# Mining Fuzzy Association Rules

*Keith C.C. Chan* *Wai-Ho Au*

Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
E-mail: {cskcchan, cswhau}@comp.polyu.edu.hk

## Abstract

*In this paper, we introduce a novel technique, called F-APACS, for mining fuzzy association rules. Existing algorithms involve 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 experts to easily obtain nontrivial knowledge from those rules discovered. Instead of using intervals, F-APACS employs linguistic terms to represent the revealed regularities and exceptions. The linguistic representation is especially useful when those rules discovered are presented to human experts for examination. The definition of linguistic terms is based on fuzzy set theory and hence we call the rules having these terms fuzzy association rules. The use of fuzzy techniques makes F-APACS resilient to noises such as inaccuracies in physical measurements of real-life entities and missing values in the databases. Furthermore, F-APACS employs adjusted difference analysis which has the advantage that it does not require any user-supplied thresholds which are often hard to determine. The fact that F-APACS is able to mine fuzzy association rules which utilize linguistic representation and that it uses an objective yet meaningful confidence measure to determine the interestingness of a rule makes it very effective at the discovery of rules from a real-life transactional database of a PBX system provided by a telecommunication corporation.*

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 [9]. The regularities or exceptions discovered from databases through data mining has enabled human decision makers to better make decisions in many different areas [8, 14].

One important topic in data mining research is concerned with the discovery of interesting *association rules* [1]. An interesting association rule describes an interesting relationship among different attributes and we refer to such relationship as an *association* in this paper. 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. [3, 16]) involve 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 experts 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 experts 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 as one of the key components of data mining systems because of the affinity with the human knowledge representation [11].

Regardless of whether the association being considered is boolean, generalized or quantitative, existing algorithms (e.g. [1-2, 10, 15-16]) 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* [3-5] analysis to identify interesting associations among attributes. Unlike other data mining algorithms (e.g. [1-2, 10, 15-16]), 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



CIKM 97 Las Vegas Nevada USA
Copyright 1997 ACM 0-89791-970-x/97/11..$3.50

attribute value (or linguistic term) will also have another attribute value (or linguistic term) whereas a negative association rule tells us that a record having certain attribute value (or linguistic term) will not have another attribute value (or linguistic term).

This paper is organized as follows. 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. We also discuss the relative strength and weaknesses of these techniques. The details of F-APACS is given in Section 3. In this same section, we also describe how F-APACS is able to overcome some of the limitations of existing algorithms. To evaluate the performance of F-APACS, we have applied it to a real-life database. The results of the experiment are discussed in Section 4. Finally, in Section 5, we provide a summary of the paper.

# 2 Related Work

Quantitative association rules are defined over quantitative and categorical attributes [16]. The statement "70% of tertiary educated people between age 25 and 30 are unmarried" is one such example. In [16], 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. And as a result, both categorical and quantitative attributes can be handled in a uniform fashion as a set of <attribute, integer value> pairs.

With the mappings defined in [16], 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 $\langle attribute_1, value_1 \rangle$ would be "1" if $attribute_1$ has $value_1$ in the original record and "0", otherwise [16]. After the mappings, the algorithms for mining boolean association rules (e.g. [1-2]) 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[1] 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 the percentage of records having both $X$ and $Y$ and confidence defined as the percentage of records having $Y$ given that they also have $X$. For the mining algorithms such as that described in [1-2, 10, 15-16] 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 experts to obtain nontrivial knowledge. *Fuzzy linguistic summaries* introduced in [17-18] 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. Although this technique can provide concise summaries which are nature for people to comprehend, 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 [13]. 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 $\Omega_1$ and $\Omega_2$ and ... and $\Omega_c$ where $\Omega_1, \Omega_2, \ldots, \Omega_c$ are the regions in the input space that are centered around the "c" prototypes for $i = 1, 2, \ldots, c$ [13]. 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 F-APACS for Mining Fuzzy Association Rules

