FUNDAÇÃO CENTRO DE ANÁLISE, PESQUISA E INOVAÇÃO TECNOLÓGICA
FACULDADE FUCAPI (INSTITUTO DE ENSINO SUPERIOR FUCAPI)
COORDENAÇÃO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uso de técnicas de *Data Mining* na classificação de tipos de dengue

Jeferson Barros Alves

Manaus - AM
Novembro de 2015

Jeferson Barros Alves

Uso de técnicas de *Data Mining* na classificação de tipos de dengue

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Faculdade Fucapi (Instituto de Ensino Superior Fucapi), como requisito parcial para a obtenção do Título de Bacharel em Ciência da Computação. Área de concentração: Banco de Dados.

Orientador: Márcio Palheta Piedade, M.Sc.

Manaus - AM

Novembro de 2015

Resumo

A tarefa de classificação em *Data Mining* consiste na predição de classe dado um determinado conjunto de atributos de uma instância. Esta técnica pode ser útil em auxílio ao diagnóstico médico, como ferramenta de apoio ao profissional da saúde. As melhores técnicas para esta tarefa na literatura, em geral, usam máquinas de vetores de suporte e árvore de decisão. Nestas técnicas é utilizado o conceito de hiperplano para separa de uma melhor forma as classes existentes num conjunto de dados, também, pode-se verificar o uso da ideia de dividir para conquistar na identificação dos melhores atributos capazes de resultados mais eficientes no processo de predição. Como resultado, obtivemos modelos classificadores, baseados em SMO e J48, que alcançaram resultados de 89,61% e 86,63% respectivamente no conjunto de dados utilizado neste trabalho. Também observamos que os modelos gerados obtiveram melhores resultados à medida que os experimentos no tamanho do conjunto de dados aumentava.

Palavras-chave: Data Mining, Classificação, Árvore de Decisão, Bayes, Support Vector Machine.

Abstract

This is the english abstract.

Keywords: Data mining, Classifier, Decision Tree, Bayes, Support Vector Machine.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

KDD	Knowledge Discovery in Databases
WEKA	Waikato Environment for Knowledge Analysis
SVM	Support Vector Machine
SMO	Sequential Minimal Optimization
SINAN	Sistema de Informação de Agravos de Notificação
SUS	Sistema Único de Saúde

Sumário

1 Introdução

Com a grande quantidade de informações geradas e armazenadas por meio do avanço tecnológico das últimas décadas, houve um acúmulo gigantesco de informações. Com isso surgiu uma abordagem de técnicas e ferramentas que buscam transformar dados, denominada Descoberta de Conhecimento em Base de Dados (knowledge Discovery in Databases - KDD), que foi proposto em 1989 com o objetivo de analisar os dados de uma base para que de alguma forma possa ser extraído conhecimento útil (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A computação tem apoiado o desenvolvimento da medicina em diversas áreas como em sistemas de apoio a coleta de dados clínicos e exames por imagens, na organização das informações obtidas, entre outras. Desta maneira, esse grande volume de dados é uma valiosa fonte de conhecimento que pode ser utilizada para o auxílio ao diagnóstico médico. Assim, é grande a importância do desenvolvimento de técnicas que permitam a descoberta de conhecimento em base de dados médicos para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e eficiência dos diagnósticos elaborados pelo especialista (COSTA, 2012).

1.1 Especificação do problema

Ao longo dos últimos anos a mineração de dados tem sido cada vez mais utilizada na literatura médica. No entanto, a sua aplicação à analise de dados médicos tem sido relativamente limitada, tendo em vista que a aplicação prática pode explorar o conhecimento disponível no contexto clínico e explicar decisões propostas, uma vez que os modelos são utilizados para apoiar decisões (BELLAZZI; ZUPAN, 2008).

É grande a importância do desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases de dados médicos para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e eficiência dos diagnósticos elaborados pelo especialista (COSTA, 2012).

A dengue é hoje uma das doenças com maior incidência no Brasil, atingindo a população de todos os estados, independente de classe social. Nesse cenário, torna-se imperioso que um conjunto de ações para a prevenção da doença seja intensificado, permitindo assim a identificação precoce dos casos de dengue, a tomada de decisões e a implementação de medidas de maneira oportuna a fim de principalmente evitar óbitos. Preservar a vida humana é obrigação de todos.

A classificação epidemiológica dos casos de dengue, que é feita habitualmente após desfecho clínico, na maioria das vezes é restrospectiva e depende de informações clínicas e laboratoriais disponíveis ao final do acompanhamento médico. Esses critérios não permitem o reconhecimento precoce de formas potencialmente graves, para as quais é crucial a instituição do tratamento imediato (SAÚDE, 2013).

1.2 Objetivos

1.2.1 Objetivo Geral

Construir um classificador que utilize dentre as diversas técnicas de classificação em Data Mining e avaliar a que mais se adequa ao conjunto de dados utilizados no SINAN dengue do município de Fortaleza.

1.2.2 Objetivos Específicos

- Realizar experimentos com classificadores mais utilizados na literatura.
- Identificar qual classificador é mais adequado para o trabalho.
- Tratar o conjunto de dados para aplicação do classificador
- Construir o modelo classificador
- Validar e interpretar os resultados obtidos no processo de classificação

1.3 Justificativa

Considera-se a dengue um dos maiores problemas de saúde pública do mundo, especialmente nos países tropicais, cujas condições sócio-ambientais favorecem o desenvolvimento e a proliferação do seu principal vetor o *Aedes aegypti*. A dengue é, hoje, uma das doenças mais frenquentes no Brasil, atingindo a população em todos os estados, independente de classe social (SAÚDE, 2008).

O desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases médicas possui grande importância para apoiar o médico em sua tarefa diária de tomada de decisões, aumentando a precisão, a confiabilidade e a eficiência dos diagnósticos elaborados pelo especialista. Esse apoio computacional pode atuar como uma junta médica virtual, ao trazer para o especialista o conhecimento armazenado em exames e diagnósticos relacionados (COSTA, 2012).

1.4 Trabalhos Relacionados

No trabalho de (SHAKIL; ANIS; ALAM, 2015), observamos que o foco principal do trabalho foi a classificação de dengue, onde inicialmente aplicou-se a classificação nos datasets iniciais para identificar qual algoritmo obteve melhor resultado. O experimentos releram que os algoritmos Naive Bayes e J48 obtiveram melhores resultados, onde as maiores contribuições do trabalho foram:

- A extração da acurácia de classificação para predição do diagnóstico de dengue.
- Comparação de diferentes algoritmos de mineração no conjunto de dados de dengue.
- Identificação dos algoritmos com melhor desempenho para previsão dos diagnósticos.

No trabalho de (THITIPRAYOONWONGSE; SURIYAPHOL; SOONTHORNPHISAJ, 2012), foi utilizado a técnica de classificação conhecida como Árvore de Decisão, aplicado a um conjunto de dados temporais contendo dados clínicos e laboratoriais. Utilizaram dois conjuntos de dados com mais de 400 atributos. Nos experimentos os dados foram divididos em: conjunto de teste e treinamento. Onde o objetivo principal deste trabalho foi identificar o dia zero da dengue, pois esta previsão torna-se crítica, tendo em vista que o dia zero tem forte relação com o melhor tratamento do paciente.

No trabalho (SANTOS; NETO,), foi desenvolvido um aplicativo para interação dos profissionais da saúde com os dados do SINAN. Onde resultou uma aplicação em java que implementa algoritmos de classificação e regras de associação da base de dados do SINAN utilizando a API WEKA. Como resultado esta ferramenta possibilitou o auxílio no diagnóstico de pacientes.

1.5 Metodologia de Desenvolvimento

Com base na metodologia utilizada no processo de Descoberta de Conhecimento em Base de Dados (Knowledge Dicovery in Databases) definida em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a realização desta pesquisa será composta das seguintes etapas:

- Na primeira etapa, ocorrerá a extração, manutenção e pré-processamento dos dados, onde será realizada a análise de base de dados para extração de informações de atendimentos de pacientes necessários à pesquisa.
- A quinta e última etapa do processo ocorrerá a análise dos resultados obtidos com a a aplicação dos algoritmos de mineração de dados. Encontrar padrões de classificação que sejam relevantes para a pesquisa em questão.

1.6 Estruturação da Monografia

Além deste capítulo introdutório, este trabalho está organizado da forma descrita à seguir. No Capítulo 2, será descrito conceitos básicos relacionados a compreensão deste trabalho, bem como referencial teórico. No Capítulo 3, será apresentado os experimentos juntamente com os resultados obtidos. Finalmente no Capítulo 4, será apresentado as conclusões observadas e sugestões de trabalhos futuros da pesquisa.

2 Referencial Teórico

Este Capitulo apresenta uma visão geral dos conceitos envolvidos nas várias técnicas de classificação utilizadas ao longo deste trabalho.

Neste capítulo, apresentamos uma visão geral dos conceitos e procedimentos envolvidos em KDD e consequentemente em Data Mining. Apresentam-se também os trabalhos encontrados na literatura sobre assunto que são relevantes para o desenvolvimento deste trabalho.

2.1 Descoberta de Conhecimento em Base de Dados

Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a Descoberta de Conhecimento em Base de Dados é um processo não trivial de identificações de novos padrões, válidos e potencialmente úteis.

Em contrapartida, no trabalho de (THOMÉ, 2002), define-se KDD como sendo a busca de extração de conhecimento de bases de dados utilizando-se de técnicas e algoritmos que realizam a mineração dos dados para trabalhar e descobrir relações.

O processo de descoberta de conhecimento ocorre quando um conjunto de padrões que são semelhantes, e que podem levar a construção de um modelo. Este processo é formado por 5 etapas: seleção, pré-processamento, transformação, a mineração de dados e a interpretação dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Além do processo, o conhecimento que se deseja buscar deve estar de acordo com três características: deve ser correto, deve ser compreensível para o usuário e deve ter de alguma forma utilidade para o usuário (FREITAS, 2000).

A seguir serão detalhadas as etapas do processo de KDD de acordo com o apresentado na figura 1.

A etapa de seleção dos dados inicia com a definição do objetivo e mapea-

mento dos grupos ou conjuntos de informações que serão utilizados.

