

ALGORITMO APRIORI VERSUS FUZZY APRIORI: ANÁLISE DE DESEMPENHO APLICADO EM SISTEMA DE DETECÇÃO DE INTRUSOS

RICARDO FERREIRA VIEIRA DE CASTRO^{1*}, NIVAL NUNES DE ALMEIDA²,
MARIA LUIZA VELLOSO³

¹ Me. em Engenharia Eletrônica, UERJ, Rio de Janeiro-RJ. Fone: (21) 2123-1119, ricardo.castro@int.gov.br

² Dr. em Engenharia Elétrica, UFRJ, Rio de Janeiro-RJ. Fone: (21) 2334-2165, nivalnunes@yahoo.com.br

³ Dr. em Engenharia Elétrica, PUC, Rio de Janeiro-RJ. Fone: (21) 2587-7633, mlfv@centroin.com.br

Apresentado no
Congresso Técnico Científico da Engenharia e da Agronomia – CONTECC' 2015
15 a 18 de setembro de 2015 - Fortaleza-CE, Brasil

RESUMO: O algoritmo APRIORI, tradicionalmente usado na extração de regras de associação, quando utilizado com atributos contínuos, ou quantitativos, é necessário discretizar os atributos criando categorias a partir dos intervalos discretos. Esses intervalos podem subestimar ou superestimar elementos próximos dos limites das partições, e portanto levar a uma representação imprecisa de semântica. Uma maneira de tratar este problema é utilizar algoritmos de mineração de regras de associação fuzzy (FARM - Fuzzy Association Rule Mining) que transforma os atributos quantitativos em partições de termos linguísticos. Neste artigo, o desempenho dos algoritmos APRIORI e FUZZY APRIORI são comparados através do percentual de acurácia alcançado à partir das regras extraídas de uma aplicação que teve como base de dados registros de conexões TCP/IP de um Sistema de Detecção de Intruso. Os resultados sugerem que o desempenho dos algoritmos tem relacionamento com o número de regras de classificação geradas.

PALAVRAS-CHAVE: Fuzzy Apriori, Extração de Regras de Associação, Detecção de Intruso.

ALGORITHM APRIORI VERSUS FUZZY APRIORI: PERFORMANCE ANALYSIS APPLIED FOR A INTRUSION DETECTION SYSTEM

ABSTRACT: The APRIORI algorithm, traditionally used in the extraction of association rules, when used with continuous attributes, or quantitative, it is necessary to discretize the attributes creating categories from discrete intervals. These intervals may underestimate or overestimate elements near the boundaries of the partitions, therefore inducing an inaccurate semantical representation. One way to address this problem is to use algorithms for mining fuzzy association rules (FARM - Fuzzy Association Rule Mining) that transforms the quantitative attributes in linguistic terms partitions. In this article, the performance of FUZZYAPRIORI and APRIORI algorithm are compared through the percentage of accuracy achieved from the extracted rules of an application based on database of records of TCP / IP of a Intrusion Detection System. The results suggest that the performance of the algorithms has relationship with the number generated classification rules.

KEYWORDS: Fuzzy Apriori, Association Rule Mining, Intruder Detection.

INTRODUÇÃO

Regras de associação são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados (Menzies, 2003). Baseado no conceito de regras fortes, Agrawal (1993) propôs o algoritmo Apriori como forma de introduzir regras de associação para descobrir regularidades entre os produtos vendidos em transações de grande porte registradas por ponto de venda (POS). Os primeiros estudos de Agrawal na extração de regras de associação estava relacionado com transações de um mesmo conjunto de dados usando valores binários. Dando continuidade a suas pesquisas, Agrawal propôs um método para extração de regras de associação de atributos com valores quantitativos. Seu método proposto transforma os itens quantitativos em itens binários através da partição dos atributos com domínio quantitativo. No entanto, esse método proposto não resolve o

problema conhecido como *sharp boundaries* que consiste em poder subestimar ou superestimar elementos próximos dos limites da partição, e portanto levar a uma representação imprecisa de semântica. Para tratar o problema de *sharp boundaries*, uma nova técnica utilizando *fuzzy sets* e *fuzzy* itens (Fuzzy Apriori), normalmente em forma de rótulos ou termos linguísticos, é usada e definida no domínio. Em Kuok (1998) descreve *fuzzy association rules mining* (FARM): "Extração de regras de associação fuzzy é o descobrimento das regras de associação usando conceito fuzzy set de tal modo que atributos quantitativos podem ser manipulados".

