A Relevance Measure for Multivalued Attributes

Mariana Tasca[1], Bianca Zadrozny[2], Alexandre Plastino[1]

[1] Universidade Federal Fluminense, Brazil
{mlobo, plastino}@ic.uff.br
[2] IBM Research, Brazil
biancaz@br.ibm.com

Abstract. An important step in the knowledge discovery in databases (KDD) process is the attribute selection procedure, which aims at choosing a subset of attributes that can represent the important information within the data. Most of the existing attribute selection methods can only handle simple attribute types, such as categorical and numerical. In particular, these methods cannot be applied to multivalued attributes, which are attributes that take multiple values simultaneously for the same instance in the dataset. In many real datasets, however, multivalued attributes are present, e.g., the types of books owned by a person may be represented by a multivalued attribute. This article proposes a relevance measure for multivalued attributes, which aims at measuring their importance for classification. The proposed measure takes into account the ability that the attribute has for determining the instance class. In order to evaluate the proposed measure, experiments were conducted with several datasets submitted to multi-relational classifiers. The experiments show that the resulting accuracy values follow, in most cases, the values of the proposed relevance measure. This is an evidence that the proposed measure can be a good indicator of the relevance of multivalued attributes for classification.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining

Keywords: attribute selection, classification, multi-relational data mining, multivalued attributes, relevance measures

1. INTRODUCTION

One of the most studied and applied task in data mining is the classification task, which aims at estimating the class of an instance based on the available set of attributes. One method to improve the performance of the classification process is to perform an attribute selection procedure, discarding those attributes that do not contribute and may even harm the performance of the task. The attribute selection procedure is a step in the data mining process, which aims at choosing a subset of attributes that can represent the important information within the data, based on some criteria [Liu and Motoda 1998]. The use of this procedure is strongly recommended, especially if the dataset has a huge dimensionality, because most of the data mining algorithms may require a large computational effort if a large number of attributes is used. The use of an attribute selection procedure can provide: (a) improvement in the performance of the classifiers, eliminating useless attributes and those that can deteriorate the results, (b) simpler classification models, reducing the computational cost of executing this models and providing a better understanding of the obtained results, (c) smaller datasets.

There are several attribute selection techniques available in the literature, some of them based on relevance measures (filter), others based on the use of the classifier itself (wrapper) and also those that are coded inside the learning algorithm (embedded) [Liu and Motoda 2008].

Given the context of conventional data mining, where the dataset under investigation is represented by a single table or a sequential file, most feature selection algorithms and relevance measures available

The development of this work was partially supported by CAPES, CNPq and FAPERJ.

in the literature take into consideration only simple attributes, such as categorical and numerical. However, many real datasets have multivalued attributes, which are characterized by being able to assume multiple values simultaneously. Examples of multivalued attributes are: the types of books a person reads (which could be "children's" and "romance" or "science fiction" and "mystery"), the research areas of a professor, etc. This type of attribute may contribute or not to the classification task, depending on the target application domain. For example, knowing which types of books a person buys (multivalued attribute) can be important to find out if this person has children or not, but it may not bring any useful information about the usage of a credit card. Thus, it is important to deal with this type of attribute so one can assess its relevance for classification.

We have not found in the literature any work that specifically addresses multivalued attributes selection or any measure to determine their importance. Hall and Holmes [2003] use a technique that transforms the k possible values of a multivalued attribute in k binary attributes, allowing the use of conventional attribute selection algorithms. But this technique is problematic for data mining since it increases the dimensionality of the original space.

With the aim of contributing to the multivalued attribute selection process, this article proposes a relevance measure for this type of attribute, which takes into account their ability to determine the class value. Our intention is that the proposed measure be used within feature selection algorithms. In our previous work [Tasca et al. 2009], we have presented some preliminary results in this research direction.

The evaluation of the proposed measure is based on the accuracy of a classifier available in a multi-relational data mining tool, Relational Weka[1], which is based on Weka[2], a well known data mining tool. Experiments are performed using real datasets – some of them obtained from public repositories and others obtained from IBGE – Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics) – and synthetic datasets, which were generated based on the real datasets. The following aspects are considered in the evaluation of the proposed measure: (a) analysis of the multivalued attribute quality by itself, disregarding the influence of any other attributes from the dataset, (b) analysis of the multivalued attribute contribution for the classification task, when it is combined with other attributes from the dataset, (c) analysis of the proposed measure behavior when the probability distribution of the attribute values within each class is modified and (d) analysis of the measure usefulness for comparing two multivalued attributes from the same dataset, trying to figure out which one contributes more for classification.

The research area which develops methods that can deal directly with databases composed of multiple tables is known as Relational Mining (also called multi-relational just to emphasize the use of multiple tables) [Dzeroski 2003]. Several algorithms have been proposed in the relational mining literature [Dehaspe and Toivonen 2001; Emde and Wettschereck 1996; Kramer et al. 2001; Laer and Raedt 2001]. Since the representation of multivalued attributes in databases is usually accomplished through a separate table to avoid redundancy within the main table [Elmasri and Navathe 2010], research about this type of attribute, which is the object of the present study, is therefore placed in the context of multi-relational mining.

This article is organized as follows. Section 2 contains a literature review on multi-relational classification, mainly about the k-NN algorithm and distance measures in this context. In Section 3, we present a brief review of feature selection, types of attribute evaluation measures and the proposed multivalued attribute relevance measure. In Section 4, we present the datasets used in the experiments, the algorithm used to create the synthetic multivalued attributes, the results obtained in the experiments and the evaluation of the proposed relevance measure. Finally, Section 5 presents the conclusions of this work and proposals for future work.

[1]http://cui.unige.ch/woznica/rel_weka/
[2]http://www.cs.waikato.ac.nz/ml/weka/

2. CLASSIFICATION USING MULTIVALUED ATTRIBUTES

Traditionally, data mining research addresses algorithms which extract information from datasets stored on a single table or a sequential file. However, the data model most commonly used is composed of multiple tables referenced through foreign keys [Elmasri and Navathe 2010]. To allow data mining in these relational databases, the research area named relational mining emerged. Multivalued attributes – which can assume multiple values simultaneously for the same instance – are represented in the relational model as separate tables referenced through foreign keys. So, the use of this type of attribute in data mining tasks requires the use of multi-relational techniques.

2.1 Multi-Relational Data Mining

Some authors name as propositional learning when the dataset is composed by a set of instances, each instance is represented by a fixed set of attributes, and each attribute has a unique value for a given instance. The learning process is called propositional because an instance is characterized by the conjunction of propositions of the form "attribute θ value", where θ is a relational operator, such as $<, \leq, >, \geq, =, \neq$ [Leiva 2002].

However, real datasets are usually organized using the relational data models, where multiple tables store information in a normalized form and are related through foreign keys. There are two ways to extract information from relational databases through data mining techniques: using multi-relational data mining algorithms, which can deal directly with multiple tables, or performing a propositionalization, i.e., transforming the multi-relational problem into a propositional problem [Leiva 2002]. This transformation, in turn, can be done in two ways: (a) joining the target tables, grouping all attributes in a single table, or (b) transforming the relational model into a single table by creating new attributes in the main table which summarize or aggregate information from the other tables.

