Prezados revisores,

Nós estamos submetendo o artigo intitulado "HCAIM: A Discretizer for the Hierarchical Classification Scenario". Ele corresponde a uma versão revisada e estendida do artigo aceito no KDMiLe 2016. As seguintes extensões foram realizadas para gerarmos esta versão do artigo:
* Nós detalhamos a descrição do método que foi utilizado como base (CAIM) para a proposta apresentada neste trabalho.
* Nós incluímos um exemplo bastante detalhado para ilustrar o funcionamento do algoritmo CAIM.
* Pseudocódigos foram incluídos no artigo com objetivo de melhorar a explicação do método proposto (HCAIM).
* Novos experimentos foram realizados com mais 8 bases de dados reais e a análise dos resultados mostrou que a utilização do método proposto continua sendo vantajosa também para esse novo conjunto de bases de dados.

Daqui em diante apresentaremos as respostas para cada um dos comentários feitos pelos revisores do KDMiLe 2016.
==========
REVIEWER 1
OVERALL EVALUATION: 2 (accept)

---------- OVERALL EVALUATION ----------
Comentário (1) O artigo apresenta uma abordagem para discretização de dados a para problemas de classificação hierárquica. A abordagem me pareceu adequada, e o artigo está bem escrito.
Resposta: OK.

Comentário (2) Creio que seria interessante incluir como baseline o método original (CAIM), considerando apenas o primeiro nível da hierarquia.
Resposta: Isso não foi realizado porque a comparação com o HCAIM ficaria injusta, dado que o método baseline estaria considerando apenas o primeiro nível da hierarquia, que teoricamente é o que apresenta o melhor desempenho preditivo. Além disso, não faria sentido compararmos a métrica F-measure com a métrica F-measure hierárquica.

Comentário (3) Também seria interessante incluir como baseline resultados sem a discretização, por exemplo, assumindo que os atributos numéricos seguem uma distribuição normal (Gaussian Naive Bayes).
Resposta: Concordamos que poderia ser uma análise interessante se o escopo do trabalho fosse outro. No entanto, o objetivo do trabalho é unicamente mostrar que, em casos onde a discretização é necessária para o processo de classificação, a utilização do método proposto é vantajosa quando comparada com as estratégias que vinham sendo utilizadas em outros trabalhos de classificação hierárquica.

Comentário (4) Seria interessante realizar um teste estatístico de comparações múltiplas, como o teste de Friedman seguido de uma comparação com o controle (método proposto).
Resposta: Como a ideia era comparar o método proposto com cada um dos métodos usados como referência, fizemos a análise estatística para cada par de métodos. Sendo assim, não vimos a necessidade de adicionar um teste estatístico de comparações múltiplas.

==========

REVIEWER 2
OVERALL EVALUATION: 3 (strong accept)

---------- OVERALL EVALUATION ----------
Comentário (1) The paper presents a new hierarchical discretisation method (possibly the first one). The ideas are clearly presented and the experimental analysis is sound. Overall, an interesting paper.
Resposta: OK.

Comentário (2) I just have a comment regarding the weight of the classes in the hierarchy. It clearly has a bias towards classes at the top of the hierarchy, since the weight will tend to be higher for those classes. What would be the effect in problems with a deeper class hierarchy?
Resposta: De fato os pesos têm como objetivo dar maior importância para as classes que estão nos primeiros níveis da hierarquia, dado que para essas classes temos mais exemplos na base de dados do que para classes de níveis inferiores da hierarquia. Desse modo, as informações associadas às classes dos primeiros níveis são mais confiáveis e, por isso, têm peso maior. Nos experimentos realizados usamos bases de dados com hierarquias de diferentes profundidades e, para todos os casos, a ponderação proposta apresentou bons resultados.

==========
REVIEWER 3
OVERALL EVALUATION: 0 (borderline paper)

---------- OVERALL EVALUATION ----------
Comentário (1) No artigo apresenta-se uma variação do método de discretização CAIM para o problema de classificação hierárquica. A proposta é comparada com o uso de dois métodos não supervisionados de discretização sobre nove bases de dados da área de bioinformática. Uma comparação de um método que leva em conta a hierarquia e é supervisionado contra métodos não supervisionados que não consideram a hierarquia pode ser questionado; por que não comparar então contra métodos não supervisionados que operam localmente, neste caso a comparação é mais justa.
Resposta: Essa avaliação comparativa com estratégias locais daria um excelente trabalho futuro, mas não corresponde ao objetivo deste trabalho. Neste trabalho propusemos um discretizador para ser utilizado com classificadores hierárquicos globais, os quais em diversas situações já se mostraram superiores a estratégias locais.

