Numbers in Multi-relational Data Mining

Arno J. Knobbe[1,2] and Eric K.Y. Ho[1]

[1] Kiminkii, Postbus 171, NL-3990 DD, Houten, The Netherlands
{a.knobbe, e.ho}@kiminkii.com
[2] Utrecht University, P.O. box 80 089, NL-3508 TB Utrecht, The Netherlands

Abstract. Numeric data has traditionally received little attention in the field of Multi-Relational Data Mining (MRDM). It is often assumed that numeric data can simply be turned into symbolic data by means of discretisation. However, very few guidelines for successfully applying discretisation in MRDM exist. Furthermore, it is unclear whether the loss of information involved is negligible. In this paper, we consider different alternatives for dealing with numeric data in MRDM. Specifically, we analyse the adequacy of discretisation by performing a number of experiments with different existing discretisation approaches, and comparing the results with a procedure that handles numeric data dynamically. The discretisation procedures considered include an algorithm that is insensitive to the multi-relational structure of the data, and two algorithms that do involve this structure. With the empirical results thus obtained, we shed some light on the applicability of both dynamic and static procedures (discretisation), and give recommendations for when and how they can best be applied.

1 Introduction

Whereas numeric data is at the core of the majority of propositional Data Mining systems, it has been largely overlooked in Multi-Relational Data Mining (MRDM). Most MRDM systems assume that the data is a mixture of symbolic and structural data, and if the source database contains numbers, they will either have to be filtered out or pre-processed into symbolic values. Apart from historical reasons – symbolic representations are popular in the logical roots of MRDM –, the full treatment of numeric data comparable to propositional approaches is mostly ignored for reasons of simplicity and efficiency. MRDM is characterised by large hypothesis spaces, and the inclusion of continuous domains that offer a large range of (very similar) refinements is thought to make MRDM intractable. Most multi-relational systems rely on so-called discretisation procedures to reduce the continuous domains to more manageable symbolic domains of low cardinality, such that the search remains realistic. The resulting loss of precision is assumed to be negligible.

In this paper, we survey a number of existing approaches to dealing with numeric data in MRDM, with the aim of empirically determining the value of each of these approaches. These approaches include a number of pre-processing procedures suggested recently [6, 2], as well as one of the few MRDM algorithms that deal with numbers dynamically, developed by the authors of this paper [2, 4]. The discretisation procedures include a simple algorithm that considers each table in isolation, and discretises each numeric attribute on the basis of the distribution of its values,

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 544 – 551, 2005.

regardless of any other tables connected to the current table. Two further discretisation procedures do involve the multi-relational structure of the database, and aim at finding good intervals, keeping in mind that the resulting symbolic attributes will be used in the context of the other tables in the database. The algorithm that deals with numbers dynamically does not require any pre-processing of the data. Rather than fixing a number of intervals prior to the analysis, it will consider the numeric data for a hypothesis at hand, and determine thresholds that are optimal for the given context. Especially at deeper levels of the search, where reasonably specific subgroups are considered, relevant thresholds will differ significantly from those determined on the whole dataset.

We test the four approaches experimentally on four well-known multi-relational datasets where numeric attributes play an important role: Mutagenesis (two varieties), Financial and Musk. With these experiments, we aim to shed some light on when and how each approach can best be applied. Furthermore, we hope to get some guidelines for important parameters of the discretisation procedures, such as the coarseness of the discretisation and the choice of representation. The experimental results are compared to those obtained on databases where all numeric information is removed, in order to get a baseline for the procedures that do (to some extent) involve the continuous domains.

2 Foundations

In the class of discrete patterns that we aim at (decision trees, rules, etc.), dealing with numeric data comes down to choosing numeric thresholds that form useful subgroups. Clearly, the distribution of numeric values, and how the target concept depends on this distribution is essential. In propositional data mining, choosing thresholds is fairly straightforward, as there is a one-to-one correspondence between occurring values and individuals. In MRDM however, we are dealing with non-determinate (i.e. one-to-many) relations between tables. In many cases, numeric attributes do not appear in the target table, and multiple values of the attribute are associated with a single structured individual. Whereas in propositional data mining, we can think of the whole database as a 'cloud' of points, in MRDM each individual forms a cloud. The majority of pattern languages in MRDM characterise such individuals by testing for the presence of values that exceed a given threshold. As the following lemma shows, only the largest and smallest values within each individual are relevant to include or exclude an individual on the basis of a single numeric test. Only these values will therefore be candidates for numeric thresholds.

