto de dados de qualquer dimensionalidade e a saída consistirá numa lista com a pontuação de *outlier* para cada elemento deste conjunto.

Essa seção tem como finalidade dar a visão geral do algoritmo, destacando como se dá o controle e parada da caminhada nos momentos de interesse para o método de detecção de *outlier*. As etapas do algoritmo que remetem à caminhada do turista podem ser mais bem compreendidas no Capítulo 3 - Caminhada Determinística do Turista.

Algoritmo 1 Algoritmo de detecção de *outlier*

Entrada: Um dado conjunto de dados composto por N elementos.

Saída: Lista de pontuação de *outlier* para os N elementos.

- 1: Construir a tabela de classificação da vizinhança.
 - 2: Iterar pelos valores de memória (μ) de 1 até μ_c ou $N-1$.
 - 3: Iterar por todos os N elementos de dados como pontos de partida.
 - 4: Procedimento de caminhada do turista:
 - 5: Visitar o elementos mais próximo que não esteja contido na janela de restrição.
 - 6: Até que: Um ciclo final seja identificado.
 - 7: Manter apenas os atratores distintos.
 - 8: Calcular a pontuação de *outlier* usando a equação (10).
-

4.5 Considerações Finais

Neste capítulo destacamos as características do método de detecção de *outliers* baseado na caminhada. Trata-se de um capítulo mais breve, pois os experimentos computacionais que o avaliam foram descritos em capítulo separado, ficando este capítulo focado em mostrar ao que o método se propõe e como se utiliza das características da caminhada para determinar os *outliers*.

Assim, com o encerramento deste capítulo e tendo descrito o funcionamento do método de detecção, passamos para o próximo capítulo, onde serão descritos os experimentos computacionais executados para a avaliação do método proposto.

Capítulo 5

EXPERIMENTOS COMPUTACIONAIS

Aqui são descritos os experimentos executados para avaliar e ressaltar o funcionamento da técnica de detecção de outliers proposta. Estes experimentos estão divididos em duas grandes seções, a primeira com cenários artificiais para avaliação das características fundamentais e a segunda com bases de dados reais para avaliar e ilustrar a robustez e generalidade da técnica.

Para avaliar o método proposto, foram executados experimentos computacionais considerando dois tipos de conjuntos de dados: artificiais e reais. Para os experimentos com dados artificiais, foram projetados cenários experimentais visando enfatizar as dinâmicas e características promissoras do novo mecanismo de identificação de *outliers*; iniciamos por cenários com configurações de *outliers* mais simples e prosseguimos para aqueles mais avançados de maneira incremental.

É importante ressaltar que a avaliação dos mecanismos de detecção de *outliers* é uma tarefa desafiadora (CAMPOS et al., 2016), dado que cada problema tem suas próprias características e cada técnica possui diferentes limitações. Ratificamos, também, que este trabalho consiste na proposta inicial do uso das dinâmicas exibidas pela caminha do turista para a identificação de *outliers*; o que consiste numa abordagem diferenciada daquelas apresentadas pela literatura e num mecanismo de detecção inteiramente novo. Portanto os experimentos realizados estão focados em aspectos de eficácia e efetividade, ao invés de eficiência.

5.1 Experimentos em Cenários de Dados Artificiais

Inicialmente, serão mostrados os resultados experimentais do método proposto em conjuntos de dados artificiais para diferentes cenários, dispostos e organizados de forma evolutiva e incremental quanto à complexidade apresentada. Cada cenário tem por intenção mostrar certos comportamentos e capacidades do método que propusemos.

A Figura 18(a) apresenta um cenário simples, onde está presente um agrupamento regular de dados e um único exemplo de dado isolado. Intuitivamente, consideramos o item de dado isolado como *outlier*, enquanto os outros constituem o padrão esperado dos dados (*inliers*).

Figura 18. Cenário trivial com apenas um *outlier*. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. A linha de asteriscos vermelhos refere-se ao item de dado isolado.

Ainda na Figura 18(a), são mostradas as pontuações de *outlier* obtidas para cada item de dado; calculadas através da equação (10). Podemos ver que o exemplar isolado apresentou uma alta pontuação de *outlier* (maior valor entre todos os elementos do espaço de dados), enquanto os outros apresentaram baixas pontuações com valores similares entre si. Portanto, o ponto isolado é um *outlier*. Dado que este cenário exibe vários elementos equidistantes (observe que o mesmo ocorre no cenário da Figura 19), é válido mencionar que o mecanismo de desempate (na escolha do vizinho mais próximo) pode influenciar fortemente os resultados finais; logo qualquer tipo de preferência direcional deve ser evitada.