MATERIAL E MÉTODOS

A escolha de uma base de dados confiável para fazer a análise de desempenho entre os dois algoritmos, buscou-se uma que já fosse amplamente utilizada nos trabalhos acadêmicos, e que tivesse características adequadas ao trabalho proposto. A base de dados que serviu como fonte para os testes foram extraídos do repositório UC Irvine Machine Learning Repository (UCI)¹, a qual foi utilizada no KDD Cup 1999 Data. Nessa competição foi escolhido o problema de detecção de intruso, e a base de dados serviu como *benchmarking* para a construção de uma rede de detecção de intrusos, em que um modelo preditivo seria capaz de distinguir entre conexões “ruins” como invasões ou ataques, e “boas” como conexões normais. Os dados são formados por um conjunto de registros de conexões de pacotes TCP/IP, onde incluem um conjunto de dados de treinamento e outro de teste. Cada conexão de rede (Tabela 1) é representada por 41 atributos, classificados como qualitativo ou quantitativo, que descrevem as diferentes características de cada conexão, e ainda uma classe nomeada por um rótulo (classe de ataque), que determina um tipo de ataque específico ou normal.

Tabela 1. Modelo básico dos registros de conexões

Atributos - características da conexão											Atributo - classe de ataque
cnx#1	1	2	3	4	40	41	C	Classe: <i>Normal ou Ataque</i>	
cnx#2	1	2	3	4	40	41	C		
cnx#3	1	2	3	4	40	41	C		
cnx#n	:	:	:	:	:	:	:	:	:	Ataques: <i>DoS/R2L/U2R/Probe</i>	
	1	2	3	4	40	41	C		
	Registros de Conexões TCP/IP										

O conjunto de dados utilizado na competição é composto por três bases de dados diferentes denominadas de 10% KDD, Corrected KDD e Whole KDD. Tais bases referem-se à base de dados utilizada para treinamento, teste e o conjunto original respectivamente. A respectiva distribuição dos tipos de ataque DoS, Probe, u2r, r2l e normal podem ser visualizadas na Tabela 2.

Tabela 2. Características básicas da base de dados KDD99

Categoria	10% KDD		Corrected KDD		Whole KDD	
Ataque	Treino	%	Teste	%	Original	%
DoS	391458	79,24	229853	73,90	3883370	79,28
Probe	4107	0,83	4166	1,34	41102	0,84
u2r	52	0,01	70	0,02	52	0,001
r2l	1126	0,23	16347	5,26	1126	0,02
Normal	97277	19,69	60593	19,48	972780	19,86
Total	494020	100	311029	100	4898430	100

Um aspecto que tem grande impacto em problemas de extração de regras de associação e classificação é a desigualdade na distribuição dos padrões entre os grupos. No caso do conjunto de dados apresentado na Tabela 1, este desbalanceamento de dados pode ser percebido. Para o desenvolvimento do trabalho foi selecionado um novo conjunto de dados denominado 10%KDD/DoS

¹University of California - Irvine - <http://archive.ics.uci.edu/ml/index.html>

onde somente os ataques da categoria DoS foram considerados, pois estes representam 79,24% dos registros. Considerando o conjunto 10%KDD/DoS, temos um sub-conjunto de ataques onde os ataques *back*, *teardrop*, *pod* e *land* juntos representam 0.7% do total dos ataques (Tabela 3). Percentual bem inferior aos 79.38% dos demais ataques smurf e neptune, o que apontou para a necessidade de uma reamostragem do conjunto DoS de forma a tratar este desbalanceamento de dados.

Conforme mencionado anteriormente, para cada conexão TCP/IP existem 41 atributos que dependendo valor atribuído a cada um desses, o atributo de saída denominado de classe de ataque, determinará que tipo de ataque estará ocorrendo. A reamostragem sugerida tem o objetivo selecionar um subconjunto de atributos, ditos relevantes, a fim de reduzir a dimensão do banco de dados. Assim, reduz-se a complexidade do banco de dados, bem como o tempo de processamento. Para seleção dos atributos relevantes, optou-se pelo conjunto de atributos sugerido segundo Wang (2008) (Tabela 4).

Tabela 3. Conjunto dados 10%KDD/DoS

Ataque	Qtd amostra	%	Categoria
Smurf	280790	57.45	dos
Neptune	107201	21.93	dos
Normal	97277	19.90	normal
Back	2203	0.45	dos
Teardrop	979	0.20	dos
Pod	264	0.05	dos
Land	21	0.004	dos
TOTAL	488735	100%	

Tabela 4. Reamostragem segundo WANG

Ataque	10% KDD/DoS	%	WANG	%
Smurf	280790	57.45	10000	17.70
Neptune	107201	21.93	5000	8.85
Normal	97277	19.90	40000	70.78
Back	2203	0.45	1000	1.77
Teardrop	979	0.20	400	0.71
Pod	264	0.05	100	0.18
Land	21	0.004	10	0.02
TOTAL	488735	100	56510	100

Os dez atributos relevantes sugeridos são: *service*, *flag*, *srcbytes*, *dstbytes*, *wrong_fragment*, *hot*, *num_compromised*, *count*, *srcvcount* e *dst_host_srv_diff_host_rate*. Para o desenvolvimento dos experimentos foram escolhidas as ferramentas WEKA (Hall, 2009) e KEEL (Alcala, 2011).

RESULTADOS E DISCUSSÃO

Cada um dos algoritmos aplicados durante a realização dos experimentos teve como objetivo extrair regras de associação de uma base de dados contendo registros de conexões TCP/IP. Uma vez gerados os resultados relativos à extração das regras de associação, procedeu-se a uma análise de desempenho das regras geradas através de algoritmos de classificação. Para análise de desempenho dos Algoritmos Apriori e Fuzzy Apriori, foram selecionados valores mínimos de suporte e confiança que apresentassem melhores resultados para essa análise. Na análise comparativa da quantidade das regras de classificação (Figura 1) e da acurácia das regras de associação (Figura 2) foram utilizados, por exemplo, os valores de confiança mínimos de 40% e 90%. As curvas dos demais valores de confiança apresentaram distribuição semelhante.

Figura 1. Distribuição de regras de classificação.

