JHI JOURNAL OF HEALTH INFORMATICS

Uma Abordagem Influenciada por Pré-processamento para Aprendizagem do Processo de Regulação Médica

An Approach Influenced to Pre-processing for Learning Medical Claim Process

Un Enfoque Influenciado por Pre-procesamiento para Aprender el Proceso de Regulación Médica

Flávio Henrique Duarte de Araújo[1], André Macedo Santana[2], Pedro de Alcântara dos Santos Neto[2]

Descritores:
Classificação; Mineração de dados; Planos de Pré-pagamento em Saúde

RESUMO

Objetivo: Apresentar uma metodologia que utiliza técnicas de pré-processamento para melhorar a qualidade dos dados presentes na base de dados de uma Operadora de Plano de Saúde para, em seguida, utilizar técnicas de aprendizado de máquina objetivando aprender o processo de regulação médica. **Métodos:** Foram utilizadas as métricas de: precisão, *recall*, acurácia, *f-measure*, área sob a curva ROC e índice *kappa* para a comparação dos algoritmos de classificação C4.5, Naive Bayes e Multi Layer Perceptron. **Resultados:** Para a validação dos resultados foi utilizado o *cross-validation* 10-*fold*. O melhor classificador foi o C4.5, com taxa de acerto superior a 91%. **Conclusão:** Demonstrou-se que o processo de regulação pode ser aprendido por algoritmos de aprendizagem de máquina, porém faz-se necessário utilizar técnicas de pré-processamento para melhorar a qualidade dos dados.

Keywords:
Classification; Data Mining; Prepaid Health Plans

ABSTRACT

Objective: Present a methodology that uses preprocessing techniques to improve the quality of the data present in Database of a health insurance company to learn the medical claim process. **Methods:** We use: precision, recall, f-measure, area under the ROC and kappa to compare classification algorithms C4.5, Naive Bayes and Multi-Layer Perceptron. **Results:** In order to validate the results we used cross-validation 10-fold. The best classification algorithm was the C4.5 with accuracy higher than 91%. **Conclusion:** We demonstrate that the medical claim process can be learned by machine learning algorithms; however it is need to use preprocessing techniques to improve quality of the data.

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

RESUMEN

Objetivo: Presentar una metodología que utiliza técnicas de pre-procesamiento para mejorar la calidad de los datos presentes en la Base de datos de un Proveedor plan de salud para aprender el proceso de la regulación médica. **Métodos**: Precisión, recall, acurácia, f-measure, área bajo la ROC y índice kappa se utilizaron comparar diferentes algoritmos de clasificación: C4.5, Naive Bayes y Multi-Layer Perceptron. **Resultados:** Con el fin de validar los resultados que hemos utilizado la validación cruzada 10 fold. C4.5 algoritmo obtenido mejor desempeño con una precisión superior al 91%. **Conclusión:** Se demuestra que el proceso de regulación médica puede ser aprendido por los algoritmos de aprendizaje de máquina, pero primero tienes que utilizar técnicas de pre-procesamiento para mejorar la calidad de los.

[1] Mestre, Universidade Federal do Piauí – UFPI, Teresina (PI), Brasil.
[2] Professor Adjunto, Universidade Federal do Piauí – UFPI, Teresina (PI), Brasil.

Autor Coorespondente: **André Macedo Santana**
e-mail: **andremacedo@ufpi.edu.br**

Artigo recebido: **10/03/2014**
Aprovado: **03/10/2014**

www.jhi-sbis.saude.ws

INTRODUÇÃO

Apesar de possuir um sistema público e universal de saúde, o Brasil tem um dos maiores mercados de saúde suplementar do mundo. A Agência Nacional de Saúde Suplementar (ANS) relata que atualmente existem no Brasil mais de 1.500 operadoras de saúde, totalizando mais de 40 milhões de pessoas assistidas[1]. Porém, segundo a própria ANS, a maioria das Operadoras de Plano de Saúde (OPS) encontra-se em situação financeira bastante complicada, as operadoras de pequeno porte possuem custos acima da sua receita, as de médio porte estão no seu limite de funcionamento e apenas as de grande porte possuem alguma estabilidade financeira.

As OPS possuem diversos mecanismos de controle que verificam constantemente se as requisições estão sendo realizadas de forma correta, evitando solicitações indevidas ou até mesmo fraudes intencionais. Dentre os mecanismos de controle destaca-se a Regulação Médica que tem como objetivo minimizar o desperdício de recursos, viabilizando um serviço justo e com custo adequado.