Both of these alternatives may lead to problems: joining all tables may result in an extremely large table, which makes it difficult to deal with it. Furthermore, the table joints will generate a great number of redundant information, which may cause statistical problems [Emde and Wettschereck 2001]. Creating attributes in the main table with aggregated information from other tables, through operations such as sum, average or count, can lead to a significant loss of information. There is some work in the literature about categorical attribute aggregation generating less loss of information. Perlich and Provost [2006] presents a framework that performs propositionalization in relational databases using conventional numerical attribute aggregation operations (sum, average, count) and more sophisticated aggregation operations on categorical attributes.

There are also several proposals in the multi-relational data mining area which can deal with tables represented in their original data model. Most of them are related with inductive logic programming (ILP) [Dzeroski and Lavrac 2001], which combine induction and logic programming [Leiva 2002]. There are also other proposals that presented good results: Knobbe [2005] presents a framework illustrated by the Warmr algorithm, a generalization of the Apriori algorithm for relational models; Knobbe et al. [1999] present a framework for multi-relational decision trees; and Kersting and De Raedt [2001] explore an approach based on Bayesian networks.

2.2 Classification Task Using Multivalued Attributes

One of the most important tasks in data mining is the classification task, which aims at estimating the class of an instance based on its attributes. A common approach for performing classification, known as eager learning, involves two steps: in the first step, a classification model is constructed based on a training dataset, where the class labels of the instances are known. Using this classification model, the unknown class label of a new instance can be predicted, based on its set of attributes. The classification step is usually fast, since there is no need to access the dataset when a new instance is classified. To

predict the class, it is enough that the attribute values are evaluated by the classification model rules. In this approach, the construction of the classification model may have high computational cost and should be performed whenever the training dataset is significantly updated.

Other important approach for classification, known as lazy learning, does not create a prior classification model based on a training dataset. In this case, the data processing is only executed when a new instance needs to be classified. The expression "lazy" is used because this type of algorithm postpones the data processing until a classification request is made. Therefore, it avoids the cost of generating a classification model; however, the classification step is longer, as the dataset must be accessed whenever a new instance is classified.

The k-NN algorithm (k Nearest Neighbours) is a well known lazy classification technique, which has been used in the multi-relational data mining context [Emde and Wettschereck 1996; Duda et al. 2001; Dzeroski 2003]. The k-NN algorithm was proposed in the '50s, but it only became popular in the data mining and relational learning areas in the beginning of the '90s [Aha 1992].

The main idea of this algorithm is to classify a new instance by comparing it with the instances in the dataset to identify the k most similar. The label of the new instance is determined by the most frequent label among the k instances which are most similar to the instance being classified. The value of k is an input parameter of the algorithm.

To compare the similarity between the instances of the dataset, one uses distance measures. The k instances selected to classify a new instance are those considered the closest to the new instance, according to a given distance measure. A popular distance measure usually used with this type of classifier when the attributes are numeric is the Euclidian distance measure. It defines that the distance between two instances, $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, is given by

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (1)$$

If the target attributes have values in different scales, it is important to normalize them, ranging between 0 and 1, so that each attribute contributes equally to the distance value.

When the attributes are categorical, a simple way to calculate the distance between them is to consider the difference $(x_i - y_i)$ in Formula 1 to equal 1 (one), when the values are different, and to equal 0 (zero), when the values are the same.

In cases where the attribute to be considered in the classification represents a set of objects (multivalued attribute), one has to use a special kind of distance measure, which can deal with set comparison. The next subsection presents some distance measures which can deal with this kind of attribute.

2.3 Distance Measures on Multivalued Attributes

To calculate the distance between instances, it is necessary to define a measure which is evaluated based on the features of those instances. When we have a multivalued attribute, it is necessary to define some way to compare the sets of objects which represent the two instances. Different measures have been proposed in the literature for defining distances between sets of objects. Among them, we can cite Average Linking, Tanimoto and RIBL, which were used in our experiments to calculate the distances between sets of objects. The next subsections describe this three sets distance measures.

2.3.1 *Average Linking.* The Average Linking (AL) distance measure between two sets A and B, represented by $D_{AL}(A, B)$, considers the distance between the elements of all the possible combinations of elements from the two sets (or multivalued attributes), i.e., it considers all the possible pairs formed from A and B. It can be defined as the average of all pairwise distances [Kalousis et al.

2005]: $D_{AL}(A,B) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_1(a_i,b_j)}{|A||B|}$, where $A = \{a_1, a_2, ..., a_n\}, n \geq 1$, and $B = \{b_1, b_2, ..., b_m\}, m \geq 1$, represent the target sets and the distance between the attribute values is defined as:

$$d_1(a_i,b_j) = \begin{cases} |a_i - b_j|, & \text{if continuous,} \\ 0, & \text{if discrete and } a_i = b_j, \\ 1, & \text{if discrete and } a_i \neq b_j. \end{cases} \quad (2)$$

This measure does not satisfy the reflexive property, which means that the distance from a set to itself may be bigger than zero. This can lead to inconsistencies, since a set may be considered more similar to another set than to itself.

2.3.2 *Tanimoto.* The Tanimoto distance between sets A and B [Duda et al. 2001], represented by $D_T(A,B)$, is defined as: $D_T(A,B) = \frac{|A|+|B|-2|A\cap B|}{|A|+|B|-|A\cap B|}$.

This measure is based on the size of the intersection between the two sets. Therefore, this distance measure is directly applicable to sets of discrete elements. For continuous elements, we apply: $d_2(a_i,b_j) = \frac{|a_i-b_j|}{|a_i+b_j|}$. If the difference $d_2(a_i,b_j)$ is lower than a threshold defined by the user, the values a_i and b_j are considered the same.

The Tanimoto distance is less sensitive to the problem of outliers since each element of a set can be matched at most once. On the other hand, in measures such as Average Linking, each element is compared to all others. Therefore, if an element is incorrect or is an outlier it may alter more significantly the value of the measure than in the case of the Tanimoto measure.

2.3.3 *RIBL.* RIBL is a more sophisticated measure, since beyond calculating the distance between two sets A and B, it can calculate the distance between two instances in a dataset, represented by a set of monovalued or multivalued attributes.

However, in the context of this article, it is just interesting to know how RIBL deals with the comparison of two sets. The detailed description about this measure, including how it deals with instances comparison, can be found in the work of Dzeroski [2003].

RIBL proceeds with the comparison of two sets A and B as follows: it takes the smaller set (or the first one, if they have the same size) and, for each element in this set, it calculates its distance to the nearest element of the other set. The value of the measure is given by the sum of these calculated distances. If the cardinality of the sets are different, a normalization with the cardinality of the larger set is done. Thus, the distance $D_{RIBL}(A,B)$ between sets A and B is given by:

$$D_{RIBL}(A,B) = \begin{cases} \frac{\sum_{i=1}^{n} min_{j=1}^{m}(d_1(a_i,b_j))}{|B|}, & \text{if } |A| < |B| \\ \frac{\sum_{j=1}^{m} min_{i=1}^{n}(d_1(a_i,b_j))}{|A|}, & \text{if } |A| \geq |B| \end{cases} \quad (3)$$

where $A = \{a_1, a_2, ..., a_n\}, n \geq 1$, and $B = \{b_1, b_2, ..., b_m\}, m \geq 1$ represent the sets being compared and $d_1(a_i,b_j)$ is defined on Section 2.3.1.

