Aplicação de Técnicas de Inteligência Computacional para Detecção de Fraude em Comércio Eletrônico

Rafael A.F. Lima [1], **Adriano C.M. Pereira** [1]

[1]Departamento de Computação (DECOM)
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Belo Horizonte - MG, Brasil

rafaelfrancalima@gmail.com, adriano@decom.cefetmg.br

Abstract. *The volume of electronic transactions has raised a lot in last years, mainly due to the popularization of e-commerce. We also observe a significant increase in the number of fraud cases, resulting in billions of dollars losses each year worldwide. Therefore it is important and necessary to develop and apply techniques that can assist in fraud detection, which motivates our research. This work aims to apply and evaluate computational intelligence techniques to identify fraud in electronic transactions, more specifically in credit card operations, using Bayesian Networks and Logistic Regression. In order to evaluate the techniques, we define a concept of economic efficiency and apply them in an actual dataset of the most popular Brazilian electronic payment service. Our results show good performance in fraud detection, presenting gains up to 35.61% compared to the actual scenario of the company.*

Resumo. *O volume de transações eletrônicas cresceu significativamente nos últimos anos, devido principalmente à popularização dos serviços de comercio eletrônico. Com essa popularização aumentou consideravelmente os casos de fraudes, resultando assim em um prejuízo de bilhões de dólares em todo o mundo. Assim, é necessário prover mecanismos capazes de auxiliar na detecção de fraude, foco principal deste projeto de pesquisa. Este trabalho de iniciação científica visa a modelagem e aplicação de técnicas de inteligência computacional para detectar fraudes em transações eletrônicas, mais especificamente em operações envolvendo cartão de crédito, utilizando as técnicas de Regressão Logística e Redes Bayesianas. Para avaliar a eficácia das técnicas, definimos o conceito de Eficiência Econômica e o aplicamos a um cenário real do serviço de pagamento eletrônico mais popular do Brasil. Nossos resultados apresentaram bom desempenho para detecção de fraude, com um ganho de 35,61% em relação ao cenário real da empresa.*

1. Introdução

Apesar de seu enorme sucesso, a Web apresenta inúmeros desafios, sendo que esse mesmo sucesso é responsável por uma grande parte deles. As pessoas, ao acessarem a Web, seja por meio das máquinas de busca, seja no contexto de novos serviços de caráter mais "social" e colaborativo, ainda encontram enormes dificuldades para alcançar seus objetivos com sucesso [J. Hendler and Weitzner 2008, Amiratti 2007]. Isso se deve não apenas à quantidade imensa de informação presente na Web e que vem aumentando consideravelmente, mas também à questão da credibilidade em sistemas computacionais [Tseng and Fogg 1999], no que diz respeito à informação, aos serviços, às fontes e à própria infra-estrutura computacional.

Entre os principais desafios encontrados na Web destaca-se o grande número de fraudes eletrônicas, impulsionadas pelo desenvolvimento de novas tecnologias, que permitiu uma maior facilidade de comunicação e disseminação de conteúdo. Essa facilidade tem atraído a atenção de criminosos cujo objetivo é obter, por meio de ações fraudulentas, vantagens financeiras, informações privilegiadas, dentre outros. Na área tecnológica, pode-se citar vários tipos de fraudes como, por exemplo, fraudes em telecomunicações e no comércio eletrônico, sendo que cada uma delas possuem suas características e necessitam de metodologias diferentes para serem combatidas.

Neste projeto de iniciação científica aplicamos e avaliamos técnicas baseadas em inteligência computacional para identificar fraudes em transações eletrônicas, mais especificamente em operações envolvendo cartão de crédito. Para validar a nossa metodologia, foi utilizado um cenário real de um sistema para pagamento e criado o conceito de Eficiência Econômica para avaliar as técnicas utilizadas. Os resultados obtidos com essa pesquisa foram bastante satisfatórios, alcançando ganhos de até 35.61% quando comparado ao cenário real da empresa na atualidade.

