Proceedings of the V Brazilian Conference on Neural Networks – V Congresso Brasileiro de Redes Neurais
pp. 397-402, April 2-5, 2001 – Rio de Janeiro – RJ - Brazil

Mineração de Dados Aplicada a CRM em uma Base de Clientes de Telefonia de Longa Distância

Hugo L. C. Azevedo[1], Marley B.R. Vellasco[2], Emmanuel P. L. Passos[3]
[1,2,3] Laboratório de Inteligência Computacional Aplicada
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marques de S. Vicente, 225, Gávea,
Rio de Janeiro, RJ, Brasil, 22453-900 Brasil
huleo@gbl.com.br, marley@ele.puc-rio.br, emmanuel@ele.puc-rio.br

Abstract

Nowadays, it is very important for retail companies to understand their customers and create relationships with them. It is then crucial to be able to segment the customers according to their buying patterns and needs. Unfortunately, this is not an easy task and several mathematical algorithms and other techniques are required to accomplish it. In this work, a database of a long distance carrier company containing client-calling patterns was clustered and characterized. First, the Kohonen algorithm was used to cluster a subset of the database. The clusters found were then characterized using rules and some multivariate statistical visualization techniques. Finally, a classification model was built to classify future costumers and the ones who were out of the dataset originally clustered.

1. Introdução

Nos mercados de hoje é imprescindível para uma empresa conhecer bem os seus clientes e conseguir criar e gerenciar relacionamentos com eles (CRM- Customer Relationship Management)[9]. Nos mercados mais competitivos, isso torna-se uma questão de sobrevivência. É importante, portanto, para uma empresa, saber agrupar os seus vários clientes em grupos com perfis semelhantes e conhecer as características de cada um desses grupos. Isso permite a ela produzir produtos e serviços específicos para cada grupo, isto é, que atendam às demandas de cada um. O bom desempenho dessa tarefa leva a um nível maior de satisfação e fidelidade dos clientes, a maiores vendas e consequentemente a maiores lucros, objetivo final de qualquer empresa. Quando a empresa possui poucos clientes, ou estes pertencem a grupos já pré definidos e bastante distintos, torna-se fácil para o departamento de marketing desempenhar a tarefa acima. Mas este não é normalmente o caso. Geralmente, as bases de clientes são muito extensas, complexas, estão distribuídas por toda a empresa e possuem muitos atributos, tornando tal tarefa complicada e requerendo técnicas sofisticadas para executá-la.

As tarefas de agrupamento, caracterização dos grupos (*clusters*) e classificação dos clientes são desempenhadas durante um processo chamado de Mineração de Dados (*data mining*), que por sua vez, é a etapa mais importante de um processo maior chamado de Descoberta de Conhecimento em Bancos de Dados (KDD)[5]. Este processo abrange várias etapas e durante a etapa de mineração, normalmente utiliza técnicas estatísticas, visualização de dados, inteligência computacional, entre outras, com o objetivo de extrair algum conhecimento útil e não trivial de uma complexa base de dados.

Neste trabalho, foi feito um processo de KDD envolvendo análise de agrupamentos (*clustering*), caracterização dos grupos encontrados e criação de um modelo de classificação (*pattern classification*) para uma complexa base de clientes de uma operadora de telefonia de longa distância.

2. Descrição do trabalho

2.1. Base de dados

A base de dados utilizada contém a média das 3 últimas faturas mensais de ligações interurbanas entre telefones fixos de clientes de uma operadora de telefonia de longa distância. A base em questão possui dados de 20.000 clientes e é um subconjunto representativo do total de clientes da empresa. A base compreende uma tabela com 49 atributos onde foram condensados os dados de várias tabelas. Os atributos estão relacionados a:

- Faixa de Distância (degrau) entre o chamador e o chamado;
- Dia da semana em que a ligação foi feita;
- Tipo de tarifa da ligação;
- Abrangência da ligação.

As medidas utilizadas para quantificar os atributos acima foram:

- Quantidade de chamadas no mês
- Quantidade de minutos no mês
- Receita no mês.