This method focuses on the set of minimum distances between the elements of one set and the elements of the other set, providing a more global measure of how similar the two sets are. This could be problematic if there is an outlier in the set A, for example, which is much closer than all other elements of set A to the elements of the set B. In this case, this minimum distance will create a distortion in the result of the measure.

It is important to observe, comparing $d_1(a_i,b_j)$ and $d_2(a_i,b_j)$, that numeric attributes are dealt with

differently in the Taminoto measure than in Average Linking and RIBL. This can cause differences in the measure results, even if only monovalued attributes are being used.

There is no rule to define which distance measure will provide the best results for classification. This will depend on the specific application and its semantics, which should mainly drive the selection of the most appropriate measure. In some cases it might be more important to consider the distance between the two most similar elements of the sets, while in other cases a more global approach that takes into account all the elements might be required [Kalousis et al. 2005].

3. THE PROPOSED RELEVANCE MEASURE

As occurs with monovalued attributes, multivalued attributes may or may not have importance for classification in a given context. Knowing the set of books bought by a bookstore customer, for example, may be useful for inferring if this customer is more probably a man or a woman and if he (or she) is likely to have children. However, this information may be useless for inferring if this customer owns a credit card. Thus, the relevance of a multivalued attribute for classification, as in the monovalued context, depends on the target application and, more specifically, on the class label to be predicted. Therefore, it is important to propose and validate a relevance measure which can be used to assess the importance of multivalued attributes for a classification task.

There are many existing ways to assess the relevance of monovalued attributes, which use different types of relevance measures, such as distance measures, dependency measures, consistency measures and precision measures [Caruana and Freitag 1994; Lee 2005]. The objective of this work is the proposal and evaluation of a relevance measure for multivalued attributes based on dependency, i.e., in how much the class label depends on the given multivalued attribute. In other words, the proposed relevance measure indicates the ability of a multivalued attribute to define the target class label.

The scope of this work is restricted to binary classification problems. The proposed relevance measure is based on the difference between the probabilities of occurrence of each class, given each domain value of the multivalued attribute. If this difference is large, it means that the attribute has some useful information for estimating the class, i.e., the class is dependent on this attribute. If this difference is small, it means that the attribute is not important to determine the class.

Let x be a value of the domain of a multivalued attribute X and let C be the class value. The occurrence probability of class C, given the value x, can be obtained by Bayes rule, as presented in: $P(C|x) = P(x \ \wedge \ C)/P(x)$. This probability can be estimated from the dataset, by dividing the number of instances labeled as C which have the value x for the multivalued attribute by the total number of instances (of any class) which have the value x for the multivalued attribute.

Using $P(C|x)$ defined previously, we can define two vectors of probabilities, P_A and P_B, for classes A and B respectively, as follows: $P_A[i] = P(A|x_i), 1 \leq i \leq n$ and $P_B[i] = P(B|x_i), 1 \leq i \leq n$, where $x_1, x_2, ..., x_n$ represent the domain values of the multivalued attribute. From the difference between these vectors, we define the vector P_D of differences: $P_D[i] = |P_A[i] - P_B[i]|, 1 \leq i \leq n$.

The relevance measure R_{MULT} for multivalued attributes is defined in Formula 4, as the weighted average of the values in P_D. The weights are important to emphasize the most frequent values in the dataset and deemphasize the less frequent values.

$$R_{MULT}(X) = \frac{\sum_{i=1}^{n} P_D[i].count(x_i)}{\sum_{i=1}^{n} count(x_i)}, \tag{4}$$

where $count(x_i)$ represents the number of occurrences of the value x_i in the dataset.

The value 1 (one) for the relevance measure represents the maximum relevance for a multivalued

```
procedure CalculateMeasure(X, domain(X), dataset)
01.  n := number of items in the multivalued attribute domain;
02.  vCount[i] := for each value x_i, the number of instances which have the item x_i;
03.  vCountA[i] := for each value x_i, the number of instances labeled as A which have the item x_i;
04.  vCountB[i] := for each value x_i, the number of instances labeled as B which have the item x_i;
05.  for i := 1 to n do begin
06.       P_A[i] := vCountA[i]/vCount[i];
07.       P_B[i] := vCountB[i]/vCount[i];
08.       P_D[i] := |P_A[i] - P_B[i]|;
09.       vSum := vSum + (P_D[i] * vCount[i]);
10.       vSumCount := vSumCount + vCount[i];
11.  end;
12.  R_MULT(X) := vSum/vSumCount;
```

Fig. 1. Pseudo-code of the procedure for calculating the relevance measure

attribute. This will occur when, for each value x_i, $P_D[i] = 1$. In other words, for any value of the multivalued attribute, the occurrence probability of one of the classes is 100% while the occurrence probability of the other class is 0%. On the other hand, if the result of the relevance measure is 0 (zero), this means that the occurrence probability of each class is the same, for each value x_i of the multivalued attribute, i.e., knowing this attribute makes no difference for classification.

Figure 1 presents the procedure pseudocode for calculating the proposed relevance measure. The calculation of the vectors $vCount$, $vCountA$ and $vCountB$, presented in lines 02, 03 and 04 of the algorithm, respectively, can be done in a single scan of all tuples in the dataset.

4. EXPERIMENTS AND RESULTS

In this section, the datasets used to perform the experiments and the results are reported. To show the validity of the relevance measure, it was applied in several datasets and those results were compared with the accuracy values obtained from a classifier. When the relevance measure value indicates that an attribute is relevant, we expect that including this attribute on the set of attributes submitted to the classifier will improve the accuracy value. On the other hand, if the relevance measure value indicates that the attribute is not relevant, we expect that the resulting accuracy remains the same (if this attribute is irrelevant) or get worse (if the attribute represents an outlier).

We used some real relational datasets and then some hybrid datasets, which are real datasets where we add some artificial multivalued attributes. Those hybrid datasets were created to make available a controlled experiment about the behavior of the proposed relevance measure when it was applied on multivalued attributes with specific features.

To perform the classification task on relational datasets, this work uses the RelIbk algorithm, a k-NN classifier from Relational Weka data mining tool. The three distance measures described in subsection 2.3 – RIBL, Tanimoto and Average Linking – were combined with the k-NN algorithm. All experiments were performed using 10 fold cross validation. In the first experiment, we used the values 1, 2 and 3 for the k-NN k parameter. But as we got better results with value 3, the experiments reported on this work were performed only with this value.