Este trabalho propiciou um grande aprendizado e amadurecimento acadêmico-científico ao bolsista, bem como uma oportunidade única de interação com a indústria, já que este projeto foi feito em cooperação com o Universo OnLine, um dos maiores provedores de serviços Web da América Latina. Dentre as principais contribuições e méritos deste trabalho, destaca-se: (i) os resultados bastante significativos alcançados, que representam ganhos de até 36% em relação ao cenário real da empresa, que foi a linha de base experimental; (ii) ter sido o único projeto aceito no Programa UOL Bolsa Pesquisa do CEFET-MG; (iii) o artigo titulado "Applying Logistic Regression to Rank Credibility in Web Applications", que foi publicado no evento WEBIST'2011 [Lima and Pereira 2011].

O restante deste artigo está organizado da seguinte forma. A Seção 2 descreve alguns trabalhos correlatos ao tema de pesquisa deste projeto. A Seção 3 apresenta uma breve descrição sobre as técnicas de redes bayesianas e de regressão logística, base dos modelos usados neste trabalho. A Seção 4 descreve a metodologia, a qual será aplicada na seção 5 em nosso estudo de caso com dados reais providos pela empresa Universo Online (UOL). Finalmente, a seção 6 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Existem diversas pesquisas sobre métodos para detecção de fraude [Fawcett and Provost 1997, Maranzato et al. 2010, Barse et al. 2003, Lundin et al. 2002], onde cada qual tem suas características, dependendo do tipo de fraude que deseja-se prevenir. Entretanto podemos perceber que técnicas baseadas em mineração de dados estão sendo muito utilizadas e desempenham um importante papel na detecção de fraude. Isso acontece devido a mineração de dados permitir extrair informações necessárias para resolução do problema a partir de um vasto conjunto de dados.

Clifton Phua et al [Phua et al. 2005] realizaram amplo estudo e publicaram diversos artigos relacionados com detecção de fraude, usando mineração de dados e suas técnicas derivadas para solucionar o problema. O autor discute diferentes estratégias para solucionar o problema, baseando em algoritmos de treinamento não supervisionado ou supervisionado, aplicando esses em dados desbalanceados, isto é dados cujo a variável a ser analisada contém uma quantidade muito maior de registros em uma das suas classes.

No treinamento supervisionado o algoritmo analisa cada transação para que matematicamente determine o padrão de uma transação de fraude e possa estimar seu risco.

Redes Neurais, *Support Vector Machine* (SVM), Árvore de Decisão e Redes Bayesianas são algumas das técnicas que adotam essa estratégia. Maes et al [Maes et al. 1993] usa o algoritmo denominado *STAGE* para redes bayesianas e o algoritmo *backpropagation* para detectar fraude em transações com cartões de crédito através de Redes Neurais. Os resultados obtidos por Maes mostram que Redes Bayesianas se mostram mais precisas e simples de serem aplicadas, entretanto são mais lentas quando se precisa inserir novos registros. Técnicas como SVM são destaque como estado-da-arte, mas o SVM não apresentou eficiência, em termos de escalabilidade, em nosso caso, demandando semanas para completar uma execução do modelo.

No treinamento não supervisionado não se tem conhecimento prévio de quais transações resultaram em fraude ou não fraude. Alguns exemplos de técnicas que utilizam treinamento não supervisionado são *Clustering* e Detecção Anômala. Netmap [Netmap 2004] descreve o funcionamento do algoritmo de *Clustering* para previsão de fraude. Bolton and Hand [Bolton and Hand 2002] propõem um método de detecção de fraude em transações de cartão de crédito utilizando a Detecção Anômala.

Os trabalhos aqui brevemente descritos sugerem a crescente necessidade de prover mecanismo que sejam capazes de detectar fraudes, para que assim usuários possam utilizar com segurança os serviços online.

