Classificação Automática das Reclamações de Clientes de uma Empresa de Telecomunicações

Dione Aparecido de Oliveira Sanga, Celso Antônio A. Kaestner

Departamento Acadêmico de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)
CEP 80.230-901 – Curitiba – Brasil

dione_sanga@hotmail.com, celsokaestner@utfpr.edu.br

Resumo. *No Brasil as empresas de telecomunicações são regulamentadas por órgãos de defesa do consumidor que recebem reclamações dos clientes das operadoras e podem penalizar as mesmas. Esse trabalho tem por objetivo avaliar se o uso de técnicas de mineração de textos para a criação de novos atributos contribui na identificação de clientes que não receberam atendimento adequado em centrais de gerenciamento do relacionamento com clientes (CRM) e evitar que estes migrem do ambiente interno de atendimento para órgãos regulamentadores. Para isso é utilizado dados extraídos de bases de CRM e aplicado diversos algoritmos de classificação. Nos experimentos o modelo que utilizou as entradas geradas pela mineração de textos se apresentou superior ao modelo tradicional, comprovando a eficácia do modelo.*

Abstract. *In Brazil telecommunications companies are regulated by consumer protection agencies that receiving customer complaints from operators and may penalize them. This paper aims to evaluate if the use of text mining techniques to create new attributes contributes to the identification of clients that did not receive adequate care in customer relationship management centers (CRM) and prevent them from migrating from the internal service environment regulatory bodies. To do so, we used data extracted from CRM databases and applied several classification algorithms. In the experiments, the model that used the entries generated by text mining presented over to the traditional model, proving the effectiveness of the model.*

1. Introdução

No Brasil as empresas de telecomunicações são regulamentadas pela Agência Nacional de Telecomunicações[1] (Anatel). Esse órgão regulamenta as regras do setor de telecomunicações no país e recebe as reclamações dos clientes das empresas de telecomunicações podendo penaliza-las por falta de qualidade no serviço prestado.

Uma forma útil que pode ser empregada para a identificação de clientes que migram de centrais de gerenciamento do relacionamento com o cliente (CRM) e procuram a Anatel é por meio da mineração de dados. O setor de telecomunicações foi um dos primeiros setores a adotar a mineração de dados e são diversas as aplicações desenvolvidas utilizando essa tecnologia. Estas aplicações podem ser divididas em três principais áreas de negócio como marketing e retenção de clientes, isolamento de falhas de rede e detecção de fraudes [Weiss et. al. 2005].

[1] http://www.anatel.gov.br/

O uso de técnicas de mineração de dados nesse contexto gera a oportunidade de identificar clientes que estão passando por problemas de atendimento na empresa e permite a tomada de decisão para evitar que os mesmos migrem para órgãos de defesa do consumidor. Um fator crucial que justifica a pesquisa é que a indústria de telecomunicações gera e armazena uma enorme quantidade de dados, essas que são informações que registram todo o ciclo de vida do cliente dentro da empresa e é o insumo básico para a mineração de dados [Huang et. al. 2010].

Nesse contexto, este artigo realiza a comparação de dois conjuntos de dados sob a execução de diferentes algoritmos de classificação de mineração de dados para identificar se o emprego de técnicas de mineração de textos em dados não-estruturados de reclamações de clientes fornecem novos atributos que possam contribuir com o aumento da acurácia dos modelos de classificação.

Em nossas pesquisas não foram encontradas aplicações em CRM que tratem o problema da migração de clientes para as agências reguladoras da área, objeto deste estudo.

Na seção 2 desse artigo é apresentado a relação e desafios encontrados pela mineração de dados em telecomunicações. A seção 3 descreve as principais técnicas e métodos encontradas no estado da arte e que são empregadas nos experimentos realizados. A seção 4 detalha os experimentos e os resultados obtidos na proposta deste artigo. Por fim, a seção 5 descreve as conclusões e melhorias que podem ser empregadas em trabalhos futuros para a continuidade do caso de estudo.

