Aggregating Interestingness Measures in Associative Classifiers

Matheus Freitas da Silva[1], **Veronica Oliveira de Carvalho**[1]

[1]Instituto de Geociências e Ciências Exatas – Universidade Estadual Paulista (Unesp)
Rio Claro – SP – Brasil

mathfreitas15@gmail.com, veronica@rc.unesp.br

Abstract. *Associative classification, which has been widely used in several domains, aims to obtain a predictive model in which the process is based on the extraction of association rules. Model generation occurs in steps, one of them aimed at ordering and pruning a set of rules. Regarding ordering, one of the solutions is to rank the rules by means of objective measures (OMs). The ordering criterion impacts the accuracy of the classifier. In the literature's works the measures are explored individually. Based on the exposed, this work aims to explore the aggregation of measures, in which several OMs are considered at the same time, in the context of associative classifiers.*

Resumo. *A classificação associativa, a qual vem sendo muito utilizada em diversos domínios, visa a obtenção de um modelo preditivo em que o processo é baseado na extração de regras de associação. A geração do modelo ocorre em etapas, sendo uma delas voltadas a ordenar e podar um conjunto de regras. No que se refere a ordenação, uma das soluções é ranquear as regras por meio de medidas objetivas (MOs). O critério de ordenação impacta a acurácia do classificador. Nos trabalhos da literatura as MOs são exploradas individualmente. Diante do exposto, este trabalho tem por objetivo explorar a agregação de medidas, em que várias MOs são consideradas ao mesmo tempo, no contexto de classificadores associativos.*

1. Introdução

Todos os dias são produzidas e coletadas uma imensa quantidade de informações. Uma empresa de inteligência de mercado disponibilizou, no início de 2018, um infográfico que apresenta o que acontece no mundo digital em um minuto[1], dentre os quais encontram-se os seguintes eventos: (i) 103.447.250 e-mails classificados como spam são enviados; (ii) 4.146.600 vídeos são assistidos no YouTube; (iii) 3.607.080 pesquisas são feitas no Google; (iv) 527.760 fotos são compartilhadas no Snapchat; (v) 456 mil tuítes são compartilhados; (vi) 154.200 chamadas são feitas pelo Skype; (vii) 120 novos profissionais são cadastrados no LinkedIn; (vii) 46.740 fotos são publicadas no Instagram.

Diante dessa enorme quantidade de dados, selecionar um subconjunto de informações relevantes para um determinado domínio, e deste, extrair conhecimento que possa ser utilizado, é de fundamental importância nos dias de hoje. Dependendo do objetivo

[1]http://itbroker.com.br/pt/o-mundo-digital-em-um-minuto/. Acessado em: 01/06/2018.

a ser alcançado, diversas técnicas podem ser utilizadas, dentre as quais classificação e associação. A classificação visa a indução de um modelo com a capacidade de predizer futuras categorias, denominadas classes, de determinados objetos. Já a associação é uma tarefa que visa a extração de regras que expressem o quanto a presença de um conjunto de itens, existentes em um conjunto de dados, implica na presença de outro(s). Maiores detalhes em [Han et al. 2013].

Uma outra técnica, denominada classificação associativa, também vem sendo muito utilizada nos últimos anos, a qual mescla características da classificação e da associação. A classificação associativa também visa a obtenção de um modelo com a capacidade de predizer futuras categorias. Contudo, o processo de geração do modelo é baseado na extração de regras de associação (RAs). A técnica vem sendo utilizada em diversos domínios, a saber: sistemas de recomendação [Yin et al. 2018, Moreno et al. 2016], detecção de spam [Nandhini et al. 2015], saúde [Alwidian et al. 2018, Singh et al. 2016], desenvolvimento de software [Shao et al. 2017, Ma et al. 2014], entre outros.

A construção de um classificador associativo ocorre em etapas, a saber: (a) extração de um conjunto de regras de associação classificativas (consequente das RAs contém apenas rótulos de classe); (b) geração do modelo via ordenação e poda das regras geradas e (c) predição. Existem diversas abordagens na literatura em relação a cada uma das etapas. No que se refere a ordenação, uma das soluções é realizar o ranqueamento das regras por meio de medidas objetivas[2] (MOs). Essas medidas são utilizadas para ordenar as regras a fim de priorizar aquelas consideradas mais relevantes. Isso se deve ao fato da tarefa de associação extrair uma quantidade de regras muita elevada, sendo muitas delas não interessantes para a construção do modelo de classificação.

O foco deste trabalho encontra-se justamente na ordenação (etapa (b)). [Abdelhamid et al. 2016] apresentam alguns desafios de pesquisa relacionados aos classificadores associativos, dentre eles a ordenação das regras. Os autores afirmam que escolher o critério de ordenação apropriado é uma tarefa crítica que impacta a acurácia do classificador. Nas propostas encontradas na literatura as MOs são exploradas separadamente. Contudo, existem trabalhos, no contexto de RAs, que investigam o uso agregado de MOs [Bouker et al. 2014, Nguyen Le et al. 2009, Yang et al. 2009]. A ideia dos trabalhos é auxiliar o usuário a decidir qual medida utilizar, uma vez que muitas delas existem. Diante do exposto, este trabalho tem por objetivo explorar a agregação de medidas, em que várias MOs são consideradas ao mesmo tempo, no contexto de classificadores associativos. A expectativa é que a ordenação obtida via agregação de medidas resulte em classificadores mais precisos.

