Rule Extraction Using Soft-Computing Techniques with Fuzzy Linguistic Representation

Nan-Chen Hsieh
Department of Information Management, National Taipei College of Nursing

Edword Wei
Department of Information Management, National Taipei College of Nursing

Abstract

Decisions for real-world problems are not always made precisely since the input data are themselves imprecise. This study presents a rough-fuzzy hybridization method to generate fuzzy if-then rules automatically from a diagnosis dataset with quantitative data values, based on fuzzy set and rough set theory. The proposed method consists of four stages: preprocessing inputs with fuzzy linguistic representation; rough set theory in finding notable reducts; candidate fuzzy if-then rules generation by data summarization, and truth evaluation the effectiveness of fuzzy if-then rules. The main contributions of the proposed method are the capability of fuzzy linguistic representation of the if-then rules, finding concise fuzzy if-then rules from diagnosis dataset, and tolerance of imprecise data.

Key words: Knowledge discovery in databases, fuzzy if-then rules, soft computing, fuzzy sets, rough sets

應用軟式計算技術擷取具模糊口語化語意之規則

謝楠楨
台北護理學院資訊管理系

魏立民
台北護理學院資訊管理系

摘要

在真實世界處理決策問題時，由於輸入資料本身即存在有不確定性，所以要做出明確的決策具有相當的困難。以模糊集合與概略集合理論為基礎，本研究提出了一個”概略-模糊混合的方法”，以從具有量化數據的診斷資料集合中，自動的產生模糊IF-THEN規則。所提出的方法包含有四個階段：輸入資料前處理使資料能呈現模糊口語化的語意、以概略集合理論找出顯著的”屬性縮減”、以資料彙總技術產生候選的IF-THEN規則、以及以真實值評估IF-THEN規則的有效性。本研究主要的貢獻，在於提出的方法所產生之IF-THEN規則具有口語化語意呈現的能力、能於診斷資料集合中找出精簡的模糊IF-THEN規則，以及具有不確定資料容忍的能力。

關鍵字： 由資料庫中發掘知識、模糊IF-THEN規則、軟式計算、模糊集合、概略集合

1. INTRODUCTION

Knowledge discovery in databases (KDD) has drawn the interest of the machine learning community. KDD generally involves statistical and data mining techniques to extract valuable knowledge in databases. Most KDD systems are designed to extract knowledge in a precise manner, rather than to provide a compact view that describes subsets of the database. Moreover, the knowledge representation is often uninformative for the user.

Fuzzy set theory, rough set theory, neural network and genetic algorithm are soft computing techniques widely used in the data mining step of the KDD process. Fuzzy set theory provides a natural framework for handling uncertainty. Rough set theory and neural network are used in rule generation and classification. Genetic algorithms are involved in various optimization and search processes. Besides, soft computing is a union methodology that works synergistically and provides flexible information processing capability for handling real-life ambiguous situations. Extending the relational databases to express knowledge, several studies (Batyrshin 2004; Bosc et al. 1999; Cubero et al. 1999; Jang 1993; Karr & Gentry 1993; Sugeno & Yasukawa 1993; Takagi & Hayashi 1991) have combined the ability of soft computing techniques with machine learning to represent and manage imprecise data together with the acceptable capacities for learning fuzzy if-then rules. Thus, fuzzy if-then rules can be generated from the fuzzy relations using linguistic representation.

Real-world data often contain imperfect information, while uncertainties, impreciseness and missing values are co-exist. The analysis of real-world data thus requires dealing with incomplete and inconsistent information, and manipulates various levels of data representation. However, soft computing techniques are based on quite strong assumptions. They cannot derive conclusions from incomplete knowledge, or manage inconsistent information. The idea of rough set was as a useful mathematical tool to deal with vague concepts and to represent ambiguity, vagueness and uncertainty.

Rough set algorithms (Pawlak 1982) do not need membership functions and prior parameter settings. It can extract knowledge from the data itself by means of indiscernibility relations, and generally needs fewer calculations than that of other soft computing techniques. Decision rules extracted by rough set are concise and valuable, which can benefit experts by revealing hidden knowledge in the dataset. The limitation of traditional rough set theory is concerned with discrete data; quantitative valued had to be discretized for rough set algorithms, which may result in some loss of information.

Many researchers proposed the hybridization of fuzzy set and rough set (Cock et al. 2007; Jensen & Shen 2004; Morsi & Yakout 1998; Qin & Pei 2005; Radzikowska & Kerre 2002). By these approaches, the comparison among objects turned from elements' indistinguishability into their similarity, and the similarity represented by a fuzzy equivalence relation. As a result, objects are categorized into classes with approximate boundaries based on their similarity to one another, and allowing an object belonging with various degrees to more than one class. But in traditional fuzzy rough set theory, fuzzy rough set theory involves only one fuzzy equivalence relation. In this study, we proposed a method that can model several similarity classes at the same time.

This study concentrates on automatically extracting the relevant fuzzy if-then rules in a dataset using fuzzy set and rough set theory. As depicted in Fig. 1, a four-stage rough-fuzzy hybridization process for learning fuzzy if-then rules in datasets was proposed. In the first stage, each input variable with a quantitative domain is automatically transformed into overlapping linguistic property sets, generating a fuzzy granulation of the feature space which contains granules with ill-defined boundaries. The linguistically terms were modeled by trapezoidal fuzzy sets defined in the appropriate attribute domains. Next, rough set theory was used to find notable reducts which model corresponding linguistic summaries in databases. Consequently, the linguistic summaries were used to generate candidate fuzzy if-then rules using the data summarization paradigm. Finally, a rule evaluation method was proposed to determine the effectiveness of the resulting fuzzy if-then rules by the fuzzy truth value judgment standards.

As stated, the core task of this four-stage process is the extraction/evaluation of fuzzy if-then rules. The extracted rules for the conventional rule generation approaches, such as association rule and classification rule, are usually very large, to the present of a huge proportion of redundant rules conveying the same information. By contrast with the conventional rule generation approaches, this study proposes a rough-fuzzy hybridization method (Cubero et al. 1999; Hsieh 2004), which generalizes concise fuzzy if-then rules. This method can be considered when the subsets of a database satisfy the linguistic summaries, then the linguistic summaries represent a set of fuzzy if-then rules explaining the subsets of data.

The rest of this study is organized as follows. Section 2 describes the analytical methodology of this study, and gives an overview of the linguistic summarization of databases and its application in extracting fuzzy if-then rules. Section 3 describes in detail the proposed rough-fuzzy hybridization method in constructing fuzzy rule-base, and shows an example of the ability to extract fuzzy if-then rules in a fuzzy database exhibiting linguistic summaries. The final section draws conclusions.

