UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Luiz Fernando Afra Brito

Mineração de dados meteorológicos a partir de monitoramento de dispositivos móveis

Uberlândia, Brasil

2015

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Luiz Fernando Afra Brito

Mineração de dados meteorológicos a partir de monitoramento de dispositivos móveis

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Marcelo Keese Albertini

Universidade Federal de Uberlândia

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2015

Luiz Fernando Afra Brito

Mineração de dados meteorológicos a partir de monitoramento de dispositivos móveis

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, XX de XX de 2015:

Marcelo Keese Albertini
Orientador

Professor

Professor

Uberlândia, Brasil
2015

Dedico a todos os meus familiares pelo incentivo e apoio que sempre me deram. Dedico, especialmente, à minha mãe e ao meu pai que, mesmo estando longe, estiveram sempre ao meu lado me dando força nos melhores e piores momentos.

Agradecimentos

Agradeço aos meus amigos mais próximos e alunos participantes do Programa de Educação Tutorial (PET) pelo apoio que me deram durante a minha jornada no curso de Bacharelado de Sistemas de Informação. Também sou grato aos professores Maurício Cunha Escarpinati e André Ricardo Backes, que sempre me ajudaram quando precisei. Agradeço, especialmente, ao professor Marcelo Keese Albertini, uma pessoa agradável e divertida que, além de ser um ótimo tutor e professor, foi peça-chave para a realização deste trabalho de conclusão de curso.

Resumo

Este trabalho de conclusão de curso apresenta estudos da utilização de técnicas de mineração de dados em dados coletados por meio de dispositivos móveis para verificar a viabilidade de seu uso em sistemas de alertas de precipitações atmosféricas. Esse tipo de sistema de alerta é útil para prevenir ou preparar para situações de alto risco e situações do cotidiano. Neste estudo, utilizou-se *smartphones* para coletar dados sobre do estado da comunicação de tecnologias tais como GSM e GPS. Também foi construído um simulador de precipitações simplificado para coletar dados em cenários controlados. Para identificar o tipo de precipitação, utilizou-se algoritmos de classificação como *k-Nearest Neighbors* baseado em instâncias, *Support Vector Machine* e árvore de decisão J48. Os resultados de classificação foram promissores e mostraram área abaixo da curva ROC de 0.95 para o algoritmo *k-Nearest Neighbors* e 0.80 para *Support Vector Machine*. Para estimar a intensidade de precipitação, utilizou-se algoritmos de regressão como linear e *Support Vector Machine* com núcleo radial do tipo *epsilon*. Diferentemente da classificação, os resultados preliminares para regressão mostraram baixa capacidade para estimação da intensidade de precipitações, os modelos criados indicaram ruídos residuais não normais. A partir dos resultados observados, sugere-se a realização de investigações aprofundadas a fim melhorar a qualidade do modelo com o emprego de técnicas de pré-processamento nos dados coletados ou utilização de algoritmos de regressão especializados.

Palavras-chave: dispositivos móveis, estado meteorológico, estimação, mineração de dados, precipitação atmosférica.

Lista de ilustrações

Figura 1 – Árvore de decisão 19
Figura 2 – Modelo de neurônio *perceptron* 20
Figura 3 – Margem máxima do algoritmo SVM 21
Figura 4 – Diagrama do fluxo de dados e informação 22
Figura 5 – Simulador de chuvas 23
Figura 6 – Simulador de chuvas e dispositivo móvel posicionado 24
Figura 7 – Valores preditos e resíduos do modelo de regressão quadrático 28
Figura 8 – Valores preditos e resíduos do modelo de regressão SVM 29
Figura 9 – Gráfico quantil-quantil do modelo de regressão quadrático 29
Figura 10 – Gráfico quantil-quantil do modelo de regressão SVM 30

Lista de tabelas

Tabela 1 – Fatores considerados para experimento de classificação 24
Tabela 2 – Fatores considerados para experimento de regressão 25
Tabela 3 – Resultados de classificação com relação ao cenário com 3 classes 26
Tabela 4 – Matriz de confusão com relação ao cenário com 3 classes e modelo SVM 26
Tabela 5 – Resultados de classificação com relação ao cenário com 12 classes . . . 26
Tabela 6 – Matriz de confusão com relação ao cenário com 12 classes e modelo SVM 27
Tabela 7 – Configurações para cada combinação de fatores utilizados no experimento com 12 classes 27
Tabela 8 – Resultado de modelos de regressão 28
Tabela 9 – Valores de coeficiente de determinação de modelos de regressão SVM para interação de até duas medidas 31
Tabela 10 – Avaliação de modelos de regressão SVM com mais de 2 medidas 31

Lista de abreviaturas e siglas

GPS	*Global Positioning System*
GSM	*Global System for Mobile Communications*, originalmente *Groupe Spécial Mobile*
ITU	*International Telecommunication Union*
IMT-2000	*International Mobile Telecommunications-2000*
WCDMA	*Wide-Band Code-Division Multiple Access*
HSPA	*High Speed Packet Access*
LTE	*Long Term Evolution*
kNN	*k-Nearest Neighbors*
SVM	*Support Vector Machine*
MLP	*Multilayer Perceptron*
SVC	*Support Vector Classification*
ROC	*Receiver Operating Characteristic*
MAE	*Mean Absolute Error*
RMSE	*Root Mean Squared Error*

Sumário

1 INTRODUÇÃO 10
1.1 Objetivos 11

2 REVISÃO BIBLIOGRÁFICA 12
2.1 Histórico de tecnologias de comunicação sem fio para dispositivos móveis . 12
2.2 Avaliação das tecnologias 13
2.3 Estudos sobre mensuração do nível pluviométrico utilizando dados de transmissão sem fio 13
2.3.1 Redes sem fio de alto desempenho e condições meteorológicas 13
2.3.2 Fatores que causam incertezas em comunicações sem fio no ambiente externo 13
2.3.3 Monitoramento pluviométrico utilizando redes de celular 14
2.3.4 Predição de intensidade da chuva utilizando *links* de micro-ondas 14
2.3.5 Coleta de temperaturas do ar urbano de temperaturas da bateria de *smartphones* 15
2.3.6 Observações da pressão de superfície em *smartphones* 16
2.3.7 Uma abordagem de *crowdsourcing* para estimação de condições meteorológicas 17
2.4 Avaliação dos trabalhos relacionados 17
2.5 Técnicas de mineração de dados 18

3 DESENVOLVIMENTO 22
3.1 Experimentos de classificação 23
3.2 Experimentos de regressão 24
3.3 Resultados dos experimentos de classificação 25
3.4 Resultados dos experimentos de regressão 27
3.5 Estudo da influência das medidas utilizadas nos modelos de regressão . . . 30

4 CONCLUSÃO 32

Referências 34

1 Introdução

Sistemas de alerta para eventos meteorológicos são de grande importância econômica e social. Esses sistemas monitoram o nível pluviométrico atual a fim de alertar eventos catastróficos. Predições de altos níveis de precipitação que possam causar problemas e perdas tais como inundações e deslizamentos, podem ser monitoradas, evitando mortes e também outras perdas. Atualmente utiliza-se imagens de satélite (EBERT; JANOWIAK; KIDD, 2007; GRIFFITH et al., 1978) e redes de sensores locais especializadas (MEISCHNER, 2004; RAMIREZ; VELHO; FERREIRA, 2005) para acompanhar tais eventos, contudo, o custo, precisão e rapidez de métodos atuais ainda podem ser melhorados. A expansão do uso de *smartphones* abre novas oportunidades nessa área (OVEREEM et al., 2013) uma vez que esses aparelhos possuem capacidade de processamento, armazenamento de dados e em alguns casos sensores especializados (YI; JIA; SANIIE, 2012).

Sabe-se que *smartphones* comunicam-se por meio de redes de transmissão de dados sem fio. Utilizando-se observações de medidas relacionadas é possível verifica a influência de precipitações atmosféricas sobre essas redes. Por exemplo, tem-se utilizado modelos matemáticos que permitiram a predição de precipitação pelo nível de sinal de torres de celulares com até 90% de acurácia em precipitações atmosféricas (ZINEVICH; MESSER; ALPERT, 2010).

Este trabalho teve como objetivo estudar a viabilidade do uso de informações obtidas por dispositivos móveis para monitorar mudanças meteorológicas para fornecer alertas de mudanças do estado meteorológico. Para isso foram estudados métodos de obtenção e análise de dados a partir do uso da atenuação de sinais, tais como da transmissão sem fio de dispositivos móveis entre células de operadoras, por exemplo, GSM, e satélites do sistema de posicionamento global (GPS). Adicionalmente, foi projetado um simulador de precipitações simplificado para avaliar a influência de precipitações nos sinais de comunicação de *smartphones*. Também foram incluídas outras fontes de dados como, por exemplo, a temperatura de bateria do dispositivo (OVEREEM et al., 2013).

Para estudar a viabilidade do uso de dados de *smartphones* para o desenvolvimento de sistemas de alerta, foram aplicadas técnicas de mineração de dados para classificar entre tipos de precipitação de acordo com sua intensidade e predizer níveis de precipitação. Os resultados obtidos são preliminares e indicam ser mais fácil e confiável classificar entre diferentes tipos de precipitação do que predizer a intensidade da precipitação.

