Uma experiência no Balanceamento Artificial de Conjuntos de Dados para Aprendizado com Classes Desbalanceadas utilizando Análise ROC

Ronaldo Cristiano Prati Gustavo E. A. P. A. Batista Maria Carolina Monard

Laboratório de Inteligência Computacional (LABIC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Av. do Trabalhador São-Carlense, 400 - Centro - Cx. Postal 668
São Carlos - São Paulo - Brasil CEP 13560-970
e-mail {prati,gbatista,mcmonard}@icmc.usp.br

Resumo

No que se refere a conjuntos de dados desbalanceados, algoritmos de aprendizado supervisionado podem encontrar dificuldades na indução de modelos de classificação. Uma das alternativas para solucionar esse problema é artificialmente balancear o conjunto de dados. Neste trabalho, utilizamos análise ROC para avaliar o desempenho de vários desses métodos em um conjunto de dados. Análise ROC é utilizada uma vez que ela é uma alternativa mais apropriada para analisar o desempenho desses modelos.

1 Introdução

Muitos aspectos podem influenciar o desempenho de um modelo de classificação criado por um sistema de aprendizado supervisionado. Um desses aspectos está correlacionado com a diferença entre o número de exemplos pertencentes a cada uma das classes. Quando essa diferença é grande, os sistemas de aprendizado podem encontrar dificuldades em induzir o conceito relacionado a classe minoritária. Nessas condições, modelos de classificação que são otimizados em relação a precisão têm tendência de criar modelos triviais, que quase sempre predizem a classe majoritária.

Entretanto, em muitos dos problemas reais, tais como em aplicações de Mineração de Dados, uma grande desproporção no número de casos pertencentes a cada uma das classes é comum. Por exemplo, na detecção de fraudes em chamadas telefônicas [Fawcett and Provost, 1997] e transações realizadas com cartões de crédito [Stolfo et al., 1997], o número de transações legítimas é muito maior que o de transações fraudulentas. Na modelagem de risco de seguros [Pednault et al., 2000], apenas uma pequena porcentagem dos segurados reclama suas apólices em um dado período. Outros exemplos de domínios com um desbalancemento intrínseco entre as classes podem ser encontrados na literatura. Além disso, em muitas aplicações, não se sabe qual é a proporção exata de exemplos pertencentes a cada classe ou essa proporção pode variar no tempo.

Muitos sistemas de aprendizado, nessas circunstâncias, não estão preparados para induzir modelos de classificação que possam predizer acertadamente as classes minoritárias. Freqüentemente, esse modelo tem uma boa precisão na classificação da classe majoritária, mas a precisão para a classe minoritária não é aceitável. O problema é ainda maior quando o custo associado a uma classificação errônea para a classe minoritária é muito maior que o custo de uma classificação errônea para a classe majoritária. Infelizmente, esta é a norma e não a exceção para a maioria das aplicações com conjuntos de dados desbalanceados, uma vez que o objetivo dessas aplicações é obter um conjunto de elementos que se destacam de uma população de casos não interessantes.

Uma das maneiras de contrabalançar essa desproporção entre as classes é utilizar sistemas de aprendizado que levem em consideração os custos da classificação errônea, estabelecendo, nesse caso, um custo maior para a classificação incorreta dos exemplos das classes minoritárias. Entretanto, essa abordagem implica tanto na restrição dos sistemas que podem ser utilizados (restringindo, nesse caso, a sistemas que tenham a habilidade de incorporar previamente o custo de classificações errôneas) quanto em um conhecimento desses custos à *priori*.

Uma alternativa é balancear artificialmente o conjunto de dados utilizado na indução. Muitos métodos foram propostos na literatura para esse fim. Esses métodos ou retiram exemplos da classe majoritária ou acrescentam exemplos artificiais para a classe minoritária. Entretanto, esses métodos são, na maioria das vezes, avaliados utilizando medidas sensíveis à distribuição de classes. A utilização dessas medidas impede que sejam feitas inferências corretas a respeito da qualidade desses métodos. Em outras palavras, é difícil decidir se o ganho de desempenho obtido pela utilização desses métodos é real ou trata-se de uma influência provocada pelo vício da nova proporção de exemplos no conjunto de treinamento.

