Classificação de estrelas de nêutrons

Bruna S. Queiroz
Estudante de Engenharia de Sitemas
Universidade Federal de Minas Gerais
Matrícula: 2016108554
Email: brunasq@ufmg.br

Hernane Braga
Estudante de Engenharia de Sitemas
Universidade Federal de Minas Gerais
Matrícula: 2014112627
Email: hernane137@ufmg.br

***Resumo*—Este relatório apresenta a aplicação do método SMOTE [1] em conjunto com o classificador SVM [2] , para identificar, dentro de uma base de dados desbalanceada, quais sinais são de uma estrela de nêutrons. O presente trabalho foi inspirado no artigo [3], onde este problema foi abordado utilizando Redes Neurais Artificiais.**

I. Problema a ser resolvido

Estrela de nêutrons, é o núcleo colapsado de uma grande estrela que, antes do colapso, teria tido um total de entre 10 e 29 massas solares [4]. Elas são as menores e mais densas estrelas que se tem conhecimento, pois tipicamente possuem um raio na ordem de 10 quilômetro e possuem uma massa que é cerca de duas vezes a do Sol [4]. Seu estudo é importante para os astrônomos, pois elas são o resultado de uma explosão da supernova de uma estrela massiva, que combinada com o colapso gravitacional, comprime seu núcleo. Caso o remanescente desta explosão possua uma massa maior do que 2,2 massas solares, a estrela de nêutrons continua a colapsar para formar um buraco negro [5].

O artigo de [3] descreve que o processo para analisar os sinais candidatos à um pulso estelar e identificar se este é ou não uma estrela de nêutrons, que é feito inteiramente por humanos, via análises gráficas. Entretanto, como a quantidade de amostras destes pulsos aumentou drasticamente devido às pesquisas utilizando raios em larga escala como o *Parkes multibeam pulsar survey* (PMS) [6], se tornou inviável a conferência de cada um dos pulsos candidatos e passou-se a adotar métodos computacionais para realizar a triagem de soluções candidatas. De acordo com (Eatough. 2010) [3], da inspeção de 40.000 raios PMS, são gerados 8 milhões de pulsos candidatos, de onde após métodos computacionais serem aplicados, apenas 200 pulsos candidatos são avaliados.

A partir dos dados obtidos da pesquisa *Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach* [7] da Universidade de Manchester, um conjunto de dados foi disponibilizado no *UC Irvine Machine Learning Repository* (UCI) [8] e este foi usado como base para o presente trabalho. Esta base de dados possui um total de 16.259 amostras, onde apenas 1.639 (10%) são pulsos de estrelas de nêutrons e cada pulso possui 8 variáveis que descrevem suas características. Para resolver este problema de classificação foram utilizadas as técnicas de desbalanceamento, como o SMOTE [1] e SVM [2], como é demonstrado nas seções seguintes.

II. Revisão bibliográfica

Neste artigo, trabalhamos com sinais usados para descrever uma estrela de nêutron, que podem ser classificados por métodos de classificação de dados. Geralmente, dados capturados para um experimento ou uma análise (como os sinais de pulsos estrelares) apresentam grandes diferenças de proporções entre os dados totais e os dados que realmente descrevem o fenômeno por serem considerados ocorrências raras. Assim, tratamos esses dados como dados *desbalanceados*. Classificar base de dados como essas podem levar a resultados e acurácias que não descrevem bem o modelo sendo necessário aplicar métodos que ajustam a distribuição dos dados. Uma dessas técnicas é chamada de SMOTE (*Synthetic Minority Oversampling Technique*- Técnica de sobreamostragem minoritária sintética) [1] que é um método que sobreamostra exemplos artificias por interpolação. Para criar dados sintéticos, em cada iteração o método utiliza de uma amostra de dados da classe minoritária e os k vizinhos desse ponto. Com os k forma-se um vetor que será multiplicado por um número aleatório que está no intervalo entre 0 e 1. Esse dado multiplicado é adicionado na amostra e assim temos um novo ponto sintético. Além disso, em algumas implementações desse método, ele diminui a classe de dados dominante retirando aleatoriamente algumas amostras a fim de chegar em uma proporção de 50%.

Para a classificação dos dados utilizamos o SVM (*Support vector machine* - Máquina de vetores de suporte) [2] que é um método supervisionado que analisa os dados para classificação binária ou análise de regressão. Esse método consiste na definição de um hiperplano de separação das classes e de amostras próximas a esse plano que são chamados de vetores de suporte. Assim, o SVM define o melhor separador das classes através da resolução de um problema de otimização não linear em que a função objetivo é maximizar a distância dos vetores de suporte ao hiperplano (margem) [9]. O propósito de utilizar esse método é investigar seu comportamento em base de dados desbalanceados e comparar quando aplica-se na base de dados o método SMOTE antes do SVM. Pretendemos avaliar se quando possuímos poucas amostras de uma classe o SVM consegue estabelecer um bom hiperplano de separação dos dados ou se ele considera a menor classe como ruído.

