A Sequential Pattern Detection and Sentiment Analysis Combined Approach to the Churn Prediction Problem in Client Relationship Management Environments

Thiago P. Pimentel[1], **Ronaldo R. Goldschmidt**[1]

[1]Seção de Engenharia da Computação (SE/8) – Instituto Militar de Engenharia (IME)
22.290-270 – Rio de Janeiro – RJ – Brazil

thiago.pimentel@ime.eb.br ,ronaldo.rgold@ime.eb.br

Abstract. *The cost of losing profitable customers in competitive markets is driving companies to engage in customer retention. Therefore, anticipating client churn (i.e., cancellation) becomes essential. Among the researches on churn prediction models, we highlight those that are based on sequential pattern detection. Although promising, such initiatives do not take into account the sentiments present in the client's interactions with the company. Given the above, this article proposes a method that generates churn prediction models from the combination of sequential pattern detection with sentiment extraction from the interactions with the clients. Experimental results confirm the adequacy of the proposed method.*

Resumo. *O custo de perder clientes rentáveis em mercados competitivos tem levado as empresas a atuar na retenção de clientes. Logo, antecipar o churn (i.e. cancelamento) de clientes torna-se essencial. Entre as pesquisas em modelos preditivos de churn, destacam-se as que se baseiam na detecção de padrões sequenciais. Embora promissoras, tais iniciativas não levam em consideração os sentimentos presentes nas interações do cliente com a empresa. Diante do exposto, o presente artigo propõe um método que gera modelos de predição de churn a partir da combinação da detecção de padrões sequenciais com a extração de sentimentos a partir das interações com clientes. Resultados experimentais comprovam a adequação do método proposto.*

1. Introdução

Devido ao crescimento da concorrência, a maior parte dos mercados está cada vez mais saturada. Como consequência, as empresas vêm percebendo que suas estratégias de negócio devem priorizar a manutenção dos clientes atuais [Coussement and Poel 2009]. Neste cenário, ganham relevância as iniciativas de gerenciamento do relacionamento com o cliente (em inglês, *CRM - Client Relationship Managemenent*) dentro das organizações. O *CRM* é um abordagem de gerenciamento que busca criar, desenvolver e aprimorar os relacionamentos com clientes cuidadosamente segmentados para maximizar o valor do cliente e a rentabilidade corporativa [A. Payne 2005]. Um dos principais desafios enfrentados pelo *CRM* é a identificação de clientes com propensão ao *churn* (i.e., cancelamento [Burez and Van den Poel 2007]) de produtos e/ou serviços [Hadden et al. 2007].

Diversos trabalhos de pesquisa têm buscado a criação de modelos que tentam detectar antecipadamente a ocorrência de *churn* [García et al. 2017]. A maioria desses

trabalhos envolve a aplicação de técnicas de aprendizado de máquina sobre dados de perfil dos clientes e de suas interações com as empresas. Algumas dessas pesquisas utilizam métodos de detecção de padrões sequenciais de comportamento [Chiang et al. 2003]. Tais métodos consideram a ordem cronológica e os tipos de interações entre o cliente e a empresa que precedem os eventos de *churn*, para identificar padrões de propensão ao desligamento. Embora promissores, tais métodos deixam de considerar um aspecto muitas vezes presente nas conversas registradas entre cliente e empresa e que pode fornecer indícios importantes para criação de modelos de prevenção de *churn*: o sentimento manifestado durante a interação. De outro lado, estão os trabalhos de pesquisa que, embora explorem os sentimentos extraídos das interações entre cliente e empresa [Coussement and Poel 2009], não consideram a cronologia de ocorrência das informações de sentimento. Se limitam a calcular estatísticas que expressam de forma consolidada, percentuais associados aos sentimentos identificados. Com a consolidação dos dados, deixam de observar possíveis padrões que precedam a ocorrência de *churn*.

Diante do exposto, o presente trabalho levanta a hipótese de que utilizar informações sobre possíveis sentimentos presentes nas interações entre cliente e empresa na ordem em que elas ocorrem ao longo do tempo pode levar à identificação de modelos mais precisos do que os do estado da arte na detecção preventiva de *churn*. Assim, este artigo tem como objetivo apresentar evidências experimentais que confirmem a hipótese levantada. Para tanto, o artigo propõe um método que combina técnicas de mineração de padrões sequenciais e de análise de sentimento, aplicando-as sobre informações acerca das interações ocorridas longitudinalmente entre clientes e empresa. Concebido para ser aplicado em qualquer contexto onde existam dados históricos sobre os contatos realizados junto aos clientes, o método proposto produziu, nos experimentos realizados, resultados que apontam para a validade da hipótese formulada.