Neste trabalho, nos concentramos na utilização da análise de curvas ROC para a avaliação de métodos de balanceamento artificial de conjuntos de dados. Alguns resultados experimentais obtidos na aplicação desses métodos também são apresentados. Este trabalho está organizado da seguinte forma: na Seção 2 são apresentados alguns conceitos preliminares e alguns trabalhos correlacionados. Na Seção 3 é apresentada a análise de curvas ROC. Na Seção 4 é apresentado o estudo de caso. Finalmente, na Seção 5, são apresentadas algumas considerações finais.

2 Alguns conceitos preliminares

Aprendizado supervisionado é o processo de (semi) automaticamente induzir um modelo a partir de um conjunto de exemplos de treinamento, o qual consiste de um conjunto de n exemplos descrito por m atributos distintos $X_1, X_2, \cdots, X_m$. O conceito induzido está relacionado a um atributo específico Y, freqüentemente chamado de classe, que pode assumir qualquer um dos k possíveis valores no domínio desse atributo, chamado de rótulos ou classes. Uma vez criado, o modelo pode ser utilizado para predizer automaticamente outros exemplos não rotulados.

Neste trabalho, nos restringimos a problemas de aprendizado de conceito, de maneira que Y apenas pode assumir um dos $k = 2$ possíveis valores mutuamente exclusivos. Serão utilizados os rótulos genéricos positivo e negativo para discriminar entre esses dois possíveis valores para as classes. Neste trabalho, consideraremos a classe positiva como a minoritária.

Existem basicamente duas classes de métodos para o balanceamento da distribuição das classes:

1. ***Under-sampling*** que tem como objetivo balancear o conjunto de dados pela eliminação de exemplos da classe majoritária, e

2. ***Over-sampling*** que replicam exemplos da classe minoritária com o objetivo de obter uma distribuição mais balanceada.

Ambos *under-sampling* e *over-sampling* têm problemas correlacionados. *Under-sampling* pode eliminar exemplos potencialmente úteis nos dados e *over-sampling* pode aumentar a probabilidade de ocorrer um *overfitting*, uma vez que a maioria dos métodos de *over-sampling* fazem cópias exatas dos exemplos da classe minoritária. Nesse sentido, um modelo simbólico de classificação, por exemplo, poderia construir regras que são aparentemente precisas, mas, na verdade, cobrem exemplos replicados.

Algumas pesquisas recentes focam-se em tentar transpor esses problemas para essas duas classes de métodos. Chawla et al. [2002] combina os métodos de *under-sampling* e *over-sampling* e, ao invés de fazer *over-sampling* pela simples replicação de exemplos da classe minoritária, o faz pela interpolação de exemplos da classe minoritária que estão próximos. Dessa forma, o *overfitting* é contornado e as fronteiras de decisão para a classe minoritária são estendidas no espaço de exemplos da classe majoritária.

Kubat and Matwin [1997], Batista et al. [2000] utilizam um método de under-sampling para minimizar a quantidade de dados potencialmente úteis descartados. Os exemplos da classe majoritária são classificados como "seguros", "redundantes", "pertencentes à borda" e "ruido". Casos de borda e ruído são excluídos utilizando ligações Tomek [Tomek, 1976]. Uma ligação Tomek pode ser definida da seguinte forma:

> *Sejam E_i e E_j dois exemplos de classes diferentes. Seja d uma função de distância entre exemplos. Um par de exemplos (E_i, E_j) constitui uma ligação Tomek se não existe um exemplo E_k, tal que a distância $d(E_k, E_i) < d(E_i, E_j)$ ou $d(E_k, E_j) < d(E_i, E_j)$.*

Se dois exemplos (E_i, E_j) formam uma ligação Tomek, então, ou E_i e Ej são exemplos próximos à borda de decisão, ou um desses exemplos é ruído.