In this section, we describe a novel algorithm, called F-APACS, which makes use of linguistic terms to represent the regularities and exceptions discovered from databases. Furthermore, F-APACS also employs *adjusted difference* [3-5] analysis to identify interesting associations among attributes. The definition of linguistic terms is presented in Section 3.1. An overview of F-APACS is then given in Section 3.2. After that, we describe how interesting associations can be identified in Section 3.3. A confidence measure, called *weight of evidence* [3-5] measure, is then defined in Section 3.4 to provide a means for representing the uncertainty associated with the fuzzy association rules. In Section 3.5, we give a discussion about the computational complexity of F-APACS.

## 3.1 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. Based on fuzzy set theory, 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

[1] The definition is adapted from [2] which is a slightly modified version of that presented in [1].

quantitative attribute, $I_v \in \mathcal{I}$. $\mathcal{L}_{vr}$ 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 $i_v \in dom(I_v)$. The degree of compatibility of some value $i_v \in dom(I_v)$ with some linguistic term $\mathcal{L}_{vr}$ is given by $\mu_{L_{vr}}(i_v)$.

As an example, let us consider the attribute *Height* and the linguistic term *Tall* represented by the fuzzy set $T$ such that

$$T = \int_{170}^{200} \frac{\frac{1}{30}(x-170)}{x} + \int_{200}^{\infty} \frac{1}{x}$$

for all $x \in dom(Height) \subseteq \Re$. If a person is 185 cm tall, then his height is compatible with the term *Tall* to a degree of $\frac{1}{30}(185-170) = 0.5$.

For any categorical attribute, $I_v \in \mathcal{I}$, let $dom(I_v) = \{i_{v1}, i_{v2}, \ldots, i_{vm_v}\}$ denote the domain of $I_v$. In order 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, $I_v \in \mathcal{I}$, where $\mathcal{L}_{vr}$ is represented by a fuzzy set, $L_{vr}$, such that

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

In addition to handling categorical and quantitative attributes in a uniform fashion, the use of linguistic terms to represent categorical attributes also allows the fuzzy nature of some real-world entities to be easily captured. For example, it may be difficult to distinguish the color orange for the color red in some situations. It is for this reason that an object which is orange in color can be perceived as red in color to certain extent. Such kind of fuzziness in attribute *Color* can be represented by the linguistic terms *Red* and *Orange*. Based on these linguistic terms, the color of an object can be compatible with the term *Red* to a degree of 0.7 and with the term *Orange* to a degree of 0.9.

Using the above technique, we can represent the original attributes, $\mathcal{I}$, 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, $\mathcal{L}_{v_1r_1}, \mathcal{L}_{v_2r_2}, \ldots, \mathcal{L}_{v_mr_m} \in \mathcal{L}$, and the corresponding fuzzy sets, $L_{v_1r_1}, L_{v_2r_2}, \ldots, L_{v_mr_m} \in L$. The degree to which $d$ is characterized by the terms $\mathcal{L}_{v_1r_1}, \mathcal{L}_{v_2r_2}, \ldots, \mathcal{L}_{v_mr_m}$ is defined as

$$\min(\mu_{L_{v_1r_1}}(d[I_{v_1}]), \mu_{L_{v_2r_2}}(d[I_{v_2}]), \ldots, \mu_{L_{v_mr_m}}(d[I_{v_m}]))$$

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. In addition to linguistic representation, the use of fuzzy techniques 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.

## 3.2 The F-APACS in Detail

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).

```
1) rules F-APACS( )
2) begin
3)     forall d ∈ 𝒟 do
4)         forall L_pq, L_jk ∈ ℒ, p ≠ j do
5)             deg_{L_pq L_jk} += min(μ_{L_pq}(d[I_p]), μ_{L_jk}(d[I_j]));
6)     forall L_pq, L_jk ∈ ℒ, p ≠ j do
7)         if interesting(L_pq, L_jk) then
8)             ℛ = ℛ ∪ rulegen(L_pq, L_jk);
9)     return(ℛ);
10) end
```

Fig. 1. Algorithm F-APACS.

