Desbalanceamento, AUC e SMOTE

ENG121 - Reconhecimento de Padrões

Autor: Matheus Castro Silva

RESUMO

O presente trabalho tem como objetivo testar, o modelo de Máquinas de Vetor de Suporte (SVM), para a base de dados *BaseCar* é originalmente disponível no repositório UCI, e foi modificada, para que tenha-se um problema de desbalanceamento, com 3.9% de observações da classe positiva. Para isto, deve-se utilizar técnicas para encontrar os melhores parâmetros da SVM, considerando o desbalanceamento. As técnicas utilizada foram; Acurácia, AUC e superamostragem por SMOTE.

Palavras Chaves **- algoritmos, SVM classificação, reconhecimento de padrões, SMOTE, AUC, Acurácia.**

I. INTRODUÇÃO

O presente trabalho busca introduzir o teste do classificador de Máquinas de Vetor de Suporte (SVM), para uma base de dados *BaseCar,* é originalmente disponível no repositório UCI, e foi modificada, para que tenha-se um problema de desbalanceamento, com 3.9% de observações da classe positiva. A base deve ser lida, a primeira coluna (ID da observação) deve ser retirada e o data frame deve ser pré-processado de acordo com critérios do aluno. Realizar uma comparação entre o modelo de SVM utilizando validação cruzada para identificação dos parâmetros da SVM utilizando os critérios de acurácia e AUC como referência. Também deve-se realizar uma reamostragem utilizando a técnica SMOTE e repetir o processo de validação cruzada para identificação dos parâmetros da SVM utilizando os critérios de acurácia e AUC como referência

II. DEFINIÇÃO DO PROBLEMA

Utilizando um conjunto de base de dados *BaseCar,* é originalmente disponível no repositório UCI, e foi modificada, para que tenha-se um problema de desbalanceamento, com 3.9% de observações da classe positiva. Para isto, deve-se utilizar técnicas para encontrar os melhores parâmetros da SVM, considerando o desbalanceamento. As técnicas utilizada foram; Acurácia, AUC e superamostragem por SMOTE. E por fim, realizar uma comparação dos resultados obtidos entre AUC e Acurácia.

I. MODELO DE SVM

Uma máquina de vetores de suporte (SVM, do inglês: support vector machine) é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Dados um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra. Um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados.

Em outras palavras, o que uma SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classe, ver imagem:

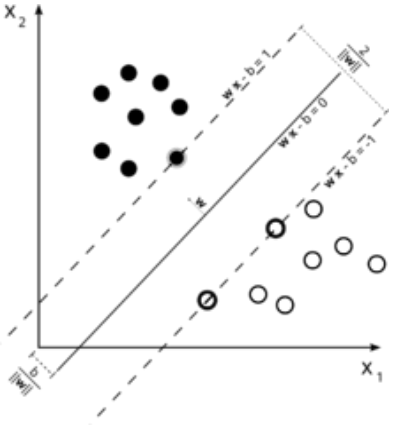

Imagem 1: Ilustração do vetor de suporte entre as duas classes no plano x1 e x2.

III. TREINAMENTO E TESTE

A base de dados foi dividida em dois conjuntos, um de treinamento e outro para os testes, sendo 70% para treinamento e 30% para teste.

IV. KFOLD

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Busca-se então estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.

1. K-FOLD

O método de validação cruzada denominado k-fold consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. A figura abaixo mostra o esquema realizado pelo k-fold.

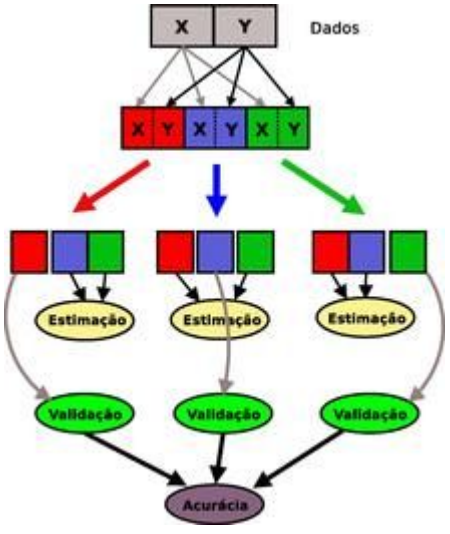

Imagem 2: Exemplo do esquema de particionamento e execução do método k-fold com k = 3.

Ao final das k iterações calcula-se a acurácia sobre os erros encontrados, através da equação descrita anteriormente, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

V. RESULTADOS DE SEPARAÇÃO

Utilizando o conjunto de amostras referente às duas classes abaixo, foi feito a superfície de separação utilizando KNN.

VI. COMPARAÇÃO DOS ERROS

a. KNN

Utilizando também o passo a passo do tópico de implementação III neste trabalho os resultados obtidos, considerando a variação das iterações foram:

	K-fold		**K-fold + Smote**	
Iterações	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC**
k = 1	98.84	99.04	100	100
k = 2	98.65	99.42	100	100
k = 3	99.42	99.42	100	100
k = 4	98.27	98.27	100	100
k = 5	99.23	99.61	100	100

O percentual de acerto perante os testes ficou entre 98% à 100%, considerando a técnica SVM utilizando a validação cruzada por K-fold, um método bastante eficiente porém que ainda tem pontos a serem levados em consideração. Pois quando realizamos uma reamostragem utilizando a técnica de SMOTE, pelo fato dos dados serem desbalanceados a melhora na porcentagem de acertos é bastante visível, mostrando que a técnica é bastante eficiente.

VII. CONCLUSÕES

O método de Máquina de vetores de suporte (SVM) utilizando todas as características é bastante satisfatório, e quase sempre se mostra eficiente na separação das classes. Mas quando utiliza-se o SVM combinado com a técnica de reamostragem (SMOTE), o algoritmo se mostra mais robusto e tende a obter resultados razoavelmente melhores que somente o SVM puro, considerando os dados desbalanceados.

O desempenho do modelo de SVM depende de dois parâmetros, cost e gamma e quanto melhor ajustados forem estes parâmetros, melhor será o resultado obtido do modelo. Como foi utilizado o método de *k-folds* como validação cruzada o algoritmo se mostrou bem mais eficiente, porém é válido ressaltar que o custo computacional quando utiliza-se validação cruzada é alto.

VIII. REFERÊNCIAS

[1] R tutorial, http://www.r-tutor.com/r-introduction, acessado 30/08/2017.

[2] Antônio de Pádua Braga - Notas de Aulas de Redes Neurais Artificiais e de Reconhecimento de Padrões, acessado 22/09/2017.

[3] SVM - Máquina de vetores de suporte https://pt.wikipedia.org/wiki/M%C3%A1quina_de_vetores_de_suporte acessado 15/11/2017.