Cada linha da Figura 18(b) ilustra a participação de um item de dado nos diferentes atratores (distintos) para cada valor de memória das caminhadas do turista (variando de 2 até quando o valor crítico de memória é alcançado). Nesta mesma figura, vemos que o item isolado (representado pela linha de asteriscos vermelhos) não participa nos atratores com tamanho de memória na faixa intermediária do gráfico, participando apenas do atrator global assim que o tamanho da memória se torna suficientemente grande. Por outro lado, os outros itens de dados do agrupamento principal (todas as linhas coloridas restantes) têm participação nos atratores para valores de memória desde os mais baixos até os mais altos.

Figura 19. Cenário com agrupamentos de diferentes densidades. (a) Conjunto de dados com a pontuação de outlier de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados.

Na Figura 19(a) estão presentes dois pequenos agrupamentos regulares de dados com diferentes densidades. Esta figura também mostra a pontuação obtida com o uso da equação (10) para cada exemplo de dados. Novamente, os pontos isolados apresentaram alto valor na pontuação de outlier, independente de quão perto ou distante estavam dos agrupamentos de dados normais (*inliers*). O que mostra que os outliers foram determinados pela estrutura apresentada pelo conjunto de dados, pois embora a distância entre os exemplares influencie o mecanismo de detecção, ela não é um fator dominante. Onde o fator dominante, de fato, é a estrutura local e global apresentada pelo próprio conjunto de dados.

A Figura 19(b) mostra as diferentes participações nos atratores relativas ao incremento no parâmetro de memória. Novamente é possível identificar comportamento semelhante para os pontos isolados (linhas de asteriscos vermelhos sobrepostas), onde fica claro que estes pontos (*outliers*) não expressam participação na faixa de valores intermediários de memória.

Observamos então que estes apenas participam das extremidades dos valores de memória, ou seja, somente nos atratores extremamente locais (com valores muito pequenos de memória) ou nos atratores globais (quando a memória é muito grande e bem próxima do valor crítico). Por outro lado, os itens de dados pertencentes aos agrupamentos regulares expressam grande participação na zona intermediária dos valores de memória. Este comportamento, evidenciado pelos experimentos já descritos, apresenta-se como uma assinatura característica dos *outliers*.

Figura 20. Cenário diverso com dois agrupamentos irregulares. (a) Conjunto de dados com a pontuação de outlier de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados.

A Figura 20(a) apresenta dois grupos irregulares majoritários de dados (*inliers*) com diferentes formatos (espalhamento) e dois grupos de dados isolados (*outliers*), um situa-se entre os dois grupos majoritários sendo composto por apenas um elemento enquanto o outro se encontra mais afastado e possui três elementos. É possível constatar que os exemplares dos grupos isolados tiveram alta pontuação de *outlier* enquanto que os grupos majoritários obtiveram pontuação baixa; comportamento desejável. Além disso, os itens de dados mais centralizados em seus agrupamentos, ou seja, com menor dispersão tiveram pontuações ainda ligeiramente mais baixas relativas ou outros elementos do mesmo agrupamento (que apresentaram maior dispersão). Como relatado nos experimentos anteriores, a Figura 20(b) nos mostra que os *outliers* deste cenário também não participaram na faixa intermediária de valores de memória, diferenciando-os do comportamento dos *inliers*.

5.1.1 Outliers Internos

O cenário ilustrado pela Figura 21 contém dois agrupamentos de dados em forma de banana e dois pontos isolados em diferentes localizações. Um encontra-se no quadrante superior esquerdo, posicionado de maneira mais externa a ambos os agrupamentos; de modo similar aos cenários previamente apresentados, mas neste contexto podemos denominar este ponto como um *outlier externo*. Em contraste, o outro ponto isolado localiza-se entre os dois agrupamentos; o qual denominamos por *outlier interno*.

