Fuzzy Partitioning of Quantitative Attribute Domains by a Cluster Goodness Index

Attila Gyenesei

Turku Centre for Computer Science

TUCS Technical Reports

No 368, October 2000

Fuzzy Partitioning of Quantitative Attribute Domains by a Cluster Goodness Index

Attila Gyenesei

University of Turku, Department of Computer Science,
Lemminkaisenkatu 14, FIN-20520 Turku, Finland

TUCS

Turku Centre for Computer Science
TUCS Technical Report No 368
October 2000

ISBN 952-12-0736-1
ISSN 1239-1891

Abstract

The problem of mining association rules for fuzzy quantitative items was introduced and an algorithm proposed in [7]. However, the algorithm assumes that fuzzy sets are given. In this paper we propose a method to find the fuzzy sets for each quantitative attribute in a database by using clustering techniques. We present a scheme for finding the optimal partitioning of a data set during the clustering process regardless of the clustering algorithm used. More specifically, we present an approach for evaluation of clustering partitions so as to find the best number of clusters for each specific data set. This is based on a goodness index, which assesses the most compact and well-separated clusters. We use these clusters to classify each quantitative attribute into fuzzy sets and define their membership functions. These steps are combined into a concise algorithm for finding the fuzzy sets. Finally, we describe the results of using this approach to generate association rules from a real-life dataset. The results show that a higher number of interesting rules can be discovered, compared to partitioning the attribute values into equal-sized sets.

Keywords: association rules, fuzzy items, clustering, linguistic terms

TUCS Research Group
Algorithmics Research Group

1 Introduction

Data mining, sometimes referred to as knowledge discovery in databases, is concerned with the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1]. Mining association rules is one of the important research problems in data mining. An interesting association rule describes an interesting relationship among different attributes. The problem of mining boolean association rules over basket data was introduced in [2], and later broadened in [3] for the case of databases consisting of categorical attributes alone. Several efficient algorithms for mining categorical association rules have been published (see [3], [4], [5] for just a few examples). In practice the information in many, if not most, databases is not limited to categorical attributes, but also contains much quantitative data. The problem of mining quantitative association rules was introduced and an algorithm proposed in [6]. The algorithm involves discretizing the domains of quantitative attributes into intervals so as reduce the domain into a categorical one. An example of a rule according to this definition would be: "10% of married people between age 50 and 70 have at least 2 cars". However, these intervals may not be concise and meaningful enough for human experts to easily obtain nontrival knowledge from those rules discovered.

In [7], we showed a method to handle quantitative attributes using a fuzzy approach. Instead of using intervals, the method employs linguistic terms to represent the revealed regularities and exceptions. We assigned each quantitative attribute several fuzzy sets which characterize it. Fuzzy sets provide a smooth transition between a member and non-member of a set. The fuzzy association rule is also easily understandable to a human because of the linguistic terms associated with the fuzzy sets. Using the fuzzy set concept, the above example could be rephrased e.g. "10% of married old people have several cars".

However, the algorithm proposed in [7] for fuzzy association rule mining suffers from the following problem. The user or an expert must provide this algorithm the required fuzzy sets of the quantitative attributes and their corresponding membership functions. It is unrealistic to assume that experts can always provide the fuzzy sets of the quantitative attributes in the database for fuzzy association rule mining. To deal with this problem, we intend to find the fuzzy sets by using clustering techniques.

In this paper, we present an approach for clustering scheme evaluation. It aims at evaluating the schemes produced by a specific clustering algorithm, assuming different input parameter values. These schemes are evaluated using a new clustering scheme validity index, which we define. Our goal it is not to propose a new clustering algorithm or to evaluate a variety of clustering algorithms, but to produce the clustering scheme with the most compact and well-separated clusters for any given algorithm.

The remainder of the paper is organized as follows. In the next section we describe the proposed goodness index for clustering scheme evaluation. In Section 3, we exploit the discovered cluster centers, to classify the quantitative attribute values into fuzzy sets, and show a method to find the corresponding membership function for each fuzzy set. Then we formulate our approach into a precise algorithm in Section 4. In Section 5 the experimental results are reported, comparing obtained association rules both qualitatively and quantitatively. The paper ends with a brief conclusion in Section 6.