De forma resumida o processo de Regulação em Saúde funciona da seguinte maneira: após a solicitação de um exame ou um procedimento por um profissional de saúde ou por um prestador associado (hospital ou clínica), a OPS verifica se a solicitação está de acordo com padrões existentes e se os protocolos clínicos estabelecidos entre as partes no ato do contrato são legais; se a análise do profissional regulador for positiva, então o exame/procedimento é autorizado; caso contrário, a seguradora questiona ao solicitante sobre a procedência da requisição e, eventualmente, não autoriza a sua execução do exame/procedimento.

O custo elevado é o principal problema em se ter um processo de regulação eficiente. Nas OPS de pequeno porte, por exemplo, existem poucos eventos por dia que precisam ser regulados, tornando muito caro manter um médico regulador 24 horas por dia analisando as solicitações. Um problema associado a não existência de regulação são os procedimentos glosados (procedimentos que foram autorizados e que não deveriam ser) que sempre geram um grande mal estar entre as partes quando do pagamento. Outro problema ocorre quando é negado um procedimentos que deveria ser autorizado, podendo gerar processos judiciais contra a empresa.

Na maioria das OPS o processo de regulação ocorre com o apoio da informática, ou seja, há registro dos dados dos pedidos e dos resultados das solicitações (autorizados/não autorizados). Logo, é possível utilizar técnicas de aprendizado supervisionado para aprender o processo de regulação a partir da análise de dados armazenados nos Bancos de Dados (BD) das OPS. No entanto, a maioria das bases de dados das OPS apresentam problemas como valores redundantes e valores não preenchidos que reduzem a qualidade da informação existente. Assim, antes de realizar o processo de aprendizagem faz-se necessário utilizar técnicas de pré-processamento para melhorar a qualidade dos dados.

Nesse sentido, o objetivo deste trabalho consiste em utilizar técnicas de pré-processamento para melhorar a qualidade dos dados presentes em um BD de uma OPS para, em seguida, utilizar técnicas de aprendizado de máquina objetivando aprender o processo de regulação médica.

Pré-processamento de Dados

A etapa de pré-processamento de dados engloba uma análise inicial dos dados para se ter uma sólida definição dos mesmos (estrutura das tabelas, valores potencias dos atributos, formatos, tipos de dados, entre outros), além de toda e qualquer operação necessária a escolha dos dados relevantes aos objetivos do usuário, limpeza e transformação dos mesmos para tornar possível que a classificação de dados seja feita pela técnica escolhida[2].

Frequentemente três atores estão envolvidos no processo de aprendizagem automática de alguma tarefa: o analista de dados, o especialista de domínio e o usuário[2].

O analista de dados é aquele que entende das técnicas computacionais envolvidas no processo. Esta pessoa tem conhecimento sobre o funcionamento dos algoritmos e das ferramentas utilizadas, mas não necessariamente conhece o domínio ao qual os dados pertencem. O especialista de domínio, por sua vez, é aquele que conhece o domínio no qual as técnicas serão aplicadas. Já o usuário é aquele que irá efetivamente utilizar o resultado do processo de aprendizagem. É importante comentar que normalmente o usuário não é somente uma pessoa, mas uma empresa ou um departamento de uma empresa. [2] Vale destacar também que embora estes três atores denotem diferentes habilidades, eles não são necessariamente pessoas diferentes. Normalmente os papéis de usuário e especialista de domínio são exercidos por uma mesma pessoa, considerando que o usuário possui conhecimento detalhado do domínio de aplicação[2].

A literatura mostra que a etapa de pré-processamento de dados é um processo semiautomático. Por semiautomático entende-se que este processo depende da capacidade do analista de dados de identificar os problemas presentes nos dados e utilizar os métodos mais apropriados para solucionar cada um deles[2].

As principais tarefas encontradas nessa fase são: tratamento de valores desconhecidos, tratamento de classes desbalanceadas e seleção de atributos.

Tratamento de valores desconhecidos

Um problema relevante em qualidade de dados é a presença de valores desconhecidos ou valores ausentes. Esse problema surge a partir do não registro dos valores de um atributo por defeitos de equipamentos, recusa por parte dos entrevistados em responder determinadas perguntas, valores de preenchimento não obrigatório, entre outras.

O uso da imputação é uma das formas mais comuns de tratamento dos valores desconhecidos. Os métodos de imputação substituem valores desconhecidos por valores estimados. Esses valores estimados são calculados por meio de alguma informação extraída do conjunto de dados como a moda ou a média. Imputação pela moda e pela média são dois exemplos dessas técnicas, nelas os valores desconhecidos são substituídos pela moda ou pela média de todos os outros valores do conjunto[2].

