MARA ANDRÉA DOTA

Modelo para a classificação da qualidade da água contaminada por solo
usando indução por árvore de decisão

São Paulo
2014

MARA ANDRÉA DOTA

Modelo para a classificação da qualidade da água contaminada por solo
usando Indução por Árvore de Decisão

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências

São Paulo
2014

MARA ANDRÉA DOTA

Modelo para a classificação da qualidade da água contaminada por solo
usando Indução por Árvore de Decisão

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências

Área de Concentração: Engenharia de
Computação

Orientador: Prof. Livre-Docente Carlos
Eduardo Cugnasca

São Paulo
2014

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 11 de novembro de 2014.

Assinatura do autor Mara Andrea Dota

Assinatura do orientador _____

Catálogo-na-publicação

Dota, Mara Andrea

Modelo para a classificação da qualidade da água contaminada por solo usando indução por árvore de decisão / M.A. Dota. -- versão corr. -
- São Paulo, 2014.
68 p.

Tese (Doutorado) - Escola Politécnica da Universidade de
São Paulo. Departamento de Engenharia de Computação e Sistemas
Digitais.

1.Inteligência artificial 2.Mineração de dados 3.Redes de sensores
sem fio 4.Fusão de sensores 5.Controle ambiental 6.Enxurradas I. Uni-
versidade de São Pau-lo. Escola Politécnica. Departamento de Enge-
nharia de Computação e Sistemas Digitais II.t.

Dedico este trabalho aos meus pais, Maria e Milton, à minha madrinha Elizabeth e à minha filha Alicia Yumi pela compreensão, dedicação, apoio e incentivo para que isso fosse possível.

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que me acompanharam nessa minha jornada de grande importância para minha vida profissional. Primeiramente, agradeço aos meus pais Maria e Milton por todo apoio, dedicação e ajuda que sempre me ofereceram, primordialmente em relação a minha filha Alicia. Agradeço a minha madrinha querida que é minha segunda mãe e amiga, Elizabeth de Almeida, quem me ofereceu todo suporte ao longo dessa caminhada.

Meu obrigada especial ao meu orientador professor Dr. Carlos Eduardo Cugnasca que me guiou nessa fase tão importante do meu crescimento profissional e pela oportunidade de ser MEU ORIENTADOR. Agradeço também aos professores do LAA Dr. Antônio Saraiva, Dr. André Hirakawa, a professora Dra. Anna Helena Reali Costa e a professora Cristina Borba. Agradeço aos técnicos do LAA: Sra. Lourdes, Edson e Suzano.

Agradecimento muito especial ao professor Dr. Domingos Sávio Barbosa, pela contribuição, ajuda constante e paciência comigo.

Aos meus colegas de pós-graduação: Clovis dos Santos Junior, Marcelo Freire de Barros, Ivairton Monteiro Santos, Juliana Saragiotto, Raul Teruel, Wesley Barbosa Thereza, Allan Kock, Wilian França, Luiz Lamardo, Felipe Leno e Leonardo Campos. Obrigada pela amizade e palavras de incentivo.

Quero agradecer as minhas amigas de coração que sempre estiveram por perto, me incentivando nessa jornada: Marcia Regina Dota Loureiro (irmã acima de tudo), a dupla Caroline Chiodini e Patrícia Carvalho, Genilce de Castro Alves, Lorryne Erika (prima). De Alicia Yumi, somente espero que me perdoe por estar ausente em tantos momentos importante da sua vida dos 3 aos 8 anos de idade.

E agradeço a Deus, por sempre me proteger e de formas enigmáticas me conduzir continuamente para o sucesso.

RESUMO

A possibilidade de avaliar remotamente e de forma instantânea alterações na qualidade das águas em função da entrada de solos permite o monitoramento de processos ecológicos como o assoreamento, perdas e solos, carreamento de pesticidas e degradação de *habitats* aquáticos. Com a utilização de um modelo automatizado, torna-se possível um monitoramento em tempo real remoto coletando dados por meio de Redes de Sensores Sem Fio. Esta pesquisa propõe um modelo de classificação da qualidade da água contaminada por solo usando técnicas de Árvore de Decisão. Com este modelo torna-se possível acompanhar alterações que venham a ocorrer em águas superficiais indicando o nível de contaminação por solo com maior rapidez do que a forma convencional que necessita de análise em laboratório e coleta de amostra manual. A classificação proposta considera sete classes de qualidade da água, conforme dados de um experimento conduzido em laboratório. Foram utilizadas técnicas de Inteligência Artificial com o intuito de realizar a Fusão de Sensores para avaliar, em tempo real, as leituras dos sensores, indicando a qual classe de qualidade a amostra se enquadra. Na verificação de quantas classes seria o ideal, utilizou-se o algoritmo k-means++. Para a construção do modelo de classificação foram usadas técnicas de Indução por Árvore de Decisão, tais como: *Best-First Decision Tree Classifier* – BFTree, *Functional Trees* – FT, *Naïve Bayes Decision Tree* – NBTree, *Grafted C4.5 Decision Tree* – J48graft, *C4.5 Decision Tree* – J48, LADTree. Os testes realizados indicam que a classificação proposta é coerente, visto que os diferentes algoritmos comprovaram uma relação estatística forte entre as instâncias das classes, garantindo que o modelo proposto irá prever saídas para entradas de dados desconhecidas com acurácia. Os algoritmos com melhores resultados foram FT, J48graft e J48.

Palavras-chave: Inteligência artificial. Mineração de dados. Redes de sensores sem fio. Fusão de sensores. Controle ambiental. Enxurradas.

ABSTRACT

The possibility to remotely and instantaneously evaluate changes in water quality due to soil contamination allows monitoring ecological processes such as siltation, soil losses, loading of pesticides and degradation of aquatic habitats. Using an automated model to classify soil-contaminated water quality allows for a remote real-time monitoring by collecting data using Wireless Sensor Networks. This study proposes a model to classify soil-contaminated water quality by using Decision Tree techniques. With this model, it is possible to track changes that may occur in surface waters indicating the level of contamination by soil faster than the conventional way, which requires laboratory analysis and manual sampling. The classification proposed considers seven classes of water quality, according to data from an experiment carried out in laboratory. Artificial Intelligence techniques were used in order to implement Sensor Fusion to evaluate, in real time, sensor readings to which class the sample quality fits. By checking how many classes would be ideal, the k-means + + algorithm was used. To build the classification model, Decision Tree Induction techniques were used, such as: Best-First Decision Tree Classifier – BFTree, Functional Trees – FT, Naïve Bayes Decision Tree – NBTree, Grafted C4.5 Decision Tree – J48graft, C4.5 Decision Tree – J48, LADTree. Tests indicated that the proposed classification is consistent because different algorithms results confirmed a strong statistical relationship between instances of classes, ensuring that this model will predict outputs to unknown inputs accurately. The algorithms with best results were FT, J48graft and J48.

Keywords: Artificial intelligence. Data mining. Wireless sensor network. Sensor fusion. Environmental control. Runoff.

LISTA DE FIGURAS

Figura 1 – Relação entre os termos Fusão de Dados, Fusão de Informação e Fusão de Sensores.....	16
Figura 2 – Fusão complementar, competitiva e cooperativa.....	18
Figura 3 – Modelo JDL.....	21
Figura 4 – Modelo em Cascata Modificado.....	22
Figura 5 - Modelo BCL.....	24
Figura 6 – Modelo Omnibus.....	25
Figura 7 – Arquitetura da RSSF do experimento realizado por Xiong e Wang (2009).....	30
Figura 8 – Esquema de descrição do modelo de classificação da qualidade da água contaminada por suspensão de solo.....	32
Figura 9 – Técnica de agrupamento na detecção de <i>outliers</i>	35
Figura 10 – Sonda usada para coletar os dados no experimento em laboratório.....	39
Figura 11 – Alguns valores estatísticos de cada variável são apresentados aqui, como os valores máximo e mínimo, média e desvio padrão. DO = dissolved oxygen; TDS = Total dissolved Solids; pH = Hydrogen Potential; ORP = Oxidation-Reduction Potential; Scm = electric conductance; MOhm = electric resistance.....	42
Figura 12 – Erro calculado para cada parâmetro de agrupamento.....	47
Figura 13 – Distribuição das instâncias em agrupamento de 7 classes. Cada grupo esperado foi colorido diferente entre si, o agrupamento definido pelo k-means++ foi atribuído ao eixo y.....	47
Figura 14 – Resultado de ICC e IIC para cada algoritmos de classificação testado usando k-fold-cross-validation. Note que os algoritmos mais a direita obtiveram menores valores de ICC.....	48
Figura 15 – Valores de kappa obtidos pela execução dos algoritmos.....	49
Figura 16 – É apresentada o fluxo de dados de entrada e a saída que será a classificação de acordo com as classes aqui definidas.....	49

LISTA DE QUADROS E TABELAS

Quadro 1 – Modelo de Fusão em Níveis.....	28
Tabela 1 – Lista das variáveis coletadas durante o experimento.....	40
Tabela 2 – Concentração nominal estimada de solo em função do tempo.....	41
Tabela 3 – Variação, Média e Desvio Padrão das variáveis referentes a qualidade da água contaminada por solo usadas para definição do modelo de classificação.....	41
Tabela 4 – ICC representam Instâncias Corretamente Classificadas e IIC representam Instâncias Incorretamente Classificadas. Kappa é um valor que representa a relação entre instâncias da mesma classe, variando entre 0 e 1, sendo 1 uma forte relação entre elas.....	48

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore de Decisão
AM	Aprendizado de Máquina
AP	Agricultura de Precisão
ARFF	Attribute Relation File Format
BCL	Modelo <i>Boyd</i>
BFTREE	<i>Best-First Decision Tree Classifier</i>
CETESB	Companhia Ambiental do Estado de São Paulo
CI	Ciclo da Inteligência
CONAMA	Conselho Nacional Do Meio Ambiente
FD	Fusão de Dados
FI	Fusão de Informação
FS	Fusão de Sensores
FT	<i>Functional Tree</i>
GPL	<i>General Public License</i>
IA	Inteligência Artificial
IAD	Indução por Árvore de Decisão
ICC	Instâncias Corretamente Classificadas
IHC	Interação Homem-Computador
IIC	Instâncias Incorretamente Classificadas
IQA	Índice de Qualidade de Água
J48	C4.5 Decision Tree
J48GRAFT	<i>Grafted C4.5 Decision Tree</i>
JDL	<i>Joint Directors of Laboratories</i>
LAA	Laboratório de Automação Agrícola
MOHM	Resistência Elétrica
NBTREE	Naïve Bayes Decision Tree
OB	Modelo <i>Omnibus</i>
OD	Oxigênio Dissolvido
OODA	<i>Observer, Orientate, Decide and Act</i>
ORP	Potencial de Oxi-Redução

PH	Potencial Hidrogeniônico
RSSF	Rede de Sensores Sem Fio
SIG	Sistemas de Informação Geográfica
SISNAMA	Sistema Nacional do Meio Ambiente
TDS	Sólidos Totais Dissolvidos
USEPA	<i>United State Environment Protection Agency</i>
WEKA	<i>Waikato Environment For Knowledge Analysis</i>
WFFM	<i>Waterfall Fusion Model</i>
μ	Condutividade Elétrica

SUMÁRIO

1.INTRODUÇÃO.....	12
1.1 OBJETIVO.....	13
1.2 MOTIVAÇÃO.....	13
1.3 JUSTIFICATIVA.....	14
1.4 ORGANIZAÇÃO DA TESE.....	15
2. REVISÃO DE LITERATURA.....	16
2.1 FUSÃO DE SENSORES.....	16
2.1.1 Modelos de Fusão de Sensores.....	18
2.1.1.1 Modelo JDL.....	19
2.1.1.2 Modelo de Fusão Cascata Modificado	22
2.1.1.3 Modelo Baseado no Ciclo da Inteligência	23
2.1.1.4 Modelo Boyd.....	24
2.1.1.5 Modelo Omnibus.....	25
2.1.1.6 Modelo de Fusão em Níveis.....	26
2.1.2 Algumas aplicações.....	29
2.1.3 Considerações finais.....	32
2.2 APRENDIZADO DE MÁQUINA.....	33
2.2.1 Aprendizagem não-supervisionada.....	33
2.2.2 Aprendizagem supervisionada.....	35
2.2.1.1 Árvore de decisão.....	35
2.3 CONSIDERAÇÕES FINAIS.....	36
3. MATERIAIS E MÉTODOS.....	37
3.1 DESCRIÇÃO DO EXPERIMENTO REALIZADO EM LABORATÓRIO.....	39
3.2 PROCESSAMENTO DOS DADOS E ANÁLISE.....	42
3.3 CONSIDERAÇÕES FINAIS.....	46
4. RESULTADOS E DISCUSSÕES.....	47
4.1 CONSIDERAÇÕES FINAIS.....	50
5. CONCLUSÕES.....	51
REFERÊNCIAS.....	54

APÊNDICE A 64

1. INTRODUÇÃO

A agricultura vem passando por diversos problemas, como a necessidade de se adaptar às mudanças climáticas, preservação do meio ambiente e também questões econômicas, políticas e sociais. É preciso aumentar a produtividade das áreas cultivadas, reduzir os custos de produção e utilizar adequadamente os recursos naturais, como a água, por exemplo, para impactar o mínimo possível no meio ambiente.

Nesse contexto, a Tecnologia da Informação e Comunicação (TIC) aplicada à Agricultura (Agricultura de Precisão - AP) pode contribuir no uso da tecnologia para melhorar o gerenciamento agrícola, como o manejo do solo, insumos e culturas.

A AP visa fazer o controle e monitoramento das áreas cultivadas em unidades físicas menores do que um talhão, pois dentro dele as características podem variar. Segundo Molin (2003), a proposta da AP é “permitir que se faça em áreas extensas o que os pequenos agricultores sempre fizeram que é o tratamento dos detalhes considerando as diferenças existentes em um talhão”. Assim, é preciso conhecer as características específicas de uma unidade de área de plantio e com exatidão as suas deficiências e necessidades (SRIVASTAVA et al., 2006). A automação na agricultura envolve a aplicação dos conceitos de TIC nos processos produtivos agrícolas, processos que fazem parte das tarefas de monitoramento e controle de culturas cultivadas.