3. Fundamentos Conceituais

3.1. Redes Bayesianas

Segundo Maes [Maes et al. 2001], Redes Bayesianas são grafos acíclicos dirigidos que representam dependência entre as variáveis de um modelo probabilístico, onde os nós são os atributos e os arcos representam os relacionamentos de influência entre as variáveis. As Redes Bayesianas podem ser representadas graficamente por um conjunto de variáveis e um conjunto de arcos que ligam essas variáveis, como mostrado na Figura 1. A seta indica o sentido de influência de um elemento sobre o outro, logo o evento *A* e *B* influenciam o evento *D*, que influencia o evento H, assim consecutivamente. Já o evento *e* é um evento independente, que acontece sem influencia de nenhum evento. [Cooper and Herskovits 1992].

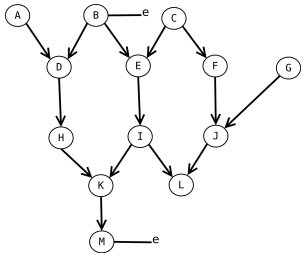


Figura 1. Exemplo conceitual de Redes Bayesianas.

Já do ponto de vista matemático Redes Bayesianas são derivadas do teorema de Bayes, o qual mostra que a probabilidade condicional de um evento A_i dado um evento B, pode ser calculado pela Equação 1.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{1}$$

, onde $P(A_i|B)$ é a probabilidade de A dado que B ocorreu.

Logo na aplicação do problema de detecção de fraudes, a partir do teorema de Bayes são calculadas a probabilidade de encontrar uma fraude, dado que outros eventos ocorreram, montando assim uma tabela de probabilidades condicionadas. Que pode ser finalmente calculada pela Equação 2.

$$P(x_i,...,x_n) = \prod_{i=0}^{n} P(x_i|Pais(X_i)) \tag{2}$$

, onde os pais de X(i) são determinados por um grafo, como da Figura 1.

3.2. Regressão Logística

Regressão logística é uma técnica estatística que produz, a partir de uma série de variáveis explicativas, um modelo que permita a predição de valores tomados por uma variável dependente categórica. Assim através de um modelo de regressão, é possível calcular a probabilidade de ocorrência de um evento, através da função de ligação, conforme descrita na Equação 3:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_i x_i)}}, \tag{3}$$

onde $\pi(x)$ é a probabilidade de sucesso quando o valor da variável preditiva é x, β_0 é uma constante usada para ajuste e β_i são os coeficientes das variáveis preditivas [Hosmer 2000]. Para encontrar a estimativa dos coeficientes β na Equação 3, é usada a técnica de máxima verossimilhança que maximiza a probabilidade de obter o grupo observado de dados, através do modelo estimado.[Casella and Berger 2002].

O modelo de regressão logística binária é um caso especial de modelo linear generalizado, que usa a função de ligação logit para obter as estimativas dos coeficientes da Equação 3 [Venables et al. 2009]. Após encontrar os coeficientes, é possível encontrar a probabilidade de sucesso, que nessa pesquisa é a probabilidade de fraude, aplicando na Equação 3 os valores das estimativas dos coeficientes encontrados.

4. Metodologia

Para a realização desse projeto foi utilizada uma metodologia comum às duas técnicas. Essa metodologia inicia-se com o entendimento e modelagem do problema de detecção de fraudes, com uma completa caracterização do cenário que será descrito na Seção 5. Antes da execução das técnicas foram feitos alguns ajustes no conjunto de dados, como retirada de dados pouco relevantes e categorização de variáveis numéricas.

Além desses pequenos ajustes foi feito uma seleção dos atributos mais significativos para detectar fraude, para isso foram utilizadas a técnica de *forward stepwise regression*, a qual baseia-se no conceito de verossimilhança e está implementada no software estatístico R [Team 2009, Version 2006]. Além da técnica *Stepwise*, foi utilizada a técnica *Infogain*, que mostra o ganho relativo com a inclusão de cada variável. Para a realização dessa técnica utilizou-se o software Weka *(Waikato Environment for Knowledge Analysis)*, um software livre para mineração de dados, que possui sua licença baseada na GPL *(General Public Licence)* [Witten and Franku 2005].