2.2. Descoberta de conhecimento

As etapas de um processo de KDD[5] compreendem: 0 - Estudo dos objetivos do processo; 1- Limpeza dos dados; 2 - Seleção dos dados; 3 - Codificação dos dados; 4 - Visualização dos dados; 5 - Mineração dos dados; 6 - Interpretação dos Resultados.

Neste trabalho, foram seguidas todas as etapas acima, com maior ou menor profundidade. A etapa 5 foi dividida em 3 fases: agrupamento, caracterização dos grupos encontrados e criação de um modelo de classificação. Vale ressaltar que o processo como um todo é interativo e iterativo já que caminha nos dois sentidos e as etapas são repetidas várias vezes até, se obter um resultado satisfatório.

Do total de 20.000 clientes, foram selecionados, durante a etapa 2, diferentes subconjuntos para se trabalhar. Na seleção dos subconjuntos, tomou-se cuidado em se garantir a representatividade da base como um todo. Cada um desses subconjuntos, após serem submetidos à fase 3, foram chamados de *datasets* (DS). A partir dos resultados obtidos com os diferentes *datasets*, chegou-se à conclusão sobre quais atributos utilizar e qual tipo de codificação empregar. Diferentes *datasets* foram também utilizados para treinamento, teste e validação dos modelos.

Na etapa 3, foram codificados os atributos selecionados. Foram testados diversos tipos de codificação[10] para os atributos, mas optou-se pela codificação mais simples, onde os valores de cada atributo eram normalizados para o intervalo [0,1].

Nas etapas 2 e 4 após algumas análises[7], foram também selecionados os atributos (e combinações de atributos) mais significativos na descrição da base. Os atributos escolhidos para descrever a base estão na Tabela 1.

Tabela 1 – Descrição dos atributos escolhidos.

ATRIBUTO	DESCRIÇÃO
RUTIL	Receita nos dias úteis
RSABADO	Receita nos sábados
RDOMINGO	Receita nos domingos
RINTER	Receita nas ligações inter-regionais
RINTRAR	Receita nas ligações intra-regionais
RINTRAS	Receita nas ligações intra-setorial
RDIFEREN	Receita no horário diferenciado
RNORMAL	Receita no horário normal
RREDUZ	Receita no horário reduzido
RSUPERED	Receita no horário super reduzido
MINUTOS	Consumo total em minutos no mês
RECEITA	Receita total no mês

Na etapa 5 - a mais importante - foi quando tentou-se efetivamente encontrar grupos de clientes (padrões) com perfis semelhantes. Para encontrar tais grupos, foi utilizado o algoritmo de Kohonen.

Nas 3 próximas seções descreve-se detalhadamente as fases de agrupamento, caracterização e classificação.

Na etapa 6 foram interpretados os resultados da mineração através da análise de gráficos e regras.

3. Análise de agrupamentos

Foram feitas mais de 20 tentativas de agrupamento com o algoritmo de Kohonen utilizando-se, diferentes *datasets* e/ou diferentes parâmetros em cada uma. Cada uma dessas tentativas foi chamada de *estudo* (S). Foram utilizadas as sugestões encontradas em [6] e [8] para a escolha de alguns dos parâmetros do algoritmo. Nesta seção, são apresentados os estudos que apresentaram os melhores resultados e que levaram ao resultado final da fase de agrupamento.

O Processo de agrupamento foi realizado no Matlab 5.3. Após cada execução do algoritmo, o resultado era analisado através de dois mapas de densidade bidimensional que continham a quantidade de padrões ativados por neurônio, isto é, a distribuição dos padrões pela matriz de neurônios, a partir do qual buscou-se identificar os grupos. O primeiro mapa continha a quantidade expressa por um valor numérico, enquanto que o segundo era composto por círculos de diferentes tamanhos que representavam diferentes faixas de valores.

Antes de se iniciar a identificação dos *clusters*, procurou-se analisar de forma geral como os padrões estavam distribuídos ao longo do mapa de acordo com as suas características principais. Por exemplo, levou-se em consideração que os padrões correspondentes a clientes que geravam receitas baixas concentravam-se no canto superior esquerdo do mapa, enquanto que os padrões correspondentes a clientes que geravam receitas altas concentravam-se no canto diametralmente oposto.

