Georgia State University
ScholarWorks @ Georgia State University

Computer Science Dissertations

Department of Computer Science

12-4-2006

Fuzzy-Granular Based Data Mining for Effective Decision Support in Biomedical Applications

Yuanchen He

Recommended Citation

He, Yuanchen, "Fuzzy-Granular Based Data Mining for Effective Decision Support in Biomedical Applications." Dissertation, Georgia State University, 2006.
https://scholarworks.gsu.edu/cs_diss/12

FUZZY-GRANULAR BASED DATA MINING FOR EFFECTIVE DECISION SUPPORT IN BIOMEDICAL APPLICATIONS

by

YUANCHEN HE

Under the Direction of Raj Sunderraman and Yan-Qing Zhang

ABSTRACT

Due to complexity of biomedical problems, adaptive and intelligent knowledge discovery and data mining systems are highly needed to help humans to understand the inherent mechanism of diseases. For biomedical classification problems, typically it is impossible to build a perfect classifier with 100% prediction accuracy. Hence a more realistic target is to build an effective Decision Support System (DSS).

In this dissertation, a novel adaptive Fuzzy Association Rules (FARs) mining algorithm, named FARM-DS, is proposed to build such a DSS for binary classification problems in the biomedical domain. Empirical studies show that FARM-DS is competitive to state-of-the-art classifiers in terms of prediction accuracy. More importantly, FARs can provide strong decision support on disease diagnoses due to their easy interpretability.

This dissertation also proposes a fuzzy-granular method to select informative and discriminative genes from huge microarray gene expression data. With fuzzy granulation, information loss in the process of gene selection is decreased. As a result, more informative genes for cancer classification are selected and more accurate classifiers can be modeled. Empirical studies show that the proposed method is more accurate than traditional algorithms for cancer classification. And hence we expect that genes being selected can be more helpful for further biological studies.

INDEX WORDS:

Data Mining, Knowledge Discovery, Computational Intelligence, Granular Computing, Fuzzy Association Rule Mining, Decision Support System, Binary Classification, Bioinformatics

FUZZY-GRANULAR BASED DATA MINING FOR EFFECTIVE DECISION SUPPORT IN BIOMEDICAL APPLICATIONS

by

YUANCHEN HE

A Dissertation Submitted in Partial Fulfillment of Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia Stage University

2006

FUZZY-GRANULAR BASED DATA MINING FOR EFFECTIVE DECISION SUPPORT IN BIOMEDICAL APPLICATIONS

by

YUANCHEN HE

Major Professor:	Rajshekhar Sunderraman
	Yan-Qing Zhang
Committee:	Saeid Belkasim
	Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2006

Acknowledgments

Firstly, my specific thanks go to my co-advisors, Dr. Rajshekhar Sunderraman and Dr. Yan-Qing Zhang, for their careful guidance and precise advisement during the process of my PhD dissertation. The dissertation would not have been possible without their helps.

Secondly, I would like to thank my committee members, Dr. Saeid Belkasim and Dr. Yichuan Zhao for their well-appreciated support and assistance.

Finally, I want to thank my family and friends for their support and beliefs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Fuzzy Association Rule Mining	FARM
Decision Support	DS

CHAPTER 1

INTRODUCTION

In the last decade, with the advent of genomic and proteomic technologies, more and more biomedical databases have been created and have been growing in an exponential rate. Developing intelligent data analysis tools is essential to extract knowledge from these databases to ease biomedical decision-making process. The knowledge extracted from these databases is expected to be as accurate as possible. However, due to complexity and huge sizes of biomedical databases, it is difficult or even impossible to find 100% accurate knowledge. Therefore, a more realistic goal is to build an intelligent data analysis tool as an effective Decision Support System (DSS). That is, the role of such a data analysis tool is not to replace human experts, but only to assist human experts to make decisions more reliably.

1.1 Problem definitions

1.1.1 Binary classification

In this dissertation, we focus on binary classification modeling. Although binary classification is the simplest classification problem, many works show that the models for it can be naturally extended to multiple classification or regression problems. (This extension itself is an interesting research topic and will not be covered in this dissertation.)

A general binary classification problem is defined as follows:

- Given l independent and identically distributed (i.i.d.) samples $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$ where $x_i \in R^d$, for $i = 1,2,\cdots,l$ is a feature vector of length d and $y_i = \{+1, -1\}$ is the class label (+1 for the positive class, and -1 for the negative class) for data point x_i,

- Assume the classes are mutually exclusive and exhaustive, which means every sample has one and only one class label,
- Find a classifier with the decision function $f(x, \theta)$ such that $y = f(x, \theta)$, where y is the class label for x, θ is a vector of unknown parameters in the function. These l samples are called "training data".

	real negatives	real positives
predicted negatives	(TN) true negatives	(FN) false negatives
predicted positives	(FP) false positives	(TP) true positives

Figure. 1.1. confusion matrix

1.1.2 Feature selection

Some binary classification problem is more natural to be modeled as a binary ranking modeling. Protein homology prediction task is a good example. The target is to predict if a protein sequence is homologous to another pre-specified natural protein sequence. Because of biological complexity, it is difficult and arbitrary to say two protein sequences are absolutely homologous or not (1 or -1 is output); an output with "confidence" may be more helpful. In this way, many protein sequences could be ranked by their confidence to be homologous to the pre-specified protein sequence. As a result, biologists could quickly prioritize a list of protein sequences for further study and thus their working efficiencies can be enhanced.

A binary ranking problem is similar to a binary classification problem. The differences are

- the output is a real number in the field of [-1,1], and
- the absolute value of the output is useless. Intuitively, a good model should rank the unseen positive samples (in case of protein homology prediction, they are homologous protein sequences) close to the top and rank unseen negative samples (in case of protein homology prediction, they are non-homologous protein sequences) close to the bottom of the list.