Tratamento de classes desbalanceadas

O problema de classes desbalanceadas representa um

domínio onde uma classe é representada por um grande número de exemplos (classe majoritária), enquanto que a outra é representada por poucos exemplos (classe minoritária). A maioria dos algoritmos de aprendizagem de máquina têm dificuldades em criar um modelo que classifique com precisão os exemplos da classe minoritária[2]. Esse problema agrava-se ainda mais quando o custo da classificação incorreta da classe minoritária é muito maior do que o custo da classificação incorreta da classe majoritária.

Vários pesquisadores têm analisado o problema de aprender a partir de conjuntos de dados com classes desbalanceadas[3]. Uma das formas mais diretas de lidar com classes desbalanceadas por meios de métodos de pré-processamento de dados é alterar a distribuição dessas classes de forma a tornar o conjunto de dados o mais balanceado possível. Existem dois métodos clássicos para balancear a distribuição das classes: remover exemplos da classe mais populosa; e inserir exemplos na classe menos populosa. Em sua versão mais simples, essa adição/remoção é feita de maneira aleatória. Nesses casos, os métodos são comumente chamados de *over-sampling* aleatório e *under-sampling* aleatório[3].

Seleção de atributos

Embora considerando que todos os atributos da base de dados possam ser utilizados, é comum que existam atributos irrelevantes aos objetivos do processo pretendido, sendo importante selecionar um determinado subconjunto de atributos.

A maioria dos autores defende a seleção de atributos relevantes na tentativa de diminuir a complexidade do problema e alcançar um bom desempenho no processo de classificação[4]. Vale destacar que os efeitos imediatos da seleção de atributos são a execução mais rápida e o aumento do desempenho do algoritmo.

As duas principais formas de seleção de atributos são: *embedded* e filtro[5]. As estratégias do tipo *embedded* são diretamente incorporadas no algoritmo responsável pela criação do modelo de classificação. Já as estratégias do tipo filtro são executadas em uma fase de pré-processamento dos dados e procuram pelo conjunto de atributos mais adequado para ser utilizado pelo algoritmo de classificação[6]. Merece ser comentado que nessa fase novos atributos podem ser construídos a partir dos atributos primitivos com o intuito de incorporar informações relevantes e não explícitas nos dados originais. Esse processo de construção é conhecido como *construção de atributos* ou *construção indutiva*[7].

Trabalhos Relacionados

Até o presente momento não foram encontrados na literatura trabalhos que tratem especificamente de ferramentas que auxiliem o processo de regulação médica em OPS, tornando esta pesquisa inovadora do ponto de vista da aplicação. No entanto, alguns trabalhos que utilizam técnicas de pré-processamento para melhorar os dados para, em seguida, realizar um processo de classificação, serviram como influência para a realização deste.

Chazard et al. 2011[8] usam técnicas de Regras de Indução e Árvore de Decisão para tentar descobrir automaticamente casos de efeitos colaterais de medicamentos. As regras encontradas são filtradas, validadas e reorganizadas por um comitê de médicos especializados.

Silachan and Tantatsanawong 2011[9] criaram um serviço informatizado de saúde a partir da extração de informações de textos médicos. Para isto eles utilizaram técnicas de melhoramento da qualidade dos dados, juntamente com Ontologia e Árvore de Decisão. Em Largeron et al. 2011[10] os autores propuseram um critério de seleção de atributos baseado na entropia dos dados. Eles fizeram uma comparação entre o critério proposto e outros critérios clássicos da literatura.

Barros et al. 2011[11] aplicaram técnicas de pré-processamento em um conjunto de dados de beneficiários de um plano de saúde suplementar objetivando prepará-los para a utilização em algoritmos de MD. Martins et al. 2012[12] usaram Árvore de Decisão e Programação Genética Difusa (PGD) em uma base de dados de beneficiários de um plano de saúde suplementar com a finalidade de extrair regras de produção que pudessem auxiliar o processo de tomada de decisão. Em Chimieski et al. 2013[13] os autores utilizaram várias bases de dados médicas para avaliar e comparar diferentes algoritmos de aprendizagem de máquina.

METODOLOGIA PROPOSTA

Neste trabalho foram utilizadas técnicas de pré-processamento para melhorar a qualidade dos dados presentes em um BD de uma OPS. Em seguida, esses dados foram utilizados, juntamente com algoritmos de aprendizagem de máquina para aprender o processo de regulação médica. O diagrama da Figura 1 apresenta todos os passos realizados nesse trabalho e detalhes sobre a execução de cada etapa são descrito a seguir.

