ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 489-494

Meaningful discretization of continuous features for association rules mining by means of a SOM.

Marco Vannucci, Valentina Colla
PERCRO - Scuola Superiore Sant'Anna
Polo Sant'Anna Valdera, Viale R. Piaggio 34, Pontedera, Pisa, Italy
e-mail: mvannucci{colla}@sssup.it

Abstract. The paper presents the problem of the unsupervised discretization of continuous attributes for association rules mining. It shows commonly used techniques for this aim and highlights their principal limitations. To overcome such limitations a method based on the use of a SOM is presented and tested over various real world datasets.

1 Introduction

Data Mining is the process of extracting useful hidden knowledge from large volumes of raw data. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either used in an automated decision support system or assessed by a human analyst. Association rules mining [1] is a branch of data mining aimed at finding interesting association relationships among large set of data items. Association rules show attributes values and conditions that frequently jointly occur in a dataset and provide informations in the form of "if-then" statements. The main problem of association rules mining is the extraction of *frequent itemsets* present in the dataset; it consists in finding all those sets of items that appear in the database with a frequency (also called *support*) higher than a prefixed threshold called *minimum support*. This problem has been solved by the *Apriori* algorithm developed by Agrawal et al. in [1] but, unluckly, this method and its numerous derivations (AprioriTID, AprioriHybrid...) work only with categorical attributes. The values of quantitative attributes have to be discretized to use their mapped version in the apriori algorithm [2]. An example of quantitative association rule is: $\langle Age[30..39]\rangle$ *and* $\langle Married[Yes]\rangle \Rightarrow \langle NumCars[2]\rangle$.

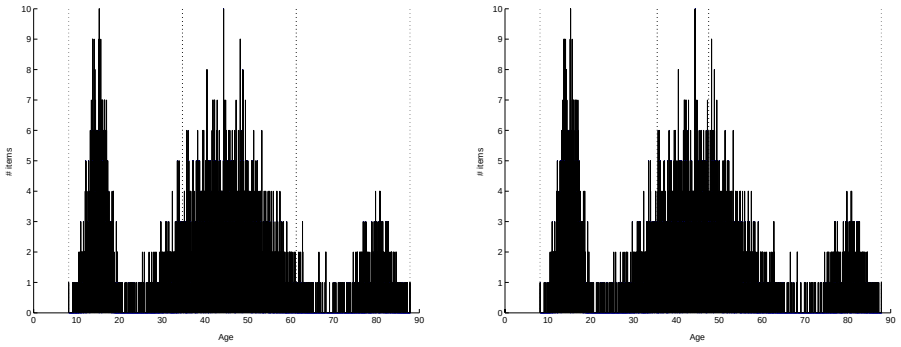
In this paper in Sec. 2 classical supervised and unsupervised discretization techniques are introduced by focusing on unsupervised ones and by discussing the weak points of classical methods. In Sec. 3 a method based on SOM that is able to overcome the limitations of classical methods is presented and its validity is assessed through numerical results shown in Sec. 4.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 489-494

2 Supervised and unsupervised discretization

Literature on discretization is vast [3] but most of works are related to a classification context, where the goal is the maximization of the accuracy on prediction of the value of a particular attribute. In association rule mining and, more generally, in data mining the emphasis is not on predictive accuracy but rather in discovering unknown and useful patterns. When coping with classification problems, each record in the dataset contains a *label-attribute* so as to indicate the class it belongs to and this information is widely used during the supervised discretization of the other attributes. On the contrary, in association rule mining there is almost never a *class-attribute* and records are not labelled, thus supervised discretization is not applicable. Unsupervised discretization methods are generally based on the distribution of attribute values. The simplest and most used discretization method divides the range of observed attribute values into k equal sized intervals. In [2] the optimal number of intervals to create is established, given an arbitrary measure of the maximum information lost due to the discretization, the so–called *partial completeness*. It is also demonstrated that, for any given number of intervals, *Equal Width* (EW) partitioning minimizes the *partial completeness*. A related method, the *equal frequency* (EF) interval, divides the continuous attribute into k intervals where, given m instances, each interval contains m/k values.

When coping with data mining problems, the EW and EF approaches do not grant the following important features: 1) discretization must reflect the original distribution of the attribute; 2) discretized intervals should not hide patterns (if intervals are too big, important discoveries that occur at a smaller resolution may be missed, but, if intervals are too small, there could not be enough data to infer patterns); 3) intervals should be semantically meaningful and must make sense to human expert.