Lemma 1. Let B be a bag of real numbers, and t some real, then

$$\exists v \in B: v \geq t \;\; \textit{iff} \;\; \max(B) \geq t,$$
$$\exists v \in B: v \leq t \;\; \textit{iff} \;\; \min(B) \leq t.$$

Lemma 1 furthermore demonstrates that there is a difference between the set of thresholds appropriate for the $\leq$ and the $\geq$ operator. This means that any procedure that selects thresholds will have to be performed separately for each operator.

Choosing thresholds can roughly be done in two ways: dynamically and statically. A *dynamic* approach (see Section 3) considers the hypothesis at hand, and determines a collection of thresholds on the basis of the information contained in the individuals covered by the hypothesis in question. A *static* approach (see Section 4) on the other hand considers the entire database prior to analysis and determines a collection of thresholds once and for all. Typically these thresholds are then used to pre-process the data, replacing the numeric data with symbolic approximations. We refer to such a pre-processing step as *discretisation*. Clearly, a dynamic approach is preferable from an accuracy standpoint, as optimal thresholds are computed for the situation at hand. On the other hand, dynamic computation of thresholds makes algorithms more complex, and less efficient.

In the context of discretisation, we refer to numeric thresholds as *cut points*. A collection of n-1 cut points splits the continuous domain into n intervals. A group of values falling in a specific interval is referred to as a *bin*.

In MRDM, it makes sense to not just consider the available numeric values in the computation of cut-points, by also the multi-relational structure of the database. In general, a table is connected to other tables by associations, some of which may be non-determinate (a single record in one table corresponds to multiple records in another table). The effect of such associations is thus that records in a table can be divided into *groups*, depending on the relation to records in the associated table. Considering the multi-relational structure in the computation of cut points is hence tantamount to considering the numeric value, as well as the group the value belongs to. In the remainder of this paper, we refer to groups as the sets implied by this multi-relational structure.

3 Dynamic Handling of Numbers

An MRDM algorithm that handles numbers dynamically considers a range of cut points for a given numeric attribute, and determines how each of these tentative cut points influences the quality of a multi-relational hypothesis under consideration. As the optimal cut point depends on the current hypothesis, and many hypotheses are considered by an MRDM algorithm, the set of relevant cut points cannot be determined from the outset. Rather, we will have to consider the subgroup at hand, and query the database for a list of relevant cut points, and associated statistics.

In general, all values for the numeric attribute that occur in the individuals covered by the hypothesis at hand can act as candidate cut points. In theory, this set of values can be quite large, which can make the dynamic generation of cut points very inefficient. The MRDM system Safarii [2, 4] uses an approach that considers only a subset of these values, thus reducing some of the work. It relies on the observation from Lemma 1 that only the extreme values within a bag of numbers are relevant in order to test the presence of values above or below a certain cut point. Safarii uses a database primitive (a predefined query template) called NumericCrossTable [2] that selects the minimum (maximum) value within each individual covered by the current hypothesis, and then groups over these extreme values to produce the desired counts. We thus get a more reasonable number of candidate refinements.

Unfortunately it is still not realistic to continue the search on the basis of each of these refinements. Safarii therefore selects from the reduced set of candidate refinements only the optimal one for further examination. Because the operators ≤ and ≥ produce two different sets of candidate refinements, we essentially get two refinements per hypothesis and numeric attribute encountered. Note that keeping only the optimal refinements introduces a certain level of greediness into the algorithm.

4 Discretisation

In this section, we briefly outline the three methods for discretising numeric data to be used in our experiments. We refer to [3] for a full description. Conceptually, discretisation entails defining a number of consecutive intervals on the domain of a numeric attribute, and replacing this attribute with a nominal attribute that represents the interval values fall into. The three methods are identical in how numeric attributes are transformed based on the intervals defined. The essential difference between the methods lies in how the cut points between intervals are computed.