O presente texto encontra-se organizado em mais cinco seções. A seção 2 apresenta fundamentos sobre detecção de padrões sequenciais, necessários à compreensão do método proposto neste artigo. Na seção 3, são apresentados os trabalhos relacionados ao tema, enfatizando suas diferenças em relação ao método proposto. A descrição formal do método proposto encontra-se na seção 4. Detalhes sobre os experimentos realizados e as análises sobre os resultados obtidos estão na seção 5. Por fim, na seção 6, são destacadas as principais contribuições deste artigo e apresentadas possíveis alternativas de trabalhos futuros.

2. Fundamentos em Mineração de Padrões Sequenciais

A tarefa de mineração de padrões sequenciais é uma extensão da tarefa clássica de mineração de regras de associação a fim de considerar a cronologia de ocorrência de eventos em um conjunto de dados. A seguir, são apresentados e exemplificados os principais formalismos envolvendo essas tarefas.

Considere D um *dataset* que contém um conjunto $\tau = \{r_1, r_2, ..., r_n\}$ de fatos observados, denominados de transações, e um conjunto de itens $\Sigma = \{i_1, i_2, ..., i_m\}$. Cada item $i \in \Sigma$ corresponde a um predicado definido sobre τ, da seguinte forma: dada uma transação $r \in \tau$, diz-se que r satisfaz i se, e somente se i é verdadeiro em r (i.e., i ocorre em r). Notação adotada: r[i] = 1 (r satisfaz i), se i é verdadeiro em r, e, r[i]= 0 (r não satisfaz i), se i é falso em r. Adicionalmente, seja C um conjunto de itens (i.e.,

$C \subseteq \Sigma$). Denomina-se suporte de C como $Sup(C) = \frac{|C|}{|\tau|}$, onde $\mid C \mid = card\{r \in \tau \mid \forall i \in C, \tau[i] = 1\}$. Assim, define-se como regra de associação toda implicação da forma R: X $\rightarrow$ Y, onde X, Y $\subset \Sigma$ e X $\cap$ Y $\neq \emptyset$. Por exemplo, $R_1 : \{Canal = Email, Sentimento = Negativo\} \rightarrow \{Churn = Sim\}$ é uma regra de associação onde $Canal = Email$, $Sentimento = Negativo$ e $Churn = Sim$ são itens. Satisfazem a R_1, todas as transações do conjunto de dados que satisfazem (i.e., tornam verdadeiros) esses três itens ao mesmo tempo. Uma regra de associação $R : X \rightarrow Y$ é dita frequente (resp. válida) se, e somente se, $Sup(R) = Sup(X \cup Y) \geq$ ***MinSup*** (resp. $Conf(R) = Sup(X \cup Y)/Sup(X) \geq$ ***MinConf***), onde ***MinSup*** e ***MinConf*** são parâmetros definidos pelo usuário. Por fim, consideradas as definições acima, a tarefa de mineração de regras de associação consiste em encontrar todas as regras de associação frequentes e válidas existentes em um conjunto de dados.

A mineração de padrões frequentes e válidos é uma extensão da tarefa de mineração de regras de associação e requer algumas definições adicionais. A primeira delas é que em D exista um conjunto de objetos $O = \{o_1, o_2, ..., o_k\}$ aos quais as transações de τ estejam vinculados. Desta forma, cada transação $r \in \tau$ passa a ser caracterizada por $r = (t, o, s)$, onde $s \subseteq \Sigma$ corresponde a um conjunto de itens, $t \in \Re$ representa um rótulo temporal e o indica o objeto associado a transação r. Sendo assim, define-se sequência de eventos associados a um objeto o como um conjunto ordenado de transações associadas a o: $S_o = \{r_i = (t_i, o_i, s_i) \mid o_i = o,\ i \in \{1, 2, ..., n\}\}$, ou, de forma simplificada, $S_o = \ll s_{k_1}, s_{k_2}, ..., s_{k_r} \gg$, onde $k_1 < k_2 < ... < k_r$. Ainda neste contexto, considere P um conjunto ordenado cujos os elementos são conjuntos de itens $P = \ll p_1, p_2, ..., p_r \gg$. Diz-se que P é um padrão sequencial se, e somente se existem $o \in O$ e S_o tais que $P = S_o$. Em seguida, definem-se conjunto suporte de um padrão sequencial P como sendo $\rho_P = \{o \in O \mid \exists S_o, P = S_o\}$, e o suporte de P como $Sup(P) = \mid \rho_P \mid$. Diz-se ainda que um padrão sequencial P é frequente se, e somente se, $Sup(P) \geq minsup$, sendo $minsup$ um limiar definido pelo usuário. Considerando $P = \ll p_1, p_2, ..., p_r \gg$ um padrão sequencial, diz-se que $\ll p_1, p_2, ..., p_{r-1} \gg$ e $\ll p_r \gg$ são prefixo ($Pref(P)$) e sufixo de P ($Suf(P)$), respectivamente. Assim, define-se confiança de um padrão sequencial P por $Conf(P) = \frac{Sup(P)}{Sup(Pref(P))}$. E, por fim, um padrão sequencial P é dito válido se, e somente se, $Conf(P) \geq minconf$, sendo $minconf$ um limiar de confiança mínima definido pelo usuário. Diante das definições acima, a tarefa de mineração de padrões sequenciais consiste em identificar padrões sequenciais frequentes e válidos em um conjunto de dados D. Entre os exemplos de algoritmos que implementam a tarefa de mineração de padrões sequenciais podem ser destacados: GSP [R. Srikant 1996], SPADE [Zaki 2001], PrefixSpan [J. Pei 2001], GSpan [X. Yan 2002] e BIDE [J. Wang 2004].

3. Trabalhos relacionados

Diversas iniciativas de pesquisa têm buscado a criação de modelos de predição de *churn* [García et al. 2017]. Em geral, elas utilizam algoritmos de aprendizado de máquina para construir modelos de classificação que, diante de informações sobre perfil de cliente e sobre as interações ocorridas entre cliente e empresa, buscam inferir uma dentre as possíveis classes: *churn* e não *churn*. Basicamente, as pesquisas em predição de *churn* podem ser organizadas em três abordagens. A tabela 1 sumariza os referidos trabalhos e abordagens. A seguir cada abordagem encontra-se indicada e comentada.

Tabela 1. Visão Resumida dos Trabalhos Relacionados

Referências	Padrões Sequenciais	Análise de Sentimentos	Algoritmos
[Chiang et al. 2003]	sim	não	Detecção de Padrões Sequenciais
[Jenamani et al. 2003]	não	não	Semi-Markov
[Jonker et al. 2004]	não	não	Algoritmos Genéticos
[Larivière and Van Den Poel 2005]	não	não	Análise de Sobrevivência
[Buckinx and Van Den Poel 2005]	não	não	Floresta Aleatória
[Goldschmidt and Passos 2005]	não	não	Floresta Aleatória
[Liu and Shih 2005]	não	não	Filtro por Preferência
[Slotnick and Sobel 2005]	não	não	Semi-Markov
[De Bock et al. 2010]	não	não	Markov
[Burez and Van den Poel 2007]	não	não	Markov e FA
[Kumar and Ravi 2008]	não	não	Floresta Aleatória
[Coussement and Van den Poel 2008]	não	não	Floresta Aleatória
[Coussement and Poel 2009]	não	sim	Regressão Logística, SVM e Floresta Aleatória
[Wu 2009]	não	não	Híbrido de Floresta Aleatória e Redes Neurais
[De Bock et al. 2010]	não	não	Classificação
[Verbeke et al. 2012]	não	não	Métodos de Conjunto
[Zhang et al. 2012]	não	não	Híbrido envolvendo Floresta Aleatória, Regressão logística e Redes Neurais
[Coussement and De Bock 2013]	não	não	Floresta Aleatória

A maioria dos trabalhos de predição de *churn* segue a primeira abordagem. Tais trabalhos desenvolvem modelos que consideram informações sobre o perfil dos clientes tais como idade, gênero, classe social, renda média, cidade onde mora, dentre outras, além de dados estatísticos consolidados sobre a relação entre cliente e empresa como, por exemplo, a quantidade de produtos adquiridos, o gasto médio mensal, e o tempo de relacionamento. São exemplos de algoritmos de aprendizado de máquina frequentemente utilizadas em trabalhos desta abordagem: redes neurais artificiais [Zhang et al. 2012], árvores de decisão [Wu 2009], floresta aleatória [Buckinx and Van Den Poel 2005] [De Bock et al. 2010], entre outros. Cabe ressaltar que nenhum dos trabalhos desta abordagem leva em consideração nem os sentimentos registrados nas interações entre os clientes e as empresas e nem o aspecto temporal quanto à ocorrência dessas interações.

Na segunda abordagem estão as pesquisas que propõem métodos de detecção de *churn* baseados em padrões sequenciais de comportamento. Tais métodos consideram a ordem cronológica e os tipos das interações entre o cliente e a empresa que precedem os eventos de *churn*. Um exemplo dessas pesquisas é o trabalho de [Chiang et al. 2003]. Nele, os autores utilizam um método de detecção de sequências inspirado nos algoritmos Apriori [Agrawal and Srikant 1994] e GSP [R. Srikant 1996] para identificar padrões de comportamento de clientes que incorreram em *churn*. O método utilizado estende os algoritmos clássicos ao utilizar o conceito de sequência invertida. Neste conceito, cada sequência frequente minerada pelo algoritmo é invertida, de forma que os eventos mais recentes passam a ser apresentados primeiro na sequência. Com isso, o método pode priorizar mais facilmente as sequências de maior relevância para identificação de *churn*. Esta abordagem tem a limitação que não leva em consideração a informação de sentimentos advindos dos registros das interações no *CRM* da empresa.