Na análise de métodos de classificação em dados desbalanceados, para avaliar o acerto de um classificador pelo cálculo

da acurácia pode não refletir bem os resultados positivos do método. Isso ocorre porque como uma classe possui muitos dados em comparação com a classe considerada "rara", a proporção de acertos das amostras da classe comum podem ser bem maiores que os acertos da classe rara e, assim, aumenta-se a média da acurácia das duas classes. Para resolver esse problema e conseguirmos avaliar bem os métodos aplicados, utilizamos a avaliação dos resultados pela curva ROC (*Receiver Operating Characteristic* - Caracterísitcas de Operação do Receptor) [10] e pelo AUC (*Area Under The Curve* - Área sob a curva). A ROC é uma representação gráfica que ilustra o desempenho de um classificador binário e o AUC é a área abaixo dessa curva que pode ter valores entre 0 e 1. Ela é a representação gráfica da especificidade de uma classificação pela sua sensibilidade. A sensibilidade pode ser definida como a quantidade de verdadeiros positivos dividido pela soma de verdadeiros positivos com falsos negativos; já a especificidade é a quantidade de verdadeiros negativos divido pela soma de verdadeiros negativos com falsos positivos. Assim, cada valor entre 0 e 1 desse limite gera um ponto falso positivo, verdadeiro positivo que define a curva ROC. Sabemos que a classificação foi boa se a curva aumenta rapidamente de 0 para 1 e o AUC for próximo de 1 [11].

III. Descrição dos dados

A base de dados utilizada para esse artigo é um conjunto de pulsos estrelares coletados durante o *High Time Resolution Universe Survey (South)* [12]. Os radiotelescópios capturam pulsos de rádio periodicamente para avaliar se esses pulsos são candidatos, ou seja, se descrevem estrelas de nêutrons. Em contrapartida, a maioria dos dados coletados são interferência de radiofrequência (**RFI**) e ruído, fazendo com que os dados que realmente representam pulsos estrelares estejam em minoria. A base de dados é composta por 17898 amostras, com 16259 representações de RFI e ruído e apenas 1639 dados são classificados como pulsos reais. As amostras possuem 8 características (variáveis contínuas), com as quatro primeiras sendo estatísticas simples obtidas a partir do perfil de pulso integrado (calculadas a partir da longitude do sinal em relação ao tempo e frequência). As últimas quatro características foram obtidas a partir da curva **DM-SNR** (medida de dispersão a partir do gráfico de período do pulso baricêntrico pela taxa de sinal-ruído) [13]. As características são:

- Média do perfil integrado;
- Desvio padrão do perfil integrado;
- Excesso de curtose do perfil integrado;
- Inclinação do perfil integrado;
- Média da curva DM-SNR;
- Desvio padrão da curva DM-SNR;
- Excesso de curtose da curva DM-SNR;
- Assimetria da curva DM-SNR;

IV. Experimentos

Para classificar se um sinal é de um pulso estelar, ou ruído, foram realizadas dois experimentos: o primeiro usando as classes desbalanceadas e o segundo aplicando a técnica SMOTE para balanceamento dos dados.

No primeiro experimento, os dados da tabela I foram separados em 10 folds, onde nove folds foram usados para treinamento e um para teste. Os dados foram classificados utilizando o método SVM, onde foi usada a implementação do pacote *kernlab*, com parâmetro de erro *c=0.5* e *kpar=2*.

Tabela I
Amostras do experimento 1

	nº de amostras	%
Ruído	16.258	90.84%
Pulso estelar	1.639	9.16%
Total	17.897	100%

Para o segundo experimento, os dados foram balanceados utilizando a técnica SMOTE [1] e os dados gerados como resultado podem ser vistos na tabela II. À partir da nova base de dados, repetiu-se os passos do experimento 1 e foi realizada a classificação usando o método SVM com validação cruzada entre os folds.

Tabela II
Amostras do experimento 2

	nº de amostras	%
Ruído	3.278	50%
Pulso estelar	3.278	50%
Total	6.556	100%

Para comparar os resultados, foi utilizado o valor da área abaixo da curva ROC (AUC). Esta métrica foi escolhida em detrimento da acurácia, pois como explicado na seção II, para bases desbalanceadas a acurácia de um modelo pode não refletir sua verdadeira classificação, pois avalia os acertos de forma desproporcional. Os resultados dos experimentos são apresentados na seção seguinte.

V. Resultados

O valor da AUC para cada um dos folds do experimento 1, pode ser visto na tabela III, enquanto os valores do experimento 2 são vistos na tabela IV. Percebe-se que os resultados do experimento 2, que utiliza a técnica SMOTE para balancear os dados, teve um melhor desempenho. Esta conclusão também pode ser observada na figura 1, onde a curva ROC do melhor fold de cada um dos experimentos são expressas no mesmo gráfico.

VI. Conclusão

Após os experimentos realizados, percebe-se melhora no valor da área abaixo da curva ROC (AUC) [10] no experimento 2, onde houve balanceamento do número de amostras de cada classe, durante a etapa de pré-processamento dos dados. Com base neste resultado, conclui-se que é importante que os dados não estejam desbalanceados ao se usar um classificador baseado na distância entre as amostras de cada classe,

Tabela III
COMPARAÇÃO DO VALOR DE AUC PARA CADA UM DOS FOLDS USANDO DADOS DESBALANCEADOS

Fold	AUC
1	0.884
2	0.899
3	0.894
4	0.885
5	0.870
6	0.885
7	0.907
8	0.904
9	0.877
10	0.905
Média	0.891

Tabela IV
COMPARAÇÃO DO VALOR DE AUC PARA CADA UM DOS FOLDS USANDO DADOS BALANCEADOS GERADOS PELA TÉCNICA SMOTE

Fold	AUC
1	0.938
2	0.937
3	0.959
4	0.946
5	0.955
6	0.959
7	0.934
8	0.939
9	0.933
10	0.945
Média	0.945

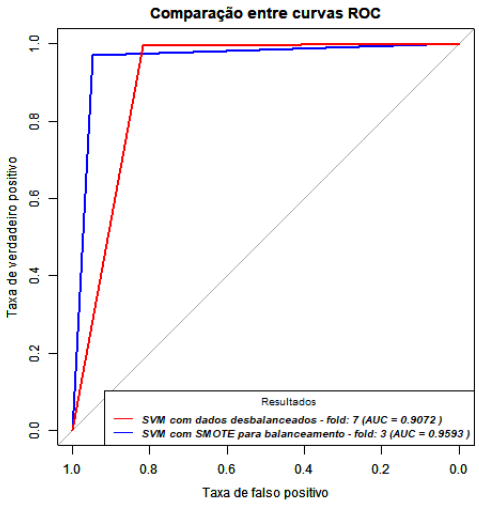

Figura 1. Comparação da curva ROC entre a classificação com dados desbalanceados e balanceados

como é o caso do método SVM [2], utilizado neste trabalho. Conclui-se também que a técnica SMOTE [1] foi efetiva no balanceamento de classes, apesar de reduzir o número total de amostras para treinamento.

REFERÊNCIAS

[1] L. O. H. N. V. Chawla, K. W. Bowyer and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[2] C. C. an Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[3] R. P. Eatough, N. Molkenthin, M. Kramer, A. Noutsos, M. J. Keith, B. W. Stappers, and A. G. Lyne, "Selection of radio pulsar candidates using artificial neural networks," *Monthly Notices of the Royal Astronomical Society*, vol. 407, no. 4, p. 2443–2450, Jul 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2010.17082.x

[4] N. K. Glendenning, *Compact stars: Nuclear physics, particle physics, and general relativity*, 1997.

[5] A. Cho, "A weight limit emerges for neutron stars," *Science (New York, N.Y.)*, vol. 359, no. 6377, p. 724—725, February 2018. [Online]. Available: https://doi.org/10.1126/science.359.6377.724

[6] R. Manchester, A. Lyne, F. Camilo, J. Bell, V. Kaspi, N. D'Amico, N. McKay, F. Crawford, I. Stairs, A. Possenti, and et al., "The parkes multi-beam pulsar survey - i. observing and data analysis systems, discovery and timing of 100 pulsars," *Monthly Notices of the Royal Astronomical Society*, vol. 328, no. 1, p. 17–35, Nov 2001. [Online]. Available: http://dx.doi.org/10.1046/j.1365-8711.2001.04751.x

[7] R. Lyon, B. Stappers, S. Cooper, J. Brooke, and J. Knowles, "Fifty years of pulsar candidate selection: From simple filters to a new principled real-time classification approach," *Royal Astronomical Society. Monthly Notices*, vol. 459, no. 1, pp. 1104–1123, 6 2016.

[8] R. Lyon, "HTRU2," 3 2016. [Online]. Available: https://figshare.com/articles/HTRU2/3080389

[9] A. de Pádua Braga, "Princípios de redes neurais artificiais e de reconhecimento de padrões," *Notas de aula*, 2017.

[10] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[11] B. Rocca, "Handling imbalanced datasets in machine learning," accessed: 2019-11-05. [Online]. Available: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

[12] M. J. Keith, A. Jameson, W. Van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. D. Bates, N. D. R. Bhat, M. Burgay, and et al., "The high time resolution universe pulsar survey - i. system configuration and initial discoveries," *Monthly Notices of the Royal Astronomical Society*, vol. 409, no. 2, p. 619–627, Sep 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2010.17325.x

[13] R. J. Lyon, "Why are pulsars hard to find?" *PhD Thesis, University of Manchester*, 2015.