2. Mineração de Dados e Telecomunicações

A grande quantidade de dados gerada por telecomunicações apresenta vários problemas interessantes para a mineração de dados. Um dos principais problemas diz respeito a escala, bases de dados de telecomunicações podem conter bilhões de registros e estão entre os maiores bancos de dados do mundo. Uma segunda questão é que os dados brutos não estão adequados para a aplicação de mineração de dados na maioria das vezes, sendo necessário a aplicação de diversas técnicas de pré-processamento para a adequação de seu uso na mineração de dados [Weiss et. al. 2005].

Muitas aplicações voltadas para a mineração de dados em telecomunicações buscam prever eventos muito raros, como falha de componentes de rede ou uma instância de fraude telefônica, portanto, raridade é outra questão que deve ser tratada. Por fim, o desempenho em tempo real é outro ponto de atenção em aplicações de mineração de dados voltadas para telecomunicações: modelos de detecção de fraude por exemplo devem executar de maneira *online* para realizar adequadamente sua tarefa [Weiss et. al. 2005].

3. Mineração de Dados e suas Etapas

A mineração de dados é a exploração e análise de forma automática ou semiautomática de grandes bases de dados com a função de descobrir padrões antes desconhecidos [Tan et. al. 2009]. O seu uso permite a automatização de processos antes executados de forma manual e lenta por métodos confiáveis e de larga escala.

No entanto, antes da execução da mineração de dados é necessário o pré-processamento dos dados para transformar dados brutos em um formato apropriado para a análise e execução de algoritmos de mineração de dados. O Descobrimento de

Conhecimento em Banco de Dados (*Knowledge Discovery Databases - KDD*) é o conjunto de técnicas e métodos que abrange desde a seleção dos dados até a análise dos resultados obtidos na etapa de mineração de dados e contempla todas as técnicas necessárias para esse trabalho. A seguir são descritos algumas dessas técnicas.

3.1. Seleção dos Dados

A seleção dos dados tem por objetivo agrupar as diversas fontes de dados dentro de um mesmo ambiente computacional, essa tarefa além de concentrar os dados utilizados no projeto em um mesmo banco de dados, separa dados transacionais de ambientes de produção para um banco de dados específico. Portanto, a compreensão das diferentes fontes de dados é importante, pois dificilmente aplicações úteis podem ser desenvolvidas caso não se tenha o conhecimento dos dados utilizados, deixando uma grande lacuna entre o que se espera e o produto entregue [Weiss et. al. 2005].

3.2. Pré-Processamento

O pré-processamento dos dados estabelece as bases para a mineração de dados, ou seja, antes da descoberta de conhecimento novo, o conjunto de dados deve ser previamente preparado, justificando portanto a importância dessa etapa. Casos onde esta etapa é ignorada ou não efetivamente executada os resultados finais normalmente são insatisfatórios. Dessa forma, os resultados obtidos com a execução dos algoritmos estão atrelados a efetiva preparação dos dados e suas características [Zhang et. al. 2007].

Os principais objetivos da etapa de pré-processamento são identificar dados corrompidos ou ruidosos, atributos irrelevantes e valores desconhecidos. Outras tarefas comumente aplicadas na etapa de pré-processamento são o uso de técnicas como discretização, binarização, construção de algoritmos de transformação e criação de variáveis e pré-processamento de dados não estruturados como acontece na mineração de textos.

Discretização é uma técnica importante para alguns algoritmos de aprendizado de máquina em especial para algoritmos de classificação que requerem a transformação de atributos contínuos em atributos categóricos. A aplicação dessa técnica na etapa de pré-processamento permite que algoritmos de classificação apresentem melhores resultados devido a melhor condição que os dados fornecem as características de alguns algoritmos [Antunes et. al. 2001].