Na identificação dos *clusters* tomou-se também o cuidado de não gerar *clusters* que viessem a corresponder a uma porcentagem muito grande ou muito pequena da base de dados. Finalmente, conforme será verificado na seção 3.1, é importante ressaltar o caráter subjetivo do processo de identificação dos *clusters*, já que a fronteira entre um possível grupo (*cluster*) e outro nem sempre fica muito clara.

3.1. Estudos

No primeiro estudo (S1) foi utilizado um mapa 20X20, 2000 e 200.000 iterações respectivamente nas fases I e II do algoritmo, um decaimento de 0.1 até 0.02 da taxa de aprendizado na fase I e vizinhança nula na fase II.

A Fig. 1, a seguir, contém o mapa encontrado em S1. Nele, parece haver 4 agrupamentos principais de padrões, sendo que apenas o agrupamento do canto inferior direito apresenta uma fronteira razoavelmente nítida ao redor.

Assumiu-se então o agrupamento do canto inferior direito (Fig. 2) como um *cluster*. Em seguida, a fim de tentar-se encontrar uma fronteira mais nítida para os possíveis *clusters* restantes e diminuir a complexidade da tarefa de agrupamento, retirou-se do *dataset* corrente os padrões pertencentes a este *cluster* e submeteu-se o novo *dataset* ao algoritmo de *Kohonen*. Após algumas

tentativas, obteve-se o resultado apresentado no estudo S2, apresentado na Fig. 3.

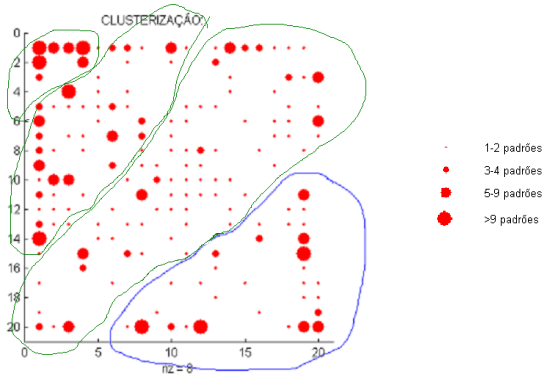

Fig. 1 - Mapa de Densidade Bidimensional para S1 com possíveis *clusters* circundados.

No segundo estudo (S2) foi utilizado um mapa 18X18, 2000 e 160.000 iterações respectivamente nas fases I e II do algoritmo, um decaimento de 0.4 até 0.02 da taxa de aprendizado na fase I e vizinhança igual a 1 na fase II.

Analisando o mapa da Fig. 3, novamente parece haver 4 agrupamentos principais de padrões, sendo que apenas os agrupamentos do canto inferior direito e do canto superior esquerdo apresentam uma fronteira razoavelmente nítida ao redor.

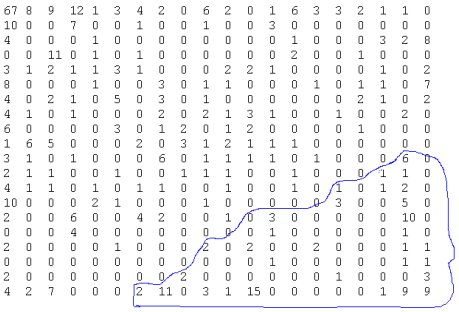

Fig. 2 - Mapa de Densidade Bidimensional para S1 com o *cluster* a ser removido circundado

Assumiu-se o agrupamento do canto superior esquerdo (destacado em azul na Fig. 3) como um *cluster*. Em seguida, retirou-se do *dataset* corrente os padrões pertencentes a este *cluster* e submeteu-se o novo *dataset* ao algoritmo de *Kohonen*. Esse processo repetiu-se nos estudos S3, S4 e finalmente no estudo S5.

No estudo S5 foi utilizado um mapa 12X12, 1000 e 70.000 iterações respectivamente nas fases I e II do algoritmo, um decaimento de 0.1 até 0.02 da taxa de aprendizado na fase I e vizinhança igual a 1 na fase II.

Neste estudo, já é possível visualizar (Fig. 4) três agrupamentos distintos de padrões com uma fronteira bem definida ao redor de cada um. Esses agrupamentos foram então assumidos como *clusters*.