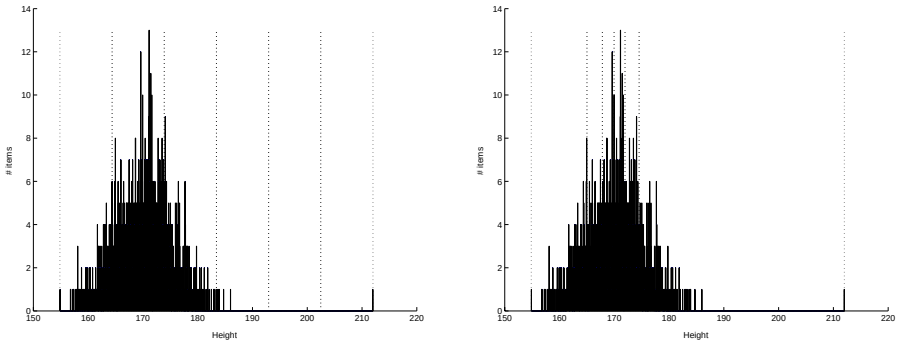
a) b)

Figure 1: EW (a) and EF (b) discretization of the *age* feature

Actually EW and EF methods can fail in discretizing a simple continuous attribute. To show this, a feature representing a sample age distribution can

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 489-494

be used. In Fig. 1 the 3-intervals discretization obtained by both methods is shown. In the EW discretization the second interval starts late and third one early, thus some items that logically belong to the second and third groups are incorrectly included in the previous one. The EF discretization behaves even worse, as it splits the second logical group by creating three completely illogical clusters. The situation is even more dramatic when this methods deal with distributions that present some isolated item placed "out of ranges": Fig. 2, referring to a sample human height distribution, shows that the results of partitioning using classical methods are illogical and unacceptable because these methods do not take care of the attribute distribution but they only consider the variability ranges width or the number of items.



a) b)

Figure 2: EW (a) and EF (b) discretization of the *height* feature

These behaviours are quite frequent (and often more exasperate) when mining rules from real world datasets and this can heavily affect the performance and correctness of every subsequent phase of rule mining process. Gyenesei in [4] tries to overcome these limitations and to satisfy the requirements listed above. His work, based on fuzzy sets theory, divides the variability ranges of each quantitative attribute in various fuzzy sets and each element belongs to a set with a *set membership value* in [0,1]. Nevertheless boundaries and shapes of the fuzzy membership functions must be established by human experts thus this method is not applicable when automatic discretisation is required.

3 SOM–based discretization method

The proposed approach to the unsupervised discretisation problem tries to preserve the original sample distribution. Firstly the k-means algorithm can be exploited, that is a *minimum square error* partitioning method. K-means generates an arbitrary number k of partitions reflecting the original distribution of the partitioned attribute. The original version of such algorithm operates on multidimensional data but it can also be used for partitioning one–dimension

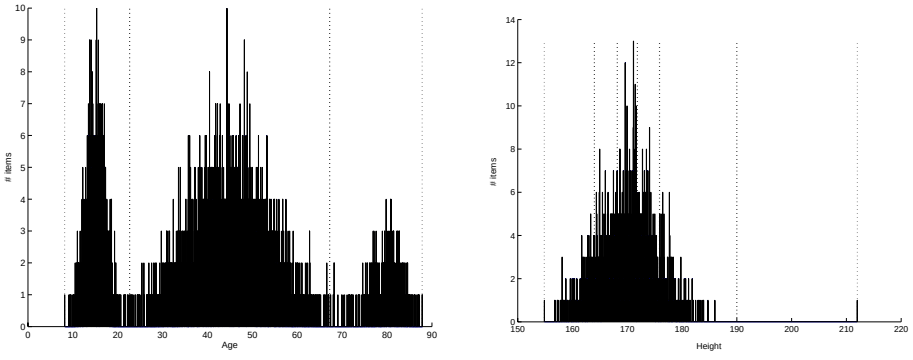
ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 489-494

vectors [5]. The obtained results were quite encouraging, but very sensitive to the value of k, that must be fixed before the computation; an uncorrect estimate of k could lead to unsatisfactory results. In order to avoid this disadvantage, a Self-Organizing Map (SOM) [6] can be used. The SOM, like k-means, partitions the variability ranges of continuous features so as to preserve the distribution; but the number of clusters to create is not a-priori required, only the *maximum* number of desired intervals must be fixed. A 1-dimension SOM is exploited, which is formed by m neurons, where m is the maximum number of desired partitions to divide the continuous attribute values. Attribute values are used as input of the SOM. During the learning process, the weights of the neurons are updated so as to follow faithfully the original attribute distribution, thus neurons play the exact role of centroids in k-means algorithm. The traditional learning algorithm [6] was adopted during the network training; some learning parameters have been tuned as shown in Tab. 1 through many test carried out in order to improve the system performances.

Parameter	**Value at time** t
Learning rate	$\varepsilon(t) = \varepsilon_0 * e^{\frac{-t}{\tau}}$
Updated zone radius	$\sigma(t) = \sigma_0 * e^{\frac{-t}{\tau}}$
Distance function	$d(t) = e^{\frac{-r^2}{2\sigma(t)^2}}$

Table 1: Values of some SOM learning parameters.

