J. Health Inform. 2012 Abril-Junho; 4(2):43-9

Artigo Original

JHI JOURNAL OF HEALTH INFORMATICS

Aplicação de algoritmos de aprendizagem de máquina para mineração de dados sobre beneficiários de planos de saúde suplementar

Application of machine learning algorithms to data mining about beneficiaries of health insurance

Aplicación de algoritmos de aprendizaje automático para la minería de datos sobre los beneficiarios de los planes de seguro de salud

Oudival Luiz Fraccaro de Marins[1], Everton Fernando Barros[2], Wesley Romão[3], Ademir Aparecido Constantino[4], Celso Lara de Souza[5]

Descritores: Base de Dados; Mineração de Dados; Planos de Pré-Pagamento em Saúde

RESUMO

Grandes bases de dados podem conter conhecimento oculto que poderia auxiliar na tomada de decisões, porém a extração de tal conhecimento não é tarefa trivial sendo necessária a utilização de técnicas de mineração de dados. Operadoras de planos de saúde suplementar normalmente possuem grande quantidade de informações armazenadas a respeito dos procedimentos realizados por seus beneficiários, o que possibilita a existência de conhecimento oculto em suas bases de dados. A literatura apresenta um algoritmo, denominado C5.0, reconhecido como eficiente para resolver a tarefa de classificação em mineração de dados. Neste artigo foram aplicados os algoritmos, de aprendizagem de máquina, C5.0 e PGD (Programação Genética Difusa) em uma base de dados sobre beneficiários de planos de saúde suplementar a fim de validar o algoritmo baseado em programação genética comparando com os resultados da aplicação do algoritmo C5.0.

Keywords: Database; Data Mining; Prepaid Health Plans

ABSTRACT

Large databases may contain hidden knowledge that could assist in making decisions, but the extraction of such knowledge is not trivial task requiring the use of data mining techniques. Operators of health insurance plans typically have large amounts of stored information about the procedures performed by its beneficiaries, which allows the existence of knowledge hidden in their databases. The literature presents an algorithm called C5.0, recognized as effective in solving the classification task in data mining. In this paper we applied the algorithms, machine learning, C5.0 and PGD (Fuzzy Genetic Programming) in a database on beneficiaries of health insurance plans in order to validate the algorithm based on genetic programming compared to the results of applying C5.0 algorithm.

Descriptores: Base de Datos; Minería de Datos; Planes de Salud de Prepago

RESUMEN

Bases de datos grandes pueden contener conocimiento oculto que podría ayudar en la toma de decisiones, pero la extracción de conocimiento no es tarea trivial y requiere el uso de técnicas de minería de datos. Los operadores de los planes de seguro de salud tienen típicamente una gran cantidad de datos almacenados acerca de los procedimientos realizados por los beneficiarios, lo que permite la existencia del conocimiento oculto en sus bases de datos. La literatura presenta un algoritmo llamado C5.0, reconocido como eficaz en la solución de la tarea de clasificación en la minería de datos. En este trabajo se aplicaron los algoritmos de aprendizaje de máquina, C5.0 y PGD (Programación Genética Fuzzy) en una base de datos sobre los beneficiarios de los planes de seguro de salud con el fin de validar el algoritmo basado en programación genética en comparación con los resultados de la aplicación C5.0 algoritmo.

[1] Bacharel em Ciência da Computação pela Universidade Estadual de Maringá – UEM – Maringá (PR), Brasil.
[2] Mestrado em Ciência da Computação pela Universidade Estadual de Maringá – UEM - Maringá (PR), Brasil.
[3] Professor Adjunto da Universidade Estadual de Maringá – UEM – Maringá (PR), Brasil.
[4] Professor Titular da Universidade Estadual de Maringá – UEM – Maringá (PR), Brasil.
[5] Especialista em Gestão de Projetos e Gestão Empresarial pela Fundação Getúlio Vargas - FGV, São paulo (SP), Brasil. Diretor da Benner Tecnologia e Sistemas de Saúde Ltda, Maringá (PR), Brasil.

