Numeric Landmarks to Obtain a Discretization Reflecting Output Qualitative Order *

N. Agell[1], X. Rovira[1], M. Sánchez[2], F. Prats[2], F. Ruiz[3]
[1] ESADE, Universitat Ramon Llull, Av. Pedralbes, 60-62, 08034 Barcelona
[2] MA2, Universitat Politècnica de Catalunya, C. Jordi Girona 1-3, 08034 Barcelona
[3] ESAII, Universitat Politècnica de Catalunya, Av. Victor Balaguer, s/n, 08800 Vilanova i la Geltrú
{nuria.agell,xari.rovira}@esade.edu, {monica.sanchez, francesc.prats, francisco.javier.ruiz}@upc.edu

Abstract

This paper lies within the domain of supervised discretization methods. The methodology aims at identifying relevant interactions between input and output variables. A new supervised discretization algorithm that takes into account the qualitative ordinal structure of the output variable is proposed. Most existing supervised discretization methods are designed for pattern recognition problems and do not take into account this ordinal structure.
A qualitative distance is constructed over the discrete structure of absolute orders of magnitude spaces. The algorithm presented implements a maximization process of this distance. A simple example allows interpretation of the process of choosing landmarks.

1 Introduction

Qualitative Reasoning applications aim at defining suitable models to automate common-sense and expert reasoning, working without numerical values. On the other hand, the design of algorithms able to automatically gather the relevant information from a set of patterns is one of the primary aims of Artificial Intelligence (AI).
When defining a model able to express simultaneously the qualitative relations between variables and the expert knowledge of a specific domain, an algorithm to find a set of landmarks capturing the essential distinctions is necessary. A suitable discretization process can carry out this objective over features of the patterns. A key point is to ensure that the discretization obtained generates a qualitative model containing only the essential changes in the application domain.
In this paper, a new supervised discretization algorithm is proposed that takes into account the qualitative ordinal structure of the output variable, to automate relevant landmark generation for each feature. Moreover, this algorithm allows the homogenizing of scales by associating qualitative labels to each pattern that reflect its values in terms of their significance in the domain considered. The methodology is based on a distance defined over an order of magnitude structure.
The next section begins with a brief overview of some classical discretization methods, and introduces the motivation of the present methodology. A distance in the orders of magnitude model is then defined in section 3. This distance, needed to build the supervised discretization algorithm, is described in detail in section 4. An application of the method is given in section 5. The last section summarizes the conclusions and outlines future work.

2 Discretization Framework

Discretization is the process of partitioning continuous variables. There are many advantages in using discrete as opposed to continuous values. In general, discretization makes learning faster, and the results obtained are more compact and easier to understand [Liu et al, 2002]. Moreover, there are some classification-learning algorithms that are only able to deal with discrete values.
The existing discretization methods in the literature can be divided into two groups: supervised and unsupervised.
The most commonly used discretization methods, based on equal-width or equal-frequency, are considered to be unsupervised methods, because they do not use class information. On the other hand, when class information is available and used, supervised discretization methods provide better results by taking into account this information to find meaningful intervals in the range of continuous input variables. These supervised methods improve the performance of the learning process [Dougherty, 1995] and, at the same time, enhance understanding of the results.
Usually, in a supervised discretization process, after sorting data in ascending or descending order with respect to the variable to be discretized, landmarks must be chosen among the whole dataset. In general, the algorithm for choosing landmarks can be either top-down, which starts with an empty list of landmarks and splits intervals, or bottom-up, which starts with the complete list of all the values as landmarks and merges intervals. In both cases there is a stopping criterion, which specifies when to stop the discretization process.

* This work has been funded by the MCyT (Spanish Ministry of Science and Technology) MERITO project (TIC2002-04371-C02).