1.1.3 Feature selection

Feature selection is another important task usually correlated with a classification problem. Given a dataset, some input features may be irrelevant to classification. Furthermore, some features may be redundant or even noise due to complex correlations among them to hide real data distribution. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant, redundant or noisy features from the learning process. In machine learning, this process is known as feature selection to filter out features, which may otherwise slow down, and possibly mislead, the learning step. Relevance analysis is closely related to binary classification. Suppose there are d input features in the original dataset, the target of feature selection is to select d_i informative features while removing d_n non-informative features. Here $d_i > 0$, $d_n >= 0$, $d_i + d_n = d$. The target is that the classifier modeled on the subset of d_i features has better performance than the classifier modeled in the original feature set.

1.2 Metrics for classification

The performance of the classifier is usually measured in terms of misclassification error on unseen “testing data” which is defined in Eq. (1.1).

$$E(y, f(x,\theta)) = \begin{cases} 0 & if \;\; y = f(x,\theta), \\ 1 & \quad otherwise \end{cases} \tag{1.1}$$

Based on the confusion matrix in Fig. 1.1, many other metrics have been used for performance evaluation on classification.

- Accuracy is the fraction of correctly classified samples over all samples.

$$accuracy = \frac{TN + TP}{TN + FN + FP + TP}\,. \tag{1.2}$$

The overall accuracy metric at Eq. (1.2) represents the same meaning as misclassification error. Both of them are used to evaluate classification performance on the whole dataset. Besides them, two other kinds of metrics have been proposed for different purposes.

The first kind of metrics is concern with balanced classification ability. Sensitivity at Eq. (1.3) and specificity at Eq. (1.4) are usually adopted to monitor classification performance on two classes, separately.

- Sensitivity is the fraction of the real positives that actually are correctly predicted as positives.
- Specificity is the fraction of the real negatives that actually are correctly predicted as negatives.

$$sensitivity = \frac{TP}{TP + FN}\,. \tag{1.3}$$

$$specificity = \frac{TN}{TN + FP}\,. \tag{1.4}$$

Notice that sensitivity is sometimes called true positive rate or positive class accuracy, while specificity called true negative rate or negative class accuracy, in different research communities. By the definitions, the combination of sensitivity and specificity can be used to evaluate a model's balance ability so that we know if a model is biased to a special class. Notice that the sum of *FP* and *FN* is the number of misclassification errors on the unseen testing dataset. Based on these two metrics, g-mean was proposed in [76] at Eq. (1.5), which is the geometric mean of classification accuracy on positive samples and classification accuracy on negative samples. Area under ROC curve (AUC-ROC) [19], as shown in Fig. 1.2, can also indicate a classifier's balance ability between sensitivity and specificity as a function of varying a classification threshold.

$$g - mean = \sqrt{sensitivity \times specificity} \tag{1.5}$$

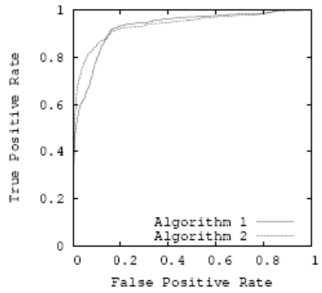

Figure. 1.2. Sample of Area under ROC curve

There is a traditional academic point system to roughly guide the performance evaluation on the AUC metric [113]:

$0.9 \leq auc \leq 1$	=	excellent	(A)
$0.8 \leq auc < 0.9$	=	good	(B)
$0.7 \leq auc < 0.8$	=	fair	(C)
$0.6 \leq auc < 0.7$	=	poor	(D)
$0.5 \leq auc < 0.6$	=	fail	(F)

On the other hand, sometimes we are interested in highly effective detection ability for only one class. For example, for credit card fraud detection problem, the target is detecting fraudulent transactions. For diagnosing a rare disease, what we are especially interested in is to find patients with this disease. For such kind of problems, another pair of metrics, precision at Eq. (1.6) and recall at Eq. (1.7), is often adopted.

- Precision is the fraction of the samples predicted as positives that really are positives.
- Recall is the fraction of the real positives that actually are correctly predicted as positives.

Notice that recall is the same as sensitivity. F-value at Eq. (1.8) is used to integrate precision and recall into a single metric for convenience of modeling. Similarly, area under precision/recall curve (AUC-PR), as show in Fig. 1.3 is also used to indicate a classifier's detection ability between precision and recall as a function of varying a classification threshold.

$$precision = \frac{TP}{TP + FP} \tag{1.6}$$

$$recall = \frac{TP}{TP + FN} \tag{1.7}$$

$$f - value = \frac{2 * precision * recall}{precision + recall} \tag{1.8}$$

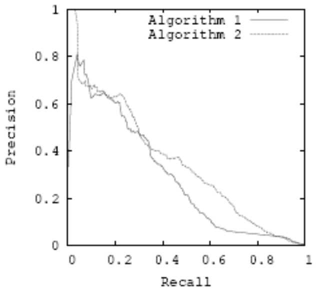

Figure. 1.3. Sample of Area under Precision/Recall

Both g-mean and AUC-ROC can be used if the target is to optimize classification performance with balanced positive class accuracy and negative class accuracy. On the other hand, either f-value or AUC-PR is a good metric if the high detection ability is more preferred.

1.3 Challenges

How to build an effective and efficient model on a huge and complex dataset is a major concern of the science of data mining and machine learning. With emergence of new machine learning application domains such as biomedical informatics, E-business and national security, more challenges are coming.

In many biomedical applications, a biologist or a clinician needs to decide whether a sample (maybe a patient, a tissue, or a tumor) is healthy or not. From the viewpoint of data mining, this problem can be modeled as a binary classification problem. If a sample is healthy, it is classified to be a negative case, and the class label is -1; otherwise it is positive and the class label is +1. For such a binary classification problem, the "effectiveness" of a DSS means that it should not only predict unseen samples accurately, but also work in a human-understandable way. Due to this reason, a desirable data analysis tool, a classifier in this context, should not only assign a class label to an unseen sample, but also provide meaningful and understandable information why it decides to assign such a class label.