As Redes de Sensores Sem Fio (RSSF) têm sido amplamente empregadas no monitoramento de ambientes (HUANG, LAN et al., 2007; XUEMEI, YUYAN et al., 2008; KHALEGHI et al., 2009). No caso de ambientes agrícolas, o seu uso proporciona uma melhor cobertura e controle da área monitorada, podendo os nós sensores serem espalhados pelo campo cobrindo grandes extensões. Uma vez que os sensores são distribuídos, o acesso físico a eles não é trivial, sendo necessário que essas redes tenham autonomia para se adaptarem a mudanças do ambiente e falhas de nós componentes da rede (MIRZA, 2008).

A grande extensão das propriedades agrícolas gera a necessidade de uma rede em larga escala composta por muitos nós sensores que monitoram o ambiente de forma contínua ou em intervalos de tempos pré-definidos, o que vem consequentemente resultar em um grande volume de dados coletados.

Em relação ao meio ambiente, o monitoramento da qualidade de corpos de água de rios, por exemplo, pode ser feita por meio de RSSF, que possibilita uma resposta mais rápida da avaliação da qualidade da água em comparação com a forma convencional, que requer coleta de amostras manuais e análise em laboratório. As redes de sensores possibilitam o monitoramento em tempo real, além da avaliação do impacto do manejo do solo na qualidade da água, como a contaminação por defensivos agrícolas entre outros. Segundo Barbosa (2008), que realizou experimentos sobre contaminação em solo e água sobre um determinado herbicida (atrazina), grande parte dos agrotóxicos utilizados em larga escala na agricultura interfere no meio ambiente de forma severa, fazendo-se necessária a tomada de medidas corretivas para a descontaminação.

No caso do monitoramento da qualidade da água, também pode-se ter a necessidade de muitos nós na rede, tornando o volume de dados elevado com a necessidade de leituras em intervalos de tempo menores, quando comparado às aplicações em ambientes agrícolas. A coleta por meio de redes de sensores permite um monitoramento em tempo real, proporcionando uma visão quase que instantânea das diversas variáveis que podem ser coletadas. A partir da fusão dos dados dos diferentes sensores de forma complementar ou cooperativa é possível obter informações mais representativas do ambiente monitorado quando comparado a leituras em um único ponto, ou analisando-se apenas um tipo de variável do ambiente.

Para auxiliar na avaliação dessa grande quantidade de dados, mecanismos de Fusão de Sensores (FS) podem ser usados. Segundo Hall e Llinas (1997), a fusão é a combinação de dados de vários sensores (e informações armazenadas em bancos de dados), para se obter uma melhor acuidade e inferências mais específicas do que poderiam ser alcançadas com a utilização de dados de um único sensor.

Tanto no monitoramento agrícola como no ambiental, mudanças no ambiente ocorrem de forma dinâmica durante o processo de coleta de dados, sendo adequado o mecanismo de fusão adaptar-se a essas mudanças em tempo real para que melhores resultados sejam alcançados.

A necessidade de identificar-se com rapidez e acurácia a contaminação de corpos de água superficiais com águas de deflúvios contaminadas por suspensão de solo, advindo por exemplo, de áreas de cultivo e áreas urbanas, torna-se importante

no monitoramento de processos ecológicos de grande relevância, como o assoreamento, perdas de solos, carregamento de pesticidas e degradação de habitats aquáticos.

O monitoramento ambiental da qualidade das águas em regiões afastadas de centros urbanos é prejudicado pelo fator logística. A necessidade de se percorrer grandes distâncias no projeto de monitoramento de uma única bacia hidrográfica tem limitado a geração de informações e a tomada de decisões a respeito da qualidade ambiental. Por exemplo, na bacia hidrográfica do Rio Miranda (Mato Grosso do Sul) a área da bacia corresponde a área de países como a Holanda, e para se avaliar 15 estações de amostragem com a aquisição de amostras em campo, gasta-se mais de 10 dias de trabalho de campo e no mínimo mais 30 dias de análises em laboratório. Depois disso, apenas após no mínimo 30 dias são realizadas outras campanhas.

Nesse contexto, o uso das Redes de Sensores Sem Fio (RSSF) podem auxiliar no monitoramento ambiental permitindo-se avaliar em tempo real mudanças ocorridas nesse ambiente. Uma RSSF é um tipo especial de rede *ad hoc* com capacidade de coletar e processar informações de maneira autônoma, estando esses sensores distribuídos em uma determinada área (AKYILDIZ et al., 2002; TUBAISHAT et al., 2003; GAJBHIYE et al., 2008). Elas apresentam grande potencial de emprego na agricultura e meio ambiente, devido à possibilidade de uma ampla área de cobertura de monitoramento (centenas ou até milhares de nós sensores), cada um com capacidade de coletar dados e transmiti-los a um nó coletor (denominado *gateway*), que por sua vez encaminha os dados para o usuário em um computador (GAJBHIYE et al., 2008). Em princípio, a proposta de usar uma RSSF não substitui o uso de protocolos de amostragem, mas geraria informações sobre o que ocorreu no sistema de monitoramento antes e após as campanhas de amostragem complementando muito as informações para os gestores ambientais.

A necessidade de muitos nós sensores e intervalos de tempos pequenos para a coleta dos dados, gera um grande volume a serem analisados. Além disso, os valores coletados devem ser analisados de forma integrada, pois o valor medido por um sensor individualmente não representará a característica do ambiente como um todo. É preciso ter uma visão integrada e global dos valores coletados pela RSSF para uma melhor exatidão da realidade estimada por esses dados (REN, 1995). Para isso, os valores dos sensores podem ser relacionados entre si para oferecerem

informações além dos dados puros coletados (SRIVASTA e BUCKMASTER, 2006). Para isso, existem técnicas conhecidas como Fusão de Sensores (FS) que permitem a combinação de duas ou mais informações diferentes advindas de sensores produzindo uma informação melhor do que se os dados dessas fontes fossem usados individualmente (HALL e LLINAS, 1997).

A ideia no monitoramento em tempo real é utilizar as RSSF para coletar os dados que serão avaliados pelo modelo de qualidade da água aqui proposto indicando aos gestores ambientais a sua classificação naquele determinado momento, podendo ser emitidos alertas para averiguação *in situ* quando necessário. A classificação proposta é composta por 7 classes, sendo elas: Excelente, Muito Boa, Boa, Pouco Boa, Pouco Ruim, Ruim, Muito Ruim, Péssima.

1.1 OBJETIVO

O objetivo deste trabalho foi a proposta de um modelo de classificação da qualidade da água contaminada por suspensão de solo em tempo real, utilizando-se para isso, RSSF e técnicas de FS para a análise dos dados coletados.

Além disso, procurou-se oferecer aos gestores ambientais uma ferramenta computacional que possa agilizar a análise da qualidade da água nesse contexto, quando comparada ao processo usual, permitindo a tomada de decisão diante de alterações na qualidade da água de forma rápida.

1.2 MOTIVAÇÃO

A possibilidade de contribuir para o monitoramento ambiental da qualidade da água contaminada por suspensão de solo foi a principal motivação deste trabalho, sendo que a proposição de um modelo computacional que avalie, em tempo real essa qualidade, diminuindo o tempo de resposta para tomada de uma ação preventiva em casos de alteração da qualidade, constitui-se em um aspecto inovador. A possibilidade de avaliar, remotamente, a estimativa de alteração da qualidade das águas em função da entrada de solos, permitirá o desenvolvimento de modelos de previsão e o monitoramento de processos ecológicos de grande relevância tais como o assoreamento, perdas de solos, carreamento de pesticidas e degradação de habitats aquáticos.

A ideia de usar técnicas de FS surgiu a partir de pesquisas na área de RSSF e aplicações em monitoramento ambiental. As técnicas de fusão se destacaram como uma forma de analisar o grande volume de dados coletado pela rede de sensores em tempo real. A proposta de monitoramento ambiental visa usar RSSF para a coleta instantânea dos dados que caracterizam a qualidade da água, sendo esses dados então analisados por técnicas de fusão para classificá-las.

A qualidade da água de leitos de rios que pode ser contaminado pelo mau uso do solo em lavouras. Com o uso desse modelo, pretende-se gerar novos índices de qualidade da água usando-se um número menor de variáveis (coletadas por meio de sensores de baixo custo), comparado com um índice usual, como o Índice de Qualidade de Água (IQA) desenvolvido pela United State Environment Protection Agency (USEPA) e é adotado no Brasil pela Companhia Ambiental do Estado de São Paulo (CETESB) e outros (FERREIRA e ALMEIDA, 2005; HADDAD, 2007). O intuito é obter um índice que represente a qualidade da água para sua aplicação na agricultura com foco na contaminação por solo, uma vez que a qualidade da água para consumo humano deve seguir normas padronizadas sendo necessário a avaliação de um conjunto específico de variáveis para garantir a qualidade desejável.

O IQA é calculado utilizando-se nove tipos de variáveis: temperatura da água, pH, oxigênio dissolvido, demanda bioquímica de oxigênio, coliformes termotolerantes, nitrogênio total, fósforo total, resíduo total e turbidez. Destes, apenas temperatura da água, pH, oxigênio dissolvido e turbidez são mensuráveis por sensores comercialmente disponíveis. Para as demais amostras d'água precisam ser coletadas manualmente para análise em laboratório. Existem outros índices que são calculados com um número menor de variáveis, que não são medidas por sensores, exigindo análise laboratorial, o que impossibilita o monitoramento em tempo real (LERMONTOV, YOKOYAMA et al., 2008). A coleta de dados por meio de uma RSSF possibilitará o monitoramento em tempo real, diferente da forma como usualmente é feita, por meio de coletas manuais e análise em laboratório, o que torna o processo demorado.

Os danos ambientais gerados pela produção agrícola tanto em água quanto no solo necessitam de ferramentas de monitoramento contínuo para gerar tomadas de decisão mais efetivas, justificando a importância de desenvolver ferramentas para avaliação da qualidade ambiental com base em RSSF.

1.3 JUSTIFICATIVA

O elevado volume de dados coletados pelas RSSF no monitoramento de ambientes deve ser analisado para que informações relevantes sobre o objeto em observação possam ser identificadas e mudanças no ambiente em tempo real possam ser imediatamente percebidas pelo sistema de monitoramento.

É necessário transformar o volume de dados coletados em informações que sejam de interesse, facilitando a tomada de decisão, o planejamento e a elaboração de soluções adequadas. Para isso, técnicas de FS podem ser usados em sistemas de tomada de decisão, nas quais a fusão oferece informações exatas e representativas do ambiente monitorado, já que em muitos casos, observar os dados individualmente é pouco representativo (HALL e LLINAS, 1997) (NAKAMURA, LOUREIRO et al., 2007).

Sabendo-se da importância do monitoramento da qualidade da água em tempo real¹, este trabalho propõem um modelo que permita a identificação da qualidade da água contaminada por suspensão de solo em tempo real. A abordagem desse modelo usa FS para a criação de uma classificação da qualidade da água voltado para aplicações na agricultura e meio ambiente, em específico essa classificação indica a contaminação por solo nas águas de rios. Para a criação dessa classificação, usou-se técnicas da FS que abrange conceitos da Inteligência Artificial (IA) em específico a Aprendizado de Máquina (AM) que estão entre as principais técnicas de fusão utilizadas, dentre as Probabilística, Estatística, Teoria da Evidência de Dempster-Shafer entre outras (KAFTANDJIAN, 2005) (SANTOS, 2007). Dentro da AM, a técnica usada foi a Indução por Árvore de Decisão. Nakamura (2009) comenta que mesmo que não tenha sido formalmente usado em RSSFs, a Indução tem um grande potencial para diversas aplicações, tais como: diagnóstico de falhas, detecção de eventos e explicação e avaliação de fenômenos ambientais.

O uso de técnicas de AM permitiram a criação do modelo de classificação, pois a partir de dados de um experimento controlado feito em laboratório, foi possível criar a classificação e validá-la por meio de algoritmos de agrupamento e

¹ O tempo real refere-se a ter uma resposta de forma instantânea que represente um determinado período de coleta.

classificação de dados. Os algoritmos de agrupamento foram importantes para definir o número de classes ideal dentro desse contexto e os algoritmos de classificação serviram para validar a coerência das classes e definir quais deles seriam melhores empregados na criação do modelo.

Qin et al. (2012) descreve um sistema de monitoramento da qualidade de águas residuais usando FS e técnicas de AM foi testado em campo, com resultados satisfatórios, demonstrando grande potencial desta técnica para o monitoramento on-line de qualidade da água. Outros trabalhos, como em Motamarri e Boccelli (2012), Olikar e Ostfeld (2014) propõem o monitoramento da qualidade da água em outros contextos de contaminação e usando outras técnicas de AM.

1.4 ORGANIZAÇÃO DA TESE

Este trabalho está organizado em 5 capítulos. O presente capítulo contém introdução, motivações e justificativas.

O Capítulo 2 apresenta a revisão de literatura necessária para o desenvolvimento desse trabalho, sendo abordados a RSSF, FS, AM. O Capítulo 3 apresenta como o trabalho foi desenvolvido expondo os materiais e métodos usados. O Capítulo 4 apresenta discussões e resultados. O Capítulo 5 apresenta as considerações finais, assim como conclusões, propostas de continuidade da pesquisa e perspectivas de trabalhos relacionados.

2. REVISÃO DE LITERATURA

A revisão de literatura aqui apresentada tem em vista proporcionar uma base teórica para fundamentar a pesquisa. Em relação às FS, buscou-se relacionar literaturas básicas para um melhor conhecimento sobre o tópico e também literaturas relacionadas a técnicas aplicados no monitoramento ambiental e seus modelos. O mesmo se aplicou a AM.

2.1 FUSÃO DE SENSORES

O termo FS é geralmente usado para especificar que os sensores são responsáveis por fornecer os dados que serão combinados resultando em uma informação mais significativa do que se os dados dessas fontes fossem usados individualmente. Existem também os termos Fusão de Dados (FD) e Fusão de Informação (FI) que são aceitos como termos gerais, apesar das questões filosóficas sobre a diferença entre os dados e informação como na Figura 1 (NAKAMURA, LOUREIRO et al., 2007). Outros autores Hall e Llinas (1997), definem a FD como sendo aquela que acontece nos dados brutos coletados e a FI aquela que acontece com dados que já passaram por algum pré-processamento.

Outros trabalhos da área, como em MSB (2010) e DAUM (2001) consideram em um nível mais abstrato da fusão a FI, frequentemente usada no contexto da Inteligência Artificial. FS mesclam dados de diferentes fontes de sensores, sendo FD a fusão que acontece próximo aos sensores (nos nós sensores), geralmente nos dados puros. Neste trabalho, é usado o termo FS relacionando a fusão sobre dados coletados por meio de Redes de Sensores Sem Fio (RSSF). Quando for necessário, será usado o termo Fusão de Dados/Informação para operações de fusão que ocorrem com dados provindos tanto de sensores e/ou banco de dados.