Este trabalho está estruturado como segue: a Seção 2 introduz alguns conceitos, necessários para compreender o trabalho, assim como discute alguns trabalhos relacionados. A metodologia proposta para se explorar a agregação de MOs em classificadores associativos é descrita na Seção 3, seguida pela seção de experimentos (Seção 4), resultados e discussão (Seção 5). A Seção 6 finaliza o artigo com as conclusões e trabalhos futuros.

[2]Neste trabalho medidas objetivas e medidas de interesse são usadas como sinônimo.

2. Definições e Trabalhos Relacionados

Como mencionado, este trabalho tem como objetivo explorar a agregação de MOs em classificadores associativos. Para tanto, faz-se necessário o entendimento de alguns conceitos (Seções 2.1 e 2.2), assim como a revisão de trabalhos relacionados (Seção 2.3), os quais são apresentados nessa seção.

2.1. Classificação Associativa

A classificação associativa visa a obtenção de um modelo preditivo em que o processo é baseado na extração de RAs. Contudo, nesse caso, as regras extraídas apresentam um padrão específico, denominadas de regras de associação classificativas (RACs), nas quais o consequente das regras contém apenas rótulos de classe. Dentre as vantagens dos classificadores associativos nota-se [Kannan 2010]: (i) capturam todas as possíveis associações, que atendam ao suporte e a confiança especificados, entre os itens e os rótulos de classe, expandindo, assim, o espaço de busca; (ii) lidam naturalmente com valores ausentes e outliers, uma vez que manipulam apenas associações estatisticamente significativas; (iii) nenhuma suposição é feita sobre a dependência ou a independência dos atributos; (iv) geralmente geram modelos mais precisos do que os classificadores tradicionais; (v) o modelo é composto por um conjunto de regras facilmente compreendido por humanos e pode ser editado.

Considere um conjunto de dados representado por uma tabela relacional, a qual é composta por m objetos (exemplos) descritos por n atributos. Cada objeto encontra-se associado a uma das p classes conhecidas (categorias a serem preditas). Tal representação é apresentada a esquerda da Tabela 1. Cada objeto O_i é, portanto, descrito por um vetor $O_i = [v_{i1}, v_{i2}, ..., v_{in}, c_i]$, em que v_{ij} representa o valor do objeto i no atributo j e c_i a classe associada ao objeto. No contexto de classificação associativa, uma vez que o modelo a ser obtido é baseado no processo de extração de RAs, os objetos são denominados de transações e cada transação é composta por um conjunto de itens, os quais representam os possíveis valores dos atributos. Cada item é considerado, portanto, um par <atributo, valor>. Essa notação é apresentada a direita da Tabela 1.

Tabela 1. Representação abstrata de um conjunto de dados utilizada para a tarefa de classificação associativa.

	A_1	A_2	...	A_n	R
O_1	v_{11}	v_{12}	...	v_{1n}	c_1
O_2	v_{21}	v_{22}	...	v_{2n}	c_2
⋮	⋮	⋮	⋱	⋮	⋮
O_m	v_{m1}	v_{m2}	...	v_{mn}	c_p

$\Longrightarrow$

t_1	$A_1 = v_{11}$	$A_2 = v_{12}$	...	$A_n = v_{1n}$	$R = c_1$
t_2	$A_1 = v_{21}$	$A_2 = v_{22}$	...	$A_n = v_{2n}$	$R = c_2$
⋮	⋮	⋮	⋱	⋮	⋮
t_m	$A_1 = v_{m1}$	$A_2 = v_{m2}$	...	$A_n = v_{mn}$	$R = c_p$

Formalmente, seja D um conjunto de dados composto por um conjunto de itens $I = \{i_1, ..., i_n\}$, por um conjunto de rótulos $R = \{c_1, ..., c_p\}$ e por um conjunto de transações $T = \{t_1, ..., t_m\}$, na qual cada transação $t_i \in T$ é composta por um subconjunto de itens $A \subseteq I$ e um rótulo de classe $c \in R$ tal que $t_i = A \cup c$. A regra de associação classificativa é uma implicação na forma $A \Rightarrow c$, em que $A \subseteq I$ e $c \in R$. A regra $A \Rightarrow c$ ocorre no conjunto de transações T com *confiança* $conf$ e *suporte* sup, em que $P(A\, c)$ representa o suporte da regra (probabilidade da ocorrência da transação $A \cup c$) e $P(c|A)$ a confiança da regra (probabilidade condicional de c dado A).