A transformação e criação de variáveis refere-se a transformação aplicada a todos os valores de um atributo, esta pode ser aplicada com métodos de normalização de dados – ajustar a escala de um atributo entre 0 e 1 – ou a criação de novos atributos à partir de atributos já existentes, esse tipo de operação é justificada pois, além de expressar relacionamentos conhecidos entre atributos existentes, pode reduzir o conjunto de dados simplificando o processamento de algoritmos [Fayyad et. al. 1996].

Portanto, na grande maioria dos projetos de mineração de dados os dados brutos devem ser processados de tal maneira que possam fornecer melhores condições ao conjunto de dados visando facilitar à compreensão dos modelos e a execução dos algoritmos [Tan et. al. 2009]. A maior parte do tempo gasto em projetos de mineração de dados é consumido com a preparação dos dados, estimando-se que 80% do tempo do projeto é gasto nesta tarefa [Zhang et. al. 2007].

3.3. Algoritmos de Classificação

A classificação pode ser definida como a tarefa de aprendizado de uma função alvo *f* que mapeie cada conjunto de valores dos atributos previsores *x* para um dos rótulos de classes pré-determinadas *y* [Tan et. al. 2009]. O processo de classificação de dados envolve duas fases a aprendizagem e o teste. Durante a fase de treinamento um modelo de classificação é gerado a partir de um conjunto de dados de treinamento. Este modelo varia de acordo com o paradigma adotado. A performance obtida na classificação é apurada na fase de teste, onde novos dados são apresentados para o modelo com sua classificação pré-definida e é calculada a taxa de acurácia do modelo obtido [Deulkar et. al. 2016].

Nos trabalhos de mineração de dados que utilizam a tarefa de classificação em conjuntos de dados de telecomunicações os principais algoritmos utilizados são árvores de decisão, Naïve Bayes, *k-nearest neighbor* (k-NN), *support vector machine* (SVM) e redes neurais [Han et. al. 2006]. Algoritmos de classificação são aplicados para se atingir diversos objetivos de áreas de conhecimento diferentes, porém cada algoritmo possui características e conceitos particulares. Dessa forma, diferentes algoritmos de classificação obtêm resultados distintos sob o mesmo conjunto de dados, fazendo com que determinados algoritmos se adequem melhor a alguns conjuntos de dados do que outros.

3.4. Mineração de Textos

O objetivo do descobrimento de conhecimento textual é extrair conceitos explícitos e implícitos e relações semânticas utilizando técnicas de processamento de linguagem natural (*Natural Language Processing* - NLP). Acredita-se que a extração de conhecimento de texto tem um potencial comercial mais elevado do que a extração de conhecimento sobre dados [Tan et. al. 1999].

Na etapa de pré-processamento de dados não-estruturados deve ser realizado a conversão dos termos em um padrão único, *lowercase* ou *uppercase*, como também a remoção de pontuação, números e caracteres não identificados. Tais tarefas eliminam ruídos e aumentam a precisão sobre as demais técnicas. Também são aplicadas técnicas como a remoção de lista de palavras (*Stop-Words*), radicalização das palavras (*Stemming*) e entre outros para que os dados tornem-se apropriados para o uso em mineração de dados.

A remoção de *Stop-Words* consiste na identificação de termos frequentes em textos e são palavras que não geram informação relevante para a base de dados. Sua remoção tem como finalidade comprimir textos, pois reduz a dimensão dos termos analisados. A definição de quais são os termos irrelevantes é feita por uma lista de *Stop-Words*, um vetor de termos irrelevantes que gera ruído ao conteúdo dos documentos.

Algoritmos de radicalização ou S*temmer* permite a remoção das variações de uma palavra, permanecendo apenas o radical da palavra. Tais variações incluem plurais, gerúndios, sufixos de terceira pessoa, sufixos de tempo passado, etc. Uma lista de sufixos (*suffixlist*) é pré-definida, dependente do idioma, contendo os sufixos das palavras. Ao final do processo o armazenamento é melhorado, pois menos termos são armazenados.