Figura 1 – Relação entre os termos Fusão de Dados, Fusão de Informação e Fusão de Sensores.

Fonte: Adaptada de Nakamura (2007).

Segundo Luo e Kay (1990), o principal objetivo das técnicas de fusão é avaliar e analisar os dados, fornecendo aos sistemas decisórios informações confiáveis, tentando reduzir taxas de erro nesses sistemas.

Algumas limitações e erros no processo de fusão são citados em LUO e KAY (1995) e Salustiano (2008):

- Alguns algoritmos de fusão necessitam de certa quantidade de sensores para que a fusão seja realizada. Essa limitação não se aplica as RSSF, pois elas se caracterizam por serem constituídas, algumas vezes, por dezenas a centenas de nós sensores.
- São mencionados erros relacionados ao: processo de fusão (determinar a correspondência e associação entre os dados); informação do sensor (ruído aleatório); operação do sistema (calibração dos sensores).

Os sistemas de FD podem ser classificados em 3 tipos básicos de configuração (Figura 2) considerando a relação existente entre as fontes de dados (NAVABZADEH RAZAVI, 2010) (SALUSTIANO, 2008):

- Competitivo: também chamada de redundante, é uma classe de aplicações de fusão na qual os dados de uma mesma grandeza serão combinados para aumentar a confiabilidade, precisão e diminuir os conflitos entre diferentes leituras. O objetivo desta configuração é reduzir os efeitos do ruído e medições errada. Pode ser de duas formas: fusão das medidas de diferentes sensores ou fusão das medidas de um mesmo sensor obtidas em instantes diferentes.
- Complementar: os dados de diferentes sensores (independentes entre si) são combinados para fornecer uma informação mais abrangente do fenômeno em observação.
- Cooperativo: combina observações de diferentes sensores dependentes entre si para derivar uma informação que não poderia ser obtida por meio de um único sensor.

No modelo aqui proposto, a configuração complementar é a utilizada. Outras classificações são apresentadas por Dasarathy (1997) e por Nakamura (2007).

Figura 2. Fusão complementar, competitiva e cooperativa.

Fonte: Adaptada de Elmenreich e Pitzek (2001).

Em Nakamura (2007), são apresentados diversas técnicas de fusão classificadas de acordo com o seu propósito. Técnicas da Inteligência Artificial (IA) estão entre as principais técnicas de fusão utilizadas, como a probabilística, estatística, Teoria da Evidência de *Dempster-Shafer* entre outras (KAFTANDJIAN, 2005) (SANTOS, 2007). Dentre as técnicas que envolvem a IA, podemos citar as que utilizam Redes Neurais, Lógica *Fuzzy* e Aprendizado de Máquina. Dentro da AM, foi escolhida a técnica de Indução por Árvore de Decisão (IAD). As Árvores de Decisão foram usadas para averiguar a classificação propostas e construir o modelo que melhor a representasse.

2.1.1 MODELO DE FUSÃO DE SENSORES

O modelo de FS depende das características individuais dos sensores, forma de integração das informações e exigências da aplicação. Alguns modelos existentes para fusão são expostos por Salustiano (2008) e Raol (2009): Modelo de Fusão *Joint Directors of Laboratories* (JDL), Modelo de Fusão em Cascata Modificado, Modelo de Fusão *Boyd* (ou *Observer, Orientate, Decide and Act* - OODA), Modelo de Fusão *Omnibus*, Modelo de Fusão em Níveis. Os modelos definem em sua estrutura os processos de fusão de acordo com a natureza do dado

a ser combinado, ou a finalidade da aplicação, classificando os tipos de fusão que podem ser empregados. A seguir são apresentados alguns desses modelos.

2.1.1.1 Modelo *Joint Directors of Laboratories* (JDL)

Um dos modelos de FD mais populares é proveniente do Departamento de Defesa dos EUA conhecido como *Joint Directors of Laboratories Fusion Model* (JDL) (RAOL, 2009). O modelo JDL tem cinco níveis de processamento de dados e uma base de dados. Esses níveis são interconectados por um barramento em comum e não necessariamente o processamento deve ser em ordem sequencial, podem ser executados concorrentemente. Existem três níveis principais (1, 2 e 3) como mostra a Figura 3. Há vários subníveis e níveis auxiliares para o processamento de dados, combinação da informação disponível, além de avaliação do desempenho do sistema de Fusão de Dados.

Este modelo é orientado a funcionalidade, e tem a finalidade de ser o mais genérico possível e útil para diferentes áreas de aplicação. A saída deste modelo (a identificação e caracterização de uma entidade individual ou objetos) é para ser a menos ambígua possível. Espera-se também, facilitar a interpretação em alto nível de certas entidades dentro do contexto do ambiente de aplicação. O processo de FD do JDL é um modelo conceitual que identifica os processos, funções, categorias de técnicas e técnicas específicas aplicáveis na FD. Esse processo é definido e conceituado por fontes de informação, interações Homem-Computador, pré-processamento dos dados puros (na fonte), níveis de FD e sistema de gerenciamento de dados, incluindo avaliação de desempenho (NAKAMURA, LOUREIRO et al., 2007; RAOL, 2009), conforme pode ser visto na Figura 3.

Informações advindas de uma variedade de fontes, tais como sensores, informações a priori em banco de dados, são normalmente requisitadas e coletadas para a fusão:

- **Nível 0:** Dados puros são pré-processados. Informações como tempo, local, tipo e assinaturas dos dados coletados são determinados nesse nível. Os dados podem ser agrupados ou classificados e algumas assinaturas podem ser adicionadas com o tipo de dado. Prioridade de processamento pode ser definida nesse nível para que no nível 1 os dados sigam essa ordem.

- **Nível 1 "Refinamento do objeto":** As tarefas realizadas nesse nível são: uniformidade dos dados (transformação dos dados em uma referência consistente - unidades), associação de dados (usando algum mecanismo de correlação), acompanhamento de situações atuais e futuras de objetos e identificação utilizando métodos de classificação. Este nível consiste de procedimentos numéricos, tais como procedimentos de estimativa, rastreamento de alvos e reconhecimento de padrões. "Refinamento do objeto" auxilia na avaliação do objeto por meio da combinação da localização, informações de identidade para atingir representação refinada dos objetos individualmente (tais como emissores, plataformas e armas) quanto ao seu tipo, identidade, posição, aceleração, velocidade, e assim por diante. Nível 1 executa as seguintes quatro funções: (1) transforma os dados em um conjunto consistente de unidades e coordenadas, (2) refina e estende-se a um tempo futuro na previsão de posição de um objeto, cinemática, dinâmica ou atributos; (3) associa dados aos objetos para permitir a aplicação de técnicas estatísticas, e (4) refina a estimativa da identidade de um objeto ou da sua classificação. A fusão no nível 1 pode ser categoricamente dividida em duas partes: (1) fusão cinemática, que envolve a fusão de informações de localização para determinar posição, velocidade e aceleração de objetos móveis, tais como mísseis, aviões, navios, pessoas e veículos terrestres e (2) a fusão de identidade, que envolve a fusão de dados parametrizados para determinar a identidade de um objeto observado, por exemplo, decidir se um objeto que está em movimento é um míssil ou um avião. Estimativa de identidade pode ser complementada por sistemas especialistas baseados em regras, nos quais vários tipos de informações (fatos, procedimentos) podem ser explorados para auxiliar a estimativa de identidade.
- **Nível 2 "Refinamento da situação":** Neste nível, é feita uma tentativa para encontrar uma descrição contextual da relação entre os objetos e os eventos observados. Após o processamento dos dados no nível 1, a situação é determinada e uma nova análise é realizada para refinar a situação, se necessário. O objetivo principal é obter uma imagem global dos objetos monitorados para fins como a aplicação de defesa (ataques). Este é um processo muito complexo.
- **Nível 3 "Refinamento da Ameaça/Aviso/Perigo":** Com base no conhecimento a priori e previsões sobre a situação futura, as inferências sobre as

vulnerabilidades e oportunidades de operação são estabelecidas. Durante a avaliação da ameaça, vários aspectos são considerados, tais como (1) estimativa do risco, (2) a indicação de advertência, e (3) alvo. O objetivo final é obter uma avaliação refinada da ameaça e seu contexto, na qual importantes decisões e ações podem ser baseadas.

- **Outros níveis:** Refinamento do processo, muitas vezes adicionado como nível 4, é um processo que monitora o desempenho do sistema, por exemplo, restrições de tempo real, e reorganiza os sensores e fontes para alcançar metas específicas ou objetivos da missão. Neste nível, não se tem a preocupação com os dados (processamento). Contudo, a gestão do sensor é um aspecto adequado para o estudo e emprego neste nível para o melhor uso do conjunto de sensores. Alguns teóricos e profissionais de FD também incluem um nível de refinamento cognitivo, o nível 5, que está entre o nível 3 e do HCI, introduzindo assim o conceito de IA, nesta fase, de forma limitada.

Figura 3 - Modelo JDL.

Fonte: Adaptada de Raol (2009).

O modelo JDL é suportado por um sistema de gerenciamento de banco de dados, que monitora, avalia, atualiza e fornece informações para o processo de fusão devido a grande quantidade de dados que está envolvida. A Interação Homem-Computador (IHC) é muito importante, permite ação humana, tais como comandos, consultas, avaliações humana de inferências e relatórios de operadores humanos. IHC é o mecanismo pelo qual um sistema de fusão comunica os

resultados através de alertas, displays, atualizações dinâmicas de ambas as informações de posicionamento e identidade em displays geográficos. Este modelo é indicado para Sistemas Multissensores usados para aplicações de defesa.

2.1.1.2 Modelo de Fusão Cascata Modificado (*Modified Waterfall Fusion Model*)

O Modelo de Fusão em Cascata Modificado enfatiza as funções de processamento dos níveis mais baixos, como mostrado na Figura 4. As etapas de processamento do modelo em cascata têm algumas semelhanças com o modelo JDL: (1) Detecção e processamento de sinais relativos ao pré-processamento na fonte (nível 0 do modelo JDL), (2) extração de características e processamento de padrões correspondentes ao refinamento do objeto (nível 1 do modelo JDL), (3) avaliação da situação (refinamento da situação) (nível 2 do modelo JDL) e (4) tomada de decisões relativa ao refinamento da ameaça (nível 3 do modelo JDL).

Figura 2 - Modelo em Cascata Modificado

Fonte: traduzida de Raol (2009).

Não há *feedback* do fluxo de dados no modelo em cascata original, porque ele é um modelo acíclicos. Na Figura 4, um modelo modificado de *Waterfall Fusion*

Model (WFFM) é proposto e apresentado, no qual: (1) a complexidade crescente da sub tarefa é indicado pela numeração crescente (a partir dos dados dos sensores, a extração de características e processamento de padrões até a tomada de decisão e ações de controle), e (2) há alguns ciclos de *feedback* (RAOL, 2009).

A mudança está no uso de *loops* de *feedback* local, como se segue: (1) A partir da tomada de decisões e ações de controle para o bloco de avaliação da situação (*situation assessment*), para refletir o refinamento da situação e seu uso como uma nova ação de controle, (2) do processamento de padrão para processamento de sinais, para refletir o reconhecimento de padrões melhorados e seu uso em refinamento da avaliação da situação (*situation assessment*), e (3) a partir das tomadas de decisão e ações de controle para o bloco de extração de características, de modo a refletir a melhor tomada de decisões e novo controle de ação, baseada na melhora das características extraídas.

2.1.1.3 Modelo Baseado no Ciclo da Inteligência (*Intelligence Cycle–Based Model*)

Como o processo de FD tem um comportamento cíclico de transformação inerente, o modelo baseado no Ciclo da Inteligência (CI) tenta capturar essas características cíclicas composta das seguintes 5 fases (RAOL, 2009):

1. Na fase de planejamento e direção, os requisitos de informação são determinados.
2. Na fase de coleta, a informação adequada é coletada.
3. Na fase de agrupamento, a informação recolhida é organizada.
4. Na fase de avaliação, a informação disponível é usada e a fusão real é realizada.
5. Na fase de divulgação, a inteligência fundida e inferências são distribuídas.

Este modelo é um modelo de fusão de dados a nível macro de processamento e parece mais um modelo em nível superior quando comparado ao modelo WFFM modificado.

As ações nos subníveis e tarefas de processamento não são definidas ou indicada neste modelo, embora essas tarefas podem ser implicitamente presumidas. Seria conveniente considerar este modelo como um macro modelo (mais abstrato).

2.1.1.4 Modelo *Boyd*

O modelo de controle cíclico *Boyd* (ou modelo cíclico Observar, Orientar, Decidir e Agir - OODA, modelo BCL), mostrado na Figura 5, representa o mecanismo de apoio clássico a decisão em operações militares e tem sido amplamente utilizada para FS e tem um ciclo de retroalimentação. Ele usa o ciclo OODA, descrito a seguir: (1) a fase de observação é, sobretudo, comparável ao passo pré-processamento (da origem) do modelo JDL e como parte da fase de coleta do modelo de IC; (2) a fase de orientação contém, talvez implicitamente, as funções de níveis 1, 2 e 3 do modelo JDL, enquanto que no modelo de IC, essa etapa corresponde aos elementos estruturados de coleta e a fase de agrupamento; (3) a fase de decisão se compara com o nível 4 e processo de refinamento do modelo JDL e ao processo de divulgação do modelo de IC; e (4) a fase de ação é comparável a uma fase de disseminação no modelo de IC.

O modelo BCL é mais comparável ao modelo WFFM modificado, e é abstrato, especificando apenas um loop OODA, sendo mais abstrato do que o modelo do IC. Uma combinação adequada dos modelos de IC e BCL é um modelo melhor para o processo de fusão, tornando o modelo da nova combinação menos abstrata. Tal combinação resultou no modelo *Omnibus*.

Figura 3 - Modelo BCL.

Fonte: traduzida de Raol (2009).

2.1.1.5 Modelo *Omnibus*

O modelo *Omnibus* (OB) integra a maioria das características benéficas de outras abordagens, como mostrado na Figura 6. O modelo OB define a ordem dos processos envolvidos para tornar a natureza cíclica mais explícita e, além disso, ele usa uma terminologia geral. A sua natureza cíclica é comparada ao modelo BCL. Este modelo fornece níveis razoavelmente detalhados de processamento em comparação com o modelo BCL.