Na próxima seção são descritos de maneira detalhada os objetivos deste trabalho. Nos seguintes capítulos são apresentados os trabalhos relacionados, a metodologia utilizada neste trabalho, os resultados obtidos, as conclusões e uma avaliação dos possíveis

trabalhos necessários.

1.1 Objetivos

A expansão do uso de dispositivos móveis apresenta diversas oportunidades para captura de monitoramento de dados utilizados na estimação do estado meteorológico. Considerando essas oportunidades e a importância socioeconômica de sistemas de alerta, neste trabalho foi abordada a seguinte hipótese:

Hipótese 1 *É possível usar dados obtidos a partir de redes de comunicação de dispositivos móveis para estimar o estado meteorológico.*

Essa hipótese motivou estabelecer o seguinte objetivo geral deste trabalho:

Objetivo 1 *Estudos e implementação de técnicas de mineração de dados que, a partir da coleta de dados de redes de comunicação de dispositivos móveis, comprovem a capacidade de estimação do estado meteorológico.*

Tendo em vista esse objetivo, definiu-se objetivos específicos a fim de organizar o desenvolvimento deste trabalho. Primeiramente, observou-se que não foi encontrada base de dados específica que possibilite a análise do estado meteorológico através de dados obtidos por celulares nos estudos de levantamento da literatura científica. Por essa razão, foi necessária uma etapa de coleta de dados, durante a qual, sob um ambiente controlado, dados foram obtidos de dispositivos móveis a fim de caracterizar o tipo de precipitação.

Objetivo Específico 1 *Coleta de dados utilizando dispositivos móveis em experimentos com ambiente controlado.*

Após a coleta de dados do Objetivo Específico 1, é necessário aplicar técnicas de pré-processamento para avaliar e, se necessário, melhorar a qualidade da base de dados. Técnicas para escolha dos atributos mais relevantes, padronização e eliminação de ruídos são exemplos de técnicas que podem ser aplicadas. Com os dados pré-processados, pode-se construir modelos por meio de técnicas de classificação que possibilitam o aprendizado do sistema para classificação de novos eventos. Também foi parte do Objetivo Específico 2, a avaliação da viabilidade de estimação do nível de volume de precipitação a partir dos dados pré-processados.

Objetivo Específico 2 *Estudos das técnicas mais adequadas para extrair informações de dados meteorológicos a partir das medições pré-processadas.*

2 Revisão Bibliográfica

Neste capítulo são apresentados conceitos sobre técnicas e tecnologias utilizadas neste trabalho. Inicialmente, apresenta-se um histórico sobre tecnologias de transmissão por ondas eletromagnéticas. Posteriormente, descreve-se estudos encontrados na literatura sobre análise meteorológica utilizando redes sem fio. Em seguida, são apresentadas técnicas de mineração de dados que são utilizadas.

2.1 Histórico de tecnologias de comunicação sem fio para dispositivos móveis

A primeira geração de tecnologias de transmissão móvel (1G) utilizou sistemas analógicos para disponibilizar os seus serviços. A grande mudança na tecnologia de geração seguinte (2G) foi a transição para sistemas digitais. Na época, três padrões destacaram-se: Global System for Mobile Communications (GSM), IS-136 e IS-95.

O padrão GSM foi formulado para padronização das comunicações entre países, o que auxiliou na melhoria de sinal e redução de custos por meio de transmissão digital. Isso favoreceu sua disseminação por todo o mundo, como, por exemplo, no Brasil. Atualmente, a maior base instalada no Brasil utiliza o GSM. A frequência base de atuação do GSM no Brasil é 1 800 MHz, na Europa são 900 MHz e 1800 MHz e nos Estados Unidos da América, 1 900 MHz.

Posteriormente, em 1989, a *International Telecommunication Union* (ITU) propôs um conjunto de especificações que futuras tecnologias deveriam contemplar. Esse documento foi chamado de *International Mobile Telecommunications 2000* (IMT-2000). Com isso, surgiu a tecnologia 3G *Wide-Band Code-Division Multiple Access* (WCDMA), a evolução do padrão GSM. Na Europa e no Brasil, as faixas de operação dessa tecnologia ocupam as frequências 1 900 MHz e 2 100 MHz. Após o padrão WCDMA, a tecnologia *High Speed Packet Access* (HSPA) foi desenvolvida e pode operar nas faixas de frequência 850 MHz, 900 MHz, 1 900 MHz e 2 100 MHz.

Em 2008, a ITU presentou um novo conjunto de especificações ao qual as tecnologias 4G deveriam atender. Esse novo documento foi chamado de IMT-*Advanced*. Naquela época a tecnologia LTE (*Long Term Evolution*) estava sendo desenvolvida e ultrapassava a velocidade da HSPA. A tecnologia LTE opera nas frequências 700 MHz, 800 MHz, 1800 MHz e 2 600 MHz. No Brasil, essa tecnologia usa faixas entre 2 500 MHz e 2 690 MHz, porém houve a alocação da frequência 700 MHz em novembro de 2014 (TELECO, 2014a; TELECO, 2014b; SILVA, 2014).

2.2 Avaliação das tecnologias

Observou-se que comunicações por micro-ondas em frequências maiores possibilitaram maior capacidade de detecção do nível de volume pluviométrico. Por isso, a utilização de tecnologias como WCDMA e LTE, que operam nas faixas de frequência de até 2 100 MHz e 2 690 MHz, respectivamente, podem proporcionar melhores resultados na estimação do estado meteorológico. O padrão GSM ainda é bastante utilizado no Brasil, por isso, essa tecnologia foi utilizada para os experimentos deste trabalho (FONG et al., 2003; MESSER, 2007; ZINEVICH; MESSER; ALPERT, 2010).

2.3 Estudos sobre mensuração do nível pluviométrico utilizando dados de transmissão sem fio

2.3.1 Redes sem fio de alto desempenho e condições meteorológicas

Otero, Yalamanchili e Braun realizaram experimentos relacionando fenômenos climáticos e condições atmosféricas e redes sem fio de alto desempenho. Duas estações foram utilizadas para coletar dados climáticos e verificar a correlação entre as informações da rede. Dos dados coletados, encontram-se gráficos para temperatura, índice de resfriamento causado pelo vento, umidade, chuva, ponto de orvalho, pressão, velocidade e direção do vento (OTERO; YALAMANCHILI; BRAUN, 2001).

Foi observado que chuvas e tempestade de neve afetam a qualidade do sinal recebido. Chuvas fortes fizeram com que o sinal sofresse atenuação de 4 dBm e tempestades de neve 14 dBm. Outros fenômenos climáticos também foram observados porém não afetam diretamente o sinal de rádio. Vento forte pode mudar a direção de antenas, fazendo com que a força do sinal diminua em pontos específicos. Umidade e temperatura são fatores que, quando em condições anormais, podem causar danos nos equipamentos ou distorção na saída de ondas eletromagnéticas.

Devido a poucas estações meteorológicas, pouco tempo e equipamentos rudimentares para coleta de dados não foi possível associar com precisão a degradação da rede e fenômenos climáticos. Os efeitos de condições solares em redes sem fio de alto desempenho podem ser uma área interessante para pesquisas futuras.

2.3.2 Fatores que causam incertezas em comunicações sem fio no ambiente externo

Fong et al. mostraram as principais variáveis que afetam comunicações sem fio no ambiente externo. A mais importante delas foi o nível de precipitação podendo atenuar o sinal, romper a conexão ou causar incertezas. Estudos foram feitos e avaliou-se o quanto

a taxa de precipitação pode atenuar o sinal. Notou-se que quanto maior o volume de precipitação, maior será a atenuação do sinal. Verificou-se também que, em transmissões em alta frequência, o sinal é severamente afetado.

2.3.3 Monitoramento pluviométrico utilizando redes de celular

Messer, estudou o monitoramento de eventos chuvosos utilizando redes de celulares. Para isso, foram utilizados pluviômetros e radares climáticos para avaliação do método (MESSER, 2007).

Nesse estudo, percebeu-se que, em frequências acima de 10 GHz, a precipitação tem grande influência na comunicação das redes de celulares. Esse fato foi observado a partir da aplicação de métodos de filtragem de sinais utilizados para determinar quais fatores são relacionados à precipitação. Os dados resultantes dessa filtragem foram utilizados para estimar parâmetros de um modelo matemático cuja saída é a taxa de precipitação. O modelo que foi utilizado é dado pela lei da potência, $A = aR^b$, onde A é a atenuação em dB/km, R é a taxa de precipitação ao longo do enlace de comunicação em mm/h e a e b são parâmetros estimados.

Esse modelo foi empregado em um método baseado em tomografia da distribuição de precipitação temporal-espacial utilizando dados armazenados em sete *links* principais da rede. Comparando com imagens de radares meteorológicos constatou-se que dados obtidos obtidos pela modelagem eram mais parecidos com as informações fornecidas por pluviômetros do que aqueles obtidos por radares. A imagem do radar mostrou cobrir uma grande área espacial, porém, algumas dessas áreas são mostradas erroneamente. Isso ocorre devido a fatores, tais como, barreiras topográficas, montanhas e regiões urbanas.