Autor Correspondente: **Oudival Luiz Fraccaro de Marins**
e-mail: **luiz.f.marins@gmail.com**

Artigo recebido: **30/03/2012**
Aprovado: **30/05/2012**

INTRODUÇÃO

O crescente avanço tecnológico da informática permite a geração, coleta e o armazenamento de uma grande quantidade de dados. Tal fato é tão evidente que no inicio da década de noventa estimava-se que a quantidade de informações armazenada se duplicaria a cada vinte meses[1].

Diante deste crescimento no volume de dados tornou-se inviável para o ser humano analisar essas grandes quantidades de dados sem a utilização de uma ferramenta computacional. Além disso, a natureza não trivial dos dados impossibilita a utilização de técnicas e ferramentas[2] tradicionais de análise de dados, mesmo em casos onde a quantidade de dados é pequena. Portanto, a extração de conhecimento de uma grande base de dados é uma tarefa complexa.

As operadoras de plano de saúde suplementar armazenam grande quantidade de dados referentes aos procedimentos realizados pelos beneficiários. O possível conhecimento oculto existente em tal base de dados pode auxiliar na tomada de decisões em programas de prevenção de doenças. Uma abordagem para extração de conhecimentos ocultos em grandes bases de dados é a Mineração de Dados (MD).

O processo de descoberta de conhecimento em bases de dados (DCBD) é composto por várias etapas: seleção, pré-processamento, transformação, MD e interpretação/avaliação dos resultados. Este trabalho é fundamentado na etapa de MD, que exige a adaptação de diversas técnicas para que o conhecimento descoberto seja correto, compreensível e relevante. A etapa de MD possui diversas tarefas: classificação, regressão, associação, agrupamento (*clustering*), detecção de anomalias (outlier). Neste trabalho focou-se na tarefa de classificação, que consiste no aprendizado de uma função (regra de previsão) que mapeie um conjunto de atributos previsores para uma classe pré-determinada[3].

A tarefa de classificação pode ser resolvida por meio de vários métodos tais como: rede neural artificial, indução em árvores de decisão, algoritmos genéticos, redes bayesianas, entre outros[4].

Este artigo propõe a aplicação de dois algoritmos de MD em uma base de dados de uma operadora de plano de saúde suplementar[5], com a finalidade de validar um dos algoritmos utilizados para extrair padrões que possam auxiliar a tomada de decisões.

MÉTODOS

Neste trabalho foram utilizados um algoritmo de programação genética difuso e um algoritmo de indução em árvore de decisão, denominados PGD e C5.0[6-7], respectivamente, com a finalidade de extrair padrões que possam auxiliar no processo de tomada de decisão.

Os dois algoritmos foram aplicados sobre dados de beneficiários de plano de saúde suplementar do ano de 2010 que foram preparados por Barros[5] utilizando os atributos meta descritos no Quadro 1 abaixo. Foi escolhida a regra com maior taxa de acerto no conjunto de treinamento, e em seguida foi avaliada no conjunto de testes, para cada atributo meta. Portanto, a avaliação da taxa de acerto das regras foi realizada, primeiramente, utilizando o método *holdout*[2] onde os dados são divididos em dois conjuntos disjuntos: treinamento e teste. Em seguida o modelo de classificação é criado a partir do conjunto de dados de treinamento e em seguida avaliado no conjunto de dados de teste[2]. Foram destinados 66% dos dados ao conjunto de treinamento e 33% ao conjunto de testes. Os resultados foram comparados apenas em relação à taxa de acerto das melhores regras.

A comparação dos resultados foi realizada em um segundo momento utilizando validação cruzada[8] em seis partes. Além da taxa de acerto, foram consideradas, na comparação, a cobertura e a média do número de antecedentes das regras. Foram selecionados três atributos para serem utilizados como atributos meta: GRUPO_CID_IIIA, GRUPO_CID_IIIB e GRUPO_EVENTO_09 de acordo com o interesse e conhecimento dos especialistas de domínio.

As seguintes subseções mostram as características da base de dados utilizada e detalham a metodologia utilizada com cada um dos algoritmos para obtenção dos resultados utilizados na comparação.

Análise da base de dados de beneficiários de plano se saúde suplementar

Com auxilio da ferramenta WEKA[9], foi analisada a distribuição de classes da base de dados. Esta distribuição de classes está ilustrada no Quadro 1.