Após a seleção dos atributos através das técnicas *Stepwise* e *Infogain*, foram realizados diversos testes de caracterização a fim de encontrar os atributos que melhor se adaptam a cada técnica. Esses atributos serão descritos na Seção 5.2.

Após selecionar os atributos, utilizou-se um conjunto de treinamento e testes para validar os modelos. Essa divisão foi realizada utilizando as três primeiras semanas do mês como treinamento e as duas últimas como teste, dessa forma fica clara a aplicação em um cenário real e preserva a generalidade dos modelos. Além dessa abordagem, utilizou-se a técnica de *K-fold-Cross-Validation*, definindo o número de subamostras como 5. A implementação das técnicas foi realizada em dois ambientes, onde cada ambiente possui diferentes parâmetros. A escolha dos melhores parâmetros para cada técnica foi feita através de diversos testes, onde foram alterados os parâmetros para cada técnica e escolhidos os que melhores se adaptaram aos modelos.

Após a execução dos algoritmos foi feito um *ranking* pelo grau de probabilidade de fraude de cada transação, onde no topo do *ranking* encontram-se os registros com maior probabilidade de serem fraudes. Após a construção do *ranking* foi feita uma análise em suas diversas faixas, verificando a precisão, revocação e uma função objetivo denominada Eficiência Econômica (EE) dos modelos.

Foi utilizado o conceito de Eficiência Econômica porque a base de dados utilizada nesse trabalho é desbalanceada tendo um volume muito maior de transações válidas. Portanto se classificássemos todas as transações como válidas (não fraude) e medíssemos a precisão, encontraríamos uma precisão muita alta independentemente do modelo utilizado, o que parecia ser uma boa precisão, mas na verdade estaria acarretando um prejuízo de milhares de reais, ao afirmar que 100% das fraudes eram legais.

A estratégia de Eficiência Econômica também tornou-se interessante, pois em transações que envolvem valores financeiros o custo de um falso positivo e um falso negativo não são o mesmo, estima-se que esse custo na grande parte das companhias é de ordem de 1:100. Ou seja, se perde bem mais prevendo que uma operação fraudulenta é legal, do que o contrário. Sendo assim, toda a análise foi feita em escalas do *ranking*, considerando a função de Eficiência Econômica que será descrita na Equação 4.

$$EE = \sum_{i=1}^{Num.Transacoes} G * k - P * (1-k) \tag{4}$$

, onde: **K** é uma constante que simboliza o percentual que o mercado eletrônico fatura em cada transação; **G** é o valor que o mercado eletrônico arrecadou após a aplicação do modelo. Ou seja valores das transações que os modelo preveram como confiável e realmente eram confiáveis; **P** é valor que o mercado eletrônico perdeu após a utilização do modelo, ou seja valores das transações que o modelo previu que eram válidas e na verdades eram fraudes (falso positivo).

Além da EE obtida pela técnica, é preciso descrever o conceito de Eficiência Econômica Máxima (EE_{Max}), que determina o maior valor que poderia ser arrecadado pelo mercado eletrônico, onde foi aplicado esta metodologia, sendo que esse ocorreria em um cenário fictício, onde um modelo ideal teria 100% de precisão e 100% de revocação. Ou seja, o modelo iria prever com sucesso todas as situações de fraudes, e não identificaria nenhuma situação de não fraude como sendo de fraude. Também foi definido o conceito de Eficiência Econômica Real(EE_{Real} que é a eficiência obtida pelo mercado eletrônico na atualidade com seus métodos para detecção de fraude.

Essa metodologia está ilustrada na Figura 2. Na próxima seção será apresentada a instanciação dessa metodologia aplicada a um estudo de caso utilizando dados reais.

