An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases

Wai-Ho Au Keith C.C. Chan
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

Abstract

In this paper, we present a novel technique, called F-APACS, for discovering fuzzy association rules in relational databases. Instead of dividing up quantitative attributes into fixed intervals and searching for rules expressed in terms of them, F-APACS employs linguistic terms to represent the revealed regularities and exceptions. The definitions of these linguistic terms are based on fuzzy set theory and the association rules expressed in them are, therefore, called fuzzy association rules here. To discover these rules, F-APACS utilizes an objective interestingness measure when determining if two attribute values are related. This measure is defined in terms of an "adjusted difference" between observed and expected frequency counts. The use of such a measure has the advantage that no user-supplied thresholds are required. In addition to this interestingness measure, F-APACS has another unique feature that it provides a mechanism to allow quantitative values be inferred from the rules. Such feature, as shown here, make F-APACS very effective at various mining tasks.

Keywords: data mining, fuzzy association rules, linguistic terms, interestingness measure.

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 [6]. One important topic in data mining research is concerned with the discovery of *association rules* [1]. An association rule describes an interesting association relationship among different attributes. A *boolean* association involves binary attributes; a *generalized* association involves attributes that are hierarchically related and a *quantitative* association involves attributes that can take on quantitative or categorical values. Existing algorithms (e.g. [2, 12]) for mining quantitative association rules require discretizing the domains of quantitative attributes into intervals so as to discover quantitative association rules. These intervals may not be concise and meaningful enough for human users to easily obtain nontrivial knowledge from those rules discovered. Instead of using intervals, we introduce a novel technique, called F-APACS, which employs *linguistic terms* to represent the revealed regularities and exceptions. The linguistic representation makes those rules discovered to be much natural for human users to understand. The definition of linguistic terms is based on fuzzy set theory and hence we call the rules having these terms *fuzzy association rules*. In fact, the use of fuzzy techniques has been considered one of the key components of data mining systems because of the affinity with the human knowledge representation [7].

Regardless of whether the association being considered is boolean, generalized or quantitative, existing algorithms (e.g. [1, 11-12]) decide if it is interesting by having a user supply two thresholds -- support and confidence. Given two attributes X and Y, the support is defined as the percentage of records having both attributes X and Y and the confidence is defined as the percentage of records having Y given that they also have X. If both support and confidence is greater than the user-supplied threshold, the association is considered interesting. A weakness of these approaches lies in the difficulty in deciding what these thresholds should be.

To overcome this problem, F-APACS utilizes *adjusted difference* [2-5] analysis to identify interesting associations among attributes. The use of this technique has the advantage that it does not require any user-supplied thresholds which are often hard to determine. Furthermore, F-APACS also has the advantage that it allows us to discover both *positive* and *negative* association rules. A positive association rule tells us that a record having certain characteristic will also have

0-7803-4863-X/98 $10.00©1998 IEEE

another characteristic whereas a negative association rule tells us that a record having certain characteristic will not have another characteristic.

Many data mining algorithms (e.g. [1-2, 4-5, 10-12]) require the class labels (conclusions of rules) to be crisp and the variables representing the class labels are therefore qualitative. This makes quantitative values are not inferred from those rules. To be more effective, F-APACS is able to deal with class boundaries that are fuzzy and to associate qualitative attribute values with quantitative attribute values. This provides a mechanism for F-APACS to allow quantitative values be inferred.

In the next section, we provide a brief description of how existing algorithms can be used for the mining of quantitative association rules and how fuzzy techniques can be applied to data mining process. The definition of linguistic terms is provided in Section 3. The details of F-APACS is given in Section 4. In Section 5, we present how to allow quantitative values be inferred from fuzzy association rules. The experimental results of F-APACS over several real-life databases are discussed in Section 6. Finally, in Section 7, we provide a summary of the paper.