Comentário (2) É importante indicar não somente o desempenho em termos de hF mas também de precision e recall hierárquicos.
Resposta: As medidas hP e hR foram adicionadas na Tabela 2.

Um Método de Discretização Supervisionado para o Contexto de Classificação Hierárquica

Valter Hugo Guandaline, Luiz Henrique de Campos Merschmann

Universidade Federal de Ouro Preto, Brasil
vhguandaline@gmail.com,luizhenrique@iceb.ufop.br

Abstract. A discretização é uma das etapas do pré-processamento de dados que tem sido objeto de pesquisas em diversos trabalhos relacionados com classificação plana. Apesar da importância da discretização de dados para a tarefa de classificação, até onde se tem conhecimento, para o cenário de classificação hierárquica, onde as classes a serem preditas estão organizadas de acordo com uma hierarquia, não existem na literatura métodos de discretização que levem em consideração a hierarquia de classes. O desenvolvimento de métodos de discretização capazes de lidar com a hierarquia de classes é de fundamental importância para viabilizar a utilização de classificadores hierárquicos globais que necessitem de dados discretizados. Portanto, neste trabalho, preenchemos essa lacuna propondo e avaliando um método de discretização supervisionado para o contexto de classificação hierárquica. Experimentos realizados com nove bases de dados de bioinformática utilizando um classificador hierárquico global mostraram que o método proposto permitiu ao classificador alcançar desempenho preditivo superior àqueles obtidos quando outros métodos de discretização não supervisionados foram utilizados.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Discretization, hierarchical classification, CAIM.

1. INTRODUÇÃO

A mineração de dados é apenas uma das etapas de um processo maior denominado Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database – KDD*), que também inclui o pré-processamento dos dados e o pós-processamento da informação minerada [Fayyad et al. 1996].

O principal objetivo da etapa de pré-processamento é preparar o conjunto de dados para que ele possa ser utilizado por alguma técnica de mineração de dados. Um dos processos que podem ser realizados nesta etapa é a discretização. O seu objetivo é transformar atributos contínuos em atributos discretos. Essa transformação é feita associando intervalos de valores contínuos à novos valores categóricos. Assim, os mêtodos de discretização reduzem e simplificam os dados, tornando o aprendizado mais rápido e os resultados mais compactos [Garcia et al. 2013].

A classificação é uma das principais tarefas de mineração de dados. O seu objetivo é, a partir de uma base de dados contendo instâncias com características e classes conhecidas, gerar modelos capazes de predizer a classe de novas instâncias a partir de suas características. A maioria dos problemas de classificação abordados na literatura são considerados problemas de classificação plana, onde as classes não possuem relacionamentos entre si. No entanto, existem problemas de classificação mais complexos, conhecidos como problemas de classificação hierárquica, onde as classes a serem preditas estão estruturadas de acordo com uma hierarquia [Freitas and de Carvalho 2007].

Apesar de as aplicações do mundo real geralmente envolverem atributos contínuos, alguns algo-

ritmos de classificação lidam somente com atributos discretos. Além disso, para alguns métodos de classificação, ainda que sejam capazes de lidar com os atributos contínuos, o seu desempenho preditivo melhora quando os atributos contínuos são previamente discretizados [Kurgan and Cios 2004].

Até onde se tem conhecimento, não existem na literatura métodos de discretização que levem em consideração os relacionamentos entre classes existentes em problemas de classificação hierárquica. Trabalhos que abordaram problemas de classificação hierárquica e necessitaram realizar a discretização dos dados, tais como [Merschmann and Freitas 2013] e [Silla Jr and Freitas 2009], tiveram que utilizar métodos de discretização não supervisionados, uma vez que métodos não supervisionados podem ser utilizados para o contexto de classificação plana ou hierárquica.

A hipótese levantada neste trabalho é que métodos de discretização supervisionados, pelo fato de levarem em consideração o atributo classe no momento da discretização, poderiam proporcionar melhoria no desempenho preditivo de classificadores hierárquicos. Isso nos motivou a propor um método de discretização supervisionado para o contexto da classificação hierárquica.

A proposta aqui apresentada corresponde a uma adaptação realizada no método CAIM [Kurgan and Cios 2004] para torná-lo capaz de considerar a hierarquia de classes existente em problemas de classificação hierárquica. Os resultados mostram que o método proposto permitiu a um classificador hierárquico alcançar desempenho preditivo superior àqueles alcançados quando a base de dados foi pré-processada por métodos não supervisionados.