Já na terceira abordagem estão as pesquisas que exploram os sentimentos extraídos das interações entre o cliente e a empresa [Coussement and Poel 2009]. Para tanto, essas

pesquisas justificam que o cliente, ao relacionar-se com a empresa, apresenta informações implícitas sobre seus sentimentos que podem agregar valor aos modelos preditivos, uma vez que essas estão diretamente ligadas à satisfação do cliente, podendo sinalizar propensão ao *churn*. No entanto, tais pesquisas não consideram o aspecto temporal de ocorrência das informações de sentimento. Se limitam a calcular estatísticas que, de forma análoga aos trabalhos da primeira abordagem, expressam, de forma consolidada, percentuais associados aos sentimentos identificados. Com a consolidação dos dados, deixam de observar possíveis padrões que precedam a ocorrência de *churns*.

Diante do exposto, cabe enfatizar que, diferentemente do método proposto neste artigo, em nenhum dos trabalhos relacionados, foi possível observar alguma iniciativa voltada à combinação simultânea de informações sobre possíveis sentimentos presentes nas interações entre cliente e empresa com a ordem em que essas interações ocorrem ao longo do tempo.

4. Método proposto

Denominado *SS-DetChurn*, o método proposto neste artigo considera simultaneamente possíveis sentimentos presentes nas interações entre cliente e empresa com a ordem em que essas interações ocorrem ao longo do tempo, a fim de construir modelos de detecção preventiva de *churn*. A Figura 1 ilustra graficamente as etapas do *SS-DetChurn*. O funcionamento das etapas desse processo será detalhado a seguir. Para tanto, serão utilizados alguns conceitos e respectivos formalismos apresentados na Seção 2.

Cabe destacar inicialmente que o *SS-DetChurn* requer a existência de $n \geq 1$ conjuntos de dados que contenham os registros históricos do conteúdo das interações realizadas entre a empresa e seus clientes por meio de canais de comunicação $C_1, C_2, ..., C_n$ tais como *E-mail*, *chat*, telefone, Portal *Online*, dentre outros. Também deverá existir uma fonte de dados que indique a situação do cliente com relação ao seu vínculo junto à empresa (i.e., *churn=sim* ou *churn=não*). O método requer ainda que o conjunto de dados associado a cada C_i possua, para cada interação, no mínimo, as seguintes informações: o momento em que a interação ocorreu, o cliente com o qual a interação ocorreu, o que assunto que levou à interação e a transcrição da conversa registrada no momento da interação.

4.1. Etapa 1 - ETC (Extração, Transformação e Carga) de Dados

Esta etapa é responsável por realizar todas as operações de seleção, adequação e integração dos dados que serão utilizados nas etapas seguintes do método proposto. Deve produzir como saída um conjunto de dados D cuja estrutura está indicada na figura 2. Ela contém quatro atributos que indicam, para cada transação de interação r ocorrida, a data t de ocorrência de r, a identificação do cliente o com o qual r ocorreu, e o conjunto de itens s associado a r. Convém notar que s pode conter diversos itens tais como o canal c que registrou r, o assunto m que motivou a ocorrência de r, a transcrição da conversa v registrada em r, dentre outros, incluindo a indicação de ocorrência de *churn*. As interações apresentadas como exemplo na figura 2 pertencem a um contexto de natureza educacional.

A informação sobre ocorrência ou não de *churn* deve ser obtida no conjunto de dados que contém os cadastros dos clientes da empresa. Assim, para cada cliente o, após