Os vários níveis são os seguintes: (1) Os passos de detecção e processamento de sinais são conduzidos pela fase de observação do modelo BCL; (2) a extração de características e processamento de padrões compreende a fase de orientação, (3) o processamento do contexto e tomada de decisões estão incluídas na fase de decisão; e (4) o controle de recursos e tarefas (tarefas de controle e recursos) é realizado pela fase de ação. A FS é a rota a partir da fase de observação para a fase de orientação. O modelo de fusão do OB é mais completo que o WFFM, IC, e BCL porque engloba muitas características importantes e os aspectos funcionais desses modelos.

Figura 4 - Modelo Omnibus.

Fonte: Adaptada de (RAOL, 2009).

2.1.1.6 Modelo de Fusão em Níveis

Destaca-se o Modelo de Fusão em Níveis, como sendo o único dentre os citados, que fazem uma distinção entre os dados a serem submetidos à fusão (Quadro 1). Neste modelo, a fusão pode acontecer em 4 níveis diferentes (LUO e KAY, 1990; SALUSTIANO, 2008):

1. Nível de sinal (*signal-level*): refere-se à combinação dos sinais provenientes de um grupo de sensores de mesma natureza (temperatura ou pressão ou altitude, coordenadas geográficas ou luminosidade) ou de naturezas diferentes, havendo uma relação de dependência entre as variáveis (posição e velocidade, por exemplo);
2. Nível de pixel (*pixel-level*): é utilizada para melhorar a informação associada a cada pixel de uma imagem resultante da associação de múltiplas imagens. O aumento da qualidade no resultado da fusão no nível de pixel está diretamente relacionado ao ganho resultante da aplicação de técnicas de processamento de imagens captadas pelos sensores. No contexto deste trabalho, esse nível de fusão não se aplica, sendo possível, no nível de característica, o uso de imagens geográficas obtidas por Sistemas de Informação Geográfica (SIG) combinados com dados coletados por sensores;
3. Nível de característica (*feature-level*): seu objetivo é potencializar as informações obtidas a partir dos dados coletados por um sensor pela análise dos dados de sua vizinhança ou fronteiras. Através desse tipo de análise é possível não somente reconhecer padrões nos dados dos sensores, mas também eliminar possíveis interpretações duvidosas efetuando fusões competitivas nos dados dos sensores. Esse nível de fusão é aplicável ao monitoramento de ambientes agrícolas, pois a observação do ambiente deve refletir a realidade da área como um todo, integrando os dados coletados por nós sensores individuais em grupos para se identificar as características daquela região (como no caso de micro climas) ou delimitar a detecção de fenômenos (detecção de pragas e doenças, por exemplo). Quando vários sensores relatam características semelhantes de um mesmo local do ambiente monitorado, a probabilidade de que essas características estejam realmente presente aumenta, podendo ser melhorada a precisão desses dados;

4. Nível de símbolo (*symbol-level*): permite que a informação dos múltiplos sensores seja realmente usada de forma conjunta no nível mais alto de abstração dentro do sistema, também chamada de nível de decisão. O nível de símbolo é o único nível em que se pode realizar fusão de informações provenientes de sensores de naturezas diferentes que não tenham nenhuma relação direta entre si ou fusão de regiões diferentes do ambiente. Esse nível de fusão pode ser aplicado ao contexto da AP proporcionando um conjunto de fatos precisos sobre a lavoura, por exemplo, que ajude no ajuste das medidas a serem tomadas para o controle de qualidade e produtividade, aplicações de defensivos entre outros.

Quadro 5 - Modelo de Fusão em Níveis.

Nível	Sinal	Pixel	Característica	Símbolo
Características				
Tipo de informação sensoriada	Sinais uni- ou multidimensionais	Imagens	Características extraídas de sinais e imagens	Símbolos que possuem características capazes de representar decisões
Grau de complexidade na representação da informação	Baixo	Baixo a médio	Médio	Alto
Mecanismos de Fusão	Média Aritmética Média Ponderada Filtro de Kalman Sensores em Consenso	Filtros Lógicos Morfologia Matemática Álgebra de Imagens Recozimento Simulado	Transformações Geométricas Segmentação Inteligência Artificial	Estimativa Bayesiana Teoria dos Conjuntos Lógica Booleana Lógica Fuzzy
Melhoria obtida pela fusão	Redução na variância Esperada	Melhoria no desempenho do processamento da Imagem	Redução no processamento e aumento da precisão nas medidas	Aumento da veracidade ou da probabilidade dos valores

Fonte: Adaptada de Luo e Kay (1990) e Salustiano (2008).

Dependendo da necessidade de cada sistema e do grau de similaridade dos sensores utilizados, um desses níveis é utilizado. Em sua maioria, os outros modelos estão mais preocupados com a organização e controle do fluxo dos dados no sistema do que com o bloco responsável pela fusão dos dados dos sensores (SALUSTIANO, 2008).

2.1.2 ALGUMAS APLICAÇÕES

A seguir são apresentadas alguns trabalho que utilizam a fusão na área de monitoramento agrícola e ambiental. Buscou-se, a seleção de trabalhos que contribuíssem em específico na definição das características que devem ser consideradas no contexto desse trabalho. Algumas referências de outros trabalhos podem ser vistos em (DOTA, 2010).

Em geral, a Agricultura de Precisão (AP) faz uso de sensoriamento remoto para identificar as condições do solo e da cultura combinando com informações detalhadas coletadas por sensores em equipamentos agrícola. Com isso, os agricultores podem planejar a variação espacial de nutrientes a serem aplicados para corrigir deficiências do solo (MACRUAIRI, KEANE et al., 2008).

O controle de pragas na qual o tempo de aplicação de pesticidas é fundamental para o combate eficiente das pragas, apresenta-se como um excelente exemplo do potencial das RSSF. As redes podem ser usadas para coletar dados em intervalos de tempo pré-definidos (minuto a minuto, hora a hora) facilitando o monitoramento em tempo real, ajudando a identificar a localização exata (unidade em metros ou centímetros) da área infestada tornando possível a aplicação de defensivos de forma precisa.

Na AP, a análise e processamento de uma grande quantidade de dados são necessários para medição e instrumentação da lavoura. A fim de aumentar a precisão da análise dos dados e da qualidade do processamento, mecanismos de fusão são cada vez mais utilizados nessa área (HUANG, LAN et al., 2007).

As aplicações encontradas em literaturas da área apresentam a aplicação de mecanismos de fusão em equipamentos agrícolas (instrumentação) sendo mais raro encontrar aplicações de FS com coleta de dados via RSSF. Geralmente os sistemas de fusão são implementados para operar em máquinas agrícolas para coleta de dados no momento em que o equipamento entra na lavoura para realizar alguma

tarefa. São sistemas que utilizam receptor GPS para localização espacial precisa dos dados coletados.

Em El Hajj et al. (2008) é descrito um experimento de FI que integra uma série de imagens via satélite com os dados de uma base de conhecimento para detecção de colheita da cana-de-açúcar. No trabalho, é empregado um mecanismo de inferência baseado em Lógica *Fuzzy* (EL HAJJ, BEGUE et al., 2008). Outra pesquisa, voltada para sistemas de navegação autônoma de veículos agrícolas (GARCIA-PEREZ, RIBEIRO et al., 2008), utiliza algoritmo de fusão (baseado em regras) para sobrepujar a falta de sinal GPS. Nesse sistema, o mecanismo de fusão é ativado quando não é possível estimar a posição via GPS.

Já Sun et al. (2007) desenvolveram um sistema de FS que por meio dos dados coletados pelos sensores, possibilita estimar a contenção de água no solo, resistência mecânica e condutividade elétrica em tempo real de monitoramento. Os autores afirmam que esse sistema, aliado a dados de localização (GPS), pode fornecer informações ainda mais valiosas sobre as propriedades do solo para a AP (SUN, ZENG et al., 2007).

Alguns exemplos da aplicação de mecanismos de FS juntamente com RSSF são apresentados na área de irrigação de precisão, como o experimento descrito por Xiong e Wang (2009) (Figura 7). Os autores apresentam no seu experimento a aplicação de algumas tecnologias em RSSF para a irrigação de precisão baseada no monitoramento de sinais acústicos do stress hídrico da lavoura (XIONG, WANG et al., 2009). Um mecanismo de FS é explorado para melhorar a precisão dos dados e garantir a tomada de decisões confiáveis.

Figura 5. Arquitetura da RSSF do experimento realizado por Xiong e Wang (2009).

Fonte: Adaptada de Xiong, Wang et al. (2009).

Sachez e Marchant (2000) realizaram um experimento cujo objetivo principal foi a construção de mapas 3D para ajudar a distinguir plantas daninhas entre culturas. Fusão de dados foi utilizada para reduzir as incertezas das estimativas de profundidade dos mapas 3D.

A compactação do solo torna as condições de crescimento das plantas ao longo do campo bastante variável. RSSF podem ser aplicadas para ajudar na obtenção de mapas georeferenciados de determinadas propriedades de estado do solo (por exemplo, resistência do solo, teor de água e permeabilidade ao ar). Hemmat e Adamchuk (2008) abordam sobre a diversidade de sensores disponíveis para medir a resistência do solo (usados nas RSSF) e enfatiza o uso de FS como uma opção capaz de melhorar a aplicabilidade futura desses sensores de compactação do solo.

Tanto sensores acústicos e pneumáticos servem como alternativa para sensores mecânicos quando se estuda a interação do solo com implementos agrícola (ADAMCHUK, HUMMEL et al., 2004). Neste caso, a relação entre a saída do sensor e o estado físico do solo é mal interpretada e consequentemente é necessária uma pesquisa adicional. Devido a princípios diferentes de medição, sensores acústicos e pneumáticos podem ser fortes candidatos para FS em que múltiplos fluxos de dados são mesclados para melhorar a previsão das características do solo.

Integrar os conceitos de medida diferentes em uma única unidade de mapeamento é um tema atual de pesquisa. Acredita-se que várias fontes de dados fundidas podem prever melhor as características agronômicas do solo e dar apoio a tomada de decisão no gerenciamento da lavoura. Um exemplo de fusão de dados em tempo real para sensores de solo é a fusão de um sensor de condutividade elétrica com um sensor ótico para determinar a diferença em matéria orgânica, textura do solo e teor de umidade em diversas condições de cultivo. Outra possibilidade é o uso de um mapa de condutividade elétrica junto com medições de pH do solo para determinar a taxa variável de aplicação de calcário (ADAMCHUK, HUMMEL et al., 2004).

2.1.3 CONSIDERAÇÕES FINAIS

Em relação a fusão, a revisão focalizou nas classificações das técnicas usadas e modelos de fusão relevantes na literatura específica. A partir desse levantamento, foi possível definir que o modelo proposto classifica-se, em relação a fonte de dados, como Complementar. Em Dota (2011) e Santos (2010), os autores apresentam conceitos importantes relacionados ao uso das RSSF juntamente com técnicas de FS para compor um modelo de fusão.

Khalifa (2009) menciona que a FS tem sido aplicada com sucesso em várias aplicações, como rastreamento de alvos, nas forças armadas e robótica (na concepção de robôs autônomos). É sua convicção de que sucesso semelhante pode ser alcançado melhorando-se a modelagem de bacia hidrográfica quando várias fontes de dados sensoriais são usadas.

Considerando o esquema de descrição do modelo proposto por este trabalho (Figura 8) e visto que o modelo de fusão está diretamente relacionado as exigências da aplicação, a opção foi pelo Modelo CI, pois seus níveis são semelhantes ao processo que envolve o monitoramento ambiental na avaliação da qualidade da água para sua classificação.

Figura 8 – Esquema de descrição do modelo de classificação da qualidade da água contaminada por suspensão de solo.

Fonte: Autor.

A seguir é apresentada a revisão de literatura referente a técnica de fusão usada no modelo proposto, Indução por Árvore de Decisão. Antes é feita uma apresentação de conceitos básicos da área de Aprendizado de Máquina, na qual a IAD se qualifica.

2.2. APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (*Machine Learning*) é uma área da IA que tem como objetivo o desenvolvimento de técnicas computacionais sobre o aprendizado, assim como a construção de sistemas capazes de adquirir conhecimento de forma automática [MONARD, 2003a]. A ideia por trás do AM é a construção de sistemas computacionais que possam tomar decisões baseadas em experiências anteriores, adquiridas por meio de soluções tomadas anteriormente com sucesso.

O AM envolve técnicas supervisionadas e não-supervisionadas. No aprendizado não-supervisionado, são analisados exemplos e tenta-se determinar se alguns deles podem ser agrupados de alguma maneira formando os chamados *clusters* ou agrupamentos. No aprendizado supervisionado, usa-se um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido, sendo cada exemplo descrito por um conjunto de atributos, assim é possível construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados (MONARD, 2003a).

No caso da qualidade da água, algoritmos de aprendizagem de máquina não-supervisionados foram usados num primeiro momento para poder-se identificar o número de classes ideal. Em seguida, essas classes foram submetidas a algoritmos supervisionados para averiguar a coerência da divisão das classes e a partir do desempenho desses algoritmos em classificar positivamente ou negativamente os exemplos de cada classe, optou-se pelos melhores resultados para a construção do modelo de classificação de entradas desconhecidas.

2.2.1 APRENDIZAGEM NÃO-SUPERVISIONADA

Aprendizagem não supervisionada (Agrupamento ou *Clustering*) busca extrair informação relevante de dados não rotulados. Uma solução mais geral consiste em definir medidas de similaridade entre dois *clusters* assim como um critério global como a soma do erro quadrático. A característica principal do Agrupamento é a não

existência de classes predefinidas, sendo os exemplos agrupados em função de sua similaridade básica.

Existem diversos algoritmos de agrupamento, cada um com suas vantagens e desvantagens. De acordo com Steinbach et al. (2000), clusterização hierárquica é retratada como a técnica de clusterização de melhor qualidade, sendo limitada pela sua complexidade quadrática, já o K-means e suas variações possuem complexidade de tempo linear, mas produzem clusters de qualidade inferior.

O K-Means é uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de k centros dado por $\chi = \{x_1, x_2, \dots, x_k\}$ de forma iterativa (LINDEN, 2009). A distância entre um ponto p_i e um conjunto de clusters, dada por $d(p_i, \chi)$, é definida como sendo a distância do ponto ao centro mais próximo dele. A função a ser minimizada então, é dada por:

$$d(P, \chi) = \frac{1}{n} \sum_{i=1}^n d(p_i, \chi)^2$$

Deve-se definir o número de classes a ser considerada no agrupamento (valor de k), o que acaba sendo um problema, pois em alguns casos não se sabe *a priori* quantas classes seria a ideal. O algoritmo funciona da seguinte forma (LINDEN, 2009):

1. Escolher k distintos valores para centros dos grupos (possivelmente, de forma aleatória).
2. Associar cada ponto ao centro mais próximo.
3. Recalcular o centro de cada grupo.
4. Repetir os passos 2-3 até nenhum elemento mudar de grupo.