Exemplos redundantes também podem ser retirados identificando subconjuntos consistentes. Um subconjunto $S \subset E$ é consistente com E se utilizando o algoritmo do 1-vizinhos-mais-próximo classifica corretamente os casos em E [Hart, 1968].

3 Análise ROC

Para problemas de duas classes, a performance de um modelo de classificação pode ser medida utilizando a matriz de confusão, mostrada na Tabela 1. Uma das medidas de desempenho mais utilizadas para avaliar a qualidade dos modelos é a *precisão na classificação*, que é definida como $Acc = \frac{TP+TN}{TP+FP+FN+TN}$, ou, equivalentemente, a *taxa de erro na classificação*, definida como $Err = 1 - acc$.

	Predição positiva	*Predição Negativa*
Classe positiva	Verdadeiro positivo (TP)	Falso negativo (FN)
Classe negativa	Falso positivo (FP)	Verdadeiro negativo (TN)

Tabela 1: Matriz de confusão para problemas de duas classes.

Nessa tabela pode ser observado que a distribuição entre as classes (a proporção entre exemplos positivos e negativos) é o relacionamento entre a primeira e a segunda linha. Assim, qualquer medida de desempenho que utilize valores de ambas as colunas será, necessariamente, sensível a desproporção entre as classes. Métricas como precisão, taxa de erro, entre outras, utilizam valores de ambas as linhas da matriz de confusão. Havendo uma mudança na distribuição de classes, os valores dessas métricas também mudarão, mesmo que o desempenho global do modelo não melhore.

Um outro fato contra o uso da precisão (ou taxa de erro) é que essas métricas consideram os diferentes erros de classificação como igualmente importantes. Por exemplo, um paciente doente diagnosticado como saudável pode ser um erro fatal, enquanto que um paciente saudável classificado como doente pode ser considerado como um erro muito menos sério, uma vez que o erro pode ser corrigido em futuros exames. Em domínios nos quais o erro de classificação é relevante, uma matriz de custo pode ser utilizada. Uma matriz de custo define os custos de classificações errôneas, isto é, uma penalidade pela incursão de uma falha para cada possível erro do modelo. Nesse caso, o objetivo do modelo é minimizar os custos de classificação ao invés da taxa de erro.

Um método alternativo para a avaliação do desempenho desses algoritmos é a análise de curvas ROC[1], a qual representa o compromisso entre a sensitividade (a taxa de verdadeiros positivos $\frac{TP}{TP+FN}$) no eixo y e falso alarme (a taxa de falsos positivos $\frac{FP}{FP+TN}$) no eixo x. Curvas ROC são obtidas pela variação do limiar de um modelo probabilístico de classificação.

Um ponto em um gráfico ROC domina outro se esse ponto estiver estiver acima e a esquerda do outro, isto é, tem uma taxa mais alta de verdadeiros positivos e uma taxa mais baixa de falsos positivos. Se um ponto A domina um ponto B, A terá um menor custo esperado do que B para todos os possíveis custos e distribuições de classes. Um modelo de classificação K domina outro modelo W se todos os pontos da curva de W estão abaixo da curva K. Caso isso não ocora, o melhor modelo para uma faixa de valores pode ser derivado utilizando-se o método *convex hull* [Provost and Fawcett, 1997].

Em resumo, um gráfico ROC permite ao analista analisar a dominância de um modelo em relação aos demais. Por outro lado, a utilização do método *convex hull* permite escolher modelos potencialmente ótimos e também permite identificar faixas de dominância.

Neste trabalho, utilizamos o método proposto por Ferri et al. [2002], o qual estima a probabilidade pela escolha de diferentes rótulos para folhas em uma árvore de decisão. As folhas são reordenadas de acordo com a precisão do ramo correspondente, utilizando a correção de Laplace para obter estimativas mais confiáveis. Cada ponto no espaço ROC é plotado variando-se de todas as folhas rotuladas com a classe negativa para todas as folhas rotuladas com a classe positiva. A cada iteração, uma nova folha é re-rotulada, seguindo-se a ordem crescente de precisão. Esse método também pode ser utilizado para quaisquer outros sistemas de aprendizado que particionam o espaço de exemplos, tais como CN2 e a maioria dos sistemas ILP [Ferri et al., 2002].