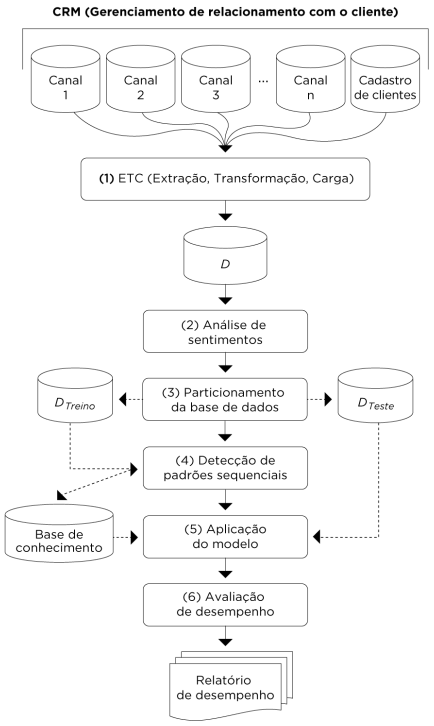

Figura 1. Visão Macro-Funcional do *SS-DetChurn*

a carga em D de todas as interações registradas nos canais de comunicação em que o participou, inclui-se uma interação adicional r vinculada a o onde o conjunto de itens associado a r contém apenas um item: *churn=sim* ou *churn=não*, retratando, assim, a situação de o no cadastro de clientes. A data t da interação r recebe a data do desligamento informada no cadastro, caso o cliente tenha se desligado. Caso contrário, t recebe a data de execução da etapa de ETC.

Transação	Data da Interação	Cliente	Conjunto de Itens
1	08/06/2017	1	{canal=email, motivo=Problemas com acesso, Conversa= "Prezado Thiago,..."}
2	03/07/2017	1	{canal=email, motivo=Reclamação sobre o curso, Conversa= "Olá Maria,..."}
3	05/07/2017	1	{canal=online, motivo=Problemas com acesso, Conversa= "Prezado Paulo,..."}
4	28/12/2017	1	{churn=sim}
5	03/02/2016	2	{canal=online, motivo=Reclamação sobre o curso, Conversa= "Caro Marcos,..."}
6	22/05/2017	2	{canal=telefone, Solicitação de 2ª via de senha, Conversa= "Alô,..."}
7	20/12/2017	2	{churn=sim}
8	30/03/2017	3	{canal=chat, Solicitação de 2ª via de senha, Conversa= "Olá Maria,..."}
9	07/11/2017	3	{churn=não}

Figura 2. Estrutura do conjunto de dados gerado pela Etapa 1 do *SS-DetChurn* ilustrada com exemplos.

4.2. Etapa 2 - Análise de sentimentos

Este ponto do processo consiste em aplicar, para cada transação r em D, um algoritmo que seja capaz de classificar o texto da conversa v associada a r quanto à polaridade do sentimento subjacente a v: positivo, negativo. Ao final do processo, o resultado da

classificação da polaridade associada à v é usado para atualizar o conjunto de itens s de r da seguinte forma: a polaridade identificada em v é inserida em s, ao mesmo tempo em que v é excluída de s. A figura 3 mostra o conjunto de dados do exemplo da figura 2 após a execução da etapa de análise de sentimentos.

Transação	Data da Interação	Cliente	Conjunto de Itens
1	08/06/2017	1	{canal=email, motivo=Problemas com acesso, Sentimento= negativo}
2	03/07/2017	1	{canal=email, motivo=Reclamação sobre o curso, Sentimento= negativo}
3	05/07/2017	1	{canal=online, motivo=Problemas com acesso, Sentimento= negativo}
4	28/12/2017	1	{churn=sim}
5	03/02/2016	2	{canal=online, motivo=Reclamação sobre o curso, Sentimento= negativo}
6	22/05/2017	2	{canal=telefone, Solicitação de 2ª via de senha, Sentimento= positivo}
7	20/12/2017	2	{churn=sim}
8	30/03/2017	3	{canal=chat, Solicitação de 2ª via de senha, Sentimento= positivo}
9	07/11/2017	3	{churn=não}

Figura 3. Exemplo da figura 2 transformado pela Etapa 2 do *SS-DetChurn*.

Potencialmente, qualquer algoritmo de classificação de polaridade como, por exemplo, o *NNLM feedforward*, entre outros [W. Medhat 2014, T. Mikolov and Zweig 2013], pode ser utilizado na implementação desta etapa. No entanto, cabe ao analista de dados a escolha de qual algoritmo deverá ser aplicado.

4.3. Etapa 3 - Particionamento da base de dados

Esta etapa consiste em dividir D aleatoriamente em dois subconjuntos, D_{Treino} e D_{Teste}, tais que não existam clientes que possuam interações em D_{Treino} e D_{Teste} simultaneamente. Para tanto, o usuário deve especificar qual deve ser a proporção resultante de clientes nos dois subconjuntos. Um método de amostragem aleatória estratificada [K. Faceli 2015] deve assegurar que a distribuição de classes *churn* e não *churn* existente em D seja preservada nos dois subconjuntos resultantes.

4.4. Etapa 4 - Detecção de padrões sequenciais

Nesta etapa, deve ser executado algum algoritmo de detecção de padrões sequenciais compatível com as definições apresentadas na seção 2. Diante da diversidade de algoritmos desta natureza, cabe ao analista de dados optar por um deles. Em essência, o algoritmo escolhido precisa identificar padrões sequenciais P frequentes e válidos de tal forma que $Suf(P) =\ll churn = sim \gg$ ou $Suf(P) =\ll$ *churn=não* $\gg$. A escolha dos parâmetros de suporte e confiança mínimos também deve ser feita pelo analista de dados. O algoritmo escolhido deve ser aplicado sobre D_{Treino}.