The identification of interesting associations is based on an objective interestingness measure, called *adjusted difference* [3-5]. F-APACS employs this measure in *interesting*($\mathcal{L}_{pq}$, $\mathcal{L}_{jk}$) to determine whether the association between $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$ is interesting. If *interesting*($\mathcal{L}_{pq}$, $\mathcal{L}_{jk}$) returns true, a fuzzy association rule is then generated by the *rulegen* function. For each rule generated, this function also returns a confidence measure called the *weight of evidence* [3-5] measure. All fuzzy association rules generated by *rulegen* are stored in $\mathcal{R}$ which will then be used later for inference or for the users to examine. The *rulegen* function takes as argument a pair of linguistic terms, $\mathcal{L}_{pq}$ and $\mathcal{L}_{jk}$, where $\mathcal{L}_{pq}$, $\mathcal{L}_{jk} \in \mathcal{L}$, $p \neq j$ to generate fuzzy association rules is the form $\mathcal{L}_{jk} \Rightarrow \mathcal{L}_{pq}$ [$w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$] where $w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ denotes the weight of evidence.

## 3.3 Identification of Interesting Associations

The details of the *interesting* function described in the last section is given as follows (Fig. 2).

```
1) bool interesting(L_pq, L_jk)
2) begin
3)     calculate d_{L_pq L_jk} using (5);
4)     if |d_{L_pq L_jk}| > 1.96 then
5)         return true;
6)     else
7)         return false;
8) end
```

Fig. 2. The *interesting* function.

In order to decide whether the association between a linguistic term, $\mathcal{L}_{jk}$, and another linguistic term, $\mathcal{L}_{pq}$, is interesting, we determine whether

$$\Pr(\mathcal{L}_{pq}|\mathcal{L}_{jk}) = \frac{\text{sum of degrees to which objects characterized by } \mathcal{L}_{pq} \text{ and } \mathcal{L}_{jk}}{\text{sum of degrees to which objects characterized by } \mathcal{L}_{jk}} \quad (1)$$

is *significantly different* from

$$\Pr(\mathcal{L}_{pq}) = \frac{\text{sum of degrees to which objects characterized by } \mathcal{L}_{pq}}{M} \quad (2)$$

where $M = \sum_{u=1}^{s_p} \sum_{i=1}^{s_j} deg_{\mathcal{L}_{pu}\mathcal{L}_{ji}}$. If this is the case, we consider the association between $\mathcal{L}_{jk}$ and $\mathcal{L}_{pq}$ interesting.

The significance of the difference can be objectively evaluated based on the *adjusted difference* [3-5] which is defined as

$$d_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}{\sqrt{\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}} \quad (3)$$

where $z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ is the *standardized difference* [3-5] given by

$$z_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{deg_{\mathcal{L}_{pq}\mathcal{L}_{jk}} - e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}{\sqrt{e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}}} \quad (4)$$

$e_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ 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

$$e_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \frac{\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pq}\mathcal{L}_{ji}} \sum_{i=1}^{s_p} deg_{\mathcal{L}_{pi}\mathcal{L}_{jk}}}{M} \quad (5)$$

and $\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ is the *maximum likelihood estimate* [3-5] of the variance of $z_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ and is given by

$$\gamma_{\mathcal{L}_{pq}\mathcal{L}_{jk}} = \left(1 - \frac{\sum_{i=1}^{s_j} deg_{\mathcal{L}_{pq}\mathcal{L}_{ji}}}{M}\right)\left(1 - \frac{\sum_{i=1}^{s_p} deg_{\mathcal{L}_{pi}\mathcal{L}_{jk}}}{M}\right) \quad (6)$$

If $|d_{\mathcal{L}_{pq}\mathcal{L}_{jk}}| > 1.96$ (the 95 percentiles of the normal distribution), we can conclude that the discrepancy between $\Pr(\mathcal{L}_{pq} \mid \mathcal{L}_{jk})$ and $\Pr(\mathcal{L}_{pq})$ is significantly different and hence 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 $\mathcal{L}_{pq}$. In other words, 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 $\mathcal{L}_{pq}$. In other words, 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 $\mathcal{L}_{pq}$.