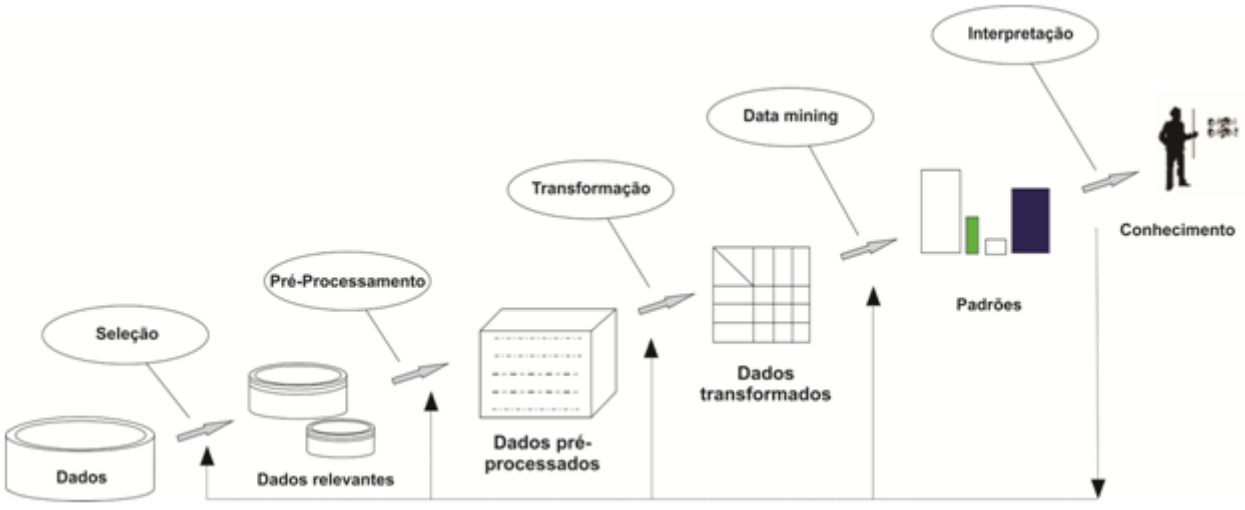
O pré-processamento é responsável pelo tratamento de ruídos e dados incompletos.

A transformação tem como objetivo selecionar as principais características que serão utilizadas para representar os dados, ou seja, os dados devem ser selecionados de modo que sejam os mais úteis para o modelo proposto.

A etapa de mineração de dados é o momento em que serão escolhidos os algoritmos que mais se ajustam ao objetivo que se quer extrair da base de dados. Além disso, nesta fase são escolhidos os melhores parâmetros para que, no momento do processamento, os resultados sejam os mais rápidos e precisos possíveis.

Ao final do processo teremos a etapa de interpretação e avaliação dos resultados, onde o conhecimento extraído da base de dados é representado por padrões.

Figura 1 – Etapas do processo KDD.



Fonte: Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Inicialmente, é necessário definir que tipo de conhecimento se deseja extrair da base de dados, pois a técnica que será utilizada para a mineração de dados depende do objetivo a que se quer chegar (DAMASCENO, 2005).

2.2 Data Mining

De acordo com (ADRIAANS; ZANTINGE, 1996), existe uma confusão entre os termos *Data Mining* e KDD, podendo ser usadas até como sinônimos em algumas situações. Em contrapartida, (BERRY; LINOFF, 1997), definiu como um processo de exploração e análise, de uma grande quantidade de dados, por meio automático ou semiautomático, com o propósito de descobrir regras e padrões significativos.

É uma técnica que faz parte de uma das etapas da descoberta de conhecimento em banco de dados. É uma área de pesquisa multidisciplinar, incluindo principalmente as tecnologias de banco de dados, inteligência artificial, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.

Em termos gerais, a técnica de Data Mining compreende os seguintes propósitos:

- Previsão – pode mostrar como certos atributos dentro dos dados irão comportar-se no futuro;
- Identificação – padrões de dados podem ser utilizados para identificar a existência de um item, um evento ou uma atividade;
- Classificação – pode repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros;
- Otimização do uso de recursos limitados, como tempo, espaço, dinheiro ou matéria-prima e maximizar variáveis de resultado como vendas ou lucros sob um determinado conjunto de restrições.

2.3 Aprendizagem Não Supervisionada

Como mostrado por (DAMASCENO, 2005) aprendizagem não supervisionada é aquela que utiliza instâncias sem a determinação do atributo classe. Este tipo de aprendizado é utilizado geralmente para análise exploratória dos dados, utilizando técnicas de agrupamento ou regras de associação. Onde agrupamentos têm como objetivo relacionar instâncias com características em comuns.

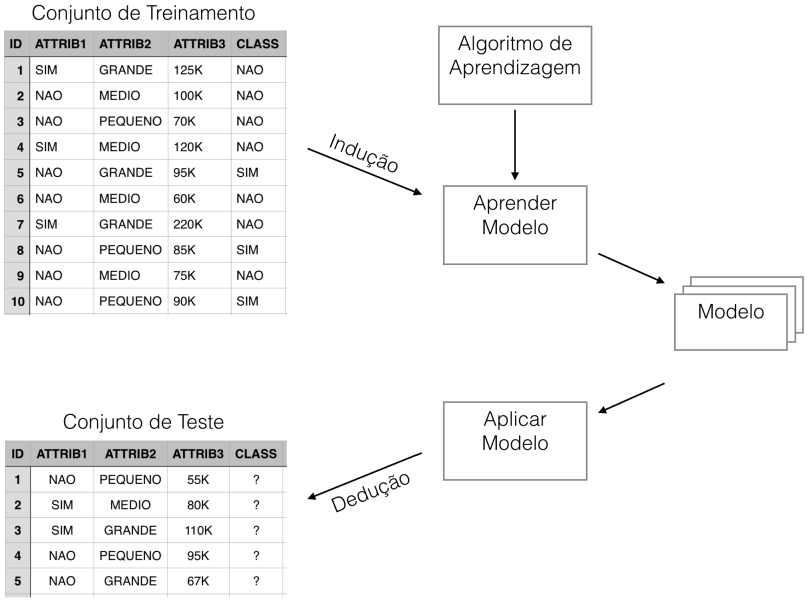
A partir da definição de uma métrica de similaridade, os dados são agrupados, dando a possibilidade de encontrar relações interessantes entre as instâncias. Assim, o cliente do conhecimento gerado pode aplicar uma determinada ação em um subconjunto de instâncias presente nos dados.