No final, o autor sugere a utilização de dispositivos móveis para monitoramento de precipitações. A atenuação nas frequências mais comuns nesses aparelhos é relativamente baixa, resultando em menor atenuação do sinal.

2.3.4 Predição de intensidade da chuva utilizando *links* de micro-ondas

Zinevich, Messer e Albert estudaram fontes de incertezas relacionadas a eventos climáticos. Dentre elas estão, a imperfeição nos instrumentos, a atenuação de base, a quantização e outras variáveis relacionadas ao ambiente externo. Percebeu-se que a variabilidade da distribuição do tamanho da gota de chuva ao longo do enlace está diretamente relacionada à medição da média e variabilidade espacial pluviométrica (ZINEVICH; MESSER; ALPERT, 2010).

Foi apresentado em (ATLAS; ULBRICH, 1977 apud ZINEVICH; MESSER; ALPERT, 2010) que *links* de micro ondas sofrem imprecisões inerentes. Foi mostrado nesse artigo que, em frequências em torno de 35 GHz, a lei da potência é aproximadamente

linear e independe do tamanho da gota de chuva e temperatura, o que resulta em um erro médio de regressão abaixo de 10%. Entretanto, para frequências abaixo de 9 GHz, o tamanho da gota de chuva passa a contribuir com mais de 20% de erro médio. Em (RINCON; LANG, 2002 apud ZINEVICH; MESSER; ALPERT, 2010) foi mostrado que estimativas baseadas na lei da potência tendem a subestimar a chuva, especialmente em altas taxas de precipitação, onde a variação das taxas afeta a lei da potência. Antenas molhadas mostraram ter grande impacto na precisão da medição (MINDA; NAKAMURA, 2005 apud ZINEVICH; MESSER; ALPERT, 2010) e esse fator não é levado em consideração na formulação inicial. As incertezas do ar, vapor, entre outros, também têm impacto direto na qualidade da medição (HOLT; KUZNETSOV; RAHIMI; RAHIMI et al.; DAVID; ALPERT; MESSER, 2003, 2003, 2009 apud ZINEVICH; MESSER; ALPERT, 2010). Os efeitos da inclinação angular da gota da chuva, temperatura, variações na microestrutura interna e externa da chuva, frequência e comprimento do enlace, estratégia de amostragem temporal, resolução de energia e umedecimento da antena foram observados em (AYDIN; DAISLEY; BERNE; UIJLENHOET, 2002, 2007 apud ZINEVICH; MESSER; ALPERT, 2010) e (LEIJNSE; UIJLENHOET; STRICKER; LEIJNSE; UIJLENHOET; STRICKER; LEIJNSE; UIJLENHOET; HOLLEMAN, 2007a, 2008a, 2008b apud ZINEVICH; MESSER; ALPERT, 2010).

Com o intuito de estimar médias pluviométricas a partir das medidas mencionadas, foi adotado um modelo de variação espacial da chuva. Esse estudo permitiu fazer predições de precipitação com até 90% de acurácia.

2.3.5 Coleta de temperaturas do ar urbano de temperaturas da bateria de *smartphones*

Overeem et al. utilizaram a temperatura da bateria de *smartphones* para mensurar a temperatura ambiente utilizando um modelo simples de transferência de calor. Fatores como custo e cobertura espacial para captação de temperatura ambiente são beneficiados com o uso desses dispositivos devido à popularização dessa tecnologia (OVEREEM et al., 2013).

A temperatura do *smartphone* foi coletada utilizando a aplicação *OpenSignal* (OPENSIGNAL, 2014) desenvolvida em tecnologia *Android* e, recentemente, para iOS. Posteriormente, essa aplicação foi espalhada por diversos países para que usuários pudessem contribuir com os estudos. Oito grandes cidades foram escolhidas para estudar o efeito de diferentes climas sobre a bateria de *smartphones*. Para comprovação do modelo obtido utilizou-se dados de estações meteorológicas nas proximidades. O modelo utilizado leva em consideração três fontes de calor, o *smartphone*, o corpo do usuário e o ambiente.

A equação obtida foi:

$$\bar{T}_{e,j,d}^{A,day} = m_j(\bar{T}_{p,j,d}^{A,day} - T_0) + T_0 + \epsilon_{j,d}, \tag{2.1}$$

onde $\bar{T}_{e,j,d}^{A,day}$ é a média diária da temperatura do ar do ambiente, $\bar{T}_{p,j,d}^{A,day}$ é a média diária da temperatura da bateria em uma determinada área A durante o período de um dia e T_0 é a constante de equilíbrio entre as temperaturas. m_j é um coeficiente e $\epsilon_{j,d}$ representa algum distúrbio aleatório, onde d denota o número do dia.

Os resultados foram satisfatórios, mostrando até 0.85 de coeficiente de determinação. Os autores mostraram que é possível estimar a temperatura do ar utilizando leituras de baterias de celulares. Também foi possível observar que existe uma forte correlação entre essas duas fontes de calor. Uma dificuldade observada é relacionada à coleta realizada em ambientes internos, o que observou-se estar relacionado a erros de estimação.

2.3.6 Observações da pressão de superfície em *smartphones*

Mass e Madaus informaram o potencial do uso de *smartphones* para predição do estado climático. Foram citados fatores que propiciam o uso desse tipo de dispositivo, como, por exemplo, a existência de diversos sensores disponíveis e o potencial de cobertura devido à crescente popularização da tecnologia. Os autores focaram em um sensor em específico: o sensor para captura da pressão atmosférica (MASS; MADAUS, 2014).

Estudos recentes, demonstraram que a pressão atmosférica provê informações para estimação de modelos atmosféricos (WHEATLEY; STENSRUD, 2010; MADAUS; HAKIM; MASS, 2014). Um ponto positivo em utilizar-se da pressão atmosférica é que, diferentemente de outras variáveis, tais como, temperatura e umidade, ela não sofre alteração quanto é coletada em ambientes internos ou perto de obstáculos que interferem no resultado e ocasionam erros de medição.

Os autores também informaram que a acurácia e resolução dos sensores de pressão atmosféricas presentes nos *smartphones* atuais têm, atualmente, erro relativo de $0,2$ hPa e erro absoluto de $2,6$ hPa. Encontram-se aplicativos que utilizam esse sensor, como é o caso do aplicativo *pressureNet* (PRESSURENET, 2014) e o aplicativo *OpenSignal*. Esses aplicativos utilizam a pressão atmosférica como uma de outras variáveis para estimar o estado meteorológico.

Foi mostrado que, para observação da pressão atmosférica utilizando *smartphones*, o aparelho não pode estar em movimento. Essa problema pode ser amenizado pela observação da posição fornecida pelo GPS. Porém, a tecnologia utilizada não é precisa em medição de altitude, produzindo até 10 metros de erro, o que é equivalente a 1 hPa de erro de pressão atmosférica.

Em estágio inicial da tecnologia e com poucos usuários, foi feito um experimento para comparação da utilização de *smartphones* com outros modelos de coleta. Os resultados obtidos sugeriram que a utilização das leituras de pressão meteorológicas utilizando *smartphones* podem mudar e melhorar previsões de curto prazo.

2.3.7 Uma abordagem de *crowdsourcing* para estimação de condições meteorológicas

Niforatos et al. mostraram uma nova abordagem para sistemas de coleta de dados que, através de *crowdsourcing*, criam um mapa de contexto com informações meteorológicas utilizando *smartphones* (NIFORATOS et al., 2014).

Foi desenvolvida uma aplicação para *smartphones* para coleta de dados sobre pressão atmosférica, temperatura, umidade e luminosidade. Também são coletados dados de usuários por meio de questionário *online*. Durante oito meses, foi possível coletar 18 mil leituras de sensores e 500 respostas a questionários *online*.

Os primeiros resultados mostraram que a correlação entre temperatura e pressão atmosférica coletada por sensores é significativamente positiva. Por outro lado, não existe correlação significante entre luminosidade e pressão atmosférica.

No questionário *online*, os usuários foram perguntados sobre a influência de condições climáticas em aspectos do cotidiano, percebeu-se três aspectos mais influenciados pelo tempo são as vestimentas, o humor e o meio de transporte. As variáveis mais importantes percebidas pelos usuários são: temperatura, vento e umidade.