Inicialmente todos os dados referentes aos procedimentos odontológicos foram disponibilizados em um banco de dados gerado pelos analistas da OPS (BD Original). Esse BD possuía 164 atributos, mas por razões éticas e legais, todos os dados que identificam pessoas individualmente, como RG, CPF, data de nascimento, endereço, telefone, etc., foram removidos dessa base. Além disso, todos os indivíduos foram identificados por uma chave artificial, gerada pelo banco de dados, permanecendo, portanto, anônimos. Neste trabalho, o banco resultante da eliminação dos atributos que identificam pessoas foi denominado BD-EL.

O BD-EL possuía 133 atributos, desse total muitos eram irrelevantes e não possuíam formato compatível com os algoritmos que foram utilizados na fase de mineração. Portanto, de posse do BD-EL, foi realizado um pré-processamento para melhorar a qualidade dos dados e diminuir a quantidade de informação irrelevante.

No primeiro passo do pré-processamento foi executada uma etapa de seleção de atributos. Esta seleção foi feita de duas formas: seleção manual e seleção automática. Na seleção manual o especialista de domínio e o analista de dados eliminaram: a) atributos que não são

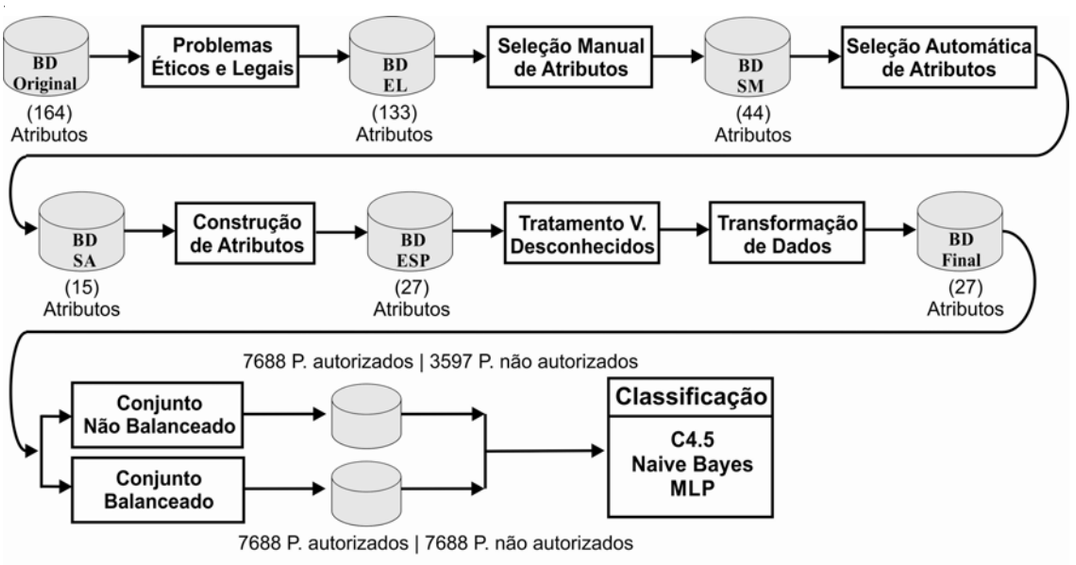

Figura 1 - Metodologia para aprendizagem do processo de regulação médica.

conhecidos e utilizados pelos profissionais no momento da regulação, b) atributos que em todas as instâncias do banco de dados não apresentavam preenchimento; e c) atributos que apresentavam em todas as instâncias do banco de dados valores iguais (preenchimento *default*). Na etapa de seleção manual foram eliminados 89 atributos, portanto somente 44 atributos continuaram na base de dados, chamada de BD-SM.

A partir de BD-SM foi realizada a seleção automática de atributos, onde foi calculado a razão de ganho[14] para cada um dos seus 44 atributos, destes, 29 obtiveram razão de ganho nulo e foram eliminados da base. Com isso, a base de dados resultante, chamada de BD-SA, continha apenas 15 atributos.

Após as etapas de seleção manual e automática o especialista de domínio da OPS criou com base no seu conhecimento sobre o processo de regulação 12 novos atributos referentes ao histórico do beneficiário (Apêndice A). Como para todos os novos atributos o valor da razão de ganho foi maior que zero adicionou-se esses atributos ao BD-SA gerando o BD-ESP.