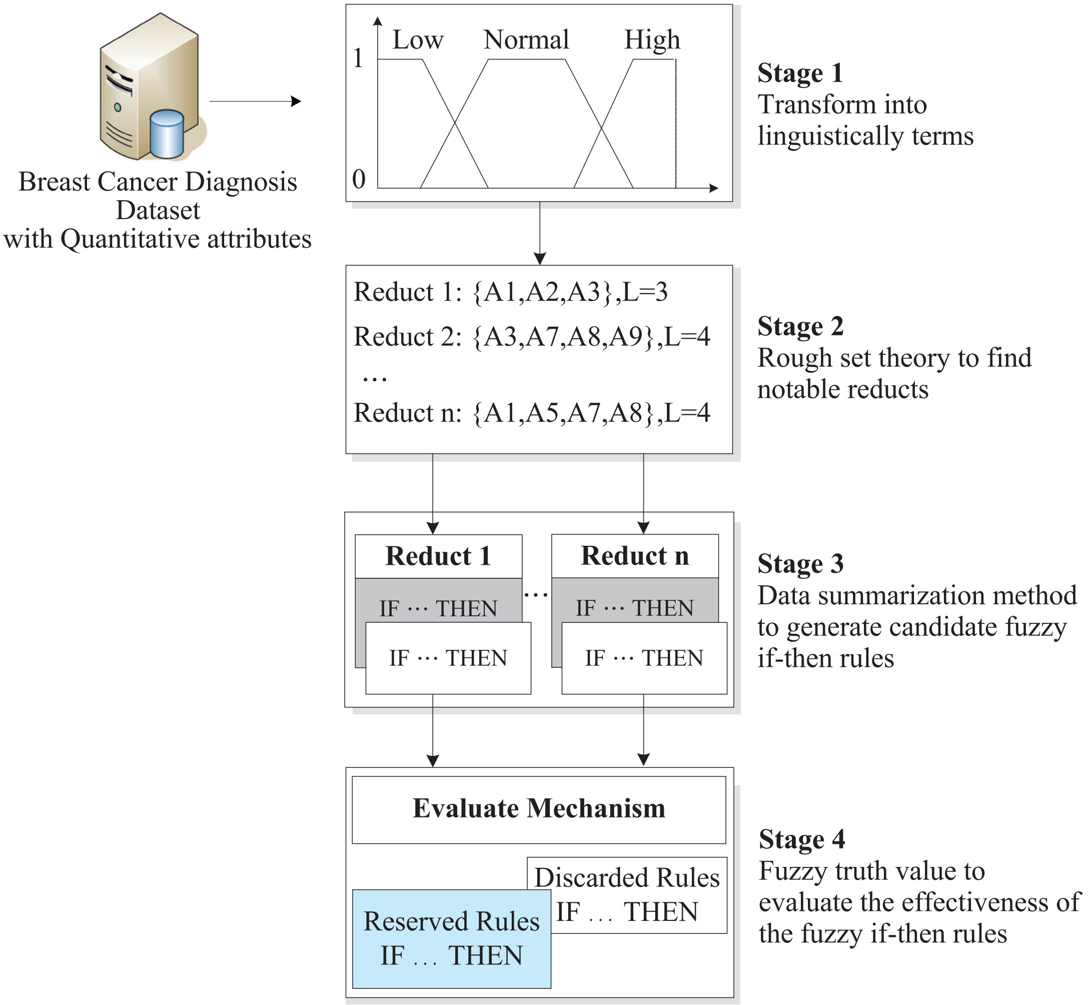

Breast Cancer Diagnosis
Dataset
with Quantitative attributes
Low
Normal
High
1
0
Stage 1
Transform into linguistically terms
Reduct 1: {A1,A2,A3},L=3
Reduct 2: {A3,A7,A8,A9},L=4
…
Reduct n: {A1,A5,A7,A8},L=4
Stage 2
Rough set theory to find notable reducts
Reduct 1
Reduct n
IF ··· THEN
IF ··· THEN
IF ··· THEN
IF ··· THEN
Stage 3
Data summarization method to generate candidate fuzzy if-then rules
Evaluate Mechanism
Discarded Rules
IF … THEN
Reserved Rules
IF … THEN
Stage 4
Fuzzy truth value to evaluate the effectiveness of the fuzzy if-then rules

Figure 1：A four-stage process for generating fuzzy if-then rules

2. THE FUNDAMENTALS OF THE ROUGH-FUZZY HYBRIDIZATION APPROACH

2.1 New fuzzy rough sets

In real-world data analysis, uncertainty is possible joined into databases. Decisions in many knowledge-intensive applications usually involve various forms of uncertainty. The values of attributes in databases may be symbolic or real-valued, and linguistic quantifiers (i.e. very, many, almost, etc.) are often used for conveying vague information in natural language (De &

Krishna 2004; Yeung et al. 2005).

Pawlak's rough set theory (Pawlak 1982; Pawlak 1991) is a method for discovering knowledge under uncertainty environment, the information systems of rough set theory contain crisp data, and the equivalence relation is a key and primitive notion. However, the equivalence relation seems to be a very strict condition that might limit the application domain of the rough set theory. To solve this problem, several authors have generalized the notion of approximation operators by using non-equivalence binary relations (Yao 1998b; Yao & Lin 1996). Alternatively, a fuzzy similarity relation was used to replace the equivalence relation. This is result a deviation of rough set theory called fuzzy rough sets (Dubois & Prade 1990).

Fuzzy rough sets contained a pair of fuzzy rough approximations of a fuzzy set by using the notions of the greatest t-norm(min), t-conorm(max), and a fuzzy similarity relation. Following Dubois and Prade' work, Morsi and Yakout (1998) developed a generalized definition of fuzzy rough sets by using a lower semi-continuous t-norm*, R-implication, and a fuzzy *-similarity relation. The axiomatic characterization of the fuzzy rough approximation operators was presented. Radzikowska and Kerre (2002) presented a general approach to fuzzy rough sets with reference to a t-norm, a special fuzzy implication, and a fuzzy similarity relation.

Besides, more researchers proposed the hybridization of fuzzy set and rough set (Cock et al. 2007; Jensen & Shen 2004; Morsi & Yakout 1998; Qin & Pei 2005; Radzikowska & Kerre 2002). By these approaches, the comparison among objects turned from elements' indistinguishability into their similarity, and the similarity represented by a fuzzy equivalence relation. As a result, objects are categorized into classes with approximate boundaries based on their similarity to one another, and allowing an object belonging with various degrees to more than one class. In these approaches, uncertainty is linked to information through the concept of granular structures (Zadeh 2005), and information is represented as a generalized constraint that is drawn from fuzzy set theory and fuzzy logic.

However, until now, most researches on rough sets and fuzzy rough sets are focused on the same universe, that is, the binary relations used in rough sets are defined on the same universe (Pawlak 1982; Pawlak 1991; Yao 1998a). Zhang and Wu (2000) proposed the approximation operators between different universes and constructed the rough set model using random sets. In this study, we followed the concept of granular structures (Liu et al. 2006) and proposed a method that can model several fuzzy similarity classes at the same time, and the approximation operators are between different universes of a new fuzzy rough sets.