Após a identificação dos padrões sequenciais frequentes e válidos em D_{Treino}, esses são armazenados em uma base de conhecimento para serem aplicados na etapa seguinte.

4.5. Etapa 5 - Aplicação do Modelo

Seguindo o processo, esta etapa tem como objetivo aplicar nos registros de dados de D_{Teste} os padrões sequenciais minerados anteriormente. Assim, para cada cliente o em D_{Teste}, verifica-se se existe alguma sequência S_o em D_{Teste} e algum padrão sequencial P

na base de conhecimento tal que $Pref(P) = Pref(S_o)$. Em caso positivo, compara-se se $Suf(P) = Suf(S_o)$. Caso a comparação seja verdadeira, é computado um acerto do modelo. Caso a comparação seja falsa, computa-se um erro. Caso não existam sequência S_o em D_{Teste} e padrão sequencial P na base de conhecimento tal que $Pref(P) = Pref(S_o)$, assume-se *churn=não* a fim de realizar a comparação com $Suf(S_o)$ e, então, computar erro/acerto.

4.6. Etapa 6 - Avaliação de desempenho

Conforme o próprio nome sugere, esta etapa tem como objetivo calcular alguma medida que expresse o grau de adequação do modelo gerado em detectar antecipadamente a ocorrência de *churn*. Tal medida deve ser função da quantidade de erros e acertos obtidos a partir da aplicação do modelo no conjunto D_{Teste}. *Precisão*, *Acurácia*, *Taxa de Falsos Positivos*, e *Área Sob a Curva*, entre outras, são exemplos de medidas popularmente utilizadas na avaliação de desempenho de problemas preditivos [K. Faceli 2015]. Também neste ponto, cabe ao analista de dados escolher a métrica de avaliação a ser utilizada.

5. Experimentos e resultados

A fim de avaliar a hipótese levantada neste trabalho, os experimentos foram realizados com os dados de *CRM* de uma universidade privada brasileira. Mais especificamente, foram considerados históricos de interação entre a universidade e seus alunos de cursos de pós-graduação ocorridos durante o ano de 2017 por meio de quatro canais de comunicação: *E-mail*, *chat*, *voz* e Portal *Online*. Cabe ressaltar que a escolha deste cenário deveu-se basicamente à disponibilidade de acesso dos autores deste trabalho aos dados necessários aos experimentos. A tabela 2 apresenta um resumo estatístico das características dos dados desse cenário. A situação de *churn* /não *churn* de cada aluno foi obtida diretamente do sistema de controle acadêmico da instituição.

Tabela 2. Sumário estatístico dos dados utilizados nos experimentos

Dados	Descrição
Amostra	49.013 alunos de pós-graduação
Período	12 meses (2017)
Distribuição	87% de não churn e 13% de churn
Número Interações por Aluno	mínimo: 1 máximo:11 médio:6

Inicialmente, na etapa de *ETC*, foram realizadas operações de seleção e limpeza dos dados disponíveis nos diferentes canais de comunicação. Em seguida, os dados foram formatados e reunidos em uma tabela única. Cabe enfatizar que a existência de um cadastro único de alunos integrado aos bancos de dados dos diferentes canais evitou a necessidade de operações de deduplicação de clientes para integração dos dados. Outro fator facilitador foi a existência de um conjunto único de motivos usado pelos canais para caracterizar de forma estruturada os assuntos tratados nas interações. Tal conjunto foi utilizado no enriquecimento dos conjuntos de itens das interações no momento de carga dos dados. A tabela 3 indica de forma sumarizada a quantidade de motivos agrupados em função da natureza da interação.

Tabela 3. Interação Cliente e Empresa - Natureza x Motivo

Natureza	Motivo
Administrativa	24
Acadêmica	16
Financeira	3

Após as operações de *ETC*, foi executada a etapa de análise de sentimentos. Nela, foi aplicado o algoritmo *NNLM feedforward* [T. Mikolov and Zweig 2013] sobre o texto da conversa de cada interação r a fim de obter a polaridade (positiva ou negativa) do sentimento associado e enriquecer com tal informação o conjunto de itens de r. Cabe destacar que, para executar a classificação, o *NNLM feedforward* utiliza o *word2vec* na representação dos textos e uma rede neural recorrente treinada a partir de um corpus específico.