O algoritmos k-means++ é idêntico ao algoritmo k-means, com diferença na seleção das condições iniciais usando uma técnica aleatória de escolha de sementes de inicialização.

A Figura 9 mostra aplicação da técnica de agrupamento para detecção de valores fora de padrões (*outliers*).

Figura 9 – Técnica de agrupamento na detecção de *outliers*.

Fonte: Extraído de Han e Kamber (2001).

2.2.2 *Aprendizagem Supervisionada*

No aprendizado supervisionado, por meio de exemplos conhecidos é possível induzir conceitos. Assim, a partir de exemplos classificados em classes conhecidas treina-se um modelo para exemplos ainda não classificados (BIGUS, 1996), sendo usado para treinamento de Redes Neurais ou Árvore de Decisão na obtenção de classificação, funções de aproximação ou modelagem e previsões baseadas no tempo. No contexto desse trabalho são usados técnicas de classificação baseadas na IAD, citando Nakamura (2007) que relata a inferência por indução usando técnicas da IA como de grande potencial para diferentes aplicações, como diagnóstico de falhas, detecção de eventos, e avaliação de fenômenos ambientais.

2.2.2.1 *Árvore de Decisão*

É um tipo de classificador que é treinado por uma seleção iterativa de características individuais que são os mais relevantes em cada nó na árvore (TARCA, 2007). Tarca (2007) explica que o espaço de entrada X é dividido em subconjuntos repetidamente descendentes, começando com o próprio X . Existem vários métodos heurísticos para a construção de classificadores de árvores de decisão. Eles são geralmente construídos de cima para baixo, começando no nó raiz e, sucessivamente, o conjunto é particionado de acordo com suas características. A construção envolve três etapas principais: selecionar uma regra de divisão para cada

nó interno, ou seja, a determinação da funcionalidade, juntamente com um limiar que vai ser usado para dividir o conjunto de dados em cada nó; determinar que os nós são nós terminais. Isso significa que para cada nó temos de decidir se quer continuar dividindo ou ao nó terminal é atribuído um rótulo de classe; atribuição de rótulos de classe para nós terminais, minimizando a taxa de erro estimada.

Dentre as técnicas usadas nesse trabalho, as que obtiveram melhores desempenho foram: *Functional Tree* (FT), J48 e J48graft. O FT usa regressão linear para criar combinações lineares dos atributos, construindo nós multivariantes durante o crescimento da árvore, e folhas durante o processo de poda. A técnica de poda é aplicada para garantir a generalização do modelo representado pela árvore. Consiste em uma análise da AD após o treinamento visando a substituição de certos conjuntos de nós e ramos por folhas, na tentativa de eliminar as especificidades geradas pelo treinamento (WITTEN, 2000). Os resultados apresentados por Gama (2004) mostram que árvores com nós e folhas multivariantes apresentam melhores resultados do que árvores univariantes, árvores multivariantes padrão e árvores modelo, principalmente para grandes conjuntos de dados.

O FT é um algoritmo de construção de árvores de decisão oblíquas, no qual as árvores são multivariadas – cada regra possui uma expressão baseada em mais de um atributo (DEBASTIANI, 2013). Vasconcellos em (2011) relata que uma AD padrão, como as geradas pelo C4.5 (QUINLAN, 1986), é chamada univariante por testar um único atributo por folha, e uma AD dita multivariante utiliza testes baseados na combinação de diversos atributos. As AD multivariantes são também chamadas de oblíquas pela capacidade de dividir o espaço de atributos com multiplanos oblíquos aos eixos que o definem (VASCONCELLOS, 2011).

O J48 é a implementação WEKA do algoritmo C4.5. O C4.5 é a evolução do algoritmo ID3 (QUINLAN, 1986), que foi um dos primeiros algoritmos de AD induzida. O J48 é um algoritmos padrão que testa um único atributo em cada nó. As árvores geradas são denominadas univariantes ou eixo-paralela, por testar um único atributo por nó (QUINLAN, 1986). Este tipo de teste divide o espaço de atributos com um multiplano perpendicular ao eixo que representa o atributo em questão, ou seja, paralelo aos demais. Logo, a denominação *axis-parallel* refere-se a esta propriedade.

O J48graft utiliza uma técnica chamada *grafting* em uma árvore treinada pelo J48 para reduzir o erro na classificação de objetos fora do conjunto de treinamento, adicionando novos testes (nós) a uma árvore já treinada, eliminando inconsistências da árvore original. Uma AD divide o espaço de atributos em um número de subespaços igual ao número de folhas da árvore (VASCONCELLOS, 2011).

2.3 CONSIDERAÇÕES FINAIS

Foi apresentada a revisão de literatura sobre Aprendizagem de Máquina. Neste trabalho, são utilizados algoritmos de Agrupamento e Classificação para a criação do modelo proposto. Alguns pontos importantes para o entendimento do trabalho foram abordados. No próximo capítulo, são descritos os materiais e métodos empregados na pesquisa

3. MATERIAIS E MÉTODOS

O início da presente pesquisa envolveu o estudo sobre RSSF focado em temas de interesse do Laboratório de Automação Agrícola da Escola Politécnica da USP (LAA). Dentro desse estudo, destacou-se o uso das RSSF no monitoramento agrícola e ambiental.

A partir desse estudo, foi observada que uma área com grandes possibilidades de contribuição é a FS, com algumas aplicações já existentes em máquinas agrícolas, porém pouco emprego em sistemas decisórios para aplicação agrícola e ambiental. Buscou-se então, um aprofundamento no estudo de técnicas de FS, que poderiam ser empregadas nas aplicações de monitoramento agrícola e ambiental.

O foco da pesquisa voltou-se ao monitoramento ambiental, no caso a qualidade da água, pois a partir de participações em eventos da área, notou-se a grande preocupação do gerenciamento da água associada a agricultura. Com os resultados da pesquisa em FS, definiu-se o foco de criar um modelo de fusão para avaliar a qualidade da água considerando seu uso na agricultura e impacto ambiental.

Por se tratar de uma pesquisa que envolve um tema multidisciplinar, a interação com outros pesquisadores, principalmente profissionais da engenharia agrícola e ambiental, foi fundamental para delimitar o tema da tese. Contou-se com a parceria do professor Dr. Domingos Sávio Barbosa pesquisador do grupo de pesquisa Tecnologias e Gestão Ambiental da Universidade Federal de Mato Grosso, Rondonópolis/MT.

Em reuniões com professor Domingos foi possível definir que avaliar com rapidez e acurácia a qualidade da água em relação a contaminação por solo, advindo por exemplo, de áreas de cultivo, tem um impacto importante no monitoramento de processos ecológicos de grande relevância tais como o assoreamento, perdas e solos, carreamento de pesticidas e degradação de *habitats* aquáticos.

Definiu-se como objetivo dessa tese, a criação de um modelo de fusão para a classificação da qualidade da água contaminada por solo em tempo real, diferente do protocolo manual que necessita visitas a rios para coleta de amostras de água e análise em laboratório. Até o presente momento, nas referências levantadas não

foram encontrados trabalhos que indique ser possível propor uma análise em tempo real para a qualidade da água tendo como referência ao Índice de Qualidade da Água (IQA) usado para classificar a água potável para consumo humano. O IQA desenvolvido pela *United State Environment Protection Agency* (USEPA) é adotado no Brasil pela Companhia Ambiental do Estado de São Paulo (CETESB) e outros (FERREIRA e ALMEIDA, 2005; HADDAD, 2007). E também, de acordo com as referências levantadas, não foram encontrados trabalhos relacionados a propostas de modelos de classificação da qualidade da água nesse contexto.

O IQA é calculado utilizando-se nove tipos de variáveis (temperatura da água, pH, oxigênio dissolvido, demanda bioquímica de oxigênio, coliformes termotolerantes, nitrogênio total, fósforo total, resíduo total e turbidez). Destes, apenas temperatura da água, pH, oxigênio dissolvido e turbidez são mensuráveis por sensores. Para as demais não existem sensores no mercado e amostras d'água precisam ser coletadas manualmente para análise em laboratório. Existem outros índices que são calculados com um número menor de variáveis, que não são medidas por sensores, exigindo análise laboratorial, o que impossibilita o monitoramento em tempo real (LERMONTOV, YOKOYAMA et al., 2008). A coleta de dados por meio de uma RSSF possibilita o monitoramento em tempo real, diferente da forma como usualmente é feita, o que torna o processo demorado.

Os estudos de técnica e modelos de fusão são, em sua maioria, aplicados nas áreas de robótica, segurança pública e militar. São poucos os exemplos de aplicações nas áreas agrícolas e ambientais. Os danos ambientais gerados pela produção agrícola tanto em água quanto no solo necessitam de ferramentas de monitoramento contínuo para gerar tomadas de decisão mais efetivas, justificando a importância de desenvolver ferramentas para avaliação da qualidade ambiental.

Com a criação desse modelo, espera-se em trabalhos futuros, definir-se um índice de qualidade da água nesse contexto, servindo de referência para o monitoramento ambiental assim como o IQA.

Para a criação do modelo foi realizada um experimento controlado em laboratório, partindo-se de uma água “pura” sendo contaminada de forma gradual. O propósito do experimento foi poder fazer alterações controladas na qualidade da água para coletar dados que indiquem mudanças na amostra. A seguir, o experimento em questão é descrito em detalhes.

3.1 DESCRIÇÃO DO EXPERIMENTO REALIZADO EM LABORATÓRIO

Os dados usados na classificação da qualidade da água foram gerados a partir de um experimento controlado em laboratório. Nesse experimento, foi preparada uma solução-teste de solo natural com a concentração da solução/suspensão inicial de 1,0 g.L⁻¹ simulando o efeito da entrada de solo em tubulação e enxurrada.

As condições do ensaio foram as seguintes: a sonda (Sonda multissensorial para água - *Hanna Instruments* HI 9828 – Figura 10) foi mergulhada em um béquer contendo 1000 ml de água de torneira e ligada. Após a estabilização deram-se início aos registros. Os primeiros 30 s de cada ensaio representam as condições iniciais. Foram coletadas onze variáveis (Tabela 1): pH, pHmV, Redox Potencial (ORP), Oxigênio Dissolvido (OD ppm e saturação), condutividade elétrica (μ _S, μ _SCM, Mohm), Sólidos Totais Dissolvidos (TDS) e salinidade. O aparato experimental adotado foi do tipo *batch*, com agitação constante, que simula uma condição de mistura completa do solo na coluna de água, sem perdas volumétricas, tipicamente observado em condições de contaminação uniforme em tubulações de distribuição de águas tratadas e em zonas de mistura completa de rios com influência de efluentes.

Figura 10 – Sonda usada para coletar os dados no experimento em laboratório.

Fonte: autor.

Tabela 1 - Lista das variáveis coletadas durante o experimento.

Variável	Descrição
Temperatura °C	Condição termal
pH	Potencial Hidrogeniônico
pHmv	Potencial Hidrogeniônico
ORP (mV)	Potencial de Oxi-redução
OD (%)	Saturação de Oxigênio Dissolvido
OD (ppm)	Concentração de Oxigênio Dissolvido
μ_s	Condutividade Elétrica
μ_{scm}	Condutividade Elétrica
Mohm	Resistência Elétrica
TDS	Sólidos Totais Dissolvidos
Salinity (‰)	Salinidade

Fonte: autor.

Nesse teste, a solução (sol) descrita anteriormente seguiu o padrão de dosagem de 1,0 ml a cada intervalo de 240 s. Após 10 dosagens sucessivas (40 min), no mesmo recipiente e sem desligar o aparelho (sonda), a dosagem era aumentada para 10 ml a cada 240 s com 10 dosagens sucessivas. Em síntese tem-se:

- Primeira etapa: 1,0 ml das sol. A cada 240 s (10 aplicações sucessivas);
- Segunda etapa: 10,0 ml das sol. A cada 240 s (10 aplicações sucessivas).

A concentração nominal estimada de solo em função do tempo (intervalo de aplicações) foi conforme pode ser visto na Tabela 2. Os parâmetros adotados como referência para o estudo são considerados como variáveis descritoras básicas do metabolismo aquático, estando relacionadas com características tais como acidez, presença de sais dissolvidos e oxigenação. (TUNDISI & TUNDISI, 2008). Desta forma, a maior parte das alterações na qualidade das águas podem ser avaliadas por meio deste conjunto básico de variáveis. Nove das onze variáveis coletadas foram consideradas: pH, pHmV, ORP, OD, ODppm, μ_{Scm} , μ_{ScmA} , MOhmcm, SDTppm. A Tabela 3 apresenta alguns dados estatísticos de cada uma.

Tabela 2 – Concentração nominal estimada de solo em função do tempo.

sequência da aplicação (nº)	Intervalo de tempo (s)	Concentração (g/L)
0		0,0000
1	0	0,0010
2	240	0,0020
3	480	0,0030
4	241	0,0040
5	481	0,0050
6	242	0,0060
7	482	0,0070
8	243	0,0079
9	483	0,0089
10	244	0,0100
11	484	0,1089
12	245	0,1176
13	485	0,1262
14	246	0,1346
15	486	0,1429
16	247	0,1509
17	487	0,1589
18	248	0,1667
19	488	0,1743
20	249	0,1818
21	489	0,2703

Fonte: autor.

Tabela 3 – Variação, Média e Desvio Padrão das variáveis referentes a qualidade da água contaminada por solo usadas para definição do modelo de classificação.

Variável	Variação	Média	Desvio Padrão
SDTppm	11-14	12.555	0.595
µScma	22-26	23.896	1.055
µScm	23-28	25.1	1.098
pHmV	20.5-46.6	33.615	8.185
pH	6.12-6.57	6.345	0.141
ORP	135.5-170.9	165.105	8.175
ODppm	6.48-6.98	6.766	0.097
OD	77.6-83.1	80.641	1.033
MOhmcm	0.036-0.044	0.04	0.002

Fonte: autor.

No presente estudo, adotou-se, por definição de qualidade ou alteração de qualidade da água, a mudança dos valores das leituras das variáveis em estudo em relação a condição inicial de qualidade da amostra de água em ensaio, considerando que a contaminação causa distúrbios nessas variáveis.