1.4 Organizations

The rest of this dissertation is organized as follows: Chapter 2 reviews related works. After that, the general idea and framework of FARM-DS is presented in Chapter 3. Chapter 4 conducts empirical studies to apply FARM-DS on real world medical data, while Chapter 5 focuses on mining FARs from microarray expression data. In Chapter 6, a fuzzy-granular based method is designed to identify marker genes from microarray expression data to support further biomedical study. Finally, we conclude this dissertation and direct the future work in Chapter 7.

CHAPTER 2

RELATED WORKS

2.1 Knowledge discovery, data mining, and data warehousing

Knowledge discovery and data mining is generally known as the science of extracting useful information from large and complex datasets or databases. A data warehousing system is targeted at integrating knowledge discovery and data mining techniques into databases for adaptive and intelligent data analysis. One important data mining task is predicting the unknown value of a variable of interest given known values of other variables. There are two important distinct kinds of problems in predictive data mining: classification if the unknown variable is categorical; and regression if the unknown variable is real-valued [52]. For a classification problem, samples of different classes are accumulated, on which a classifier is modeled to predict future samples.

2.2 Association rule mining

Association rule mining is one of the best studied models for data mining. In recent years, the discovery of association rules from databases is an important and highly active research topic in the data mining field. Association rule mining searches for interesting association or correlation relationships among items in a given dataset.

2.2.1 Basic concepts

Agrawal et al [3] proposed the first association rule mining algorithm in 1993 to discover patterns in transactional databases from the retail industry and business. The idea to discover association rules is also named “market basket analysis” because it looks for associations among items that a customer purchases in a retail shop. For example, when a customer buys item *A*, there is 90% probability he or she will also buy item *B*.

With a transaction database ***D*** = $\{T_1, T_2, ..., T_n\}$ where each T_i $(1 \le i \le n)$represents a transaction and a set of items ***I*** = $\{I_1, I_2, ..., I_m\}$ where each I_j $(1 \le j \le m)$ represents one kind of item, each transaction T_i records the items purchased by the corresponding customer, i.e., $T_i \subseteq$ ***I***. An association rule on this database is formatted as $X \Rightarrow Y$, where *X* and *Y* are called itemsets, which are non-empty subsets of ***I***, *X* and *Y* are disjoint. Two metrics are usually used to measure the reliability and accuracy of the mined association rule:

- The support *s* of the rule is the prior probability of X and Y, $s = \sup(X \cup Y) = \frac{|X \cup Y|}{n}$, and
- The confidence *c* of the rule is the conditional probability of Y given X, $c = \frac{\sup(X \cup Y)}{\sup(X)} = \frac{|X \cup Y|}{|X|}$.

Intuitively, *s* can be viewed as the occurrence frequency of *X* in the whole transaction database ***D***, while *c* indicates that when X is true, *Y* is also true with the probability of *c*. Two thresholds, minimum confidence and minimum support, are used by the mining algorithm to find all association rules whose support and confidence are above the corresponding thresholds.

Usually, an association rule mining algorithm consists of two steps:

1) Finding the frequent itemsets which have support above the predetermined minimum support.
2) Deriving all rules, based on each frequent itemset, which have more than predetermined minimum confidence.

2.2.2 The Apriori Algorithm

The Apriori algorithm is proposed in [3] for finding frequent itemsets. It generates the candidate itemsets in one pass through only the itemsets with large support in the previous pass, without considering the transactions in the database.

An itemset with support larger than or equal to the minimum support is called a frequent itemset. The idea of the Apriori algorithm lies in the “downward-closed” property of support, which means if an itemset is a frequent itemset, then each of its subsets is also a frequent itemset. The candidate itemsets having *k* items can be generated by joining frequent itemsets having *k-1* items, and removing all subsets that are not frequent.

The Apriori algorithm starts by finding all frequent 1-itemsets (itemsets with 1 item); then consider 2-itemsets, and so forth. During each iteration only candidates found to be frequent in the previous iteration are used to generate a new candidate set during the next iteration. The algorithm terminates when there are no frequent *k*-itemsets.

Figure 2.2 sketches the idea of the Apriori algorithm with the notation given at Table 2.1.

k-itemset	An itemset having k items
L_k	Set of frequent k-itemset (those with minimum support)
C_k	Set of candidate k-itemset (potentially frequent itemsets)

Table. 2.1. Notation for mining algorithm

```
$L_1$ = { frequent 1-itemsets };
for (k =2; $L_{k-1} \neq \emptyset$; k++ ) do begin
   $C_k$ = apriori-gen ($L_{k-1}$ );  // New candidates
   forall transactions $t \in D$ do begin
      $C_t$  = subset ($C_k$, t);       // Candidates contained in t
      forall condidates $c \in C_t$  do
         c.count ++;
   end
   $L_k$  =  { $c \in C_k$ | c.count $\geq$ minsup }
end
Answer = $\cup$ k $L_k$ ;
```

Figure. 2.1 Apriori algorithm

The apriori-gen function takes as an input parameter L_{k-1} and returns a superset of the set of all frequent k-itemsets. It consists of a join step and a prune step. In the join step, L_{k-1} is joined with itself:

```
insert into $C_k$
select p.item1, p.item2, …, p.itemk-1, q.itemk-1
from $L_{k-1}$  p, $L_{k-1}$  q
where p.item1 = q.item1, …, p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1
```

In the prune step, all itemsets $c\in C_k$ such that some (k-1)-subset of c is not in L_{k-1} are deleted.

The subset function finds all candidate k-itemsets in the transaction database using a hash tree.

To improve the efficiency of the Apriori algorithm, many variations of the Apriori algorithm have been designed including hashing [97], transaction reduction [4, 51, 97], partitioning the data (mining on each partition and then combining the results) [107], and sampling the data (mining on a subset of the data) [117].