2.4 Avaliação dos trabalhos relacionados

Percebeu-se que, na literatura, existem vários estudos sobre redes de comunicação sem fio com o intuito de monitorar eventos meteorológicos. Essas redes constituíam da comunicação entre torres que operavam em frequências acima de 10 Ghz. Alguns estudos como (OTERO; YALAMANCHILI; BRAUN, 2001; FONG et al., 2003; MESSER, 2007) mostraram que precipitações atmosféricas afetam a qualidade do sinal na comunicação sem fio e, em (ZINEVICH; MESSER; ALPERT, 2010), uma análise mais profunda foi feita resultando em modelos de regressão para realizar predições de precipitações atmosféricas com até 90% de acurácia. Porém, sabe-se que, em situações reais, essas redes não são otimizadas para monitoramento de dados referentes à sua comunicação, o que dificulta a sua utilização com o intuito de desenvolver sistemas de alertas para precipitações atmosféricas. Ainda, atualmente, essa tecnologia vem sendo substituída pelo emprego de *links* físicos, como fibras ópticas, que não sofrem influência direta de precipitações atmosféricas. Porém, na topologia dessas redes, existem torres que se comunicam com dispositivos

móveis, tais como *smartphones* e *wearables*, por meio de *links* sem fio. Em (MESSER, 2007) o autor sugere a utilização desses dispositivos para monitoramento de precipitações atmosféricas mas adverte que as frequências mais comuns nesses dispositivos são baixas em relação à comunicação entre torres, o que pode resultar em uma menor atenuação do sinal.

Nota-se que poucos estudos foram realizados a fim de analisar o impacto provocado por precipitações atmosféricas na comunicação entre dispositivos móveis e torres de comunicação sem fio. Por esse motivo, neste trabalho, utilizou-se dados referentes a comunicação entre *smartphones* e torres de comunicação sem fio para verificar a viabilidade de desenvolvimento de sistemas de alertas para precipitações atmosféricas. Além disso, outras variáveis como: dados da comunicação de satélites e a temperatura da bateria de *smartphone* foram utilizadas. Adicionalmente, variáveis alternativas foram apresentadas nos trabalhos correlatos (OVEREEM et al., 2013; MASS; MADAUS, 2014; NIFORATOS et al., 2014), tais como, sensores e informações dos próprios usuários não fizeram parte do escopo deste trabalho de conclusão de curso porém possuem potencial para monitoramento do estado meteorológico.

2.5 Técnicas de mineração de dados

Técnicas de mineração de dados auxiliam na recuperação e extração de informações em uma grande quantidade de dados. Neste trabalho foram utilizadas técnicas de pré-processamento e de aprendizado supervisionado dos dados. Algoritmos para aprendizado supervisionado são aqueles em que são necessárias informações fornecidas por especialistas para um conjunto de dados. Dois tipos distintos de algoritmos podem ser encontradas na literatura para aprendizado supervisionado: algoritmos de classificação e de regressão. Nos algoritmos de classificação, a partir de um conjunto de dados fornecido é gerado um modelo que pode ser utilizado para, posteriormente, classificar instâncias futuras de dados entre tipos especificados previamente na criação do modelo. Exemplos de algoritmos de classificação são: C4.5, *k-Nearest Neighbors* (k-NN), *Multilayer Perceptron* (MLP) e *Support Vector Machine* (SVM). Nos algoritmos de regressão, também é utilizado um conjunto de dados para criação de um modelo. Entretanto, esse tipo de algoritmo preocupa-se em predizer valores contínuos ao invés de encontrar uma classe para uma instância dada. Exemplos de algoritmos de regressão são: regressão linear utilizando mínimos quadrados e SVM (implementação específica para regressão) (WITTEN; FRANK; HALL, 2011).

O algoritmo C4.5 constrói modelos baseados em árvores de decisão. Uma árvore de decisão é construída selecionando, em cada iteração, o atributo e um valor limiar que melhor separam as instâncias do conjunto de dados entre as classes previamente

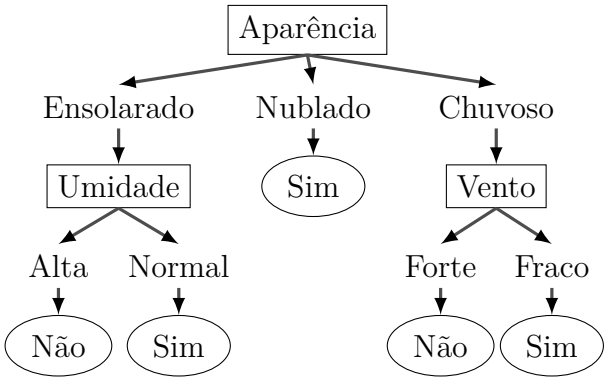

Figura 1 – Árvore de decisão para o problema de jogar tênis: irei ou não jogar tênis? (MICHALSKI; CARBONELL; MITCHELL, 2013)

estabelecidas. Por exemplo, caso a instância contenha o atributo `força de sinal` menor que -80 dBm, essa instância é conduzida ao nó filho da esquerda, caso contrário ao nó filho da direita. Outro exemplo é ilustrado na Figura 1. Para a seleção do atributo e o seu valor limiar, são utilizados conceitos de teoria da informação (MICHALSKI; CARBONELL; MITCHELL, 2013), tal como o ganho de informação, que representa a capacidade de separação das instâncias em classes distintas. Essa é uma técnica caixa-branca, ou seja, é possível explicar de maneira simplificada a classificação de uma instância de acordo com as variáveis envolvidas.

O algoritmo k-NN baseia-se na comparação com instâncias de treino. De maneira simplificada, esse algoritmo armazena uma matriz de distâncias entre os elementos do conjunto de dados. Para a classificação de uma nova instância, o algoritmo a introduz na matriz de distâncias e, observando-se a classe majoritária dos k vizinhos mais próximos, a nova instância é classificada. Esse tipo de classificação é do tipo *lazy* pois, não é construído um modelo explícito com base nas instâncias de exemplo a fim de fazer futuras classificações.

O algoritmo MLP constrói modelos inspirando-se em fenômenos observados em redes neurais naturais. Sabe-se que um neurônio recebe estímulos de outros neurônios por meio de dendritos, processa a informação recebida e transmite uma resposta por meio de impulsos nervosos para outros neurônios por meio de axônios. O algoritmo MLP utiliza modelos computacionais de neurônios denominados *perceptron*. A Figura 2 ilustra as partes de um *perceptron*, onde X_i representa um atributo das instâncias coletadas, W_i um peso associado a X_i e $o(x_1, x_2, \dots)$ o resultado obtido por meio de uma função de ativação. O objetivo do algoritmo é encontrar o conjunto de pesos W_i que minimizam o erro da classificação no conjunto de dados. A cada iteração do algoritmo, todos os elementos do conjunto de dados são introduzidos na rede neural e, ao fim de cada iteração, ajusta-se os pesos. O algoritmo para quando os valores de W_i convergem ou quando se chega a uma taxa erro aceitável. O neurônio encontrado, assim como no algoritmo J48, divide o espaço linearmente, entretanto, misturando-se vários neurônios e utilizando funções de ativação

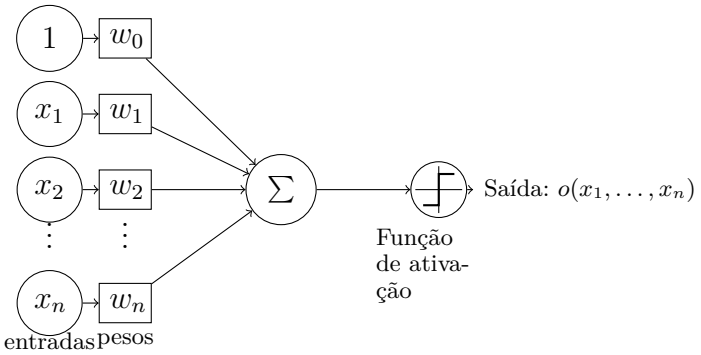

Figura 2 – Modelo de neurônio *perceptron*

não linear, por exemplo, $\frac{1}{1+e^{-x}}$, é possível separar classes não-linearmente separáveis.

O algoritmo SVM pode ser interpretado como uma mistura de técnicas presentes no k-NN e na MLP pois busca a semelhança entre instâncias estratégicas, denominados vetores-suporte, e utiliza funções *kernel* para melhorar a capacidade de separabilidade como as funções de ativação não lineares similares às da MLP. Diferentemente do algoritmo k-NN, a SVM utiliza somente as instâncias de exemplo que melhor descrevem a fronteira que representa a passagem de uma classe para outra como, por exemplo, os círculos destacados são utilizados para definir fronteiras entre os círculos cheios ou vazios na Figura 3. Essas instâncias são chamadas de vetores de suporte e, para serem encontradas, o algoritmo utiliza técnicas de otimização para encontrar a maior margem que separa duas classes. A função utilizada na classificação de novas instâncias é chamada de núcleo (*kernel*) e pode ter diferentes formas, tais como, linear, polinomial ou radial. Utilizando-se núcleos não-lineares é possível a separação de classes não-linearmente separáveis assim como o algoritmo MLP. A Figura 3 ilustra a margem encontrada utilizando a maximização de $\frac{2}{||w||}$ que representa a largura da margem que separa duas classes. Onde w é o vetor da margem, x_i as dimensões relacionadas aos atributos de cada instância e b uma variável de deslocamento que compõe a margem. O algoritmo tem como objetivo encontrar w tal que a função $\frac{2}{||w||}$ deve ser maximizada utilizando como restrições as funções $w \cdot x_i - b = 1$ e $w \cdot x_i - b = -1$ representando os vetores das bordas da margem e w será composto de uma combinação desses vetores-suporte. Os pontos que tocam as bordas da margem são os vetores de suporte utilizados para classificação. Algumas implementações do algoritmo SVM utilizam os vetores de suporte para regressão utilizando ponderações entre a instância a ser classificada e os vetores de suporte juntamente com suas distâncias.