[1]ROC é um acrônimo para *Receiver Operating Characteristic*, um termo utilizado na detecção de sinais que caracterizam o compromisso entre a taxa de êxitos e de alarmes falso em um canal com ruído.

A partir de um gráfico ROC, é possível calcular uma medida da qualidade global do modelo: a área abaixo da curva ROC (AUC[2]). A AUC é a fração da área total que está abaixo da curva ROC. Essa media é equivalente a várias outras medidas estatísticas para a avaliação e ranqueamento de modelos de classificação [Hand, 1997]. A AUC efetivamente fatora o desempenho de um modelo de classificação sobre todas as distribuições de custo. Entretanto, é importante notar que para matrizes de custo específicas, o modelo que maximiza a AUC pode não ser o melhor modelo.

4 Estudo de Caso

Os experimentos foram realizados utilizando o conjunto de dados Pima, disponível no repositório de dados da UCI [Merz and Murphy, 1998]. Esse conjunto de dados contém 768 exemplos divididos em duas classes. Os dados são utilizados para identificar os casos positivos de diabetes em uma população indígena perto de Phoenix, Arizona. O número de casos positivos é de somente 268. Uma qualidade desejável no modelo de classificação induzido é uma boa sensitividade na detecção da diabetes.

Com o objetivo de balancear artificialmente o conjunto de exemplos de treinamento, foram aplicados as duas classes (*under-sampling* and *over-sampling*) de métodos. Para uma avaliação mais realista das taxas de erro, foi utilizado 10-*fold cross-validation*. A proporção de exemplos nos conjuntos de teste não foram alteradas.

Dentre os métodos de *under-sampling* foi aplicado nos conjuntos de treinamento o algoritmo para a geração de subconjuntos consistentes, descrito em Batista et al. [2000], para reduzir o conjunto de dados para a proporção de 2 exemplo da classe minoritária para 3 exemplos da classe majoritária. Além disso, no subconjunto consistente resultante, foi aplicado o algoritmo para a remoção de ligações Tomek. O conjunto resultante têm uma distribuição aproximada de 1 exemplo da classe minoritária para 1 exemplo da classe majoritária. Também foi aplicado, no conjunto de dados original, um algoritmo que aleatoriamente retira exemplos da classe majoritária até que a proporção de exemplos fosse de 1 para 1.

Quanto aos métodos de *over-sampling* foi aplicado, nos conjuntos de treinamento, o algoritmo que gera exemplos novos exemplos pela interpolação de exemplos da classe minoritária próximos. Exemplos foram adicionados até que a proporção de exemplos fosse de 1 para 1. Também foi aplicado um algoritmo que aleatoriamente replica exemplos da classe minoritária até que a proporção dos exemplos fosse de 1 para 1.

Na Tabela 2 são apresentados os resultados da aplicação, no conjunto de dados original e nos conjuntos artificialmente balanceados, do algoritmo de indução de árvores de decisão C4.5 [Quinlan, 1993]. Nessa tabela estão apresentadas as taxas de erro global, taxas de erro da classe positiva (sensitividade), taxas de erro da classe negativa (especificidade), e a área abaixo da curva ROC (AUC). Também são apresentados, entre parênteses, os respectivos erros padrão para cada uma dessas medidas.