O restante deste artigo encontra-se organizado como descrito a seguir. A Seção 2 apresenta uma breve revisão da literatura sobre classificação hierárquica e discretização de dados. Em seguida, o método proposto é detalhado na Seção 3 e os experimentos computacionais com os resultados obtidos são descritos na Seção 4. Por fim, a Seção 5 apresenta as conclusões deste trabalho.

2. REFERENCIAL TEÓRICO

2.1 Classificação Hierárquica

Em um problema de classificação hierárquica, os relacionamentos entre as classes são representados por uma estrutura hierárquica, que pode ser uma árvore ou um grafo acíclico direcionado (*Directed Acyclic Graph – DAG*). A principal diferença entre essas estruturas é que, enquanto em uma árvore um nó (classe) está associado a no máximo um nó (classe) pai, em um DAG um nó pode ter mais do que um nó pai.

De acordo com [Freitas and de Carvalho 2007], os métodos de classificação hierárquica diferem em uma série de aspectos. O primeiro aspecto refere-se ao tipo de estrutura com a qual o método é capaz de lidar. No caso deste trabalho, a estrutura hierárquica das classes corresponde a uma árvore.

O segundo aspecto está relacionado à profundidade da execução da classificação na hierarquia. Um método pode realizar predições usando somente classes dos nós folha da hierarquia (*Mandatory Leaf-Node Prediction – MLNP*) ou classes referentes a qualquer nó (interno ou folha) da hierarquia (*Non-Mandatory Leaf-Node Prediction – NMLNP*). Neste trabalho, considera-se o cenário *NMLNP*.

O terceiro aspecto está relacionado ao número de classes (ramos da estrutura hierárquica) que um método é capaz de atribuir a uma instância. Um método pode ser capaz de predizer múltiplas classes para uma determinada instância (multirrótulo), desse modo envolvendo múltiplos ramos da hierarquia de classes (*multiple paths of labels*), ou somente uma classe (monorrótulo), a qual estará vinculada a um único ramo da hierarquia de classes (*single path of labels*). O método proposto neste trabalho lida com a classificação monorrótulo.

Por fim, o quarto aspecto está relacionado ao tipo de abordagem que o classificador utiliza para explorar a estrutura hierárquica. Segundo [Silla Jr and Freitas 2011] existem três tipos de abordagens: (i) abordagem por classificação plana, na qual a hierarquia de classes é ignorada e as predições são

realizadas considerando-se somente as classes dos nós folha da estrutura hierárquica; (ii) abordagem local, onde são utilizados diversos classificadores planos tradicionais, cada um com uma visão local da estrutura hierárquica; (iii) abordagem global, onde um único modelo de classificação é construído levando em consideração toda a hierarquia de classes de uma só vez. O método de discretização proposto neste trabalho tem como objetivo adequar as bases de dados para serem utilizadas por classificadores hierárquicos globais, dado que para a abordagem local, podem ser utilizados métodos de discretização supervisionados projetados para o cenário de classificação plana.

2.2 Discretização de Dados

Discretização é uma estratégia de redução de dados amplamente utilizada na etapa de pré-processamento dos dados [Garcia et al. 2013]. O processo de discretização transforma atributos contínuos em atributos discretos dividindo o atributo contínuo em intervalos de valores e associando cada um desses intervalos a um valor discreto.

De acordo com [Garcia et al. 2013], os métodos de discretização podem ser categorizados como supervisionados ou não supervisionados. O método é denominado supervisionado quando ele leva em consideração os valores do atributo classe. Por outro lado, se o atributo classe não é considerado no processo de discretização, o método é dito não supervisionado.

Diferentes critérios podem ser usados para avaliar os algoritmos de discretização, tais como o número de intervalos gerados, o nível de inconsistência e a taxa de acerto de classificadores. Neste trabalho, os métodos de discretização foram avaliados a partir do método de classificação hierárquica global denominado *Global Model Naive Bayes – GMNB*, proposto em [Silla Jr and Freitas 2009].

Em [Garcia et al. 2013] os autores avaliaram 30 discretizadores sobre 40 bases de dados utilizando 6 classificadores planos. Essa avaliação mostrou que o CAIM foi um dos métodos de discretização mais eficientes. Por isso, esse foi o método escolhido neste trabalho para ser adaptado para o contexto hierárquico. Além disso, dada a inexistência de métodos supervisionados para o contexto hierárquico, os métodos não supervisionados *EqualWidth* e *EqualFrequency* (adotado em [Merschmann and Freitas 2013] e [Silla Jr and Freitas 2009]) foram utilizados como base de comparação com o método aqui proposto. A seguir, serão apresentados mais detalhes do método de discretização (CAIM) que foi adaptado para o contexto de classificação hierárquica.