Nota-se que, para a maioria dos atributos, tem-se maior concentração de registros para uma classe e um número muito pequeno pertencente à outra classe. Esta má distribuição tem influência no resultado obtido pelos algoritmos de MD, dificultando a indução de regras.

Algoritmo C5.0

Os resultados obtidos pelo C5.0 diferem dos resultados do PGD, pois o C5.0 cria um modelo de classificação no formato de árvore de decisão[2], enquanto o PGD gera um conjunto de regras de produção. Além disso, a árvore de decisão muitas vezes assume grandes tamanhos, o que dificulta o entendimento do modelo de classificação.

Porém, uma árvore de decisão pode ser convertida facilmente para um conjunto de regras de produção equivalente[7]. Portanto, o C5.0 foi aplicado a base de dados utilizando sua funcionalidade de converter a árvore de decisão gerada em um conjunto de regras de produção equivalentes, obtendo assim resultados de fácil entendimento e no formato equivalente ao outro algoritmo (PGD).

Além disso, o C5.0 foi aplicado na base de dados utilizando a técnica de custo para erro de classificação[10-11]. Com a utilização desta técnica, o algoritmo foi configurado de maneira que o custo de classificar um registro como pertencente à classe SIM, quando na verdade ele pertença à classe NÃO, seja maior do que classificar um registro como sendo da classe NÃO, quando na verdade ele pertença à classe SIM.

Com o objetivo de descobrir a partir de qual custo o algoritmo é capaz de induzir árvores de decisão com mais de uma folha, o C5.0 foi aplicado variando o custo entre os valores 2 (dois) e 50 (cinquenta). Foram considerados os resultados utilizando custo maior ou igual a 10 (dez), pois em média este custo obteve melhores resultados.

Quadro 1 - Distribuição de classes dos dados dos beneficiários do plano de saúde suplementar, 2010

Atributo meta	Descrição	Sim (%)	Não (%)
GRUPO_CID_I	Doenças Infecciosas e parasitárias	2,857	97,143
GRUPO_CID_II	Neoplasias	4,072	95,928
GRUPO_CID_IIIA	Doenças Endócrinas	2,257	97,743
GRUPO_CID_IIIB	Doenças Nutricionais e metabólicas	1,845	98,155
GRUPO_CID_V	Doenças Mentais	2,422	97,578
GRUPO_CID_VIA	Sistema Nervoso	1,869	98,131
GRUPO_CID_VIB	Olhos e anexos, ouvido e apófise Mastóide	4,696	95,304
GRUPO_CID_VIIA	Circulatório, menos Veias e Linfáticos	10,12	89,880
GRUPO_CID_VIIIA	Infecções Respiratórias Agudas	4,666	95,334
GRUPO_CID_VIIIB	Outras Doenças Respiratórias	1,725	98,275
GRUPO_CID_IXA	Outras Doenças Respiratórias	1,808	98,192
GRUPO_CID_XII	Pele e tecido subcutâneo	2,712	97,288
GRUPO_CID_XIII	Osteomuscular e tecido conjuntivo	5,466	94,534
GRUPO_CID_XA	Doenças Urinárias	3,003	96,997
GRUPO_CID_XB	Doenças genitais masculinas	1,729	98,271
GRUPO_CID_XC	Doenças genitais femininas	5,893	94,107
GRUPO_CID_XVI	Sintomas, sinais e afecções mal Definidas	10,554	89,446
GRUPO_CID_XXI	Lesões, envenenamentos e causas externas	17,754	82,246
ESPECIALIDADE_04	Análises Clínicas	3,512	96,488
ESPECIALIDADE_05	Anatomia Patológica	0,515	99,485
ESPECIALIDADE_06	Anestesioliga	1,167	98,833
ESPECIALIDADE_07	Angiologia	0,641	99,359
ESPECIALIDADE_09	Cardiologia	6,022	93,978
ESPECIALIDADE_13	Cirurgia Geral	2,730	97,270
ESPECIALIDADE_18	Clínica Médica	37,127	62,873
ESPECIALIDADE_24	Dermatologia	2,857	97,143
ESPECIALIDADE_27	Endocrinologia	1,867	98,133
ESPECIALIDADE_31	Gastroenterologia	1,519	98,481
ESPECIALIDADE_33	Ginecologia	6,459	93,541
ESPECIALIDADE_39	Internações	2,239	97,761
ESPECIALIDADE_47	Medicina Preventiva	2,766	97,234
ESPECIALIDADE_50	Neurologia	1,382	98,618
ESPECIALIDADE_53	Oftalmologia	6,821	93,179
ESPECIALIDADE_54	Ortopedia	5,783	94,217
ESPECIALIDADE_55	Otorrinolaringologia	3,537	96,463
ESPECIALIDADE_56	Patologia	3,137	96,863
ESPECIALIDADE_57	Patologia Clínica	9,136	90,864
ESPECIALIDADE_58	Pediatria	6,687	93,313
ESPECIALIDADE_59	Pneumologia	1,281	98,719
ESPECIALIDADE_62	Psiquiatria	2,215	97,785
ESPECIALIDADE_63	Radiodiagnostico	0,968	99,032
ESPECIALIDADE_64	Radiologia	4,849	95,151
ESPECIALIDADE_66	Reumatologia	0,713	99,287
ESPECIALIDADE_69	Urologia	2,072	97,928
GRUPO_EVENTO_01	Procedimentos – Diabetes Mellitus	0,668	99,332
GRUPO_EVENTO_04	Procedimentos Sentinelas – Diabetes Mellitus	0,443	99,557
GRUPO_EVENTO_09	Procedimentos – Câncer de Mama	3,615	96,385

Muitas das regras obtidas para a classe SIM possuem baixa cobertura no conjunto de treinamento, muitas vezes menores do que 10 (dez) registros. Com o intuito de obter resultados melhores, foi definido empiricamente que as regras selecionadas devem ter cobertura mínima de 25 (vinte e cinco) registros.

Além disso, algumas regras possuem apenas um antecedente, o que não expressa um conhecimento interessante. Sendo assim, também foi definido que as regras devem ter no mínimo dois antecedentes.

Resumindo, o C5.0 foi configurado para converter a árvore de decisão gerada para um conjunto de regras de produção equivalente. Para cada possível valor de cada atributo meta (SIM e NÃO), foi escolhida a regra com maior taxa de acerto no conjunto de treinamento, que possui cobertura maior que 25 (vinte e cinco) e com pelo menos dois atributos no antecedente. Para 16 (dezesseis) dos atributos citados acima, o C5.0 foi aplicado em sua configuração padrão. Para o restante dos atributos, o C5.0 foi aplicado utilizando a técnica de custo para erro de classificação.

Programação Genética

A Programação Genética (PG) utilizada é baseada em uma população de "programas" que são representados em forma de árvores (indivíduos), no qual os nós internos são funções e os nós folhas são constantes. Essas árvores utilizam como funções (ou nós internos) os operadores lógicos E e OU, e como folhas (ou terminais) o valor booleano de termos como "Sexo = Masculino", "Idade = 10_a_19" etc. Assim, com a combinação de terminais e funções, podem ser criadas árvores do tipo:

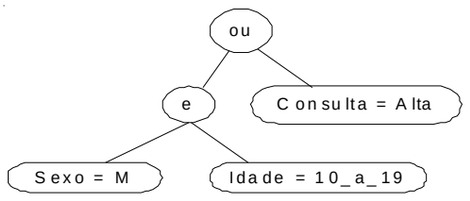

Figura1 - Representação de um indivíduo

Essas árvores representam a parte do antecedente da regra (parte SE):

SE (Sexo = M) E (Idade=10_A_19)) OU (Consulta = Alta) (1)

O consequente da regra (parte do ENTÃO) é determinada previamente para todos os indivíduos da população. Após o processo de evolução (aplicação dos operadores de mutação, cruzamento e reprodução) é aplicada a Forma Normal Disjuntiva (FND) para se obter regras como:

SE (Sexo=M) E (Idade = 10_A_19) ENTÃO (Internação = mais_que_uma) (2)

SE (Consulta = Alta) ENTÃO (Internação = mais_que_uma)(3)

Para os valores contínuos como número de consultas a PG utiliza lógica difusa (portanto PGD) para facilitar a compreensão do conhecimento descoberto e simplificar o processo de preparação de dados, pois dessa maneira esses valores não precisam ser categorizados ou discretizados. Os valores linguísticos utilizados foram "baixo", "médio" e "alto", no qual os parâmetros difusos são determinados pelo especialista de domínio, incorporando assim conhecimento do especialista na descoberta das regras.