Some representative supervised discretization methods are described below. First, those based on the entropy measure, among others MDLP [Fayyad and Irani, 1993] and D2 [Catlett, 1991]. These define a function measuring the entropy of each possible discretization to be optimized. There are also some decision tree induction algorithms that use entropy measurement, such as ID3 [Quinlan, 1986] and C4.5 [Quinlan, 1993], to implement the discretization process.
Other methods are based on statistical techniques such as χ^2 test or classical clustering techniques, for instance Chi-Merge [Kerber, 1992], Chi2 and ConMerge [Wang and Liu, 1998].
Finally, there are some methods based on a strength association measurement between the class and a feature, for example Zeta [Ho and Scott, 1997], CADD [Ching et al 1995], CAIR [Wong and Liu 1975], and CAIM [Kurgan and Cios 2004].
Most of these methods are based on the optimization of a coefficient, which depends solely on the contingency table between the discretized variable (*D*) obtained in the discretization process and the output variable (*O*). See Table 1, in which q_{ir} represents the number of elements in the interval $(d_{r-1},d_r]$ classified in class C_i.

Class	*Interval*					*total classes*
	$[d_0,d_1]$	…	$(d_{r-1},d_r]$	…	$(d_{n-1}, d_n]$	
C_1	q_{11}	...	q_{1r}	...	q_{1n}	M_{1+}
...	...	...	...	...	...	...
C_i	q_{i1}	...	q_{ir}	...	q_{in}	M_{i+}
...	...	...	...	...	...	...
C_C	q_{C1}	...	q_{Cr}	...	q_{Cn}	M_{C+}
total intervals	M_{+1}	...	M_{+r}	...	M_{+n}	M

Table 1. Contingency table.

These methods do not consider the eventual order in the set $\{C_1,...C_C\}$ of the possible output values. However, the method proposed and presented in this paper, which could be classified as belonging to this last group of discretization methods, takes this order into account, and is based on the concept of distance between the ordered output labels introduced in the next section. The algorithm is suitable when the output is described in terms of a qualitative ordered variable and is neither top-down nor bottom-up: it allows all the landmarks to be found simultaneously.

3 Building a distance in the absolute orders of magnitude space OM(n)

Absolute orders of magnitude models [Travé, 2003] work with a finite set of ordered symbols, or qualitative ordered labels. These models provide a mathematical structure that unifies sign algebra and interval algebra through a continuum of qualitative structures. In this section, the absolute orders of magnitude model is briefly described [Agell, 1998], and a methodology to build distances between labels is considered.

The absolute orders of magnitude model of granularity n, OM(n), is defined from a symmetric partition of the real line in 2n+1 classes:

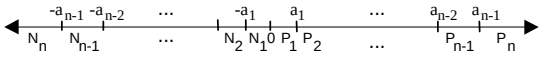


Fig.1. Partition of the real line

where $N_i=[-a_i,-a_{i-1})$, $0=\{0\}$ and $P_i=(a_{i-1},a_i]$.
In the absolute orders of magnitude model of granularity n, OM(n), the *basic elements*, using the notation introduced in [Agell, 1998], are represented by the ordered labels of the set S_1:

$$S_1=\{N_n, N_{n-1}, ..., N_1, 0, P_1, ..., P_{n-1}, P_n\}.$$

The *quantity space* $S= OM(n)$, is the set of labels in the form $[X,Y]$ for all $X,Y\in S_1$, with $X\leq Y$, i.e. $x\leq y$ for all $x\in X$, $y\in Y$. The interval $[X,Y]$ stands for:

$$[X,Y]=\bigcup_{X\leq L\leq Y} L,$$

which is the union of all labels in S_1 between X and Y. Obviously $S_1\subset S$.