Conjunto de Dados	Taxa Erro Global (%)	Taxa Erro + (%)	Taxa Erro - (%)	AUC (%)
Original	26,30 (1,21)	40,57 (4,28)	18,60 (2,17)	88,42 (1,15)
Under. Sub. Consis.	27,37 (2,22)	31,32 (2,22)	25,20 (3,28)	88,92 (2,17)
Under. Sub. Consis. + Tomek	29,44 (1,15)	29,87 (1,78)	29,20 (1,67)	91,23 (1,58)
Under. Sub. Randômico	27,08 (1,54)	24,27 (1,62)	28,60 (2,65)	89,00 (1,33)
Over. Interpolação	24,62 (1,22)	32,85 (1,47)	20,20 (1,82)	90,53 (1,57)
Over. Randômico	26,26 (1,53)	31,72 (2,02)	23,40 (2,23)	76,82 (1,79)

Tabela 2: Resultados obtidos a partir do conjunto de dados Pima

De maneira geral, fazendo-se uma análise de variância dos resultados, nenhuma das taxas de erro global é significativamente melhor que outra (com um grau de confiança de 95%). O mesmo ocorre para os valores das taxas de erro positiva e negativa. Entretanto, para os valores da AUC, o valor correspondente ao conjunto de dados cujo método utilizado para balanceamento é um método de over-sampling com replicação randômica de exemplos, é estatisticamente menos eficiente que os outros. Uma análise somente das taxas de erro poderia levar a conclusão errônea de que a aplicação de *over-sampling* randômico é estatisticamente equivalente à aplicação dos outros métodos.

[2] *Area Under Curve*

Esses resultados podem ser melhor entendidos analisando os gráficos das curvas ROC para esse 6 conjuntos de dados, mostrados na Figura 1. A aplicação do método de conjunto consistente mais ligações Tomek praticamente é dominante em relação as outras curvas. Apesar dessa dominância não ser estatisticamente significativa, a instanciação do problema para uma dada distribuição de classes ou custos de classificação pode levar à escolha de um modelo estatisticamente mais significativo que o conjunto de dados original. Por outro lado, a curva correspondente ao *over-sampling* randômico é dominada por todas as outras curvas. Nesse caso, esse método pode ser descartado em futuras análises.

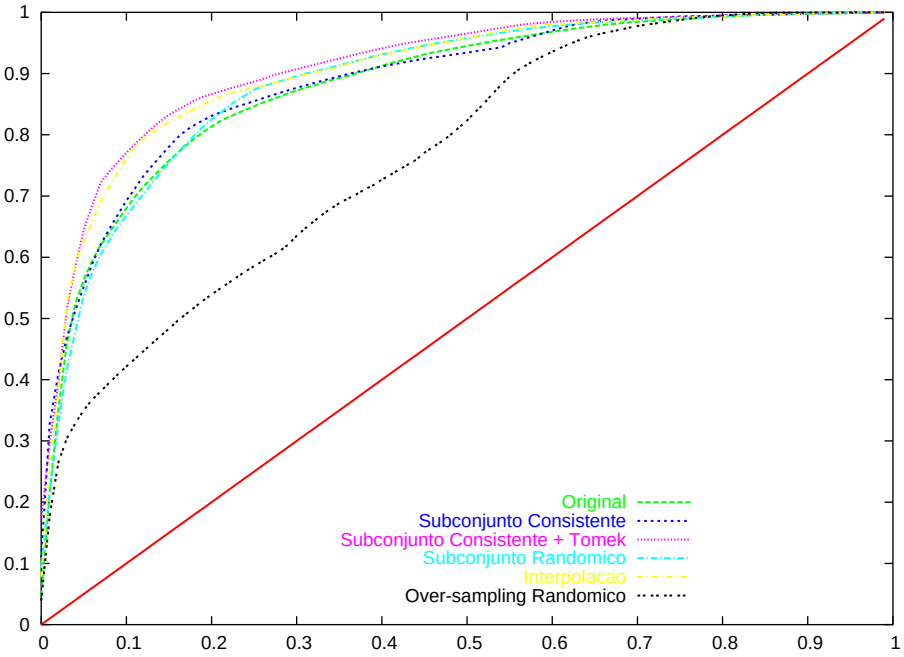

Figura 1: Curvas ROC obtidas a partir do conjunto de dados Pima

Um outro ponto interessante, mas pouco abordado nos trabalhos de balanceamento artificial de conjuntos de dados, refere-se ao tamanho dos modelos induzidos. Um resumo das características dos modelos é mostrado na Tabela 3, na qual os números entre parênteses indicam os erros padrões. De maneira geral, os modelos que foram induzidos a partir dos conjuntos de dados utilizando *under-sampling* apresentam características semelhantes, no que se refere ao número de ramos e altura (número de condições por ramo) da árvore. Nos modelos induzidos a partir dos conjuntos de dados utilizando *over-sampling*, houve um aumento significativo tanto no número de ramos quanto na altura da árvore.