Figura 21. Cenário com *outlier interno* e *externo*. (a) Conjunto de dados com a pontuação de *outlier* de cada ponto. (b) Quantidade de participação nos atratores em relação ao tamanho da memória, cada linha representa um determinado item do conjunto de dados. A linha de asteriscos vermelhos e linha de asteriscos azuis (parcialmente sobrepostas) representam os itens de dado isolados interno e externo, respectivamente.

Após a execução do experimento, ambos *outliers* foram identificados com os maiores valores de pontuação; comportamento desejável. Tendo o *outlier externo* o maior valor de pontuação de *outlier* entre todos os elementos presentes no conjunto de dados e o *outlier interno* o segundo maior valor. Este experimento foi realizado para diferentes estratégias de aceitação de atratores, que impactaram apenas superficialmente nas pontuações calculadas. Independentemente da estratégia adotada, os *outliers externo* e *interno* se mantiveram sempre no topo da classificação.

Podemos observar na Figura 21(b) a participação nos atratores contra o tamanho da memória até que seu valor crítico seja alcançado. Para esta figura, todos os atratores extraídos foram mantidos, para facilitar a análise e diferenciação da dinâmica de detecção de *outliers externos* e *internos*. Novamente os pontos isolados (linha de asteriscos vermelhos e linha de

asteriscos azuis, que representam os *outliers interno* e *externo* respectivamente) não participaram dos atratores extraídos para faixas de valores intermediários de memória. Mas é possível verificar que se diferenciam entre si, pois o *outlier interno* participou em vários atratores extraídos com grandes valores de memória (muitos dos quais eram repetições idênticas; sendo apenas 5 deles distintos). Enquanto que, o *outlier externo* participou apenas do atrator global que encerrou o processo de caminhada.

O mesmo experimento apresentado na Figura 21(a) foi reproduzido na Figura 22, porém sem o identificador de atratores com cruzamento; de modo a podermos ilustrar o impacto desses na dinâmica do método de detecção. O procedimento algorítmico utilizado no experimento é basicamente o mesmo, tendo ainda como base caminhantes turista que seguem as mesmas condições apresentadas nos experimentos anteriores. A diferença é que neste caso serão extraídos apenas atratores sem pontos de interseção em suas trajetórias. Para os casos em que um caminhante deveria ficar preso num atrator com cruzamento, é extraído apenas o primeiro semiciclo encontrado. Os resultados obtidos mostram que não houve impacto direto na identificação dos *outliers* reais (visto que ambos os *outliers* presentes ainda situam-se no topo da lista de classificação), mas claramente houve uma transformação no cenário geral do experimento. Os agrupamentos que eram claramente diferenciados na Figura 21(a) agora possuem pontuações de *outlier* bastante misturadas, o que certamente impacta na acurácia da técnica em cenários maiores e mais complexos, já que os detalhes estruturais não foram preservados e detectados precisamente (como no experimento anterior). Este experimento reforça a relevância da dinâmica complexa dos atratores com cruzamento para o melhor desempenho do método de detecção proposto.

Figura 22. Cenário com *outlier interno* e *externo*, desconsiderando atratores com cruzamento. As cores representam a pontuação de *outlier* para cada ponto.

Os cenários apresentados nos permitem ilustrar que a técnica de detecção de *outliers* proposta é capaz de identificar corretamente os *outliers* em diversas situações; das mais simples e óbvias até aquelas menos triviais. Mostrando assim, que essa nova abordagem de detecção é funcional e robusta.

5.2 Experimentos com Dados Reais

Nessa seção descrevemos os experimentos realizados com conjuntos de dados reais, os quais foram obtidos no repositório de aprendizado de máquinas da UCI (FRANK; ASUNCION, 2010).

Wisconsin Breast Cancer (Original): O conjunto completo possui 683 instâncias clínicas divididas em duas classes: benigna (444 instâncias) e maligna (239 instâncias). Para realização dos experimentos, removemos os atributos de *ID* e *classe* e produzimos um subconjunto dos dados originais com 454 instâncias, 10 das quais foram selecionadas aleatoriamente da classe maligna. Portanto, mantivemos inalterada a classe majoritária (benigna) de dados normais e reduzimos a classe de dados malignos; para considerarmos como *outlier*.