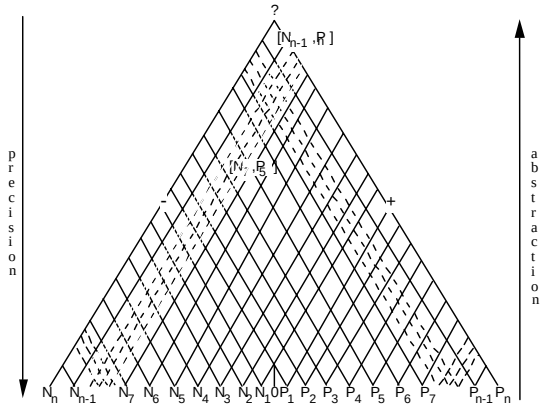


Figure 1. The quantity space

The relation $\leq_P$, *to be more precise than* (given that $X,Y\in S$, X is more precise than Y ($X\leq_P Y$) if $X\subseteq Y$) is an order relation in S, due to inclusion properties.
For all $X\in S-\{0\}$, *the basis of* X is the set:

$$B_X=\{B\in S_1-\{0\}: B\leq_p X\},$$

and, given a basic element $U\in S_1$, the *U-expansion* of X is:

$$X_U=Min\{Y\in S: X\leq_P Y \text{ and } U\leq_P Y\},$$

the minimum label that is *less precise* than X and U, i.e., the smallest interval with respect to the inclusion containing X and U.
In order to define a distance in an OM(n) structure, the strategy proposed in this paper is split into two steps:

- First a *location function* is considered, to associate a k-dimensional real vector to each label
- Then a metric defined in a Euclidean space R^k is used.

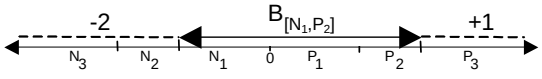
This location function and the metric must be chosen to capture the intrinsic values and significance of labels in the qualitative space, depending on the scenario defined by the application domain.
As an example of this methodology, the location function defined in [Rovira et al, 2004] is first considered. By this function, each element X in S is codified by a pair $(l_1(X), l_2(X))$ of integers: $l_1(X)$ is the number of basic elements in S_1-{0} that are "between" the basis of X and N_n, and $l_2(X)$ is the number of basic elements in S_1-{0} that are "between" the basis of X and P_n.
These numbers permit us to "locate" each element in S, where all different levels of precision are considered.
As an example to illustrate this process, let us consider the absolute orders of magnitude model with granularity 3, OM(3):

$$S_1=\{N_3, N_2, N_1, 0, P_1, P_2, P_3\},$$

and

$$S=S_1 \cup \{[N_3,N_2],[N_2,N_1],[N_1,P_1],[P_1,P_2],[P_2,P_3],[N_3,N_1],[N_2,P_1],[N_1,P_2],[P_1,P_3],[N_3,P_1],[N_2,P_2],[N_1,P_3],[N_3,P_2],[N_2,P_3],?\}.$$

The location of the label $X=[N_1,P_2]$ is the pair (-2,1), because we are left with two basic elements to the left of X and only one to its right:



The formal definition of the *location function* is $l: S \to Z^2$ such that:

$$l(X) = (l_1(X), l_2(X)) =$$
$$(-Card(B_{X_{N_n}}) + Card(B_X), Card(B_{X_{P_n}}) - Card(B_X))$$

This is a way of codifying labels by points in a Euclidian plane, in such a manner that the Euclidean distance between them will allow the definition of a distance between labels.
Let us define:

$$D: S \times S \to [0, +\infty)$$
$$(X,Y) \to \sqrt{(l_1(X)-l_1(Y))^2 + (l_2(X)-l_2(Y))^2}$$

This function D inherits all properties of the distance in R^2, and therefore satisfies the three axioms of a distance.
The distance D between two labels measures the similarity between them, in the sense that the more similar labels are, the smaller the distance between their codifications, and so the smaller their distance $D(X,Y)$.