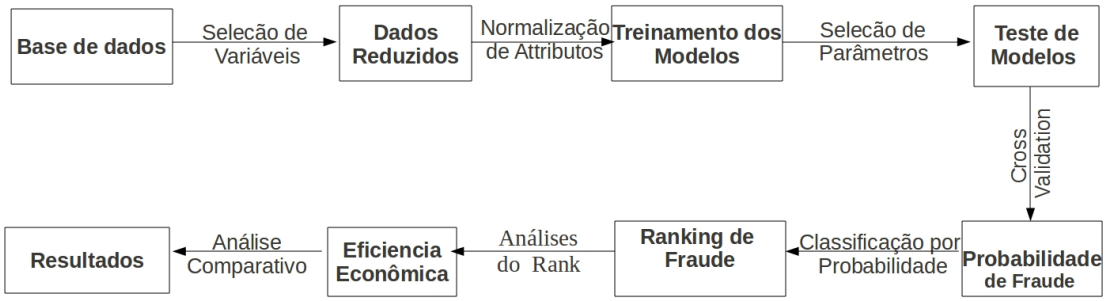


Figura 2. Metodologia para construção dos modelos de detecção de fraude

5. Estudo de Caso

Esta seção apresenta o estudo de caso, onde foi aplicada a metodologia para detectar fraudes em transações eletrônicas, mais especificamente em operações com cartão de crédito, evidenciadas através das transações que ocorreram *charge back* (estorno no cartão de crédito quando a transação não é reconhecida como válida).

5.1. Visão geral da base de dados

Nesse estudo de caso estamos utilizando a base de dados proveniente do sistema para pagamentos online PagSeguro [1] do grupo Universo Online Inc. (UOL)[2] . O PagSeguro hoje é o sistema de pagamento mais utilizado no Brasil, ficando clara a necessidade de previsão de fraudes nesse sistema, pretendendo assim trazer mais segurança para usuários que o utilizam e reduzir o déficit econômico ocasionados por esse tipo de transações.

No PagSeguro cada transação é composta por dezenas de atributos, de mais diferentes tipos e um desses atributos refere-se ao status da transações, podendo acarretar em transações seguras e transações que ocorreram *charge back*.

A Tabela 1 mostra um pequeno resumo referente aos valores médios de transações no PagSeguro, a partir dessa tabela nota-se que as transações que ocorreram *charge back* tem, na média, maiores valores do que as transações válidas. A cobertura desta quanto dos demais resultados apresentados nesse artigo é de outubro de 2010 a março de 2011.

Cobertura (tempo)	Out/2010 a Mar/2011
Valor Médio de transação (R$)	75,11
Valor Médio de transação fraudulentas (R$)	167,99
Valor Médio de transação válidas (R$)	74,77

Tabela 1. Dataset do UOL PagSeguro - Informações Gerais

A Figura 3 apresenta o gráfico sobre a quantidade relativa de *charge back* para cada mês. Esse gráfico mostra um fator impactante para essa pesquisa, já que em meses que possuem maior quantidade de *charge back*, as técnicas de mineração de dados tendem a ser mais eficazes. Neste gráfico também é possivel perceber que existe uma proporção muito maior de transações legais, já que a maior porcentagem de *charge back* em um mês é de 0.5%. Essa diferença deixa os dados desbalanceados, com a classe de transações válidas muito maior, o que foi um dos maiores desafios dessa pesquisa. .

Através de técnicas para seleção de atributos e análise visual de distribuições e valores semânticos explicadas na Seção 4, foi selecionado um conjunto de 11 atributos que têm maior potencial para caracterizar a variável resposta como fraude. São eles:

[1] http://pagseguro.uol.com.br

[2] http://www.uol.com.br

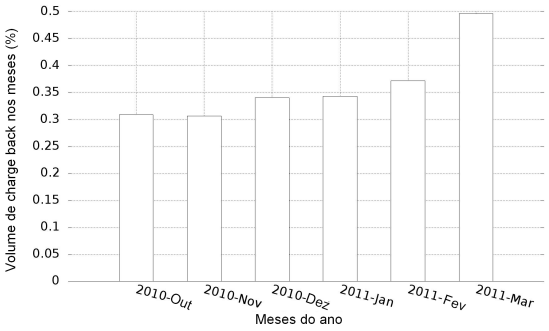


Figura 3. Quantidade relativa de *charge back* ao longo dos meses.