A construção de um classificador associativo ocorre em etapas, a saber: (a) extração de um conjunto de regras de associação classificativas; (b) geração do modelo via ordenação e poda das regras geradas em (a) e (c) predição. Em relação a parte (a), algoritmos de extração de RAs podem ser adaptados para extrair apenas as regras cujo consequente contenha um rótulo de classe. Em [Liu et al. 1998], por exemplo, os autores apresentam uma adaptação do algoritmo Apriori para obtenção das regras em questão. Por ser algo intuitivo, este trabalho não apresentará tal descrição. Tendo o conjunto de RACs, é necessário selecionar, das regras geradas, aquelas que irão compor o classificador (etapa (b)). Isso se deve ao fato da problemática relacionada a extração de RAs, a qual tende a gerar um número elevado de regras. Assim, nem todas as regras geradas são relevantes ao modelo. Para realizar tal seleção, em geral, as regras passam por um processo de ordenação e poda. Ao final, as regras selecionadas compõem o modelo. Além disso, é necessário definir a estratégia a ser adotada para predição (etapa (c)) da classe de novos objetos. Uma boa revisão sobre classificadores associativos é apresentada em [Thabtah 2007].

Existem diversos algoritmos de classificação associativa disponíveis na literatura. O algoritmo CBA [Liu et al. 1998], apresentado no Algoritmo 1, é geralmente o utilizado como baseline, sendo aqui o utilizado (vide Seção 3). Em linhas gerais, o algoritmo funciona da seguinte maneira: na linha 1 obtém-se as RACs. Na linha 2 as RACs são ordenadas segundo três critérios: confiança, suporte e ordem de geração. Desse modo, uma regra r_i precede uma regra r_j, em uma lista ordenada, se confiança(r_i) > confiança(r_j); se as confianças são iguais, mas suporte(r_i) > suporte(r_j); se os suportes são iguais, mas r_i foi gerada antes de r_j. É exatamente neste ponto que este trabalho se insere (linha em destaque). Na sequência, linhas 3 a 17, inicia-se a seleção das RACs que comporão o modelo considerando a ordenação realizada. Para cada regra r verificam-se as transações que a mesma cobre e se a mesma cobre corretamente pelo menos uma transação (linhas 5 a 10). Em seguida (linhas 11 a 16), verifica-se se a regra cobriu corretamente ao menos uma transação e, em caso afirmativo, a insere no modelo e retira de D todas as transações cobertas pela mesma. A cada regra inserida computa-se a classe majoritária atual e o erro atual do classificador construído até o momento. Ao final (linhas 18 a 20), descartam-se todas as regras após a primeira regra q que produza o menor número de erros e associa-se a classe majoritária atrelada a q em C como regra *default*. Em relação a predição, dado um novo objeto, utiliza-se a classe associada a primeira regra que satisfizer as características do objeto.

2.2. Medidas Objetivas

As medidas objetivas, como o suporte e a confiança, são comumente empregadas na etapa de ordenação dos classificadores associativos. De fato, elas são a estratégia mais utilizada para pós-processar RAs. Uma MO computa a relevância de uma regra considerando a informação disponível no conjunto de dados. 61 MOs são definidas e discutidas em [Tew et al. 2014], o qual apresenta uma boa revisão sobre o tema; portanto, as mesmas não serão aqui descritas. As MOs são usualmente computadas para cada regra. As regras podem ser ranqueadas, tendo como base esses valores, para se obter uma lista ordenada de regras. Em geral, quanto maior o valor de uma dada MO melhor ranqueada é a regra.

Uma vez que muitas MOs existem, algumas soluções foram propostas de modo a auxiliar o usuário a decidir qual delas utilizar. Um review discutindo um série de abor-

Algoritmo 1 Algoritmo CBA. Adaptado de [Liu et al. 1998].

Entrada: Conjunto de dados D (como na Tabela 1)
Saída: Classificador C

1: $R \leftarrow RACs(D)$ ▷ Obtenção das Regras de Associação Classificativas
2: $R \leftarrow Ordena(R)$ ▷ Ordenação
3: **para todo** $r \in R$ **faça** ▷ Poda / Construção do Modelo
4: $temp \leftarrow \varnothing$
5: **para todo** $d \in D$ **faça**
6: **se** d satisfaz as condições de r **então**
7: armazena d.id em $temp$
8: marca r se a mesma classifica corretamente d
9: **fim se**
10: **fim para**
11: **se** r está marcada **então**
12: insere r no fim de C
13: deleta todos os casos com os ids em $temp$ de D
14: seleciona uma classe majoritária para o classificador atual C
15: computa o número total de erros de C
16: **fim se**
17: **fim para**
▷ Finalização do Modelo
18: encontra a primeira regra q em C com o menor número total de erros
19: descarta todas as regras depois de q em C
20: adiciona a classe majoritária associada à q no fim de C
21: **retorna** C

dagens pode ser visto em [Bong 2014]. Dentre as soluções existentes existem aquelas voltadas a agregar os valores de duas ou mais MOs visando não se ter que selecionar uma medida para ordenar as regras. Em [Nguyen Le et al. 2009] os autores agregam MOs usando uma Integral de Choquet, a qual consiste em fazer uma somatória dos valores de várias medidas utilizando-se pesos. Em [Yang et al. 2009] os autores utilizam um algoritmo genético que a cada geração cria diversas equações contendo um conjunto de medidas agregadas. Em [Bouker et al. 2014] os autores propõem uma solução de agregação, que por ser determinística e dependente apenas de informações contidas nos dados, foi a escolhida para ser utilizada neste trabalho (vide Seção 3).