4 The new supervised discretization algorithm

In this section, a new supervised discretization algorithm is presented. This algorithm is suitable specifically when the output is described in terms of qualitative orders of magnitude or in any interval-based domain. It is based on the concept of distance between qualitative labels. This method is neither top-down nor bottom-up, and it allows all the landmarks to be found simultaneously. This fact improves the algorithmic efficiency.
Let us consider M input features $F_1,...,F_M$, and a training set $\{X_1,..., X_N\}$ of N patterns. Each pattern X_i is characterized by a set of values of the M input features together with the output: $X_i = (x_{i1},...,x_{iM}, y_i)$. Each x_{ij} is the value of F_j for the pattern X_i, and the output y_i is a value of a variable described in a qualitative orders of magnitude space OM(n).
The following discretization algorithm will be applied to each of the input variables separately. The case in which the landmarks of the various input variables are not independent, which leads to dependent discretizations, is not dealt with in this paper.
Let $F = F_j$ be one of the M continuous input features $F_1,...,F_M$ to be discretized. A discretization D of this variable F consists of a set of disjoint intervals:

$$D=\{[d_0, d_1], (d_1,d_2],...(d_{m-1}, d_m]\}$$

where d_0 and d_m are the extreme values of F, and the rest of the landmarks d_k are chosen from among the pattern values $x_{1j},...x_{Nj}$ of F.
This new method takes as possible landmarks all the pattern values of F. The criterion applied considers a landmark d as *suitable* when it splits a neighbourhood of d into two meaningful different intervals, in order to ensure that the discretization obtained generates a qualitative model containing the essential changes in the application domain.
The key idea is to implement a criterion to distinguish meaningful different adjacent intervals consisting of a maximization process of the qualitative distance. This criterion is based on five concepts: labels of the outputs of the neighbours, size of the neighbourhood previously fixed via a parameter L, percentage P of neighbours considered when computing qualitative labels, qualitative distances between output labels, and biggest local maxima.

First, for each pattern X_i only its value x_{ij} corresponding to F and its output value y_i are considered. Secondly, the values $x_{1j},...x_{Nj}$ of F are ordered in a non-decreasing sequence: $x_{i(1)j} \leq ... \leq x_{i(N)j}$ (note that the extreme values are $d_0 = x_{i(1)j}$ and $d_m = x_{i(N)j}$). Let us rename this sequence as $d_0 \leq ... \leq d_N$. The candidates for landmarks are $d_1,...,d_{N-1}$.
The steps needed to determine suitable landmarks among these candidates are the following:

- A positive integer $L < N/2$ and a positive parameter $P < 100$ are previously fixed.
- For a candidate landmark d_i, let us consider the two sets of pattern values at the left and right sides of d_i, respectively: $D_- = \{d_{i-L},...,d_{i-1}\}$, with the convention that if at the left side of d_i there are less than L values, D_- contains these, and $D_+ = \{d_{i+1},...,d_{i+L}\}$, with a similar convention.
- Let $S_P(D_-)$ be the most precise qualitative expression of the outputs corresponding to, at least, $P\%$ of the patterns whose values are in D_-
- If different qualitative expressions of the same precision satisfy the latter condition, the one corresponding to a larger number of patterns is chosen.
- The same process applied to D_+ gives the qualitative expression $S_P(D_+)$.
- The qualitative distance, defined in section 3, between $S_P(D_-)$ and $S_P(D_+)$ is associated to the candidate landmark d.
- The candidate landmarks chosen are those associated to the biggest local maxima of the distance function. These are the suitable landmarks of the discretization.

Before presenting the pseudo-code of this algorithm, some observations about the different steps must be made. The fixed parameter L is associated to the desired level of discretization. That is to say, a large value of L leads to a reduced number of landmarks, and a small L will increase this number.
The association of each landmark to the qualitative expressions, related with the outputs, gives them a representation that is directly linked to the proposed learning problem.
The pseudo code used to obtain $S_P(D_-)$ is the following:

left: l=1
right: r=2n
$S_P(D_-)=[l,r]$
(corresponds to the smallest possible precision interval).