A avaliação dos indivíduos para o processo evolutivo é realizado com base na média da precisão de todas as regras pertencentes a árvore[12].

Na execução dos experimentos do PGD utilizou-se o método *holdout*[2] na proporção 66% para treinamento e 33% para teste como no C5.0. No processo de treinamento o PGD considerou regras com mais de 25 casos de verdadeiro positivo. Após o treinamento foram escolhidas as regras com a maior taxa de acerto da população e que possuíssem cobertura mínima de 25 casos sobre o conjunto de teste.

As regras obtidas possuem no mínimo dois e no máximo cinco antecedentes, pois acredita-se que regras com apenas um antecedente não expressa conhecimento interessante e com mais de cinco antecedentes as torna de difícil compreensão. Portanto, o PGD possui restrições de tamanho mínimo de dois e no máximo cinco antecedentes.

RESULTADOS E DISCUSSÃO

Foram realizados dois experimentos, um utilizando o método *holdout* para validação e outro utilizando validação cruzada. Utilizando-se os algoritmos propostos, no primeiro experimento considerou-se como atributo meta cada atributo do Quadro 1, enquanto que no segundo experimento foi considerado apenas três atributos de maior interesse do usuário.

Holdout

No Quadro 2 são exibidos os resultados dos dois algoritmos. Os valores em negrito representam a maior taxa de acerto entre os algoritmos para cada meta, apenas para a classe SIM. Destaca-se, também, a frequência da distribuição de classes para o atributo meta SIM.

Quando aplicado à base de dados em sua configuração padrão (sem a utilização de *misclassification cost*), o C5.0 foi capaz de induzir uma árvore de decisão com mais de uma folha apenas para os atributos: GRUPO_CID_IIA, GRUPO_CID_XVI, GRUPO_CID_XXI, ESPECIALIDADE_04, ESPECIALIDADE_06, ESPECIALIDADE_09, ESPECIALIDADE_18, ESPECIALIDADE_33, ESPECIALIDADE_39, ESPECIALIDADE_47, ESPECIALIDADE_56, ESPECIALIDADE_57, ESPECIALIDADE_58, ESPECIALIDADE_63, ESPECIALIDADE_64, GRUPO_EVENTO_09.

Para o restante dos atributos meta, o resultado obtido foi uma árvore de decisão com apenas uma folha, rotulada com a classe NÃO. Isto significa que todos os registros são classificados como pertencentes à classe NÃO, independente dos valores de seus atributos previsores. Esta situação ocorre porque, para estes atributos, a maior parte dos registros pertence à classe NÃO, em média 97,3%, enquanto uma parcela muito pequena dos registros pertence à classe SIM.

Analisando e comparando os resultados obtidos pelos algoritmos, observa-se que ambos alcançaram altas taxas de acerto para todos os atributos meta, considerando a classe NÃO. No entanto esta alta taxa de acerto deve-se ao fato da maioria dos registros pretencerem a classe NÃO. Sendo assim, qualquer regra gerada classificando um registro como pertencente à classe NÃO, tem em média 97,3% de chance de estar correta. O caso extremo pode ser observado no algoritmo PGD, onde praticamente todas as regras obtêm 100% de taxa de acerto.

Um resultado mais interessante é observado em relação à classe SIM. Pode-se observar por meio do Quadro 2 que o C5.0 obteve as melhores taxas de acerto para sete atributos, e o PGD obteve os melhores resultados para trinta e nove atributos. Constata-se que, as taxas de acerto obtidas pelo C5.0 diferem do PGD em média 28,8%.

A partir deste experimento, observou-se que as regras induzidas pelo C5.0 favorecem (estão focadas) na classe que possui maior quantidade de registros, fato já observado por Carvalho[13].

