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

XXVIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2024

ANÁLISE DE ALGORITMOS NA GERAÇÃO DE CLASSIFICADORES BASEADOS EM REGRAS DE ASSOCIAÇÃO FUZZY PARA BASES DE DADOS DESBALANCEADA

João Gabriel Lima Almeida[1]; Matheus Giovanni Pires[2]
1. Bolsista PIBIC/CNPq, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: gabriel.lima.almeida@gmail.com
2. Orientador, DEXA, Universidade Estadual de Feira de Santana, e-mail: mgpires@uefs.br

PALAVRAS-CHAVE: dados desbalanceados; classificadores fuzzy; regras de associação

INTRODUÇÃO

Na computação, o uso de classificadores baseados em regras de associação é bastante útil em vários contextos, uma vez que esses algoritmos podem imitar a habilidade humana de tomar decisões diante de uma situação com base em um conjunto de regras. Entretanto, um problema sobre o uso desses tipos de classificadores é que para problemas do mundo real, a maioria das grandezas e valores não são discretos, mas sim contínuos e quebrados, o que impossibilita uma abordagem com regras de associação clássica.

Pensando nisso, os Sistemas Classificadores Baseados em Regras Fuzzy (SCBRF), utilizam a Lógica Fuzzy para lidar com esses tipos de problema, pois ela é capaz de processar valores contínuos e fracionados. O modelo mais conhecido na geração de um SCBRF é o FARC-HD (*Fuzzy Association Rule-based Classification Model for High-Dimensional Problems*) proposto por Alcala-Fdez, Alcala e Herrera (2011), que utiliza uma variação do algoritmo Apriori para gerar as regras do classificador fuzzy, o Apriori Fuzzy.

O desempenho de um SCBRF depende especialmente da base de regras utilizada, que, por sua vez, depende da forma como essas regras foram geradas. Um fator importante a respeito da geração é a base de dados utilizada. Uma base de dados balanceada e com bons exemplos consegue gerar boas regras, porém, quando há desequilíbrio na quantidade de exemplos por classe, as classes minoritárias tendem a ser ignoradas, ficando com pouca ou nenhuma cobertura pelas regras.

Diante disso, este trabalho tem como objetivo realizar uma análise da arquitetura do FARC-HD e do Apriori Fuzzy, observando suas potencialidades e defeitos na geração de regras com base de dados desbalanceados, além de elencar possíveis melhorias e mudanças no algoritmo.

METODOLOGIA

Inicialmente foi realizada a fase de investigação da literatura, com a leitura de artigos que circundam a temática e estudo da teoria de Conjuntos Fuzzy. Além do artigo

sobre o FARC-HD, outro artigo muito importante foi "*A fuzzy association rule-based classifier for imbalanced classification problems*", publicado por SANZ et al. (2021). Neste artigo é proposto uma versão modificada do FARC-HD, nomeada FARCI, que apresenta melhor desempenho na geração de regras para bases de dados desbalanceadas. Outro artigo importante foi "*How to Determine Minimum Support in Association Rule*", publicado por Hikmawati et al. (2020). O suporte é uma das métricas mais importantes utilizadas no Apriori Fuzzy e que possui grande impacto na geração das regras para as classes minoritárias.

A implementação do FARC-HD teve bastantes problemas devido à complexidade do algoritmo e ao tempo curto disponível para realizar a implementação (já que a pesquisa teve o tempo de execução de apenas cinco meses). Por esse motivo, foi utilizado o código implementado do FARCI em linguagem Java, já que ele é o mesmo algoritmo do FARC-HD, porém com alterações que já eram alinhadas com a nossa pesquisa. Para a utilização e análise do FARCI, foi necessário o desenvolvimento de *scripts* auxiliares para gerenciar a adição de novos conjuntos de dados, criação de arquivos de configuração e automação de tarefas.