Após o estudo S5, haviam sido encontrados 7 *clusters* – 1 em cada um dos estudos de S1 a S4 e 3 no estudo S5. Foi feita uma análise final sobre todo o processo de agrupamento realizado, e verificou-se que o *cluster* encontrado em S1 poderia ser redividido, já que este possuía três agrupamentos de padrões dentro dele, um mais à direita e embaixo, outro mais à esquerda e em cima e um embaixo e à direita. Após as primeiras análises da fase de caracterização, descobriu-se que os 3 agrupamentos possuíam características ligeiramente diferentes. Resolveu-se então redividir este *cluster* em outros três.

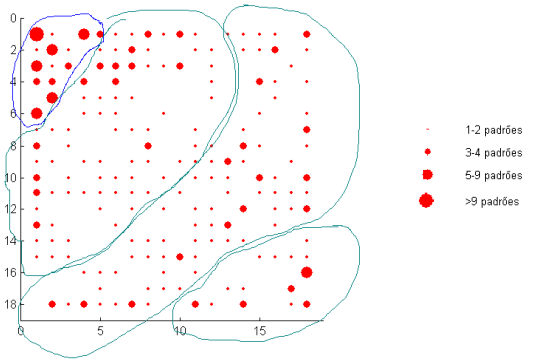

Fig. 3 - Mapa de Densidade Bidimensional para S2 com possíveis *clusters* circundados.

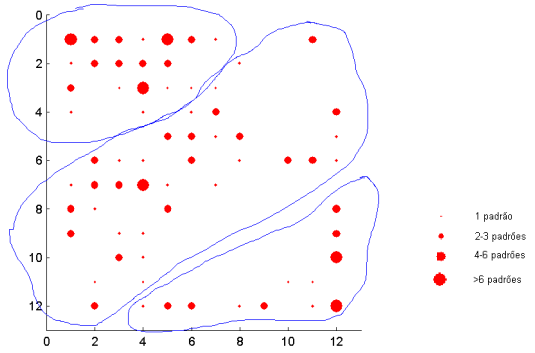

Fig. 4 - Mapa de Densidade Bidimensional para S5 com os *clusters* encontrados envolvidos em azul.

3.2. Resultados

Ao final dos estudos da fase de agrupamento, foram encontrados 9 *clusters*. As porcentagens de cada um no *dataset* utilizado encontram-se na Tabela 2. A fim de se poder visualizar o resultado final do agrupamento realizado, submeteu-se novamente o *dataset* utilizado ao algoritmo de *Kohonen* e desta vez coloriu-se os padrões de acordo com os *clusters* a que pertenciam, obtendo-se o resultado mostrado na Fig. 5. Analisando-se a figura, verifica-se que o processo de agrupamento conseguiu separar relativamente bem os *clusters*, havendo pouca superposição entre eles, mesmo utilizando uma única rede para visualizar todos os *clusters*.

Tabela 2 – Porcentagens de cada *cluster* no *dataset*.

Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**Total**
% do total	20,0	22,6	7,8	13,6	6,0	9,4	9,2	6,2	5,2	100

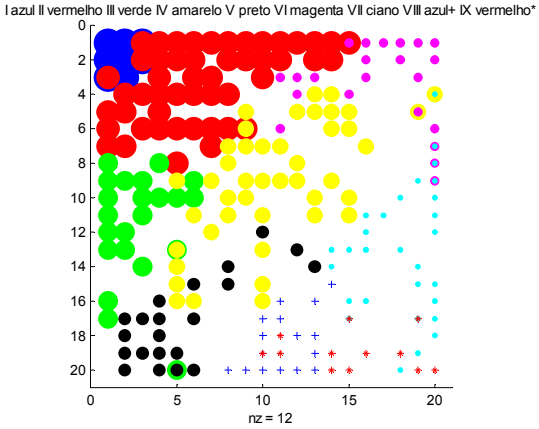

Fig. 5 – Agrupamento com o *Kohonen*.