De posse do BD-ESP, foi necessário tratar o problema de valores desconhecidos, pois alguns dos atributos selecionados possuíam valores não preenchidos. Neste caso, optou-se por realizar a imputação pela média para preencher os valores vazios.

A transformação nos dados foi necessária por que muitas vezes os dados não estavam no formato compatível com os algoritmos de classificação utilizados. Normalização, conversão de atributos qualitativos em quantitativos e extração de informações de atributos de tipo de dados complexos (extração do mês de um atributo do tipo data) foram utilizados para gerar o BD-Final.

Uma característica dos dados do BD-Final era que a quantidade de procedimentos autorizados (7.688) era bem maior que a de procedimentos não autorizados (3.597). Logo, fez-se necessário tratar o problema do desbalanceamento das classes. Para contorná-lo utilizou-se a estratégia de replicação dos dados da classe minoritária para posterior utilização nos algoritmos de aprendizado. Assim, o conjunto balanceado ficou com a mesma quantidade de procedimentos autorizados e não autorizados (7.688). Vencidas todas as fases anteriores, o passo seguinte consistiu em utilizar algoritmos de classificação para tentar aprender a partir do BD-Final o comportamento dos profissionais reguladores.

Com o objetivo de obter o melhor desempenho na aprendizagem do processo de regulação, foram comparados os resultados obtidos por três classificadores: C4.5 (AD), Naive Bayes (NB) e Multi-Layer Perceptro (MLP)[15-17]. Esses classificadores foram escolhidos por serem de diferentes paradigmas: o primeiro é do paradigma simbólico, o segundo é do paradigma estatístico e o terceiro é do paradigma conexionista[16]. As métricas de avaliação são apresentadas a seguir.

Critérios de Avaliação dos Classificadores

A maioria dos critérios de análise dos resultados de uma classificação parte de uma matriz de confusão, que indica a quantidade de classificações corretas e incorretas para cada uma das classes. Uma matriz de confusão é criada baseada em quatro valores: Verdadeiro Positivo (VP), número de procedimentos corretamente classificados como não autorizados; Falso Positivo (FP) número de procedimentos classificados como não autorizados quando, na realidade, eram autorizados; Falso Negativo (FN), número de procedimentos classificados como autorizados quando, na realidade, eram não autorizados e Verdadeiro Negativo (VN), número de procedimentos classificados corretamente como autorizados[13].

A partir destas quantidades algumas taxas estatísticas, exibidas na Tabela 1, podem ser calculadas para avaliar o desempenho dos classificadores.

Outra métrica utilizada foi o índice *Kappa*[18], que é um coeficiente de concordância para escalas nominais que mede o relacionamento entre a concordância, além da casualidade, e a discordância esperada[18]. O índice *Kappa*

Tabela 1 - Taxas estatísticas para a avaliação dos algoritmos.

Nome	Fórmula	Significado
Precisão	$P = \frac{VP}{VP + FP}$	Proporção de verdadeiros positivos em relação a todas as predições positivas.
Recall	$R = \frac{VP}{VP + FN}$	Proporção de verdadeiros positivos em relação a suas predições positivas e suas incorretas predições negativas.
Acurácia	$A = \frac{VP + VN}{VP + FP + FN + VN}$	Proporção de predições corretas, ou seja, o acerto total.
F-measure	$FM = \frac{2 * VP}{2 * VP + FN + FP}$	Proporção de verdadeiros positivos em relação as predições positivas e todas as suas predições negativas.
Área sob a curva ROC (AUC)	-	Representação gráfica da sensibilidade de um classificador dada pela taxa de VP em função da taxa de FP.

pode ser encontrado com base na Equação

$$K = \frac{(observado - esperado)}{1 - esperado}$$

Neste caso, entenda-se por "observado" o valor global para a percentagem correta, ou seja, o somatório da diagonal principal da matriz dividido pela quantidade de elementos. Por "esperado" entenda-se o os valores calculados usando-se os totais de cada linha e cada coluna da matriz.

O nível de exatidão do índice *Kappa* foi classificado conforme a Tabela 2, de acordo com o estabelecido por Landis et al. 1977[19].

Tabela 2 - Nível de exatidão de uma classificação conforme o índice *Kappa*.