2.2 Fuzzy queries involving linguistic summaries as rules mining tool

To learn rules from examples, fuzzy queries involving linguistic summaries (Bosc et al. 1999; Lee & Kim 1997) can be regarded as part of the data mining technique based on the association rules. The quantitative/categorical interface provided by fuzzy set theory, and reducts provided by rough set theory, are considered as fundamental to linguistic summary. The general form of a linguistic summary is “Q X objects in DB are S” , where Q denotes the fuzzy linguistic quantifier, X denotes a class of objects, DB denotes the database, and S denotes a property of the class or a linguistic summary that applies to the class quantified by Q. For example, a fuzzy rule “tall people in DB are heavy” can be justified using the validity of the linguistic summary “most tall people in DB are heavy” . Rasmussen and Yager (1999) proposed a SummarySQL for computing linguistic summaries with truth values as fuzzy predicates, which can be regarded as a method for generating fuzzy if-then rules in databases.

Hence, the discovery of fuzzy if-then rules is closely related to the validation of the corresponding linguistic summaries. However, notable linguistic summaries are difficult to find in a database. This study proposes a rough-fuzzy hybridization method for generating fuzzy if-then rules in databases. The candidate linguistic summaries were built by the reducts through rough set theory, and the fuzzy if-then rules were generated through the linguistic summary by the reducts. That is, the reducts infer the fuzzy if-then rules using the linguistic summary, enabling the database to be partitioned with less data than crisp summary for generating fuzzy if-then rules. Furthermore, the linguistic representation of data can handle vague, uncertain or imprecise information, as well as improve the accuracy and robustness of the linguistic summary construction process.

This study focuses on generating fuzzy if-then rules for the classification problem. For each fuzzy if-then rule, when the linguistic summaries in the antecedent are given, the consequent class membership and degree of truth can be determined.

To concretize the presentation, an example of the linguistic summary method used in generating fuzzy if-then rules is presented. Assume a classification problem with two features (Height, Weight), where Height={short, medium, tall} and Weight={light, normal, heavy}. As depicted in Fig. 2, the linguistic variables are described by two trapezoidal fuzzy sets. In Table 1, DB denotes a database with quantitative values, and DB_1 denotes a fuzzy database after the linguistic transformation. Moreover, DB_2 and denote two fuzzy databases with different threshold values associated with “Height” and “Weight” . Herein, if most data exhibit such linguistic summaries, then the fuzzy if-then rules can be validated by the corresponding linguistic summaries. For example, if the rule's antecedent is described by Height × Weight, then the linguistic expression, IF {Height=(tall,1)} AND Weight={(heavy,0.9)∨(heavy,1)} THEN Class=1, defines roughly the class membership of an object.

Table 1：The adult database

DB (database with quantitative attributes)			DB_1 (fuzzy database)		
Height	Weight	Class	Height	Weight	Class
177	83	0	(medium,0.8)	(normal,0.7)	0
170	75	0	(medium,1)	(normal,1)	0
170	83	0	(medium,1)	(normal,0.7)	0
185	89	1	(tall,1)	(heavy,0.9)	1
185	90	1	(tall,1)	(heavy,1)	1
…	…	…	…	…	…
DB_2, $\alpha_{\text{height}} = 1.0$			$DB_2^{'}$, $\alpha_{\text{height}} = 0.8$		
Height	Weight	Class	Height	Weight	Class
(medium, 0.8)	(normal, 0.7)	0	(medium, 0.8)	(normal,0.7) ∨ (normal,1)	0
(medium, 1)	(normal, 1) ∨ (normal, 0.7)	0			
(tall, 1)	(heavy,0.9) ∨ (heavy, 1)	1	(tall, 1)	(heavy,0.9) ∨ (heavy,1)	1
…	…	…	…	…	…

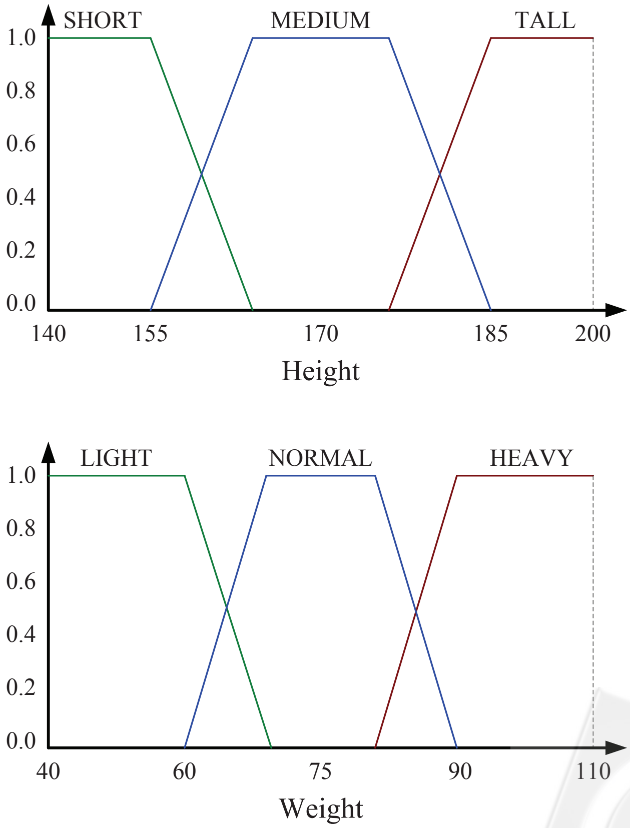


Figure 2：The linguistic variables “Height” and “Weight”

3. ASSESSING SOFT COMPUTING TECHNIQUES FOR GENERATING FUZZY IF-THEN RULES

3.1 The experimental dataset

The experimental dataset used in this study is a breast cancer diagnosis database obtained from the UCI machine learning repository at http://www.ics.uci.edu/~mlearn/databases/breast-cancer-wisconsin/. The Wisconsin diagnostic breast cancer (WDBC) dataset was collected at different periods of time with different characteristic of attributes. The WDBC base dataset consists of ten attributes for the cancer cell nuclear, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The data values of each attribute are quantitative. The mean, worst, and standard error of each attribute were computed from the base dataset, resulting in a total of thirty attributes. In this study, we only considered the mean values in the learning process. Besides, each sample was associated with a diagnosis class label, either “benign” or “malignant” . However, the diagnostics of this dataset did not provide additional information about the degree of benignity or malignancy. The dataset include 357 benign examples and 212 malignant examples. The classification problem is to learn rules from benign or malignant examples from the physical attributes of cell given in the dataset.

Several studies are based on this dataset. Peña-Reyes and Sipper (1999) proposed a fuzzy-genetic hybrid approach to produce a fuzzy rule-based diagnostic system, and later used fuzzy modeling and cooperative co-evolutionally techniques (Peña-Reyes & Sipper 2001) to predicate the class membership of examples. Setiono (2000) proposed a rule extraction technique to generate concise and accurate classification rules in a trained neural network. Tan et al. (2003) proposed a two-phase hybrid evolutionary classification technique to extract classification rules to be applied in clinical practice for better understanding and prevention of unwanted medical events. Chou et al. (2004) used neural network and MARS techniques to discover the breast cancer pattern.

Among these approaches, even with high classification accuracy, the diagnostic decisions are black boxes and the extracted knowledge is difficult to understand. Hence, a rough-fuzzy hybridization process could be developed to obtain informative and interpretable fuzzy if-then rules in databases.