3.2 PROCESSAMENTO DOS DADOS E ANÁLISE

Os dados coletados durante o experimento em laboratório foram submetidos a diferentes algoritmos de agrupamento e classificação para verificar qual seria a melhor divisão em classes e em seguida averiguar a coerência dessa divisão, ou seja, se cada os dados de uma mesma classe apresentavam relação estatística forte entre si. A partir disso, definiu-se os algoritmos a serem usados no modelo de classificação para identificar dados não conhecidos por ele. A Figura 11 apresenta alguns valores estatísticos de cada variável considerada.

Figura 11 – Alguns valores estatísticos de cada variável são apresentados aqui, como os valores máximo e mínimo, média e desvio padrão. DO = dissolved oxygen; TDS = Total dissolved Solids; pH = Hydrogen Potential; ORP = Oxidation-Reduction Potential; Scm = electric conductance; MOhm = electric resistance.

Fonte: autor.

Os algoritmos de Árvore de Decisão usados foram: Best-First Decision Tree Classifier – BFTree, Functional Trees – FT, Naïve Bayes Decision Tree – NBTree,

Grafted C4.5 Decision Tree – J48graft, C4.5 Decision Tree – J48 e LADTree. Esses algoritmos fazem parte da coleção de algoritmos de aprendizagem de máquina para mineração de dados implementados no software WEKA 3.6.10 (HALL, 2009). O software WEKA (*Waikato Environment for Knowledge Analysis*) começou a ser escrito em 1993, usando Java, na Universidade de Waikato, Nova Zelândia e atualmente encontra-se licenciado sob a *General Public License* (GPL). O WEKA agregar um conjunto de algoritmos da área de aprendizagem de máquina para a mineração de dados.

O conjunto de dados gerados no experimento em laboratório, gerou uma amostra de 5100 leituras. Esse conjunto foi dividido em dois: um de treinamento (3400 dados) e um de teste (1700 dados) (ABERNETHY, 2010). Esses dados foram preparados com rotinas em linguagem de programação C, sendo feita a preparação para o arquivo de entrada (.ARFF) para o WEKA e a divisão em dois conjuntos. A divisão em dois arquivos foi feita de forma aleatória, removendo linhas de dados do arquivo original (com 5100 dados) para construir o arquivo de dados testes.

É necessário fazer essa divisão para se evitar o problema de *overfitting*, que acontece quando fornece-se muitos dados para a criação de um modelo (usando árvores de decisão) e este se torna perfeito para esse único conjunto de dados (ABERNETHY, 2010). Na verdade, pretende-se criar um modelo para prever a saída de outros conjuntos de dados diferente dos usados na sua construção, por isso com o conjunto de teste pode-se averiguar se a exatidão do modelo não diminui com um conjunto diferente de dados. Isso garante que o modelo vai prever com exatidão as saídas para valores desconhecidos .

O conjunto de dados completo (5100) foi separado em um primeiro momento em doze classes de 425 dados, classificados como Muito Ótima, Ótima, Pouco Ótima, Muito Boa, Boa, Pouco Boa, Pouco Ruim, Ruim, Muito Ruim, Pouco Péssima, Péssima, Muito Péssima. Sendo Muito Ótima a classe do primeiro intervalo de tempo, e a Muito Péssima, a classe de tempo final da amostragem, pois de acordo com a condução do experimento, sabe-se que a primeira classe de dados tem nenhuma contaminação e a ultima classe tem a maior contaminação do experimento.

Para avaliar a coerência da classificação em questão, os algoritmos selecionados foram executados usando-se o procedimento experimental como sugerido pelo WEKA (usando uma amostra de treinamento e outra amostra de

teste). A amostra de treinamento é usada para a construção do modelo. A amostra de teste verifica se o modelo está correto. Com essa classificação de 12 classes alguns testes foram realizados para averiguar a coerência da classificação, e foram publicados em (DOTA, 2014).

Num segundo momento, houve a necessidade de averiguar se realmente essa divisão em 12 classes era a melhor, já que não há na literatura um consenso de quantas classes são necessárias para classificar a qualidade da água contaminada por solo. Foi utilizado uma abordagem de agrupamento para definir o numero ótimo de classes.

O método de agrupamento k-means++ (ARTHUR, 2007) foi usado para construir classes variando de 2 ate 20, e as seguintes métricas foram adotadas para escolher o melhor numero de classes. Como já citado anteriormente, no experimento em laboratório, a amostra de água foi contaminada sistematicamente, sendo os dados coletados ordenados de uma água pouco poluída até muito poluída e consequentemente, um bom agrupamento manterá instâncias similares juntas e instâncias diferentes em classes separadas.

Considerando que as classes sejam nomeadas como $l=0,1,...,n$, onde n é o número de classes definidas pelo agrupamento. O erro quadrático para esse agrupamento é definido por:

$$E_{dist}(i) = l_i - v_i$$

$$E = \sum_{i=1}^N \frac{(E_{dist}(i))^2}{n}$$

onde v_i é classe correta para a instância i , e N é o numero total de instâncias na amostra (5100). Para estimar a classe correta para cada instância, a amostra foi manualmente dividida em n classes, atribuindo um numero de N/n instâncias para cada classe, sendo zero a amostra mais pura e a classe n com a amostra mais poluída. Este erro irá assumir um grande valor quando a amostra pura for atribuída como uma poluída, mas um valor pequeno quando sua atribuição for perto do pressuposto. Essa partição foi executada para cada agrupamento (de 2 ate 20 classes).

Além de uma distância mínima do nosso agrupamento de referência, é também conveniente que as classes sejam particionadas em conjunto de dados (aproximadamente) de mesmo número de instâncias por classe, assim foi incluída uma métrica baseada na entropia para medir se as partições estão bem equilibradas:

$$H = \sum_{i=1}^n p(l=i) \log_n p(l=i)$$

onde, $p(l=i)$ é a probabilidade de uma instância ser nomeada como i . H irá assumir o valor 1 se o conjunto de dados for perfeitamente particionado e valor igual a zero se todas as instâncias forem associadas a uma só classe.

Finalmente, o melhor valor para n foi escolhido como o agrupamento que minimiza a seguinte métrica:

$$M = \frac{E}{H}$$

Para avaliar a coerência da classificação, alguns algoritmos foram executados seguindo o procedimento experimental *cross-validation* (*k-fold-cross-validation*) com $k = 10$. Os algoritmos de Árvore de Decisão usados foram:

- BFTree (Best-First Decision Tree Classifier),
- FT (Functional Trees),
- NBTree (Naïve Bayes Decision Tree),
- J48graft (Grafted C4.5 Decision Tree),
- J48 (C4.5 Decision Tree),
- LADTree.

A execução usando *Cross-Validation* consiste em dividir a amostra em k vezes de teste, geralmente com $k = 10$. Considerando $k = 10$, a amostra é dividida em 10 partes, sendo 9 delas usadas para treino e uma para teste, repetindo isso 10 vezes e calculando no final a média das 10 execuções. O objetivo é criar um modelo para prever a produção de outros conjuntos de dados diferentes das utilizadas para a sua construção; o arquivo de dados de treinamento é usado para construir um modelo e o arquivo de dados de teste é usado para verificar se a sua precisão não diminui com diferentes conjuntos de dados, garantindo que o modelo vai prever a saída para valores desconhecidos com precisão.

Após os testes com o algoritmos de agrupamento, a amostra foi separada em 07 classes e nomeadas como: Excelente, Muito Boa, Boa, Pouco Boa, Pouco Ruim, Ruim e Péssima. Excelente é a primeira classe contendo as primeiras leituras, já que de acordo com a condução do experimento, é sabido que as primeiras leituras têm pouca contaminação e as últimas maiores contaminações.

3.3 CONSIDERAÇÕES FINAIS

Foi apresentada a metodologia usada no trabalho, como os algoritmos de agrupamento e classificação, configuração do experimento em laboratório, processamento dos dados. No próximo capítulo são mostrados e discutidos os resultados obtidos.

4. RESULTADOS E DISCUSSÕES

A partir das dos ensaios laboratoriais e modelos apresentadas no capítulo anterior, os resultados estão aqui apresentados. A Figura 12 mostra o valor do erro M para cada agrupamento executado (2 a 20). O agrupamento com sete classes apresenta-se com o melhor resultado comparado aos demais, sendo a com 12 classes a segunda melhor. Portanto, a divisão em classes de qualidade da água foi feita em sete classes. A Figura 13 mostra a saída do melhor agrupamento, a maioria dos casos de instâncias associadas de forma incorreta, ocorreram na borda das classes, em que o *cluster* em que a instância se encaixa melhor é incerto. Alguns erros de classificação ocorreram com a classe 5, mas ainda assim as instâncias não ficaram distância da classe esperada.

Figura 12 – Erro calculado para cada parâmetro de agrupamento.

Fonte: autor.

Figura 13 – Distribuição das instâncias em agrupamento de 7 classes. Cada grupo esperado foi colorido diferente entre si, o agrupamento definido pelo k-means++ foi atribuído ao eixo y.

Fonte: autor.

A Tabela 4 apresenta os resultados dos testes com os algoritmos de classificação e a Figura 14 apresenta os resultados em um gráfico de barra. Os algoritmos estão colocados em ordem decrescente de Instâncias Corretamente Classificadas (ICC). Uma instância nesse caso é um conjunto de leituras em um determinado tempo das nove variáveis consideradas. O valor de ICC indica quantas instâncias foram classificadas de forma apropriada a sua classe, e pode-se notar que os valores de ICC foram altos. Os últimos dois algoritmos foram os que obtiveram melhor desempenho, ou seja, menor valor de ICC.

Tabela 4 – ICC representam Instâncias Corretamente Classificadas e IIC representam Instâncias Incorretamente Classificadas. Kappa é um valor que representa a relação entre instâncias da mesma classe, variando entre 0 e 1, sendo 1 uma forte relação entre elas.

Algoritmo	ICC (%)	IIC (%)	kappa
FT	99,6667	0,3333	0,9961
J48graft	99,2157	0,7843	0,9908
J48	99,2157	0,7843	0,9908
NBTree	99,0392	0,9608	0,9887
BFTree	98,9608	1,0392	0,9877
LADTree	96,4314	3,5686	0,9579

Fonte: autor.

Figura 14 – Resultado de ICC e IIC para cada algoritmos de classificação testado usando *k-fold-cross-validation*. Note que os algoritmos mais a direita obtiveram menores valores de ICC.

Fonte: autor.

Outro resultado interessante está relacionado ao valor de Kappa. Kappa é um índice análogo a um coeficiente de correlação. Kappa igual a zero significa ausência de qualquer relação estatística entre as instâncias de uma classe, e próximo a 1 uma forte relação. Pode-se notar na Tabela 4 e também pela Figura 15, que os mesmos algoritmos que apresentaram altos valores para ICC, também apresentaram valores

de Kappa próximos a um, demonstrando uma relação estatística forte entre as diferentes instâncias dentro de uma mesma classe classificada corretamente.

Figura 15 – Valores de kappa obtidos pela execução dos algoritmos.

Fonte: autor.

Estes resultados mostram que a classificação proposta é coerente, já que os diferentes algoritmos mostraram uma forte relação estatística entre as instâncias dentro da classe. Os algoritmos que melhor avaliaram a amostra foram FT, J48graft e J48. Assim, qualquer um destes três algoritmos podem ser utilizados no modelo de classificação de qualidade da água contaminada com o solo, garantindo que o modelo preverá a saída para valores desconhecidos com acurácia.

Para aplicar o modelo proposto, deve-se receber como entrada os dados dos sensores (advindos por exemplo de uma RSSF) fazer a fusão por meio dos métodos de AD, segundo a arquitetura apresentada na Figura 16.

Figura 16 – É apresentada o fluxo de dados de entrada e a saída que será a classificação de acordo com as classes aqui definidas.

Fonte: adaptado de Bar-Shalom, 2000.

4.1 CONSIDERAÇÕES FINAIS

Foram apresentados os resultados e discussão referente aos testes realizados que foram necessários para a criação do modelo de classificação. Os resultados mostram coerência na classificação do modelo proposto, de forma que ele possa avaliar entradas desconhecidas e classifica-las dentro das sete classes de qualidade da água contaminada por solo aqui proposta. Os algoritmos de classificação com melhores resultados foram FT, J48graft e J48. No APÊNDICE A, encontram-se as árvores geradas por cada um desses classificadores.

5. CONCLUSÕES

Neste capítulo são apresentadas as conclusões do trabalho com referência nos resultados obtidos. Seguem também sugestões de trabalhos futuros para a continuidade da pesquisa.

Este trabalho apresentou como contribuição original o modelo de classificação da qualidade da água contaminada por solo usando técnicas de IAD. Com este modelo, torna-se possível acompanhar alterações que venham a ocorrer em águas superficiais, indicando o nível de contaminação por solo com maior rapidez do que a forma convencional, que necessita de análise em laboratório e coleta de amostra manual. Utilizando sensores para medir em tempo real as condições da água, como as RSSFs, e a partir desses dados, inferir-se-á ou não se há contaminação, classificando-a dentro das sete classes definidas no trabalho. O objetivo em abordar temas como FS, RSSF e AM foi para ilustrar a potencialidade deles para aplicações agrícolas e ambientais.

Mesmo com diversas outras técnicas dentro da área de IA que podem ser aplicadas a FS, como Redes Neurais e Lógica *Fuzzy*, a Indução por Árvore de Decisão tem sido considerada como um modelo adequado para a extração de conhecimento dentro da AM, pois são de fácil compreensão e aplicação (MICHIE et al., 1995). Além disso, os resultados dos testes obtidos com a AD foram satisfatórios em relação aos objetivos pretendidos, não havendo, assim, a necessidade de usar classificadores mais poderosos como Redes Neurais e *Support Vector Machine* (SVM), dando preferência a uma solução parcimoniosa. A configuração dos testes como *cross-validation* ($k=10$) valida os resultados alcançados, pois o número de instâncias classificadas corretamente foi produzido de uma média de 10 execuções. Um ponto interessante das AD é a criação das regras, podendo essas serem usadas pelos gestores ambientais tanto em uma ferramenta automatizada como anotada em papel.

Os testes realizados com os algoritmos de classificação indicaram, por meio da taxa de classificação correta (ICC) de cada um, que os algoritmos FT, J48 e J48 são os melhores para compor o modelo de classificação. O presente trabalho abordou a fusão seguindo a premissa que a contaminação causa mudanças em variáveis que podem ser medidas no ambiente monitorado.