Índice *Kappa* (K)	Qualidade
$K \leq 0.2$	Ruim
$0.2 < K \leq 0.4$	Razoável
$0.4 < K < 0.6$	Bom
$0.6 < K \leq 0.8$	Muito Boa
$K > 0.8$	Excelente

RESULTADOS E DISCUSSÕES

Para a aprendizagem automática do comportamento do médico regulador foi utilizada a parte odontológica da base de dados de uma OPS sem fins lucrativos contendo registros coletados desde janeiro de 2007 a janeiro de 2013. Estes dados foram originalmente recebidos como um dump do banco de dados PostgreSQL, acompanhados de um dicionário de dados no formato de planilha Excel. O banco de dados foi importado normalmente em um servidor PostgreSQL 8.4 e manipulado neste ambiente através da linguagem SQL.

Para mostrar a importância do pré-processamento foram realizados três diferentes testes para cada um dos três classificadores especificados anteriormente. Os dois conjuntos de dados formados (replicado e não replicado) também foram utilizados em cada um dos testes. No primeiro teste foram utilizados todos os atributos resultantes após a Seleção Manual (BD-SM), no segundo teste foram utilizados somente os atributos resultantes após a Seleção Automática (BD-SA) e no terceiro teste foram utilizados os atributos selecionados de forma automática juntamente com os atributos criados pelo especialista após o tratamento dos valores desconhecidos e das transformações (BD-Final).

Os algoritmos utilizados foram testados pelo WEKA[20] e o método de avaliação utilizado foi o *cross-validation* 10-folds[21].

A Tabela 3 mostra uma comparação entre os resultados de Precisão e *Recall* para os algoritmos testados utilizando os atributos obtidos após a seleção manual (BD-SM), a seleção automática (BD-SA) e os atributos representados por (BD-Final).

Analisando a Tabela 3 percebe-se que os resultados obtidos com o BD-SM e o BD-SA foram praticamente idênticos. Isso se justifica pela diferença entre essas bases ser apenas os atributos com ganho de informação nulo, que não influenciam no processo de aprendizagem. Porém, para o BD-Final, o resultado para todos os classificadores são melhorados. No tocante ao balanceamento das classes percebeu-se que os melhores resultados são obtidos após o balanceamento e o melhor classificador para estas métricas foi a C4.5.

As Tabelas 4 e 5 mostram respectivamente Acurácia vs *F-Measure e AUC vs Kappa*. Conclusões análogas às obtidas na análise da Tabela 3 podem ser tiradas.

Para avaliar estatisticamente os resultados comparativos entre os classificadores, foi realizado um teste de hipótese[22], com significância de 5%, para verificar se os classificadores testados são significativamente diferentes. Com base nesse teste foi concluído que o classificador C4.5 possui desempenho significativamente diferente e maior que o Naive Bayes e a MLP. Pela análise da Tabela 2 e dos valores de índice *Kappa* apresentados na Tabela 5 foi possível concluir que o Naive Bayes, a MLP e a C4.5 obtiveram, utilizando os atributos de BD-Final, respectivamente o desempenho "Razoável", "Bom" e "Muito bom" para o conjunto de dados não replicado. Para o conjunto replicado o Naive Bayes e a MLP obtiveram o mesmo desempenho, já a C4.5 obteve desempenho "Excelente".

A Figura 2 mostra a comparação entre cada uma das métricas definidas considerando o melhor classificador, o algoritmo C4.5.

Tabela 3 - Comparação entre a Precisão e o *Recall*.

		Precisão			Recall		
		BD SM	**BD SA**	**BD Final**	**BD SM**	**BD SA**	**BD Final**
Não Replicado	Naive Bayes	0,39	0,39	0,57	0,69	0,69	0,75
	MLP	0,53	0,53	0,66	0,71	0,70	0,81
	C4.5	0,74*	0,74*	0,84*	0,84*	0,84*	0,91*
Replicado	Naive Bayes	0,57	0,57	0,74	0,62	0,60	0,76
	MLP	0,62	0,60	0,76	0,63	0,62	0,74
	C4.5	0,81*	0,81*	0,89**	0,86*	0,86*	0,93**

(*) Melhor resultado para cada base de Dados. (**) Melhor resultado por medida.

Tabela 4 - Comparação entre a Acurácia e o *F-Measure*.

		Acurácia			F-Measure		
		BD SM	**BD SA**	**BD Final**	**BD-SM**	**BD-SA**	**BD Final**
Não Replicado	Naive Bayes	0,66	0,66	0,71	0,21	0,20	0,47
	MLP	0,69	0,69	0,77	0,31	0,30	0,61
	C4.5	0,84*	0,84*	0,89*	0,68*	0,68*	0,82*
Replicado	Naive Bayes	0,55	0,55	0,64	0,48	0,48	0,54
	MLP	0,63	0,61	0,75	0,64	0,62	0,74
	C4.5	0,84*	0,84*	0,91**	0,84*	0,84*	0,91**

(*) Melhor resultado para cada base de Dados. (**) Melhor resultado por medida.