2 Related Work

Quantitative association rules are defined over quantitative and categorical attributes [12]. The statement 70% of tertiary educated people between age 25 and 30 are unmarried" is one such example. The values of categorical attributes are mapped to a set of consecutive integers and the values of quantitative attributes are first discretized into intervals using *equi-depth partitioning*, if necessary, and then mapped to consecutive integers to preserve the order of the values/intervals [12]. Both categorical and quantitative attributes can then be handled in a uniform fashion as a set of <attribute, integer value>.

With the mappings defined in [12], a quantitative association rule is mapped to a set of boolean association rules. In other words, therefore, rather than having just one field for each attribute, there is a need to use as many fields as the number of different attribute values. For example, the value of a boolean field corresponding to <*attribute*$_1$, *value*$_1$> would be "1" if *attribute*$_1$ has *value*$_1$ in the original record and "0", otherwise [12]. After the mappings, the algorithms for mining boolean association rules (e.g. [1]) is then applied to the transformed data set.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of binary attributes called items and T be a set of transactions. Each transaction $t \in T$ is represented as a binary vector with $t[k] = 1$ if t contains item i_k and $t[k] = 0$, otherwise, for $k = 1, 2, \ldots, m$. An association rule is defined as an implication of the form $X \Rightarrow Y$ where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in T with support defined as $\Pr(X \cup Y)$ and confidence defined as $\Pr(Y|X)$. For the mining algorithms such as that described in [1, 11-12] to determine if an association is interesting, its support and confidence have to be greater than some user-supplied thresholds. A weakness of such approach is that many users do not have any idea what the thresholds should be. If it is set too high, a user may miss some useful rules but if it is set too low, the user may be overwhelmed by many irrelevant ones.

Furthermore, the intervals involved in quantitative association rules may not be concise and meaningful enough for human users to obtain nontrivial knowledge. *Fuzzy linguistic summaries* introduced in [13-14] express knowledge in linguistic representation which is natural for people to comprehend. An example of linguistic summaries is the statement "about half of people in the database are middle aged." In contrast to association rules which involve implications between different attributes, the fuzzy linguistic summaries only provide summarization on different attributes. There is no idea of implication in fuzzy linguistic summaries. As a result, this technique which provides a means for data analysis is not developed for the task of rule discovery.

In addition to fuzzy linguistic summaries, the applicability of fuzzy modeling techniques to data mining has been discussed in [8]. Given a series of fuzzy sets, $\mathcal{A}_1$, $\mathcal{A}_2, \ldots, \mathcal{A}_c$, *context-sensitive Fuzzy C-Means* (FCM) method is used to construct the rule-based models with the rules y is $\mathcal{A}_i$ if Ω_1 and Ω_2 and ..and Ω_c where $\Omega_1, \Omega_2, \ldots, \Omega_c$ are the regions in the input space that are centered around the c" prototypes for $i = 1, 2, \ldots, c$ [8]. Nevertheless, the context-sensitive FCM method can only manipulate quantitative attributes and it is for this reason that this technique is inadequate to deal with most real-life databases which consist of both quantitative and categorical attributes.

3 Linguistic Terms

Given a set of records, $\mathcal{D}$, each of which consists of a set of attributes $\mathcal{J} = \{I_1, I_2, \ldots, I_n\}$, where I_v, $v = 1, 2, \ldots, n$ can be quantitative or categorical. For any record, $d \in \mathcal{D}$, $d[I_v]$ denotes the value i_v in d for attribute I_v. For any quantitative attribute, $I_v \in \mathcal{J}$, let $dom(I_v) = [l_v, u_v] \subseteq \Re$ denote the domain of the attribute. A set of linguistic terms can be defined over the domain of each quantitative attribute. Let $\mathcal{L}_{vr}$, $r = 1, 2, \ldots, s_v$ be linguistic terms associated with some quantitative attribute, $I_v \in \mathcal{J}$. $\mathcal{L}_{vr}$ is represented by a fuzzy set, L_{vr}, defined on $dom(I_v)$ whose membership function is $\mu_{L_{vr}}$ such that