A Tabela 1 resume o percentual dos *outliers* detectados por três técnicas clássicas: *LOF* (BREUNIG et al., 2000), *NDoT* (HUBBALLI; PATRA; NANDI, 2011) e *SSOD* (GAO; CHENG; TAN, 2006); e compara com o resultado obtido com a técnica proposta. Os resultados numéricos apresentados são a média para 20 execuções com 10 *outliers* selecionados aleatoriamente da classe maligna (20 configurações de dados distintas), sendo a acurácia definida em termos da classificação dos pontos que obtiveram maior pontuação de *outlier*; onde são considerados os 10 maiores valores e é calculada a proporção dos acertos, ou seja, quantos dos elementos dessa lista são realmente *outliers* esperados. Ao lado do valor da acurácia é exibido o intervalo de confiança das medidas, calculado com base numa significância de 5%. Ratificamos que a técnica proposta é determinística, isto é, para cada configuração de dados distinta, é necessária apenas uma execução.

Tabela 1. Conjunto de dados Wisconsin breast cancer: Comparação entre as técnicas de detecção de outliers mais conhecidas e a técnica proposta.

<i>LOF</i>	<i>NDoT</i>	<i>SSOD</i>	<i>Técnica proposta</i>
81.50 ± 11.37	80.00 ± 13.76	81.50 ± 9.88	84.00 ± 7.83

No conjunto de dados *Wisconsin breast cancer*, nossa técnica pode encontrar a maioria das instâncias da classe maligna, alcançando uma acurácia de 84.00%, isto é, ela frequentemente encontra 8 em 10 *outliers*. Este resultado é melhor que todas as outras técnicas em comparação. É importante observar que nenhuma das técnicas obteve acurácia máxima provavelmente por conta de que algumas instâncias normais do conjunto de dados situam-se em regiões de baixa densidade ou certas instâncias da classe maligna estão muito próximas dos agrupamentos de dados normais (podendo inclusive apresentar certo nível de sobreposição). A Figura 23 remete a uma das 20 configurações de dados (na qual foi obtida uma acurácia de 90%), nela podemos verificar o mesmo comportamento distinto exibido pelos *outliers* (linhas de asteriscos vermelhos sobrepostas), também identificado e evidenciado nos experimentos com cenários de dados artificiais.

Figura 23. Experimento com a base *Wisconsin breast câncer*. Participação das amostras de dados nos atratores conforme crescimento do valor de memória. As linhas de asteriscos vermelhos sobrepostas referem-se aos itens de dado isolados.

Ionosphere: Este conjunto de dados possui 351 instâncias para 34 atributos distintos, os elementos de dados dividem-se em duas classes: boa (225 instâncias) e ruim (126 instâncias); sem valores faltantes. Para o experimento, foram removidos atributos de *ID* e *classe*. Neste contexto, a classe ruim refere-se ao retorno de sinais anômalos (que atravessaram a ionosfera), ou seja, *outliers*, enquanto que a classe boa é composta por dados obtidos a partir dos sinais que evidenciam algum tipo de estrutura ionosférica e, portanto são considerados *inliers*. Para este experimento foram mantidas as proporções originalmente apresentadas pelo conjunto de dados.

No conjunto de dados *Ionosphere*, apenas uma execução foi necessária, dado que as proporções originais dos dados foram mantidas; não sendo necessário nenhum tipo de seleção aleatória para redução da classe de *outliers*. Para comparar os resultados obtidos selecionamos na literatura aquelas técnicas que apresentaram os melhores resultados quando aplicadas neste conjunto de dados (CAMPOS et al., 2016). Estes resultados estão organizados na Tabela 2 em ordem crescente de melhor acurácia obtida, sendo FastABOD (KRIEGEL; S HUBERT; ZIMEK, 2008) a técnica que obteve melhor acurácia com de 85.71%, seguida pela KNNW (ANGIULLI; PIZZUTI, 2005), KNN (RAMASWAMY; RASTOGI; SHIM, 2000), COF (TANG et al., 2002) e LOF; todas com resultados inferiores à primeira. Neste experimento, a técnica proposta também encontrou a maioria dos *outliers* pertencentes à classe ruim, alcançando uma acurácia de 87.30%; o melhor resultado dentre todas as técnicas comparadas na Tabela 2.

Figura 24. Experimento com a base *Ionosphere*. (a) Participação dos dados normais conforme o crescimento do valor de memória. (b) Participação dos *outliers* esperados conforme o crescimento do valor de memória.