2.3 Association rule mining for classification

There are two kinds of data mining problems: descriptive data mining and predictive data mining [54]. Up to now, most of association rule mining algorithms are designed for descriptive data mining problems. That is, they are used to describe interesting relationships among items in a given dataset. Because of their easy interpretability, the mined association rules may also be utilized for predictive data mining including supervised classification problems.

Some research works have been carried out to utilize "crisp" association rules for classification.

In 1997, Lent et al proposed a method, Association Rule Clustering System, or ARCS, to mine association rules based on clustering and then employ the rules for classification [77]. The ARCS, mined association rules of the form $A_{quan1} \wedge A_{quan2} \Rightarrow A_{cat}$, where A_{quan1} and A_{quan2} are tests on quantitative attribute ranges, and A_{cat} assigns a class label for a categorical attribute from the given training data. The clustered association rules generated by ARCS were applied to classification, and their accuracy was compared to C4.5 [105]. ARCS algorithm is found to be slightly more accurate than C4.5.

The classification by aggregating emerging patterns, called CAEP, is proposed by Dong et al [44]. CAEP uses the notion of itemset support to mine emerging patterns (EPs), which are used to construct a classifier. An EP is defined as an itemset whose support increases significantly from one class to another. CAEP has been found to be more accurate than C4.5 and association-based classification on several data sets.

Association based decision tree [120], called ADT, is a different classification algorithm based on association rules, combined with decision tree pruning techniques. All rules with a confidence greater or equal to a given threshold are extracted and more specific rules are pruned. A decision tree is created based on the remaining association rules, on which classical decision tree pruning techniques are applies.

Baralis et al [12] proposed "Live and Let Live" (L^3), for associative classification. In this algorithm, classification is performed in two steps. Initially, rules which have already correctly classified at least one training case, sorted by confidence, are considered. If the case is still unclassified, the remaining rules (unused during the training phase) are considered, again sorted by confidence.

Liu et al proposed a framework, named associative classification, to integrate association rule mining and classification [84]. The integration is done by focusing on mining a special subset of association rules whose consequent parts are restricted to the classification class labels, called "Class Association Rules" (CARs). This algorithm first generates all the association rules and then selects a small set of rules to form the classifiers. When predicting the class label for a coming sample, the best rule is chosen.

Li et al proposed an algorithm "Classification based on Multiple Association Rules" (CMAR), which utilizes multiple class-association rules for accurate and efficient classification [78]. This method extends an efficient mining algorithm, FP-growth [53], constructs a class distribution- associated FP-trees, and predicts the unseen sample within multiple rules, using weighted χ^2.

Liu and Li's approaches generate the complete set of association rules as the first step, and then select a small set of high quality rules for prediction. These two approaches achieve higher accuracy than traditional classification approaches such as C4.5. However, they often generate a very large number of rules in association rule mining, and take efforts to select high quality rules from among them. Yin et al proposed “Classification based on Predictive Association Rules” (CPAR) [126], which combines the advantages of both associative classification and traditional rule-based classification. CPAR adopts a greedy algorithm to generate rules directly from training data, and hence generates and tests more rules than traditional rule-based classifiers to avoid missing important rules, and uses expected accuracy to evaluate each rule and uses the best *k* rules in prediction to avoid overfitting.

Using association rules for classification helps to solve the understandability problem [32, 100] in classification rule mining. Many rules produced by standard classification systems are difficult to understand because these systems use domain independent biases and heuristics to generate a small set of rules to form a classifier. However, these biases may not be in agreement with the knowledge of the human user, result in that many generated rules are meaningless to user, while many understandable and meaningful rules are left undiscovered.

2.4 Soft computing and fuzzy logic

The basic ideas underlying soft computing in its current incarnation have links to many earlier influences, among them Prof. Zadeh’s 1965 paper on fuzzy sets [130]; the 1973 paper on the analysis of complex systems and decision processes [131].

The principal constituents of soft computing (SC) are fuzzy logic (FL), neural network theory (NN) and probabilistic reasoning (PR), with the latter subsuming belief networks, evolutionary computing including DNA computing, chaos theory and parts of learning theory. For more detailed information and latest news on the soft computing, please refer to The Berkeley Initiative in Soft Computing (BISC) program (http://www-bisc.cs.berkeley.edu/).

2.4.1 Fuzzy concept in the data mining domain.

Real world data often comes with impreciseness and uncertainty. Such data needs to be transformed to be well-defined and unambiguous so that it can be handled with a standard relational data model. For example, many extensions to a standard relational model have been proposed [21, 89, and 4] to support quantitative data.

The fuzzy approach clearly represents a robust solution for the transformation. Instead of defining special "null values" or specific relational algebra operators or first order predicates, fuzzy sets and fuzzy databases are used [132, 106].

Knowledge presented by fuzzy sets is not only more human-understandable but also usually more compact and robust. Furthermore, mining association rules based on fuzzy sets can handle quantitative data, not only just providing the necessary support to use uncertain data types with existing algorithms; but also creating smoother transition boundaries between partitions for numerical values [75]. As a result, fuzzy approaches constitute a good solution for both well-defined and imprecise data.

2.4.2 Fuzzy data modeling

The use of fuzzy logic in the relational model provides an effective way to handle quantitative data with imprecise, uncertain or incomplete information. Fuzzy set theory is more and more frequently used in intelligent systems because of its affinity to human reasoning and the simplicity of the concept [34, 62, and 129].

Some early works [106, 21, and 89], have demonstrated the superior performance of fuzzy logic on data mining and data warehousing as an extension to the relational model. In order to fuzzify a relational data model, structural modifications are introduced to represent and manage quantitative data. There are two major approaches: the proximity relation model [21, 89] and a probability distribution based model [1, 89].

2.4.2.1 Fuzzy sets

A fuzzy set F in a universe of discourse U (classical set of objects) is characterized by a membership function:

$$\mu F: U \rightarrow [0,1]$$

where $\mu F(U)$ for each $u \in U$ denotes the membership value of u in the fuzzy set F.