The first method presented computes a (user-determined) number of cut points based on the distribution of values of the numeric attribute. It ignores the fact that data in a particular table will generally be considered in the context of that in other tables. The remaining two methods do consider the multi-relational structure of the data, and compute cut points assuming that discretised values will be considered after joining with tables that are directly attached to the table at hand.

Because the numeric data typically appears in tables other than the target table, it is not always straightforward to assign a class (which is related to the target table) to the value. All three methods are therefore class-blind (or *unsupervised*): the methods do not consider a predefined target concept. As a result, the transformed data can be used on a range of class-definitions.

Equal Height Histogram. The first algorithm computes cut points regardless of any multi-relational structure. It simply considers every numeric attribute in every table in turn and replaces it by a nominal attribute that preserves as much of the information in the original attribute as possible. A collection of cut points is computed that produces bins of (approximately) equal size. Such a procedure is known as *equal interval frequency*, or *equal height histogram*, which is the term we will adopt.

Equal Weight Histogram. The second discretisation procedure involves an idea proposed by Van Laer et al. [6]. The algorithm considers not only the distribution of numeric values present, but also the groups they appear in. It is observed that larger groups have a larger impact on the choice of cut points because they have more contributing numeric values. In order to compensate for this, numeric values are weighted with the inverse of the size of the group they belong to. Rather than producing bins of equal size, we now compute cut points to obtain bins of equal weight.

Aggregated Equal Height Histogram. Like the EqualWeight algorithm, the AggregatedEqualHeight algorithm proposed in [2] takes the multi-relational structure of the database into account in the computation of the cut points. The algorithm is centred around the idea that not all values within a group are relevant when inquiring about the presence of numeric values above or below some threshold. As was outlined

in Section 2, it suffices to consider the minimum and maximum value within a group. The idea of the AggregatedEqualHeight algorithm is hence to take the minimum value per group and compute an equal height histogram on these values, in order to discretise all values. The process is then repeated for the maximum per group. We thus get two new attributes per numeric attribute.

Representation. In our discussion of the different discretisation procedures, we have assumed that the outcome is a collection of nominal attributes, where each value represents one of the computed intervals. In fact when we produce n nominal values, we do not only lose some amount of precision (which we assume to be minimal), but also the inherent order between intervals. Although the inability to handle ordered domains (numeric or ordinal) is part of our motivation for applying discretisation, we can choose a representation that preserves the order information without having to accommodate for it explicitly. This representation involves n-1 binary attributes per original numeric attribute, one for each cut point. Rather than representing each individual interval, the binary attributes represent overlapping intervals of increasing size. By adding such attributes as conjuncts to the hypothesis through repeated refinements, a range of intervals can be considered. A further advantage of this representation is that the accuracy is less sensitive to the number of intervals as the size of the intervals does not decrease with the number of intervals. An important disadvantage of this representation is the space it requires. Especially with larger numbers of intervals, having n-1 new binary attributes per original attribute can become prohibitive.

In our experiments, we will consider both the nominal and the binary representation, and compare the results to determine the optimal choice. We will refer to the latter representation as *cumulative binary*.

5 Experiments

Although we have multiple approaches to dealing with numeric data to test, we have chosen to apply a single mining algorithm. This allows us to sensibly compare results. The algorithm of choice is the Rule Discovery algorithm contained in the Safarii MRDM package produced by the authors [2, 4]. This algorithm produces a set of independent multi-relational rules. The algorithm includes the dynamic strategy for dealing with numbers described in Section 3. In order to test the discretisation procedures, we have pre-processed the different databases by generating the desired discretised attributes, and removing the original numeric attributes. The different discretisation procedures were implemented in the pre-processing companion to Safarii, known as ProSafarii.