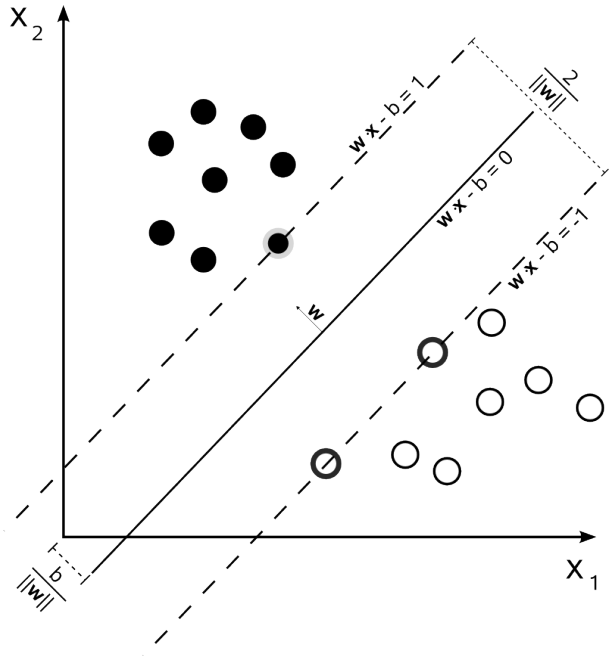

Figura 3 – Margem máxima do algoritmo SVM
Fonte: *Wikipedia - Support vector machine*

3 Desenvolvimento

Foi considerada uma arquitetura de *feedback* para o ciclo dos dados envolvidos, como é mostrado na Figura 4. A arquitetura foi definida levando em consideração aspectos como: coleta de dados e pré-processamento de dados; análise de dados utilizando algoritmos de mineração de dados, tais como, classificação e regressão; previsão de curto prazo de precipitações atmosféricas; e desenvolvimento de aplicação *web* utilizando técnicas de ludificação e *crowdsourcing*.

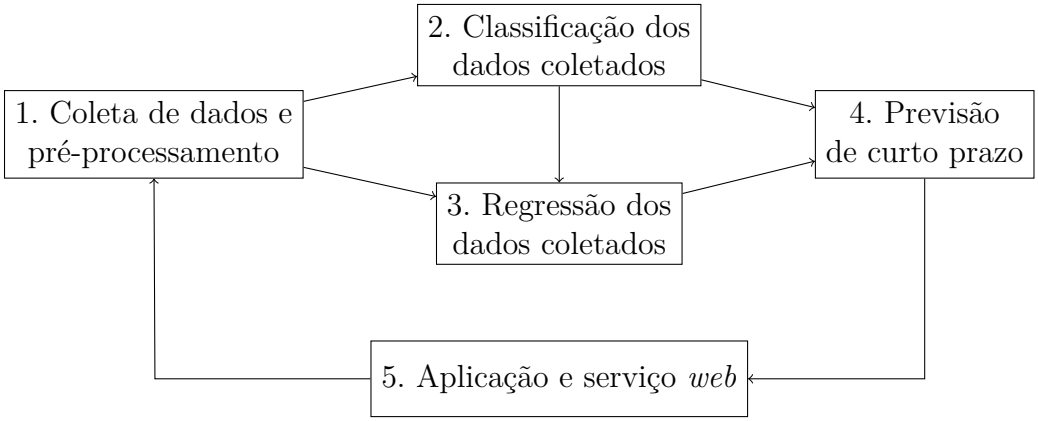

Figura 4 – Diagrama do fluxo de dados e informação

O primeiro bloco da Figura 4 representa o estágio de coleta e pré-processamento dos dados, no qual um protótipo de aplicação *mobile* coleta dados relacionados à tecnologias de telecomunicação. Além disso, a temperatura da bateria também foi coletada, pois, conforme descrito em (OVEREEM et al., 2013), existe correlação entre a temperatura de bateria do *smartphone* e a temperatura do ambiente.

O segundo e terceiro blocos da Figura 4 representam o estágio de mineração de dados. Nesse estágio, algoritmos para classificação e regressão utilizam os dados obtidos do primeiro estágio para estimar o estado atual de precipitação atmosférica. Na seção seguinte são descritos dois experimentos em relação a esse estágio. O objetivo do primeiro experimento foi aplicar algoritmos de classificação para identificar o tipo de precipitação sob a influência de vários fatores. O objetivo do segundo experimento foi estimar o volume de saída de precipitações utilizando algoritmos de regressão.

Os blocos 1, 2 e 3 foram desenvolvidos durante este trabalho. Os demais blocos, 4 e 5, não fazem parte do escopo deste trabalho de conclusão de curso, porém a seguir os estágios que completam a arquitetura definida são descritos. O quarto bloco da Figura 4 representa o estágio de previsão de curto prazo de precipitações atmosféricas. Nesse estágio, as informações obtidas no segundo e terceiro estágios, juntamente com a localização dos usuários, serão utilizadas por um algoritmo para realizar previsões de curto prazo. O

quinto bloco representa o estágio da continuação do ciclo do diagrama de *feedback*. Nesse estágio, uma aplicação *mobile* será desenvolvida com a intenção de apresentar informações relevantes para usuários e coletar dados em sua localização.

Nas próximas seções, serão descritos experimentos acerca dos primeiros três estágios. Para realizar esses experimentos, um simulador de precipitações atmosféricas simplificado foi construído para coletar dados. O simulador ajudou isolar fatores como tipo de precipitação, variação do sinal de acordo com o dia e manter fixado a fonte de precipitação e o dispositivo móvel. O simulador também proporcionou um modo para verificação da influência de precipitação artificial sobre os dados coletados pelo aplicativo para dispositivo móvel em um ambiente controlado.

O simulador de precipitações é mostrado nas Figuras 5 e 6. Ele consiste de um tripé que suporta o dispositivo móvel e uma mangueira apontada para o mesmo. A mangueira é equipada com vários tipos de jatos representando tipos distintos de precipitação. Uma pequena rede ampara o dispositivo móvel no centro do tripé direcionando-o no caminho da água. Coletou-se leituras das seguintes variáveis: força de sinal GSM, acurácia de geolocalização, número de satélites GPS disponíveis, média e desvio padrão das forças de sinais de satélites GPS e temperatura de bateria. Após a coleta, as leituras foram normalizadas subtraindo a média dos valores encontrados para cada variável e dividindo pelo desvio padrão.

Figura 5 – Simulador de chuvas

3.1 Experimentos de classificação

O objetivo dos primeiros experimentos foi utilizar algoritmos de classificação para identificar tipos de precipitação e avaliar a influência de três tipos de fatores experimentais. O primeiro fator foi o tipo de saída de água do simulador de precipitações, podendo ser um dos três níveis: alto, baixo ou sem precipitação. Para considerar a forma e a intensidade das gotas de água, utilizou-se duas distâncias diferentes entre o dispositivo móvel e a fonte

Figura 6 – Simulador de chuvas e dispositivo móvel posicionado

de água: 40 cm e 80 cm. O último fator considerou a existência de uma cobertura sobre o *smartphone* evitando contato direto com as gotas de água. Combinando esses três fatores, foi possível gerar 12 configurações distintas e, ao todo, foram coletadas 6 883 instâncias.

Tabela 1 – Fatores considerados para experimento de classificação

Fator 1	Fator 2	Fator 3
precipitação alta	40 cm	com cobertura
precipitação baixa	80 cm	sem cobertura
sem precipitação		

Os dados coletados foram utilizados em duas análises. Na primeira análise, agrupou-se as instâncias baseando-se somente no tipo de saída de água. Em outras palavras, houveram três classes diferentes para identificação: precipitação alta, precipitação baixa e sem precipitação. Na segunda análise, agrupou-se as instâncias baseado em todos os fatores. Ou seja, utilizou-se cada uma das configuração, totalizando 12 classes diferentes.

Os algoritmos de classificação C4.5, k-NN com $k = 5$ e SVM com núcleo radial, utilizando o algoritmo C-SVC (*Support Vector Classification*), e MLP providos pelo *software Weka* (WAIKATO, 2014), foram executados com parametrização padrão nos dados organizados com relação a cada análise. Também utilizou-se a estratégia de validação cruzada com 10 partes em cada método.

3.2 Experimentos de regressão

O objetivo desses experimentos foi utilizar algoritmos de regressão para estimar a intensidade de precipitações atmosféricas. Dois fatores experimentais foram considerados: o primeiro fator compreendeu dois tipos de saída de água, precipitação alta e baixa. O se-

gundo fator mensurou o volume de saída de água, coletou-se dados com relação a 5 níveis de abertura da torneira: 0% (completamente fechado), 25%, 50%, 75% e 100% (completamente aberto). Um hidrômetro foi conectado na saída da torneira e forneceu medidas para regressão. Em comparação com o primeiro experimento, a razão para não considerar a terceira opção, sem precipitação, é porque o segundo fator foi projetado contemplando esses fins. Os dois fatores combinados resultaram em 10 configurações diferentes. Assim como, além dos dados obtidos no primeiro experimento, coletou-se a temperatura de bateria do *smartphone* e, ao todo, foram coletadas 5 661 instâncias.