2.2.1 *CAIM. Class-Attribute Interdependency Maximization*

é um método de discretização supervisionado que independe de outros métodos de aprendizagem. Ele utiliza uma métrica para avaliar a interdependência entre o atributo classe e o atributo em processo de discretização [Kurgan and Cios 2004].

Considere uma base de dados com um conjunto de instâncias M, um conjunto de atributos numéricos F e atributo classe S, onde $|M|$, $|F|$ e $|S|$ são, respectivamente, o número de instâncias, número de atributos e o número de classes. Além disso, cada instância M_k está associada à uma classe S_i, onde $k \in \{1, 2, ..., |M|\}$ e $i \in \{1, 2, \ldots, |S|\}$.

Para cada atributo contínuo F_j, $j \in \{1, 2, \ldots, |F|\}$, o CAIM ordena os seus valores em ordem crescente e, em seguida, divide-os em n intervalos da seguinte forma: $D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{n-1}, d_n]\}$, onde d_0 e d_n são, respectivamente, os valores mínimo e máximo do atributo F_j e $d_i < d_{i+1}$ para $i \in \{0, 1, \ldots, n-1\}$. Cada par de valores (d_i, d_{i+1}) define um intervalo do atributo F_j, sendo que o resultado da discretização D, chamado de esquema de discretização do atributo F_j, define o seguinte conjunto de pontos de corte $P = \{d_1, d_2, \ldots, d_{n-1}\}$.

A interdependência entre o atributo classe S e um esquema de discretização D de um atributo F_j é calculada utilizando-se a métrica CAIM (Equação 1), que faz uso de uma matriz de frequência denominada matriz de contingência, apresentada na Figura 1. Nessa matriz, considerando-se um esquema de discretização $D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{n-1}, d_n]\}$ do atributo em processo de discretização, q_{ir}

é a quantidade de instâncias pertencentes à i-ésima classe que estão contidas no r-ésimo intervalo, M_{i+} é a quantidade total de instâncias pertencentes à i-ésima classe e M_{+r} é a quantidade total de instâncias contidas no r-ésimo intervalo.

$$CAIM(S, D|F_j) = \frac{\sum_{r=1}^{n} \frac{max_r^2}{M_{+r}}}{n}, \quad (1)$$

onde n é o número de intervalos e max_r é o número máximo de instâncias contidas no intervalo r pertencentes a uma mesma classe. Essa equação é utilizada pelo método CAIM para se escolher o melhor ponto de corte a ser inserido em um determinado esquema de discretização. Quanto maior o valor retornado por essa métrica, maior é dependência entre o atributo F_j (discretizado segundo o esquema D) e o atributo classe S.

Classes	Intervalos					Instâncias por Classe
	$[d_0, d_1]$	...	$(d_{r-1}, d_r]$	...	$(d_{n-1}, d_n]$	
C_1	q_{11}	...	q_{1r}	...	q_{1n}	M_{1+}
...	...	...	...	...	...	...
C_i	q_{i1}	...	q_{ir}	...	q_{in}	M_{i+}
...	...	...	...	...	...	...
C_s	q_{s1}	...	q_{sr}	...	q_{sn}	M_{s+}
Instâncias por Intervalo	M_{+1}	...	M_{+r}	...	M_{+n}	M

Fig. 1. Matriz de contingência para o atributo F_j e esquema de discretização D

3. MÉTODO PROPOSTO

O principal problema na utilização de métodos de discretização supervisionados tradicionais (utilizados em conjunto com classificadores planos) para bases de dados relacionadas com o contexto de classificação hierárquica está no fato de esses discretizadores não serem capazes de considerar as informações dos relacionamentos entre as classes do problema. Neste trabalho, parte-se da hipótese de que esse tipo de informação, se considerada ao longo do processo de discretização, pode contribuir na geração de uma base de dados discretizada de melhor qualidade para a tarefa de classificação.

Portanto, o método de discretização aqui proposto, denominado HCAIM (Hierarchical CAIM), considera a hierarquia de classes enquanto realiza o processo de discretização. O HCAIM corresponde a uma adaptação do método de discretização CAIM para o contexto hierárquico, cuja principal diferença está na métrica de avaliação utilizada pelo método para a definição dos pontos de corte de um esquema de discretização.