$C=L$ *(number of patterns in $S_P(D_-)$)*

If $F_l < F_r$
  If $(C-F_l)\cdot 100/L > P$ *then*
    $l=l+1$ *(eliminates the label on the left)*
    $C=C-F_l$
  else end
else
  If $(C-F_r)\cdot 100/L > P$ *then*
    $r=r-1$ *(eliminates the label on the right)*
    $C=C-F_r$
  else end
end if

where the qualitative basic labels are:
$N_n=1, N_{n-1}=2, ..., P_n=2n,$
and F_1, $F_{2,...}F_{2n}$ *are the frequencies of patterns with output values of, respectively,* $N_n, N_{n-1,...}$ P_n.

In the next section, the effects of the parameters L and P in the determination of the landmarks is heuristically discussed. Some graphs of the distance function are shown in order to observe the chosen local maxima.

5 An example

To illustrate the proposed discretization method, a set of 200 patterns is considered. The patterns are characterized by a continuous input variable F and a qualitative ordered output O in an OM(3). Data have been generated to obtain ordered values of F. Values between 0 and 50 in F are associated to output values from N_3 to P_3; between 50 and 100 are associated with values between N_3 and P_1; from 100 to 150 the outputs are between N_1 and P_3, and the last 50 patterns take values N_3 and N_2. The discretization method will be able to determine these three landmarks (50, 100 and 150).

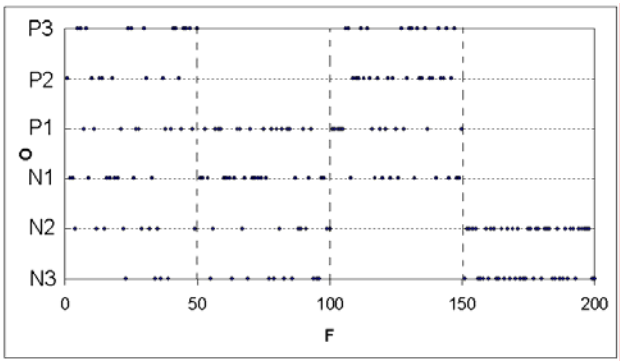


Figure 2. Distribution of the data.

Distances associated to each possible landmark are shown in figures associated to different values of L and P. Figure 3 shows results for the same value of parameter P and three different values of L. Better results are observed by avoiding small values of L. A reduced value of this parameter leads to too many non-desired maxima, hiding the suitable landmarks.

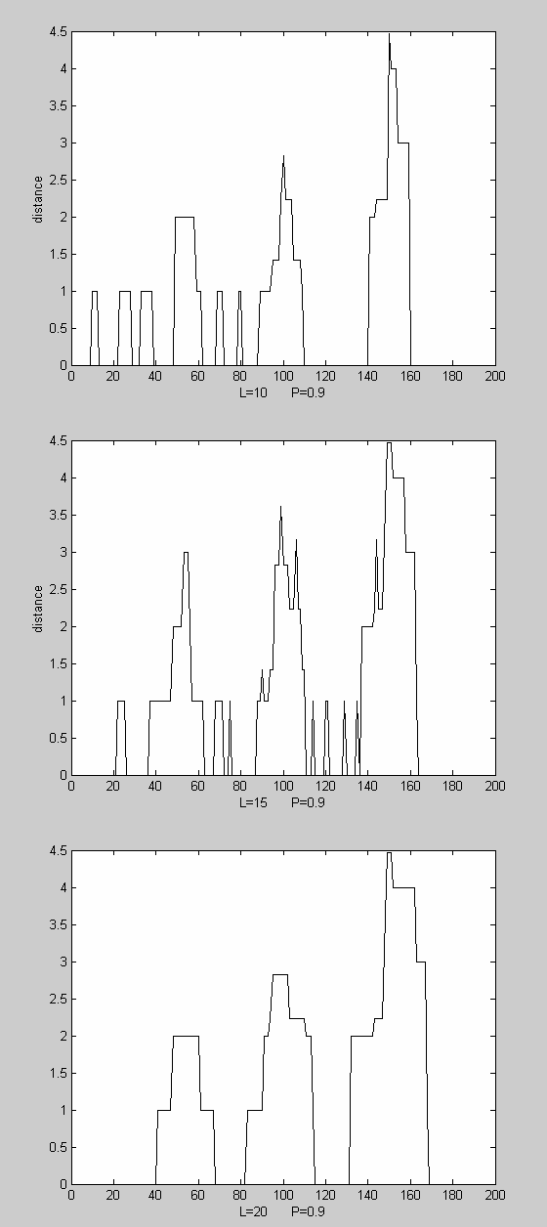


Figure 3. Results varying parameter *L*.

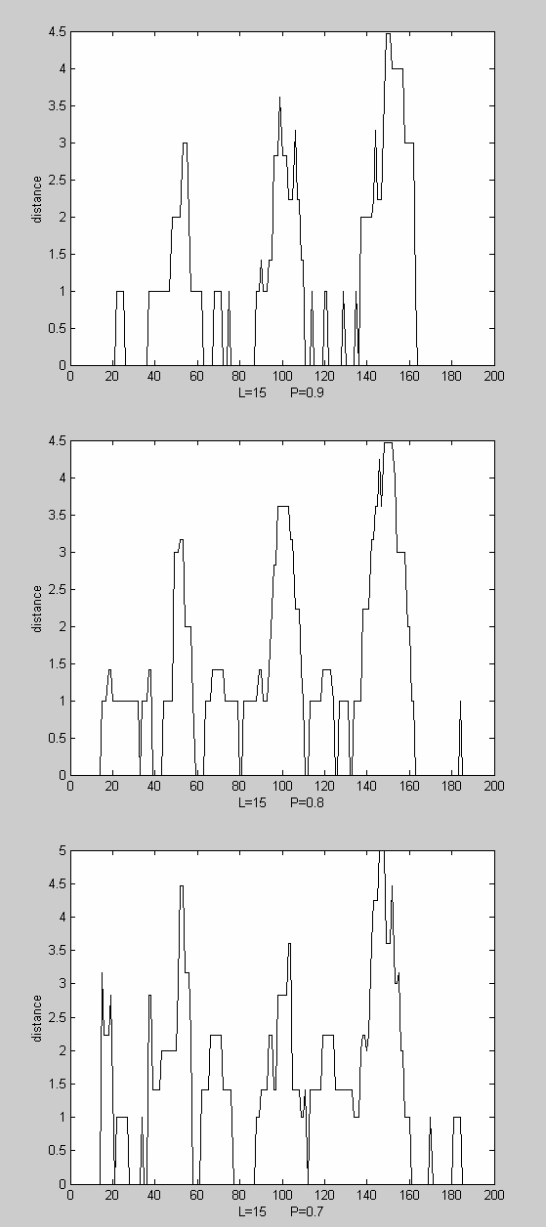


Figure 4. Results varying parameter *P*.

Regarding parameter *P*, which refers to the frequency of patterns associated to the qualitative expressions, figure 4 shows results for the same value of *L* and three different values of *P*. The best results are obtained with values between 0.7 and 1. Nevertheless, the value 1 means considering the less precise qualitative expression and therefore taking into account the outliers in the process of discretization.

6 Conclusions and future work

The present work aims at motivating, defining and analysing the use of supervised discretization algorithms in models based on orders of magnitude spaces.

The focus of this paper is the design of an automatic algorithm to be used in problems for which the output variable is described in terms of qualitative values of orders of magnitude. For this reason a distance between qualitative values is introduced and used to maximize the distinction between contiguous labels.

The following discretization algorithm will be applied to each of the input variables separately. The case in which the landmarks of the various input variables are not independent, which has to lead to dependent discretizations, is a matter of on-going research.

Although this paper has focused on the discretization of each of the variables separately, the methodological aspects considered can be used in a more complex situation.
When considering an OM(n), different approaches can be analysed to measure the distance between labels. It seems to be reasonable to look for other suitable distances in these kinds of sets.
With regard to open problems and future work, the following comments can be made:

- To define new criteria for choosing landmarks related to landmarks in other input variables.
- To define new distances between qualitative orders of magnitude labels.
- To apply the given method in real problems of ranking or ordered multi-classification might be considered.