RESULTADOS E DISCUSSÃO

O FARC-HD demonstrou alguns pontos que limitam o algoritmo em bases de dados desbalanceadas, alguns deles, inclusive, elencados e melhorados na variação do FARCI. Um dos pontos mais impactantes que gera o problema da ausência de regras para classes minoritárias são as métricas de seleção dos *itemsets* frequentes, que posteriormente são transformados em regras. No Apriori Fuzzy do FARC-HD são utilizadas as métricas *Support* e *Confidence*. No trabalho desenvolvido por SANZ et al. (2021) é proposto a substituição de *Confidence* por outra métrica chamada *Lift*, que é definda como: *Lift*(A→B) = *Confidence*(A→B) / *Support*(B). O *Lift* compara a confiança de uma regra com a expectativa de ocorrência do consequente, considerando a probabilidade desse consequente. Para regras que envolvem classes ou eventos raros, o *Lift* pode destacar regras que, embora tenham suporte baixo, ainda assim representam associações fortes

O FARCI implementa o uso de *Lift*, porém, o *Support* ainda é necessário e consequentemente, um valor de suporte mínimo precisa ser definido pelo usuário. Isso é problemático para o escopo da pesquisa, já que pequenas variações no valor do suporte mínimo causam grandes variações na quantidade de regras geradas. Por isso, uma alteração ideal seria a substituição por uma métrica independente de parâmetros. Outro detalhe importante do FARCI é o uso de métodos diferentes no cálculo do suporte. Originalmente, o cálculo é feito utilizando o produto dos graus de pertinência, mas SANZ et al. (2021) propõem outras formas alternativas ao produto, como a média aritmética, geométrica e harmônica.

Apesar do Apriori Fuzzy do FARCI ser de fato melhor para dados desbalanceados, o código é experimental, tem um uso complicado, não prático e difícil de adaptá-lo para outras necessidades. Outro problema é a ausência de documentação do código, apesar do artigo do FARC-HD descrever as especificações do algoritmo e o do FARCI destacar as mudanças, a codificação possui diversos detalhes importantes que, como não há documentação das classes e métodos, não ficam claros e tornam o

entendimento e uso do código bastante dificultoso. Outro problema é que o código é bastante acoplado, o que impossibilita a substituição do Apriori Fuzzy por outros algoritmos de geração de regras e o uso do Apriori Fuzzy em outros algoritmos.

Devido a esses problemas, sugerimos uma remodelagem e reimplementação do Apriori Fuzzy presente no FARC-HD e melhorado no FARCI, uma vez que ele ainda possui aspectos que podem ser aprimorados, principalmente na questão do desempenho computacional, que é bastante útil no contexto de geração de regras. A reimplementação tem como principais objetivos:

- Baixo acoplamento do Apriori Fuzzy;
- Entrada de dados mais flexível, separando a leitura do algoritmo para permitir uso de padrões além do ARFF;
- Documentação do código;
- A melhoria do algoritmo para torná-lo computacionalmente eficiente, como a mudança no cálculo do grau de pertinência dos itens que, da forma atualmente implementada, realiza cálculos repetidos para cada itemset, mas poderiam ser armazenados por label e reutilizados já que o valor nas transações não vai mudar;
- A implementação de novas métricas para melhorar o desempenho do algoritmo em relação a geração de regras para classes minoritárias, como por exemplo, a substituição das métricas de *Support* e *Confidence* por *Lift* e outras métricas que sejam independentes de entrada e valorizem mais a qualidade e eficiência da regra voltada para a classe;
- A não limitação para conjuntos de dados binários, já que a estrutura original do FARC-HD é feita para suportar dados com múltiplas classes, mas a implementação feita do FARCI é limitada apenas a datasets binários.
- Cálculo de suporte via média geométrica, a mais eficiente para dados desbalanceados como demonstrado por SANZ et al. (2021).
- A implementação visando a construção de uma biblioteca ou *framework*, uma vez que não há bibliotecas que ofereçam o algoritmo Apriori Fuzzy para geração de regras de associação Fuzzy, e a implementação nesse formato facilitaria o uso do algoritmo e padronização das entradas e saídas.

A Figura 1 ilustra o diagrama UML da modelagem das classes e seus métodos. Sobre o algoritmo, ele deve seguir a mesma estrutura do presente no FARC-HD, sendo dividido nesse caso em duas partes:

1. Leitura: leitura do conjunto de entrada e fornecimento das informações sobre as transações tratadas; informações de cada atributo e seus limites/valores e atributos de saída e seus valores.
2. Apriori Fuzzy: geração das funções de pertinência de cada atributo e das regras de associação fuzzy.