With the membership function, a fuzzy set F is represented as

$F = \{\mu(u1)/u1, \mu(u2)/u2, \ldots, \mu(un)/un\}$

where $ui \in U$, $1 \leq i \leq n$.

To deal with a fuzzy set, classical set theory operations have been extended to deal with fuzzy sets. One example extension is as follows [RM 88]:

$$\mu A \cup B\ (u) = \max(\mu A(u), \mu B(u))$$

$$\mu A \cap B\ (u) = \min(\mu A(u), \mu B(u))$$

$$\mu \bar{A}(u) = 1 - \mu A(u)$$

where A and B are two fuzzy subsets in a universe of discourse U with membership functions μA and μB respectively .

Based on these definitions, most of the properties that hold for classical set operations, such as DeMorgan's Laws, have been shown to also hold for fuzzy sets. The only law of classical set theory that is no longer true is the law of excluded middle, ie., $A \cap \bar{A} \neq \emptyset$ and

$A \cup \bar{A} \neq U$. where $\varnothing$ is the null set for all $u \in U$. Two fuzzy sets are defined to be equal if $A \supseteq B$ and $A \subseteq B$.

The Cartesian product A1xA2x…An (n universes) is defined to be the fuzzy set U1xU2x…Un where μA1xμA2x…μAn(u1…un) = min(μA1(u1), μA2(u2),… μAn(un))

2.4.2.2 Probability distribution and fuzzy sets

Instead of considering μF(u) to be the membership value of u in F, it can also be considered as a measure of the possibility that a variable X has a value u, where X takes values in U.

$$P(X = u) = \mu F(u) \text{ for all } u \in U.$$

A distribution function of the previous probability equation can be defined with classical statistical definitions [106] to provide a very powerful analysis tool.

2.4.3 Data mining and quantitative data

Data mining, or knowledge discovery in databases, is the extraction of hidden relationships among data items. A Boolean Association Rule problem [3] is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. It can be conceptually reduced to find all matching values in different categories belonging to a given database, which appear together with certain frequency. Since the problem of discovering association rules was introduced [3], many algorithms have been proposed to find association rules in large databases with binary attributes. However, the binary association rule restricts the application area to a binary one and real data usually contains quantitative data that cannot be directly treated with classical binary mining algorithms.

2.4.3.1 Transforming quantitative data

In order to deal with quantitative data, the quantitative association rule was proposed as an extension to the boolean association rule [4], where boolean features can be considered a special case of categorical features.

Several partitioning methods based on classical set theory have been proposed to accomplish this task [4] but all of them are susceptible to the effect of sharp boundaries and sensitive loose of intrinsic relational data information.

The discrete interval method divides a feature domain into discrete intervals and measures the importance of an interval based on the frequency of items appeared in the interval. However, there is a potential risk of information loss because of excluding some potential elements near the crisp boundaries. (Fig. 2.2).

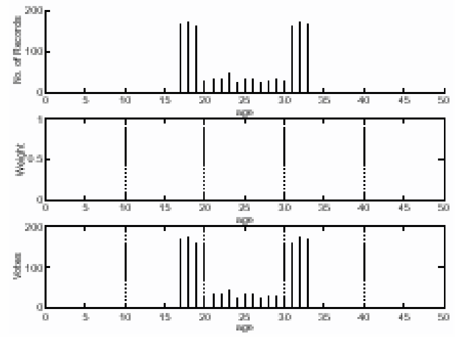

Figure. 2.2 discrete interval method

Another feature partitioning method tries to minimize this effect creating overlapping regions but this causes that the near boundary elements become more important, overemphasizing the important of some intervals (Fig. 2.3).

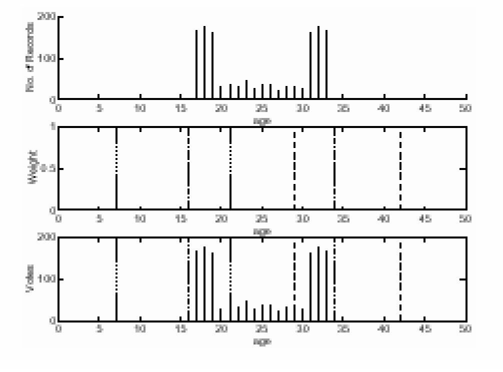

Figure. 2.3 creating overlapping regions

In the fuzzy theory set, an element can belong to a set with a set membership value between 0 and 1 that is assigned by the membership function associated with each fuzzy set. As such, an interval membership is no longer defined by an absolute true/false binary statement but by a probabilistic degree of membership specified by the membership function. As a result, fuzzy sets provide a smooth change between boundaries and the effect is represented by the curve of a traditional fuzzy set (Fig. 2.4)

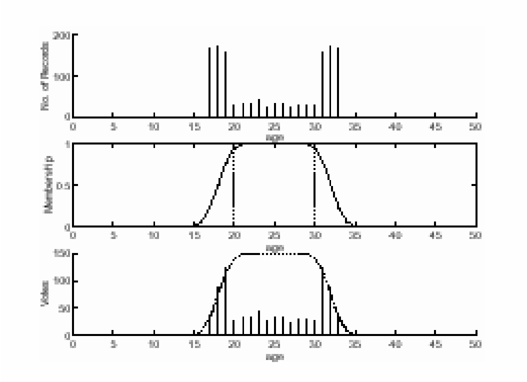

Fig. 2.4 fuzzy partition

With the fuzzy approach, quantitative data can be defined and specified without introducing crisp partition boundaries, side-effects of conventional partitioning algorithms.

2.4.3.2 Fuzzy data mining

Although current quantitative association rule mining algorithms can solve some of the problems introduced by quantitative features, they also introduce some other problems [75]. The use of a crisp partition is also not reasonable with respect to human perception. Fuzzy sets provide an intuitive and understandable solution to handle quantitative data by providing a valid data abstraction to use boolean association rule mining algorithms. Several fuzzy learning algorithms have been successfully applied to specific domains [3, 129, and 34], where strategies based on decision trees [129] can be found in conjunction with space learning [62] and some other classical machine learning algorithms [34].