2.4 Técnicas de Classificação

Uma técnica de classificação é uma abordagem sistemática para construção de modelos de classificação a partir de um conjunto de dados de entrada. Exemplos incluem classificadores de árvores de decisão, classificadores baseados em regras, redes neurais, máquinas de vetores de suporte e classificadores Bayes simples. Cada técnica emprega um algoritmo de aprendizagem para identificar um modelo que seja mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo da classe dos dados de entrada. O modelos gerado pelo algoritmo de aprendizagem deve se adaptar bem aos dados de entrada e prever corretamente os rótulos de classes de registros que ele nunca viu antes. Portanto, o objetivo chave do algoritmo de aprendizagem é construir modelos com boa capacidade de generalização (TAN; STEINBACH; KUMAR, 2009).

A Figura 2 mostra uma abordagem geral para resolver problemas de classificação. Primeiro, um conjunto de treinamento consistindo de registros cujos rótulos sejam conhecidos e devem ser fornecidos. O conjunto de treinamento é usado para construir um modelo de classificação, que é subsequentemente aplicado ao conjunto de teste, que consiste de registros com rótulos de classes desconhecidas (TAN; STEINBACH; KUMAR, 2009).

Figura 2 – Abordagem geral para a construção de um modelo de classificação



Fonte: Adaptado de (TAN; STEINBACH; KUMAR, 2009).

2.4.1 Árvore de Decisão

Em uma árvore de decisão, cada nodo folha recebe um rótulo de classe. os nodos não terminais, que incluem o nodo raiz e outros nodos internos, contêm condições de testes de atributos para separar registros que possuem características diferentes.

Classificar um registro de testes é direto, assim que uma árvore de decisão tenha sido construída. Começando do nodo raiz, aplicamos a condição de teste ao registro e seguimos a ramificação apropriada baseados no resultado do teste. Isto nos levará a um outro nodo interno, para o qual uma nova condição de teste é aplicada, ou a um nodo folha (TAN; STEINBACH; KUMAR, 2009).

2.4.1.1 Algoritmo J48

2.4.1.2 Algoritmo RandomTree

2.4.1.3 Algoritmo REPTree

2.4.2 Teorema de Bayes

É um teorema usado para calcular a probabilidade condicional , usando em estatística, probabilidade e outras aplicações

Em muitas aplicações, o relacionamento entre o conjunto de atributos e a variável classe é não determinístico. O rótulo da classe de um registro de teste não pode ser previsto com certeza embora seu conjunto de atributos seja idêntico a alguns dos exemplos de treinamento. Esta situação pode surgir por causa de dados com ruídos ou da presença de determinados fatores de confusão que afetam a classificação mas que não são incluídos na análise.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{2.1}$$

Sendo $P(A \mid B)$ a probabilidade **posteriori** condicional de A a B, tem-se que:

- $P(B \mid A)$ a probabilidade **posteiori** condicional de B a A.
- $P(A)$ probabilidade **apriori** de A.
- $P(B)$ probabilidade **apriori** de B.

2.4.2.1 Algoritmo NaiveBayes

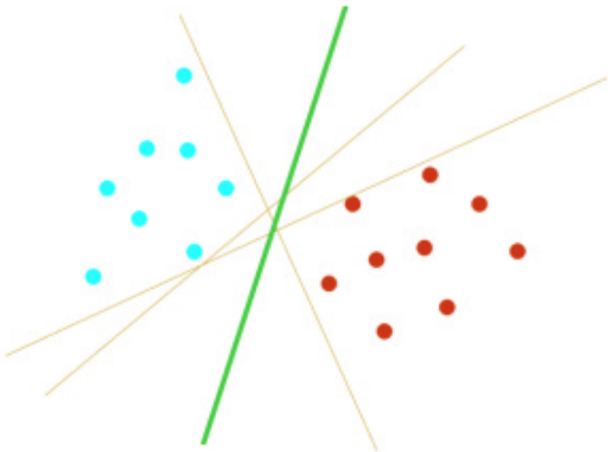
O Classificador *NaiveBayes* é uma técnica probabilística baseada no teorema de *Bayes*, expressão, para calcular a probabilidade Posteriori da classe C Em (JOHN; LANGLEY, 1995)

2.4.3 Máquinas de Vetores de Suporte (SVM)

Uma técnica de classificação que tem recebido considerável atenção pois esta técnica possui seus fundamentos na teoria de aprendizagem estatística e tem mostrado resultados empíricos promissores em muitas aplicações práticas, desde o reconhecimento de dígitos escritos à mão até a categorização de textos. SVM também funciona bem com dados de alta dimensionalidade e evita o problema da dimensionalidade. Outro aspecto único desta abordagem é que ela representa o limite da decisão usando um subconjunto dos exemplos de treinamento, conhecido com vetores de suporte (TAN; STEINBACH; KUMAR, 2009).

Na Figura 3, existe um conjunto de classificadores lineares que separam duas classes, mas apenas um (em destaque) que maximiza a margem de separação (distância da instância mais próxima ao hiperplano de separação das classes). O hiperplano com margem máxima é chamado de hiperplano ótimo (JUNIOR, 2010).

Figura 3 – Conjunto de hiperplanos possíveis



Fonte: (JUNIOR, 2010).

2.4.3.1 Algoritmo SMO

2.4.4 Métricas para avaliação de modelo

As métricas mais utilizadas na literatura...