5 Considerações Finais

O aprendizado com classes desbalanceadas é uma características importante em aplicações práticas de algoritmos de aprendizado. Um método direto para contornar o problema de desbalanceamento dos dados é de artificialmente balancear o conjunto de dados. Esse balanceamento pode tanto ser feito retirando exemplos da classe majoritária, replicando exemplos da classe minoritária, ou ambos.

A análise de curvas ROC é um importante mecanismo com o qual é possível avaliar o desempenho de métodos de balanceamento artificial de conjuntos de dados. Neste trabalho, foram aplicados vários métodos de balanceamento

Conjunto de Dados	Número de Ramos	Altura Média
Original	26,40 (3,65)	5,59 (0,34)
Under. Sub. Consis.	28,00 (4,05)	5,80 (0,34)
Under. Sub. Consis. + Tomek	26,60 (2,43)	5,76 (0,21)
Under. Sub. Randômico	21,80 (1,55)	5,56 (0,17)
Over. Interpolação	43,40 (1,60)	6.95 (0,17)
Over. Randômico	68,60 (4,11)	7.06 (0,11)

Tabela 3: Características dos modelos

artificial para o conjunto de dados Pima. Esses métodos aplicam seleção aleatória e heurísticas para obter uma proporção equilibrada entre as classes. Também foi feita uma análise do tamanho do modelo induzido, por meio da análise da altura e largura da árvore de decisão do modelo induzido.

Como trabalhos futuros, pretende-se aplicar os métodos aqui apresentados para outros conjuntos de dados, a fim de verificar se os resultados obtidos podem ser generalizados. Uma outra extensão natural deste trabalho é verificar, para custos específicos, se os métodos apresentados melhoram o desempenho do modelo induzido.

Referências

G. E. A. P. A. Batista, A. Carvalho, and M. C. Monard. Applying One-sided Selection to Unbalanced Datasets. In O. Cairo, L. E. Sucar, and F. J. Cantu, editors, *Proceedings of the Mexican International Conference on Artificial Intelligence – MICAI 2000*, pages 315–325. Springer-Verlag, April 2000.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Jounal of Artificial Intelligence Research*, 16:321–357, 2002.

T. Fawcett and F. J. Provost. Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

C. Ferri, P. Flach, and J. Hernández-Orallo. Learning decision trees using the area under the ROC curve. In C. S. A. Hoffman, editor, *Nineteenth International Conference on Machine Learning (ICML-2002)*, pages 139–146. Morgan Kaufmann Publishers, 2002.

D. J. Hand. *Construction and Assessment of Classification Rules*. John Wiley and Sons, 1997.

P. E. Hart. The Condensed Nearest Neighbor Rule. *IEEE Transactions on Information Theory*, IT-14:515–516, 1968.

M. Kubat and S. Matwin. Addressing the Course of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of 14th International Conference in Machine Learning*, pages 179–186, San Francisco, CA, 1997. Morgan Kaufmann.

C. J. Merz and P. M. Murphy. UCI Repository of Machine Learning Datasets, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

E. P. D. Pednault, B. K. Rosen, and C. Apte. Handling Imbalanced Data Sets in Insurance Risk Modeling. Technical Report RC-21731, IBM Research Report, March 2000.

F. J. Provost and T. Fawcett. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Knowledge Discovery and Data Mining*, pages 43–48, 1997.

J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

S. J. Stolfo, D. W. Fan, W. Lee, A. L. Prodromidis, and P. K. Chan. Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results. In *AAAI-97 Workshop on AI Methods in Fraud and Risk Management*, 1997.

I. Tomek. An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Communications*, SMC-6(6):448–452, June 1976.