2 Clustering Scheme Evaluation

The objective of the clustering methods is to provide in some sense optimal partitions of a data set. In general, they should search for well separated clusters whose members are close to each other. Another problem in clustering is to decide the optimal number of clusters that fits best a data set. The majority of clustering algorithms produce a partitioning based on the input parameters (e.g. number of clusters, minimum density) that finally lead to a finite number of clusters. Thus, the application of an algorithm assuming different input parameter values results in different partitions of a particular data set, which are not easily comparable. A solution to this problem is to run the algorithm repetitively with different input parameter values and compare the results against a well-defined validity index.

A number of cluster validity indices are described in the literature. A cluster validity index for crisp (i.e. non-fuzzy) clustering proposed in [8], attemps to identify "compact and separated clusters". Other validity indices for crisp clustering have been proposed in [9] and [10]. The implementation of most of these measures is very expensive computationally, especially when the number of clusters and number of objects in the data set grow very large [11]. Other validity measures are proposed in [12], [13]. We should mention that the evaluation of proposed measures and the analysis of their reliability have been quite limited.

In the following, we define a goodness index for evaluating clustering schemes based on the validity index defined for the fuzzy c-means method (FCM) in [13]. We use the same concepts for validation, but the goodness index can be used for any clustering method, not just for FCM. Assume that we study a quantitative attribute X.

Definition 1. *The* ***variance of an attribute X****, denoted $\sigma^2(X) \in \mathbb{R}$, is defined as*

$$\sigma^2(X) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2,$$

where $x_1, \ldots, x_n$ are the attribute instances, and $\overline{x}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Definition 2. *The* ***variance of cluster i*** *containing elements $X_i = \{x_{i_1}, \ldots, x_{i_n}\}$ is given by*

$$\sigma^2(X_i, r_i) = \frac{\sum_{k=1}^{n_i}(x_{i_k} - r_i)^2}{n_i},$$

where r_i is the center of cluster i, having n_i elements.

Definition 3. *The* ***average scattering (separation) for c clusters*** *is defined as*

$$Scat(X, R) = \frac{\frac{1}{c}\sum_{i=1}^{c}\sigma^2(X_i, r_i)}{\sigma^2(X)},$$

where R is the set of c cluster centers.

$Scat(X, R)$ indicates the average compactness of clusters. A small value for this term indicates compact clusters and as the scattering within clusters increases (they become less compact) the value of $Scat(X, R)$ also increases.

Definition 4. *The* ***total separation between clusters*** *is given by*

$$Dis(R) = \frac{D_{max}}{D_{min}}\sum_{i=1}^{c}\left(\frac{1}{\sum_{j=1}^{c}|r_i - r_j|}\right),$$

where D_{max} is the maximum, and D_{min} is the minimum distance between cluster centers.

The term 'total separation' sounds like a measure that we want to maximize. However, here the opposite holds: a smaller value is better. $Dis(R)$ indicates the total separation (scattering) between the c clusters, and generally, this term will increase with the number of clusters.
Now, we can define our goodness index based on the last two definitions.

Definition 5. *The* ***goodness index for clusters R*** *within set X is as follows:*

$$G(X, R) = \alpha \cdot Scat(X, R) + Dis(R),$$

where α is a weighting factor equal to $Dis(c_{max})$, c_{max} is the maximum number of input clusters.

The goodness index uses cluster separation as an indication of the average scattering between clusters. Minimizing the separation thus also tends to minimize the possibility to select a cluster scheme with significant differences in cluster distances. Since the two terms of goodness index are of different ranges, a weighting factor is needed in order to incorporate both terms in a balanced way. (Note that the influence of the weighting factor is an issue for further study as mentioned in [13].)

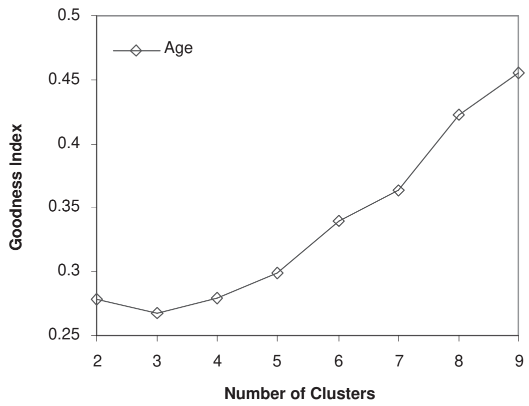

Figure 1: Example of Goodness Index for the Attribute Age

For example, Fig. 1 shows the values of the goodness index as a function of the number of clusters for attribute Age, which is given in the Section 5. We can see that the best number of clusters is three for this dataset.