- **Valor:** valor em reais da transação.
- **Pontos da Transação:** pontos gerados por uma transação, os envolvidos na transação atribuem pontos para transações bem sucedidas, mal sucedidas e bem sucedidas sem pontuação (1,-1,0), respectivamente.
- **Hora da transação**: horas em formato inteiro na qual a transação foi realizada.
- **TipoConta**: define o tipo de conta do usuário, (empresarial, vendedor, comprador).
- **Tempo Cadastro Comprador**: tempo em que o comprador está registrado em dias.
- **Tempo Cadastro Vendedor**: tempo em que o vendedor está registrado em dias.
- **Categoria Principal**: categoria em que o produto comercializado pertence.
- **Bandeira do Cartão**: bandeira do cartão de crédito, que foi realizada a operação.
- **Idade**: idade do comprador da transação.
- **Flag DDD**: compara se o código DDD do usuário do PagSeguro é similar ao portador do cartão de crédito.
- **Estado de Cadastro**: campo referente à unidade federal do usuário.

5.2. Resultados

Após a implementação descrita na Seção 4, obteve-se dois modelos para gerar os testes, o *Modelo de Regressão*, que utiliza a técnica de Regressão Logística para predição de valores de fraudes, e o Modelo Bayesiano, que utiliza a técnica de Redes Bayesianas. Nesta seção será feita uma comparação entre os dois modelos quanto ao grau de precisão, revocação e Eficiência Econômica, descrita pela Equação 4.

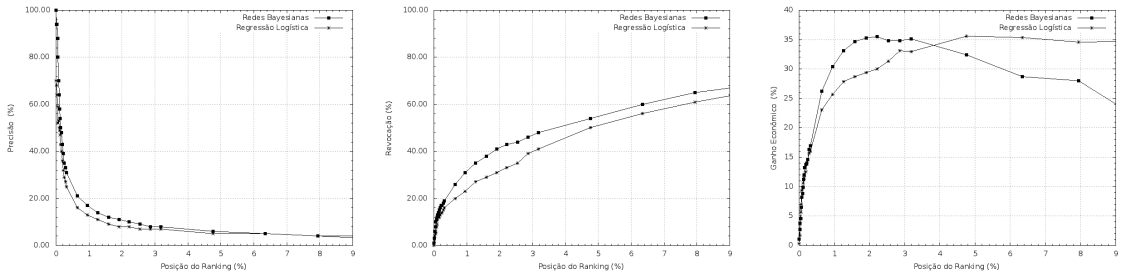
A Tabela 2 apresenta um resultado comparativo entre o Modelo Regressão e o Modelo Bayesiano referente à precisão, revocação e Eficiência Econômica em um conjunto de dados de outubro/2010 a março de 2011, usando um treinamento das três primeiras semanas de um mês e testando na quarta e quinta semanas. Analisando essa tabela nota-se que no geral o Modelo Bayesiano foi mais eficaz que o Modelo de Regressão Logística, ambos destacaram-se no mês de março com uma Eficiência Econômica de 35,53% e 35,61%, respectivamente, comparados ao cenário real do UOL PagSeguro.

Para analisar o comportamento dos modelos em diferentes faixas do *ranking* foram construídos gráficos de precisão, revocação e Eficiência Econômica para os diferentes períodos. Devido à restrição de espaço, destaque-se apenas o mês de março/2011, uma vez que foi o que apresentou os melhores resultados.

		Bayes	Logist
Out.	**Prec.**	7,05	4,10
	Rev.	18,93	27,52
	Posição do Rank.	0,79	1,98
	Eficiência Econômica	**14,33**	12,03
Nov.	**Prec.**	14,70	8,33
	Rev.	32,38	36,67
	Posição do Rank.	0,73	1,47
	Eficiência Econômica	**29,70**	28,73
Dez.	**Prec.**	7,40	3,53
	Rev.	21,08	30,20
	Rank.	1,16	3,49
	Eficiência Econômica	**16,61**	10,64
Jan.	**Prec.**	8,78	9,70
	Rev.	25,56	21,19
	Posição do Rank.	1,30	0,98
	Eficiência Econômica	**16,57**	15,54
Fev.	**Prec.**	7,78	6,06
	Rev.	42,96	44,62
	Posição do Rank.	3,10	4,13
	Eficiência Econômica	**27,40**	25,75
Mar.	**Prec.**	9,93	5,38
	Rev.	43,01	49,94
	Posição do Rank.	2,22	4,76
	Eficiência Econômica	35,53	**35,61**