- *Correctly Classified Instances*: Mostra o percentual de precisão das instâncias corretamente classificados.
- *Incorrectly Classified Instances* : Mostra o percentual de precisão das instâncias incorretamente classificados.
- *TP Rate*: Taxa de verdadeiros positivos.
- *FP Rate*: Taxa de falsos positivos.
- *Precision*: Percentual de instâncias positivas classificadas corretamente sobre o total de instâncias classificadas como positivas.
- *Recall*: Percentual de instâncias positivas classificadas corretamente sobre o total de instâncias positivas.
- *F-Measure*: É uma média ponderada de *Precision* e *Recall*.
- *ROC Area*:

2.5 WEKA

A ferramenta WEKA (Waikato Environment for Knowledge Analysis) conforme descrito em (HALL et al., 2009), visa proporcionar uma coleção abrangente de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento de dados tanto para pesquisadores como para profissionais. Permite aos usuários experimentar rapidamente e comparar diferentes métodos de aprendizagem de máquina em diferentes conjuntos de dados. Sua arquitetura modular, extensível, permite que os processos de mineração de dados pode ser construido a partir da vasta coleção de algoritmos de aprendizado e diversas ferramentas oferecidas.

2.6 Dados Abertos

Utilizamos a base de dados disponibilizada pelo portal de dados abertos da prefeitura do Rio de Janeiro de casos notificados de dengue. O portal de dados abertos de Fortaleza é um espaço desenvolvido pela Coordenadoria de Ciência, Tecnologia e Inovação da Prefeitura de Fortaleza (CITINOVA) para que a sociedade possa encontrar e utlizar os dados e informações públicas da cidade de Fortaleza. Os dados são publicados em formatos abertos que permitem sua reutilização em aplicativos digitais desenvolvidos por e para qualquer pessoa. Além disso, o portal serve como uma ferramenta de interlocução com a sociedade fortalezense para pensar e promover a inovação e a criatividade em prol da melhoria de serviços e da vida na cidade de Fortaleza. O portal de dados abertos de Fortaleza está em conformidade com os princípios da administração pública e observâncias às recomendações aceitas internacionalmente, como as emitidas pela Open Knowledge Foundation e pelo Consórcio W3C Internacional.

2.7 SINAN

O Sistema de Informações de Agravos de Notificação - SINAN é o principal sistema de informações que tem como objetivo os dados referentes a morbidade, sendo fundamental no processo de trabalho da Vigilância em Saúde, estando envolvido não somente nas ações de Vigilância Epidemiológica, mas também na Vigilância Ambiental em Saúde e Vigilância em Saúde do Trabalhador (NOTA. . . , 2015). O SINAN foi desenvolvido no início da década de 90, com o objetivo de padronizar a coleta e o processamento de dados sobre agravos de notificação obrigatória em todo território nacional. Construido de maneira hierarquizada, mantendo coerência com a organização do SUS (Sistem Único de Saúde), pretende ser suficiente ágil na viabilização de análises de situações de saúde em curto espaço de tempo. O SINAN fornece dados para a análise do perfil da morbidade e contribui para a tomada de decisões nos níveis municipal, estadual e federal (SAÚDE, 2008). A dengue é uma das doenças

de notificação compulsória, devendo todo caso suspeito ou confirmado ser notificado ao Serviço de Vigilância Epidemiológica, por meio do SINAN nas fichas de notificação de investigação (SAÚDE, 2008).

3 Experimentos

Neste capítulo, avaliamos os modelos gerados através de uma série de experimentos. Apresentaremos a base de dados utilizada nestes experimentos, nossa metodologia de experimentação e os experimentos em si realizados.

3.1 Base de Dados

Para avaliar o objetivo geral neste trabalho, usamos a base de dados do Município de Fortaleza, diposnibilizada pelo portal de dados abertos. Os dados são publicados em formatos abertos que permitem sua reutilização em aplicativos digitais desenvolvidos por e para qualquer pessoa (NOTIFICAÇÕES. . . , 2015). Este conjunto de dados contém notificações de dengue do Município de Fortaleza no período de janeiro à junho de 2015, contendo 26.568 instâncias. Na Tabela **??**, podemos observar os atributos utilizados neste trabalho.

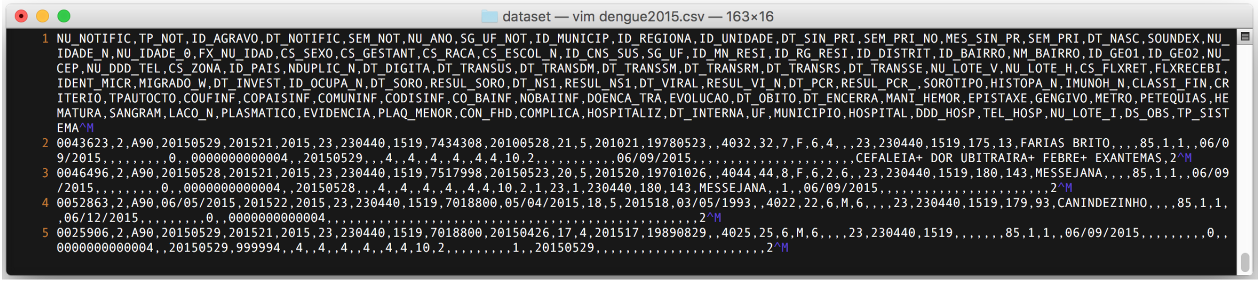
Neste Capítulo é descrita a metodologia utilizada para classificação do tipo de dengue em ocorrências do SINAN do município de Fortaleza. Basicamente, pode ser apresentada da seguinte forma:

- Definição de Atributos utilizados:
- Seleção dos Dados:
- Particionamento do Conjunto de Dados:
- Treinamento dos Classificadores:

3.2 Pré-processamento

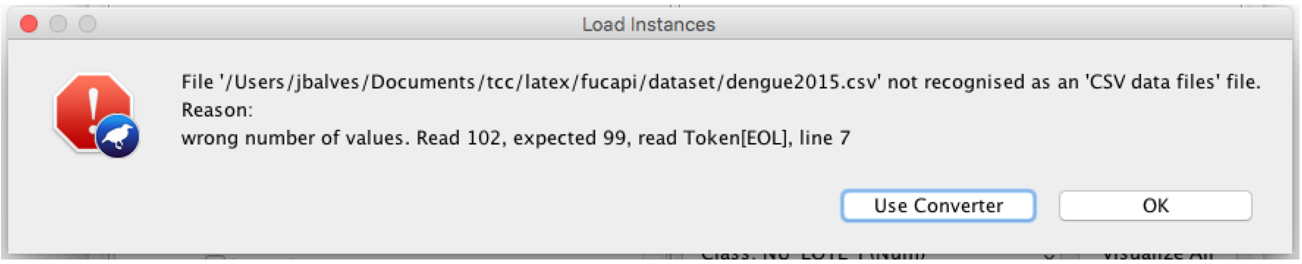
Após o download da base de dados em (NOTIFICAÇÕES. . . , 2015), iniciamos a importação do arquivo no formato CSV (*Comma Separated Value*), conforme mostrado na Figura 4.

Figura 4 – Data Set Original



Fonte: Próprio Autor.

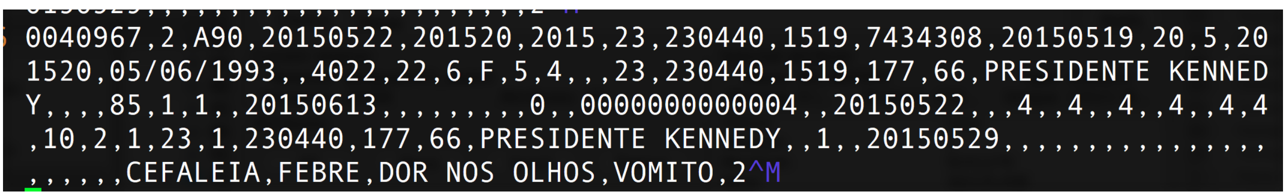
Figura 5 – Erro ao Carregar Instâncias



Fonte: Próprio Autor.

Onde nos deparamos com o erros no campos DS_OBS, onde o WEKA informa que está lendo 102 atributos ao invés de 99, conforme mostrado na Figura 5. Já na Figura 6, podemos observar que houve falha na inserção dos dados no atributo DS_OBS, pois o mesmo foi preenchido com vírgula em seu conteúdo da seguinte ”CEFALEIA, FEBRE, DOR NOS OLHOS, VOMITO”, além do caracter ”$\hat{M}$” que representa quebra de linha. Pelo fato do arquivo ser contruído no formato CSV (*Comma Separated Value*), essas vírgulas excedentes serão consideradas como início de novos atributos.

Figura 6 – Erro Mencionado pelo WEKA



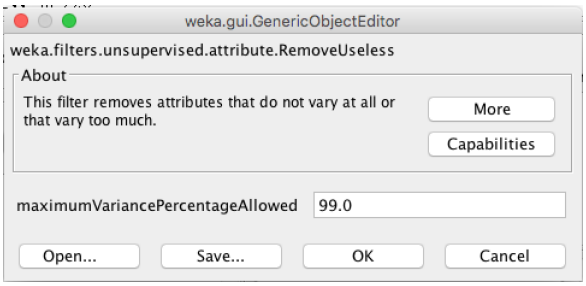
Fonte: Próprio Autor.

Para resolver este problema editamos o arquivo de tal forma que os atributos ”DS_OBS” e ”TP_SISTEMA” foram removidos.

Em seguida aplicamos os seguintes filtros:

- *RemoveUseless*: Este filtro remove atributos que não variam em tudo ou que variam muito. Todos os atributos constantes são excluídos automaticamente, juntamente com qualquer que exceder o percentual máximo de parâmetro de variação. O teste de variância máxima só é aplicada aos atributos nominais. Utilizamos os parâmetros padrões conforme Figura 7.

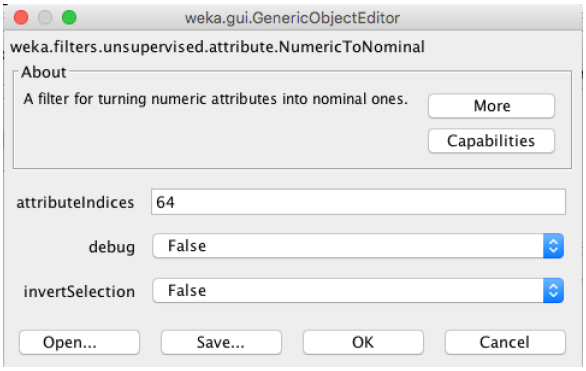
Figura 7 – Filtro *RemoveUseless*



Fonte: Próprio Autor.

- *NumericToNominal*: Um filtro para transformar numérico atributos em uns nominais. Ao contrário de discretização, apenas toma todos os valores numéricos e os adiciona à lista de valores nominais do que atributo. Útil após a importação CSV, para impor certos atributos para se tornar nominal. Utilizamos o parâmetro mostrado na Figura 8.

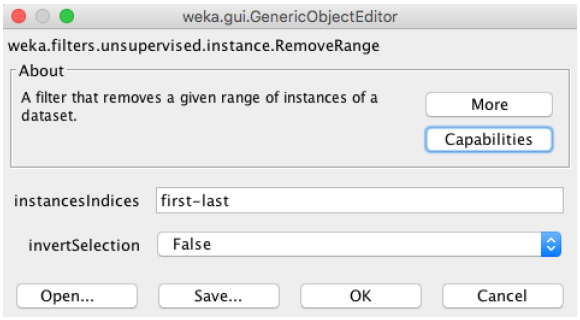
Figura 8 – Filtro *NumericToNominal*



Fonte: Próprio Autor.