3 Determining Fuzzy Sets by Using the Discovered Cluster Scheme

After we have obtained the best cluster scheme (i.e. centers of clusters), we can use this to classify the quantitative attribute values into c fuzzy sets. We divide the attribute interval into c sub-intervals by using the discovered r_i values, with a coverage of p percent between two adjacent ones, and give each subinterval a symbolic name related to its position (Fig. 2).

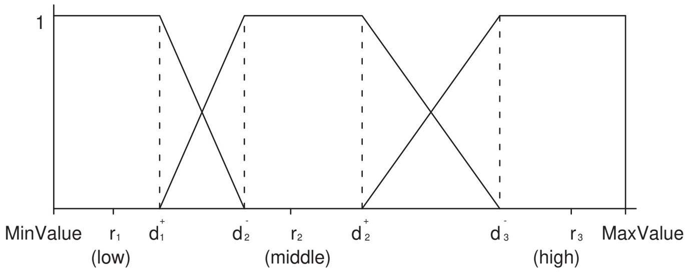

Figure 2: Example of the proposed fuzzy partitions

To specify our heuristic method, we give the following definitions.

Definition 6. *The* ***effective upper bound****, denoted d_i^+ for fuzzy set i, is given by:*

$$d_i^+ = r_i + 0.005(100 - p)(r_{i+1} - r_i),$$

where p is the overlap parameter in %, and r_i is the center of cluster i, $i = \{1, 2, \ldots, c-1\}$. d_i^+ is also the ***fuzzy lower bound*** *of cluster $i+1$.*

Definition 7. *The* ***effective lower bound****, denoted d_j^- for fuzzy set j, is as follows:*

$$d_j^- = r_j - 0.005(100 - p)(r_j - r_{j-1}),$$

where p is the overlap parameter in %, and r_j is the center of cluster j, $j = \{2, 3, \ldots, c\}$. d_j^- is also the ***fuzzy upper bound*** *of cluster $j-1$.*

Notice that $0.005(100 - p) = \frac{1}{100}\left(\frac{100-p}{2}\right)$. These definitions become clear by inspecting Fig. 2. To quote an example, we classify the attribute *Age* into three fuzzy sets as given in Table 1, where *Age* ranges from 15 to 90.

Table 1: The ranges of fuzzy set Age ($p = 30\%$)

Fuzzy set	**Range**	**Cluster center**
(Age,young)	15 to 43.95	31.65
(Age,middle)	38.78 to 65.80	50.58
(Age,old)	58.77 to 90	73.99

In the following, we describe how to generate the corresponding membership function for each fuzzy set of a quantitative attribute.

Let $\{r_1, r_2, \ldots, r_i, \ldots, r_c\}$ be the cluster centers for a quantitative attribute. We use the following formulas to define the required membership functions for each fuzzy set.

For the fuzzy set with cluster center r_1, the membership function for element x is given by

$$f(r_1, x) = \begin{cases} 1 & \text{if } x \le d_1^+ \\ \frac{d_2^- - x}{d_2^- - d_1^+} & \text{if } d_1^+ < x < d_2^- \\ 0 & \text{if } x \ge d_2^- \end{cases}$$

For the fuzzy set with the cluster center r_i, where $2 \le i < c$, the membership function for element x is given by

$$f(r_i, x) = \begin{cases} 0 & \text{if } x \le d_{i-1}^+ \\ \frac{d_{i-1}^+ - x}{d_{i-1}^+ - d_i^-} & \text{if } d_{i-1}^+ < x < d_i^- \\ 1 & \text{if } d_i^- \le x \le d_i^+ \\ \frac{d_{i+1}^- - x}{d_{i+1}^- - d_i^+} & \text{if } d_i^+ < x < d_{i+1}^- \\ 0 & \text{if } x \ge d_{i+1}^- \end{cases}$$