4.1 Real Datasets

We used for the experiments four real datasets: KDD Cup 2000, Ebooks, IBGE POF 1999 and IBGE POF 2002. All those datasets have one or more tables which implements a multivalued attribute. KDD Cup 2000 and EBooks datasets were used previously on some work about multi-relational data mining [Perlich and Provost 2006]. We did not find any reference, in the literature, about IBGE

Table I. Number of instances and attributes, by type, for each real dataset

Dataset	# instances	# categorical monovalued attributes	# numerical monovalued attributes	# multivalued attributes	Average of set cardinality
KDD	795	49	6	2	Product: 1.35 and Collection: 1.21
EBooks	808	6	0	2	Book: 21.30 and Category: 1.60
IBGE 2002	822	7	2	2	Product: 13.08 and Category: 7.91
IBGE 1999	859	3	1	1	Product: 47.87

POF dataset (obtained from IBGE website: `http://loja.ibge.gov.br/`) been used on data mining research area. Table I shows the characteristics of these datasets.

4.1.1 *KDD Cup 2000.* KDD Cup 2000 dataset describes sales transactions collected from the website Gazelle.com through Blue Martini software. This website sells leg care products. The data were collected by permission directly from the website (`http://cobweb.ecn.purdue.edu/KDDCUP/data/`), where some basic cleaning treatment had already been done. The class labels are "True" for customers which spend more than 12 dollars mean per sale, and "False" for other customers. The distribution of instances is: 164 records (21%) are labeled as "True" and 631 records (79%) are labeled as "False". On this dataset, each customer is represented by 55 monovalued attributes and the customer can purchase a set of products, which are organized by collections. Therefore, we use two multivalued attributes on this dataset: one of them is represented by the set of products purchased by the customers and the other one is represented by the set of collection labels related to the products purchased.

4.1.2 *EBooks.* EBooks dataset contains data about online book sells of a Korean company. Details about this dataset can be found in the work of Perlich and Provost [2006]. We used two different attributes ("Sex" and "Kids") in the class role, one at a time, in our experiments. The distribution of instances is: for "Sex" class, 606 records (75%) are labeled as "1" and 202 records (25%) are labeled as "2"; for "Kids" class, 480 records (59%) are labeled as "0" and 328 records (41%) are labeled as "1". The dataset documentation does not define the meaning of the values {1,2} for "Sex" attribute and {0,1} for "Kids" attribute. On this dataset, each customer is represented by six monovalued attributes and the customer can purchase a set of books, which are organized by categories. Therefore, we use two multivalued attributes on this dataset: one of them is represented by the set of books purchased by the consumers and the other one is represented by the set of category labels related to the books purchased by the consumers.

4.1.3 *IBGE POF 2002.* IBGE POF 2002 dataset describes a family budget research. The dataset instances represent people from several regions of Brazil. These people have individual features (monovalued attributes) and purchase a set of products, which are represented by a multivalued attribute. We used two different attributes ("Credit Card" and "Sex") in the class role, one at a time, in our experiments. The distribution of instances is: for "Credit Card" class, 286 records (35%) are labeled as "Yes" and 536 records (65%) are labeled as "No"; for "Sex" class, 615 records (75%) are labeled as "1" and 207 records (25%) are labeled as "2". The dataset documentation does not define the meaning of the values {1,2} for "Sex" attribute. On this dataset, each person is represented by nine monovalued attributes and the person can purchase a set of products, which are organized by categories. Therefore, we use two multivalued attributes on this dataset: one of them is represented by the set of products purchased by people and the other one is represented by the set of category labels related to the products purchased by people.

4.1.4 *IBGE POF 1999.* IBGE POF 1999 dataset describes a family budget research, made by IBGE in 1999. The dataset instances represent families from several regions of Brazil. These families have individual features (monovalued attributes) and purchase a set of products, which are represented by a multivalued attribute. We used two different attributes ("Income" and "Family size") in the class

role, one at a time, in our experiments. The distribution of instances is: for "Income" class, 455 records (52%) are labeled as "Low income" and 414 records (48%) are labeled as "High income"; for "Family size" class, 360 records (42%) are labeled as "Small family" and 499 records (58%) are labeled as "Large family". On this dataset, each family is represented by four monovalued attributes and the family can purchase a set of products, represented by the single multivalued attribute.

4.2 Real Datasets – Results and Reviews

This section presents the experimental results using the real datasets. For each dataset and class attribute evaluated, the following results are showed: (a) relevance measure values for evaluated multivalued attributes, and (b) tables with accuracy values as a result of submitting those attributes to a specific classifier. We used a relational data mining tool with k-NN (RelIbk) classifier to obtain this accuracy values. The k-NN k parameter was set as 3. The distance measures used to perform k-NN algorithm were: Average Linking, Tanimoto and RIBL.

In the first scenario, we will analyze if the relevance measure value, for each attribute, is compatible with the accuracy value as a result of submitting this single attribute to the classifier. We expect that, for high relevance measure values (>0.7), the accuracy is high. In other hand, for low relevance measure values, we expect accuracy values next to the most frequent class percentage. Results in Table II confirm that the relevance measure can be a good index about the multivalued attribute relevance for classification. In most cases, high relevance measure values (>0.7) correspond to significant accuracy values. It is important to highlight that an accuracy value is considered significant when its value is higher than the most frequent class percentage.

For KDD/Spend dataset, the relevance measure values for both multivalued attributes are high (>0.720) and the accuracy values obtained for all distance measures are higher than the most frequent class percentage. On EBooks/Sex dataset, the relevance measure for Book attribute is high (0.795) and the accuracy value for all distance measures are substantially higher than the most frequent class percentage (75%). For Category attribute, the relevance measure is low (0.576) and the accuracy values are not significant (the values are around the most frequent class percentage). In this case, the relevance measure could indicate which is the better multivalued attribute for classification.

On EBooks/Kids, IBGE 2002 and IBGE 1999 datasets, the relevance measure values were low for all multivalued attributes (<0.61). In most cases, the accuracy values were lower or around the most frequent class percentage. So, the behaviour of relevance measure and accuracy were consistent. Only on IBGE 1999/Income we observe a contradiction: the Product attribute has a low relevance measure, but its accuracy for Tanimoto distance was well higher than the most frequent class percentage.

The results showed in Table II for multivalued attributes, do not consider the influence of monova-

Table II. Relevance measure values for multivalued attributes and accuracies obtained.

Dataset / Class / Most frequent class percentage	**Multivalued Attribute**	**Relevance Measure**	**Accuracy for each Distance Measure**
KDD / Spend / 79%	Product	0.741	RIBL: 81.89 | TA: 81.13 | AL: 81.00
	Collection	0.721	RIBL: 82.39 | TA: 83.27 | AL: 82.89
EBooks / Sex / 75%	Book	0.795	RIBL: 79.33 | TA: 84.03 | AL: 84.40
	Category	0.576	RIBL: 75.00 | TA: 75.25 | AL: 75.62
EBooks / Kids / 59%	Book	0.508	RIBL: 60.64 | TA: 61.14 | AL: 60.77
	Category	0.235	RIBL: 59.40 | TA: 60.15 | AL: 59.40
IBGE 2002 / Credit Card / 65%	Product	0.348	RIBL: 62.65 | TA: 65.94 | AL: 65.08
	Category	0.259	RIBL: 63.87 | TA: 64.35 | AL: 65.81
IBGE 2002 / Sex / 75%	Product	0.603	RIBL: 75.79 | TA: 77.00 | AL: 76.64
	Category	0.508	RIBL: 74.69 | TA: 77.00 | AL: 73.84
IBGE 1999 / Income / 52%	Product	0.229	RIBL: 42.26 | TA: 65.31 | AL: 54.48
IBGE 1999 / Family / 58%	Product	0.226	RIBL: 41.91 | TA: 60.42 | AL: 50.41

lued attributes. The accuracy values were obtained from the submission of each multivalued attribute to the classifier, one by one. In the next analysis, we consider the monovalued attributes as well.