Em [Bouker et al. 2014] os autores agregam as MOs computadas para todas as RAs em uma regra denominada *regra referência*. Tal regra contém o melhor valor de cada MO no conjunto de regras utilizado. A *regra referência* é então utilizada em um algoritmo denominado *SkyRule*. Esse algoritmo encontra a RA mais similar a *regra referência* e, na sequência, as regras não dominadas por ela. Define-se que uma regra X domina uma regra Y quando todas os valores das MOs de X são melhores do que os valores das MOs de Y. O algoritmo *SkyRule* encontra-se inserido em um laço de tal modo que a cada iteração a regra mais próxima a *regra referência* é extraída, juntamente com todas as regras não dominadas por ela. Considerando a execução completa do laço, cria-se um ranking das RAs por ordem de dominância. As regras menos dominadas são consideradas as melhores. Para maiores detalhes vide [Bouker et al. 2014].

2.3. Trabalhos Relacionados

Embora muitos trabalhos referentes a classificação associativa existam, poucos deles exploram o efeito das MOs na etapa de ordenação. Inicialmente, [Azevedo and Jorge 2007]

investigam o uso de 10 MOs em 17 conjuntos de dados em relação a ordenação e a predição em classificadores associativos. Os autores concluem que a medida *conviction* é a mais adequada a ser utilizada. Em [Jalali-Heravi and Zaïane 2010] avalia-se o impacto de 53 MOs em cada fase do processo de geração do modelo (etapas (a), (b) e (c) (Seção 2.1)). Além disso, a combinação das melhores medidas obtidas nas fases (b) e (c) também é avaliada. Em relação a etapa (a) apenas o suporte é utilizado, tanto em uma visão global (em todo o conjunto de dados) quanto em uma local (na classe). Em relação a etapa (b) os autores exploram a ordenação com poda e sem poda. Em relação a etapa (c) os autores exploram duas estratégias: (i) selecionar a regra melhor ranqueada que cubra o exemplo e (ii) separar as regras que cobrem o exemplo em grupos, de acordo com a classe, e realizar a média da MO em análise em cada grupo. O grupo com melhor média define a classe. Em relação a precedência de ordenação das regras, os autores consideram o seguinte: valor da medida em análise, suporte, tamanho da regra (regras mais gerais são preferidas). Os autores avaliam os resultados em 20 conjuntos de dados e concluem que muitas das MOs avaliadas melhoraram o desempenho do classificador. Contudo, que não existe uma medida que seja mais adequada para todos os conjuntos de dados, nem para todas as fases do processo de geração do classificador.

[Kannan 2010] apresentam um estudo para avaliar a influência de 39 MOs tanto em relação a poda quanto em relação a ordenação. Os autores testam 3 alternativas de ordenação: (i) somente uma dada medida, (ii) uma dada medida mais critérios de desempate como no CBA (confiança, suporte e regra gerada primeiro) e (iii) ordenação por uma dada medida seguida da reordenação, por meio da estratégia (ii), das k melhores regras selecionadas. Os autores realizaram o estudo em um único conjunto de dados referente a estudantes em um programa de ensino a distância. Os autores concluem que a precisão do classificador associativo pode ser melhorada usando a medida de interesse apropriada, tanto para a poda quanto para a ordenação.

Já em [Yang and Cui 2015] os autores visam melhorar o desempenho dos classificadores associativos em conjuntos de dados desbalanceados por meio do estudo de 55 MOs em 9 conjuntos de dados. Os autores destacam que se trata de um assunto relevante uma vez que as medidas podem ser aplicadas tanto na fase de geração de regras como na filtragem e ordenação das mesmas. Os autores realizam dois tipos de análise: (i) uma para encontrar grupos de medidas similares em dados desbalanceados; (ii) uma para identificar as medidas mais apropriadas a serem utilizadas no contexto apresentado. Em relação a (i), os autores usam agrupamento baseado em grafo, assim como mineração de padrões frequentes. Em relação a (ii), os autores fazem uso do CBA para computar o desempenho do mesmo, via AUC (área sob a curva ROC), quando o processo de ordenação é realizado por cada uma das medidas individualmente. Os autores sugerem o uso de 26 MOs divididas em dois grupos: aquelas voltadas para dados extremamente desbalanceados e outras voltadas para dados levemente desbalanceados. Os autores afirmam que o estudo ajuda os usuários a decidirem pela melhor medida a ser utilizada.

Nota-se, diante dos trabalhos expostos, que o assunto é relevante e que em nenhum deles as medidas são avaliadas em conjunto, de maneira agregada, o que este trabalho se propõe a fazer. Outros trabalhos relacionados ao tema podem ser vistos em [Abdelhamid et al. 2012] e [Hernández-León et al. 2014].

3. Metodologia Proposta

Essa seção descreve a metodologia proposta visando analisar a agregação de MOs em classificadores associativos. A Figura 1 apresenta o fluxo da metodologia adotada, a qual consiste em:

Selecionar Classificador Associativo ([A]). É necessário selecionar o algoritmo ao qual as MOs agregadas serão exploradas. Embora diversos algoritmos possam ser utilizados, este trabalho, visando um estudo inicial, optou por selecionar o CBA (Algoritmo 1, Seção 2.1) por ser geralmente o utilizado como baseline nos trabalhos da literatura.