For the fuzzy set with cluster center r_c, the membership function for element x is given by

$$f(r_c, x) = \begin{cases} 0 & \text{if } x \le d_{c-1}^+ \\ \frac{d_{c-1}^+ - x}{d_{c-1}^+ - d_c^-} & \text{if } d_{c-1}^+ < x < d_c^- \\ 1 & \text{if } x \ge d_c^- \end{cases}$$

4 An Algorithm for Finding Fuzzy Sets by Using a Clustering Scheme Goodness Index

In Section 2 we have defined a goodness index for clustering scheme evaluation. We exploit this index during the clustering process in order to define the optimal number of clusters for a quantitative attribute. More specifically, we first define the range of input parameters (e.g. number of clusters) of a clustering algorithm. Let parameter c denote the number of clusters, to be optimized. The range of values for c is defined by an expert, so that the clustering schemes produced are compatible with expected attribute partitions. Then, a clustering algorithm is performed for each value c and the results of clustering are evaluated using goodness index G. We use the discovered most compact and well-separated clusters to classify each quantitative attribute into fuzzy sets. After that, we can generate the corresponding membership function for each fuzzy set.

The steps for finding fuzzy sets can be summarized as: (1) Finding the best clustering scheme by using a goodness index for each quantitative attribute, (2) constructing fuzzy sets with the c cluster centers, and (3) deriving the corresponding membership functions.

Main algorithm (C_{alg}, X, c_{min}, c_{max}, p)
    (*First phase: finding the optimal number of clusters and its centers*)
    Initialize: $c \leftarrow c_{max}$
    repeat
        Run the clustering algorithm C_{alg} for data set X to produce c cluster centers R
        Compute the goodness index $G(X, R)$
        if $(c = c_{max})$ **then**
            $\alpha \leftarrow Dis(c_{max})$
            $G_{opt} \leftarrow G(c)$
            $c_{opt} \leftarrow c$
        endif
        else if $(G(c) < G_{opt})$ **then**
            $c_{opt} \leftarrow c$
            $G_{opt} \leftarrow G(c)$
        endif
        $c \leftarrow c - 1$
    until $c = c_{min} - 1$

    (*Second phase: constructing fuzzy sets with the c cluster centers*)
    for $i := 1$ **to** c_{opt} **do**
        if $i < c_{opt}$ **then** determine d_i^+ by using overlap percentage p
        if $i \geq 2$ **then** determine d_i^- by using overlap percentage p
    endfor

    (*Third phase: generating membership function for each fuzzy set*)
    for each $x \in X$ **do**
        for each $r_i \in R$ **do**
            compute the corresponding membership function $f(r_i, x)$
        endfor
    endfor
End algorithm

Parameters:

C_{alg} = the clustering algorithm
$X = \{x_1, x_2, \ldots, x_n\}$ the set of attribute values to be clustered
c_{min} = the minimum number of clusters
c_{max} = the maximum number of clusters
p = overlap parameter in %

5 Experimental Results

We assessed the effectiveness of our approach by experimenting with a real-life dataset. The data set comes from a research by the U.S. Census Bureau (<http://www.census.gov/main/www/access.html>). The data had 6 quantitative attributes for 63756 families: age of family head in years ("head" is the reference person in a family), number of persons, children in family, education level of head, head's personal income and family income.

Table 2: The best number of clusters with cluster centers

Attribute	**No. of clusters**	**Cluster centers**
Age	3	31.65, 50.58, 73.99
FamPers	2	1.73, 4.48
NumKids	3	0.16, 2.24, 4.25
EdHead	3	34.81, 39.39, 43.35
IncHead	4	13436.1, 40107.8, 84932.9, 171354
IncFam	4	16056.3, 47395.5, 889382.8, 161793.7

First, we evaluate the proposed approach for finding the optimal clustering scheme using the above data set. The clustering schemes are discovered using the C-means algorithm while its input parameters (number of clusters) take values between 2 and 5 for the attributes $FamPers$, $NumKids$, and between 2 and 9 for the others (see Fig. 3 for attributes $IncHead$ and $IncFam$). Applying the first phase of our algorithm (see in Section 4.), Table 2 shows the best number of clusters for different attributes.

After finding the best cluster scheme, we can create the fuzzy sets for each quantitative attribute by using the discovered cluster centers.