## 3.4 Confidence Measure of Association Rules

As described in Section 3.2, the *rulegen* function uses the *weight of evidence* [4-5] measure as a confidence measure for fuzzy association rules. If the association between $\mathcal{L}_{jk}$ and $\mathcal{L}_{pq}$ is found to be interesting, there is some evidence for or against a record having $\mathcal{L}_{pq}$ given it has $\mathcal{L}_{jk}$. The weight of evidence measure is defined in terms of an information theoretic concept known as *mutual information*. Mutual information measures the change of uncertainty about the presence of $\mathcal{L}_{pq}$ in a record given that it has $\mathcal{L}_{jk}$ is and in turn defined as

$$I(\mathcal{L}_{pq}:\mathcal{L}_{jk}) = \log \frac{\Pr(\mathcal{L}_{pq}|\mathcal{L}_{jk})}{\Pr(\mathcal{L}_{pq})} \quad (7)$$

Based on mutual information, the weight of evidence measure is defined in [4] as

$$\begin{aligned} w_{\mathcal{L}_{pq}\mathcal{L}_{jk}} &= I(\mathcal{L}_{pq}:\mathcal{L}_{jk}) - I(\bigcup_{i\neq q}(\mathcal{L}_{pi}:\mathcal{L}_{jk})) \\ &= \log \frac{\Pr(\mathcal{L}_{jk}|\mathcal{L}_{pq})}{\Pr(\mathcal{L}_{jk}|\bigcup_{i\neq q}\mathcal{L}_{pi})} \end{aligned} \quad (8)$$

$w_{\mathcal{L}_{pq}\mathcal{L}_{jk}}$ can be intuitively interpreted as 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 $\mathcal{L}_{pi}$, $i \neq q$. The weight of evidence is positive if $\mathcal{L}_{jk}$ is positively associated with $\mathcal{L}_{pq}$ whereas the weight of evidence is negative if $\mathcal{L}_{jk}$ is negatively associated with $\mathcal{L}_{pq}$.

### 4.5 On the Scalability of F-APACS

Given a database containing $N$ records such that each record is characterized by $n$ attributes and each attribute is represented by $m$ linguistic terms, there are $(n-1)m$ possible pairs of linguistic terms involving a certain term. The total number, $r$, of pairs of linguistic terms is therefore given by

$$\begin{aligned} r &= (nm)((n-1)m) \\ &= n(n-1)m^2 \end{aligned} \quad (9)$$

Each record in the database has $n$ attribute values (one for each attribute) and each of them is represented by $m$ linguistic terms. Moreover, there are $(n-1)m$ linguistic terms for it to have associations. F-APACS therefore calculate the cumulative degree for $(nm)((n-1)m)$ times. Consequently, F-APACS takes $(n(n-1)m^2)N$ operations to scan through the database. After obtaining the cumulative degree of all pairs of linguistic terms, F-APACS determines whether the association between each pair is interesting and it takes another $r$ operations. As a result, F-APACS takes $(n(n-1)m^2)N + r$ operations in total. Hence the computational complexity of F-APACS, $O$(F-APACS), is given by

$$\begin{aligned} O(\text{F-APACS}) &= O((n(n-1)m^2)N + r) \\ &= O((n(n-1))m^2N + n(n-1)m^2) \\ &= O(n^2m^2(N+1)) \\ &= O(n^2m^2N) \end{aligned} \quad (10)$$

Let us note that the F-APACS algorithm scales up linearly in terms of number of records in databases. This scalability of F-APACS is indeed desirable because there are usually tremendous amount, say, millions, of records in real-life databases. In addition to number of records in databases, the performance of F-APACS also depends on the total number of attributes and the number of linguistic terms used to represent an attribute. Specifically, F-APACS possesses quadratic scalability in terms of number of attributes and number of linguistic terms associated with each attribute respectively.