Adaptar Classificador Associativo: MOs Agregadas, Executá-lo e Computar Acurácia ([B]). Considerando o algoritmo selecionado, é necessário identificar o trecho em que as regras são ordenadas. Em seguida, é necessário substituir o critério de ordenação pelas MOs agregadas. Como neste trabalho adotou-se o CBA como baseline, a alteração realizada ocorre na linha 2 do Algoritmo 1 (linha em destaque). Assim como na etapa [A], embora diversas abordagens de agregação de MOs possam ser utilizadas, este trabalho, visando um estudo inicial, optou por selecionar a abordagem de [Bouker et al. 2014] (Seção 2.2) por ser determinística e dependente apenas de informações contidas nos dados. Desse modo, a função $Ordena(R)$ é modificada para retornar as regras ordenadas segundo a abordagem de [Bouker et al. 2014]. Assim, uma regra r_i precede uma regra r_j, em uma lista ordenada, se o valor das MOs agregadas de r_i for maior que as de r_j; em caso de empate, r_i foi gerada antes de r_j. Após alteração, o classificador associativo "agregado" é executado e a acurácia computada.

Adaptar Classificador Associativo: MOs Individuais, Executá-lo e Computar Acurácia ([C]). Assim como na etapa [B], é necessário identificar o trecho em que as regras são ordenadas. Em seguida, é necessário substituir o critério de ordenação pelas MOs individuais. Desse modo, a alteração realizada também ocorre na linha 2 do Algoritmo 1 (linha em destaque). Assim, a função $Ordena(R)$ é modificada para retornar as regras ordenadas segundo cada medida a ser utilizada (as mesmas a serem utilizadas na etapa [B]). Nesse caso, uma regra r_i precede uma regra r_j, em uma lista ordenada, se o valor da MO individual de r_i for maior que a de r_j; em caso de empate, r_i foi gerada antes de r_j. Após alteração, o classificador associativo "individual" é executado e a acurácia computada. Essa etapa é necessária para se verificar se o resultado obtido pelo classificador "agregado" é melhor que o resultado das medidas individuais.

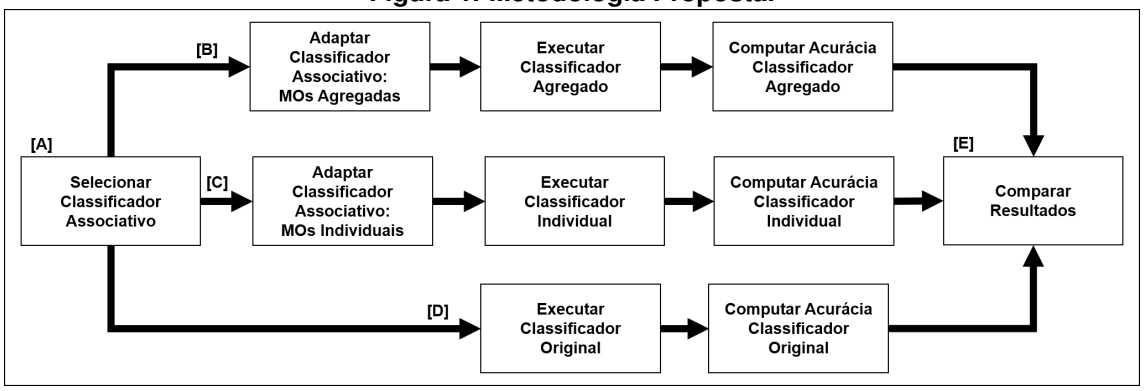
Executar Classificador Original e Computar Acurácia ([D]). Nessa etapa executá-se o classificador selecionado, neste caso, o CBA, e computa-se a acurácia. Essa etapa é necessária para se verificar se o resultado obtido pelo classificador "agregado" é melhor que o resultado tradicional.

Comparar Resultados ([E]). Nessa etapa as acurácias obtidas nas etapas [B], [C] e [D] são comparadas. Espera-se que o classificador "agregado" produza os melhores resultados.

4. Experimentos

A fim de avaliar o impacto da agregação de MOs em classificadores associativos, neste caso, no CBA, experimentos foram realizados. Para tanto, apresenta-se, a seguir, os requisitos necessários para a realização dos experimentos.

Figura 1. Metodologia Proposta.



Conjuntos de Dados. Utilizou-se, neste trabalho, 10 conjuntos de dados disponíveis na UCI[3]. As características dos conjuntos de dados, desconsiderando colunas de identificação e considerando a classe, podem ser vistas na Tabela 2. Em função dos conjuntos de dados voltados para classificação associativa necessitarem de discretização, este trabalho fez uso do algoritmo proposto em [Fayyad and Irani 1993] por meio da implementação disponível no Weka[4]. Além disso, foi necessária a implementação de um *script*[5] de pré-processamento a fim de adequar o formato de entrada dos conjuntos de dados para o formato utilizado na implementação do CBA adotada disponível em [Coenen 2004]. Nesse *script* os atributos e a classe passam a ser numéricos, ordenados de forma crescente e sequencial, começando pelo número 1, sendo a última coluna a classe. Valores ausentes e atributos sem valores distintos após a discretização são ignorados pelo *script*.

