Determining Fuzzy Sets for Quantitative Attributes in Data Mining Problems

ATTILA GYENESEI
Turku Centre for Computer Science (TUCS)
University of Turku, Department of Computer Science
Lemminkäisenkatu 14A, FIN-20520 Turku
FINLAND

Abstract: The problem of mining association rules for fuzzy quantitative items was introduced and an algorithm proposed in [5]. 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.

Key-Words: association rules, fuzzy items, quantitative attributes, clustering

1 Introduction
Since knowledge can often be expressed in a more natural way by using fuzzy sets, many decision support problems can be greatly simplified. We attempt to take advantage of fuzzy sets in knowledge discovery from databases. One important topic in knowledge discovery and decision support research is concerned with the discovery of interesting association rules [1]. An interesting association rule describes an interesting relationship among different attributes. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form \( X \Rightarrow Y \), where \( X \) and \( Y \) are sets of items. An example of an association rule is: “40% of transactions that contain beer and potato chips also contain diapers; 5% of all transactions contain all of these items”. Here 40% is called the confidence of the rule, and 5% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

The problem of mining boolean association rules over supermarket data was introduced in [2], and later broadened in [3], for the case of databases consisting of categorical attributes alone. 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 [4]. The algorithm involves discretizing the domains of quantitative attributes into intervals in order to 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 nontrivial knowledge from those rules discovered.

In [5], 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 [5] 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 clustering proposed in [6], attempts to identify “compact and separated clusters”. Other validity indices for crisp clustering have been proposed in [7] and [8]. 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 [9]. Other validity measures are proposed in [10], [11]. 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 [11]. 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 $\sigma^2(X)$, is defined as

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

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

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

Definition 2 The variance of cluster $i$ containing elements $X_i = \{x_{i1}, \ldots, x_{in} \}$ is given by

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

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

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

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

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

where $D_{\text{max}}$ is the maximum, and $D_{\text{min}}$ 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. $\text{Dis}(R)$ 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 cluster $R$ within set $X$ is as follows:

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

where $\alpha$ is a weighting factor equal to $\text{Dis}(c_{\text{max}})$, $c_{\text{max}}$ 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 [11].)

![Figure 1: Example of Goodness Index for the Attribute Age](image1)

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 Section 5. We can see that the best number of clusters is three for this dataset.

![Figure 2: Example of the proposed fuzzy partitions](image2)

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).

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%)](image3)

<table>
<thead>
<tr>
<th>Fuzzy set</th>
<th>Range</th>
<th>Cluster center</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Age,young)</td>
<td>15 to 43.95</td>
<td>31.65</td>
</tr>
<tr>
<td>(Age,middle)</td>
<td>38.28 to 65.80</td>
<td>50.58</td>
</tr>
<tr>
<td>(Age,high)</td>
<td>58.77 to 90</td>
<td>73.99</td>
</tr>
</tbody>
</table>

In the following, we describe how to generate the corresponding membership function for each fuzzy set of a quantitative attribute. Let $\{r_1, r_2, \ldots, r_c\}$ 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 $r_i$, the membership function for element $x$ is given by
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 \((C_{\text{alg}}, X, c_{\text{min}}, c_{\text{max}}, p)\)

(*First phase: finding the optimal number of clusters and cluster centers*)

Initialize: \( c \leftarrow c_{\text{max}} \)

repeat

Run the clustering algorithm \( C_{\text{alg}} \) for data set \( X \) to produce \( c \) cluster centers \( R \)

Compute the goodness index \( G(X, R) \)

if \( (c = c_{\text{max}}) \) then

\[ \alpha \leftarrow \text{Dis}(c_{\text{max}}) \]

\[ G_{\text{opt}} \leftarrow G(c) \]

\[ c_{\text{opt}} \leftarrow c \]

endif

else if \( G(c) < G_{\text{opt}} \) then

\[ c_{\text{opt}} \leftarrow c \]

\[ G_{\text{opt}} \leftarrow G(c) \]

endif

\( c \leftarrow c-1 \)

until \( c = c_{\text{min}} \cdot 1 \)

(*Second phase: constructing fuzzy sets with the \( c \) cluster centers*)

for \( i=1 \) to \( c_{\text{opt}} \) do

if \( i < c_{\text{opt}} \) then determine \( d^+_i \) by using \( p \)

if \( i \geq 2 \) then determine \( d^-_i \) by using \( p \)

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:

\( C_{\text{alg}} = \) the clustering algorithm

\( X = \{x_1, x_2, \ldots, x_n\} \) the set of attribute values to be clustered

\( c_{\text{min}} = \) the minimum number of clusters

\( c_{\text{max}} = \) 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. 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.
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 $\text{FamPers}$, $\text{NumKids}$, and between 2 and 9 for the others (see Fig.3 for attributes $\text{IncHead}$ and $\text{IncFam}$). 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. For example, the ranges of fuzzy set of $\text{Age}$ is shown in the Table 1. These ranges include all values where the membership function is positive.

<table>
<thead>
<tr>
<th>Attr.</th>
<th>No. of clust.</th>
<th>Cluster centers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>3</td>
<td>31.65, 50.58, 73.99</td>
</tr>
<tr>
<td>$\text{FamPers}$</td>
<td>2</td>
<td>1.73, 4.48</td>
</tr>
<tr>
<td>$\text{NumKids}$</td>
<td>3</td>
<td>0.16, 2.24, 4.25</td>
</tr>
<tr>
<td>$\text{EdHead}$</td>
<td>3</td>
<td>34.81, 39.39, 43.35</td>
</tr>
<tr>
<td>$\text{IncHead}$</td>
<td>4</td>
<td>13436, 40108, 84933, 171354</td>
</tr>
<tr>
<td>$\text{IncFam}$</td>
<td>4</td>
<td>16056, 47396, 88938, 161794</td>
</tr>
</tbody>
</table>

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 [5], 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 decreases 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.

In deriving the association rules, we apply the algorithm described in [5], 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 decreases 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.
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 IncFam is low

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 [5]. 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 [11], 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.

References: