Aplicação de Técnicas de Balanceamento Baseadas em Classificadores de Larga Margem no Problema de Multiclassificação

Matheus Franklin Rodrigues Silva (mfranklin@ice.ufjf.br)
Carlos Cristiano Hasenclever Borges (cchborges@ice.ufjf.br)
Saulo Moraes Villela (saulo.moraes@ufjf.edu.br)
Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora

Introdução

Classificadores de larga margem, como as Máquinas de Vetores Suporte (*Support Vector Machines* – SVMs), têm sido usados com muita eficácia em problemas de classificação binária. Porém, quando aplicados em problemas de multiclassificação, os resultados não se mostram muito satisfatórios ou apresentam um alto custo computacional.

Uma das estratégias mais utilizadas com esse propósito é a um contra todos, que acaba por gerar um desbalanceamento de uma base balanceada. É avaliada nesse trabalho uma possibilidade de explorar a utilização de técnicas de balanceamento para contornar esse problema.

Multiclassificação

O problema de multiclassificação consiste em, dado um conjunto de treinamento dividido em mais de 2 classes, gerar um classificador capaz de prever a qual classe pertence uma nova amostra dada.

Uma abordagem muito comum é a utilização de classificadores de larga margem, fazendo assim uma adaptação do problema de multiclassificação para subproblemas de classificação binária. As duas estratégias mais utilizadas nesse método são a um-contra-todos (*one-against-all*) e a um-contra-um (*one-against-one*).

A abordagem um-contra-todos realiza a solução de um problema de classificação binária de cada classe do problema contra todas as demais, gerando n hiperplanos solução, onde n representa o número de classes do problema. Para a determinação da classe de uma nova amostra, utiliza-se uma função de decisão que pondera cada uma das soluções.

Já a um-contra-um realiza a solução de um problema de classificação binária de cada classe contra cada uma das outras . Essa estratégia tende a ter resultados melhores que a primeira, porém gera $n \times (n - 1) / 2$ hiperplanos solução, o que demanda um custo computacional elevado.

Um dos principais motivos pelo baixo desempenho da primeira abordagem (um contra todos) é devido ao desbalanceamento gerado quando se faz a classificação binária de uma classe contra as demais. Esse desbalanceamento é caracterizado pela presença de poucas amostras da classe em questão, enquanto a junção das demais passa a possuir muitos exemplos. Para contornar esse problema de desbalanceamento entre as classes em um problema de classificação binária, diferentes abordagens já foram propostas [1].

Técnicas de Balanceamento

Uma das técnicas mais populares para contornar (ou reduzir) o problema de classes desbalanceadas baseia-se na modificação do conjunto de dados originais. Tais técnicas são conhecidas como reamostragem de dados.

Essa reamostragem pode ser realizada basicamente por dois métodos: *undersampling* (redução das amostras da classe majoritária) ou *oversampling* (aumento das instâncias da classe minoritária).

Nesse trabalho são utilizadas técnicas de *oversampling*, mais especificamente os métodos *Synthetic Minority Over-Sampling Technique* (SMOTE) e Algoritmo de Balanceamento Sintético Incremental (*Incremental Synthetic Balancing Algorithm* – ISBA) [2].

O funcionamento do SMOTE consiste em, para cada amostra da classe minoritária, escolher aleatoriamente um dos k vizinhos mais próximos e gerar uma nova instância baseando-se nos atributos dos vizinhos escolhidos.

Já no ISBA, que foi baseado no SMOTE, a geração de novas amostras artificiais utiliza os vetores suportes obtidos por um classificador de larga margem, sendo esse um processo incremental, pois, a cada passo, são escolhidos novos vetores suportes. Em seu processo construtivo, o método permite a possibilidade de extrapolação da amostra sintética em relação às amostras de referência para geração, além do descarte de exemplos artificiais considerados potencialmente nocivos ao processo de aprendizagem.

Abordagem proposta

Esse projeto será desenvolvido através da aplicação de técnicas de geração de amostras artificiais para balancear problemas de classificação binária com intuito de melhorar o poder (acurácia) de um classificador de larga margem quando aplicado a problemas de multiclassificação, além de um estudo sobre a parametrização dos algoritmos e a função de ponderação na abordagem um contra todos.

Considerações finais

Experimentos realizados indicam que o método ISBA apresenta uma melhora no desempenho de classificadores de larga margem em bases desbalanceadas. Tem-se a expectativa que este desempenho também seja obtido na aplicação do método em problemas de classificação multiclasses.

Referências bibliográficas

[1] Hsu, C. W.; Lin, C. J. **A comparison of methods for multiclass support vector machines**. Trans. Neur. Netw. 13 (2), 415-425, 2002.

[2] Marques, M. L.; Villela, S. M.; Borges, C. C. H. **Uma estratégia de geração de dados artificiais para classificadores de larga margem aplicada em bases de dados desbalanceadas**. IV Symposium on Knowledge Discovery, Mining and Learning, 2016.