With fuzzy transformation, most of classical algorithms for mining boolean association rule can be directly used to handle quantitative data [75, 43, 61, and 62], without the need to discover new techniques.

2.4.3.3 Finding Fuzzy Sets

As mentioned in [43], most of the proposed fuzzy mining algorithms relieve the creation of fuzzy sets on quantitative features and defining corresponding membership functions to an end user or an expert. As a result, the performance of these algorithms relies crucially on the appropriateness of the fuzzy sets to a given dataset. Unfortunately, in the real word applications, it is usually difficult to know a priori which fuzzy sets will be the most suitable. Moreover, human experts can not always provide the fuzzy sets of the quantitative features in the database for fuzzy association rule mining.

Some researches have demonstrated that fuzzy sets can be determined automatically from the data using clustering techniques. Some of them are integrated in the fuzzy mining algorithm [118, 61], while others are fully external to completely reuse preexistent algorithms [43]. Once defined the fuzzy sets, the membership functions can be efficiently calculated [43].

2.5 Fuzzy association rule mining

Traditional association rule mining algorithms can only be applied to data mining problems with categorical features. For a data mining problem with quantitative features, it is necessary to transform each quantitative feature into discrete intervals. Many discretization algorithms have been proposed for this purpose. Kamber et al proposed one such algorithm to mine multidimensional association rules using statistically discretization of quantitative features and data cubes based on predefined concept hierarchies [70]. The ARCS [77] algorithm mines quantitative association rules by dynamically discretizing quantitative attributes based on binding, where “adjacent” association rules may be combined by clustering. Techniques for mining quantitative rules based on x-monotone and rectilinear regions were presented by Fukuda et al [44], and Yoda et al. [128]. A non-grid-based technique for mining quantitative association rules, which uses a measure of partial completeness, has been proposed by Srikant and Agrawal [110]. The distance-based association rule mining algorithm [91] can mine distance-based association rules to capture the semantics of interval data, where intervals are defined by clustering. But these approaches have the disadvantage that they involve crisp cutoffs for quantitative features. Fuzzy logic can be introduced into the system to allow “fuzzy” thresholds or boundaries to be defined. Fuzzy logic is demonstrated to be a superior mechanism to enhance interpretability of these discrete intervals.

Many fuzzy association rule mining algorithms have been proposed in recent research works.

[76] uses a membership threshold to transform fuzzy transactions into crisp ones before looking for binary association rules in the set of crisp transactions. This algorithm can diminish the granularity of quantitative features. Chan et al introduced F-APACS to employ linguistic terms for representing the reveal regularities and exceptions for mining fuzzy association rules [23]. The linguistic representation is especially useful when those rules discovered are presented to human experts for examination. In order to avoiding the usage of user-supplied thresholds such as minimum support and minimum confidence, which are often difficult to determine, F-APACS utilizes adjusted difference analysis to identify interesting associations among attributes. Moreover, a confidence measure, called weight of evidence measure, is used to provide a way for representing the uncertainty associated with the fuzzy association rules. In [7, 8 and 24], Au et al also proposed a series of algorithms to employ a set of predefined linguistic labels using adjusted difference and weight of evidence to measure the importance and accuracy of fuzzy association rules. These two measures can avoid the need for a user to provide importance thresholds, but has the drawback of making symmetric the adjusted difference and thus, when a rule $A \Rightarrow C$ is found to be interesting, then $C \Rightarrow A$ will be too.

In [chun1998hongkong], the usefulness of itemsets and rules is measured by means of a significance factor, which is defined as a generalization of support based on sigma-counts (to count the percentage of transactions where the item is) and the product. The accuracy is based on a kind of certainty factor (with different formulation and semantics). In [tp1999], only one item per feature is considered: the pair <feature, label> with greater support among those items based on the same feature. The model is the usual generalization of support and confidence based on sigma-counts. The proposed mining algorithm first transforms each quantitative value into a fuzzy set in linguistic terms. The algorithm then calculates the scalar cardinalities of all linguistic terms in the transaction data. Now the linguistic term with maximal cardinality is used for each feature and thus the number of items keeps. The algorithm therefore focuses on the most important linguistic terms and hence speeds up finding frequent itemsets. The mining process is then performed by using fuzzy counts.

Chien et al [30] proposed an efficient hierarchical clustering algorithm based on variation of density to solve the problem of interval partition. For this purpose, two main characteristics of clustering numerical data: relative inter-connectivity and relative closeness are defined. By giving a proper parameter to determine the importance between relative closeness and relative inter-connectivity, the proposed approach can generate a reasonable interval automatically for data transformation.

Bosc et al [16, 18] introduced another approach to the linguistic summarization of databases. The basic ideas are to use fuzzy partitions on feature domains, which are meaningful for the users, to perform a “soft compression” of the database, and then explore it for evaluating potential summaries. The evaluation is made by computing fuzzy cardinalities which account for the possible variations of the interpretation of the labels.

To cope with the task of diminishing the granularity in quantitative feature representations to obtain useful and natural association rules, some researchers opted for using crisp grid partition or clustering based approaches/ algorithms like Partial Completeness [110], Optimized Association Rules [45] or CLIQUE [2]. Hu et al. [64] have extended the ideas of using crisp grid partition or clustering based approaches to allow non-empty intersections between neighborhood sets in partitions and to describe that by fuzzy sets. They construct an effective algorithm Fuzzy Grid Based Rules Mining Algorithm, called FGBRMA. This algorithm deals with both quantitative and categorical features in a similar manner. The concepts of large fuzzy grid and effective fuzzy association rule are introduced by using special fuzzy support and fuzzy confidence measures. FGBRMA generates large fuzzy grids and fuzzy association rules.