In the second scenario, we will compare, for each dataset, two different situations: (a) accuracy values obtained using monovalued attributes and one multivalued attribute and (b) accuracy values using only monovalued attributes, without the contribution of any multivalued attribute. If the relevance measure related to the multivalued attribute is high, we expect that the inclusion of this attribute on the set of (monovalued) attributes will improve the accuracy; otherwise, we expect same or lower accuracy values on situation (b).

Table III shows the results about the second scenario. In each subtable, the first line presents the dataset and class identification; on the second line we can see the multivalued attributes and its relevance measure values, identified by R_{MULT}; the third line identifies the distance measures used on k-NN algorithm; the other lines are filled with the accuracy values obtained from classifier, for each distance measure. The first column of each subtable identifies the groups of attributes used on the classifier: "MV" indicates that only the multivalued attribute was submitted to the classifier; "MN" indicates that only a set of monovalued attributes was submitted to the classifier. "Good MN" indicates that we choose the "best" set of monovalued attributes and "Poor MN" indicates that we choose the "worst" group of monovalued attributes; "MV + MN" indicates that the set of attributes submitted to the classifier considers both the multivalued attribute and the monovalued attributes. An attribute selection algorithm available on Weka data mining tool (*Attribute Evaluator: InfoGainAttributeEval; Search Method: Ranker*) was used to create the "worst" and "best" sets of attributes. From the ranking generated by *InfoGainAttributeEval* algorithm, the best set of monovalued attributes was created by choosing those attributes with higher scores in the ranking; and the worst set of monovalued attributes was created by choosing those attributes with lower scores in the ranking. The strategy to create the worst set of attributes was different to KDD dataset, because the number of attributes was too big and the lower scores in the ranking was zero. So, to create the worst set of attributes with a non-zero individual score, we perform a random selection among attributes which did not belong to the best set. The size of this sets depends on the number of available attributes on datasets. For KDD dataset, the sets were created with ten elements each; for EBooks dataset, the size of attribute sets is only two elements; and for IBGE 2002 dataset, the attribute sets contain three elements each.

To show the influence of multivalued attributes when combined with monovalued attributes, we perform two different analysis: in the lines identified by "MV + Poor MN", the multivalued attribute was combined with the worst set of monovalued attributes. In the lines identified by "MV + Good MN", the multivalued attribute was combined with the best set of monovalued attributes.

We can observe that the multivalued attribute contribution tends to improve the accuracy of the previous set of monovalued attributes when its relevance measure is high ($>$0.7). However, this contribution is more relevant when the multivalued attribute is combined with the "worst" attributes, as we can see in subtables A and B. If the accuracy value of the initial set of monovalued attributes is already high, the inclusion of the multivalued attribute tends not to give a relevant contribution as we can see comparing the values from the two last lines of subtable A for the attribute "Product", and subtable B. In subtable A we can see an exception for the attribute "Collection". Even though the relevance measure of the multivalued attribute is high (0.721), its contribution to the initial set of monovalued attributes was irrelevant.

On the other hand, if the relevance measure of the multivalued attribute is low, its contribution to the accuracy value of the initial set of monovalued attributes tends to be irrelevant or it may even deteriorate the accuracy values. This situation can be observed in subtable B for the attribute "Category", subtable C for the attribute "Category" combined with "Poor MN" and subtable D. In subtable C, excepting for "Category" combined with "Poor MN", even though the relevance measure of the multivalued attribute is low, the inclusion of the multivalued attribute to the initial set of monovalued attributes improved the accuracy values.

Table III. Contribution of multivalued attributes (MV) when they are combined with monovalued ones (MN).

Subtable A: KDD - Class: Spend

MV Attribute: "Product" – R_{MULT}: 0.741				MV Attribute: "Collection" – R_{MULT}: 0.721			
	RIBL	TA	AL		RIBL	TA	AL
MV	81.89	81.13	81.00	MV	82.39	83.27	82.89
Random MN	74.97	77.23	74.97	Random MN	74.97	77.23	74.97
MV + Random MN	77.36	79.37	78.87	MV + Random MN	75.60	78.87	77.86
Good MN	83.02	83.02	83.02	Good MN	83.02	83.02	83.02
MV + Good MN	82.89	84.40	84.02	MV + Good MN	82.26	83.65	83.14

Subtable B: EBooks - Class: Sex

MV Attribute: "Category" – R_{MULT}: 0.567				MV Attribute: "Book" – R_{MULT}: 0.795			
	RIBL	TA	AL		RIBL	TA	AL
MV	75.00	75.25	75.62	MV	79.33	84.03	84.40
Poor MN	75.00	75.00	75.00	Poor MN	75.00	75.00	75.00
MV + Poor MN	75.37	75.25	75.37	MV + Poor MN	74.63	84.28	82.67
Good MN	75.00	75.00	75.00	Good MN	75.00	75.00	75.00
MV + Good MN	75.00	74.38	74.75	MV + Good MN	78.09	84.28	83.17

Subtable C: EBooks - Class: Kids

MV Attribute: "Category" – R_{MULT}: 0.235				MV Attribute: "Book" – R_{MULT}: 0.508			
	RIBL	TA	AL		RIBL	TA	AL
MV	59.40	60.15	59.40	MV	60.64	61.14	60.77
Poor MN	59.03	59.03	59.03	Poor MN	59.03	59.03	59.03
MV + Poor MN	60.02	59.90	60.27	MV + Poor MN	62.62	62.25	62.00
Good MN	62.50	62.50	62.50	Good MN	62.50	62.50	62.50
MV + Good MN	68.56	73.39	68.93	MV + Good MN	67.32	69.43	71.66

Subtable D: IBGE 2002 - Class: Credit Card

MV Attribute: "Product" – R_{MULT}: 0.348				MV Attribute: "Category" – R_{MULT}: 0.259			
	RIBL	TA	AL		RIBL	TA	AL
MV	62.65	65.94	65.08	MV	63.87	64.35	65.81
Poor MN	61.31	61.19	61.31	Poor MN	61.31	61.19	61.31
MV + Poor MN	58.27	60.83	60.34	MV + Poor MN	59.49	58.88	60.46
Good MN	69.83	69.10	69.83	Good MN	69.83	69.10	69.83
MV + Good MN	69.95	69.59	70.32	MV + Good MN	69.71	69.71	70.19

Subtable E: IBGE 2002 - Class: Sex

MV Attribute: "Product" – R_{MULT}: 0.603				MV Attribute: "Category" – R_{MULT}: 0.508			
	RIBL	TA	AL		RIBL	TA	AL
MV	75.79	77.00	76.64	MV	74.69	77.00	73.84
Poor MN	65.33	65.21	65.33	Poor MN	65.33	65.21	65.33
MV + Poor MN	68.61	69.95	68.86	MV + Poor MN	68.37	69.83	68.86
Good MN	80.78	80.78	80.78	Good MN	80.78	80.78	80.78
MV + Good MN	81.63	81.87	81.63	MV + Good MN	80.90	81.51	81.51

In subtable E, the relevance values are around 0.5 and 0.6. On these cases, the accuracies obtained with the single multivalued attribute are worse than those obtained with the "best" monovalued attribute set, but they are better than those obtained with the "worst" monovalued attribute set. The accuracies obtained with the combination of these multivalued attributes of "average" quality with the monovalued attributes were better than those obtained only with the use of monovalued attributes. This effect is more pronounced when we use the "worst" monovalued attributes.