3.2 Automatically transform quantitative data values into linguistic terms

Learning rules from examples is an active research topic in machine learning. However, most algorithms for learning rules from examples only accept categorical values, or sharp divide quantitative values into intervals. For example, the algorithm proposed in (Srikant & Agrawal 1996) initially partitions the quantitative attribute domain into small intervals and merges adjacent intervals into a larger interval, such that the combined intervals have sufficient supports. Then, the original attribute values are replaced by attribute-interval pairs, and the quantitative problem can be transformed into a Boolean one.

However, the sharp division of quantitative values either ignores or over-emphasizes the elements near the interval boundary during data mining. For example, the interval method may classify a person as “young” if age under 30 and “old” if age over 30, obviously does not correspond to human perceptions of “young” and “old”. Furthermore, sharp divisions do not easily distinguish the degree of membership. For example, ages of 60 and 80 might both be classified as “old”. However, people intuitively know that 80 is much older than 60. Fuzzy set theory can be used to solve this problem.

Kuck et al. (1998) proposed a fuzzy rule learning algorithm using pre-defined fuzzy sets as input data. Although such a fuzzy rule learning algorithm can solve the problem introduced by sharp division, it has some other problems. First, since the fuzzy sets are pre-defined, the interval definition for the categorical value is subjective. Second, such a transformation seems to be unnatural because the quantitative values, unlike categorical values, have a linear order. Moreover, the categorical values may have overlapping linguistic meanings which cannot be manipulated in the learning process. Shape division is also unintuitive with respect to human perception. To overcome these problems, this study investigates whether the intervals can be defined automatically from the dataset itself, rather than only by domain experts.

In this study, fuzzy set theory was employed for linguistic representation of quantitative data, thereby producing a fuzzy granulated of the attribute domain. A self-organizing map (SOM) algorithm was used to obtain k midpoints of the granular feature space from each quantitative attribute domain, following by the judgment of domain experts. Next, using fuzzy linguistic representation technique, each attribute domain was characterized as a trapezoidal fuzzy set with individually linguistic terms. The transformed terms are more closely than the linguistic meaning of quantitative data. Moreover, the extracted fuzzy if-then rules can be represented the learned knowledge in terms of human thinking, and tolerated imprecise information more robustly.

The steps for automatically finding fuzzy sets from a given dataset are described herein. Assume that the domain of a quantitative attribute ranges from v_1 to v_2, and $\{m_1, m_2, \cdots, m_k\}$ denote the k midpoints obtained by the SOM algorithm. Using these k midpoints, $k/2+1$ linguistic terms or membership functions can be determined for a trapezoidal fuzzy set. The first membership function is computed as:

$$f_{first}(x)=\begin{cases} 1 & \text{if} \quad v_1 \le x \le m_1 \\ (m_2-x)/(m_2-m_1) & \text{if} \quad m_1 < x < m_2 \\ 0 & \text{if} \quad x \ge m_2 \end{cases}.$$

Generally, the middle *p*th, *p* =2,.., *k*/2, membership function is computed as:

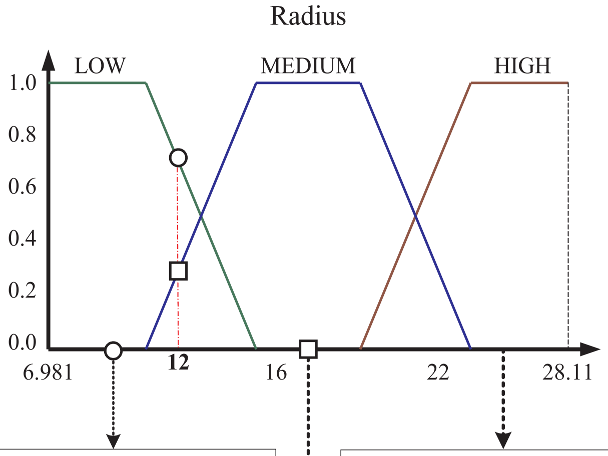
$$f_p(x)=\begin{cases} 0 & \text{if} \quad x \le m_{2p-3} \\ (x-m_{2p-3})/(m_{2p-2}-m_{2p-3}) & \text{if} \quad m_{2p-3} < x < m_{2p-2} \\ 1 & \text{if} \quad m_{2p-2} \le x \le m_{2p-1} \\ (m_{2p}-x)/(m_{2p}-m_{2p-1}) & \text{if} \quad m_{2p-1} < x < m_{2p} \\ 0 & \text{if} \quad x \ge m_{2p} \end{cases}.$$

The final membership function is computed as:

$$f_{final}(x)=\begin{cases} 0 & \text{if} \quad x \le m_{k-1} \\ (m_k-x)/(m_k-m_{k-1}) & \text{if} \quad m_{k-1} < x < m_k \\ 1 & \text{if} \quad m_k \le x \le v_2 \end{cases}.$$

For example, Radius is a quantitative attribute from the WDBC dataset. The domain of Radius ranges from 6.981 to 28.11. Four midpoints, (10.984, 14.084, 18.76, 23.39), were obtained using the SOM algorithm. As depicted in Fig. 3, a trapezoidal fuzzy set was obtained with three membership functions corresponding to the linguistic terms Low, Medium, and High. These linguistic terms were modeled by three membership functions defined in the appropriate attribute domains. If Radius={12} is a quantitative value, then it can be automatically transformed to a trapezoidal fuzzy set, {(Low,0.67), (Medium,0.33), (High,0)}, which signifies that the value of Radius can be either (Low,0.67), (Medium,0.33) or (High,0). This datum is a fuzzy information with overlapped fuzzy similarity classes.

As suggested by Medina et al. (1995) and Cubero et al. (1999), an ad hoc database system is not necessary built to store fuzzy information. A general relational database management system with special relations and dictionaries can be constructed to manage fuzzy information, thus enabling efficient data management.



$$f_{LOW}(x)=\begin{cases} 1 & \text{if } 6.981\le x\le 10.984 \\ (14.084-x)\big/(14.084-10.984) & \text{if } 10.984<x<14.084 \\ 0 & \text{if } x\ge 14.084 \end{cases}$$

$$f_{HIGH}(x)=\begin{cases} 0 & \text{if } x\le 18.76 \\ (23.39-x)/(23.39-18.76) & \text{if } 18.76<x<23.39 \\ 1 & \text{if } 23.39\le x\le 28.11 \end{cases}$$

$$f_{MEDIUM}(x)=\begin{cases} 0 & \text{if } x\le 10.984 \\ (x-10.984)/(14.084-10.984) & \text{if } 10.984<x<14.084 \\ 1 & \text{if } 14.084\le x\le 18.76 \\ (23.39-x)/(23.39-18.76) & \text{if } 18.76<x<23.39 \\ 0 & \text{if } x\ge 23.39 \end{cases}$$

Figure 3：The trapezoidal fuzzy set of Radius

3.3 Rough set theory for finding candidate linguistic summaries

Data summarization is a knowledge discovery technique providing the user with comprehensive information for grasping the essence from a large amount of examples in a database. Fuzzy set theory works well for data summarization in real-world application, and the generalized partition of tuples are as a representational form of a linguistic summary including fuzzy concepts. Since the interpretation and exploration of linguistic summaries are the main goals of data summarization, the reducts provided by the rough set theory could be served as the candidate linguistic summaries to extract fuzzy if-then rules.

