UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2021

GERAÇÃO GENÉTICA DE CLASSIFICADORES FUZZY PARA BASES DE DADOS DESBALANCEADAS

Allan Pereira da Silva[1] e Matheus Giovanni Pires[2]

1. Bolsista PIBIC/FAPESB, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: allanpereira016@gmail.com

2. Orientador, Departamento de nome, Universidade Estadual de Feira de Santana, e-mail: mgpires@ecomp.uefs.br

PALAVRAS-CHAVE: Aprendizado genético de sistemas fuzzy, algoritmos evolutivos multiobjetivo, dados desbalanceados.

INTRODUÇÃO

Eventos raros, padrões não usuais e comportamentos anormais são difíceis de serem detectados e frequentemente exigem respostas em tempo hábil (HAIXIANG et al, 2017). Eventos raros se referem aos eventos que acontecem com uma frequência muito menor em relação aos eventos comuns (MAALOUF; TRAFALIS, 2011). Exemplos de eventos raros são detecção de defeitos em software (RODRIGUEZ et al, 2014), desastres naturais (HAIXIANG et al, 2017), detecção de fraudes em transações financeiras (PANIGRAHI, S. et al, 2009), dentre outros.

Na área de Mineração de Dados, a detecção de eventos é um problema de predição, ou tipicamente, um problema de classificação. A classificação de eventos raros é uma tarefa difícil devido à baixa frequência e casualidade dos dados (HAIXIANG et al, 2017), resultando em bases de dados desbalanceadas. Estas bases possuem muitos exemplos (ou instâncias) de uma classe, chamada de classe majoritária, e poucos exemplos de outra classe, chamada de classe minoritária (YIJING, 2016).

Os Sistemas Baseados em Regras Fuzzy (SBRF) são amplamente utilizados em várias aplicações. Para os problemas de classificação, são aplicados os Sistemas Classificadores Baseados em Regras Fuzzy (SCBRF) (ALCALÁ-FDEZ, 2009). O grande sucesso no uso de SBRF é sua capacidade de modelar o modo aproximado de raciocínio, permitindo o desenvolvimento de sistemas que imitem a habilidade humana de tomar decisões racionais em um ambiente de incerteza e imprecisão (MENDEL, 1995).

Uma das principais lacunas que envolvem a classificação de dados desbalanceados é a necessidade do uso de algoritmos de pré-processamento, como técnicas para balanceamento dos dados, ou abordagens sensíveis ao custo. Esses algoritmos podem tornar uma classificação que antes era computacionalmente custosa, cada vez mais custosa, dependendo dos algoritmos utilizados.

Sabendo dessa lacuna na classificação de bases de dados desbalanceadas, o objetivo desse projeto é utilizar algoritmos genéticos para otimizar a base de dados e a bases de regras de um SCBRF ao mesmo tempo, e assim, classificar os dados desbalanceados de forma que os resultados se aproximem da abordagem com pré-processamento, mesmo sem usar esses algoritmos. Assim, gerando um sistema com menor custo computacional que obtenha uma resposta em tempo hábil e com um resultado próximo ou até mesmo melhor para algumas bases de dados.

METODOLOGIA

Para a definição dos conjuntos Fuzzy que compõem a base de conhecimento, o universo de discurso de cada atributo foi dividido em três conjuntos Fuzzy, distribuídos de maneira uniforme, onde os conjuntos das extremidades possuem funções de pertinências trapezoidais e o conjunto central triangular. O Algoritmo Genético (AG) utilizado foi o *Non-Dominated Sorting Genetic* (NSGA-II), que é um Algoritmo Evolutivo Multiobjetivo (AEMO) encontrado na biblioteca JMetal desenvolvida na linguagem de programação Java. A aprendizagem da base de conhecimento do SCBRF foi feita utilizando a sinergia da Base de Dados (BD) com a Base de Regras (BR), ou seja, o cromossomo do AG foi codificado de forma a comportar as regras, as classes determinadas por cada regra e os pontos centrais da BD.