Although a range of evaluation measures and search strategies is available in Safarii, we have opted for rules of high *novelty*, discovered by means of *beam search* (beam width 100, maximum depth 6). A time limit of 30 minutes per experiment was selected. The algorithm offers filtering of rules by means of a computed convex hull in ROC space [2]. The area under the ROC curve gives a good measure of the quality of the discovered rule set, as it is insensitive to copies or redundant combinations of rules. We will use this measure (values between 0.5 and 1) to compare results.

We will test the different algorithms on the following three well-known multi-relational databases:

- **Mutagenesis [5].** A database containing structural descriptions of molecules. We use two varieties, called B2 and B3. B2 contains symbolic and structural information as well as a single numeric attribute describing the charge of each atom. B3 contains two additional attributes on the molecule-level.
- **Financial [7, 2].** A database containing seven tables, describing various activities of customers of a Czech bank.
- **Musk [1].** A database describing 166 continuous features of different conformations molecules may appear in.

In [3] we present a detailed overview of the results obtained. We summarize the main conclusions in the paragraphs below.

Discretisation Procedures. Let us begin by considering how well the discretisation procedures perform. The table below summarises how often each procedure is involved in a win or a tie (no other procedure is superior). Procedures are compared per setting of the number of bins, in order to get comparable results. It turns out that AggregatedEqualHeight is clearly the best choice for Financial and Musk. Surprisingly, the propositional procedure EqualHeight performs quite well on Mutagenesis B2. The results for EqualHeight and EqualWeight on Mutagenesis B3 are virtually identical, which should come as no surprise, as this database contains two powerful attributes in the target table. The multi-relational data is mostly ignored.

In every case, the use of discretised attributes is better than not using the numeric information altogether, although in a few cases the advantage was minimal.

	EqualHeight	EqualWeight	AggregatedEqualHeight
Mutagenesis B2	62.5%	50.0%	37.5%
Mutagenesis B3	75.0%	87.5%	75.0%
Financial	0%	12.5%	87.5%
Musk	0%	25%	75.0%

Discretisation vs. Dynamic Handling. So can the discretisation procedures compete with the dynamic approach to numeric data, or is it always best to use the latter? In the table below, we compare the performance of the collection of discretisation procedures to dynamic handling of numbers. Each row shows in how many of the $3\times4\times2=24$ runs discretisation outperforms the dynamic approach. In the majority of cases, the dynamic approach outperforms the discretisation procedures, as was expected. However, for every database, there are a number of choices of algorithm, representation and number of bins, for which discretisation can compete, or even give slightly better results (see [3] for details).

If the set of cut points considered by the dynamic approach in theory is a superset of that considered by any discretisation procedure, how can we explain the moderate performance of the dynamic algorithm in such cases? The main reason is that the dynamic algorithm is more greedy than the discretisation procedures, because of the way numeric attributes are treated. Of the many refinements made possible by the numeric attribute, only the optimal pattern is kept for future refinements. Therefore, good rules involving two or more numeric conditions may be overlooked. On the

other hand, the nominal attributes resulting from discretisation produce a candidate for each occurring value, rather than only the optimal one. Because beam search allows several candidates to be considered, it may occur that sub-optimal initial choices may lead to optimal results in more complex rules.

	discretisation	dynamic
Mutagenesis B2	5	19
Mutagenesis B3	9	15
Financial	0	24
Musk	1	23

Choice of representation. The comparison between the two proposed representations is clear-cut: the cumulative binary representation generally gives the best results (see table below). The few cases where the nominal representation was (slightly) superior can be largely attributed to lower efficiency caused by the larger hypothesis space of the cumulative binary approach.

Although the cumulative binary representation is very desirable from an accuracy point of view, in terms of computing resources and disk space, the cumulative binary approach can become quite impractical, especially with many bins. Particularly in the Musk database, which contains 166 numeric attributes, several limits of the database technology used were encountered.