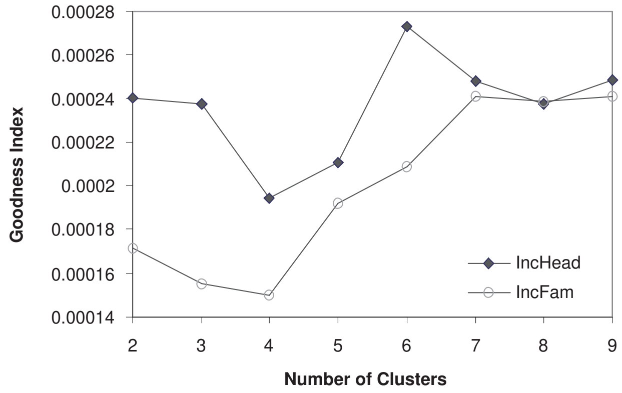

Figure 3: Goodness index as a function of the number of clusters

For example, the ranges of fuzzy sets of *Age* and *IncHead* are shown in the Table 3. These ranges include all values where the membership function is positive.

Table 3: Example ranges of fuzzy sets

Fuzzy set	**Range**
(Age,young)	15 to 43.95
(Age,middle)	38.78 to 65.80
(Age,old)	58.77 to 90
(IncHead,low)	285 to 30772.7
(IncHead,medium)	22771.2 to 69244.1
(IncHead,high)	55796.6 to 141106.6
(IncHead,very-high)	115180.3 to 304998

In the following, we illustrate how the above concept (clustering-based partitioning) gives a larger number of frequent itemsets and interesting association rules than the case when we don't use the proposed approach. In the latter case we use the same number of attribute elements for each interval (quantile-based partitioning). Note that the same definitions of membership functions are used for both methods as described in Section 3.

In deriving the association rules, we apply the algorithm described in [7], developed for fuzzy attributes. It is an extension of the well-known technique based on incrementally finding the frequent sets [3].

Fig. 4(a) and Fig. 4(b) show the average support and the number of frequent itemsets for different minimum support thresholds. As expected, the average support increases and the number of frequent itemset decreses as the minimum support increases from 10% to 50%. We can see that the clustering-based partitioning gives a higher number of frequent itemsets. However, the quantile-based partitioning gives higher average support values if the minimum support is between 0.35 and 0.5. Note, however, that it is generated by only two frequent itemsets.

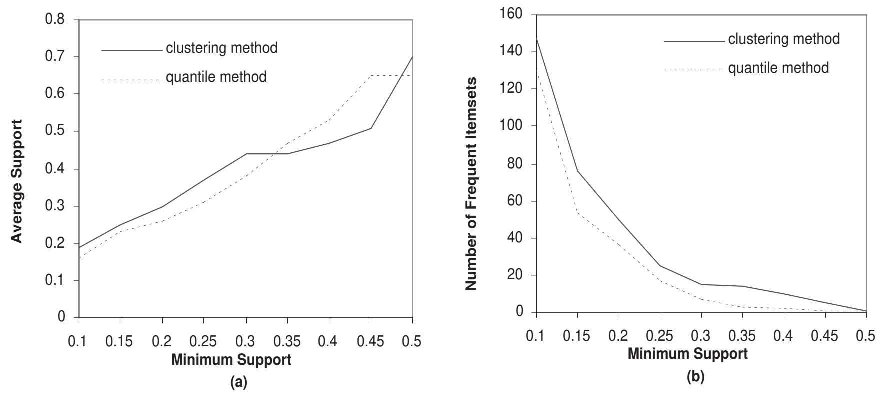

Figure 4: Left: Average Support. Right: Number of Frequent Itemsets.

Fig. 5(a) shows the average confidence for different minimum confidence thresholds. The result is quite similar to that Fig. 4(a). We can see that the clustering-based partitioning gives in most cases higher average confidence values.

Fig. 5(b) shows the number of generated rules as a function of minimum confidence threshold, for both clustering and quantile case. The minimum support was set to 30%. The results are as expected: the numbers of rules for the clustering-based partitioning case are larger, but both decrease with increasing confidence threshold.

Finally, we show some interesting rules. The minimum support was set to 30% and the minimum confidence to 50%.