$$\mu_{L_{vr}} : dom(I_v) \rightarrow [0,1]$$

The fuzzy sets L_{vr}, $r = 1, 2, \ldots, s_v$ are then defined as

$$L_{vr} = \begin{cases} \sum_{dom(I_v)} \frac{\mu_{L_{vr}}(i_v)}{i_v} & \text{if } I_v \text{ is discrete} \\ \int_{dom(I_v)} \frac{\mu_{L_{vr}}(i_v)}{i_v} & \text{if } I_v \text{ is continuous} \end{cases}$$

for all $i_v \in dom(I_v)$. Some value $i_v \in dom(I_v)$ is compatible with some linguistic term $\mathcal{L}_{vr}$ to a degree of $\mu_{L_{vr}}(i_v)$.

For any categorical attribute, $I_v \in \mathcal{J}$, let $dom(I_v) = \{i_{v1}, i_{v2}, \ldots, i_{vm_v}\}$ denote the domain of I_v. To handle categorical and quantitative attributes in a uniform fashion, we can also define a set of linguistic terms, $\mathcal{L}_{vr}$, $r = 1, 2, \ldots, m_v$, for each categorical attribute, $I_v \in \mathcal{J}$, where $\mathcal{L}_{vr}$ is represented by a fuzzy set, L_{vr}, such that

$$L_{vr} = \frac{1}{i_{vr}}$$

Using the above technique, we can represent the original attributes, $\mathcal{J}$, using a set of linguistic terms, $\mathcal{L} = \{\mathcal{L}_{vr} | v = 1,2,\ldots,n, r = 1,2,\ldots,s_v\}$ where $s_v = m_v$ for categorical attributes. Each linguistic term is represented by a fuzzy set and hence we have a set of fuzzy sets, $L = \{L_{vr} | v = 1,2,\ldots,n, r = 1,2,\ldots,s_v\}$. Given a record, $d \in \mathcal{D}$, and a linguistic term, $\mathcal{L}_{vr} \in \mathcal{L}$, which is, in turn, represented by a fuzzy set, $L_{vr} \in L$, the degree of membership of the values in d with respect to L_{vr} is given by $\mu_{L_{vr}}(d[I_v])$. In order words, d is characterized by the term $\mathcal{L}_{vr}$ to a degree of $\mu_{L_{vr}}(d[I_v])$. If $\mu_{L_{vr}}(d[I_v]) = 1$, d is completely characterized by the term $\mathcal{L}_{vr}$. If $\mu_{L_{vr}}(d[I_v]) = 0$, d is undoubtedly not characterized by the term $\mathcal{L}_{vr}$. If $0 < \mu_{L_{vr}}(d[I_v]) < 1$, d is partially characterized by the term $\mathcal{L}_{vr}$. In case that $d[I_v]$ is unknown, $\mu_{L_{vr}}(d[I_v]) = 0.5$ which indicates that there is no information available concerning whether d is characterized by the term $\mathcal{L}_{vr}$ or not.

In fact, d can also be characterized by more than one linguistic terms. Let us consider the linguistic terms, $\mathcal{L}_{v_1 r_1}, \mathcal{L}_{v_2 r_2}, \ldots, \mathcal{L}_{v_m r_m} \in \mathcal{L}$, and the corresponding fuzzy sets, $L_{v_1 r_1}, L_{v_2 r_2}, \ldots, L_{v_m r_m} \in L$. d is characterized by the terms $\mathcal{L}_{v_1 r_1}, \mathcal{L}_{v_2 r_2}, \ldots, \mathcal{L}_{v_m r_m}$ to a degree of

$$\min(\mu_{L_{v_1 r_1}}(d[I_{v_1}]), \mu_{L_{v_2 r_2}}(d[I_{v_2}]), \ldots, \mu_{L_{v_m r_m}}(d[I_{v_m}]))$$

Based on linguistic terms, we can apply F-APACS to discover fuzzy association rules which are presented in a manner that is natural for human experts to understand. The use of fuzzy techniques also buries the boundaries of adjacent intervals of numeric qualities. This makes F-APACS resilient to noises such as inaccuracies in physical measurements of real-life entities. Furthermore, the fact that 0.5 is the fuzziest degree of membership of an element in a fuzzy set provides a new means for F-APACS to deal with missing values in databases.