Na etapa de detecção de padrões sequenciais, foi utilizado o algoritmo BIDE [J. Wang 2004] com os suporte e confiança mínimos de 30% e 80%, respectivamente. Tal algoritmo toma como base os conceitos apresentados na seção 2 a fim de identificar padrões sequenciais frequentes e válidos no conjunto de dados.

A fim de permitir a comparação do efeito da utilização dos sentimentos na geração dos modelos preditivos, foram considerados dois cenários. No primeiro, o método proposto foi integralmente aplicado. Considerou-se, portanto, o uso dos sentimentos na detecção de padrões sequenciais para prevenção de *churn*. No segundo cenário, as informações de sentimento identificadas na etapa 3 não foram consideradas na construção dos modelos preditivos. Em ambos os cenários, a técnica de validação cruzada com *k-conjuntos* foi utilizada na avaliação dos modelos preditivos gerados. Duas métricas de avaliação foram calculadas: acurácia e taxa de falsos positivos. A tabela 4 apresenta os desempenhos dos modelos obtidos na validação cruzada com $k = 5$. A seguir, encontram-se comentados os resultados nesses cenários.

Observa-se que, de uma maneira geral, o modelo obtido sem considerar o atributo de sentimento, obteve um resultado significativo de 70,1% de acurácia. Entretanto, quando inserido o atributo de sentimento, o resultado aumentou em 14,5 p.p., obtendo a acurácia de 84,6%, corroborando para a confirmação da hipótese inicial levantada neste trabalho. Estes números podem ser considerados como resultados satisfatórios, uma vez que a taxa de falsos positivos ficou em torno de 5,5%.

Tabela 4. Desempenho do *SS-DetChurn* em dois cenários: sem e com atributo de sentimento

Métrica de desempenho	Sem atributo de sentimento	Com atributo de sentimento
Taxa de Falsos Positivos	7,2%	5,5%
Acurácia Total	70,1%	84,6%

Outro aspecto interessante a ser comentado e que também aponta para a adequação do método proposto e para a confirmação da hipótese deste trabalho pode ser observado a partir das tabelas 5 e 6. A tabela 5 contém os seis padrões sequenciais válidos com maior frequência identificada pelo algoritmo de mineração (note que o sufixo $churn = sim$ de

cada padrão foi omitido por questão de simplificação, assim como o nome do atributo de cada item). A tabela 6 indica a acurácia de cada um dos seis padrões em duas situações: uma em sua versão integral, com o item sentimento em cada conjunto de itens, e a outra em sua versão simplificada, onde o item sentimento foi removido de cada conjunto de itens. Pode-se perceber que a acurácia de cada padrão apresenta piora quando o atributo de polaridade de sentimento é removido em todos os casos, chegando a uma perda de 39,23% para o padrão 2. Tal fato reforça a importância do uso da informação sobre o sentimento associado a interações em ambientes de CRM na construção de modelos preditivos de *churn*.

Tabela 5. Padrões sequenciais válidos com maior frequência identificados na Etapa 3 do método proposto

Padrão	Descrição	Suporte	Confiança
1	(chat; nota; negativo); (email; solicitação de senha; negativo); (voz; ouvidoria; negativo)	41,56%	86,66%
2	(voz; ouvidoria; negativo); (voz, ouvidoria; negativo); (voz, ouvidoria; negativo)	31,23%	83,20%
3	(chat; nota; negativo); (email; nota; negativo); (voz; 2ª via do boleto; negativo)	34,21%	80,10%
4	(email; 2ª via do boleto; negativo); (email; bilhete atrasado; negativo); (voz; ouvidoria; negativo)	39,47%	86,73%
5	(email; 2ª via do boleto; negativo); (email; ouvidoria; negativo)	47,57%	85,91%
6	(voz; solicitação de senha; negativo); (email; solicitação de senha; negativo)	30,28%	89,33%

Tabela 6. Acurácia dos padrões da tabela 5 em dois cenários: com e sem o atributo de sentimento.

Padrão Sequencial	Sem atributo de sentimento	Com atributo de sentimento	Variação %
1	32,21%	43,20%	34,16%
2	15,30%	21,31%	39,23%
3	9,35%	11,35%	21,39%
4	6,42%	7,43%	15,59%
5	2,29%	2,89%	26,20%
6	1,58%	1,78%	12,66%

6. Considerações finais

Entre as pesquisas em modelos preditivos de *churn* (i.e., cancelamento) de clientes, destacam-se as que se baseiam na detecção de padrões sequenciais. Embora promissoras, tais pesquisas deixam de considerar um aspecto muitas vezes presente nas conversas registradas entre cliente e empresa e que pode fornecer indícios importantes para criação de modelos de prevenção de *churn*: o sentimento manifestado durante a interação. Nesse contexto, este trabalho levantou a hipótese de que utilizar informações sobre possíveis sentimentos presentes nas interações entre cliente e empresa na ordem em que elas ocorrem ao longo do tempo pode levar à identificação de modelos mais precisos do que os do estado da arte na detecção preventiva de *churn*. A fim de buscar evidências experimentais que confirmassem a hipótese levantada, o presente trabalho propôs e avaliou um método de predição de *churn* que considera a combinação da detecção de padrões sequenciais com os sentimentos extraídos a partir das interações com os clientes. Os resultados obtidos apontaram para a adequação do método proposto e para confirmação da hipótese levantada.