A população inicial é gerada de maneira aleatória com exceção de um único indivíduo, denominado de semente, sua necessidade surgiu para otimizar o tempo de convergência do algoritmo e minimizar os recursos computacionais necessários, como: alocação de memória e processamento. A semente é um modelo padrão utilizado para geração dos demais indivíduos da população, a partir da mesma é possível definir um valor máximo para a quantidade de regras e ajustar o espaço vetorial. O algoritmo de Wang-Mendel foi utilizado para gerar o primeiro SCBRF.

Para fazer o cálculo da acurácia do SCBRF, primeiro foi montada uma matriz de confusão, visto que os 22 *datasets* do site Keel Datasets com um índice de desbalanceamento superior a 9, têm apenas duas possíveis classes para a classificação de uma instância sendo 0 (negativa) e 1 (positiva). Após a matriz de confusão ser formada, é calculada a Área Sobre a Curva ROC, do inglês, *Area Under the Curve* (AUC) que é uma medida utilizada para medir corretamente a acurácia de sistemas de classificação em bases desbalanceadas. Outro objetivo do AG é a interpretabilidade, calculada como a quantidade de antecedentes das regras fuzzy ignorando as condições que não importam.

O algoritmo foi parametrizado com 100 indivíduos de população, uma taxa de cruzamento de 95% e 5% de mutação. O critério de parada definido para o AG foi de 30 gerações sem ocorrer aprendizagem. A seleção foi feita através de elitismo.

Antes de finalizar a execução dos 22 *datasets* desbalanceados, percebeu-se que essa configuração não era a mais adequada para obter os melhores resultados se comparado com a literatura. Logo, foram feitas variações nos parâmetros para encontrar a melhor configuração para o AG, assim, foi determinado uma população de 100 indivíduos, 75% de cruzamento, 15% de mutação e o método de seleção anterior foi mantida. Ainda assim, devido ao alto desbalanceamento dos dados e a abordagem de não utilizar métodos de pré-processamento para balanceá-los, os resultados ainda não foram satisfatórios pois alguns *folds*, de alguns *datasets*, estavam com um desempenho muito baixo, principalmente em termos de acurácia. Isso foi identificado como uma anomalia porque esses *folds* se distanciavam muito da média dos outros *folds* que se mostravam mais condizentes com a realidade da classificação. Sabendo disso, foi necessário encontrar a melhor configuração para cada *fold* especificamente. Vale lembrar que cada um dos 5 *folds* foram executados 5 vezes.

RESULTADOS

As tabelas a seguir mostrar os resultados obtidos em cada configuração descrita na metodologia. Sendo a coluna INTERP equivalente à interpretabilidade, AUC equivalente a acurácia através da Área Sobre a Curva ROC.

Tabela 1 – Primeira configuração

	Teste		Desvio Padrão	
DATASET	**AUC**	**INTERP**	**AUC**	**INTERP**
abalone19	0,688	0,976	0,127	0,002
abalone918	0,812	0,935	0,086	0,008
ecoli4	0,870	0,856	0,105	0,019
ecoli0137vs26	0,793	0,809	0,217	0,018
glass2	0,665	0,782	0,160	0,027
glass4	0,815	0,798	0,176	0,030
glass5	0,876	0,839	0,177	0,021
glass016vs2	0,630	0,806	0,135	0,031
glass016vs5	0,873	0,872	0,151	0,024
pageblocks13vs4	0,938	0,946	0,071	0,013
shuttlec0vsc4	0,924	0,998	0,156	0,001
shuttlec2vsc4	0,947	0,983	0,144	0,005
vowel0	0,980	0,583	0,029	0,007
yeast1vs7	0,701	0,829	0,083	0,014
yeast2vs4	0,888	0,884	0,052	0,016
Média	**0,808**	**0,874**	**0,108**	**0,013**