3.1 Métrica de avaliação

Para avaliar um esquema de discretização $D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{n-1}, d_n]\}$ para um atributo F_j, o método CAIM verifica o quão bons são os intervalos contidos nesse esquema. Por meio da métrica também denominada CAIM (ver Equação 1), cada intervalo contido em D é avaliado medindo-se a correlação entre os valores do atributo F_j existentes naquele intervalo e as classes nele contidas. Essa correlação é dada por $(\frac{max_r^2}{M_{r+}})$, onde max_r é o número de ocorrências da classe mais frequente no intervalo r e M_{r+} é a quantidade de instâncias contidas nesse mesmo intervalo. Esse cálculo permite ao método CAIM: (i) considerar o grau de pureza do intervalo (quanto mais próximo max_r for de M_{r+}, mais puro é o intervalo) e (ii) priorizar intervalos com maior número de instâncias.

Entretanto, essa métrica CAIM não leva em consideração a hierarquia de classes existente em um problema onde as classes estão hierarquicamente organizadas. Por exemplo, dado um intervalo r contendo 9 instâncias ($M_{r+} = 9$), sendo 3 instâncias da classe *R.2* e 6 da classe *R.2.1*, a avaliação

desse intervalo segundo a métrica CAIM é dada por $6^2/9 = 4$, uma vez que a classe majoritária $R.2.1$ é considerada como completamente distinta da $R.2$. No entanto, no contexto hierárquico, instâncias da classe *R.2.1* também pertencem à classe *R.2*, já que $R.2.1$ é uma classe filha da $R.2$.

Portanto, neste trabalho, a métrica CAIM foi adaptada para calcular o grau de pureza de cada intervalo considerando a hierarquia de classes. Na adaptação proposta, denominada HCAIM (*Hierarchical CAIM*), o cálculo do grau de pureza de um intervalo é realizado para cada nível hierárquico, sendo o seu valor final uma média ponderada dos valores calculados para cada um dos níveis. Desse modo, a dependência entre o atributo classe S e o esquema de discretização D para um dado atributo F_j levando em consideração a hierarquia de classes é dada por:

$$HCAIM(S, D|F_j) = \frac{\sum_{r=1}^{n} \sum_{l=1}^{H_r} \frac{max_{r,l}^2}{M_{+r}} . W_{l,r}}{n}, \quad (2)$$

onde n é o número de intervalos, H_r é a profundidade da hierarquia de classes referente ao intervalo r, $max_{r,l}$ é o número de ocorrências da classe mais frequente no intervalo r considerando-se a hierarquia de classes até o nível l, M_{+r} é o total de instâncias contidas no intervalo r e $W_{l,r}$ é peso associado ao nível l da hierarquia de classes referente ao intervalo r.

No cálculo da métrica HCAIM para um determinado intervalo r, além das matrizes de contingência para cada nível hierárquico, há necessidade de se calcular os pesos $W_{l,r}$ que serão aplicados de acordo com o nível hierárquico l e a profundidade da hierarquia naquele intervalo H_r. O valor do peso para cada um dos níveis hierárquicos é dado por $W(l, r) = (H_r - l + 1)\frac{2}{H_r \times (H_r+1)}$, sendo que $\sum_{l=1}^{H_r} W_{l,r} = 1$.

Retomando o exemplo anterior, onde consideramos um único intervalo r contendo 9 instâncias ($M_{r+} = 9$), sendo 3 instâncias da classe *R.2* e 6 da classe *R.2.1*, a avaliação desse intervalo segundo a métrica HCAIM é dada por $9^2/9 \times W_{1,r} + 6^2/9 \times W_{2,r}$. A primeira parcela ($9^2/9$) deve-se ao fato de considerarmos todas as classes presentes no intervalo r somente até o primeiro nível da hierarquia, ou seja, todas as instância estão associadas à classe $R.2$. Já a segunda parcela ($6^2/9$) é calculada considerando-se todas as classes até o segundo nível hierárquico, onde passamos a ter 3 instâncias da classe *R.2* e 6 da classe *R.2.1*. Considerando que os pesos associados aos níveis hierárquicos são $W_{1,r} = 2/3$ e $W_{2,r} = 1/3$, o valor final da métrica para o exemplo em questão é $HCAIM = 7,33$.

3.2 Algoritmo HCAIM

O algoritmo HCAIM pode ser divido em três etapas: Inicialização, Avaliação e Verificação. Essas etapas são aplicadas a cada atributo contínuo F_j. A seguir, tem-se o detalhamento de cada uma delas.