Em geral, os principais trabalhos da área abordam a questão da fusão no sentido de inferir, por meio de variáveis conhecidas, o valor de uma variável não medida, como Motamarri e Boccelli (2012) apresentam a aplicação de um método para inferir a quantidade de coliformes fecais na qualidade da água para recreação baseado na turbidez, fluxo de água e dados referente a precipitação. Já o presente trabalho teve o objetivo de criar classes para qualificar a água com base na realidade, testá-la e assim propor o modelo de classificação.

A relevância também está relacionada à necessidade em se identificar, com rapidez e acurácia, a contaminação de corpos de água superficiais com fluxos de águas contaminadas por suspensão de solo. O processo usado para identificar essas alterações foi por meio de amostragem em campo, e a maior parte dos métodos de monitoramento *online* envolvem altos custos.

Por meio desse modelo, quando alguma alteração for detectada, alertas poderão ser emitidos aos gestores ambientais para que as devidas ações sejam tomadas, ou uma averiguação *in situ* para se confirmar ou não o alerta emitido. Também, gestores ambientais poderão planejar políticas ambientais, como a elaboração de práticas de manejo da terra adequadas, minimizando a contaminação de corpos de água por perda de pesticidas e fertilizantes devido ao escoamento e enxurrada superficial desses produtos químicos de áreas agrícolas. Além dessas práticas adequadas auxiliarem na redução da contaminação, podem também aumentar a produtividade e proporcionar melhor relação custo/benefício aos agricultores a longo prazo.

Como trabalhos futuros em continuidade a este trabalho, a partir do modelo criado, pretende-se implementar uma ferramenta que seja de fácil uso dos gestores ambientais, oferecendo um sistema automatizado que receba as leituras dos sensores, avalie os dados, por meio do modelo, e produza saídas indicando a classificação da água.

Também pretende-se realizar novos experimentos em laboratório e ampliar a aplicação dos métodos de IAD para outros contexto da qualidade da água, buscando averiguar a possibilidade de inferir informações por meio das leituras dos sensores de variáveis para as quais ainda há a necessidade de análise em laboratório, estendendo o emprego das técnicas de FS.

Outra proposta para trabalho futuro é a de definir um padrão de qualidade para contaminação da água no contexto desse trabalho, tendo como modelos o IQA

e também tratando questões abordadas pela Resolução CONAMA Nº 357/2005, que dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões de lançamento de efluentes, e de outras providências (CONAMA, 2005). O Conselho Nacional do Meio Ambiente (CONAMA), é um órgão consultivo e deliberativo do Sistema Nacional do Meio Ambiente (SISNAMA), instituído pela Lei 6.938/81, que dispõe sobre a Política Nacional do Meio Ambiente, regulamentada pelo Decreto 99.274/90. É um colegiado representativo de cinco setores, a saber: órgãos federais, estaduais e municipais, setor empresarial e sociedade civil. A criação do modelo de classificação é o primeiro passo para se chegar a definição de um índice padrão.

Com esse trabalho também espera-se ampliar a aplicação das RSSF no monitoramento de processos ambientais, podendo contribuir de forma interdisciplinar nas áreas de Redes, Inteligência Artificial e Ambiental.

REFERÊNCIAS

ADAMCHUK, V. I. et al. On-the-go soil sensors for precision agriculture. **Computers and Electronics in Agriculture**, v. 44, n. 1, p. 71-91, 2004.

ABERNETHY, M. **Data mining with WEKA, Part 2: Classification and clustering**. Series: Data Mining with WEKA, Part2. DeveloperWorks, IBM Technical Library. 2010. Disponível em: <<http://www.ibm.com/developerworks/library/os-weka2/os-weka2-pdf.pdf>>. Acesso em: 10 dez. 2012.

AKYILDIZ, I. F.; SU, W.; SANKARASUBRAMANIAM, Y.; CAYIRCI, E. Wireless sensor networks: a survey. **Computer Networks**. v. 38, n. 4, p. 393-422, 2002.

ARTHUR, D. et al. k-means++: the advantages of careful seeding. In: ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS, 18., 2007. New Orleans, **Proceedings ...** Philadelphia: SIAM, 2007, p. 1027-1035.

BARBOSA, D. S. **Análise da interação entre substâncias húmicas e xenobióticas através de estudos ecotoxicológicos: propostas para a geração de tecnologias de detoxicação aquática**. 2008. 123p. Tese (Doutorado). Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2008.

BAR-SHALOM, Y.; BLAIR, W. D. **Multitarget-Multisensor Tracking: Applications and Advances**. vol. III, Norwood: Artech House, 2000, 608 p.

BIGUS, J. P. **Data mining with neural networks: solving business problems-from application development to decision support**. Hightstown: McGraw-Hill, 1996, 220 p.

BITENCORT, B. R. **Fusão de Dados Paralela em Redes de Sensores Sem Fio Densas Utilizando Algoritmo Genético**. 2008. 85p. Dissertação (Mestrado). Engenharia de Automação e Sistemas, Universidade Federal de Santa Catarina, Florianópolis. 2008.

CONAMA – CONSELHO NACIONAL DO MEIO AMBIENTE. 2005. **Resolução Conama no 357**. Disponível em: <www.mma.conama.gov.br/conama>. Acessado em 25 de maio de 2014.

DASARATHY, B. V. Sensor fusion potential exploitation-innovative architectures and illustrative applications. **Proceedings of IEEE**, vol. 85, n. 1, 1997, pp. 24-38. doi:10.1109/5.554206.

DAUM, F. Handbook of multisensor data fusion [Book Review]. **IEEE Aerospace and Electronic Systems Magazine**, v. 16, n. 10, p. 15-16, 2001. ISSN 0885-8985.

DEBASTIANI, A. B. et al. Árvore de Decisão aplicada na identificação de áreas de risco de deslizamento: comparação dos métodos J48, Functional Trees e Random Tree. In: COLÓQUIO BRASILEIRO DE CIÊNCIAS GEODÉSICAS, 8., 2013, Curitiba. **Anais ...** Curitiba: UFRP, 2013. p.001-002.

DOTA, M. A.; SANTOS, I. M.; CUGNASCA, C. E. Fusão de sensores na agricultura de precisão. In: CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO - CONBAP 2010, 2010, Ribeirão Preto. **Anais ...** Jaboticabal: SBEA, 2010. p.1-9.

DOTA, M. A.; CUGNASCA, C. E. Proposta de um modelo de Fusão Adaptativa de Sensores para o monitoramento agrícola e ambiental. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA - SBIAGRO 2011, 8., 2011, Bento Gonçalves. **Anais ...** Bento Gonçalves: SBIAgro, 2011. p.1-6.

DOTA, M. A.; CUGNASCA, C. E.; BARBOSA, D. S. Comparative analysis of decision tree algorithms for water quality classification with soil-contaminated. **Revista Ciência Rural**. 2014. No prelo.

EL HAJJ, M. et al. Combining multi-source information for crop monitoring. In: INTERNATIONAL CONFERENCE ON INFORMATION FUSION, 11., Cologne. **Proceedings ...** Cologne: IEEE, 2008. p.1-7.

ELMENREICH, W.; PITZEK, S. The Time-Triggered Sensor Fusion Model. In: IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT ENGINEERING SYSTEMS (INES), 5., Helsinki. **Proceedings ...** Helsinki : IEEE, 2001, p. 297-300.

FERREIRA, E. C. F.; ALMEIDA, M. C. D. **Sistema de Cálculo da Qualidade da Água (SCQA) - Estabelecimento das Equações do índice de Qualidade das Águas (IQA)**. Ministério do Meio Ambiente - Secretaria de Estado do Meio Ambiente e Desenvolvimento Sustentável de Minas Gerais. Belo Horizonte, 2005. p.19.

GAJBHIYE, P.; MAHAJAN, A. A survey of architecture and node deployment in Wireless Sensor Network. In: INTERNATIONAL CONFERENCE ON THE APPLICATIONS OF DIGITAL INFORMATION AND WEB TECHNOLOGIES - ICADIWT 2008, 1., 2008. Ostrava. **Proceedings ...** Ostrava: IEEE, 2008. p. 426-430. doi: 10.1109/ICADIWT.2008.4664386.

GAMA, J. A. Functional trees. **Journal Machine Learning**, v. 55, p. 219–250, 2004.

GARCIA-PEREZ, L.; RIBEIRO, A.; GUINEA, D. An agent of behaviour architecture for unmanned control of a farming vehicle. **Computers and Electronics in Agriculture**, v. 60, n. 1, p. 39-48, 2008. ISSN 0168-1699.

HADDAD, E. A. **Influência Antrópica na Qualidade da Água da Bacia Hidrográfica do Rio São Miguel, Carste do Alto São Francisco, Minas**

Gerais. 2007. 141p. Dissertação (Mestrado). Instituto de Geociências, Universidade Federal de Minas Gerais, Belo Horizonte. 2007.

HALL, D. L.e LLINAS, J. An introduction to multisensor data fusion. **Proceedings of the IEEE**, v. 85, n. 1, p. 6-23, 1997. ISSN 0018-9219.

HALL, M. et al. The WEKA Data Mining Software: An Update. **ACM SIGKDD Explorations Newsletter**. v.11, n.1, p.10-18. Jun. 2009. doi: 10.1145/1656274.1656278.

HAN, J; KAMBER, M. **Data Mining – Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers, Inc, 2001.

HEMMAT, A.; ADAMCHUK, V. I. Review: Sensor systems for measuring soil compaction: Review and analysis. **Computers and Electronics in Agriculture**, v. 63, n. 2, p. 89-103, 2008. ISSN 0168-1699.

HUANG, Y. et al. Multisensor Data Fusion for High Quality Data Analysis and Processing in Measurement and Instrumentation. **Journal of Bionic Engineering**, v. 4, n. 1, p. 53-62, 2007. ISSN 1672-6529.

JI, W. et al. **A Remote Wireless Sensor Networks for Water Quality Monitoring**. Innovative Computing & Communication, 2010. Intl Conf on and Information Technology & Ocean Engineering, 2010 Asia-Pacific Conf on (CICC-ITOE), 2010. 30-31 Jan. 2010. p.7-12. doi: 10.1109/CICC-ITOE.2010.9.

KAFTANDJIAN, V. et al. The Combined Use of the Evidence Theory and Fuzzy Logic for Improving Multimodal NDT Systems. **IEEE Transactions on Instrumentation and Measurement**. v. 54, n. 5, p. 1968-1977, 2005. doi: 10.1109/TIM.2005.854255.

KHALEGHI, B. et al. **Multisensor data fusion: Antecedents and directions**. Signals, Circuits and Systems (SCS), In: **Proceedings of the 2009 3rd International Conference on**, 2009. p.1-6. doi: 10.1109/ICSCS.2009.5412296.

KHALIFA, F. A. Survey of watershed modeling and sensor data fusion. 2009. SSST 2009. In: SOUTHEASTERN SYMPOSIUM ON SYSTEM THEORY, 2009, 41, Tullahoma. **Proceedings ...** Tullahoma: IEEE, 2009. p.39-43, 2009. doi: 10.1109/SSST.2009.4806786

LERMONTOV, A. et al. Aplicação da Lógica Nebulosa na Parametrização de um Novo Índice de Qualidade das Águas. **Revista ENGEVISTA**. Rio de Janeiro: Universidade Federal Fluminense, p 10-20. 2008.

LINDEN, R. Técnicas de Agrupamento. **Revista de Sistemas de Informação da FSMA**. n. 4, p. 18-36. 2009.

LUO, R. C.e KAY, M. G. **A tutorial on multisensor integration and fusion**. Industrial Electronics Society, Annual Conference of IEEE, IECON '90., 1990, 16. Pacific Grove. **Proceedings ...** Pacific Grove: IEEE 1990. p.707-722 vol. 1.

_____, (eds.). **Multisensor integration and fusion for intelligent machines and systems**. Norwood, N.J: Ablex Pubed. 1995. 671p. ISBN:0-89391-863-6

MACRUAIRI, R.; KEANE, M. T.e COLEMAN, G. A Wireless Sensor Network Application Requirements Taxonomy. In: International Conference on Sensor Technologies and Applications, 2008 - SENSORCOMM'08, 2. **Proceeding ...** Cap Esterel: IEEE, 2008. p.209-216.

MARTÍN-VIDE, J.P et al. Collapse of the Pilcomayo River. **Geomorphology**. v. 205, n. 15, Jan. 2014, p. 155-163. ISSN 0169-555X.

MSB. **Sensor data fusion: state of the art survey**. Swedish Civil Contingencies Agency 2010. Disponível em: https://www.msb.se/upload/om%20msb/forskning/kunskapsversikt/sensor_data_fusion_survey.pdf. Acessado em: 10 de dez. de 2011.

MICHIE, D. et al. **Machine learning, Neural and Statistical Classification**. 1994. NJ, USA: Ellis Horwood, Upper Saddle River. 290 p.

MIRZA, A. R. **Data Fusion Architectures for Sensor Platforms**. In: IEEE AEROSPACE CONFERENCE, 2008. Big Sky. **Proceedings ...** Big Sky: IEEE, 2008. p 1-13. doi: 10.1109/AERO.2008.4526434.

MOLIN, J. P. Agricultura de Precisão: situação atual e perspectivas. In: FANCELLI, A. L. e DOURADO-NETO, D. (Ed.). **Milho: Estratégias de Manejo para Alta Produtividade**. Piracicaba, 2003. p.89-98.

MONARD, M. C. ; BARANAUSKAS, J. A. **Conceitos sobre Aprendizado de Máquina**. In: Solange Oliveira Rezende. (Org.). **Sistemas Inteligentes - Fundamentos e Aplicações**. 1 ed. Barueri: Editora Manole Ltda, 2003a. p. 89-114.

MOTAMARRI, S.; BOCCELLI, D. L. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. **Journal of the Water Research**, v. 46, n. 14, Sep. 2012, p. 4508-4520.

NAKAMURA, E. F.; LOUREIRO, A. A. F.e FRERY, A. C. Information fusion for wireless sensor networks: Methods, models, and classifications. **ACM Computing. Surveys**, v. 39, n. 3, p.1-9, 2007. ISSN 0360-0300.

NAVABZADEH RAZAVI, S. **Data Fusion for Materials Location Estimation in Construction**. 2010. Tese (PhD). Civil and Environmental Engineering, University of Waterloo, Faculty of Engineering.

OLIKER, N.; OSTFELD, A. A coupled classification – Evolutionary optimization model for contamination event detection in water distribution systems. **Journal of the Water Research**, v.51, Mar. 2014, p.234-245. doi: 10.1016/ j.watres. 2013. 10.060

QIN, X.;GAO, F.; CHEN, G. Wastewater quality monitoring system using sensor fusion and machine learning techniques. **Journal of the Water Research**, v.46, n.4, Mar. 2012, p.1133–1144.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v.1, n.1, p. 81-106, 1986.