A similar method is developed in [65] for inductive machine learning problems to extract classification rules from a set of examples. They proposed a new fuzzy data mining technique consisting of two phases to find fuzzy if–then rules for classification problems. The first phase is used to find frequent fuzzy grids by using a pre-specified simple fuzzy partition method to divide each quantitative feature, and then the second phase is for generating fuzzy classification rules from frequent fuzzy grids. Another interesting work in [43] finds the fuzzy sets to represent suitable linguistic labels for data by using fuzzy clustering techniques. This way, fuzzy sets can be automatically extracted but may be hard to fit to meaningful labels.

Kaya et al. [73] proposed a clustering method that employs multi-objective Genetic Algorithm for the automatic discovery of membership functions used in determining fuzzy quantitative association rules. This approach optimizes the number of fuzzy sets and their ranges according to multi-objective criteria in a way to maximize the number of large itemsets with respect to a given minimum support value.

Chen et al [27, 28] have considered the case in which there are certain fuzzy taxonomic structures reflecting partial belonging of one item to another in the hierarchy. To deal with these situations, association rules are requested to be of the form $X \Rightarrow Y$ where either X or Y is a collection of fuzzy sets. The model is based on a generalization of support and confidence by means of sigma-counts, and the algorithms are again extensions of the classic Apriori algorithm.

Delgado et al define "fuzzy transactions", which can be applied to quantitative features. They also propose an algorithm to mine "fuzzy association rules" based on these "fuzzy transactions" [35]. The model can be employed in mining distinct types of patterns, from ordinary association rules to fuzzy and approximate functional dependencies and gradual rules.

2.6 Fuzzy association rule mining for classification

In recent years, many research works haven been conducted for fuzzy association rules mining. However, to out best knowledge, there are very few works focusing on fuzzy association rule mining on supervised classification problems. Hu et al proposed to extract "fuzzy associative classification rules" in "fuzzy grids" that are generated by fuzzy partitioning on each input feature [63]. A fuzzy associative rule is defined as a fuzzy if-then rule, whose consequent part is one class label. They divide both quantitative and categorical features into many fuzzy partitions by the concept of the fuzzy grids, resulting from fuzzy partitioning in the feature space, and a linguistic interpretation is easily obtained for each fuzzy partition, since each fuzzy partition is a fuzzy number. After fuzzy partition for each feature, these partitions are viewed as candidates of one-dimension fuzzy grid used to generate large k-dimension fuzzy grids, and then the fuzzy associative classification rules are generated from these large fuzzy grids. In their work, they limit the application of mined fuzzy association rules in the domain of industrial engineering. Moreover, their algorithm faces the "combinatorial rule explosion" problem [37] in that the number of "fuzzy grids" increases exponentially with the dimension of a dataset. Chatterjee et al propose a fuzzy pattern classifier named Influential Rule Search Scheme (IRSS) [26]. This fuzzy classification algorithm is used for automatic construction of the membership functions (MFs) and the fuzzy rule base from an input-output data set. IRSS constructs MFs for each input attribute individually, applying fuzzy C-means (FCM) algorithm. And shapes of all the input MFs are generic in nature and depend entirely on data. This method adaptively modifies the fuzzy rule base, after each epoch, by identifying those rules which are mostly influential in contributing to the system error and subsequently punishing them to improve performance. This coarse adjustment scheme can be followed by another fine adjustment scheme where output MFs are adapted depending on system cumulative error after each epoch. The entire adaptation process stops when system rms error falls below maximum allowable limit. The proposed IRSS, developed as a pattern classifier, has four basic development stages. In stage 1, initial construction of the membership functions for input and output variables from the input-output data set is achieved. In stage 2, initial construction of the fuzzy rule base, from MFs constructed in stage 1 and the input-output data set, us done. Stage 3 contains the defuzzification method to generate crisp output value from fuzzified consequence. Stage 4 contains the proposed approach for tuning of both the fuzzy rule base and the output MFs to achieve desired performance of the IRSS, hence constructed. However, some parameters in IRSS need to be decided by human experts in advance. As a result, it is difficult to be applied to mine FARs on real biomedical datasets due to absent of this kind of prior knowledge.

2.7 Granular computing

Granular computing represents information in the form of some aggregates (called "information granules") such as subsets, classes, and clusters of a universe and then solves the targeted problem in each information granule [11, 80-83, 124-125]. On one hand, for a huge and complicated problem, it embodies Divide-and-Conquer principle to split the original task into a sequence of more manageable and smaller subtasks. On the other hand, for a sequence of similar little tasks, it comprehends the problem at hand without getting buried in all unnecessary details. As opposed to traditional data-oriented numeric computing, granular computing is knowledge-oriented [124]. From the data mining viewpoint, if built reasonably, information granules can make the mining algorithms more effective and at the same time avoid the notorious noise problem.

Many previous works have reported that the frequent patterns occurred in the training dataset of a complex and huge classification problem could lead to measured improvement on testing accuracy [126]. The idea was named "association classification" [126].

For a binary classification problem with continuous features, an association rule is usually formed as:

$$if\ a_1 \in [v_{11}, v_{12}]\ and\ \ a_2 \in [v_{21}, v_{22}]\ and \ldots a_n \in [v_{n1}, v_{n2}],\ then\ \ \mathrm{y} = 1\,(\mathrm{or}\ \text{-}1) \quad (2.1)$$

The support and confidence of an association rule for a binary classification problem are defined in Equations2.2-2.3:

$$SUP(AR) = S_{PG} / S_W \qquad (2.2)$$

$$COF(AR) = S_{PG} / S_G \qquad (2.3)$$

where S_W is the size of training data with the same class label as the THEN-part of the association rule, S_G is the size of training data that satisfy the IF-part, while S_{PG} is the size of training data correctly classified by the association rule. Notice that S_W is defined in such a way that the support and confidence of an association rule are calculated based on a single class. As a result, the association rule mining will not be biased for major class in an unbalanced binary classification problem.