The relevance measure proposed on this article evaluates multivalued attributes in a single way, not considering its combination with other attributes. In many cases, even when only monovalued attributes are considered on classification, single attributes may not be useful to discriminate the class label, however, when they are combined in a set of attributes, they can be useful to estimate the class label. Subtable C can show this situation, excepting for attribute "Category" combined with "Poor MN". Both the accuracy values obtained with the single multivalued attribute and the accuracy values obtained with the set of monovalued attributes were low. However, when those attributes

```
procedure CreateSyntheticAttribute(n,V)
1.  for i := 1 to n do begin
2.    qtdClasseA := round(V[i].Qtd * V[i].ProbA);
3.    qtdClasseB := round(V[i].Qtd * V[i].ProbB);
4.    for j := 1 to qtdClasseA do begin
5.      randomly select an instance from A class label which does not have the item V[i].Item;
6.      add the item V[i].Item to the synthetic multivalued attribute of the chosen instance;
7.    end;
8.    for j := 1 to qtdClasseB do begin
9.      randomly select an instance from B class label which does not have the item V[i].Item;
10.     add the item V[i].Item to the synthetic multivalued attribute of the chosen instance;
11.   end;
12. end;
```

Fig. 2. Pseudo-code of the procedure to create an artificial multivalued attribute

were combined, they could produce a much better accuracy values on classification, mainly on cases when multivalued attributes were combined with "Good MN". The evaluation of this combination of attributes is not included on the scope of this work, but it is an important issue to be considered on future works about the use of multivalued attributes on classification task.

4.3 Hybrid Datasets

The hybrid datasets used in the experiments were generated from the real datasets described in Section 4.1. We call them hybrid because the class value and all monovalued attributes (numerical and categorical) come from the real datasets, while the multivalued attributes are generated synthetically, to allow a better control of the experiments. We discard the multivalued attributes of the original real datasets, and keep only the ones that have been generated synthetically.

The pseudo-code of the algorithm that we use for generating a synthetic multivalued attribute, in a dataset with two classes, A and B, is shown in Figure 2. It takes as input the following parameters: size of the domain of the attribute (n) and, for each element i in the domain ($V[i] : Item$), the number of occurrences of i in the dataset ($V[i] : Qtd$), the probability of the instance belonging to class A given that it has item i ($V[i] : ProbA$) and the probability that it belongs to class B given that it has item i ($V[i] : ProbB$). The sum of the probabilities $V[i] : ProbA$ and $V[i] : ProbB$ should be 100%. In lines 2 and 3, the algorithm calculates, for each item i, the number of times that i should occur in classes A and B, as the integer that is closest to the products ($V[i] : Qtd * V[i] : ProbA$) and ($V[i] : Qtd * V[i] : ProbB$), respectively. In lines 5 and 9, the algorithm randomly selects instances in which to include these items. This random selection is done without replacement, since we do not allow repeated items for the same instance.

For each real dataset, we generate five hybrid datasets. The first has the original multivalued attribute substituted by a synthetic multivalued attribute with the same domain, the same number of occurrences for each element of the domain and the same probability values of each class given the occurrence of each item in the domain. The other four datasets have the original multivalued attributes substituted by a synthetic multivalued attribute with the same properties of the one generated in the first dataset, with the exception of the probability values. For the first dataset, Dataset 1, these values are represented by $V[i] : ProbA$ and $V[i] : ProbB$, for each element i of the domain. For the second dataset, Dataset 2, these values are recalculated such that the absolute value of the difference between $V[i] : ProbA$ and $V[i] : ProbB$, for each i, is increased by 20% with respect to the difference in Dataset 1. For the third dataset, Dataset 3, these values are recalculated such that the absolute value of the difference between $V[i] : ProbA$ and $V[i] : ProbB$, for each i, is increased by 40%. The fourth and fifth datasets, Dataset 4 and Dataset 5, are generated analogously to Dataset 2 and Dataset 3. However, for Datasets 4 and 5, the values of the probabilities are recalculated such that the absolute value of

the difference between $V[i] : ProbA$ and $V[i] : ProbB$ is reduced by 20% and 40%, respectively.

Note that the actual set of items that belongs to each instance varies from one synthetic attribute to another. But, if in the original dataset there is a pattern of co-occurrence of items in the domain, this pattern is likely to disappear, since correlations between items are not taken into account in the generation process.

It is also important to note that increasing/decreasing the difference between the probabilities is limited by the fact that each one of the probabilities cannot be below 0 (zero) or above 1 (one). And, the sum of $V[i] : ProbA$ and $V[i] : ProbB$ must always be 1 (one). Therefore, to increase the difference between the probabilities, the smallest of $V[i] : ProbA$ and $V[i] : ProbB$ is decreased by a factor, while the largest is increased by the same factor, respecting the lower and upper bounds. To decrease the difference, the inverse operation is performed, that is, the lowest of $V[i] : ProbA$ and $V[i] : ProbB$ is increased by a factor, while the largest is decreased by the same factor.

Because of the lower and upper bound restriction for the probabilities, the average difference, in some cases, will not attain exactly the desired value. This is not a problem for the evaluation we are doing, since the intention is just to determine the impact on the relevance measure of having multivalued attributes that are more descriptive or less descriptive of the class.

We expect that the relevance measure of the multivalued attribute in Dataset 1 be approximately equal to the relevance measure of the multivalued attribute in the original dataset. We also expect that the relevance measure of the multivalued attribute in Dataset 3 be larger than that of the multivalued attribute in Dataset 2, which should be larger than that of Dataset 1. Similarly, we expect that the relevance measure of the multivalued attribute in Dataset 5 be smaller than that of the multivalued attribute in Dataset 4, which should be smaller than that of Dataset 1. Furthermore, we expect that the classification accuracies obtained with the classifier, using the different datasets, follow the results predicted by the relevance measure of their multivalued attributes.

4.4 Hybrid Datasets - Results and Reviews

In this section, we present the results of experiments using the hybrid datasets described in previous section. As the experiments with real datasets, we used three distance measures to execute k-NN algorithm: RIBL, Tanimoto and Average Linking. The results are presented in the same way: relevance measure values for evaluated multivalued attributes and tables with accuracy values as a result of submitting those attributes to a specific classifier (for each distance measures used).

Two new scenarios were designed specifically to hybrid datasets, considering the possibility to vary the class probabilities for each attribute value. The first scenario takes into account the following issue. When we vary the class probabilities on hybrid datasets, the relevance measure values are modified. We will analyze if the accuracy values of the classifier – when those different multivalued attributes are used – follow the same behaviour than the relevance measure. In other words, we expect that for high relevance measures we have good accuracy values and vice-versa.