4. Caracterização dos grupos

Nesta fase da etapa de mineração procurou-se descobrir características dos padrões de cada um dos *clusters* encontrados. Foi utilizado o software WizRule[11] para gerar regras que descrevessem as características de cada *cluster*. Foram feitas também consultas em SQL[3], gráficos de distribuição das variáveis no SPSS e uma análise da distribuição dos *clusters* no Mapa de Kohonen, resultante da fase anterior. Num primeiro momento, foram analisados os gráficos[7] e as consultas em SQL e a partir delas formuladas hipóteses sobre o perfil de cada *cluster*. Em seguida, procurou-se confirmar estas hipóteses com base nas regras descobertas pelo WizRule e chegar-se a uma conclusão.

A identificação das características de cada *cluster*, foi feita com base nas seguintes análises, sugeridas pelo Dept. de Marketing da empresa que forneceu a base de dados:

- Nível de receita gerada;
- Nível de consumo em minutos;
- Receita em dias úteis X receita em fins de semana;
- Receita em ligações de longa distâncias X receita em ligações de curta distância;
- Receita nos diferentes tipos de tarifação.

Na seção 4.1 encontram-se alguns dos histogramas, boxplots e scatterplots gerados com a ajuda do SPSS. Na seção 4.2 estão algumas das regras geradas pelo WizRule. Finalmente, na seção 4.3 encontram-se os resultados das análises feitas.

4.1. Gráficos no SPSS

O histograma da Fig.6 dá uma idéia da distribuição da receita, e de quais valores podem ser considerados como receita baixa, média ou alta. No boxplot da Fig. 7, pode-se observar, por exemplo, que os clientes pertencentes aos *clusters* III, V e VIII tem uma considerável parte do seu consumo concentrado nos fins de semana (Domingo pode ser considerado como indicador para o fim de semana), enquanto que os clientes pertencentes aos *clusters* II, IV, VI e VII tem seu consumo mais concentrado em dias de semana.

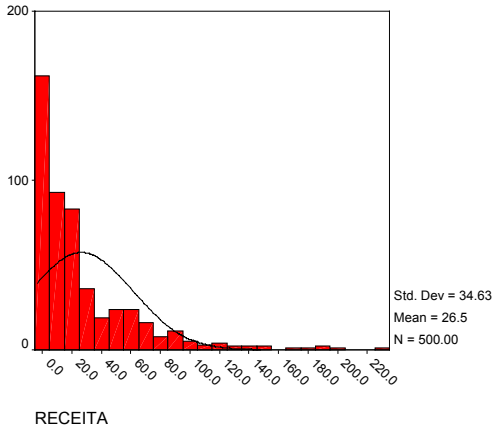

Fig. 6 – Histogramas da variável Receita.

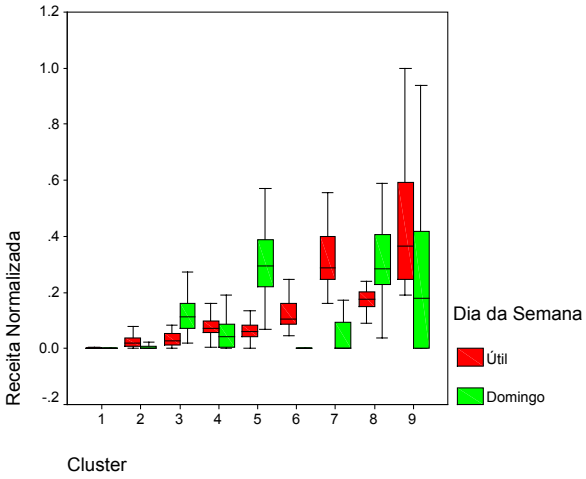

Fig. 7 – Alguns dos Histogramas e Boxplots gerados para tentar caracterizar os *clusters*.

Analisando-se o scatterplot da Fig. 8, observa-se, por exemplo, que os clientes pertencentes ao *cluster* VI, tem sua receita mais compreendida nos horários normal e diferenciado, especialmente neste último.