From the viewpoint of data summarization, fuzzy query involving linguistic summaries are a useful utility, since they allow one to express relationship among attributes in the form of fuzzy if-then rules which are valid on a given subset of examples in a database. For example, Cubero et al. (1999) and Rasmussen and Yager (1999) defined linguistic summary in terms of fuzzy functional dependencies and showed how fuzzy rules can be extracted from a database. Bosc et al. (1999) proposed a data mining algorithm to extract fuzzy functional dependencies using gradual rules.

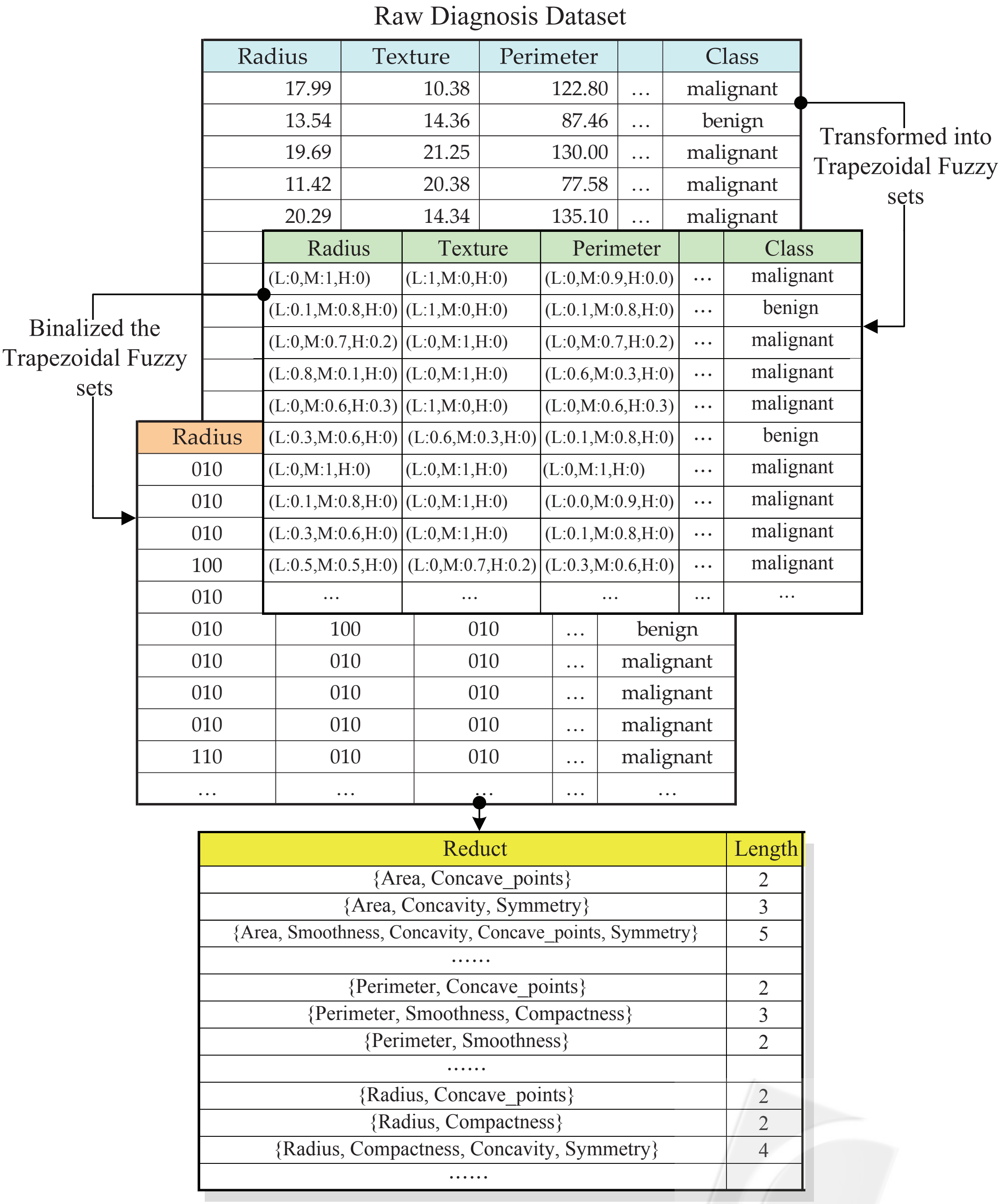
Since rough set theory is focused on the ambiguity caused by limited discernibility of objects in the domain of discourse, therefore, this study employs rough set theory to find candidate linguistic summaries. The core of rough set theory is finding reducts. A reduct contains a clump of objects in the universal of discourse drawn together by the indistinguishability relation. Hence, rough set theory can be employed to reduce the data by identifying indistinguishability classes through indistinguishability relation. The reduction step in rough set evolutional processes keeps only those attributes which preserve the indiscernibility relation. Therefore, minimal subsets of attributes that induce the partitions on the same target attributes with higher support are concerned. In other words, the essence of information remains intact, and superfluous attributes are removed. The remaining sets of attributes, called reducts, represent the fundamental integrity constraint of the database.

However, rough set theory can only handle precise values. Therefore, the trapezoidal fuzzy sets were binalized according to their membership values. For example, the trapezoidal fuzzy set {(Low,0.8), (Medium,0.6), (High,0.3)} is binalized as “100”. Fig. 4 shows the transformed WDBC dataset and the extracted reducts. By applying rough set theory, several reducts can be generated from the experimental dataset.

3.4 Using fuzzy truth value to evaluate the confidence of fuzzy if-then rules

For generating fuzzy if-then rules, fuzzy queries involving linguistic summaries are derived as linguistically quantified propositions with a truth value representing the effectiveness of the generated fuzzy if-then rules. Generally, each fuzzy if-then rule often involves several linguistically quantified propositions, indicating a search problem for determining the most appropriate combinational linguistic summaries from all possible combinations of attributes. For this purpose, Kacprzyk et al. (1989) developed a fuzzy query language for interactive linguistic summarization using natural terms and comprehensible quantifiers. Rasmussen and Yager (1999) defined searching processes for fuzzy and gradual functional dependencies in the light of linguistic summaries which can also be used for knowledge discovery. In these approaches, the interesting linguistic summaries are difficult to generate and generally user interaction is required.

As stated by Yager (1996), the linguistic summary is a linguistically quantified proposition containing meta-knowledge about a set of particular objects, and is useful in knowledge discovery. This study aims to consider the notable subsets of tuples in the fuzzy relational database, and to construct linguistic summaries in which attribute values are fuzzy linguistic labels describing each subset of tuples. Thus, for each notable linguistic summary “Q X objects in DB are S”, the attributes S were determined using rough set theory to the class of objects X, and the validation process is to test the truth of the association between the X and S with respect to the quantifier, Q. Fig. 5 shows an example of two fuzzy if-then rules generated by linguistic summary. Finally, the truth value of a linguistic summary is a number in the unit interval, such that a value close to one indicates that the proposed linguistic summary is likely to be truthful.