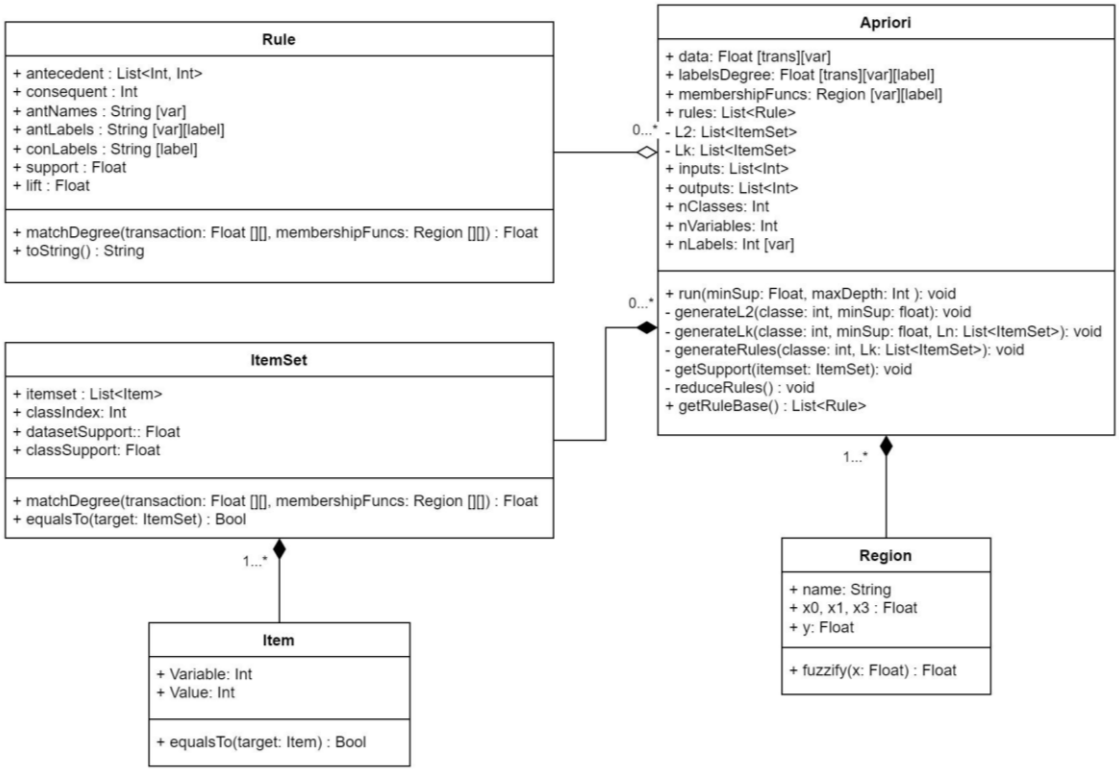

Figura 1: Diagrama de classes da modelagem do Apriori Fuzzy.

CONSIDERAÇÕES FINAIS

Ao fim desse trabalho, espera-se que os problemas levantados e a reimplementação proposta possam levar a uma implementação do Apriori Fuzzy independente, com baixo acoplamento e com bom desempenho na geração de regras de associação fuzzy, melhor que o do Apriori Fuzzy original.

Além disso, espera-se obter um desempenho computacional melhor, através das mudanças sugeridas no cálculo do suporte e na geração da árvore de combinações. Também é esperado que, com a implementação focada em se tornar um *framework*, o código tenha a usabilidade facilitada e mais prática, permitindo que ele possa ser utilizado tanto em usos práticos quanto ser inserido como um passo de outros algoritmos de geração de SCBRFs.

REFERÊNCIAS

ALCALA-FDEZ, J.; ALCALA, R.; HERRERA, F. A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems with Genetic Rule Selection and Lateral Tuning. **IEEE Transactions on Fuzzy Systems**, vol.19, n.5, pp.857–872, 2011.

SANZ, J.; SESMA-SARA, M.; BUSTINCE, H. A fuzzy association rule-based classifier for imbalanced classification problems. **Information Sciences**, vol.577, pp.265–279, 2021.

ENY HIKMAWATI; KRIDANTO SURENDRO. **How to Determine Minimum Support in Association Rule**. Proceedings of the 9th International Conference on Software and Computer Applications, 2020.