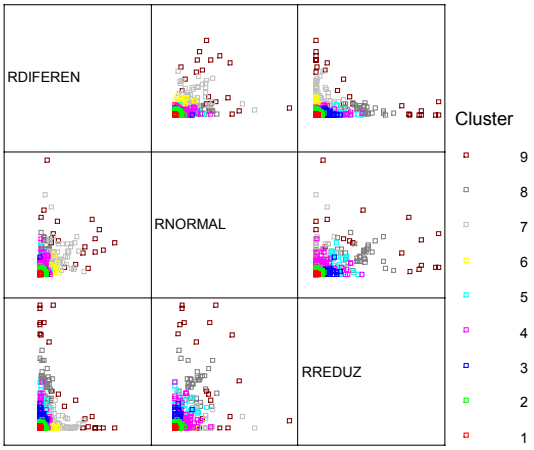

Fig. 8 – Scatterplots dos *clusters*.

4.2. Geração de regras com o WizRule

O WizRule é um software de mineração de dados que gera regras para descrever uma base de dados. Foram feitas várias tentativas (estudos) de se encontrar regras para caracterizar os *clusters*. Em cada uma, foram utilizados graus de suporte e confiança diferentes[1]. Na Fig. 9, a seguir, apresenta-se um subconjunto reduzido das regras mais significativas encontradas nos dois principais estudos. No primeiro, foi utilizado um suporte de 10% e um grau de confiança de 70%. Com esses parâmetros, foram encontradas regras para os *clusters* I, II, VIII e IX. No segundo estudo, foi utilizado um suporte de 4% e um grau de confiança de 50%. Nesse último, relaxou-se os parâmetros de suporte para tentar-se encontrar regras para os *clusters* restantes.

3) *If* **RECEITA** *is* **0.05 ... 1.52** (average = **0.60**)
Then **CLUSTER** *is* **1.00**

78) *If* **RECEITA** *is* **1.58 ... 3.60** (average = **2.72**)
Then **CLUSTER** *is* **2.00**

93) *If* **RUTIL** *is* **1.12 ... 2.99** (average = **2.00**)
and **RDOMINGO** *is* **0.00 ... 0.09** (average = **0.01**)
Then **CLUSTER** *is* **2.00**

346) *If* **RUTIL** *is* **0.00 ... 1.03**(average = **0.32**)
and **RINTER** *is* **1.94 ... 3.17**(average = **2.68**)
Then **CLUSTER** *is* **3.00**

373) *If* **RUTIL** *is* **17.06 ... 18.68**(average = **18.04**)
and **RINTRAS** *is* **0.00**
Then **CLUSTER** *is* **4.00**

391) *If* **RSABADO** *is* **0.00**
and **RDOMINGO** *is* **8.03 ... 12.94**(average = **10.73**)
and **RDIFEREN** *is* **0.00 ... 1.48**(average = **0.47**)
Then **CLUSTER** *is* **5.00**

393) *If* **RINTRAR** *is* **0.00**
and **RINTRAS** *is* **0.00**
and **MINUTOS** *is* **205.00 ... 214.00**(average = **209.17**)
Then **CLUSTER** *is* **5.00**

279) *If* **RSABADO** *is* **0.00**
and **RDOMINGO** *is* **0.00**
and **MINUTOS** *is* **78.00 ... 87.00**(average = **83.33**)
Then **CLUSTER** *is* **6.00**

433) *If* **RUTIL** *is* **18.98 ... 20.59**(average = **19.97**)
and **RREDUZ** *is* **0.00 ... 0.09**(average = **0.01**)
Then **CLUSTER** *is* **6.00**

333) *If* **RDIFEREN** *is* **61.62 ... 160.41**(average = **101.59**)
and **RNORMAL** *is* **22.00 ... 37.00**(average = **27.50**)
and **RINTRAR** *is* **0.00 ... 0.15**(average = **0.02**)
Then **CLUSTER** *is* **7.00**

77) *If* **RREDUZ** *is* **19.81 ... 66.80** (average = **34.09**)
Then **CLUSTER** *is* **8.00 ... 9.00** (average = **8.32**)

85) *If* **RSABADO** *is* **7.00 ... 33.00** (average = **11.35**)
Then **CLUSTER** *is* **8.00 ... 9.00** (average = **8.29**)

Fig. 9 – Algumas das regras geradas pelo WizRule

As regras geradas pelo WizRule levaram à conclusões semelhantes às obtidas com os gráficos gerados pelo SPSS. Por exemplo, as regras de número 279 e 433 estão em conformidade com as suposições feitas sobre o perfil da receita gerada pelos clientes do *cluster* VI apresentadas na seção anterior. O mesmo pode ser dito para a regra de número 93, relativa ao *cluster* II.

4.3. Resultados

Com base nas regras geradas pelo WizRule e nos gráficos gerados pelo SPSS conseguiu-se traçar um perfil para os clientes de cada *cluster*. Os perfis encontrados, estão listados a seguir:

Cluster I: receita muito baixa, sem padrão definido para as chamadas.
Cluster II: receita baixa, mais caracterizado por ligações em dia úteis.
Cluster III: receita intermediária, mais caracterizado por ligações inter-regionais nos fins de semana e no horário reduzido.
Cluster IV: receita intermediária, mais caracterizado por ligações inter-regionais em dias úteis e no horário normal.
Cluster V: receita intermediária, ligações longas, inter-regionais nos fins de semana e no horário reduzido.
Cluster VI: receita intermediária, ligações curtas, intra-setoriais, em dias úteis e no horário diferenciado.
Cluster VII: receita alta, ligações inter-regionais, em dias úteis nos horários diferenciado e normal.
Cluster VIII: receita alta, ligações longas, inter-regionais, nos fins de semana nos horários normal e reduzido.
Cluster IX: receita muito alta, mais caracterizado por ligações inter-regionais em todos os dias e horários.

5. Classificação

Na última fase da etapa de mineração, tentou-se construir um modelo de classificação para os clientes que ficaram de fora do *dataset* utilizado na fase de agrupamento e para futuros clientes que a empresa viesse a adquirir. Ao invés de utilizar como modelo de classificação a própria rede treinada com o algoritmo de *Kohonen*, optou-se por criar e verificar o desempenho de outros dois modelos, um baseado em regras e o outro baseado em redes neurais com o algoritmo Back Propagation. Os softwares utilizados para a construção desses dois modelos foram, respectivamente, o WizWhy[11] e o NeuralSIM.

5.1 Classificação com o NeuralSIM

O NeuralSIM é um software de redes neurais que antes de treinar e executar uma rede, também faz um pré processamento sobre os dados e tenta encontrar a melhor arquitetura para a rede. O NeuralSIM implementa o BackPropagation com a possibilidade de ajuste de diversos de seus parâmetros. Por implementar uma rede neural, o software tem a desvantagem de não gerar regras para justificar as respostas dadas.

Foram realizados três estudos para a construção do modelo de classificação com o NeuralSIM. Em cada estudo, foram testadas várias redes. No primeiro (S0), foram utilizados os parâmetros *default* do software e, nos outros dois, tentou-se melhorar o desempenho ajustando alguns parâmetros (Tabela 3). Todas as redes

criadas foram treinadas utilizando a regra de aprendizado baseada no gradiente adaptativo.

Tabela 3 – Principais parâmetros utilizados nos estudos com o NeuralSIM.

Parâmetro	S0	S1	S2
Hidden Layer Function	Tanh	Tanh	Sigmoid
Output Layer Function	Softmax	Softmax	Softmax
Network Evaluation Function	Average Classification Rate	Average Classification Rate	Accuracy
Number of Nets to Train	1	3	4
	S0	**S1**	**S2**
Arquitetura Encontrada	16-4-9	16-8-9	14-15-9

Dentre os modelos gerados com o NeuralSIM, escolheu-se o S1 por apresentar um desempenho mais homogêneo para todos os *clusters* (Tabela 4).

Tabela 4 – Perfomance quanto a falsos positivos (A) e falsos negativos (B).

	S0		S1		S2	
Cluster	A	B	A	B	A	B
I	100%	96,8%	100%	96,8%	96,7%	96,7%
II	91,2%	100%	94,1%	100%	97,0%	97,1%
III	91,7%	64,7%	83,3%	71,4%	91,7%	84,6%
IV	66,7%	100%	76,2%	100%	90,5%	100%
V	100%	75%	100%	69,2%	100%	81,8%
VI	85,7%	80%	85,7%	92,3%	92,9%	100%
VII	85,7%	85,7%	100%	87,5%	100%	77,8%
VIII	100%	100%	100%	100%	100%	90%
IX	87,5%	87,5%	87,5%	100%	37,5%	100%
Total	89,4%	89,4%	92,1%	92,1%	92,7%	92,7%

5.2. Classificação com o WizWhy

O WizWhy é um software que possui um funcionamento bastante semelhante ao WizRule e que utiliza otimização combinatória, regressão e teste de hipótese para gerar regras que descrevem uma base de dados e fazer previsões com base nessas regras. Tem a vantagem de apresentar uma justificativa (regras) para as respostas dadas, algo que não acontece com o modelo de redes neurais. O algoritmo implementado, no entanto, esta apto apenas a fazer previsões booleanas. Para fazer a previsão, ele calcula a probabilidade dela ser verdadeira e caso esta probabilidade esteja acima de um limiar, dá verdadeiro como resposta. Para contornar esse problema, foram geradas regras específicas para cada *cluster*, utilizando graus de confiança e suporte apropriados a cada um. Em seguida, para cada cliente do conjunto de teste, foram feitas 9 previsões diferentes, uma com cada conjunto de regras gerado. Cada uma das nove previsões forneceu a probabilidade do cliente pertencer a um dado *cluster*, semelhante a um classificador *bayesiano*[4]. Classificou-se cada cliente como pertencendo ao *cluster* que apresentou a maior probabilidade associada.

5.3. Avaliação dos Resultados

Para avaliação dos modelos, foi verificado o desempenho quanto a falsos negativos e falsos positivo. Embora a classificação feita com NeuralSIM tenha a desvantagem de não gerar regras, dificultando uma justificativa para as classificações obtidas, ela obteve um desempenho consideravelmente melhor (melhor percentual médio de acerto de 92,1% contra 72,3% do WizWhy). O desempenho inferior apresentado pelo WizWhy pode em parte ser causado pela dificuldade em se escolher os parâmetros de suporte para os diversos conjuntos de regras de maneira homogênea, isto é, tal que a previsão de um cliente pertencer ou não a um *cluster* feita por um conjunto de regras não fosse mais conservadora do que a feita por um outro, acabando por influenciar no resultado.

6. Conclusões

Apesar da complexidade da base, os resultados obtidos foram satisfatórios e úteis para a empresa de telefonia. Como foram utilizadas diferentes técnicas para realizar uma mesma tarefa, foi possível ainda, obter uma idéia geral das vantagens e desvantagens das técnicas utilizadas.

Pretende-se, em trabalhos futuros, realizar outros estudos de caso, procurando integrar mais o algoritmo de *Kohonen* com técnicas de visualização durante o processo de agrupamento, a fim de se fazer cortes mais "conscientes" no mapa. Pretende-se também dar pesos diferentes a cada atributo na regra de aprendizado, conforme a importância de cada um. Além disso, pretende-se ainda utilizar sistemas neuro-fuzzy na fase de classificação com o objetivo de conjugar as vantagens das redes neurais com a capacidade de extração de regras dos sistemas fuzzy.

Referências

[1] Agrawal, R., Imielinski, T., Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings, ACM SIGMOD Conference on Management of Data, Washington, D.C.

[2] Bigus, J. P., Data Mining with Neural Networks, McGraw Hill, 1996

[3] Date, C. J., A Guide to the SQL Standard, Addison Wesley, 2nd edition, 1989

[4] Duda, R. Hart, P., Pattern Classification and Scene Analysis, John Wiley & Sons, 1973

[5] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, MIT Press, 1996

[6] Haykin, S., Neural Networks – A Comprehensive Foundation, Prentice Hall, 2nd edition, 1999

[7] Johnson, R. Wichern, D., Applied Multivariate Statistical Analysis, Prentice Hall, 4th edition, 1999

[8] Kohonen, T., Self-Organizing Maps, Springer, 2nd edition, 1997

[9] Mckenna, R., Marketing de Relacionamento, Ed. Campus, 8ª edição, 1992

[10] Weiss, S. M., Indurkhya, N., Predictive Data Mining. Morgan Kaufmann Publishers, Inc, 1998

[11] WizRule e WizWhy, www.wizsoft.com