Tabela 5 - Comparação entre a AUC e o índice *Kappa*.

		AUC			Kappa		
		BD SM	**BD SA**	**BD Final**	**BD SM**	**BD SA**	**BD Final**
Não Replicado	Naive Bayes	0,57	0,57	0,73	0,10	0,10	0,30
	MLP	0,64	0,63	0,80	0,15	0,14	0,45
	C4.5	0,84*	0,84*	0,90**	0,54*	0,54*	0,73**
Replicado	Naive Bayes	0,58	0,58	0,73	0,10	0,10	0,27
	MLP	0,67	0,65	0,82	0,26	0,24	0,49
	C4.5	0,90*	0,90*	0,95**	0,67*	0,67*	0,82**

(*) Melhor resultado para cada base de Dados. (**) Melhor resultado por medida.

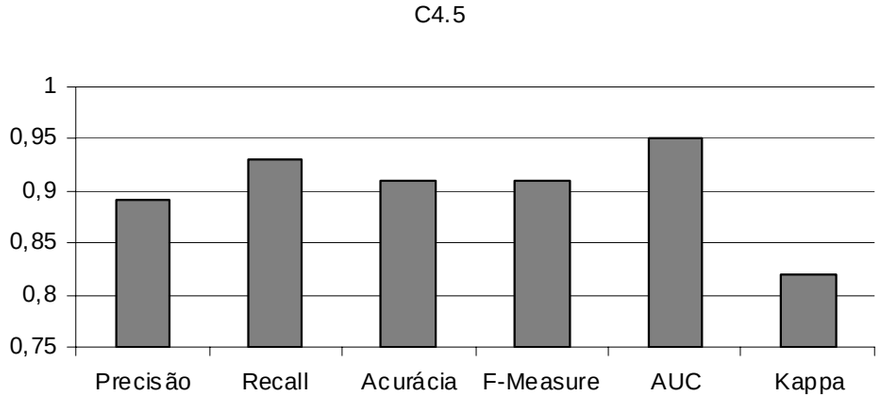

Figura 2 - Métricas obtidas para o melhor classificador.

CONCLUSÃO

Este trabalho mostra a possibilidade de utilizar técnicas de inteligência computacional para modelar o comportamento de profissionais reguladores (profissionais que avaliam se as solicitações médicas devem ou não ser autorizada) utilizando como base os dados de autorizações e não autorizações anteriores de uma determinada OPS.

A principal contribuição está no fato de considerar que a existência de uma ferramenta que auxilie um profissional regulador melhora a qualidade das análises

por parte destes profissionais, uma vez que um número muito grande de solicitações são feitas dia a dia e uma grande quantidade de regras devem ser levadas em consideração para o julgamento das solicitações.

Para o desenvolvimento do sistema de regulação automático foi necessário utilizar técnicas de pré-processamento para melhorar a qualidade dos dados utilizados. A partir dos dados contidos no banco de dados da operadora de saúde foi utilizado um processo de seleção de atributos, em seguida foi necessário o tratamento de valores desconhecidos para determinados atributos. Antes da etapa de classificação também foi necessário tratar o problema de classes desbalanceadas. Na etapa de classificação foram testados três classificadores de diferentes paradigmas (AD, NB e MLP), onde a AD obteve o melhor desempenho na classificação dos procedimentos em Autorizados e Não Autorizados.

A evolução deste trabalho consiste em avaliar outros algoritmos baseados em Árvores de Decisão, visto que a C4.5 obteve desempenho bem superior aos demais classificadores utilizados. Além disso, outras técnicas relacionadas aos problemas de seleção de atributos, tratamento de valores desconhecidos e classes desbalanceadas também devem ser consideradas.

AGRADECIMENTO

Ao CNPq (projeto 560.128/2010) e ao Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (INES), por apoiar parcialmente este trabalho. Também a empresa Infoway pelo incentivo e importantes contribuições para a pesquisa.