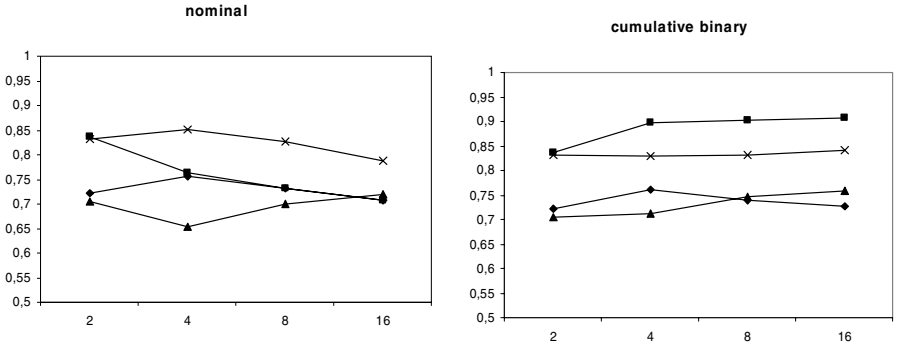
	nominal	cumulative binary	ties
Mutagenesis B2	3	5	4
Mutagenesis B3	0	9	3
Financial	2	6	4
Musk	5	4	7

Effect of Number of Bins. As has become clear, the number of bins is an important parameter of the discretisation procedures considered. Can we say something about the optimal value for this parameter? It turns out that the answer to this question depends on the choice of representation. Let us consider the cumulative binary representation. The performance roughly increases as more cut points are added (see the diagrams on the next page). This is because extra cut points just add extra opportunities for refinement and thus extra precision. The only exception to this rule is when severe time constraints are present. Because of the larger search space, there may be no time to reach the optimal result. For the nominal representation, there appears to be an optimal number of cut points that depends on specifics of the database in question. Having fewer cut points has a negative effect on the precision, whereas too many cut points results in rules of low support, because each nominal value only represents a small interval. For the Mutagenesis and Musk database, the optimal value is relatively low: between 2 and 4. The optimal value for Financial is less clear.

6 Conclusion

In general, we can say that the dynamic approach to dealing with numbers outperforms discretisation. This should come as no surprise, as the dynamic approach

is more precise in choosing the optimal numeric cut points. It is surprising however to observe that in some cases, it is possible to choose parameters and set up the discretisation process such that it is superior. Unfortunately, it is not immediately clear when faced with a new database what choice of algorithm, representation and



coarseness produces the desired result. Essentially, it is a matter of some experimentation to come up with the right settings. Even then, there is no guarantee that the extra effort of pre-processing the data provides a substantial improvement over the dynamic approach. Of course, when working with a purely symbolic MRDM system, discretisation is mandatory.

For discretisation, we recommend that the AggregatedEqualHeight procedure be tried first, as it has proven to give good results. It is worth the effort to consider EqualHeight as an alternative. The added value of the EqualWeight procedure over EqualHeight is negligible, and can therefore be ignored.

Our experimentation shows that in general, the simple nominal representation commonly used in MRDM projects is sub-optimal. Moreover, this representation is rather sensitive to the selected number of bins. In most cases the cumulative binary representation is preferable. This representation should be applied with as many bins as is realistic, given space and time limitations. Only when time restrictions can be expected to have a detrimental effect on the search depth, should lower numbers be considered.

References

1. Dietterich, T., Lathrop, R., Lozano-Pérez, T. *Solving the multiple-instance problem with axis-parallel rectangles*, Artificial Intelligence, 89(1-2):31-71, 1997
2. Knobbe, A.J. *Multi-Relational Data Mining*, Ph.D. dissertation, 2004, http://www.kiminkii.com/thesis.pdf
3. Knobbe, A.J., Ho, E.K.Y. *Numbers in Multi-Relational Data Mining*, 2005, http://www.kiminkii.com/publications/pkdd2005long.pdf
4. *Safarii, the Multi-Relational Data Mining engine*, Kiminkii, 2005, http://www.kiminkii.com/safarii.html
5. Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E., King, R.D., *Theories for mutagenicity: A study in first-order and feature-based induction*, Artificial Intelligence, 85(1,2), 1996
6. Van Laer, W., De Raedt, L., Džeroski, S., *On multi-class problems and discretization in inductive logic programming*, In Proceedings ISMIS '97, LNAI 1325, Springer-Verlag, 1997
7. Workshop notes on Discovery Challenge PKDD '99, 1999