Nota-se ainda que para os dados utilizados o PGD obteve melhores resultados que o C5.0 para a maioria dos atributos meta para classe SIM e muito similares para a classe NÃO, com isso conclui-se que o PGD tende a obter resultados melhores para bases de dados com distribuição de classes ruim, do que algoritmos baseados em árvore de decisão, principalmente para a classe com menos registros.

Quadro 2 - Taxas de acerto obtidos por meio do algoritmo C5.0 e PGD utilizando *houldout*, sobre os dados dos beneficiários do plano de saúde suplementar, 2010

Atributo meta / Valor	Taxa de Acerto C5.0		Frequência	Taxa de Acerto PGD	
	SIM (%)	NÃO (%)	SIM (%)	SIM (%)	NÃO (%)
GRUPO_CID_I	-	99,2	2,857	20,8	100
GRUPO_CID_II	11,2	100,0	4,072	40,9	100
GRUPO_CID_IIIA	14,2	98,5	2,257	46,8	100
GRUPO_CID_IIIB	5,0	99,2	1,845	11,9	100
GRUPO_CID_V	72,7	99,7	2,422	60,7	100
GRUPO_CID_VIA	18,5	99,2	1,869	26,8	100
GRUPO_CID_VIB	16,6	100,0	4,696	30,3	100
GRUPO_CID_VIIA	30,7	100,0	10,120	67,6	100
GRUPO_CID_VIIIA	14,2	98,5	4,666	35,2	100
GRUPO_CID_VIIIB	5,0	99,2	1,725	10,3	100
GRUPO_CID_IXA	13,8	98,6	1,808	91,9	100
GRUPO_CID_XII	40,0	95,8	2,712	26,2	100
GRUPO_CID_XIII	40,0	100,0	5,466	53,1	100
GRUPO_CID_XA	11,5	100,0	3,003	39,7	100
GRUPO_CID_XB	13,0	98,8	1,729	26,6	100
GRUPO_CID_XC	16,6	99,3	5,893	27,7	100
GRUPO_CID_XVI	27,8	98,0	10,554	66,0	100
GRUPO_CID_XXI	88,1	100,0	17,754	100,0	100
ESPECIALIDADE_04	64,6	99,8	3,512	93,5	100
ESPECIALIDADE_05	50,0	99,7	0,515	19,6	100
ESPECIALIDADE_06	25,0	99,6	1,167	49,2	100
ESPECIALIDADE_07	22,2	99,6	0,641	18,1	100
ESPECIALIDADE_09	58,3	99,8	6,022	81,3	100
ESPECIALIDADE_13	30,7	99,7	2,730	45,3	100
ESPECIALIDADE_18	57,1	98,8	37,127	100,0	100
ESPECIALIDADE_24	12,5	98,5	2,857	21,6	100
ESPECIALIDADE_27	25,0	99,3	1,867	37,9	100
ESPECIALIDADE_31	11,5	99,2	1,519	45,3	100
ESPECIALIDADE_33	21,4	99,9	6,459	100,0	100
ESPECIALIDADE_39	90,9	99,9	2,239	86,7	100
ESPECIALIDADE_47	75,0	99,9	2,766	54,3	100
ESPECIALIDADE_50	13,0	98,8	1,382	25,5	100
ESPECIALIDADE_53	-	100,0	6,821	43,8	100
ESPECIALIDADE_54	-	99,0	5,783	54,3	100
ESPECIALIDADE_55	-	97,4	3,537	18,8	100
ESPECIALIDADE_56	44,0	100,0	3,137	79,5	100
ESPECIALIDADE_57	31,3	99,3	9,136	95,2	100
ESPECIALIDADE_58	80,0	100,0	6,687	94,9	100
ESPECIALIDADE_59	-	99,2	1,281	15,4	100
ESPECIALIDADE_62	25,6	100,0	2,215	36,8	100
ESPECIALIDADE_63	84,6	99,5	0,968	70,9	100
ESPECIALIDADE_64	-	99,9	4,849	85,3	100
ESPECIALIDADE_66	5,7	99,6	0,713	50,3	100
ESPECIALIDADE_69	10,0	99,7	2,072	38,0	100
GRUPO_EVENTO_01	4,2	100,0	0,668	100,0	100
GRUPO_EVENTO_04	2,9	100,0	0,443	100,0	100
GRUPO_EVENTO_09	5,0	99,9	3,615	93,4	100