## 4 Experimental Results

In order to evaluate the effectiveness of F-APACS, we applied it to a real-life transactional database of a PBX system provided by a telecommunication corporation. This database contains the phone calling records of clients of the corporation using the PBX system. There are 3,009 records in the database. Each record is represented by a string of 132 characters long from which 13 attributes can be extracted. These 13 attributes are *Time-of-call-origination*, *Duration-of-call*, *Calling-party-identification*, *Originating-extension-number*, *Trunk-identification*, *Trunk-number*, *Trunk-access-code*, *Directory-number-dialed*, *Account-code*, *Tonant-number*, *Metering-group*, *Call-charge*, and *Modem-identification-number*. Except the two categorical attributes *Calling-party-identification* and *Trunk-identification*, all the remaining attributes are quantitative. The domain of *Calling-party-identification* is {*ST*, *AT*, *TI*, *DD*, *DS*, *DT*, *CO*, *LN*, *TL*} whereas the domain of *Trunk-identification* is {–, *C*, *F*, *L*, *W*, *T*}.

As an illustration, let us consider attributes *Time-of-call-origination* and *Duration-of-call* in detail. We define the linguistic terms *Mid-night*, *Morning*, *Afternoon*, *Evening*, and *Night* for *Time-of-call-origination*. These linguistic terms are shown in Fig. 3.

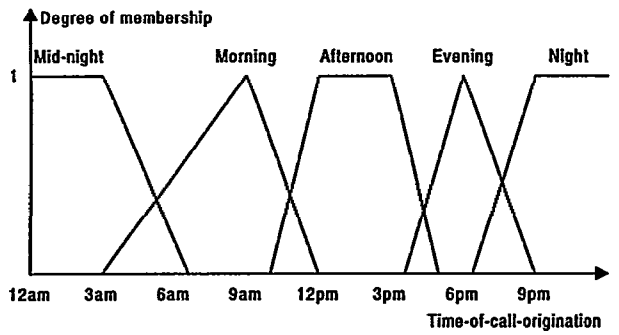


Fig. 3. Definition of linguistic terms for attribute *Time-of-call-origination*.

For *Duration-of-call*, we define the linguistic terms *Very-short*, *Short*, *Moderate*, *Long*, and *Very-long*. These terms are given in Fig. 4.

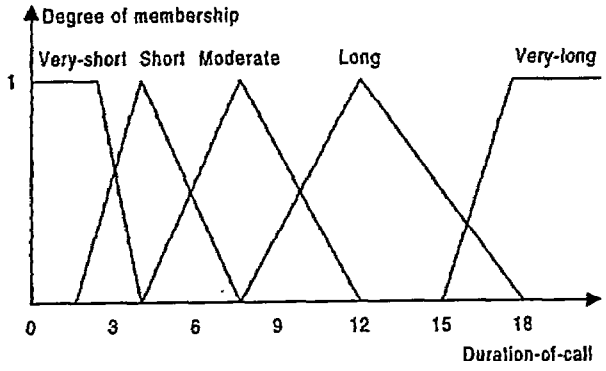


Fig. 4. Definition of linguistic terms for attribute *Duration-of-call.*

Using the linguistic terms described in Section 3.1, we applied F-APACS to the *PBX* database. The results of the experiment demonstrate that F-APACS is able to discover meaningful fuzzy association rules from the data. Among the fuzzy association rules discovered by F-APACS, the following rules are particularly interesting.