Para a representação dos termos o modelo espaço-vetorial é amplamente utilizado, onde cada termo é representado por um vetor e cada termo possui um valor associado que indica o grau de importância do mesmo. No vetor estão todos os termos da coleção e não

somente aqueles presentes no documento. Os termos que o documento não contém recebem grau de importância zero [Tan et. al. 1999].

O peso de um termo pode ser calculado de diversas formas, uma forma comum é a frequência do termo (*Term Frequency*), essa medida de peso soma a quantidade vezes que o termo é encontrado no documento. Outras formas é quando identificado a existência do termo sua representação é dada pelo número 1 e quando não houver sua incidência no documento sua representação é dada pelo número 0. Outra técnica é o TF-IDF (*Term Frequency – Inverse Document Frequency*) que é a frequência do termo, inverso a frequência dos termos nos documentos, entre outros.

3.5. Avaliação

A última etapa do processo de KDD é realizar a interpretação e avaliação dos resultados obtidos a fim de identificar se os objetivos iniciais foram alcançados. Com a interpretação podem surgir padrões, relacionamentos e descoberta de novos fatos antes desconhecidos. Caso os resultados obtidos não satisfaçam os objetivos iniciais é possível retornar as etapas anteriores para a realização de ajustes e correções, caso contrário os resultados podem ser incorporados a outros sistemas, documentados ou utilizados em processos de tomada de decisão [Fayyad et. al. 1996].

4. Classificação das Reclamações de Clientes em Telecomunicações

Essa seção apresenta os experimentos desenvolvidos com dados de uma empresa de telecomunicações que fornece serviços de banda larga, comunicação por voz e TV. Os experimentos comparam dois modelos de dados com os principais algoritmos de classificação, identificando clientes que migraram do ambiente interno de atendimento para a Anatel. O objetivo do experimento é comparar se a inclusão de informações textuais das reclamações dos clientes, aumenta a acurácia do modelo, por meio da mineração de textos.

4.1. Dados

Os dados utilizados nos experimentos são compreendidos por duas bases, a primeira é formada por informações de registros de reclamações no ambiente de CRM da empresa e informações cadastrais dos clientes no período 1 a 31 de julho de 2015. A segunda base é disponibilizada pela Anatel e contêm as reclamações dos clientes no período de 1 a 31 de agosto de 2015.

Tabela 1 – Atributos utilizados

Nome do Atributo	Tipo
Reclamações no CRM	Textual
Número de reclamações na Anatel	Discreto
Tempo de instalação em meses	Discreto
Linha de outra operadora	Discreto
Idade	Discreto
Sexo	Categórico
Ocupação	Categórico
Dependentes	Discreto
Estado civil	Categórico
Classe social	Categórico

Tipo de residência	Categórico
Escolaridade	Categórico
Contexto das reclamações	Discreto
Peso da reclamação mais critica	Discreto
Soma do peso das reclamações	Discreto
Cidade	Nominal
Total	**16**

Para os experimentos foram selecionadas dez mil instâncias já balanceadas em duas classes alvos – Migrou e Não Migrou –, ou seja, cinco mil instâncias de clientes que migraram para a Anatel e cinco mil de clientes que não migraram para o órgão.

4.2. Experimentos

Durante os experimentos os dois modelos de dados foram submetidos a execução nos principais algoritmos de classificação (árvores de decisão (J48), Naïve Bayes, K-NN, SVM e Redes Neurais), a fim de identificar qual algoritmo obtêm os melhores resultados nessa configuração de dados.

Devido os bons resultados encontrados no estado da arte e sua ampla utilização é adotado nos experimentos a técnica de validação cruzada, onde o conjunto de dados foi dividido em *k* partes iguais, no experimento *k* é igual a 10, onde *k*-1 partes são utilizadas para o treinamento e a *k* parte selecionada é utilizada para o teste, repetindo esse processo até que as *k* partes tenham sido testadas.