Tabela 2. Resultados - outubro/2010 a março/2011

A Figura 4(a) apresenta um gráfico comparativo das duas técnicas, analisando a precisão. Pode-se notar que o Modelo Bayesiano foi mais eficaz que o Modelo Regressão, obtendo precisão de 100% para as 0.01% transações que esse classificou como maior probabilidade de fraude e aproximadamente 60% de precisão até os 0,11% registros mais bem classificados para detectar fraude. Já a técnica de Regressão Logística atingiu 70% nos 0,08% primeiros registros, e 50% nos 0,10% registros a partir do topo do *ranking*.



(a) Precisão (b) Revocação (c) Eficiência Econômica

Figura 4. Resultados Comparativos

A Figura 4(b) apresenta um gráfico comparativo do desempenho das técnicas relativo à revocação. Pode-se observar que o Modelo Bayesiano obteve as maiores coberturas classificando menos registros como fraude, quando comparado com o Modelo de Regressão Logística. A Figura 4(c) mostra um gráfico comparativo do desempenho das técnicas quanto a seu ganho sobre o modelo atual. Esse ganho é calculado pela Equação 5.

$$EE_{Relativa} = (EE - EE_{Real})/(EE_{Max} - EE_{Real}) \tag{5}$$

É possível verificar na Figura 4(c) que na maioria das faixas o Modelo Bayesiano obteve melhores resultados (Eficiência Econômica) do que o Modelo de Regressão. O máximo de desempenho obtido pelo Modelo Bayesiano foi classificando 2,22% dos registros como fraude, onde o modelo obteve um ganho de 35,53% de ganho sobre o modelo real atual do PagSeguro, já no Modelo de Regressão Logística a maior Eficiência Econômica foi classificando 4,76% dos registros como fraude, onde o modelo obteve ganho de 35,61%, superando levemente o Modelo Bayesiano. Os resultados aqui apresentados indicam um ganho bastante considerável quando comparados com o cenário atual utilizado

pelo PagSeguro, com o potencial de reduzir em até 36% o prejuízo causado pelas fraudes eletrônicas neste serviço da empresa.

6. Conclusão

Neste projeto de iniciação científica, foram aplicadas técnicas de inteligência computacional para modelar e detectar preventivamente fraudes em transações Web, mais especificamente em transações de comércio eletrônico que utilizam cartão de crédito. Utilizou-se diferentes técnicas de inteligência computacional, dando destaque neste trabalho a um comparativo entre as técnicas de Regressão Logística e Redes Bayesianas.

Para comparação desses modelos, foi utilizado um conjunto de dados providos pelo serviço de pagamento eletrônico UOL PagSeguro, contendo milhões de transações entre os meses de outubro de 2010 e março de 2011. Nota-se nessa pesquisa que o desbalanceamento das classes de fraude e não fraude foi um fator determinante para a previsão de fraudes. Devido a termos uma proporção muito maior de transações válidas, as técnicas em geral encontram dificuldades para reconhecer o padrão de não fraude, assim os piores resultados foram encontrados em meses com menor volume de transações fraudulentas. Apesar dessa dificuldade os resultados apresentam ganhos bastante satisfatórios quando comparado ao cenário real do PagSeguro na atualidade.

A fim de deixar a comparação entre os modelos mais fidedigna, foi utilizado o conceito de Eficiência Econômica (EE), que descreve a melhoria financeira obtida pelo modelo, sendo que a melhor Eficiência Econômica obtida foi de 35,61%. Conclui-se que o Modelo Bayesiano foi mais eficaz na maioria dos experimentos, obtendo em março/2011 sua melhor Eficiência Econômica, com um ganho de 35,53% sobre o cenário real. Já o Modelo de Regressão também apresentou em março/2011 o seu melhor ganho, chegando a superar o Modelo Bayesiano, com Eficiência Econômica de 35,61%.