*Calling-party-identification* = "*DD*"
⇒ *Trunk-identification* = "–" [*infinity*]
*Calling-party-identification* = "*CO*"
⇒ *Trunk-identification* = "–" [*infinity*]
*Calling-party-identification* = "*ST*"
⇒ *Trunk-identification* = "*C*" [*infinity*]

All of these rules has a weight of evidence of infinity and this means that we are extremely certain that they are true all the time. In fact, it is reasonable that a trunk is dedicated to some specific calling parties. Hence which trunk is used depends on what the calling parties are and this is what the above rules imply.

Let us consider the following fuzzy association rules as well.

*Duration-of-call* = *Short*
⇒ *Calling-party-identification* = "*ST*" [2.80]
*Calling-party-identification* = "*ST*"
⇒ *Time-of-call-origination* = *Morning* [-2.36]

The first rule states that it is more likely that the calling party is "*ST*" if the call duration is short whereas the later rule states that a calling party which is "*ST*" tends not to make a phone call in the morning. The later one is an example of negative association rules. It should be noted that existing algorithms (e.g. [1-2, 10, 15-16]) are unable to discover negative associations of such type. The ability of F-APACS to discover negative association rules is another unique feature of it.

F-APACS can also discover association rules between linguistic terms and the following are examples of rules concerning the relationships between linguistic terms.

*Duration-of-call* = *Very-long*
⇒ *Time-of-call-origination* = *Afternoon* [1.94]
*Time-of-call-origination* = *Mid-night*
⇒ *Duration-of-call* = *Long* [-2.30]

The first one says that those phone calls with very long duration tend to be made at afternoon whereas the second one says that the clients who make phone calls at mid-night will not have talks of long duration. It is recognized that those fuzzy association rules involving linguistic terms are more natural for human experts to understand when compared to quantitative association rules which involves intervals.

# 5 Conclusions

Quantitative association rules discovery algorithms involve discretizing the domains of quantitative attributes into intervals. These intervals may not be concise and meaningful enough for human experts to easily obtain nontrivial knowledge from those rules discovered. Instead of using intervals, 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 experts for examination. The definition of linguistic terms is based on fuzzy set theory and hence we call the rules having these terms fuzzy association rules. The use of fuzzy techniques makes F-APACS resilient to noises such as inaccuracies in physical measurements of real-life entities and missing values in the databases. Unlike other algorithms which discover association rules based on the use of some user-supplied threshold such as the minimum support and minimum confidence, F-APACS employs adjusted difference analysis to identify interesting associations among attributes. This makes F-APACS to be able to avoid the use of some user-supplied thresholds which are often difficult to determine. F-APACS also has the unique features that it is able to discover both positive and negative associations and it uses a confidence measure, called the weight of evidence measure, to represent the uncertainty of the association rules. Experiments on a real-life database containing calling records of a PBX system showed that F-APACS is able to discover meaningful 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 the ACM SIGMOD Int'l Conf. on Management of Data*, Washington D.C., May 1993, pp. 207-216.

[2] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proc. of the 20th VLDB Conf.*, Santiago, Chile, 1994, pp. 487-499.

[3] 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.

[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 [14], pp. 107-123.

[6] J.Y. Ching, A.K.C. Wong, and K.C.C. Chan, "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 1-11, June 1995.

[7] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," in [8], pp. 1-34.

[8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[9] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus, "Knowledge Discovery in Databases: An Overview," in [14], pp. 1-27.

[10] J. Han, and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in *Proc. of the 21st VLDB Conf.*, Zurich, Switzerland, 1995, pp. 420-431.

[11] 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.

[12] J.M. Mendel, "Fuzzy Logic Systems for Engineering: A Tutorial," *Proc. of the IEEE*, vol. 83, no. 3, pp. 345-377, Mar. 1995.

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

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

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

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

[17] R.R. Yager, "On Linguistic Summaries of Data," in [14], pp. 347-363.

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