A Figura 24(a) mostra o comportamento dos *inliers* no conjunto de dados *Ionosphere* conforme é incrementado o valor de memória das caminhadas. Enquanto que na Figura 24(b) é ilustrado o comportamento dos *outliers* esperados com relação a esta mesma evolução das caminhadas. Ao compararmos as duas imagens, é possível distinguir claramente os diferentes comportamentos. Os *outliers* da Figura 24(b) se concentram no quadrante inferior direito, ou seja, possuem a tendência de se manifestar apenas para grandes valores de memória, já próximos ao valor crítico; participação mais tardia nos atratores. Enquanto os dados normais tendem a ser mais ativos na zona de valores intermediários de memória. As linhas da Figura

24(b) que se destacam de maneira mais independente (como a linha laranja que já se inicia próxima ao valor 50) correspondem justamente aos *outliers* esperados que não foram corretamente identificados, já que se desviam do comportamento esperado pela técnica proposta.

Tabela 2. Conjunto de dados *Ionosphere*: Comparação entre as técnicas de detecção de *outliers* mais conhecidas e a técnica proposta.

<i>FastABOD</i>	<i>KNNW</i>	<i>KNN</i>	<i>COF</i>	<i>LOF</i>	<i>Técnica proposta</i>
85.71	85.71	84.92	84.92	83.33	87.30

Em suma, nossa técnica mostrou resultados promissores, tendo detectado a maioria dos *outliers* presentes nas bases reais acima ilustradas.

5.3 Considerações Finais

Neste capítulo apresentamos diversos experimentos computacionais para avaliar a técnica de detecção de *outliers* baseada no algoritmo de caminha do turista em diversos cenários de dados artificiais e reais.

Os cenários foram dispostos, na medida do possível, em ordem crescente de dificuldade para determinação dos *outliers* e a técnica se mostrou bastante robusta e de boa acurácia, sendo capaz de identificar todos os *outliers* esperados nos cenários artificiais.

Nos experimentos com conjuntos de dados reais, a técnica alcançou resultados compatíveis com o estado da arte, identificando a maioria dos *outliers* esperados e mostrando ser robusta, também, nestes cenários.

Assim, finalizamos o capítulo com a impressão de que se trata de uma nova abordagem para de detecção de *outliers* bastante promissora e seguimos para o capítulo final com as conclusões obtidas com a execução da pesquisa.

Capítulo 6

CONCLUSÕES

Aqui são descritas as conclusões do presente trabalho, com enfoque para as características que foram evidenciadas após a análise dos experimentos computacionais com o método de detecção de outlier que propusemos. E, por fim, são apresentados os trabalhos futuros.

Os resultados apresentados através dos experimentos computacionais descritos neste trabalho mostraram que o método de detecção de *outlier* que aqui propusemos é funcional e robusto, sendo capaz de identificar *outliers* presentes em cenários de dados bastante desafiadores e, inclusive, em bases de dados reais de alta dimensionalidade sem exigir qualquer conhecimento prévio sobre as mesmas. Em certos cenários nossa técnica foi capaz, inclusive, de superar os resultados das técnicas no estado da arte em detecção de *outliers*.

É importante mencionarmos, ainda, que a técnica de detecção proposta, além de apresentar os resultados positivos acima destacados configura-se como uma inovação conceitual na abordagem de identificação de anomalias. Isso, por ter como base um mecanismo de detecção de estruturas de dados de alto nível, ao invés de utilizar-se apenas de características físicas dos espaços de dados (como distância, similaridade ou densidade).

Foram apresentadas evidências de que a caminhada do turista pode gerar atratores complexos exibindo formas variadas com pontos de cruzamento. Achado este que revela estruturas de dados ainda mais detalhadas e, conseqüentemente, melhora a acurácia do método de detecção de *outlier* proposto.

Por fim, propusemos, também, uma nova subclassificação dos *outliers* pontuais em: *outliers externos* e *outliers internos*. Sendo os *outliers internos* aqueles que se situam em regiões entre dois ou mais agrupamentos de dados; geralmente mais difíceis de serem detectados e os *outliers externos* aqueles mais intuitivos e comumente abordados nos diversos

trabalhos da área. Mostramos através dos experimentos computacionais que o método proposto foi, também, capaz de detectar estes *outliers internos*.