4 Mining Fuzzy Association Rules

F-APACS begins the data mining process by calculating the sum of degrees to which those records in databases are characterized by the linguistic terms. Given a record, $d \in \mathcal{D}$, and linguistic terms, $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \in \mathcal{L}$, $p \neq j$ which are, in turn, represented by fuzzy sets, L_{pq}, $L_{jk} \in L$, $p \neq j$ respectively, the degree to which d is characterized by $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$ is accumulated in $deg_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$. F-APACS then determines if an interesting association relationship exists between $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \in \mathcal{L}$, $p \neq j$. More specifically, F-APACS can be described as follows (Fig. 1).

```
rules F-APACS( )
begin
    forall $d \in \mathcal{D}$ do
        forall $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \in \mathcal{L}$, $p \neq j$ do
            $deg_{\mathcal{L}_{pq}\mathcal{L}_{jk}} += \min(\mu_{L_{pq}}(d[I_p]), \mu_{L_{jk}}(d[I_j]))$ ;
    forall $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \in \mathcal{L}$, $p \neq j$ do
        if interesting($\mathcal{L}_{pq}$, $\mathcal{L}_{jk}$) then
            $\mathcal{R} = \mathcal{R} \bigcup rulegen(\mathcal{L}_{pq}, \mathcal{L}_{jk})$ ;
    return($\mathcal{R}$);
end
```

Fig. 1. Algorithm F-APACS.

The identification of interesting associations is based on an objective interestingness measure, called *adjusted difference* [2-5]. This measure is defined as

$$d_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}{\sqrt{\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}} \tag{1}$$

where $z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ is the *standardized difference* given by

$$z_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{deg_{\mathcal{L}_{pq}\mathcal{L}_{jk}} - e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}{\sqrt{e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}} \tag{2}$$

$e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ is the sum of degrees to which records are expected to be characterized by $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$ and is calculated by

$$e_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pq}\mathcal{L}_{ji}} \sum_{i=1}^{s_p} deg_{\mathcal{L}_{pi}\mathcal{L}_{jk}}}{\sum_{u=1}^{s_p} \sum_{i=1}^{s_j} deg_{\mathcal{L}_{pu}\mathcal{L}_{ji}}} \tag{3}$$

and $\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ is the *maximum likelihood estimate* of the variance of $z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ and is given by

$$\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \left(1 - \frac{\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pq}\mathcal{L}_{ji}}}{\sum_{u=1}^{s_p}\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pu}\mathcal{L}_{ji}}}\right)\left(1 - \frac{\sum_{i=1}^{s_p} deg_{\mathcal{L}_{pi}\mathcal{L}_{jk}}}{\sum_{u=1}^{s_p}\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pu}\mathcal{L}_{ji}}}\right) \tag{4}$$

If $|d_{\mathcal{L}_{pq}\mathcal{L}_{jk}}| > 1.96$ (the 95 percentiles of the normal distribution), we can conclude that the association between $\mathcal{L}_{jk}$ and $\mathcal{L}_{pq}$ is interesting. If $d_{\mathcal{L}_{pq}\mathcal{L}_{jk}} > +1.96$, the presence of $\mathcal{L}_{jk}$ implies the presence of $\mathcal{L}_{pq}$. It is more *likely* for a record having both $\mathcal{L}_{jk}$ and $\mathcal{L}_{pq}$. We say that $\mathcal{L}_{jk}$ is *positively* associated with $\mathcal{L}_{pq}$. If $d_{\mathcal{L}_{pq}\mathcal{L}_{jk}} < -1.96$, the absence of $\mathcal{L}_{jk}$ implies the presence of $\mathcal{L}_{pq}$. It is more *unlikely* for a record having $\mathcal{L}_{jk}$ and $\mathcal{L}_{pq}$ at the same time. We say that $\mathcal{L}_{jk}$ is *negatively* associated with $\mathcal{L}_{pq}$.