Tabela 2. Conjuntos de dados utilizados.

Conjunto de Dados	Transações	Atributos	Itens Distintos	Número de Classes
Australian Credit Approval	690	15	51	2
Breast Cancer Wisconsin	699	10	38	2
Glass Identification	241	10	29	7
Heart	270	14	30	2
Hepatitis	155	20	39	2
Iris	150	5	17	3
Tic-Tac-Toe	958	10	29	2
Vehicle Silhouettes	946	19	36	4
Wine	178	14	41	3
Zoo	101	17	41	7

Medidas Objetivas Utilizadas. [Tew et al. 2014] analisaram um grande número de MOs, em uma grande variedade de conjuntos de dados, e apresentaram uma discussão sobre quais delas poderiam ser agrupadas em relação a similaridade de ordenação das regras. Desse modo, não seria coerente utilizar um grande número de MOs se há estudos que demonstram que apenas um subconjunto delas são suficientes. Assim, selecionou-se uma medida representativa de cada um dos grupos descritos em [Tew et al. 2014], a saber: *Support*, *Prevalence*, *K-Measure*, *Least Contradiction*, *Confidence*, *EII1*, *Leverage*,

[3]`https://archive.ics.uci.edu/ml`.
[4]`https://www.cs.waikato.ac.nz/ml/weka/`.
[5]`https://github.com/mgillory/arff2ac`.

DIR, *Certainty Factor*, *Odds Ratio*, *Dilated Q2*, *Added Value*, *Cosine*, *Lift*, *J-Measure*, *Recall*, *Specificity*, *Conditional Entropy* e *Coverage*. A escolha, dentro de cada grupo, foi realizada considerando o custo computacional para se computar a MO; neste caso, as de menor custo foram as escolhidas.

Especificação de Parâmetros. O suporte mínimo e a confiança mínima foram especificados, respectivamente, em 5% e 50%. Os valores foram definidos empiricamente. Além disso, os seguintes limites foram considerados na implementação do CBA adotada disponível em [Coenen 2004]: (i) quantidade máxima de itens no antecedente: 6; (ii) quantidade máxima de itemsets frequentes a serem gerados: 5.000.000; (iii) quantidade máxima de regras a serem geradas: 10.000.

Critério de Avaliação. Como descrito na metodologia (Seção 3), utilizou-se a acurácia como medida de desempenho em todas as etapas ([B], [C] e [D]). Para tanto, executou-se 10 vez o 10-fold cross-validation estratificado. Assim, o valor de acurácia apresentado representa a média das 10 execuções.

Síntese da Configuração Experimental. Considerando a metodologia apresentada (Seção 3), assim como as definições acima descritas, cada conjunto de dados foi executado em 21 configurações distintas, a saber: (1°) *CBA*, (2°) *Support*, (3°) *Prevalence*, (4°) *K-Measure*, (5°) *Least Contradiction*, (6°) *Confidence*, (7°) *EII1*, (8°) *Leverage*, (9°) *DIR*, (10°) *Certainty Factor*, (11°) *Odds Ratio*, (12°) *Dilated Q2*, (13°) *Added Value*, (14°) *Cosine*, (15°) *Lift*, (16°) *J-Measure*, (17°) *Recall*, (18°) *Specificity*, (19°) *Conditional Entropy*, (20°) *Coverage*, (21°) *MOs Agregadas* (MOs.A). A rodada de número *1* refere-se ao *CBA* tradicional (etapa [C]). As 19 rodadas subsequentes (*2-20*) ordenam as regras de acordo com a MO estipulada (etapa [B]). Por fim, a rodada de número *21* ordena as regras segundo a abordagem de [Bouker et al. 2014] (etapa [A]).

5. Resultados e Discussão

Os resultados obtidos encontram-se na Tabela 3. A primeira coluna apresenta os conjuntos de dados utilizados. A segunda coluna (CBA:1°) apresenta a acurácia observada para o CBA tradicional (1° configuração). Da terceira até a penúltima coluna encontram-se as acurácias observadas para cada classificador "individual" (configurações de 2° a 20°). Por fim, a última coluna (*MOs.A:21*) exibe a acurácia observada para o classificador "agregado". A cada conjunto de dados (linha) destaca-se o melhor resultado obtido. Para o conjunto de dados *hepatitis*, por exemplo, a melhor acurácia (*85.50%*) ocorreu com o classificador "individual" por meio da MO *EII1* (coluna 7°). Na linha "**Média**" apresenta-se a acurácia média de cada configuração em todos os conjuntos de dados. Nessa linha destaca-se também a configuração que apresentou o melhor desempenho.