Tabela 2 – Segunda configuração

	Teste		Desvio Padrão	
DATASET	**AUC**	**INTERP**	**AUC**	**INTERP**
abalone19	0,721	0,977	0,096	0,002
abalone918	0,812	0,935	0,086	0,008
ecoli4	0,888	0,859	0,061	0,020
ecoli0137vs26	0,859	0,820	0,157	0,027
glass2	0,705	0,797	0,160	0,046
glass4	0,892	0,804	0,118	0,039
glass5	0,876	0,839	0,177	0,021
glass016vs2	0,681	0,812	0,126	0,035
glass016vs5	0,936	0,879	0,107	0,018
pageblocks13vs4	0,950	0,946	0,059	0,014
shuttlec0vsc4	0,964	0,998	0,113	0,001
shuttlec2vsc4	0,979	0,985	0,069	0,005
vowel0	0,980	0,583	0,029	0,007
yeast1vs7	0,701	0,829	0,083	0,014
yeast2vs4	0,888	0,884	0,052	0,016
Média	**0,831**	**0,878**	**0,089**	**0,016**

Foram feitas comparações estatísticas através do teste de Wilcoxon com o sistema da literatura visto em (CÁRDENAS et al, 2016), para a primeira configuração, o p-valor da acurácia foi de $1.1920928955078125e^{-05}$ e o da interpretabilidade 0.05869007110595703, ou seja, em termos de acurácia existe diferença estatística significante e em termos de interpretabilidade não é possível afirmar. Já na segunda configuração, o p-valor da

acurácia foi de $4.1961669921875e^{-05}$ e o da interpretabilidade 0.03587532043457031, ou seja, nos dois casos existem diferença estatística significante. A média, em termos de acurácia e interpretabilidade, de todos os datasets da literatura foram, respectivamente, 86,6% e 82,7% (lembrando que na literatura a interpretabilidade está calculada como o inverso dessa porcentagem). Na primeira configuração, a média da nossa acurácia e interpretabilidade foram, respectivamente, 80,8% e 87,4%, já na segunda configuração 83,1% e 87,8% como mostram as Tabelas 1 e 2. Além disso, a média de gerações que foram necessárias para chegar nesses resultados foi de aproximadamente 231, sendo que no trabalho comparado foi de 1000.

CONSIDERAÇÕES FINAIS

Através dos resultados obtidos, é possível concluir que a abordagem que não utiliza pré-processamento para classificar dados desbalanceados, tem uma leve perda na acurácia, porém, a interpretabilidade é igual ou melhor estatisticamente. Além de, computacionalmente ser mais barata por não usar esses métodos para balancear os dados, comparando a quantidade média de gerações desse trabalho com outros, o AGMO converge mais cedo do que os demais.

Assim, fica claro que dependendo de qual seja a maior necessidade de quem utilizar esses sistemas computacionais, pode ser mais viável usar a nossa abordagem ou usar abordagens com pré-processamento.

REFERÊNCIAS

ALCALÁ-FDEZ, J. et al. Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms. **Fuzzy Sets and Systems**, v. 160, n. 7, p. 905-921, 2009.

HAIXIANG, G. et al. (2017). Learning from class-imbalanced data: Review of methods and applications. **Expert Systems with Applications**, 73, 220-239.

MAALOUF, M.; TRAFALIS, T. B. Robust weighted kernel logistic regression in imbalanced and rare events data. **Computational Statistics & Data Analysis**, v. 55, n. 1, p. 168-183, 2011.

MENDEL, J. M. **Fuzzy logic systems for engineering: a tutorial**. Proceedings of the IEEE, v.83, n. 3, p. 345–377, 1995.

PANIGRAHI, S. et al. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. **Information Fusion**, v. 10, n. 4, p. 354-363, 2009.

RODRIGUEZ, D. et al. **Preliminary comparison of techniques for dealing with imbalance in software defect prediction**. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 2014. p. 1-10.

YIJING, L. et al. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. **Knowledge-Based Systems**, v.94, p. 88-104, 2016.

CÁRDENAS, E. H., CAMARGO, H. A., & TÚPAC, Y. J. **Imbalanced datasets in the generation of fuzzy classification systems-an investigation using a multiobjective evolutionary algorithm based on decomposition**. In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp.1445-1452, 2016.