Quadro 3 - Taxa de acerto e desvio padrão obtidos por meio do algoritmo C5.0 e PGD utilizando validação cruzada, sobre os dados dos beneficiários do plano de saúde suplementar, 2010

Atributo meta / Valor	C5.0 (x% ± DP%)	PGD (x% ± DP%)
GRUPO_CID_IIIA: SIM	59,7% ± 8,2%	34,5% ± 2,8%
GRUPO_CID_IIIB: SIM	17,1% ± 13,3%	5,9% ± 1,2%
GRUPO_EVENTO_09: SIM	75,4% ± 7,6%	81,3% ± 4,4%

Validação cruzada

Nesta seção são apresentados os resultados do PGD e do C5.0 no qual utilizou-se o método de validação cruzada de *k* partições para avaliar o desempenho de ambas abordagens, no qual *k* foi definido com valor 6 (empiricamente). É importante salientar que o PGD foi executado com restrição no tamanho das regras e o C5.0 não.

Os atributos escolhidos para os experimentos são apresentados no Quadro 3, os quais foram selecionados de acordo com o interesse dos especialistas de domínio. O Quadro 3 apresenta o cálculo da média x e desvio padrão (**DP**) da taxa de acerto sobre o conjunto de teste das execuções em todas as partições para a classe SIM.

O Quadro 4 apresenta o cálculo da média (â) e desvio padrão (**DP**) da cobertura (*recall*) sobre o conjunto de

Quadro 4 - Cobertura e desvio obtidos por meio do algoritmo C5.0 e PGD utilizando validação cruzada, sobre os dados dos beneficiários do plano de saúde suplementar, 2010

Atributo meta / Valor	C5.0 (x% ± DP%)	PGD (x% ± DP%)
GRUPO_CID_IIIA: SIM	0,73% ± 0,14%	1,80% ± 0,80%
GRUPO_CID_IIIB: SIM	0,31% ± 0,09%	0,76% ± 0,33%
GRUPO_EVENTO_09: SIM	0,56% ± 0,14%	0,58% ± 0,25%

Quadro 5 - Número de antecedentes obtidos por meio algoritmo C5.0 e PGD utilizando validação cruzada, sobre os dados dos beneficiários do plano de saúde suplementar, 2010

Atributo meta / Valor	C5.0 (Média Ant*)	PGD (Média Ant*)
GRUPO_CID_IIIA: SIM	5	4,6
GRUPO_CID_IIIB: SIM	5	3,8
GRUPO_EVENTO_09: SIM	5,8	5

*Ant = Número de antecedentes por regra

teste das execuções em todas as partições para a classe SIM.

Nota-se no Quadro 3 que, considerando todas as iterações, o C5.0 apresenta a melhor média de taxa de acerto em dois dos três atributos. No entanto, o Quadro 4 mostra que a PGD obtém maior cobertura.

O Quadro 5 apresenta o tamanho médio dos antecedentes da melhor regra em cada execução, e pode-se observar que no PGD o número de antecedentes é menor devido a restrição do tamanho das regras. O C5.0 apesar de não ter restrição de tamanho para as regras, em média encontra regras pequenas e portanto de fácil compreensão.

CONCLUSÕES

Neste trabalho foram definidos os algoritmos C4.5 e PGD para servir como base para os experimentos. Tais algoritmos foram aplicados a uma base de dados de beneficiários de plano de saúde suplementar com a finalidade de extrair regras de produção que pudessem auxiliar o processo de tomada de decisão.

Aplicando o algoritmo C5.0 na base de dados, em sua configuração padrão, não foi possível induzir modelos de classificação com mais de uma folha para diversos atributos. Portanto, visando obter melhores resultados, o C5.0 foi aplicado a base de dados utilizando a técnica de custo para erro de classificação.

Com a utilização da técnica de custo para erro de classificação foi possível induzir modelos de classificação com mais de uma folha. Portanto, os resultados obtidos com esta técnica foram utilizados para comparar com os resultados obtidos pelo algoritmo PGD.