REFERÊNCIAS

1. ANS. Caderno de informação da saúde suplementar [Online]. [citado 2013 Dez 13. Disponível em: HYPERLINK "http://www.ans.gov.br" http://www.ans.gov.br .
2. Batista GEAPA. Pré-processamento de dados em aprendizado de máquina supervisionado [tese]. São Paulo (SP): Universidade de São Paulo, Instituto de Ciências Mátematicas e de Computação; 2003.
3. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell data anal. 2002;6(5):429-49.
4. Neves RCD. Pré-processamento no processo de descoberta de conhecimento em banco de dados [dissertação]. Porto Alegre (RS): Universidade Federal do Rio Grande do Sul; 2003.
5. kohavi R. Wrappers for feature subset selection. Artif Intell. 1997; 97: 273-324.
6. Pereira RB. Seleção Lazy de atributos para a tarefa de classificação [dissertação]. Rio De Janeiro (RJ): Universidade Federal Fluminense; 2009.
7. Bloedorn E, Michalski RS. Data-Driven construtive induction. IEEE Intelligence Systems. 1998;13(2):30-7.
8. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. Information Technology in Biomedicine. HYPERLINK "http://www.ncbi.nlm.nih.gov/pubmed/?term=Chazard+E%2C+Ficheur+G%2C+Bernonville+S%2C+Luyckx+M%2C+Beuscart+R.+Data+mining+to+generate+adverse+drug+events+detection+rules.+Information+Technology+in+Biomedicine.+2011" \o "IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society." IEEE Trans Inf Technol Biomed. 2011;15(6):823-30.
9. Silachan K, Tantatsanawong P. Domain ontology health informatics service from text medical data classification. SRII Global Conference. 2011 Mar 29 -Apr 2, San Jose, CA; p. 357-62.
10. Largeron C, Moulin C, Gery M. Entropy based feature selection for text categorization. In: ChuWC, Wong EW, Palakal MJ, Hung CC, editores. Proceedings of the ACM Symposium on Applied Computing; 2011 Mar 21-25; TaiChung, Taiwan. p. 924-8.
11. Barros FE, 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):19-26.
12. Martins OLF, Barros FE, Romão W, Constantino AA, Souza CL. Aplicação de algoritmos de aprendizagem de máquina para mineração de dados sobre beneficiários de planos de saúde suplementar. J. Health Inform. 2012; 4(2):43-9.
13. Chimieski BF, Fagundes RDR. Association and classification data mining algorithms comparicion over medical datasets. J. Health Inform. 2013;5(2): 44-51.
14. Quinlan JR. Induction of decision tree. Machine Learning. 1986;1(1):81-106.
15. Cageiro JN, Pestana MH. Análise categórica, árvores de decisão e análise de conteúdo. Lisboa: Lidel; 2009.
16. Haykin S. Neural networks: a comprehensive foundation. 2nd ed. New Jersey: Prentice Hall; 2001.
17. Matsubara ET. Relações entre ranking, análise ROC e calibração em aprendizado de máquina [tese]. São Carlos (SP): Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação; 2008.
18. Rosenfield GH, Fitzpatrick-lins KA. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing. 1986;(52)2:223-7.
19. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
20. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington, MA : Morgan Kaufmann; 2011.
21. Dietterich TG. Statistical tests for comparing supervised classification learning algorithms. Oregon State University, Technical report; 1997.
22. Congalton RG, Green K. Assessing the accuracy of remotely sensed data: principles and practices. 2a ed. London: CRC Press; 2009.

Apêndice A

Tabela com os atributos utilizados na fase de mineração de dados

Atributos Selecionados (BD)	1) Sexo do beneficiário 2) Idade do beneficiário 3) Tipo do beneficiário 4) Mês de solicitação do exame 5) Código da doença na tabela CBHPM 6) Periodicidade do procedimento 7) Idade mínima para a realização do procedimento 8) Idade máxima para a realização do procedimento 9) Carência 10) Valor do procedimento 11) Valor moderação 12) Nível do procedimento 13) Ordem do procedimento 14) Se o procedimento é especial 15) Se o procedimento passou por perícia inicial
Criados pelo Especialista	16) Número de exames realizados pelo beneficiário naquele mês 17) Número de exames realizados pelo beneficiário naquele semestre 18) Número de exames realizados pelo beneficiário naquele ano 19) Número de exames realizados pelo beneficiário no mês e iguais ao solicitado 20) Número de exames realizados pelo beneficiário no semestre e iguais ao solicitado 21) Número de exames realizados pelo beneficiário no ano e iguais ao solicitado 22) Número de exames de mesmo nível de complexidade realizado no mês 23) Número de exames de mesmo nível de complexidade realizado no semestre 24) Número de exames de mesmo nível de complexidade realizado no ano 25) Número de guias solicitadas pelo beneficiário no mês 26) Número de guias solicitadas pelo beneficiário no semestre 27) Número de guias solicitados pelo beneficiário no ano