F-APACS employs the adjusted difference analysis in *interesting*($\mathcal{L}_{pq}$, $\mathcal{L}_{jk}$) to determine whether the association between $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$ is interesting. When such association is found to be interesting, the *interesting* function returns true; otherwise, it returns false. If *interesting*($\mathcal{L}_{pq}$, $\mathcal{L}_{jk}$) returns true, a fuzzy association rule is then generated by the *rulegen* function. The *rulegen* function takes as argument a pair of linguistic terms, $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$, to generate a fuzzy association rule such that it is in the form of $\mathcal{L}_{jk} \Rightarrow \mathcal{L}_{pq}\ [w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}]$ where $w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ denotes *weight of evidence* [3-5] measure which is defined as

$$w_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \log\frac{\Pr(\mathcal{L}_{jk} | \mathcal{L}_{pq})}{\Pr(\mathcal{L}_{jk} | \bigcup_{i \neq q} \mathcal{L}_{pi})} \tag{5}$$

$w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ provides a measure of the difference in the gain in information when a record with $\mathcal{L}_{jk}$ characterized by $\mathcal{L}_{pq}$ and when characterized by $\mathcal{L}_{pi}$, $i \neq q$. It is positive if $\mathcal{L}_{jk}$ is positively associated with $\mathcal{L}_{pq}$ whereas it is negative if $\mathcal{L}_{jk}$ is negatively associated with $\mathcal{L}_{pq}$.

5 Inferring Previously Unknown Values Using Fuzzy Association Rules

Using discovered fuzzy association rules, F-APACS is able to predict the values of some characteristics of previously unseen records. The results can be quantitative or categorical depending on the nature of the attributes whose values are to be predicted. Unlike other classification techniques which classify records into distinct classes, F-APACS allows quantitative values be inferred from fuzzy association rules.

Given a record, $t \in dom(I_1) \times \cdots \times dom(I_p) \times \cdots \times dom(I_n)$, let t be characterized by n attribute values, $\alpha_1, \ldots, \alpha_p, \ldots, \alpha_n$, where α_p is the value to be predicted. Let $\mathcal{L}_p$, $p = 1, 2, \ldots, s_p$, be the linguistic terms corresponding to the class attribute, I_p. We further let l_p be a linguistic term with domain $dom(l_p) = \{\mathcal{L}_{p1}, \mathcal{L}_{p2}, \ldots, \mathcal{L}_{ps_p}\}$. The value of α_p is assigned according to the value of l_p. To predict the correct value of l_p, F-APACS searches the association rules with $\mathcal{L}_{pq} \in dom(l_p)$ as consequents. If some attribute value, say α_j, $j \neq p$, of t, is characterized by the linguistic term in the antecedent of a rule which implies $\mathcal{L}_{pq}$, it can be considered as providing some evidence for or against the value of l_p being assigned to $\mathcal{L}_{pq}$. By repeating this procedure, that is, by matching each attribute value of t against the rules, F-APACS can determine the value of l_p by computing the total evidence measure.