IF *EdHead* is *medium* THEN *IncHead* is *low*
IF *IncHead* is *medium* THEN *IncFam* is *medium*
IF *FamPers* is *low* AND *NumKids* is *low* THEN *EdHead* is *low*
IF *FamPers* is *low* AND *NumKids* is *low* AND *IncHead* is *low* THEN *Inc-Fam* is *low*

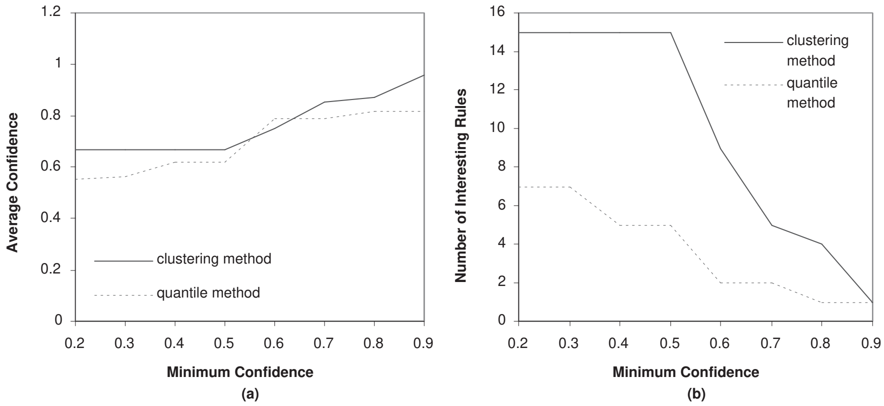

Figure 5: Left: Average Confidence. Right: Number of Interesting Rules.

We see that the rules are very easy to read and understand for anyone. This is our main goal in using fuzzy partitions for attributes. Of course, the usefulness of the rules can only be judged by a human.

6 Conclusion

The problem of mining association rules for fuzzy quantitative items was introduced in [7]. However, the algorithm assumes that the fuzzy sets are given. In this paper we have proposed a method to find the fuzzy sets for each quantitative attribute in a database by using clustering techniques. We defined the goodness index G for clustering scheme evaluation, based on two criteria: compactness and separation. The goodness index is a variant of the indices defined for the fuzzy c-means algorithm in [13], adapted to crisp clustering algorithms. Our approach is independent of the clustering algorithm used to partition the data set.

After having obtained the best cluster scheme, we exploited the discovered cluster centers, to classify the quantitative attribute values into fuzzy sets, and showed a method to find the corresponding membership function for each fuzzy set discovered. Then we combined the different steps into an explicit algorithm.

The experimental results demonstrated that by using the goodness index G as a basis for generating clusters (and thereby fuzzy sets), a higher number of fuzzy association rules can be discovered. According to our observations, we claim that the generated rules are very meaningful for real-life data sets.

Acknowledgment

I wish to thank Jukka Teuhola for useful discussion and his suggestions.

References

[1] Piatetsky-Shapiro, G., Frawley, W.J.: Knowledge Discovery in Databases. AAAI Press/The MIT Press, Menlo Park, California (1991)

[2] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proceedings of ACM SIGMOD (1993) 207–216

[3] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. Proceedings of the 20th VLDB Conference (1994) 487–499

[4] Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient Algorithms for discovering association rules. KDD-94: AAAI Workshop on KDD (1994) 181–192.

[5] Toivonen, H.: Sampling large databases for association rules. Proceedings of the 20th VLDB Conference (1996)

[6] Srikant, R., Agrawal, R.: Mining quantitative association rules in large relation tables. Proceedings of ACM SIGMOD (1996) 1–12.

[7] Gyenesei, A.: Mining Weighted Association Rules for Fuzzy Quantitative Items. Proceedings of PKDD Conference (2000) 416–423.

[8] Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. J. Cybern. Vol.4, (1974) 95–104.

[9] Dave, R.N.: Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letters, Vol.17, (1996) 613–623.

[10] Huang, Z.: A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining. DMKD (1997)

[11] Xie, X.L., Beni, G.: A Validity measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and machine Intelligence, Vol.13, No.4, August (1991)

[12] Gath, Geva, B.: Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on Pattern Analysis and machine Intelligence, Vol.11, No.7, July (1989)

[13] Reazee, R., Lelieveldt, B.P.F., Reiber, J.H.C.: A new cluster validity index for the fuzzy c-mean. Pattern Recognition Letters, 19, (1998) 237–246.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**