6.1 Publicações

O trabalho aqui apresentado gerou um artigo de conferência, intitulado “*Structural Outlier Detection: A Tourist Walk Approach*”, publicado em “*Proceedings of 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017)*”, 2017, v.1, p. 366-371; em colaboração com os coautores: ZHAO, LIANG; ZHENG, Q.; ZHANG, J.

Esse artigo foi selecionado pelo comitê da referida conferência, sendo recomendado para publicação da versão estendida numa edição especial da revista *Neurocomputing*; para a qual submetemos o manuscrito “*A Tourist Walk Applied to Internal and External Outlier Detection*”, ainda em processo de revisão.

No primeiro artigo apresentamos a nova abordagem de detecção de *outliers* com uso da caminhada determinística do turista e no último relatamos o achado relacionando o fenômeno de identificação de atratores com cruzamento observado na dinâmica das caminhadas e caracterizamos os *outliers internos*.

6.2 Trabalhos Futuros

Uma vez que os resultados apresentados pela técnica proposta foram bastante positivos, pretendemos estender os conceitos identificados e adaptá-los para o problema de detecção de *outliers* em ambientes semissupervisionados, verificando também seu comportamento para uma abordagem hierárquica na detecção de *outliers*.

Além disso, pretendemos conduzir estudos teóricos mais aprofundados sobre os fenômenos dos atratores com cruzamento e avaliar a aplicação dessa dinâmica em outras tarefas de mineração de dados e aprendizado de máquina; como caracterização de dados, agrupamento e classificação.

Outros aspectos que merecem atenção e podem ser trabalhados no futuro remetem ao ajuste e refinamento da técnica para apresentar uma execução mais eficiente, avaliar diferentes critérios de parada para as caminhadas do turista e comparar diferentes mecanismos de aceitação dos atratores obtidos.

Por fim, gostaríamos ainda de avaliar a técnica em cenários de dados incrementais e verificar a possibilidade de estabelecer critérios específicos para a caminhada nesse caso, de modo a extrair as medidas para as novas amostras recém-incluídas, sem a necessidade de executar todos os ciclos de caminhadas.

REFERÊNCIAS

AGGARWAL, C. C. **Outlier Analysis**. New York: Springer, 2013.

AGGARWAL, C. C.; SATHE, S. **Outlier Ensembles: An Introduction**. [s.l.] Springer, 2017.

ANGIULLI, F. et al. GPU Strategies for Distance-Based Outlier Detection. **IEEE Transactions on Parallel and Distributed Systems**, v. 27, n. 11, p. 3256–3268, nov. 2016.

ANGIULLI, F.; PIZZUTI, C. **Fast Outlier Detection in High Dimensional Spaces**. Principles of Data Mining and Knowledge Discovery. **Anais...: Lecture Notes in Computer Science**. In: EUROPEAN CONFERENCE ON PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY. Springer, Berlin, Heidelberg, 19 ago. 2002.

ANGIULLI, F.; PIZZUTI, C. Outlier mining in large high-dimensional data sets. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 2, p. 203–215, fev. 2005.

BACKES, A. R. et al. Texture analysis and classification using deterministic tourist walk. **Pattern Recognition**, v. 43, n. 3, p. 685–694, mar. 2010.

BACKES, A. R.; MARTINEZ, A. S.; BRUNO, O. M. Texture analysis using graphs generated by deterministic partially self-avoiding walks. **Pattern Recognition**, v. 44, n. 8, p. 1684–1689, 1 ago. 2011.

BARNETT, V.; LEWIS, T. **Outliers in Statistical Data**. Edição: 3 ed. Chichester; New York: Wiley-Blackwell, 1994.

BREUNIG, M. M. et al. **LOF: Identifying Density-based Local Outliers**. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. **Anais...: SIGMOD '00**. New York, NY, USA: ACM, 2000.

BUNIMOVICH, L. A. Deterministic walks in random environments. **Physica D: Nonlinear Phenomena**, Microscopic Chaos and Transport in Many-Particle Systems. v. 187, n. 1, p. 20–29, 1 jan. 2004.

CAMPITELI, M. G. et al. Deterministic walks as an algorithm of pattern recognition. **Physical Review E**, v. 74, n. 2, p. 026703, 4 ago. 2006.

CAMPOS, G. O. et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. **Data Mining and Knowledge Discovery**, v. 30, n. 4, p. 891–927, 1 jul. 2016.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly Detection: A Survey. **ACM Computing Surveys**, v. 41, n. 3, p. 15:1–15:58, jul. 2009.