Since each of the attributes of t may or may not provide evidence, and for those that do, they may support the assignment of different values, the different pieces of evidence are quantitatively measured and combined for comparison in order to find the most suitable value of l_p. For any attribute value α_j, $j \neq p$, of t, it is characterized by a linguistic term, $\mathcal{L}_{jk}$, to a degree of compatibility, $\mu_{L_{jk}}(\alpha_j)$, for each $k \in \{1, 2, \ldots, s_j\}$. Given those rules implying the assignment of $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \Rightarrow \mathcal{L}_{pq}\ [w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}]$, for all $k \in \zeta \subseteq \{1, 2, \ldots, s_j\}$, the evidence provided by α_j for or against such assignment is given by

$$w_{\mathcal{L}_{pq}\alpha_j} = \sum_{k \in \zeta} w_{\mathcal{L}_{pq}\mathcal{L}_{jk}} \cdot \mu_{L_{jk}}(\alpha_j) \tag{6}$$

Suppose that, of the $n - 1$ attribute values excluding α_p, only some of them, $\alpha_{[1]}, \ldots, \alpha_{[j]}, \ldots, \alpha_{[m]}$ with $\alpha_{[j]} = \{\alpha_i \mid i \in \{1, 2, \ldots, n\} - \{p\}\}$, are found to match one or more rules, then the overall weight of evidence for or against the value of l_p to be assigned to $\mathcal{L}_{pq}$ is given by

$$w_q = \sum_{j=1}^{m} w_{\mathcal{L}_{pq}\alpha_{[j]}} \tag{7}$$

In case that I_p is categorical, l_p is assigned to $\mathcal{L}_{pc}$ if

$$w_c > w_h,\ h = 1, 2, \ldots, s'_p \text{ and } h \neq c \tag{8}$$

where s'_p ($\leq s_p$) denotes the number of linguistic terms implied by the rules. α_p is therefore assigned to $i_{pc} \in dom(I_p)$.

If I_p is quantitative, a novel method is used to assign an appropriate value to α_p. Given the linguistic terms, $\mathcal{L}_{p1}, \mathcal{L}_{p2}, \ldots, \mathcal{L}_{ps_p}$, and their overall weights of evidence,

$w_1, w_2, \ldots, w_{s_p}$, let $\mu'_{L_{pu}}(i_p)$ be the weighted degree of membership of $i_p \in dom(I_p)$ to the fuzzy set L_{pu}, $u \in \{1, 2, \ldots s_p\}$. $\mu'_{L_{pu}}(i_p)$ is given by

$$\mu'_{L_{pu}}(i_p) = w_u \cdot \mu_{L_{pu}}(i_p) \tag{9}$$

where $i_p \in dom(I_p)$ and $u = 1, 2, \ldots, s_p$. The predicted value, α, is then defined as

$$\alpha = \frac{\int_{dom(I_p)} \mu'_{L_{p1} \cup L_{p2} \cup \cdots \cup L_{ps_p}}(i_p) \cdot i_p \, di_p}{\int_{dom(I_p)} \mu'_{L_{p1} \cup L_{p2} \cup \cdots \cup L_{ps_p}}(i_p) \, di_p} \tag{10}$$

where $\mu'_{X \cup Y}(i) = \max(\mu'_X(i), \mu'_Y(i))$ for any fuzzy sets X and Y. This prediction, α, provides an appropriate value for α_p.

For quantitative predictions, we use the *root-mean-squared error* as a performance measure. Given a set of testing records, D, let n be number of records in D. For any record, $r \in D$, let $[l, u] \subset \Re$ denote the domain of the class attribute. We further let t_r be the target value of the class attribute in r and o_r be the predicted value given by F-APACS. The root-mean-squared error, *rms*, is defined as

$$rms = \sqrt{\frac{1}{n} \sum_{r \in D} \left(\frac{t_r - l}{u - l} - \frac{o_r - l}{u - l} \right)^2} \tag{11}$$

6 Experimental Results

To evaluate the effectiveness of F-APACS, we used several real-life databases including (i) a transactional database of a PBX system; and (ii) a database concerning with 800 industrial enterprises in mainland China.