Primeiramente, os algoritmos foram aplicados a base de dados utilizando a técnica *houdout* para avaliar a taxa de acerto das regras geradas. Neste experimento, todos os algoritmos obtiveram altas taxas de acerto para a classe NÃO de todos os atributos meta. Por outro lado, ambos os algoritmos obtiveram taxas de acerto baixas para a classe SIM da maioria dos atributos meta. Pode-se concluir que este método não foi eficiênte na extração de regras de classificação com alta taxa de acerto para os algoritmos utilizados.

A partir deste primeiro experimento, observou-se que para os dados utilizados o PGD obteve melhores resultados que o C5.0 para a maioria dos atributos meta para classe SIM e muito similares para a classe NÃO, com isso conclui-se que o PGD tende a obter resultados melhores para bases de dados com distribuição de classes ruim, do que algoritmos baseados em árvore de decisão, principalmente para a classe com menos registros.

Um segundo experimento foi realizado com a mesma base de dados, porém utilizando validação cruzada para avaliar a taxa de acerto dos algoritmos e considerando apenas os três principais atributos meta. Neste experimento o C5.0 obteve uma taxa de acerto média superior em dois dos três atributos meta. Para o outro atributo a maior taxa de acerto média foi obtida pela PGD. Ainda assim, as taxas de acerto obtidas neste experimento não podem ser consideradas satisfatórias, ou seja, as regras geradas não podem ser consideradas conhecimento que possa auxiliar na tomada de decisão.

O C5.0 é um algoritmo consolidado na literatura, ou seja, já possui sua validade confirmada. Desta maneira, como os resultados obtidos pelo PGD foram similares para alguns atributos meta e superiores em outros, pode-se considerar que os resultados obtidos são válidos e portanto o algoritmo PGD é uma abordagem adequada para resolver a tarefa de classificação em mineração de dados com a característica de desbalanceamento de classes.

Por fim, pode-se considerar que os resultados obtidos pela PGD se mostraram melhores do que os resultados obtidos pelo C5.0, pois foi capaz de encontrar regras com maior taxa de acerto para a maioria dos atributos meta.

REFERÊNCIAS

1. Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: An Overview. American Association for Artificial Intelligence. 1992; 13(3):70-57.
2. Tan PN, Steinbach M, Kumar V. Introdução ao DATAMINING mineração de dados. Rio de Janeiro: Ciência Moderna Ltda; 2009.
3. Fayyad U, Shapiro GP, Smyth P. From data mining to knowledge discovery in databases. American Association for Artificial Intelligence. 1996; 13(3):54- 37.
4. Pinheiro CAR. Inteligência Analítica - Mineração de Dados e descoberta de conhecimento. Rio de Janeiro: Ciência Moderna Ltda; 2008.
5. Barros EF, Romão W, Constantino AA, Souza CL. Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar. J. Health Inform. 2011; 3(1)26-19.

6. RuleQuest [Internet]. Australia: Data Mining Tools See5 and C5.0; 1997. [cited 2011 out 21]. Available from: http://www.rulequest.com/see5-info.html
7. Quinlan R. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann Publishers; 1993.
8. Freitas AA. Data mining and knowledge discovery with evolutionary algorithms. Berlin: Springer Publishing Company; 1998.
9. Hall M, Frak E, Holmes G, Pfahringer B, Reutemann P, et. al. The WEKA data mining software: An update. SIGKDD Explorations. 2009; 11(1):10-8.
10. Raines T, Tambe M, Marsella S. Automated assistants to aid humans in understanding team behaviors. In: Proceedings of the 4th International Conference on Autonomous Agents; 2000 jun 3-7; Barcelona, Catalonia, Spain.
11. Ting KM. Inducing cost-sensitive trees via instance-weighting. In: Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery; 1998 set 23-6; Nantes, France.
12. Barros EF, Romão W, Constantino A A, and Souza CL. Programação genética aplicada à mineração de dados sobre beneficiários de planos de saúde suplementar. In: XXXVII Conferencia Latinoamericana de Informática CLEI; 2011 Oct 10-4; Quito Equador. p.1061-77/1.
13. Carvalho DR. Árvore de decisão / Algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados [Tese]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2005.