Tabela 2 – Fatores considerados para experimento de regressão

Fator 1	Fator 2
precipitação alta	0%
precipitação baixa	25%
	50%
	75%
	100%

Dois modelos de regressão foram utilizados com base no algoritmo dos mínimos quadrados (GELADI; KOWALSKI, 1986): modelo linear, $f(x_1, x_2, \ldots, x_n) = (ax_1 + bx_2 + \ldots)$, e quadrático, $f(x_1, x_2, \ldots, x_n) = (ax_1 + bx_2 + \ldots)^2$. O terceiro modelo foi criado baseando-se na implementação libSVM do algoritmo SVM (CORTES; VAPNIK, 1995). Esses modelos foram utilizados para avaliar a viabilidade de estimação da intensidade de precipitação. Depois, foram feitos testes para verificar a importância de cada variável coletada.

3.3 Resultados dos experimentos de classificação

Na primeira análise, onde os dados foram agrupados em três classes, o algoritmo C4.5 alcançou os melhores resultados. O algoritmo obteve taxa de recuperação de 0.86, que indica a fração de instâncias classificadas corretamente de acordo com cada classe, e coeficiente Kappa de 0.79, que representa a concordância entre os resultados obtidos pelas partes da validação cruzada. A área abaixo da curva *Receiver Operating Characteristic* (ROC) dos algoritmos k-NN e C4.5 foram 0.95 e 0.97, o que pode sugerir que essas técnicas sobreajustaram-se aos dados. Entretanto, o algoritmo SVM com núcleo radial obteve área ROC de 0.87 indicando um não-sobreajustamento e performance de classificação razoável. A matriz de confusão dos resultados obtidos com o algoritmo SVM são mostrados na Tabela 4, na qual as linhas indicam as classes corretas e as colunas as classes preditas. Observou-se que as classes preditas mais corretamente foram **sem precipitação** e **pre-**

cipitação alta. O resultado foi considerado relevante pelo fato que precipitação alta é o cenário mais importante para um sistema de alerta.

Tabela 3 – Resultados de classificação com relação ao cenário com 3 classes

Algoritmo	Coeficiente Kappa	Taxa de recuperação •	Área ROC •
C4.5	0.79	0.86	0.97
k-NN	0.76	0.84	0.95
SVM	0.60	0.73	0.80
MLP	0.52	0.68	0.82

• Média ponderada

Tabela 4 – Matriz de confusão com relação ao cenário com 3 classes e modelo SVM

		predito como		
		sem precipitação	precipitação baixa	precipitação alta
classes	sem precipitação	2 013	165	95
	precipitação baixa	962	1 208	86
	precipitação alta	435	78	1 813

No segundo cenário, considerou-se todas as combinações de fatores como uma classe, o que resultou em 12 classes. O desempenho do algoritmo SVM melhorou ligeiramente a medida de área ROC para 0.85, por outro lado, a taxa de recuperação diminuiu ligeiramente para 0.72. Observando a matriz de confusão das predições do algoritmo SVM na Tabela 6, observou-se que a maior parte dos erros aconteceram nas classes sem precipitação e precipitação baixa, indicadas pelas linhas de **a** a **g**. Enquanto que, para a classe mais importante, precipitação alta, representadas pelas linhas de **i** a **l**, poucos erros predição foram encontrados. Esses erros ocorreram devido à não proporcionalidade da interferência da chuva nas medidas mais perceptíveis no cenário de precipitação alta. Observa-se que as coletas dos dados nas configurações de **a** a **l** foram realizadas em ordem aleatória, e dessa forma, excluindo a possibilidade de correlação ou erros de predição sequencial na classificação de **a** a **g**.

Tabela 5 – Resultados de classificação com relação ao cenário com 12 classes

Algoritmo	Coeficiente Kappa	Taxa de recuperação •	Área ROC •
C4.5	0.80	0.82	0.98
k-NN	0.74	0.76	0.97
SVM	0.70	0.72	0.85
MLP	0.60	0.63	0.93

• Média ponderada

Tabela 6 – Matriz de confusão com relação ao cenário com 12 classes e modelo SVM

		predito como											
		a	b	c	d	e	f	g	h	i	j	k	l
classes	a	363	66	2	10	67	0	27	0	0	31	0	0
	b	65	327	14	32	66	2	53	0	0	13	0	0
	c	0	6	391	24	2	46	17	0	0	0	80	0
	d	14	30	79	307	33	8	49	0	0	2	47	0
	e	63	61	25	54	281	2	48	0	0	21	4	0
	f	19	9	39	22	0	413	16	0	0	9	37	0
	g	52	49	14	44	40	6	335	0	0	19	10	0
	h	0	0	0	0	0	0	0	558	0	0	0	6
	i	0	0	0	0	0	0	0	0	616	0	0	0
	j	48	44	0	14	20	0	5	0	0	438	0	0
	k	1	3	98	28	4	35	14	0	0	0	391	0
	l	0	0	0	0	0	2	0	2	0	0	0	563

Cada configuração de **a** a **l** são explicadas na Tabela 7

Tabela 7 – Configurações para cada combinação de fatores utilizados no experimento com 12 classes

Letra	Fator 1	Fator 2	Fator 3
a	sem precipitação	40 cm	com cobertura
b	sem precipitação	80 cm	sem cobertura
c	sem precipitação	40 cm	sem cobertura
d	sem precipitação	80 cm	com cobertura
e	precipitação baixa	80 cm	com cobertura
f	precipitação baixa	40 cm	sem cobertura
g	precipitação baixa	80 cm	sem cobertura
h	precipitação baixa	40 cm	com cobertura
i	precipitação alta	40 cm	com cobertura
j	precipitação alta	80 cm	sem cobertura
k	precipitação alta	80 cm	com cobertura
l	precipitação alta	40 cm	sem cobertura

3.4 Resultados dos experimentos de regressão

Nesses experimentos foram construídos modelos de regressão para predizer a intensidade de chuva. Foram utilizados o algoritmo dos mínimos quadrados com modelos lineares e quadráticas e o algoritmo SVM. Como pode ser visto na Tabela 8, o modelo construído a partir da equação linear das medidas coletadas produziu resultados de baixa qualidade. Obteve-se a medida *Root Mean Squared Error* (RMSE) de 29.95 em um intervalo de valores para predição de $[0, 120]$ ml/s. O alto valor para RMSE foi atribuído à falta de consideração da interação entre variáveis. Além disso, o coeficiente de determinação R^2 foi 0.17, indicando um modelo não apropriado. O modelo construído a partir da equação quadrática obteve melhores resultados indicados por coeficiente de determinação de

0.41. Entretanto, esse modelo obteve as medidas *Mean Absolute Error* (MAE) de 17.54 e RMSE de 25.21 caracterizando predições de baixa qualidade. O modelo construído a partir do algoritmo SVM resultou em melhores resultados com coeficiente de determinação $R^2 = 0.96$ e erros de predição MAE = 4.64 e RMSE = 6.10.

Tabela 8 – Resultado de modelos de regressão

Modelo	R^2	MAE	$RMSE$
linear	0.17	20.49	29.95
quadrático	0.41	17.54	25.21
SVM	0.96	4.64	6.10

Entretanto, mesmo considerando o modelo construído a partir do algoritmo SVM, alguns problemas foram encontrados. Nota-se que os resíduos encontrados são irregulares para precipitações de alta intensidade sendo a maior fonte de erro. Uma razão possível para esses resultados pode ser o fato de que o conjunto de dados coletados continha poucas instâncias de precipitações de alta intensidade quando comparado com o resto dos cenários. As Figuras 7 e 8 mostram os valores preditos e os resíduos encontrados utilizando o modelo construído a partir da equação quadrática e o modelo construído a partir do algoritmo SVM. Nos dois casos existem tendências de resíduos com relação aos valores preditos e, especialmente no modelo SVM, existe correlação linear entre os valores preditos e os erros residuais.

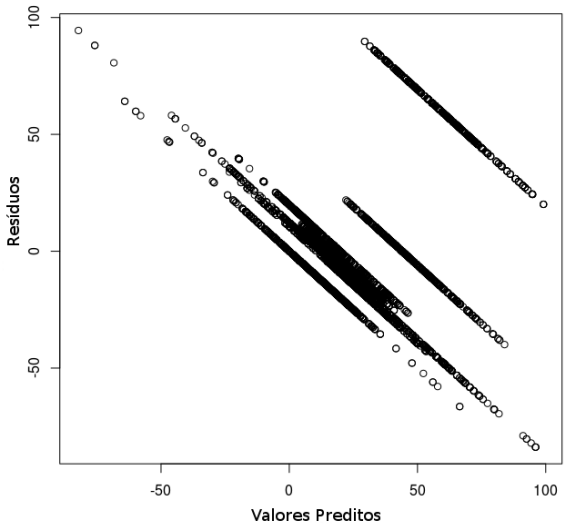

Figura 7 – Valores preditos e resíduos do modelo de regressão quadrático

Como mostrado nas Figuras 9 e 10, os gráficos apresentam os erros residuais dos modelos quadrático e SVM e, nos dois gráficos, é apresentado distribuição não normal. No modelo baseado no algoritmo SVM, mostrado na Figura 10, o gráfico mostra resíduos com assimetria negativa e bimodalidade, o que sugere uma capacidade fraca para predição.