In Table IV, the rows identified by "RIBL", "TA" and "AL" shows the accuracy values obtained from classifier with all different hybrid datasets used on this experiment; the rows identified by "R_{MULT}" shows the relevance measure values for each multivalued attribute. To analyze the behavior of the relevance measure, values were organized on an intuitive order from the lowest to the highest relevance measure value. The lowest values are on that datasets where the probabilities differences were decreased (Dataset 5: -40% e Dataset 4: -20%) in relation to the original dataset; and the highest values are on that datasets where the probabilities differences were increased (Dataset 2: +20% e Dataset 3: +40%) in relation to the original dataset. The Dataset 1 values, which were created from the original dataset, are in the columns identified by "orig".

We can observe, in Table IV, that in most cases the accuracy variation is consistent with the

Table IV. Accuracy and Relevance Measure variation from the variation of the differences between classes probabilities

KDD - Class: Spend						**KDD - Class: Spend**					
Multivalued attribute: **"Product"**						Multivalued attribute: **"Collection"**					
	-40%	**-20%**	**orig**	**+20%**	**+40%**		**-40%**	**-20%**	**orig**	**+20%**	**+40%**
RIBL	66.03	73.08	82.26	85.28	84.90	RIBL	70.94	77.48	82.89	87.42	86.79
TA	76.35	77.61	83.65	84.28	85.16	TA	76.73	77.36	82.89	87.04	86.54
AL	75.34	77.61	83.65	84.28	85.03	AL	76.23	78.11	82.89	87.04	86.41
R_{MULT}	0.188	0.467	0.741	0.831	0.864	R_{MULT}	0.148	0.435	0.721	0.881	0.916
EBooks - Class: Sex						**EBooks - Class: Kids**					
Multivalued attribute: **"Book"**						Multivalued attribute: **"Book"**					
	-40%	**-20%**	**orig**	**+20%**	**+40%**		**-40%**	**-20%**	**orig**	**+20%**	**+40%**
RIBL	14.85	48.64	75.87	36.39	33.91	RIBL	44.68	61.01	48.14	44.68	36.76
TA	74.88	80.44	98.76	99.88	99.75	TA	40.22	50.87	59.65	68.07	73.27
AL	75.00	76.11	98.76	99.75	99.75	AL	48.89	51.73	56.06	62.38	62.50
R_{MULT}	0.293	0.564	0.795	0.865	0.900	R_{MULT}	0.251	0.412	0.508	0.559	0.611
IBGE 2002 - Class: Credit Card						**IBGE 2002 - Class: Credit Card**					
Multivalued attribute: **"Product"**						Multivalued attribute: **"Category"**					
	-40%	**-20%**	**orig**	**+20%**	**+40%**		**-40%**	**-20%**	**orig**	**+20%**	**+40%**
RIBL	47.08	52.80	58.27	65.08	71.41	RIBL	58.76	63.38	64.11	65.57	64.72
TA	61.19	58.51	64.23	65.45	71.41	TA	64.48	61.19	59.61	61.80	64.23
AL	59.49	57.78	63.38	64.96	70.92	AL	63.14	63.38	62.77	63.99	64.47
R_{MULT}	0.127	0.251	0.348	0.426	0.506	R_{MULT}	0.050	0.156	0.258	0.361	0.467
IBGE 2002 - Class: Sex						**IBGE 2002 - Class: Sex**					
Multivalued attribute: **"Product"**						Multivalued attribute: **"Category"**					
	-40%	**-20%**	**orig**	**+20%**	**+40%**		**-40%**	**-20%**	**orig**	**+20%**	**+40%**
RIBL	56.69	65.57	74.45	76.88	79.44	RIBL	72.26	74.57	76.15	76.52	83.09
TA	74.94	74.09	74.33	78.71	83.70	TA	75.18	73.36	74.09	77.37	83.82
AL	73.36	73.72	75.06	77.49	83.09	AL	74.94	74.57	75.06	77.49	84.18
R_{MULT}	0.172	0.396	0.603	0.729	0.804	R_{MULT}	0.100	0.301	0.508	0.680	0.823

relevance measure behavior. Only for EBooks/Sex dataset, the accuracy obtained with RIBL shows an inconsistent situation. The multivalued attribute of this dataset has a very large domain (more than 5000 items) and, as consequence, the range of item set sizes is too large. A detailed analysis can show that RIBL is quite sensitive to the difference between the item set sizes of the sets which are been compared [Tasca 2008]. RIBL sums the minimum distances among the distances from the elements of the smallest set to the elements of the largest set, and then, divides it by the size of the largest set. Thus, the smaller the smallest set is, the smaller will be the numerator; the larger the largest set is, the larger will be the denominator and, consequently, the smaller will be distance value.

In the second scenario of experiments with hybrid datasets, we intend to verify if the proposed relevance measure can be used to compare two multivalued attributes from the same dataset. For highly-rated attributes we expect better accuracy values and vice-versa. It is important to know that the accuracy values for these experiments were obtained with the multivalued attribute by itself, disregarding the influence of any other attributes from the dataset.

In Table V we compare two multivalued attributes ("MV" column) from the same dataset, with different relevance measure values ("R_{MULT}" column). In each real dataset used on this experiment, there are two multivalued attributes represented here by a and b. For each real dataset, from each of these multivalued attributes a and b, we generate five synthetic multivalued attributes: $\{a_1, a_2, a_3, a_4, a_5\}$ e $\{b_1, b_2, b_3, b_4, b_5\}$, for a total of 10 synthetic attributes. The choice of attributes to be compared in this analysis was done by choosing pairs (a_i and b_j) of attributes with the largest differences in the values of their relevance measures.

In all analyzed cases, the better quality attribute, in terms of relevance measure, presented better accuracy. Thus, we can say that the proposed relevance measure also can be used to compare attributes

Table V. Comparing two multivalued attributes from the same dataset

Dataset / Class	MV	R_{MULT}	Accuracy for each Distance Measure
IBGE 2002 / Sex	a_i	0.804	RIBL: 79.44 | TA: 83.70 | AL: 83.09
	b_j	0.301	RIBL: 74.57 | TA: 73.36 | AL: 74.57
IBGE 2002 / Credit Card	a_i	0.505	RIBL: 71.41 | TA: 71.41 | AL: 70.92
	b_j	0.156	RIBL: 63.38 | TA: 61.19 | AL: 63.38
EBooks / Kids	a_i	0.508	RIBL: 60.64 | TA: 61.14 | AL: 60.77
	b_j	0.235	RIBL: 59.40 | TA: 60.15 | AL: 59.40
EBooks / Sex	a_i	0.795	RIBL: 79.33 | TA: 84.03 | AL: 84.40
	b_j	0.576	RIBL: 75.00 | TA: 75.25 | AL: 75.62
KDD / Spend	a_i	0.864	RIBL: 84.90 | TA: 85.16 | AL: 85.03
	b_j	0.148	RIBL: 82.89 | TA: 82.89 | AL: 82.89

from the same dataset, contributing to select the best set of attributes for classification task.