Transformed into Trapezoidal Fuzzy sets
Binalized the Trapezoidal Fuzzy sets

Raw Diagnosis Dataset

Radius	Texture	Perimeter		Class
17.99	10.38	122.80	…	malignant
13.54	14.36	87.46	…	benign
19.69	21.25	130.00	…	malignant
11.42	20.38	77.58	…	malignant
20.29	14.34	135.10	…	malignant

Radius	Texture	Perimeter		Class
(L:0,M:1,H:0)	(L:1,M:0,H:0)	(L:0,M:0.9,H:0.0)	…	malignant
(L:0.1,M:0.8,H:0)	(L:1,M:0,H:0)	(L:0.1,M:0.8,H:0)	…	benign
(L:0,M:0.7,H:0.2)	(L:0,M:1,H:0)	(L:0,M:0.7,H:0.2)	…	malignant
(L:0.8,M:0.1,H:0)	(L:0,M:1,H:0)	(L:0.6,M:0.3,H:0)	…	malignant
(L:0,M:0.6,H:0.3)	(L:1,M:0,H:0)	(L:0,M:0.6,H:0.3)	…	malignant
(L:0.3,M:0.6,H:0)	(L:0.6,M:0.3,H:0)	(L:0.1,M:0.8,H:0)	…	benign
(L:0,M:1,H:0)	(L:0,M:1,H:0)	(L:0,M:1,H:0)	…	malignant
(L:0.1,M:0.8,H:0)	(L:0,M:1,H:0)	(L:0.0,M:0.9,H:0)	…	malignant
(L:0.3,M:0.6,H:0)	(L:0,M:1,H:0)	(L:0.1,M:0.8,H:0)	…	malignant
(L:0.5,M:0.5,H:0)	(L:0,M:0.7,H:0.2)	(L:0.3,M:0.6,H:0)	…	malignant
…	…	…	…	…

Radius				
010				
010				
010				
100				
010				
010	100	010	…	benign
010	010	010	…	malignant
010	010	010	…	malignant
010	010	010	…	malignant
110	010	010	…	malignant
…	…	…	…	…

Reduct	Length
{Area, Concave_points}	2
{Area, Concavity, Symmetry}	3
{Area, Smoothness, Concavity, Concave_points, Symmetry}	5
……	
{Perimeter, Concave_points}	2
{Perimeter, Smoothness, Compactness}	3
{Perimeter, Smoothness}	2
……	
{Radius, Concave_points}	2
{Radius, Compactness}	2
{Radius, Compactness, Concavity, Symmetry}	4
……	

Figure 4：The transformed WDBC dataset and the extracted reducts

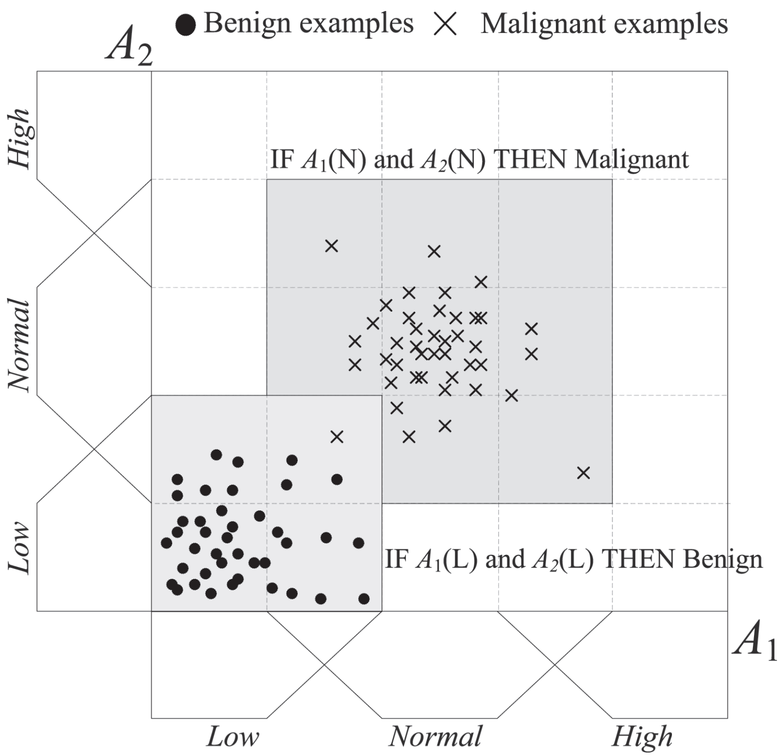

● Benign examples × Malignant examples
A_2
High
Normal
Low
IF A_1(N) and A_2(N) THEN Malignant
IF A_1(L) and A_2(L) THEN Benign
A_1
Low
Normal
High

Figure 5：Fuzzy if-then rules generated by linguistic summary

The fuzzy logic based calculus provides the interpreting and validating of the truth statement involving complex linguistic quantifiers, such as “many” , “some” and “few” . Let “Q $\{t_1,\ldots,t_n\}$are S” denotes a linguistically quantified statement, and let $\{t_1,\ldots,t_n\}$ denotes a set of fuzzy tuples in the fuzzy database, DB. The procedure for determining the truth value of a linguistically quantified statement is as follows. If the summary S involves an attribute A, and t_i denotes a tuple that satisfies the summary S, then the membership value of t_i to S is given by:

$$S(t_i) = \max_{\forall k}\left(\mu_{EQ}(a_k, b_k)\right), \text{ for all } a_k \in t_i[A],\ b_k \in S,$$

where $S(t_i)$ denotes the degree to which t_i satisfies the summary S, and the function $\mu_{EQ}(a_k, b_k) = 0$ if and only if $(a_k \neq b_k) \vee (a_k = b_k \wedge \mu(b_k) = 0)$; $\mu_{EQ}(a_k, b_k) = 1 - |\mu(a_k) - \mu(b_k)|$ if and only if $(a_k = b_k \wedge \mu(b_k) \neq 0)$. Then, the individual truth value of “ $\{t_1,\ldots,t_n\}$ are S” for an attribute A over DB is computed as:

$$Truth(\{t_1,\ldots,t_n\} \text{ are } S) = \left(\frac{1}{n}\sum_{i=1}^{n} S(t_i)\right).$$

Moreover, when the linguistic summaries are distributed over m attributes with *ANDed* conditions, that is, $S = S_1 \wedge \ldots \wedge S_m$, then the total truth value $Truth(\{t_1,\ldots,t_n\}$

are S) = are $S) = \min_{j=1}^{m}(Truth(\{t_1,...,t_n\}$ are $S_j)$. When the linguistic summaries are distributed over m attributes with *ORed* conditions, that is, $S = S_1 \vee \ldots \vee S_m$, then the total truth value $Truth(\{t_1,\ldots,t_n\}$ are $S)$ = $\max_{j=1}^{m}(Truth(\{t_1,...,t_n\}$ are $S_j)$ .Finally, T= $Q(Truth(\{t_1,\ldots,t_n\}$ are $S))$ denotes the truth value of the linguistically quantified statement "Q $\{t_1,\ldots,t_n\}$ are S" to the fuzzy quantifier Q in agreement. Fig. 5 shows the procedure in obtaining the truth value of the linguistic summaries.