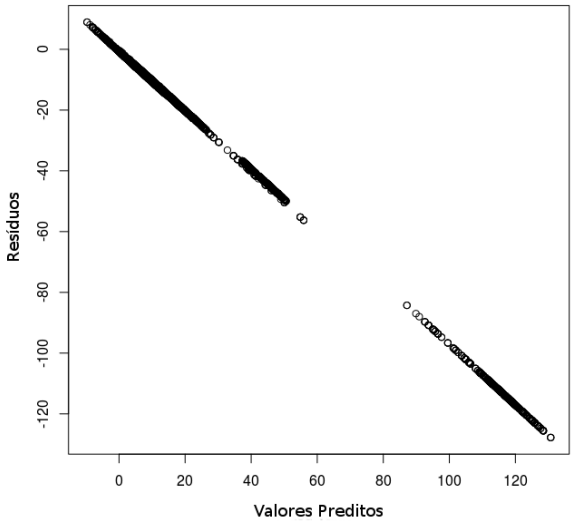

Figura 8 – Valores preditos e resíduos do modelo de regressão SVM

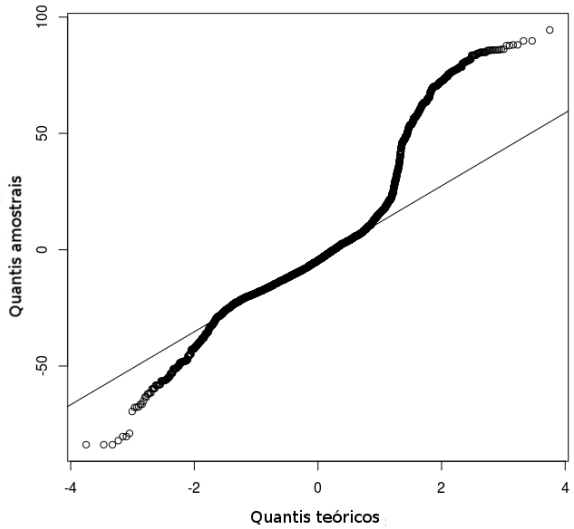

Figura 9 – Gráfico quantil-quantil do modelo de regressão quadrático

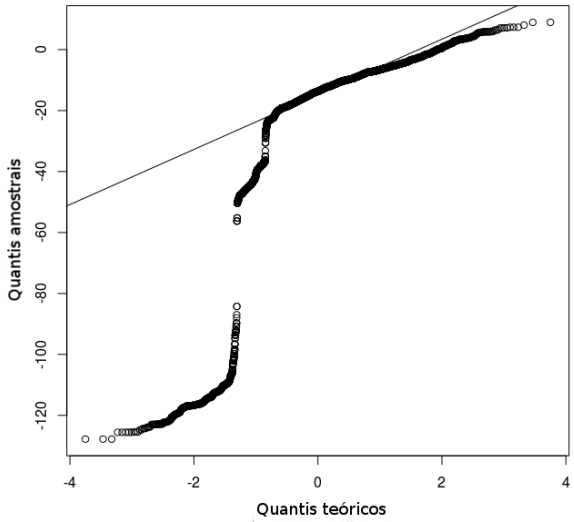

Figura 10 – Gráfico quantil-quantil do modelo de regressão SVM

3.5 Estudo da influência das medidas utilizadas nos modelos de regressão

Para verificar a influências das medidas coletadas para dos modelos de regressão para intensidade de precipitação, construiu-se novos modelos utilizando o algoritmo SVM com uma medida por vez e, também, sua interação com outra variável. A Tabela 9 mostra o coeficiente determinante R^2 para cada modelo, nota-se que a força de sinal GSM e temperatura da bateria foram as principais medidas para predição da intensidade de precipitação. As duas variáveis obtiveram melhores resultados do coeficiente determinante quando sozinhas ou quando interagindo com outras medidas. Sugere-se que a temperatura da bateria foi importante para o modelo de regressão devido a influência da água sobre a temperatura em volta do *smartphone* que, por sua vez, reduz a temperatura da bateria. A interação da força de sinal GSM (`sinal GSM`) e da temperatura da bateria (`TEMP`) forneceu o melhor modelo considerando apenas duas medidas com coeficiente determinante de 0.92. As outras medidas como acurácia da geolocalização (`GEOACC`), número de satélites GPS disponíveis (`nSAT`), média e desvio padrão das forças de sinais dos satélites GPS (média and desvio `SNR`) não tiveram relevância quando sozinhas porém, quando interagindo com outras medidas, elas melhoraram o coeficiente ligeiramente dos modelos SVM. Por exemplo, a acurácia de geolocalização e média das forças de sinais de satélites GPS disponíveis melhoraram o coeficiente de determinação quando houve interação com a força de sinal GSM. Similarmente, o desvio padrão das forças de sinais de satélites GPS disponíveis interagindo com a temperatura da bateria também melhorou o coeficiente de determinação.

Considerando que o melhor modelo encontrado para duas variáveis utiliza a força

Tabela 9 – Valores de coeficiente de determinação de modelos de regressão SVM para interação de até duas medidas

	GEOACC	sinal GSM	nSat	média SNR	desvio SNR	TEMP
GEOACC	0.04	0.58	0.07	0.17	0.09	0.19
sinal GSM		0.30	0.49	0.34	0.55	0.92
nSat			−0.09	0.03	0.19	0.50
média SNR				−0.04	0.08	0.73
desvio SNR					−0.03	0.55
TEMP						0.36

do sinal GSM e a temperatura a bateria do dispositivo móvel, criou-se novos modelos adicionando-se uma medida por vez para observar qual a próxima melhor variável. A Tabela 10 mostra os resultados obtidos desse experimento. Notou-se que a acurácia de geolocalização foi a medida que mais melhorou o modelo SVM, obtendo o valor 0.92 para R^2. Todas as outras medidas melhoraram ligeiramente o modelo, o que sugere que todas as variáveis podem ser utilizadas para a criação do modelo de regressão.

Tabela 10 – Avaliação de modelos de regressão SVM com mais de 2 medidas

Medidas	R^2
sinal GSM, TEMP e GEOACC	0.96
sinal GSM, TEMP e nSat	0.93
sinal GSM, TEMP e média SNR	0.93
sinal GSM, TEMP e padrão SNR	0.93

4 Conclusão

O objetivo deste trabalho de conclusão de curso foi estudar e implementar de técnicas de mineração de dados que, a partir da coleta de dados de redes de comunicação de dispositivos móveis, comprovem a capacidade de estimação do estado meteorológico. A fim de cumprir esse objetivo, utilizou-se técnicas de mineração de dados a fim de estudar a viabilidade do uso dados coletados em *smartphones* para identificação de tipos de precipitações e estimar a intensidade da precipitação em ambiente controlado.

O Objetivo Específico 1 referente à coleta de dados utilizando dispositivos móveis de ambientes controlados foi contemplado pela construção de um simulador de precipitações atmosféricas simples que possibilitou a reprodução de vários cenários em ambiente controlado. Por meio de um *smartphone*, coletou-se leituras utilizando uma aplicação desenvolvida. Ao todo, foram coletadas 6 883 instâncias para os experimentos de classificação e 5 661 instâncias para regressão.

O Objetivo Específico 2 referente aos estudos de técnicas mais adequadas para extrair informações meteorológicas das medidas coletadas foi contemplado utilizando alguns algoritmos para classificação e regressão. De acordo com os resultados, observou-se a possibilidade de identificação de precipitação em três tipos: precipitação alta, precipitação baixa e sem precipitação. O modelo SVM obteve performance de classificação razoável nas análises atingindo área abaixo da curva ROC de até 0.80. Observou-se que as classes preditas com mais acertos foi sem precipitação e precipitação alta, o cenário mais importante para um sistema de alerta. A maioria dos erros no processo de classificação foram atribuídos à classe de baixa intensidade de precipitação. Suspeita-se que esses erros são devido uma interferência não proporcional da precipitação nas medidas coletadas, as quais foram perceptíveis para o cenário da classe de intensidade alta de precipitação. Também foi observado que os modelos de regressão construídos para estimação da intensidade de precipitação com os dados coletados não foram suficientemente adequados. Os modelos baseados nos mínimos quadrados utilizando equações linear e quadrática obtiveram resultados ruins na predição. Considerando todas as medidas coletadas do *smartphone*, o modelo de regressão baseado no algoritmo SVM obteve coeficiente determinante de 0.96 e erro RMSE de 6.10, entretanto os resíduos encontrados apresentam correlação linear com os valores preditos. Uma possível razão pode ser o fato de que o conjunto de dados coletados continham poucas instâncias representando o cenário de alta precipitação com relação ao resto dos cenários.