Inicialização: a primeira inicialização é a do conjunto dos possíveis pontos de corte B. Considerando-se que o atributo a ser discretizado F_j encontra-se ordenado, os pontos de corte inseridos no conjunto B correspondem à média dos valores do atributo F_j para cada par de instâncias vizinhas que encontram-se associadas a classes diferentes e possuem valores distintos para o atributo em questão. Em seguida, o esquema de discretização D é inicializado com um único intervalo, $D = \{[-\infty, +\infty]\}$. Por fim, as variáveis $GlobalHCaim$ (armazena os melhores valores da métrica HCAIM ao longo do processo de discretização) e k (número de intervalos no esquema D) são inicializadas com os valores 0 e 1, respectivamente. Após isso, o método passa para a etapa seguinte (etapa de avaliação).

Avaliação: é uma etapa iterativa que consiste em avaliar todos os pontos de corte contidos em B enquanto o critério de parada não for satisfeito. Para cada possível ponto de corte p do conjunto B, o método cria um novo esquema de discretização D' inserindo o ponto de corte p no esquema de discretização D. Em seguida, o esquema D' é avaliado pela métrica HCAIM(S, $D'|F_j$). Após avaliar todos os possíveis pontos de corte contidos em B, o método armazena o ponto de corte $p*$ que obteve o maior valor para o critério de avaliação HCAIM. Essa informação é utilizada na etapa seguinte (etapa de verificação).

Verificação: nessa etapa é realizada a verificação do critério de parada. O algoritmo encerra a sua execução quando as duas condições a seguir são falsas: i) se o número de intervalos gerados até o momento (k) é menor do que o número de classes ($|S|$); ii) se o valor da métrica HCAIM para o ponto de corte $p*$ é maior do que o obtido na iteração anterior ($GlobalHCaim$). Caso o contrário, o algoritmo remove o ponto de corte $p*$ do conjunto B e adiciona-o ao esquema D, incrementa o número de intervalos criados ($k = k + 1$), atualiza o valor da métrica HCAIM para o esquema D ($GlobalHCaim = HCAIM$) e, por fim, volta para a etapa de avaliação, onde a inserção de novos pontos de corte será avaliada.

4. EXPERIMENTOS COMPUTACIONAIS

4.1 Bases de Dados

Todos os experimentos foram conduzidos a partir de nove bases de dados relacionadas com a classificação de funções de genes. Nessas bases, os atributos preditores incluem diversos tipos de dados da área de bioinformática, tais como: estrutura secundária da sequência, fenótipo, homologia, estatísticas da sequência e outros. Essas bases de dados, inicialmente utilizadas em [Clare and King 2003], são multirrótulo. Como o foco deste trabalho é a classificação hierárquica monorrótulo, essas bases de dados foram transformadas para o contexto monorrótulo selecionando-se, para cada instância, a classe mais frequente na base de dados original.

A partir das bases de dados monorrótulo, o pré-processamento descrito a seguir foi realizado para substituição dos valores ausentes de atributos. Ao identificar um valor ausente para um determinado atributo F_j de uma instância associada à classe C_i, calcula-se a média dos valores conhecidos do atributo F_j de todas as demais instâncias da base associadas à classe C_i e, em seguida, utiliza-se essa média para substituir o valor ausente. Se para a classe C_i nenhuma instância possuir valor conhecido para o atributo F_j, calcula-se a média dos valores conhecidos do atributo F_j de todas as instâncias da base associadas a classes descendentes de C_i na hierarquia e, em seguida, utiliza-se essa média para substituição do valor ausente. Em último caso, se a classe C_i não possuir classes descendentes ou se para as classes descendentes de C_i nenhuma instância possuir valor conhecido para o atributo F_j, então substitui-se o valor ausente utilizando-se a média global do atributo F_j.

A Tabela I mostra as principais características das bases de dados após o pré-processamento. Essa tabela apresenta, para cada base de dados, o seu número de instâncias, de atributos preditores, de classes e a distribuição das mesmas pelos níveis da hierarquia (1°|2°|3°|4°|5°|6°).

4.2 Configuração Experimental

Os métodos de discretização não supervisionados *EqualFrequency* e *EqualWidth* foram utilizados como referência para comparação com método proposto, o HCAIM. Esses métodos foram escolhidos para

Table I. Características das bases de dados

Bases	# Instâncias	# Atributos Contínuos / Categóricos	# Classes	# Classes por Nível
Church	3755	26 / 1	190	7|37|72|47|25|2
Eisen	2424	79 / 0	143	4|26|55|34|22|2
Cellcycle	3757	77 / 0	190	7|37|73|46|25|2
Gasch2	3779	52 / 0	191	7|37|73|46|26|2
Gasch1	3764	173 / 0	191	7|37|73|46|26|2
Derisi	3725	63 / 0	190	7|37|72|47|25|2
Spo	3703	77 / 3	191	7|37|73|46|26|2
Seq	3919	473 / 5	192	7|37|73|47|26|2
Expr	3779	547 / 4	191	7|37|72|47|26|2

as comparações pelo fato de já terem sido adotados em trabalhos de classificação hierárquica, uma vez que não há métodos de discretização supervisionados para esse contexto.