A Rede Neural utilizada no experimento foi configurada com taxa de aprendizado 0,3 e um número de épocas igual a 500, o momentum empregado foi 0,2 e possui 5 neurônios na camada oculta da rede neural. A execução dos algoritmos foi dividida em duas etapas, denominadas de etapas de treinamento e teste. Os conjuntos de dados foram divididos igualmente para o treinamento e teste de forma aleatória.

Por fim, para auxiliar o processo de execução dos algoritmos é utilizado o software Weka[2] (*Waikato Enviroment for Knowledge Analysis*) que é desenvolvido pela Universidade de Waikato sendo escrito em Java e publicado sob a licença GPL[3]. O Weka suporta a execução diversos algoritmos de mineração de dados, onde posteriormente é possível realizar a visualização e interpretação dos dados.

4.3. Resultados

Os resultados apresentados referem-se à aplicação da etapa de mineração de dados do processo de KDD, que tem por objetivo identificar clientes que migraram do ambiente interno de atendimento (CRM) da empresa para a Anatel, por meio dos principais algoritmos de classificação sob dois conjuntos de dados distintos.

A aplicação dos algoritmos de mineração dados sob os dois conjuntos de dados resultou que a base de dados que utilizou as reclamações dos clientes enriquecendo-a por meio da mineração de textos, denominada base 1, obteve resultados superiores de acurácia do que o modelo que não utilizou esses novos atributos, essa base é definida como base 2. Sendo assim, os resultados para a acurácia no primeiro conjunto de dados

[2] http://www.cs.waikato.ac.nz/
[3] https://www.gnu.org/licenses/gpl-3.0.en.html

foram: para as árvores de decisão 80,50%, para o SVM 79,45%, para o Naïve Bayes 75,65%, para o K-NN 69,65% e para as redes neurais 52,30%. O conjunto de dados que não utilizou os novos atributos (base 2) obteve os seguintes resultados: para as árvores de decisão 77,90%, para o SVM 72,25%, para o Naïve Bayes 68,45%, para o K-NN 66,65% e para as redes neurais 53,90%.