Nota-se, por meio dos resultados obtidos, que:

- além de obter a maior média de acurácia dos classificadores avaliados (tradicional, "individual" e "agregado"), o "agregado" obteve uma maior acurácia em 40% dos conjuntos de dados (*heart*, *tic-tac-toe*, *wine* e *zoo*).
- avaliando a acurácia média (última linha da Tabela 3), os 3 melhores resultados foram, do melhor para o pior: MOs.A:21°, CBA:1° e *EII1* (7°). Os 3 piores resultados foram, do pior para o melhor: *Prevalence* (3°), *Conditional Entropy* (19°) e *Coverage* (20°).

Tabela 3. Resultados Obtidos.

Conjunto de Dados	**CBA:**1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°
Australian	85.93	55.51	55.51	80.61	84.97	**86.33**	85.97	86.07	86.22	86.14	85.29
Breast-C-W	96.01	77.97	73.10	95.74	88.01	95.91	**96.27**	95.92	95.75	95.92	95.99
Glass	63.79	60.84	56.42	63.75	61.42	63.19	64.32	62.83	63.34	63.60	65.80
Heart	80.48	55.56	55.59	80.11	76.07	81.19	80.33	80.33	79.78	80.41	81.30
Hepatitis	84.70	80.18	79.25	80.11	80.34	83.09	**85.50**	78.61	82.71	80.67	81.03
Iris	95.80	**96.00**	78.40	95.33	**96.00**	94.67	95.73	95.00	94.33	95.13	95.07
Tic-Tac-Toe	**100.00**	69.49	65.34	74.15	70.43	**100.00**	99.47	98.97	98.90	99.58	**100.00**
Vehicle	57.37	52.64	52.43	**62.53**	58.78	58.01	58.18	58.78	57.41	56.78	50.68
Wine	98.87	61.25	58.25	92.28	93.94	86.25	96.53	86.21	86.62	86.50	64.76
Zoo	90.75	53.65	49.04	83.69	87.45	81.13	87.85	81.70	81.96	82.72	70.58
Média	85.37	66.31	62.33	80.83	79.74	82.98	85.01	82.44	82.70	82.74	79.05

Conjunto de Dados	12°	13°	14°	15°	16°	17°	18°	19°	20°	**MOs.A:**21°
Australian	85.72	85.62	85.32	84.97	85.32	85.30	85.51	54.91	55.51	85.78
Breast-C-W	95.75	94.84	87.78	94.38	96.24	77.96	79.78	77.35	78.01	96.15
Glass	64.82	64.53	58.44	65.44	**66.14**	55.70	60.72	58.89	58.79	64.25
Heart	80.70	80.48	67.78	80.93	83.00	55.56	69.00	57.19	55.56	**83.52**
Hepatitis	82.55	82.24	81.16	81.75	83.77	80.14	81.40	76.97	79.34	85.17
Iris	**96.00**	94.93	**96.00**	94.80	**96.00**	**96.00**	**96.00**	73.13	79.53	95.80
Tic-Tac-Toe	96.55	97.15	70.75	97.27	89.27	70.97	69.40	61.75	72.33	**100.00**
Vehicle	57.43	57.52	56.53	57.93	58.20	52.24	52.92	52.38	54.33	59.16
Wine	84.75	93.11	94.39	93.10	96.79	63.31	64.60	59.11	52.75	**98.94**
Zoo	75.56	84.35	86.86	84.35	88.62	69.64	69.30	56.08	54.00	**91.70**
Média	81.98	83.48	78.50	83.49	84.33	70.68	72.86	62.78	64.01	**86.23**

- a Tabela 4 apresenta as diferenças entre a melhor acurácia observada em cada conjunto de dados e os valores observados nas configurações MOs.A:21° e CBA:1°. Os valores em destaque correspondem a menor diferença entre a melhor acurácia e a respectiva configuração. Nota-se que mesmo nos conjuntos de dados que nenhuma das duas configurações obtiveram os melhores resultados, o classificador "agregado" tende a gerar acurácias mais altas do que o CBA tradicional.

Tabela 4. Diferenças observadas entre as configurações MOs.A:21° **e CBA:**1° **em relação a melhor acurácia observada em cada conjunto de dados.**

Conjunto de Dados	MOs.A:21°	CBA:1°
Australian	0.55	**0.40**
Breast-C-W	**0.12**	0.26
Glass	**1.89**	2.35
Heart	**0.00**	3.04
Hepatitis	**0.33**	0.80
Iris	0.20	0.20
Tic-Tac-Toe	0.00	0.00
Vehicle	**3.37**	5.16
Wine	**0.00**	0.07
Zoo	**0.00**	0.95
Média	**0.65**	1.32

- 8 MOs (*Prevalence* (3°), *Leverage* (8°), *DIR* (9°), *Certainty Factor* (10°), *Added Value* (13°), *Lift* (15°), *Conditional Entropy* (19°) e *Coverage* (21°)) não obtiveram em nenhum conjunto de dados a melhor acurácia. Isso pode indicar que essas MOs podem estar influenciando negativamente a precisão do classificador "agregado".

6. Conclusão

Neste trabalho investigou-se a agregação de MOs, tendo como base o trabalho de [Bouker et al. 2014], no CBA [Liu et al. 1998]. Para tanto, uma metodologia de investigação foi proposta a fim de explorar a influência da agregação de MOs na etapa de ordenação do CBA. De acordo com os resultados apresentados, observa-se que a agregação de MOs tende a elevar a precisão dos classificadores, especificamente do CBA; porém, um estudo mais aprofundado é necessário.

Visando uma análise mais detalhada dos resultados, pode-se citar como trabalhos futuros: (i) utilizar outros algoritmos de classificação associativa; (ii) utilizar outras abordagens de agregação de MOs; (iii) utilizar outros conjuntos de dados.

Referências

[Abdelhamid et al. 2012] Abdelhamid, N., Ayesh, A., and Thabtah, F. (2012). An experimental study of three different rule ranking formulas in associative classification. In *2012 International Conference for Internet Technology and Secured Transactions*, pages 795–800.

[Abdelhamid et al. 2016] Abdelhamid, N., Jabbar, A. A., and Thabtah, F. (2016). Associative classification common research challenges. In *45th International Conference on Parallel Processing Workshops*, pages 432–437.

[Alwidian et al. 2018] Alwidian, J., Hammo, B. H., and Obeid, N. (2018). WCBA: Weighted classification based on association rules algorithm for breast cancer disease. *Applied Soft Computing*, 62:536–549.

[Azevedo and Jorge 2007] Azevedo, P. J. and Jorge, A. M. (2007). Comparing rule measures for predictive association rules. In *Machine Learning: ECML 2007*, pages 510–517.

[Bong 2014] Bong, K. K., J. M. Q. C. A. T. M. S. (2014). Selection and aggregation of interestingnes measures: A review. 59(1):146–166.

[Bouker et al. 2014] Bouker, S., Saidi, R., Yahia, S. B., and Nguifo, E. M. (2014). Mining undominated association rules through interestingness measures. *International Journal on Artificial Intelligence Tools*, 23(4):22p.

[Coenen 2004] Coenen, F. (2004). LUCS KDD implementation of CBA (Classification Based on Associations). [Online. Acesso em 05-06-2018].

[Fayyad and Irani 1993] Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Artificial Intelligence*, pages 1022–1029.

[Han et al. 2013] Han, J., Kamber, M., and Pei, J. (2013). *Data Mining: Concepts and Techniques*. 3 edition.

[Hernández-León et al. 2014] Hernández-León, R., Hernández-Palancar, J., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2014). Studying netconf in hybrid rule ordering strategies for associative classification. In *Pattern Recognition*, pages 51–60.

[Jalali-Heravi and Zaïane 2010] Jalali-Heravi, M. and Zaïane, O. R. (2010). A study on interestingness measures for associative classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1039–1046.

[Kannan 2010] Kannan, S. (2010). *An Integration of Association Rules and Classification: An Empirical Analysis*. PhD thesis, Madurai Kamaraj University.

[Liu et al. 1998] Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 80–86.

[Ma et al. 2014] Ma, B., Zhang, H., Chen, G., Zhao, Y., and Baesens, B. (2014). Investigating associative classification for software fault prediction: An experimental perspective. *International Journal of Software Engineering and Knowledge Engineering*, 24(1):61–90.

[Moreno et al. 2016] Moreno, M. N., Segrera, S., López, V. F., Muñoz, M. D., and Sánchez, A. L. (2016). Web mining based framework for solving usual problems in recommender systems. A case study for movies' recommendation. *Neurocomputing*, 176:72–80.

[Nandhini et al. 2015] Nandhini, M., Sivanandam, S. N., Rajalakshmi, M., and Sidheswaran, D. (2015). Enhancing the spam email classification accuracy using post processing techniques. 10(15):35125–35130.

[Nguyen Le et al. 2009] Nguyen Le, T. T., Huynh, H. X., and Guillet, F. (2009). Knowledge acquisition: Approaches, algorithms and applications. chapter Finding the Most Interesting Association Rules by Aggregating Objective Interestingness Measures, pages 40–49.

[Shao et al. 2017] Shao, Y., Liu, B., Li, G., and Wang, S. (2017). Software defect prediction based on class-association rules. In *2nd International Conference on Reliability Systems Engineering*, pages 1–5.

[Singh et al. 2016] Singh, J., Kamra, A., and Singh, H. (2016). Prediction of heart diseases using associative classification. In *5th International Conference on Wireless Networks and Embedded Systems*, pages 1–7.

[Tew et al. 2014] Tew, C., Giraud-Carrier, C., Tanner, K., and Burton, S. (2014). Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 28(4):1004–1045.

[Thabtah 2007] Thabtah, F. (2007). A review of associative classification mining. *Knowledge Engineering Review*, 22(1):37–65.

[Yang and Cui 2015] Yang, G. and Cui, X. (2015). A study of interestingness measures for associative classification on imbalanced data. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 141–151.

[Yang et al. 2009] Yang, G., Shimada, K., Mabu, S., and Hirasawa, K. (2009). A nonlinear model to rank association rules based on semantic similarity and genetic network programing. *IEEJ Transactions on Electrical and Electronic Engineering*, 4(2):248–256.

[Yin et al. 2018] Yin, C., Guo, Y., Yang, J., and Ren, X. (2018). A new recommendation system on the basis of consumer initiative decision based on an associative classification approach. *Industrial Management and Data Systems*, 118(1):188–203.