6.1 The *PBX* Database

The *PBX* database is collected from a private branch exchange (PBX) system used in a telecommunication company. It consists of phone call records about the usage of the PBX system. There are 3,009 records and each record is characterized by 13 attributes. Among these attributes, two are categorical and all the remaining attributes are quantitative. There are 98.4% of records having missing values in at least one of the attributes.

As an illustration, let us consider attributes *Time-of-call* and *Duration-of-call* in detail. We define the linguistic terms (i) *Mid-night*, *Morning*, *Afternoon*, *Evening*, and *Night* for *Time-of-call*; and (ii) *Very-short*, *Short*, *Moderate*, *Long*, and *Very-long* for *Duration-of-call*. These linguistic terms are shown in Fig. 2 and Fig. 3 respectively.

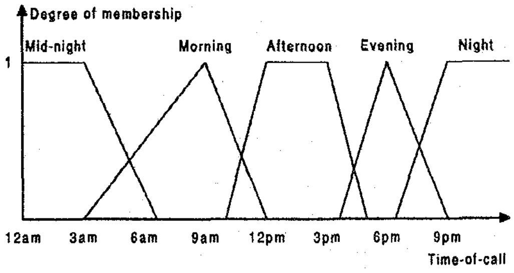

Fig. 2. Linguistic terms for *Time-of-call*.

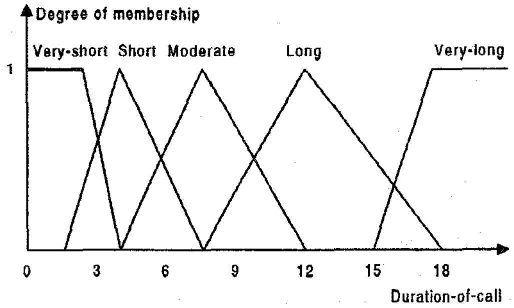

Fig. 3. Linguistic terms for *Duration-of-call*.

In [3], F-APACS has been shown to be able to discover meaningful fuzzy association rules in databases. In this experiment, *Calling-party-id* is chosen as the class attribute. Using the linguistic terms, we applied F-APACS to the *PBX* database. For comparison, we also applied C4.5 [10], a decision-tree based classification technique, to this database. Of all the records, 2,000 are randomly selected for training and the others are used for testing. Fig. 4 shows the classification accuracy.

	Classification Accuracy
F-APACS	99.5%
C4.5	94.6%

Fig. 4. Classification accuracy for the *PBX* database.

As shown in Fig. 4, F-APACS outperformed C4.5. This revealed that F-APACS is able to discover some important knowledge that cannot be found by C4.5.

6.2 The *china* Database

The *china* database is provided by the China Business Center of the Hong Kong Polytechnic University. It contains the data collected in a survey performed by the census bureau of the government in People' s Republic of China. This survey is concerned with the situations of production and operation of 800 industrial enterprises in mainland China in 1992. The *china* database contains 800 records representing these industrial enterprises. Each record is characterized by 56 quantitative attributes.

The percentage of missing values in each attribute ranges from 0.5% to 97.5%.

As an illustration, let us consider attribute *A1* which represents "Total value of industrial output (constant price of 1990)" in detail. We define the linguistic terms *Very-low*, *Low*, *Moderate*, *High*, and *Very-high* for *A1*. These linguistic terms are shown in Fig. 5.

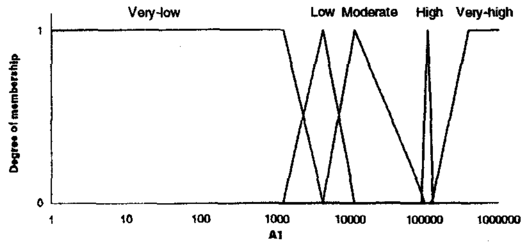

Fig. 5. Linguistic terms for *A1*.

Among all the attributes, domain experts identified attribute *A1* as the class attribute. It should be noted that there are 30 records whose values of attribute *A1* are missing or unknown. These records are ignored in our experiment. Of the remaining records, 580 are randomly selected for training and the others are used for testing.