- *RemoveRange*: Um filtro que remove uma determinado intervalo de ocorrências de um conjunto de dados. Utilizamos o parâmetro mostrado na Figura 9. O parâmetro "instancesIndices" foi utilizado cinco vezes para criarmos 05 (cinco) conjuntos de dados com as seguintes quantidades de instâncias: 5.000, 10.000, 15.000, 20.000 e 25.000. Que foram utilizados durantes os experimentos.

Figura 9 – Filtro *RemoveRange*



Fonte: Próprio Autor.

3.3 Metodologia

Nossos experimentos foram realizar a partir dos conjuntos de dados divididos nas seguintes quantidades de instâncias: 5.000, 10.000, 15.000, 20.000 e 25.000. A cada conjunto de instâncias, aplicamos os processos de classificação

utilizando 5 classificadores que estão contidos em 3 técnicas de classificação. Como resultado dos classificadores obtivemos 25 modelos. Todos os modelos utilizara o processo de validação conhecido como *Cross Validation*. Cada modelo gerado disponibiliza um conjunto de informações para avaliação que serão utilizadas para comparação entre si.

3.4 Resultados

Nesta seção apresentamos os resultados obtidos na aplicação dos classificadores em 5 conjuntos de dados e as medidas resultantes que utilizamos:

Tabela 1 – Resultados dos Experimentos com 5.000 instâncias

Classificador	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**FP Rate**
SMO	89,2866%	10,7134%	0.863	0.346
J48	86,6530%	13,3470%	0.867	0.701
REPTree	83,9683%	16,0317%	0.84	0.815
NaiveBayes	83,5592%	16,4408%	0.836	0.619
ZeroR	84,2496%	15,7504%	0.842	0.842
RandomTree	82,0762%	17,9238%	0.821	0.623

Tabela 2 – Resultados dos Experimentos com 10.000 instâncias

Classificador	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**FP Rate**
SMO	88.5345%	11.4655%	0.885	0.218
J48	86,2315%	13,7685%	0.862	0.328
REPTree	79,7414%	20,2586%	0.797	0.497
RandomTree	77,2906%	22,7094%	0.773	0.408
NaiveBayes	76,8103%	23,1897%	0.768	0.420
ZeroR	75,6773%	24,3227%	0.757	0.757

Tabela 3 – Resultados dos Experimentos com 15.000 instâncias

Classificador	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**FP Rate**
SMO	89,6195%	10,3805%	0.896	0.207
J48	86,6322%	13,3678%	0.866	0.335
REPTree	80,7076%	19,2924%	0.807	0.507
ZeroR	76,8608%	23,1392%	0.769	0.769
RandomTree	78,8551%	21,1449%	0.789	0.435
NaiveBayes	76,1682%	23,8318%	0.762	0.448

Tabela 4 – Resultados dos Experimentos com 20.000 instâncias

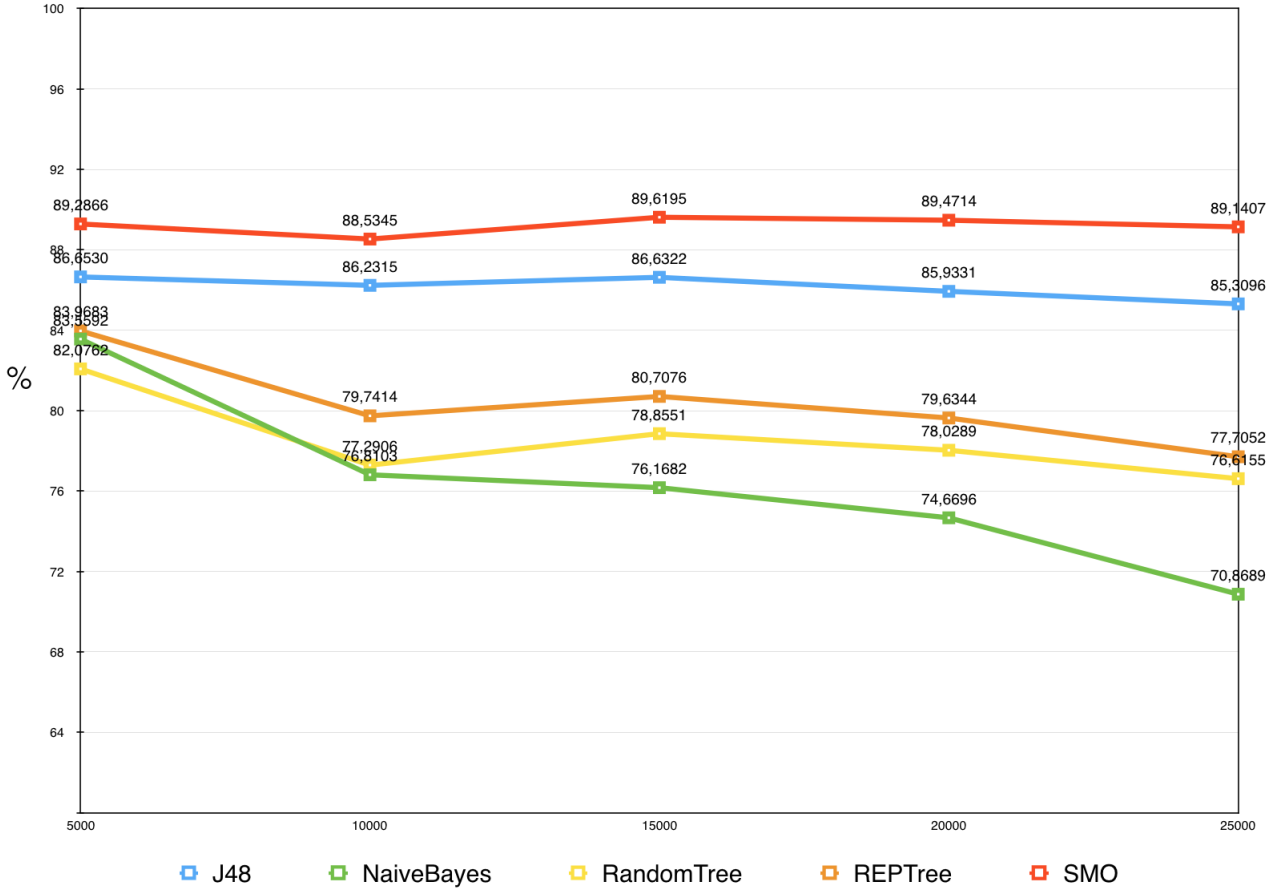
Classificador	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**FP Rate**
SMO	89,4714%	10,5286%	0.895	0.199
J48	85,9331%	14,0669%	0.859	0.324
REPTree	79,6344%	20,3656%	0.796	0.499
RandomTree	78,0289%	21,9711%	0.780	0.426
ZeroR	75,9664%	24,0336%	0.760	0.760
NaiveBayes	74,6696%	25,3304%	0.747	0.417