The trained SOM will return as output the class of any input attribute value. During the competitive learning phase some neurons could almost never be updated, because each time other neurons are more similar to the value currently taken as input. Neurons that are seldom updated can be pruned in a successive phase as there are no items to classify in their classes. Thus, in practice, an optimal number of classes is formed.



a) b)

Figure 3: Discretizations with the SOM method of a) *Age* and b) *Height* attributes.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 489-494

Figure 3.a shows how the proposed approach manages the *age* attribute presented in Sec. 2. The SOM–based method, by using the same number of intervals as EW and EF methods (such number has been evaluated as optimal by SOM method itself by setting an arbitrary high value, for instance 25, as maximum number of clusters), correctly partitions the attribute values by exactly detecting the 3 logical groups. The discretization of the *height* attribute introduced in Sec. 2 is reported in Fig. 3.b. The SOM method appropriately isolates the item that is *out of ranges* and sharply partition the other ones.

4 Results and comparison

All the discussed discretization methods have been tested over different commonly used real world datasets (see Tab. 2) coming from the *UCI machine learning repository*. Unfortunately an official and widely used measure of the goodness of unsupervised discretization does not exist; here it is proposed as an error measure the weighted mean of the standard deviation of items classified in the same cluster, defined as:

$$\epsilon = \frac{1}{N} \sum_{i=1}^{M} \sigma_i \cdot n_i \tag{1}$$

where M is the number of clusters, σ_i is the standard deviation of the items of class i, n_i is the number of items in class i and N is the total item number.

SOM and k-means methods generally achieve better results than classical methods by reducing the ϵ value up to 63% and by leading to a partitioning more similar to the original distribution of the attribute and more intuitive. The error difference is more pronounced when standard deviation of attribute values is high; in these cases classical methods provide a solution that is unacceptable if compared to the one provided by SOM and k-means methods. Qualitative results of SOM and k-means are very similar, but the SOM–based method seems more efficient than k-means as it reduces the average error to the 50% by respect to classical methods and to the 65% by respect to the k-means based method. In Tab 2 tests marked by † have been performed by using for test a sample distribution different from the one used during the training phase of the system. We note that in these cases the improvement achieved by SOM method by respect to k-means method is higher (53%), probably due to the well-known properties of generalization of neural networks.

5 Conclusions and future work

The problem of unsupervised discretization of continuous attributes for association rules mining has been faced. Widely used EW and EF techniques can lead to a discretization that does not reflect the original distribution of the attribute and this can heavily affect the subsequent process of rule mining by making it incapable of achieving useful results. The k-means and a SOM based

Dataset	No. Data	Std	Clusters	ϵ_{som}	ϵ_{km}	ϵ_{ef}	ϵ_{ew}
Age(1)	15000	17.90	4	3.74	3.80	3.82	5.00
Age(2)†	12000	17.60	3	2.26	3.25	3.24	3.04
Hpw	15000	12.60	5	2.80	3.21	3.12	4.27
Hpw	15000	12.60	3	5.08	8.88	6.91	6.85
Gain	15000	7393	2	1886	2575	7393	2575
Gain†	12000	7394	3	956	2575	2479	7394
Rain rate	52596	77.77	4	12.62	32.69	31.69	49.68
Rain rate	52596	77.77	3	17.80	21.40	77.77	59.38
Rain rate	52596	77.77	6	8.30	21.41	27.00	36.60

Table 2: Results obtained by different discretization methods over real world continuous attributes. Datasets marked by † are those whose distribution is different from the one used for the training.

methods seem capable of overcoming such limitation. Both techniques have been deeply tested over classical real-world datasets. Results prove that both methods obtain good results. In particular the SOM based one seems to be the best performing showing also an interesting generalization capability. Future work will concern the integration of SOM and k-means methods in order to improve the results; more tests will be carried out also with databases obtained from industrial applications. Moreover the *a priori* algorithm will be used for the extraction of association rules based on the obtained discretizations in order to evaluate this way the goodness of the obtained rules.

Acknowledgments

The authors wish to thank Prof. Beatrice Lazzerini Bosio and Dr. Pablo Rossi for their fruitful discussions which led to the present analysis.

References

[1] R.Agrawal, T.Imielinski,A.Swami *Mining association rules between sets of items in large databases.* Proc. ACM SIGMOD Conf. on management of data, pages 207-216, Washington, D.C, May 1993.

[2] R.Srikant, R.Agrawal *Mining quantitative association rules in large relational tables* Proc. 1996 ACM SIGMOD Int. Conf. on Management of Data.

[3] J.Dougherty, R.Kohavi, M.Sahami *Supervised and unsupervised discretization of continuous features* In A. Prieditis and S. Russell, eds., Machine learning Proc. 12th Int. Conf., 1995, Morgan Kaufmann Publishers, San Francisco, CA.

[4] A.Gyenesei A *fuzzy approach for mining quantitative association rules* Acta cybernetica vol.15(2), 2001.

[5] W.Dillon, M.Goldstein *Multivariate analysis* New York: Wiley, 1984.

[6] T.Kohonen, *The self-organizing map* Proc. IEEE, Vol. 78, No. 9, pp. 1464-1480, 1990.