For example, a notable reduct, {Area, Concave_points}, was determined using the rough set theory. The fuzzy quantifier sets up as "*some*", which is defined as the membership function $\mu_{some}(x) = 1/e^{(x-0.38)/0.08*(x-0.38)/0.08}$. Then the linguistically quantified statement for validating the fuzzy if-then rule is given by:

"*Some* of the benign breast's cells have *low* Area and *low* Concave_points."

The method of computing the individual truth value for the attribute Area is described as follows. Let the linguistic summary, S_1(*low* Area)=(Low:1, Medium:0, High:0) and a tuple t_i =(Low:0.8, Medium:0.1, High:0). Then, the membership value of t_i to S_1 is given by S(t_1)= max(0.8,0,0). For n tuple $\{t_1,\ldots,t_n\}$, the individual truth value "$Truth(\{t_1,\ldots,t_n\}$ are S_1" over *DB* is computed as $\left(\frac{1}{n}\sum_{i=1}^{n} S(t_i)\right)$=0.437, where n=188. Similarly, for linguistic summary, S_2(low Concave_points), the individual truth value "*Truth*({t1,⋯,tn} are S_2" over *DB* is 0.473. Since the linguistic quantified statement is *ANDed* with linguistic summaries, the total truth value is min(0.437,0.473)=0.437. After applying the linguistic quantifier "*some*", a fuzzy if-then rule with truth is obtained in the following form:

IF Area *is Low and* Concave_points *is Low* THEN Breast_Cancer *is benign,* 0.602.

With the judgment standards of support and total truth value, Table 2 shows the fuzzy if-then rules generated by the proposed rough-fuzzy hybridization process. Suitable linguistic quantifier can be employed to interpret linguistic confidence.

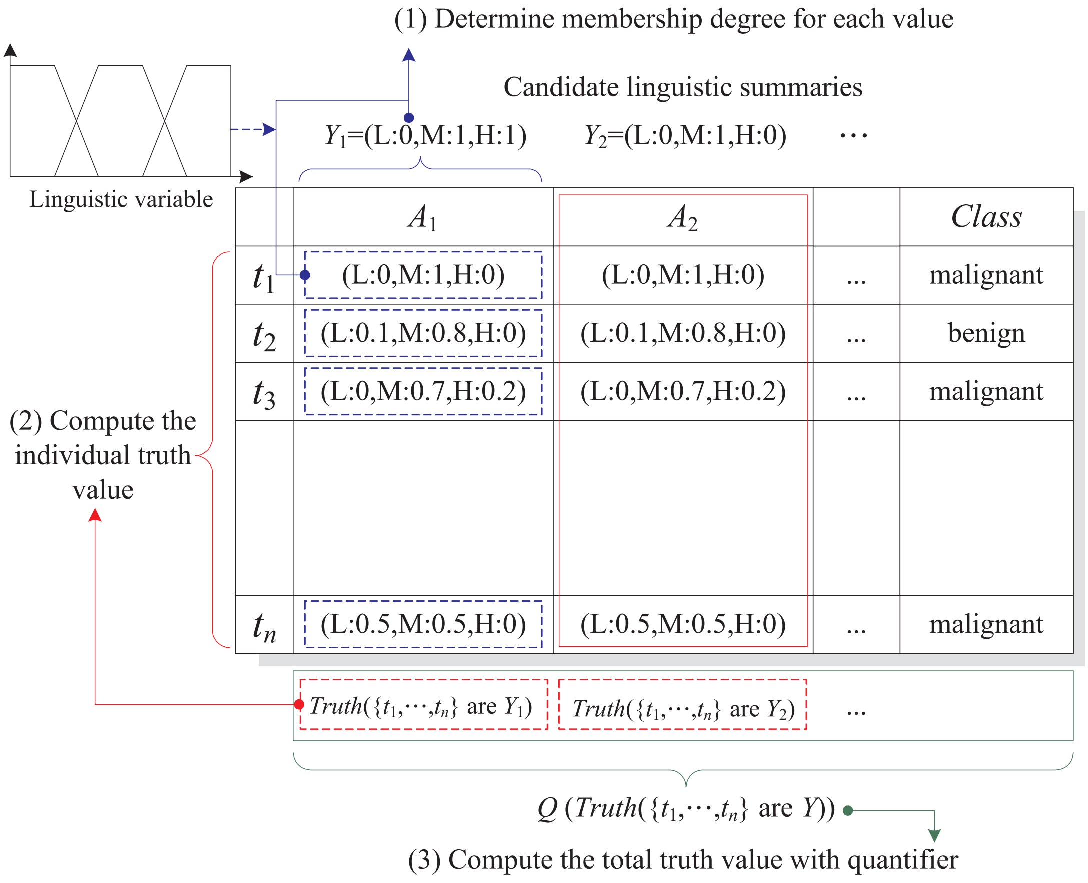

(1) Determine membership degree for each value
Candidate linguistic summaries
Y_1=(L:0,M:1,H:1)
Y_2=(L:0,M:1,H:0)
···
Linguistic variable
A_1
A_2
Class
t_1
(L:0,M:1,H:0)
(L:0,M:1,H:0)
...
malignant
t_2
(L:0.1,M:0.8,H:0)
(L:0.1,M:0.8,H:0)
...
benign
t_3
(L:0,M:0.7,H:0.2)
(L:0,M:0.7,H:0.2)
...
malignant
t_n
(L:0.5,M:0.5,H:0)
(L:0.5,M:0.5,H:0)
...
malignant
(2) Compute the individual truth value
$Truth(\{t_1,\cdots,t_n\}$ are $Y_1)$
$Truth(\{t_1,\cdots,t_n\}$ are $Y_2)$
...
$Q\ (Truth(\{t_1,\cdots,t_n\}$ are $Y))$
(3) Compute the total truth value with quantifier

Figure 5：The procedure for obtaining the truth value of the linguistic summaries

Table 2：The fuzzy if-then rules generated by the rough-fuzzy hybridization method

ID	IF	THEN	Support	Accuracy	Total Truth Value
1	Area(Low) AND Concave_points(Low)	benign	0.3304	100%	0.601
2	Perimeter(Low) AND Concave_points(Low)	benign	0.2689	100%	0.852
3	Radius(Low) AND Concave_points(Low)	benign	0.2742	100%	0.928
4	Perimeter(Low) AND Smoothness(Normal) AND Compactness(Normal)	benign	0.0580	100%	0.852
5	Area(Normal) AND Concavity(Normal) AND Symmetry(Low, Normal)	malignant	0.0053	100%	0.805
	…	…	…		…

4. CONCLUSION

The extracted rules for the conventional association rule method are usually very large, to the present of a huge proportion of redundant rules conveying the same information. Many of the rules may contain redundant, irrelevant information or describe trivial knowledge. This study proposes a rough-fuzzy hybridization method for learning informative and concise fuzzy if-then rules from examples. The quantitative/categorical interface provided by fuzzy set theory is used for the linguistic representation of examples, and balances the expert perception and system automation. Besides, the reducts provided by rough set theory were found to be a useful tool for finding candidate linguistic summaries. Hence, the discovery of fuzzy if-then rules is similar to the validation of the corresponding linguistic summaries, and the generated fuzzy if-then rules are on the basis of equivalence relation to enhance its readability. Moreover, this study proposed to use fuzzy truth value to evaluate the confidence of fuzzy if-then rules. In contrast to the conventional rule mining method, the proposed method can handle imprecise information and improve the efficiency and robustness of the rule base construction process. Important future work is deployed the proposed method into real system.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 95-2416-H-227 -001.