It should also be noted that attribute *A1* is quantitative and the predictions on this attribute are therefore expected to be of quantitative values. Unfortunately, other classification techniques which can only classify records into distinct classes are unable to give quantitative values as predictions. It is for this reason that they are not readily applicable to this task. On the contrary, by representing attribute *A1* with the linguistic terms, F-APACS is able to inferring quantitative values using fuzzy association rules.

In our experiment, the quantitative attributes are represented by appropriate linguistic terms. The root-mean-squared error over the class attribute *A1* for the *china* database is equal to 4.4%. In other words, the predictions produced by F-APACS deviated from the target values by 4.4% in average.

7 Conclusions

We presented a novel algorithm, called F-APACS, which employs linguistic terms to represent the revealed regularities and exceptions in this paper. The linguistic representation is especially useful when those rules discovered are presented to human users for examination. Unlike other algorithms which discover association rules based on the use of some user-supplied thresholds, F-APACS employs adjusted difference analysis to identify interesting fuzzy associations among attributes. This makes it to be able to avoid the use of some user-supplied thresholds which are often difficult to determine. Using adjusted difference analysis, F-APACS is able to discover positive and negative association rules. Furthermore, unlike other classification techniques which classify records into distinct classes, F-APACS allows quantitative values be inferred from fuzzy association rules.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proc. of 1993 ACM SIGMOD Int'l Conf. on Management of Data*, Washington D.C., May 1993, pp. 207-216.

[2] K.C.C. Chan, and W.-H. Au, "An Effective Algorithm for Mining Interesting Quantitative Association Rules," in *Proc. of the 12th ACM Symp. on Applied Computing (SAC'97)*, San Jose, CA, Feb. 1997.

[3] K.C.C. Chan, and W.-H. Au, "Mining Fuzzy Association Rules," to appear in *Proc. of the 6th ACM Int'l Conf. on Information and Knowledge Management (CIKM'97)*, Las Vegas, Nevada, Nov. 1997.

[4] K.C.C. Chan, and A.K.C. Wong, "APACS: A System for the Automatic Analysis and Classification of Conceptual Patterns," *Computational Intelligence*, vol. 6, pp. 119-131, 1990.

[5] K.C.C. Chan, and A.K.C. Wong, "A Statistical Technique for Extracting Classificatory Knowledge from Databases," in [9], pp. 107-123.

[6] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus, Knowledge Discovery in Databases: An Overview," in [9], pp. 1-27.

[7] A. Maeda, H. Ashida, Y. Taniguchi, and Y. Takahashi, Data Mining System using Fuzzy Rule Induction," in *Proc. of 1995 IEEE Int'l Conf. on Fuzzy Systems*, Yokohama, Japan, Mar. 1995, pp. 45-46.

[8] W. Pedrycz, "Data Mining and Fuzzy Modeling," in *Proc. of 1996 Biennial Conf. of the North American Fuzzy Information Processing Society (NAFIS)*, Berkeley, CA, June 1996, pp. 263-267.

[9] G. Piatetsky-Shapiro, and W.J. Frawley (Eds.), *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.

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

[11] R. Srikant, and R. Agrawal, "Mining Generalized Association Rules," in *Proc. of the 21st VLDB Conf.*, Zurich, Switzerland, 1995, pp. 407-419.

[12] R. Srikant, and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables," in *Proc. of 1996 ACM SIGMOD Int'l Conf. on Management of Data*, Monreal, Canada, June 1996, pp. 1-12.

[13] R.R. Yager, "On Linguistic Summaries of Data," in [9], pp. 347-363.

[14] R.R. Yager, "Fuzzy Summaries in Database Mining," in *Proc. of the 11th Conf. on Artificial Intelligence for Application*, Los Angeles, CA, Feb. 1995, pp. 265-269.