CHRISTIE, M. **The Ozone Layer: A Philosophy of Science Perspective**. 1 edition ed. Cambridge, UK ; New York: Cambridge University Press, 2001.

ESTER, M. et al. **A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. **Anais...**: KDD'96.Portland, Oregon: AAAI Press, 1996.

FARMAN, J. C.; GARDINER, B. G.; SHANKLIN, J. D. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. **Nature**, v. 315, n. 6016, p. 207–210, maio 1985.

FRANK, A.; ASUNCION, A. **UCI Machine Learning Repository**. Disponível em: <<http://archive.ics.uci.edu/ml/>>.

FRÉDAY, B.; VERLEYSEN, M. Reinforced Extreme Learning Machines for Fast Robust Regression in the Presence of Outliers. **IEEE Transactions on Cybernetics**, v. 46, n. 12, p. 3351–3363, dez. 2016.

FREUND, H.; GRASSBERGER, P. The red queen's walk. **Physica A: Statistical Mechanics and its Applications**, v. 190, n. 3, p. 218–237, 1 dez. 1992.

GAO, J.; CHENG, H.; TAN, P.-N. **Semi-supervised Outlier Detection**. Proceedings of the 2006 ACM Symposium on Applied Computing. **Anais...**: SAC '06.New York, NY, USA: ACM, 2006.

GRUBBS, F. E. Procedures for Detecting Outlying Observations in Samples. **Technometrics**, v. 11, n. 1, p. 1–21, 1 fev. 1969.

HAWKINS, D. **Identification of Outliers**. Dordrecht: Springer, 1980.

HILBE, J. M. (ED.). **Astrostatistical Challenges for the New Astronomy**. 2013 edition ed. Place of publication not identified: Springer, 2014.

HODGE, V. J.; AUSTIN, J. A Survey of Outlier Detection Methodologies. **Artificial Intelligence Review**, v. 22, n. 2, p. 85–126, 1 out. 2004.

HUBBALLI, N.; PATRA, B. K.; NANDI, S. **NDoT: Nearest Neighbor Distance Based Outlier Detection Technique**. Pattern Recognition and Machine Intelligence. **Anais...** In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION AND MACHINE INTELLIGENCE. Springer, Berlin, Heidelberg, 27 jun. 2011.

HUBERTY, C. J.; OLEJNIK, S. **Applied MANOVA and Discriminant Analysis**. Hoboken, N.J: Wiley-Interscience, 2006.

JIN, W. et al. **Ranking Outliers Using Symmetric Neighborhood Relationship**. Advances in Knowledge Discovery and Data Mining. **Anais...**: Lecture Notes in Computer Science. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Springer, Berlin, Heidelberg, 9 abr. 2006.

KANG, M. et al. A Hybrid Feature Selection Scheme for Reducing Diagnostic Performance Deterioration Caused by Outliers in Data-Driven Diagnostics. **IEEE Transactions on Industrial Electronics**, v. 63, n. 5, p. 3299–3310, maio 2016.

KNORR, E. M.; NG, R. T. **A Unified Approach for Mining Outliers**. Proceedings of the 1997 Conference of the Centre for Advanced Studies on Collaborative Research. **Anais...**: CASCON '97.Toronto, Ontario, Canada: IBM Press, 1997.

KNORR, E. M.; NG, R. T. **Algorithms for Mining Distance-Based Outliers in Large Datasets**. Proceedings of the 24rd International Conference on Very Large Data Bases. **Anais...**: VLDB '98.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998.

KNORR, E. M.; NG, R. T.; TUCAKOV, V. Distance-based outliers: algorithms and applications. **The VLDB Journal**, v. 8, n. 3–4, p. 237–253, 1 fev. 2000.

KOU, Y. et al. **Survey of fraud detection techniques**. IEEE International Conference on Networking, Sensing and Control, 2004. **Anais...** In: IEEE INTERNATIONAL CONFERENCE ON NETWORKING, SENSING AND CONTROL, 2004. 2004

KRIEGEL, H.-P. et al. **LoOP: Local Outlier Probabilities**. Proceedings of the 18th ACM Conference on Information and Knowledge Management. **Anais...**: CIKM '09.New York, NY, USA: ACM, 2009.

KRIEGEL, H.-P.; S HUBERT, M.; ZIMEK, A. **Angle-based Outlier Detection in High-dimensional Data**. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais...**: KDD '08.New York, NY, USA: ACM, 2008.

LATECKI, L. J.; LAZAREVIC, A.; POKRAJAC, D. **Outlier Detection with Kernel Density Functions**. Machine Learning and Data Mining in Pattern Recognition. **Anais...**: Lecture Notes in Computer Science. In: INTERNATIONAL WORKSHOP ON MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION. Springer, Berlin, Heidelberg, 18 jul. 2007.

LIMA, G. F.; MARTINEZ, A. S.; KINOCHI, O. Deterministic Walks in Random Media. **Physical Review Letters**, v. 87, n. 1, p. 010603, 19 jun. 2001.

PATHAK, M. A. **Beginning Data Science with R**. 2014 edition ed. Cham: Springer, 2014.

PREPARATA, F. P.; SHAMOS, M. I. **Computational Geometry: An Introduction**. F First Edition edition ed. New York: Springer, 1993.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. **Efficient Algorithms for Mining Outliers from Large Data Sets**. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. **Anais...**: SIGMOD '00.New York, NY, USA: ACM, 2000.

- SALEHI, M. et al. Fast Memory Efficient Local Outlier Detection in Data Streams. **IEEE Transactions on Knowledge and Data Engineering**, v. 28, n. 12, p. 3246–3260, dez. 2016.
- SCHUBERT, E. et al. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. **ACM Trans. Database Syst.**, v. 42, n. 3, p. 19:1–19:21, jul. 2017.
- SCHUBERT, E.; ZIMEK, A.; KRIEGEL, H. Generalized Outlier Detection with Flexible Kernel Density Estimates. In: **Proceedings of the 2014 SIAM International Conference on Data Mining**. Proceedings. [s.l.] Society for Industrial and Applied Mathematics, 2014a. p. 542–550.
- SCHUBERT, E.; ZIMEK, A.; KRIEGEL, H.-P. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. **Data Mining and Knowledge Discovery**, v. 28, n. 1, p. 190–237, 1 jan. 2014b.
- SILVA, T. C.; ZHAO, L. High-level pattern-based classification via tourist walks in networks. **Information Sciences**, Innovative Applications of Artificial Neural Networks in Engineering. v. 294, p. 109–126, 10 fev. 2015.
- SILVA, T. C.; ZHAO, L. **Machine Learning in Complex Networks**. 1. ed. [s.l.] Springer, 2016.
- STANLEY, H. E.; BULDYREV, S. V. Statistical physics: The salesman and the tourist. **Nature**, v. 413, n. 6854, p. 373–374, set. 2001.
- STEIN, R. B.; GOSSEN, E. R.; JONES, K. E. Neuronal variability: noise or part of the signal? **Nature Reviews Neuroscience**, v. 6, n. 5, p. 389–397, maio 2005.
- STOLARSKI, R. S. et al. Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease. **Nature**, v. 322, n. 6082, p. 808–811, ago. 1986.
- SU, X.; TSAI, C.-L. Outlier detection. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 3, p. 261–268, 1 maio 2011.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 1 edition ed. Boston: Pearson, 2005.
- TANG, J. et al. **Enhancing Effectiveness of Outlier Detections for Low Density Patterns**. Advances in Knowledge Discovery and Data Mining. **Anais...**: Lecture Notes in Computer Science. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Springer, Berlin, Heidelberg, 6 maio 2002.
- TUKEY, J. W. **Exploratory Data Analysis**. 1 edition ed. Reading, Mass.: Pearson, 1977.
- WONG, W.-K. et al. **Bayesian Network Anomaly Pattern Detection for Disease Outbreaks**. Proceedings of the Twentieth International Conference on International Conference on Machine Learning. **Anais...**: ICML'03. Washington, DC, USA: AAAI Press, 2003.

ZHANG, K.; HUTTER, M.; JIN, H. **A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data**. Advances in Knowledge Discovery and Data Mining. **Anais...**: Lecture Notes in Computer Science. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Springer, Berlin, Heidelberg, 27 abr. 2009.

ZIMEK, A.; SCHUBERT, E.; KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. **Statistical Analysis and Data Mining**, v. 5, n. 5, p. 363–387, 1 out. 2012.