REFERENCES

1. Batyrshin, I. “On linguistic representation of quantitative dependencies,” *Expert Systems with Applications* (26), 2004, pp. 95-104.
2. Bosc, P., Pivert, L., and Ughetto, L. “On data summaries based on gradual rules,” *in: Proc. Internat. Conf. on Computational Intelligence, 6th Dortmund Fuzzy Days (DFD'99), Lecture Notes in Computer Science, Springer, Dortmund, Germany, 25-28 May* (1625), 1999, pp. 512-521.
3. Chou, S.-M., Lee, T.-S., Shao, Y.E., and Chen, I.-F. “Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines,” *Expert Systems with Applications* (27), 2004, pp. 133-142.
4. Cock, M.D., Cornelis, C., and Kerre, E.E. “Fuzzy rough sets: The forgotten step,” *IEEE Transactions on Fuzzy Systems* (15), 2007, pp. 121-130.
5. Cubero, J.C., Medina, J.M., Pons, O., and Vila, M.A. “Data summarization in relational

databases through fuzzy dependencies," *Information Sciences* (121), 1999, pp. 233-270.

6. De, S.K., and Krishna, P.R. "A new approach to mining fuzzy databases using nearest neighbor classification by exploiting attribute hierarchies," *International Journal of Intelligent Systems* (19), 2004, pp. 1277-1290.
7. Dubois, D., and Prade, H. "Rough fuzzy sets and fuzzy rough sets," I*nternational Journal of General Systems* (17), 1990, pp. 191-209.
8. Hsieh, N.C. "Handling indefinite and maybe information in logical fuzzy relational databases," *International Journal of Intelligent Systems* (19:3), 2004, pp. 257-276.
9. Jang, J.S.R. "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics* (23), 1993, pp. 665-685.
10. Jensen, R., and Shen, Q. "Semantics-preserving dimensionality reduction: rough and fuzzy-rough based approaches," *IEEE Transactions on Knowledge and Data Engineering* (16:12), 2004, pp. 1457-1471.
11. Kacprzyk, J., Zadrozny, S., and Zi?lkowski, A. "FQUERY III+: a "human-consistent" database querying system based on fuzzy logic with linguistic quantifiers," *Information Systems* (14:6), 1989, pp. 443-453.
12. Karr, C.L., and Gentry, E.J. "Fuzzy control of pH using genetic algorithms," *IEEE Transactions on Fuzzy Systems* (1), 1993, pp. 46-53.
13. Kuck, C.M., Fu, A., and Wong, M.H. "Fuzzy association rules in large databases with quantitative attributes," ACM SIGMOD Records, 1998.
14. Lee, D.H., and Kim, M.H. "Database summarization using fuzzy ISA hierarchies," *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics* (27), 1997, pp. 68-78.
15. Liu, M., Chen, D., Wu, C., and Li, H. "Fuzzy reasoning based on a new fuzzy rough set and its application to scheduling problems," *Computers and Mathematics with Applications* (51), 2006, pp. 1507-1518.
16. Medina, J.M., Vila, M.A., Cubero, J.C., and Pons, O. "Toward the implementation of a generalized fuzzy relational database model," *Fuzzy Sets and Systems* (75), 1995, pp. 273-289.
17. Morsi, N.N., and Yakout, M.M. "Axiomatics for fuzzy rough sets," *Fuzzy Sets and Systems* (100:1-3), 1998, pp. 327-342.
18. Pawlak, Z. "Rough sets," *International Journal of Computer and Information Sciences* (11:5), 1982, pp. 341-356.
19. Pawlak, Z. *Rough sets, theoretical Aspects of reasoning about data* Kluwer Academic Publishers, Boston, 1991.
20. Peña-Reyes, C.A., and Sipper, M. "A fuzzy-genetic approach to breast cancer diagnosis," *Artificial Intelligence in Medicine* (17), 1999, pp. 131-155.
21. Peña-Reyes, C.A., and Sipper, M. "Fuzzy CoCo: A cooperative coevolutionary approach

to fuzzy modeling,” *IEEE Transactions on Fuzzy Systems* (9:5), 2001, pp. 727-737.

22. Qin, K., and Pei, Z. “On the topological properties of fuzzy rough sets,” *Fuzzy Sets and Systems* (151:3), 2005, pp. 601-613.
23. Radzikowska, A.M., and Kerre, E.E. “A comparative study of fuzzy rough sets,” *Fuzzy Sets and Systems* (126), 2002, pp. 137-155.
24. Rasmussen, D., and Yager, R.R. “Finding fuzzy and gradual functional dependencies with SummarySQL,” *Fuzzy Sets and Systems* (106), 1999, pp. 131-142.
25. Setiono, R. “Generating concise and accurate classification rules for breast cancer diagnosis,” *Artificial Intelligence in Medicine* (18:3), 2000, pp. 205-217.
26. Srikant, R., and Agrawal, R. “Mining quantitative association rules in large relational tables,” Proceedings of ACM SIGMOD, 1996, pp. 1-12.
27. Sugeno, M., and Yasukawa, T. “A fuzzy-logic-based approach to qualitative modeling,” *IEEE Transactions on Fuzzy Systems*, 1993, pp. 7-31.
28. Takagi, H., and Hayashi, I. “NN-driven fuzzy reasoning,” *Approximate Reasoning* (5), 1991, pp. 191-212.
29. Tan, K.C., Yu, Q., Heng, C.M., and Lee, T.H. “Evolutionary computing for knowledge discovery in medical diagnosis,” *Artificial Intelligence in Medicine* (27), 2003, pp. 129-154.
30. Yager, R.R. “Database discovery using fuzzy sets,” *International Journal of Intelligent Systems* (11:9), 1996, pp. 691-712.
31. Yao, Y.Y. “Constructive and algebraic methods of the theory of rough sets,” *Information Sciences* (109), 1998a, pp. 21-47.
32. Yao, Y.Y. “Relational interpretations of neighborhood operators and rough set approximation operators,” *Information Sciences* (111), 1998b, pp. 239-259.
33. Yao, Y.Y., and Lin, T.Y. “Generalization of rough sets using modal logic,” *Intelligent Automation and Soft Computing, an International Journal* (2), 1996, pp. 103-120.
34. Yeung, D.S., Chen, D.G., Tsang, C.C., Lee, W.T., and Wang, X.Z. “On the generalization of fuzzy rough sets,” *IEEE Transactions on Fuzzy Systems* (13), 2005, pp. 343-361.
35. Zadeh, L.A. “Towards a generalized theory of uncertainty (GTU)-an outline,” *Information Sciences* (172), 2005, pp. 1-40.
36. Zhang, W.X., and Wu, W.Z. “The rough set model based on the random set,” *Journal of Xi'an Jiaotong University* (34:12), 2000, pp. 15-19.