5. CONCLUSIONS

This work proposes a relevance measure for multivalued attributes, which aims at measuring their importance for classification. As we could not find in the literature any other relevance measure specific for multivalued attribute, it was not possible to compare our proposed measure with other ones, but just verify if this measure may be useful to select a set of good attributes for classification task. To perform this analysis we used the k-NN classification algorithm in conjunction with three different distance measures for multivalued attributes: RIBL, Tanimoto and Average Linking. We performed analysis in four different scenarios, using real and hybrid datasets, and we could show that in most cases the proposed relevance measure can be a good index about the quality of multivalued attributes. This measure could be quite useful in conjunction with algorithms for attribute selection which uses filter approach. For each evaluated multivalued attribute, we could verify that the relevance measure value is coherent with the accuracy value generated by the classifier when this attributes were used by itself, as the single attribute used for classification. For high relevance measure values, in most cases the accuracy was higher than the most frequent class percentage, and for low relevance measure values, the accuracy remained lower or around the most frequent class percentage.

We also evaluated these two distinct situations: (a) submitting to the classifier a set of monovalued attributes and one multivalued attribute and (b) submitting to the classifier only a set of monovalued attributes. The relevance measure value for the multivalued attribute was consistent: when this value was high, the inclusion of the multivalued attribute on the set of monovalued attribute improved the accuracy. It is important to highlight that the result of this experiment is quite sensitive to the quality of the monovalued attributes combined with the multivalued attribute.

By changing the difference of class probabilities on hybrid datasets, the relevance measure value varies. We could verify with this experiment that the accuracy values follow the behavior of the relevance measures, i.e., for high relevance measures we obtained better accuracies and vice-versa. We also could conclude that the proposed relevance measure can be used to compare multivalued attributes in the same dataset. For highly-rated attributes we obtained better accuracies on classification. It is important to observe that the accuracy values for these experiments were obtained with the multivalued attribute by itself, disregarding the influence of any other attributes from the dataset.

Although it was not the focus of this work, we could analyze the behavior of the distance measures used on experiments: RIBL, Tanimoto and Average Linking. RIBL presents a serious problem with multivalued attributes, because it is quite sensitive to the difference between the item set sizes. This can generate distorted results and even contradictory situations. Average Linking presented a good performance, although it also has some problems, since it sometimes considers greater than zero the distance between two identical sets. Tanimoto seems to fit better for comparing multivalued attributes. However, each measure has different features which may fit better in different contexts and objectives.

The relevance measure proposed in this work takes into account the quality of multivalued attributes by themselves. A future research could be performed to study some kind of relevance measure which could take into account the influence of sets of attributes on predicting the class label. Another suggestion for future work is to extend the proposed relevance measure to be used on datasets with more than two class labels. The relevance measure proposed is based on the difference between the probabilities of the two class labels. One way to apply this concept to many class labels would be by using a standard deviation among the average probabilities of each class. A multivalued attribute could be considered important when the relevance measure value was higher than the standard deviation for one of the classes and lower than the standard deviation for the other ones.

REFERENCES

Aha, D. W. Tolerating Noisy, Irrelevant and Novel Attributes in Instance-Based Learning Algorithms. *International Journal of Man-Machine Studies* 36 (2): 267–287, 1992.

Caruana, R. and Freitag, D. How Useful is Relevance? In *Relevance, Papers from the 1994 AAAI Fall Symposium*. AAAI, 1994. Technical Report FS-94-02.

Dehaspe, L. and Toivonen, H. Discovery of Relational Association Rules. In *Relational Data Mining*, S. Dĕzeroski (Ed.). Springer-Verlag, New York, USA, pp. 189–208, 2001.

Duda, R., Hart, P., and Stork, D. *Pattern Classification and Scene Analysis*. John Willey and Sons, New York, USA, 2001.

Dzeroski, S. Multi-Relational Data Mining: an Introduction. *SIGKDD Explorations* 5 (1): 1–16, 2003.

Dzeroski, S. and Lavrac, N. *Relational Data Mining*. Springer-Verlag, Secaucus, USA, 2001.

Elmasri, R. and Navathe, S. B. *Fundamentals of Database System*. Addison-Wesley, USA, 2010.

Emde, W. and Wettschereck, D. Relational Instance Based Learning. In *Proceedings of the International Conference on Machine Learning*. San Francisco, USA, pp. 122–130, 1996.

Emde, W. and Wettschereck, D. Multi-relational Data Mining Using Probabilistic Relational Models: research Summary. In *Proceedings of the Workshop in Multi-relational Data Mining*. Freiburg, Germany, 2001.

Hall, M. A. and Holmes, G. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 15 (3): 1437–1447, 2003.

Kalousis, A., Woznica, A., and Hilario, M. A Unifying Framework for Relational Distance-Based Learning. Tech. rep., University of Geneva, Geneva, Switzerland, 2005.

Kersting, K. and De Raedt, L. Interpreting Bayesian Logic Programs. In *Proceedings of the Work-in-Progress Track at the International Conference on Inductive Logic Programming*. Szeged, Hungary, pp. 138–155, 2001.

Knobbe, A. J. Multi-Relational Data Mining. In *Proceedings of the Conference on Multi-Relational Data Mining*. Amsterdam, The Netherlands, pp. 1–118, 2005.

Knobbe, J., Siebes, A., and Van der Wallen, D. M. G. Multi-Relational Decision Tree Induction. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*. London, UK, pp. 378–383, 1999.

Kramer, S., Lavrac, N., and Flach, P. Propositionalization Approaches to Relational Data Mining. In *Relational data mining*, S. Dĕzeroski (Ed.). Springer-Verlag, New York, USA, pp. 262–286, 2001.

Laer, V. V. and Raedt, L. How to Upgrade Propositional Learners to First Order Logic: a Case Study. In *Relational data mining*, S. Dĕzeroski (Ed.). Springer-Verlag, New York, USA, pp. 235–261, 2001.

Lee, H. *Selection of Important Attributes for Knowledge Discovery from Databases (in portuguese)*. Ph.D. thesis, USP - São Carlos, SP-Brazil, 2005.

Leiva, H. *MRDTL: a multi-relational decision tree learning algorithm*. M.S. thesis, Iowa State University, Ames, USA, 2002.

Liu, H. and Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, USA, 1998.

Liu, H. and Motoda, H. Less Is More. In *Computational Methods of Feature Selection*, H. Liu and e. Motoda, H. (Eds.). Chapman and Hall/CRC, pp. 3–17, 2008.

Perlich, C. and Provost, F. Distribution-Based Aggregation for Relational Learning from Identifier Attributes. *Machine Learning* 62 (1-2): 65–105, 2006.

Tasca, M. *A Proposal of Relevance Measure for Classification Using Multivalued Attributes (in portuguese)*. M.S. thesis, UFF - Niterói, RJ-Brazil, 2008.

Tasca, M., Plastino, A., and Zadrozny, B. A Proposal of Relevance Measure for Multivalued Attributes in Classification (in portuguese). In *Proceedings of the Workshop on Data Mining Algorithms and Applications, in conjunction with the Brazilian Symposium on DataBases*. Fortaleza, Brazil, 2009.