Adicionalmente, para verificar a influência das medidas coletadas para experimentos de regressão, construiu-se modelos baseados no algoritmo SVM utilizando diferentes

combinações de medidas. Os resultados obtidos sugerem que a força de sinal GSM e a temperatura da bateria são as medidas mais relevantes para estimação da intensidade precipitação. Outras medidas tais como acurácia de geolocalização, número de satélites GPS disponíveis e outras medidas com relação à força de sinal dos satélites melhoraram o modelo. Esse experimento mostrou que todas as medidas utilizadas para estimação da intensidade de precipitação foram relevantes na análise de regressão.

Sugere-se que estudos devem ainda ser realizados a fim de melhorar a qualidade da regressão incluindo conhecimento de domínio no modelo, além de coletas outras variáveis. Além disso, os experimentos de classificação e regressão devem ser avaliados em um ambiente menos controlado. Para os próximos estudos, será desenvolvido uma aplicação móvel para coletar dados de usuários reais. A aplicação permitirá coletar dados sobre o tempo tais como percepção do usuário sobre o tempo e outras variáveis com informações indiretas tais como humor, vestimenta e meio de transporte do usuário para o dia. Antecipa-se que a coleta de dados de dispositivos de usuários reais podem trazer novos desafios, tais como tendências variáveis relacionadas aonde o dispositivo móvel está (dentro ou fora de uma bolsa) e como ele está sendo utilizado (se está carregado ou está durante uma ligação dentro de um prédio).

Referências

ATLAS, D.; ULBRICH, C. W. Path-and area-integrated rainfall measurement by microwave attenuation in the 1-3 cm band. *Journal of Applied Meteorology*, v. 16, n. 12, p. 1322–1331, 1977. Citado na página 14.

AYDIN, K.; DAISLEY, S. E. Relationships between rainfall rate and 35-ghz attenuation and differential attenuation: modeling the effects of raindrop size distribution, canting, and oscillation. *Geoscience and Remote Sensing, IEEE Transactions on*, IEEE, v. 40, n. 11, p. 2343–2352, 2002. Citado na página 15.

BERNE, A.; UIJLENHOET, R. Path-averaged rainfall estimation using microwave links: Uncertainty due to spatial rainfall variability. *Geophysical research letters*, Wiley Online Library, v. 34, n. 7, 2007. Citado na página 15.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 25.

DAVID, N.; ALPERT, P.; MESSER, H. Technical note: Novel method for water vapour monitoring using wireless communication networks measurements. *Atmospheric Chemistry and Physics*, Copernicus GmbH, v. 9, n. 7, p. 2413–2418, 2009. Citado na página 15.

EBERT, E. E.; JANOWIAK, J. E.; KIDD, C. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bulletin of the American Meteorological Society*, v. 88, n. 1, p. 47–64, 2007. Citado na página 10.

FONG, B. et al. Factors causing uncertainties in outdoor wireless wearable communications. *Pervasive Computing, IEEE*, IEEE, v. 2, n. 2, p. 16–19, 2003. Citado 2 vezes nas páginas 13 e 17.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Analytica chimica acta*, Elsevier, v. 185, p. 1–17, 1986. Citado na página 25.

GRIFFITH, C. G. et al. Rain estimation from geosynchronous satellite imagery-visible and infrared studies. *Monthly Weather Review*, v. 106, n. 8, p. 1153–1171, 1978. Citado na página 10.

HOLT, A.; KUZNETSOV, G.; RAHIMI, A. Comparison of the use of dual-frequency and single-frequency attenuation for the measurement of path-averaged rainfall along a microwave link. *IEE Proceedings-Microwaves, Antennas and Propagation*, IET, v. 150, n. 5, p. 315–320, 2003. Citado na página 15.

LEIJNSE, H.; UIJLENHOET, R.; HOLLEMAN, I. Sources of error in microwave link rainfall estimation. In: *International Symposium on Weather Radar and Hydrology, Grenoble, France*. [S.l.: s.n.], 2008b. Citado na página 15.

LEIJNSE, H.; UIJLENHOET, R.; STRICKER, J. Hydrometeorological application of a microwave link: 2. precipitation. *Water resources research*, Wiley Online Library, v. 43, n. 4, 2007a. Citado na página 15.

LEIJNSE, H.; UIJLENHOET, R.; STRICKER, J. Microwave link rainfall estimation: Effects of link length and frequency, temporal sampling, power resolution, and wet antenna attenuation. *Advances in Water Resources*, Elsevier, v. 31, n. 11, p. 1481–1493, 2008a. Citado na página 15.

MADAUS, L. E.; HAKIM, G. J.; MASS, C. F. Utility of dense pressure observations for improving mesoscale analyses and forecasts. *Monthly Weather Review*, n. 2014, 2014. Citado na página 16.

MASS, C. F.; MADAUS, L. E. Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bulletin of the American Meteorological Society*, 2014. Citado 2 vezes nas páginas 16 e 18.

MEISCHNER, P. *Weather radar: principles and advanced applications*. [S.l.]: Springer Science & Business Media, 2004. Citado na página 10.

MESSER, H. Rainfall monitoring using cellular networks. *IEEE Signal Proc. Mag*, v. 24, 2007. Citado 4 vezes nas páginas 13, 14, 17 e 18.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 19.

MINDA, H.; NAKAMURA, K. High temporal resolution path-average rain gauge with 50-ghz band microwave. *Journal of Atmospheric and Oceanic Technology*, v. 22, n. 2, p. 165–179, 2005. Citado na página 15.

NIFORATOS, E. et al. Atmos: a hybrid crowdsourcing approach to weather estimation. In: ACM. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. [S.l.], 2014. p. 135–138. Citado 2 vezes nas páginas 17 e 18.

OPENSIGNAL. *OpenSignal: 3G and 4G LTE Cell Coverage Map*. 2014. <http://opensignal.com>. [Online; acessado em 16/11/2014]. Citado na página 15.

OTERO, J.; YALAMANCHILI, P.; BRAUN, H.-W. High performance wireless networking and weather. *High Performance Wireless Research and Education Network*, 2001. Citado 2 vezes nas páginas 13 e 17.

OVEREEM, A. et al. Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophysical Research Letters*, Wiley Online Library, v. 40, n. 15, p. 4081–4085, 2013. Citado 4 vezes nas páginas 10, 15, 18 e 22.

PRESSURENET. *The Weather's Future*. 2014. [Online; acessado em 16/11/2014]. Disponível em: <http://pressurenet.io/>. Citado na página 16.

RAHIMI, A. et al. Use of dual-frequency microwave links for measuring path-averaged rainfall. *Journal of Geophysical Research: Atmospheres (1984–2012)*, Wiley Online Library, v. 108, n. D15, 2003. Citado na página 15.

RAMIREZ, M. C. V.; VELHO, H. F. de C.; FERREIRA, N. J. Artificial neural network technique for rainfall forecasting applied to the sao paulo region. *Journal of Hydrology*, Elsevier, v. 301, n. 1, p. 146–162, 2005. Citado na página 10.

RINCON, R. F.; LANG, R. H. Microwave link dual-wavelength measurements of path-average attenuation for the estimation of drop size distributions and rainfall. *Geoscience and Remote Sensing, IEEE Transactions on*, IEEE, v. 40, n. 4, p. 760–770, 2002. Citado na página 15.

SILVA, P. B. *Portaria nº 14, de 6 de Fevereiro de 2013*. 2014. <http://www.mc.gov.br/documentos/Portaria/Portaria_700_versao_final_faixa_700_megaherz.pdf>. Citado na página 12.

TELECO, I. e. T. *4G: Frequências e licitações*. 2014. <http://www.teleco.com.br/4g_freq.asp>. [Online; acessedo em 16/11/2014]. Citado na página 12.

TELECO, I. e. T. *Áreas e Bandas de Celular*. 2014. <http://www.teleco.com.br/bandac.asp>. [Online; acessado em 16/11/2014]. Citado na página 12.

WAIKATO, M. L. G. at the University of. *Weka 3: Data Mining Software in Java*. 2014. [Online; acessado em 16/11/2014]. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Citado na página 24.

WHEATLEY, D. M.; STENSRUD, D. J. The impact of assimilating surface pressure observations on severe weather events in a wrf mesoscale ensemble system. *Monthly Weather Review*, v. 138, n. 5, p. 1673–1694, 2010. Citado na página 16.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560. Citado na página 18.

YI, W.-J.; JIA, W.; SANIIE, J. Mobile sensor data collector using android smartphone. In: IEEE. *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on*. [S.l.], 2012. p. 956–959. Citado na página 10.

ZINEVICH, A.; MESSER, H.; ALPERT, P. Prediction of rainfall intensity measurement errors using commercial microwave communication links. *Atmospheric Measurement Techniques*, Copernicus GmbH, v. 3, n. 5, p. 1385–1402, 2010. Citado 5 vezes nas páginas 10, 13, 14, 15 e 17.