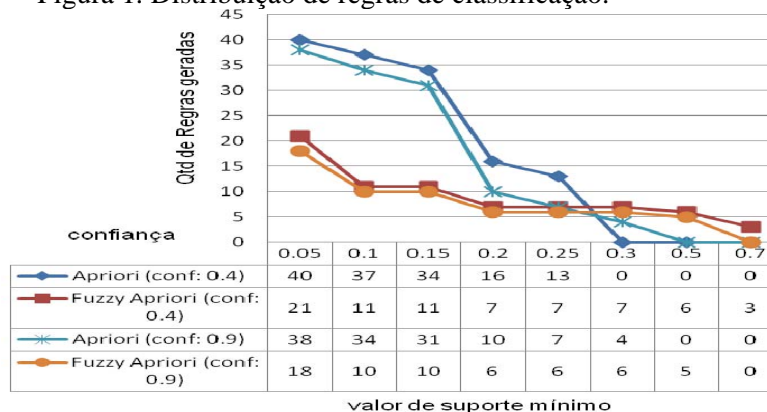
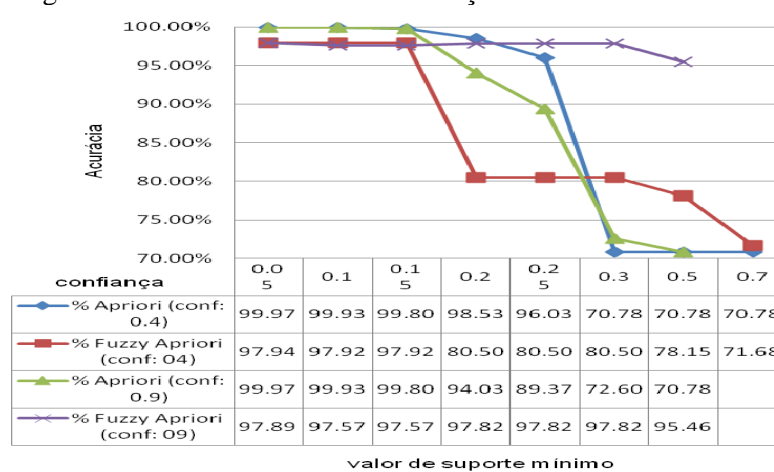


Figura 2. Percentual de acurácia alcançado.



CONCLUSÕES

Com relação à quantidade do número de regras de classificação associativa, para ambos os algoritmos, como já era esperado, conforme o valor de suporte mínimo aumentava o número de regras de associação classificativa diminuía.

Quanto às taxas de acurácia de acertos da classe de ataque, observa-se que com valor de suporte mínimo de confiança de 40% o percentual de acerto no Fuzzy Apriori (80,50%) foi superior ao Apriori (70,78%) a partir do valor de suporte mínimo de 30%. Neste momento, o subconjunto de regras de classificação associativa do algoritmo Apriori deixou de ser gerado, enquanto que o Fuzzy Apriori ainda gerou 7 (sete) regras de classificação.

Com relação ao valor de suporte mínimo de confiança de 90% o percentual de acerto no Fuzzy Apriori (97,82%) foi superior ao Apriori (94,03%) a partir do valor de suporte mínimo de 20%, mesmo tendo um subconjunto de regras de classificação associativa inferior (6 regras) ao do algoritmo Apriori (10).

Pelos experimentos, observamos que o algoritmo Fuzzy Apriori tende a ter o percentual de acurácia de acertos da classe de ataque maior que o algoritmo Apriori conforme a quantidade de regras de classificação no Apriori tende a zero e as regras de classificação do algoritmo Fuzzy Apriori continuam a ser geradas. Essa constatação deve-se ao fato que o algoritmo Fuzzy Apriori exerce tratativa diferenciada sobre os valores limites das partições.

REFERÊNCIAS

- Agrawal, Rakesh; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases, SIGMOD 5/93, 207-216, Washington, USA, 1993.
- Alcala, J. ; FERNANDEZ, A.; LUENGO, J.; DERRAC, J.; GARCIA, S.; SANCHEZ, L.; HERRERA, F. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287
- Fayyad, U; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI Magazine 17 (1996) 37-54.
- Hall, Mark; FRANK, Eibe; HOLMES, Geoffrey; PFAHRINGR, Bernhard; REUTEMAN, Peter; WITTEN, Ian H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009, Volume 11, Issue 1.
- Kuok, C. M.; FU, A.; WONG, M. H. Mining Fuzzy Association Rules in Databases, SIGMOD Record, pp. 41-46, Vol. 27, No. 1, 1998.
- Menzies, Tim; HU, Ying. Data Mining For Busy People. IEEE Computer, Outubro de 2003, pgs. 18-25.
- Wang, Wei; GOMBAULT, S.; GUYET, T. Towards Fast Detecting Intrusions: Using Key Attributes of Network Traffic. Internet Monitoring and Protection, 2008. ICIMP '08. The Third International Conference on, janeiro/2008.