RAOL, J. R., Ed. **Multi-Sensor Data Fusion with MATLAB**. Bangalore: CRC PRESS, 2009. 568p.

RASEKH, A.; BRUMBELOW, K. Drinking water distribution systems contamination management to reduce public health impacts and system service interruptions. **Environmental Modelling & Software archive**, v. 51, Jan. 2014, p. 12-25. doi: 10.1016/j.envsoft.2013.09.019.

REN, C. L.; MICHAEL, G. K. **Multisensor integration and fusion for intelligent machines and systems**. Norwood: Ablex Publishing Corp., 1995. 704 p. ISBN:0-89391-863-6.

SANTOS, T. G. et al. Fusão de Dados em Ensaios Não Destrutivos Utilizando Decisão Fuzzy para a Avaliação de Soldas Obtidas pelo Processo de Fricção Linear. **Soldagem & Inspeção**, v.12, n.3, Jul. 2007, p.124-132. ISSN 0104-9224.

SANTOS, I. M.; DOTA, M. A.; CUGNASCA, C. E. Visão geral da aplicabilidade de redes de sensores sem fio no monitoramento agrícola no estado de Mato Grosso. In: CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO - CONBAP 2010, 2010, **Anais ...** Ribeirão Preto: SBEA, 2010. p.1-9.

SALUSTIANO, R. E. **Aplicações de técnicas de fusão de sensores no monitoramento de ambientes**. 2008. 162p. Dissertação (Mestrado). Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas Campinas. 2008.

SRIVASTAVA, A. K.; GOERING, C. E.; ROHRBACH, R. P.; BUCKMASTER, D. R. **Precision agriculture**. In: McCann, P. (ed). Engineering Principles of Agricultural Machines. Michigan: ASABE, 2006. Chapter 6, p.123-138.

STEINBACH, M., KARYPIS, G., and KUMAR, V. (2000). A COMPARISON OF DOCUMENT CLUSTERING TECHNIQUES. In: KDD WORKSHOP ON TEXT MINING. 2000, **Proceedings ...** Karypis Lab, 2000. p.142-146. p.1-20.

SUN, Y.; ZENG, Q.e ZHU, Z. Measuring Soil Physical Properties by Sensor Fusion Technique. In: IEEE CONFERENCE ON INDUSTRIAL ELECTRONICS AND APPLICATIONS, 2007 - ICIEA 2007, 2., 2007, **Proceedings ...** Harbin: IEEE, 2007. p.142-146.

TARCA, A. L. et al. Machine learning and its applications to biology. **PLoS Computational Biology**, v.3, n.6, p.0954-0963, 2007. doi: 10.1371/ journal. pcbi. 0030116

TUBAISHAT, M.; MADRIA, S. Sensor networks: an overview. **Potentials, IEEE**, v. 22, n. 2, p. 20-23. doi: 10.1109/MP.2003.1197877.

TUNDISI, J. G.; MATSUMURA-TUNDISI, T. **Limnologia**. São Paulo: Oficina de Textos. 2008. 632 p.

VASCONCELLOS, E. C. **Árvore de Decisão aplicadas ao problema da separação estrela/galáxia**. 2011. 72 p. Dissertação (Mestrado). Instituto Nacional de Pesquisas Espaciais – INPE, São José dos Campos, 2011.

XIONG, S. M. et al. Application Research of WSN in Precise Agriculture Irrigation. In: INTERNATIONAL CONFERENCE ON ENVIRONMENTAL SCIENCE AND INFORMATION APPLICATION TECHNOLOGY - ESIAT 2009, 2009, **Proceedings ...** Wuhan: IEEE, 2009. v.2, p.297-300. doi: 10.1109/ ESIAT. 2009. 231.

XUEMEI, L.; YUYAN, D.; LIXING, D. Study on precision agriculture monitoring framework based on WSN. In: ANTI-COUNTERFEITING, SECURITY AND IDENTIFICATION, 2008 - ASID, 2008, 2., **Proceedings ...** Guiyang City: IEEE, 2008. p.182-185.

WITTEN, I. H.; FRANK, E. **Data Mining: practical machine learning tools and techniques with java implementations**. [S.l.]: Morgan Kaufmann, 2000.

WITTEN, I. H. et al. **Data mining: Practical machine learning tools and techniques**. 3ed. San Francisco: Morgan Kaufmann. 2011. 629 p.

ZHOU, J; ZHANG, M. Coarse sediment and lower Yellow River siltation. **Journal of Hydro-environment Research**, v. 6, n. 4, p. 267-273, dez. 2012. ISSN 1570-6443.

APÊNDICE A

A seguir as árvores são apresentadas para os algoritmos FT, J48 e J48GRAFT, respectivamente.

FT tree

```
pHmV <= 31.7
|   SDTppm <= 12: FT_1:15/45 (1736)
|   SDTppm > 12: Class=5
pHmV > 31.7
|   N0#4 <= 0.408749
|   |   N3#5 <= 0.452317: FT_3:15/60 (1387)
|   |   N3#5 > 0.452317: Class=3
|   N0#4 > 0.408749: FT_5:15/45 (671)
```

Number of Leaves : 5

Size of the Tree : 9

J48 pruned tree

```

pHmV <= 31.7
|
| SDTppm <= 12
| |
| | microScmA <= 22: cluster2 (634.0)
| | microScmA > 22
| | |
| | | pH <= 6.52: cluster7 (936.0/1.0)
| | | pH > 6.52
| | | |
| | | | microScm <= 24
| | | | |
| | | | | OD <= 81.8
| | | | | |
| | | | | | ORP <= 166.4: cluster2 (12.0)
| | | | | | ORP > 166.4
| | | | | | |
| | | | | | | ODppm <= 6.86: cluster7 (27.0/1.0)
| | | | | | | ODppm > 6.86
| | | | | | | |
| | | | | | | | pHmV <= 21.5: cluster2 (4.0)
| | | | | | | | pHmV > 21.5: cluster7 (8.0/1.0)
| | | | |
| | | | | OD > 81.8
| | | | | |
| | | | | | MOhmcm <= 0.0411
| | | | | | |
| | | | | | | OD <= 82.3: cluster7 (5.0)
| | | | | | | OD > 82.3: cluster2 (5.0)
| | | | | | | MOhmcm > 0.0411: cluster2 (73.0/2.0)
| | | |
| | | | microScm > 24: cluster7 (32.0)
|
| SDTppm > 12: cluster4 (458.0)
pHmV > 31.7
|
| pHmV <= 37.9
| |
| | pHmV <= 32.1
| | |
| | | OD <= 80.6
| | | |
| | | | microScm <= 24
| | | | |
| | | | | OD <= 80.1: cluster5 (5.0)
| | | | | OD > 80.1: cluster7 (5.0)
| | | |
| | | | microScm > 24
| | | | |
| | | | | ORP <= 169.8: cluster5 (58.0)
| | | | | ORP > 169.8
| | | | | |
| | | | | | pH <= 6.37: cluster5 (10.0)
| | | | | | pH > 6.37: cluster7 (2.0)
| | | |
| | | | OD > 80.6
| | | | |
| | | | | microScmA <= 23: cluster7 (23.0)
| | | | | microScmA > 23: cluster5 (2.0)
| |
| | pHmV > 32.1
| | |
| | | OD <= 79.6
| | | |
| | | | microScm <= 25: cluster5 (12.0)
| | | | microScm > 25: cluster1 (3.0)
| | | | OD > 79.6: cluster5 (742.0)
|
| pHmV > 37.9
| |
| | ORP <= 157.4
| | |
| | | pHmV <= 41.9: cluster1 (608.0)
| | | pHmV > 41.9
| | | |
| | | | OD <= 79.7
| | | | |
| | | | | ODppm <= 6.63: cluster1 (26.0)
| | | | | ODppm > 6.63
| | | | | |
| | | | | | pHmV <= 42.2: cluster1 (16.0)
| | | | | | pHmV > 42.2
| | | | | | |
| | | | | | | microScm <= 26: cluster6 (6.0/1.0)
| | | | | | | microScm > 26
| | | | | | | |
| | | | | | | | OD <= 79.5: cluster1 (11.0)
| | | | | | | | OD > 79.5: cluster6 (6.0/2.0)
| | | | |
| | | | | OD > 79.7
| | | | | |
| | | | | | ORP <= 154.8: cluster1 (3.0/1.0)
| | | | | | ORP > 154.8: cluster6 (15.0)
| |
| | ORP > 157.4
| | |
| | | pH <= 6.18
| | | |
| | | | ORP <= 169.9
| | | | |
| | | | | ODppm <= 6.73: cluster6 (652.0/1.0)
| | | | | ODppm > 6.73
| | | | | |
| | | | | | microScmA <= 24: cluster3 (3.0)
| | | | | | microScmA > 24: cluster6 (12.0)
| | | | |
| | | | | ORP > 169.9
| | | | | |
| | | | | | microScmA <= 24
| | | | | | |
| | | | | | | ODppm <= 6.64: cluster6 (8.0)
| | | | | | | ODppm > 6.64: cluster3 (85.0)
| | | | | |
| | | | | | microScmA > 24
| | | | | | |
| | | | | | | pH <= 6.17: cluster6 (119.0)
| | | | | | | pH > 6.17
| | | | | | | |
| | | | | | | | OD <= 80.4: cluster6 (7.0)
| | | | | | | | OD > 80.4: cluster3 (3.0)
| | | |
| | | | pH > 6.18
| | | | |
| | | | | MOhmcm <= 0.0394
| | | | | |
| | | | | | microScm <= 25
| | | | | | |
| | | | | | | ORP <= 169.7: cluster3 (42.0)
| | | | | | | ORP > 169.7
| | | | | | | |
| | | | | | | | ODppm <= 6.72: cluster3 (8.0)
| | | | | | | | ODppm > 6.72: cluster5 (6.0)
| | | | | |
| | | | | | microScm > 25
| | | | | | |
| | | | | | | pHmV <= 42.6: cluster3 (380.0)
| | | | | | | pHmV > 42.6
| | | | | | | |
| | | | | | | | microScm <= 26: cluster3 (12.0)
| | | | | | | | microScm > 26: cluster6 (3.0)
| | | | |
| | | | | MOhmcm > 0.0394: cluster5 (13.0)

```

Number of Leaves : 44

Size of the tree : 87

J48graft pruned tree

```

pHmV <= 31.7
  SDTppm <= 12
    microScmA <= 22: cluster2 (634.0)
    microScmA > 22
      pH <= 6.52: cluster7 (936.0/1.0)
      pH > 6.52
        microScm <= 24
          OD <= 81.8
            ORP <= 166.4: cluster2 (12.0)
            ORP > 166.4
              ODppm <= 6.86
                pHmV <= 23.3
                  ORP <= 167.4
                    MOhmcm <= 0.0421
                      microScm <= 23.5: cluster2 (0.0|78.0)
                      microScm > 23.5: cluster7 (27.0/1.0)
                    MOhmcm > 0.0421: cluster2 (0.0|90.0)
                  ORP > 167.4: cluster7 (0.0|398.0)
                pHmV > 23.3: cluster7 (0.0|444.0)
              ODppm > 6.86
                pHmV <= 21.5: cluster2 (4.0)
                pHmV > 21.5
                  pHmV <= 23.25
                    ORP <= 167.4
                      MOhmcm <= 0.0421
                        microScm <= 23.5: cluster2 (0.0|20.0)
                        microScm > 23.5: cluster7 (8.0/1.0)
                      MOhmcm > 0.0421: cluster2 (0.0|21.0)
                    ORP > 167.4: cluster7 (0.0|75.0)
                  pHmV > 23.25: cluster7 (0.0|84.0)
          OD > 81.8
            MOhmcm <= 0.0411
              OD <= 82.3: cluster7 (5.0)
              OD > 82.3: cluster2 (5.0)
            MOhmcm > 0.0411: cluster2 (73.0/2.0)
          microScm > 24: cluster7 (32.0)
      SDTppm > 12: cluster4 (458.0)
  pHmV > 31.7
    pHmV <= 37.9
      pHmV <= 32.1
        OD <= 80.6
          microScm <= 24
            OD <= 80.1: cluster5 (5.0)
            OD > 80.1
              MOhmcm <= 0.0405: cluster5 (0.0|33.0)
              MOhmcm > 0.0405
                microScmA <= 23.5: cluster7 (5.0)
                microScmA > 23.5: cluster5 (0.0|24.0)
          microScm > 24
            ORP <= 169.8: cluster5 (58.0)
            ORP > 169.8
              pH <= 6.37: cluster5 (10.0)
              pH > 6.37
                ODppm <= 6.765: cluster5 (0.0|28.0)
                ODppm > 6.765
                  OD <= 80.45: cluster5 (0.0|28.0)
                  OD > 80.45
                    MOhmcm <= 0.0403: cluster5 (0.0|13.0)
                    MOhmcm > 0.0403
                      microScmA <= 23.5: cluster7 (2.0)
                      microScmA > 23.5: cluster5 (0.0|13.0)
        OD > 80.6
          microScmA <= 23
            MOhmcm <= 0.0421
              microScmA <= 22.5: cluster2 (0.0|633.0)
              microScmA > 22.5
                microScm <= 23.5: cluster2 (0.0|569.0)
                microScm > 23.5
                  SDTppm <= 11.5: cluster2 (0.0|193.0)
                  SDTppm > 11.5
                    ORP <= 166.45: cluster2 (0.0|152.0)
                    ORP > 166.45
                      OD <= 82.55
                        ODppm <= 6.935
                          pH <= 6.565: cluster7 (23.0)
                          pH > 6.565: cluster2 (0.0|117.0)
                        ODppm > 6.935: cluster2 (0.0|134.0)
                      OD > 82.55: cluster2 (0.0|143.0)
                    MOhmcm > 0.0421: cluster2 (0.0|633.0)
            microScmA > 23
              MOhmcm <= 0.0393: cluster4 (0.0|302.0)
              MOhmcm > 0.0393
                microScm <= 25.5
                  microScmA <= 24.5
                    pH <= 6.365: cluster5 (0.0|116.0)
                    pH > 6.365
                      ORP <= 170.05
                        ODppm <= 6.775: cluster5 (0.0|43.0)
                        ODppm > 6.775
                          SDTppm <= 13.5
                            MOhmcm <= 0.04035
                              ORP <= 166.8: cluster7 (0.0|12.0)
                              ORP > 166.8

```