Entre as iniciativas de trabalhos futuros estão: comparar *SS-DetChurn* aos trabalhos que modelam o problema de churn usando perfil de clientes e a investigação da

aplicabilidade do método proposto em outras áreas tais como telecomunicações e bancos, a fim de buscar mais evidências de validade da hipótese formulada; o enriquecimento dos conjuntos de itens das interações entre clientes e empresas com outras informações que possam ser captadas de outras fontes de dados do CRM; e a investigação dos impactos de se utilizar outros algoritmos de mineração de padrões frequentes e de análise de sentimentos na implementação do método proposto.

Referências

A. Payne, P. F. (2005). A strategic framework for customer relationship management. pages 167—176. Journal of Marketing Research, 69.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. pages 478–479. Proceedings of the 20th international conference on very large data bases.

Buckinx, W. and Van Den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. volume 164, pages 252–268. European journal of operational research.

Burez, J. and Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. volume 32, pages 277–288. Expert Systems with Applications.

Chiang, D. A., Wang, Y. F., Lee, S. L., and Lin, C. J. (2003). Goal-oriented sequential pattern for network banking churn analysis. volume 25, pages 293–302. Expert Systems with Applications.

Coussement, K. and De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. volume 66, pages 1629–1636. Journal of Business Research.

Coussement, K. and Poel, D. V. d. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. volume 36, pages 6127–6134. Expert Systems with Applications.

Coussement, K. and Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. volume 34, pages 313–327. Expert Systems with Applications.

De Bock, K. W., Coussement, K., and Van Den Poel, D. (2010). Computational Statistics and Data Analysis Ensemble classification based on generalized additive models. volume 54, pages 1535–1546. Computational Statistics and Data Analysis.

García, D. L., Nebot, , and Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. volume 51, pages 719–774. Knowledge and Information Systems.

Goldschmidt, R. R. and Passos, E. (2005). *Data Mining: Um guia prático*. Campus, 2nd edition.

Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. volume 34, pages 2902–2917. Computers and Operations Research.

J. Pei, J. Han, B. M.-A. H. P. Q. C. U. D. C. H. (2001). PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. pages 215—224. IEEE Press.

J. Wang, J. H. (2004). BIDE: Efficient mining of frequent closed sequences. pages 79––90. IEEE Press.

Jenamani, M., Mohapatra, P. K., and Ghose, S. (2003). A stochastic model of e-customer behavior. volume 2, pages 81–94. Electronic Commerce Research and Applications, 1 edition.

Jonker, J. J., Piersma, N., and Van Den Poel, D. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. volume 27, pages 159–168. Expert Systems with Applications.

K. Faceli, A. C. Lorena, J. G. A. d. C. (2015). Inteligência Artificial. Uma Abordagem de Aprendizado de Máquina. LTC.

Kumar, D. A. and Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. pages 4–28. International Journal of Information and Decision Sciences.

Larivière, B. and Van Den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. volume 29, pages 472–484. Expert Systems with Applications.

Liu, D.-R. and Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. volume 42, pages 387–400. Information & Management.

R. Srikant, R. A. (1996). Mining sequential patterns: generalizations and performance improvements. pages 3—17. 5th international conference on extending database technology (EDBT).

Slotnick, S. A. and Sobel, M. J. (2005). Manufacturing lead-time rules: Customer retention versus tardiness costs. volume 163, pages 825–856. European Journal of Operational Research.

T. Mikolov, W. Y. and Zweig, G. (2013). Linguistic regularities in continuous space word representations. pages 746—-751. HLT-NAACL.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. volume 218, pages 211–229. European Journal of Operational Research.

W. Medhat, A. Hassan, H. K. (2014). . pages 1093–1113. Ain Shams Engineering Journal, v5.

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. volume 36, pages 9105–9112. Expert Systems with Applications.

X. Yan, j. H. (2002). gspan: Graph-based substructure pattern mining. ICDM.

Zaki, M. (2001). Spade: an efficient algorithm for mining frequent sequences. pages 31—60. Mach Learn 42.

Zhang, X., Zhu, J., Xu, S., and Wan, Y. (2012). Predicting customer churn through interpersonal influence. volume 28, pages 94–104. Knowledge-Based Systems.