Os métodos *EqualFrequency* e *EqualWidth* possuem um parâmetro k, que define o número de intervalos a serem criados no processo de discretização. Os experimentos foram executados para diferentes valores de k, a saber, 5, 10, 15 e 20. Esses métodos foram executados a partir das suas implementações disponíveis na ferramenta *WEKA* [Hall et al. 2009].

Para avaliar a qualidade da discretização realizada por cada um dos métodos foi utilizado o classificador hierárquico *Global Model Naive Bayes – GMNB* [Silla Jr and Freitas 2009]. Para expressar o desempenho preditivo do classificador hierárquico *GMNB* adotou-se a métrica *F-measure* hierárquica (hF) posposta em [Kiritchenko et al. 2005]. Além disso, o método k-validação cruzada (k=10) foi utilizado na avaliação do desempenho do *GMNB*. Vale ressaltar também que a discretização dos dados ocorreu somente após o particionamento da base pelo método 10-validação cruzada, ou seja, para cada base, ela foi aplicada considerando-se cada uma das 10 partições de treinamento.

4.3 Resultados

A Tabela II apresenta o hF médio (com desvio padrão entre parênteses) obtido pelo classificador *GMNB* para cada base de dados discretizada utilizando-se o HCAIM e os outros dois métodos utilizados como referência, a saber, *EqualFrequency* (EF) e *EqualWidth* (EW). No caso dos métodos de discretização EF e EW, o nome da coluna é formado pelo nome do método acrescido, entre parênteses, do valor do parâmetro k utilizado. Para cada base de dados, com objetivo de verificar se há diferença com significância estatística entre os desempenhos (hF) do classificador *GMNB* ao processar a base de dados discretizada pelo HCAIM e por outro método de referência, utilizou-se o teste estatístico de Wilcoxon com a correção de Bonferroni devido às múltipas comparações entre o HCAIM e cada método de referência [Japkowicz and Shah 2011]. Esse teste estatístico foi executado com nível de confiança de 95%. Os valores em negrito indicam o melhor resultado obtido para cada base de dados. Além disso, o símbolo • indica que há diferença com significância estatística entre o método de referência em questão e o HCAIM. Por fim, a última linha dessa tabela resume o resultado do teste estatístico, ou seja, para cada método de referência, mostra-se o número de vezes em que o HCAIM o superou apresentando um melhor desempenho preditivo do classificador *GMNB*.

Os resultados apresentados na Tabela II mostram que o método de discretização proposto neste trabalho (HCAIM) proporcionou o maior desempenho preditivo ao GMNB para 6 das 9 bases de dados utilizadas nos experimentos (valores em negrito). Além disso, os testes estatísticos revelam que,

Table II. Valores médios de hF obtidos pelo GMNB após discretização das bases.

Base	EF(5)	EF(10)	EF(15)	EF(20)	EW(5)	EW(10)	EW(15)	EW(20)	HCAIM
Cellcycle	20.83 • (1.39)	24.56 • (2.69)	26.03 • (2.21)	26.56 • (2.13)	15.41 • (1.75)	17.08 • (1.37)	18.96 • (1.63)	19.10 • (2.03)	**31.85** (**1.69**)
Church	10.07 • (1.25)	11.57 • (1.31)	12.00 • (1.63)	13.14 • (1.45)	10.90 • (1.01)	11.85 • (1.51)	11.76 • (1.41)	12.63 • (1.58)	**18.63** (**1.13**)
Derisi	9.36 • (1.00)	10.98 (1.20)	11.73 (1.07)	11.52 (1.07)	8.92 • (0.75)	9.31 • (1.28)	9.69 (1.30)	9.91 • 1.14)	**12.42** (**0.82**)
Eisen	**22.56** (**1.33**)	22.52 (2.10)	21.86 (2.23)	21.78 (1.70)	20.74 (2.35)	22.17 (1.42)	22.40 (1.27)	22.49 (1.97)	21.28 (1.43)
Expr	43.91 • (1.39)	45.82 (1.88)	45.67 (2.35)	45.54 (2.49)	26.82 • (1.69)	29.62 • (1.88)	32.60• (1.49)	34.46 • (1.27)	**46.41** (**1.66**)
Gasch1	18.64 • (1.97)	21.75 • (2.33)	22.39 • (1.51)	22.84 • (1.75)	16.80 • (1.47)	18.37 • (1.61)	19.62 • (2.27)	19.36 • (2.06)	**26.86** (**1.16**)
Gasch2	16.48 • (1.70)	17.59 • (1.45)	19.37 • (1.90)	19.69 • (1.68)	14.75 • (1.17)	16.11 • (1.47)	15.77 • (1.32)	16.61 • (1.47)	**25.51** (**1.74**)
SPO	13.62 (1.41)	14.31 (1.25)	**14.84** (**1.37**)	14.30 (0.78)	13.77 (1.43)	13.17 (1.19)	13.02 (1.63)	13.62 (0.85)	13.14 (1.73)
Seq	21.39 • (0.87)	19.58 (1.03)	18.83 (1.32)	18.75 (1.15)	24.02 • (1.67)	**24.91** • (**1.47**)	24.05 • (1.11)	24.05 • (1.26)	18.10 (1.08)
# Vitórias do HCAIM	6	4	4	4	6	6	5	6	