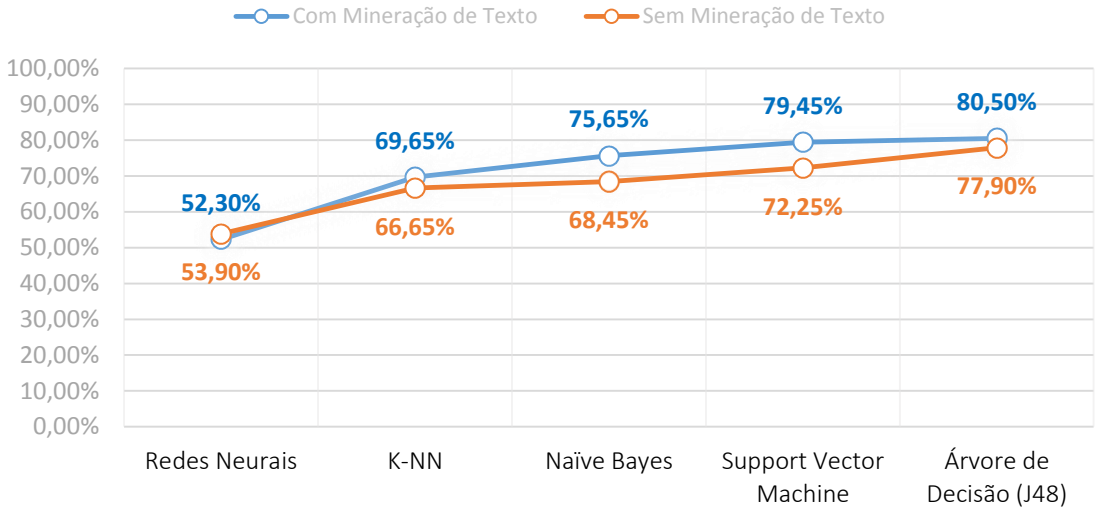

Figura 1 – Acurácia (%) dos experimentos com dados textuais

Na comparação individual dos algoritmos sob os dois modelos de dados, apenas redes neurais obteve resultado superior no conjunto de dados com as entradas tradicionais (base 2), todos os outros algoritmos obtiveram resultados superiores no modelo de dados que utilizou os novos atributos gerados pela mineração de textos (base 1). Sendo assim, na avaliação dos dois modelos de dados é constatado que o modelo que utiliza os novos atributos é superior ao modelo que não os utiliza, exceto redes neurais que obteve resultado superior na base 2 mas que no geral possui a pior taxa de acurácia de todos os modelos.

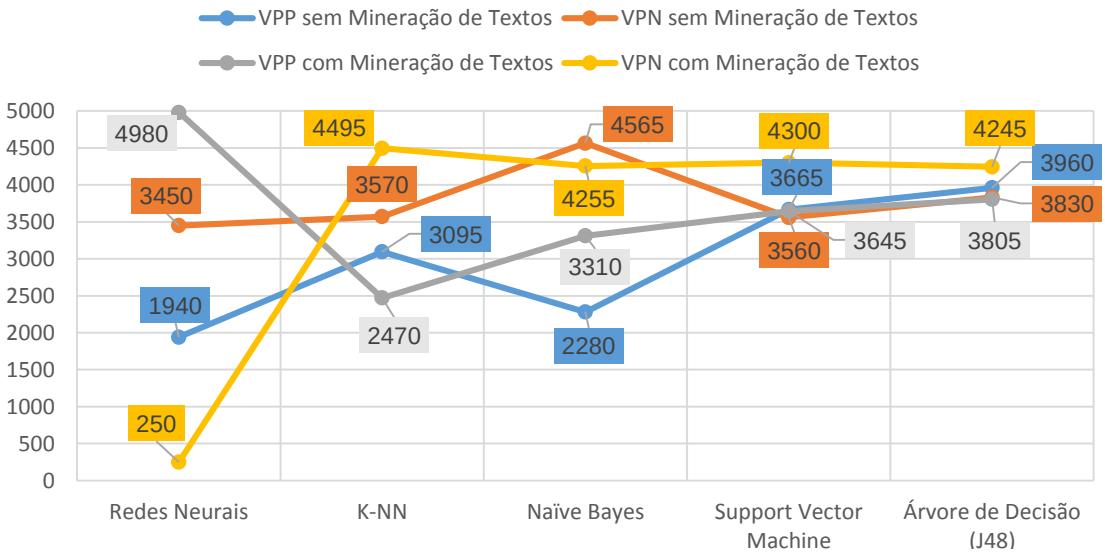

Figura 2 – Valor Preditivo Positivo e Valor Preditivo Negativo obtido para cada classificador

A figura 2 apresenta os valores dos preditivos positivos (VPP) e dos valores preditivos negativos (VPN) obtidos com o uso de cada classificador, onde VPP é dado pela fórmula: VPP = acertos positivos / total de predições positivas e VPN por VPN = acertos negativos / total de predições negativas. Essa análise detalhada dos resultados da mineração de dados permite identificar se houve ou não melhora nos resultados da execução dos algoritmos ao modificar os conjuntos de dados ou a configuração dos algoritmos.

A partir dos resultados apresentados na Figura 2 é possível observar que redes neurais no segundo conjunto de dados possui VPP de 4980 para a classe que identifica que o cliente migrou para a Anatel, em contrapartida, o VPN para a classe que identifica que o cliente não migrou para a Anatel é extremamente baixa com 250, tornando o classificador pouco interessante para o experimento.