Um dos maiores desafios dessa pesquisa foi a natureza dos dados, já que não foi utilizada nenhuma técnica de balanceamento simples,preservando assim as características do cenário, e assim consequentemente trabalhos com um volume de fraude com menos de 1% do volume total. Assim, como trabalhos futuros surge a necessidade de utilizar técnicas para lidar com dados desbalanceados sem perder a generalidade dos dados. Uma possível solução seria atribuir maiores pesos para as classe com menor volume e menor pesos para as classes com maior volume [Chen et al. 2004]. Outra solução que pretendemos estudar é a utilização de modelos híbridos que compreendem a união de mais de uma técnica de inteligência computacional.

Agradecimentos

Este trabalho foi parcialmente patrocinado pelo Universo OnLine S. A. - UOL (www.uol.com.br), através do programa UOL Bolsa Pesquisa (processo número 20110212153600) e parcialmente patrocinado pelo Instituto Nacional de Ciência e Tecnologia para a Web (CNPq no. 573871/2008-6), CAPES, CNPq, Finep, e Fapemig.

Referências

Amiratti, S. (2007). Google's udi manber- search is a hard problem. In *ReadWriteWeb Blog*.

Barse, E. L., Kvarnström, H., and Jonsson, E. (2003). Synthesizing test data for fraud detection systems. In *Proceedings of the 19th Annual Computer Security Applications Conference*, ACSAC '03, pages 384–, Washington, DC, USA. IEEE Computer Society.

Bolton, R. J. and Hand, D. J. (2002). Unsupervised Profiling Methods for Fraud Detection. *Statistical Science*, 17(3):235–255.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, 2nd edition.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *Discovery*, (1999):1–12.

Cooper, G. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. data mining and knowledge discovery.

Hosmer, D. W. (2000). *Applied Logistic Regression*. Wiley, New York, 2nd edition.

J. Hendler, N. Shadbolt, W. H. T. B.-L. and Weitzner, D. (2008). Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51:60–69.

Lima, R. and Pereira, A. (2011). Applying logistic regression to rank credibility in web applications. In *7th Intl Conference on Web Information Systems and Technologies*.

Lundin, E., Kvarnström, H., and Jonsson, E. (2002). A synthetic fraud data generation methodology. In *Proceedings of the 4th International Conference on Information and Communications Security*, ICICS '02, pages 265–277, London, UK. Springer-Verlag.

Maes, S., karl Tuyls, Vanschoenwinkel, B., and Manderick, B. (2001). Credit card fraud detection using bayesian and neural networks. *Vrije Universiteir Brussel*.

Maes, S., Tuyls, K., Vanschoenwinkel, B., and Manderick, B. (1993). Credit card fraud detection using bayesian and neural networks. In *Maciunas RJ, editor. Interactive image-guided neurosurgery. American Assoc. Neurological Surgeons*, pages 261–270.

Maranzato, R., Pereira, A., Neubert, M., and do Lago, A. P. (2010). Fraud detection in reputation systems in e-markets using logistic regression and stepwise optimization. *SIGAPP Appl. Comput. Rev.*, 11:14–26.

Netmap (2004). Fraud and crime example brochure.

Phua, C., Lee, V., Smith-Miles, K., and Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research.

Team, R. (2009). R: A language and environment for statistical computing. `http://www.r-project.org`. The R Development Core Team.

Tseng, S. and Fogg, B. J. (1999). Credibility and computing technology. *ACM?42*, 429:39–44.

Venables, W. N., Smith, D. M., and the R Development Core Team (2009). An introduction to r. `http://www.cran.r-project.org`.

Version, T. R. D. C. T. (2006). The r project for statistical computing. `http://www.r-project.org`.

Witten, I. H. and Franku, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.