Tabela 5 – Resultados dos Experimentos com 25.000 instâncias

Classificador	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**FP Rate**
SMO	89,1407%	10,8593%	0.891	0.169
J48	85.3096%	14,6904%	0.853	0.287
REPTree	77,7052%	22,2948%	0.777	0.431
RandomTree	76,6155%	23,3845%	0.766	0.377
ZeroR	72,1555%	27,8445%	0.722	0.722
NaiveBayes	70,8689%	29,1311%	0.709	0.319

Como podemos observar na Figura 10, os modelos com melhores resultados SMO e J48.

Figura 10 – Instâncias Classificadas Corretamente



Fonte: Próprio Autor.

4 Conclusões e Trabalhos Futuros

Neste Capítulo, apresentamos as conclusões do nosso trabalho, incluindo suas limitações de nossa pesquisa e futuros estudos que possam ser explorados.

4.1 Resultados Obtidos

As medidas de desepenho resultantes do experimentos na classificação foram

4.2 Limitações

Neste trabalho não estamos levando em consideração o tempo de criação de cada modelo gerado.

4.3 Trabalhos Futuros

Como trabalhos futuros verificamos a oportunidade de estudos na área de geração de modelos em tempo real, para que desta forma o modelo possa ser melhorado a cada nova clasificação.

A junção dods conjuntos de dados utilizados com índices de precipitação fluvial da região de onde os dados foram gerados.

Referências

ADRIAANS, P.; ZANTINGE, D. Data mining. *Addision-Wesley, Harlow*, 1996. páginas 15

BELLAZZI, R.; ZUPAN, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, Elsevier, v. 77, n. 2, p. 81–97, 2008. páginas 10

BERRY, M. J.; LINOFF, G. *Data mining techniques: for marketing, sales, and customer support*. [S.l.]: John Wiley & Sons, Inc., 1997. páginas 15

COSTA, A. F. *Mineração de imagens médicas utilizando características de forma*. Tese (Doutorado) — Universidade de São Paulo, 2012. páginas 10, 11

DAMASCENO, M. Introdução a mineração de dados utilizando o weka. *Instituto AAAFederal de Educação, Ciência e Tecnologia do Rio Grande do Norte*, 2005. páginas 15, 16

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. páginas 9, 13, 14, 15

FREITAS, A. Uma introdução a data mining. *Informática Brasileira em Análise, Centro de Estudos e Sistemas Avançados do Recife (CESAR), Recife, Pe, ano II*, n. 32, 2000. páginas 14

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. páginas 20

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. [S.l.], 1995. p. 338–345. páginas 19

JUNIOR, G. M. d. O. Máquina de vetores suporte: estudo e análise de parâmetros para otimização de resultado, 2010. *Trabalho de Graduação em Ciência da Computação*, 2010. páginas 19, 20

NOTA Técnica - 45 - 2013. 2015. <http://www.conass.org.br>. Acessado em Novembro de 2015. páginas 22

NOTIFICAÇÕES de Dengue - Ref. 01 a 06/2015. 2015. <http://dados.fortaleza.ce.gov.br/catalogo/dataset/notificacoes-de-dengue-ref-01-06-2015>. Acessado em 2015. páginas 23

SANTOS, M. S.; NETO, J. C. da C. Amagodis: Algoritmos de mineração para apoio à gerência de ocorrências de dengue a partir de informações presentes na base dados do sinan. páginas 12

SAÚDE, M. da. Vigilância em saúde: Dengue, esquistossomose, hanseníase, malária, tracoma e tuberculose. *Departamento de Atenção Básica*, 2008. páginas 11, 22

SAÚDE, M. da. Dengue: diagnóstico e manejo clínico: adulto e criança. *Departamento de Atenção Básica*, 2013. páginas 10

SHAKIL, K. A.; ANIS, S.; ALAM, M. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv:1502.05167*, 2015. páginas 12

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciencia Moderna, 2009. páginas 17, 18, 19

THITIPRAYOONWONGSE, D.; SURIYAPHOL, P.; SOONTHORNPHISAJ, N. Data mining of dengue infection using decision tree. *Entropy*, v. 2, p. 2, 2012. páginas 12

THOMÉ, A. C. G. Redes neurais: uma ferramenta para kdd e data mining. *Material Didático http://equipe. nce. ufrj. br/thome/grad/nn/mat_didatico/apostila_kdd_mbi. pdf, Outubro*, 2002. páginas 14