In particular, the methodology given in this paper is going to be used within the MERITO project, supported by the Spanish Ministry of Science and Technology. The project addresses the prediction and measurement of financial credit risk.

References

[Abelson *et al*., 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation of Computer Programs.* MIT Press, Cambridge, Massachusetts, 1985.

[Agell, 1998] Núria Agell. *Estructures matemàtiques per al model qualitatiu d'ordres de magnitud absoluts*. Ph. D. Thesis. Universitat Politècnica de Catalunya, 1998.

[Agell et al., 2000] Agell, N., Rovira, X., Ansotegui, C., Sánchez, M. and Prats, F. Homogenising References in Orders of Magnitude Spaces: An Application to Credit Risk Prediction. In *Proceedings of the 14th International Workshop on Qualitative Reasoning (QR'00).* Morelia, Mexico, 2000.

[Catlett, 1991] Catlett J. On changing continuous attributes into ordered discrete attributes. *Proc. Fifth European Working Session on Learning*. Berlin: SpringerVerlag, pp. 164-177, 1991.

[Ching, 1995] Ching, J.Y., Wong, A.K.C. and Chan, K.C.C. Class-Dependent Discretization for Inductive Learning from Continuous and Mixed Mode Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 17, no. 7, pp. 641-651, 1995.

[Dougherty, 1995] Dougherty, J., Kohavi R. and Sahami, M. Supervised and Unsupervised Discretization of continuous-valued attributes for classification learning *In International Conference on Machine Learning*, pp.194-202, 1995.

[Fayyad and Irani, 1993] Fayyad, U. and Irani, K. Multi-interval discretization of continuous-valued attributes for classification learning. *Proc. 13th International Joint Conference on Artificial Intelligence.* San Mateo, California. Morgan Kaufmann Ed. pp. 1022-1027, 1993.

[Ho and Scott, 1997] Ho, K.M and Scott, P.D. Zeta: A global method for discretization of continuous variables. *In KDD97: 3rd International Conference of Knowledge Discovery and Data mining.* Newport Beach, CA, pp. 191-194, 1997.

[Kerber, 1992] Kerber, R. ChiMerge: Discretization of Numeric Attributes. *Proc. 10th National Conference on Artificial Intelligence.* MIT Press, pp. 123-128, 1992.

[Kurgan and Cios, 2004] Kurgan, L.A and Cios, K.J CAIM Discretization Algorithm. *IEEE Transactions on Knowledge and Data Engineering,* vol. 16, no.2. pp. 145-153, 2004.

[Liu et al., 2002] Liu, H., Hussain, F., Lim Tan, C., Dash, M., Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery* 6, Kluwer Academic Publisher, pp. 393-423, 2002.

[Quinlan, 1986] Quinlan, J.R. *Induction of decision trees.* Machine Learning, 1, pp. 81-106, 1986.

[Quinlan, 1993] Quinlan, J.R. 1993. C4.5: *Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[Rovira et al., 2004] Rovira, X., Agell, N., Sánchez, M., Prats, F. and Parra, X. An Approach to Qualitative Radial Basis Function Networks over Orders of Magnitude. *Proceedings of 18th International Workshop on Qualitative Reasoning*. 2004.

[Travé-Massuyès and Dague, 2003] Travé-Massuyès, L. and Dague, P. *Modèles et raisonnements qualitatifs.* Hermès, 2003.

[Wang and Liu, 1998] Wang, K., and Liu, B. Concurrent discretization of multiple attributes. *Pacific-Rim International Conference on AI.* pp. 250-259, 1998.

[Wong and Liu, 1975] Wong, A.K.C. and Liu, T.S.: *Typicality, diversity and feature pattern of an ensemble*, IEEE Trans. Computers, vol. 24, pp.158-181, 1975.