As matrizes de confusão dos algoritmos K-NN e Naïve Bayes apresentaram valores de VPN e VPP menos discrepantes do que os encontrados em redes neurais, porém a variação é bem distinta. SVM e árvores de decisão foram os algoritmos com o melhor desempenho nos experimentos, os resultados de suas matrizes de confusão mostram que não existe discrepância entre os resultados, ou seja, VPN e VPP são similares, confirmando assim sua regularidade e consistência nos resultados.

5. Conclusão e Trabalhos Futuros

Os experimentos realizados nessa pesquisa buscam comprovar a melhoria dos resultados com a inserção de novos atributos gerados pela mineração de textos sob as reclamações de clientes em ambientes de centrais de gerenciamento do relacionamento com o cliente.

Na comparação dos dois conjuntos de dados ficou comprovado que o uso da mineração de textos contribui para o aumento da acurácia dos algoritmos de classificação, devido a geração de novos atributos a partir das reclamações dos clientes. Dentro dos experimentos realizados os melhores resultados ficaram por conta de árvores de decisão J48 e SVM, que apresentaram-se consistentes e confiáveis. Contudo, sabemos que ainda é possível atuar na fase de pré-processamento da etapa de mineração de textos, testando e avaliando novas formas de ponderação dos termos, buscando identificar os atributos mais relevantes para o problema em questão.

O KDD e a mineração de dados se mostrou comprovadamente uma técnica útil para a extração de conhecimento novo para este problema, onde por meio da aplicação de diversas técnicas foi possível o desenvolvimento de um modelo de classificação que realizasse a predição da classe alvo do modelo em questão. A classificação dessas instâncias permite que tomadores de decisão possam agir de forma a identificar e tentar prevenir que clientes saiam do ambiente interno de atendimento de empresas de telecomunicações e migrem para ambientes externos de atendimento como a Anatel.

Referências

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis. *Life Science Journal*, 11(3), 75-81.

Almana A. M., Aksoy M. S., Alzahrani R. (2014). A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. *Journal of Engineering Research and Applications,* (4):165-171.

Antunes, C. M., & Oliveira, A. L. (2001, August). Temporal data mining: An overview. In *KDD workshop on temporal data mining* (Vol. 1, p. 13).

Bouckaert, Remco R., et al. "WEKA Manual for Version 3-7-8." *Hamilton, New Zealand* (2013).

Chang, C. W., Lin, C. T., Wang, L. Q. (2009). Mining the text information to optimizing the customer relationship management. *Expert Systems with applications*, 36(2), 1433-1443.

Dogan, N., & Tanrikulu, Z. (2013). A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management*, 14(2), 105-124.

Deulkar, Miss Deepa S., and R. R. Deshmukh. Data Mining Classification. *Imperial Journal of Interdisciplinary Research* 2.4 (2016).

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *Al magazine,* 17(3):37.

Femina, B. T., Sudheep, E. M. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science*, (46):725-731.

Hadden, J., Tiwari, A. Roy, R., Ruta, D. (2006). Churn Prediction using complaints data. In *Proceedings Of World Academy Of Science*, Engineering and Technology.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques.* Elsiever.

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A Brief Survey of Text Mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*, 40(14), 5635-5647.

Hung, S. Y., Yen D. C., Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, (31):515-524.

Maimon, O., & Rokach, L. (2010). *Data mining and Knowledge discovery handbook*. New York: Springer.

Ngai, E. W., Xiu, L., Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert system with applications*, 36(2), 2592-2602

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70).

Tan, P. N., Steinback, M., Kumar, V. (2009). *Introdução ao datamining: mineração de dados.* Ciência Moderna.

Zhang, N., & Lu, W. F. (2007, June). An Efficient Data Preprocessing Method for Mining Customer Survey Data. In *Industrial Informatics, 2007 5th IEEE International Conference on* (Vol. 1, pp. 573-578). IEEE.

Weiss, G. M. (2005). Data Mining in Telecommunications. *Data Mining and Knowledge Discovery Handbook*, pages 1189-1201. Springer.