para quatro base de dados (Cellcycle, Church,Gasch1 e Gasch2), o HCAIM superou todos os métodos de referência utilizados na avaliação comparativa. Para apenas uma base de dados (Seq) o HCAIM obteve desempenho inferior, com significância estatística, ao de alguns métodos de referência.

Os testes estatísticos também mostram que na comparação do HCAIM com cada um dos outros métodos utilizados nos experimentos, ele apresenta um desempenho estatisticamente superior ou igual ao dos demais métodos para a maioria das bases de dados avaliadas. Por exemplo, quando comparado com o método *EqualFrequency* com $k = 5$ (EF(5)), o HCAIM é superior para 6 bases de dados e equivalente para as 3 bases restantes.

5. CONCLUSÃO

Apesar da importância dos métodos de discretização para o pré-processamento das bases de dados utilizadas por técnicas de classificação, até onde se tem conhecimento, não existem na literatura propostas de métodos de discretização supervisionados que possam ser utilizados em conjunto com classificadores hierárquicos globais. Portanto, este trabalho propôs um método de discretização supervisionado para o contexto de classificação hierárquica monorrótulo. A proposta apresentada, denominada HCAIM, corresponde a uma adaptação do método de discretização supervisionado CAIM.

Os experimentos computacionais realizados mostraram que o método HCAIM, para a maioria das bases avaliadas, permitiu ao classificador hierárquico GMNB alcançar desempenho preditivo superior àqueles alcançados quando as bases de dados foram pré-processadas pelos métodos não supervisionados *EqualWidth* e *EqualFrequency*. Esse resultado confirma o potencial de aplicação do método proposto para a realização da discretização de bases utilizadas em trabalhos de classificação hierárquica.

AGRADECIMENTOS

Os autores agradecem a UFOP, a FAPEMIG e o CNPq pelo apoio financeiro concedido.

REFERENCES

Clare, A. and King, R. D. Predicting gene function in saccharomyces cerevisiae. *Bioinformatics* 19 (suppl 2): ii42–ii49, 2003.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37, 1996.

Freitas, A. A. and de Carvalho, A. C. A tutorial on hierarchical classification with applications in bioinformatics. In *D. Taniar (Ed.) Research and Trends in Data Mining Technologies and Applications*. Idea Group, pp. 175–208, 2007.

Garcia, S., Luengo, J., Sáez, J. A., López, V., and Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 734–750, 2013.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: An update. *SIGKDD Explorations* 11 (1), 2009.

Japkowicz, N. and Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, New York, NY, USA, 2011.

Kiritchenko, S., Matwin, S., and Famili, A. F. Functional annotation of genes using hierarchical text categorization. In *Proceedings of the ACL workshop on linking biological literature, ontologies and databases: mining biological semantics*, 2005.

Kurgan, L. A. and Cios, K. J. Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering* 16 (2): 145–153, 2004.

Merschmann, L. H. C. and Freitas, A. A. An extended local hierarchical classifier for prediction of protein and gene functions. In *Data Warehousing and Knowledge Discovery*. pp. 159–171, 2013.

Silla Jr, C. N. and Freitas, A. A. A global-model naive bayes approach to the hierarchical prediction of protein functions. In *IEEE International Conference on Data Mining*. pp. 992–997, 2009.

Silla Jr, C. N. and Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22 (1-2): 31–72, 2011.