From Eq. 2.1, an association rule (or a set of association rules combined disjunctively) could be used to partition the feature space to find an information granule. So association rules mining is a possible solution for granulation. The realization of a successful "association granulation" depends on the following two issues:

An association rule with high enough confidence could deduce a "pure" granule, in which it is unnecessary to build a classifier because of its high purity. If its support is also high, it could significantly simplify and speed up classification because it decreases the size of the training dataset.

A more general association rule with a shorter IF-part should be more possible to avoid overfitting training dataset. A short IF-part means a low model complication, which in turn means a good generalization capability.

2.8 Clustering and data abstraction

2.8.1 Clustering

2.8.1.1 Basic concepts

Clustering is a division of data into groups of similar objects. Each group, called a cluster, is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Representing data by fewer clusters loses certain details, but achieves simplification.

Clustering analysis has wide applications including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, Web document classification, and many others.

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution for descriptive data mining, or serve as a preprocessing step for predictive data mining algorithms operating on the detected clusters.

There are a large number of clustering algorithms in the literature. In general, most of clustering methods can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, or model-based methods. Among them partitioning and hierarchical methods are most popular ones. A partitioning method first creates an initial set of k partitions, where k is the number of partitions to construct; then it iteratively moves objects from one group to another to improve the partitioning. Typical partitioning methods include k-means [87], k-medoids [71], CLARANS [95, 42], and so on. A hierarchical method creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. The quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (such as in Cure [47] and Chameleon [72]) or integrating other clustering techniques, such as iterative relocation (as in BIRCH [133]).

Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function. The out-put of such algorithms is a clustering, but not a partition. In fuzzy clustering, each cluster is a fuzzy set of all the patterns. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be obtained from a fuzzy partition by thresholding the membership value. The most popular fuzzy clustering algorithm is the fuzzy c-means (FCM [40, 14]) algorithm.

2.8.1.2 Representation of clusters

In applications where the number of classes or clusters in a data set must be discovered, a partition of the data set is the end product. Here, a partition gives an idea about the separability of the data points into clusters and whether it is meaningful to employ a supervised classifier that assumes a given number of classes in the data set. However, in many other applications that involve decision making, the resulting clusters have to be represented or described in a compact form to achieve data abstraction. Even though the construction of a cluster representation is an important step in decision making, it has not been examined closely by researchers. The notion of cluster representation was introduced and was subsequently studied. The followings are three popular representation schemes:

1. Represent a cluster of points by their centroid or by a set of distant points in the cluster.
2. Represent clusters using nodes in a classification tree.
3. Represent clusters by using conjunctive logical expressions

Use of the centroid to represent a cluster is the most popular scheme. It works well when the clusters are compact or isotropic. However, when the clusters are elongated or non-isotropic, then this scheme fails to represent them properly. In such a case, the use of a collection of boundary points in a cluster captures its shape well. The number of points used to represent a cluster should increase as the complexity of its shape increases. Every path in a classification tree from the root node to a leaf node corresponds to a conjunctive statement. An important limitation of the typical use of the simple conjunctive concept representations is that they can describe only rectangular or isotropic clusters in the feature space.

2.8.2 Data abstraction

Data abstraction is useful in decision making because of the following reasons:

- It gives a simple and intuitive description of clusters which is easy for human comprehension. In both conceptual clustering and symbolic clustering this representation is obtained without using an additional step. These algorithms generate the clusters as well as their descriptions. A set of fuzzy rules can be obtained from fuzzy clusters of a data set. These rules can be used to build fuzzy classifiers and fuzzy controllers. It helps in achieving data compression that can be exploited further by a computer. A partition clustering like the k-means algorithm cannot separate these two structures properly. The single-link algorithm works well on this data, but is computationally expensive. So a hybrid approach may be used to exploit the desirable properties of both these algorithms. We obtain 8 subclusters of the data using the (computationally efficient) k-means algorithm.

- It increases the efficiency of the decision making task. In a cluster-based document retrieval technique, a large collection of documents is clustered and each of the clusters is represented using its centroid. In order to retrieve documents relevant to a query, the query is matched with the cluster centroids rather than with all the documents. This helps in retrieving relevant documents efficiently. Also in several applications involving large data sets, clustering is used to per-form indexing, which helps in efficient decision making.

Chapter 3 Fuzzy Association Rule Mining for Decision Support

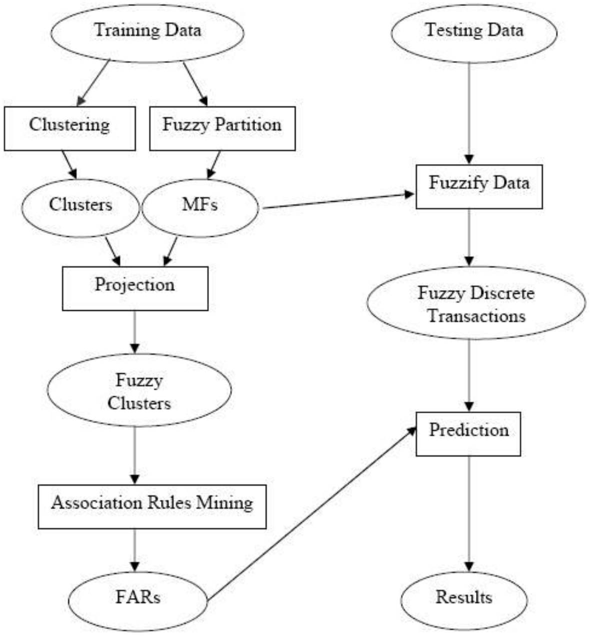

Figure. 3.1. a sketch of FARM-DS

The new FARM-DS algorithm consists of two phases: the training phase and the testing phase. In the training phase, four steps are executed to mine fuzzy association rules. At step 1, a 1-in-1-out ANFIS system is used to generate fuzzy internals on each input feature. Each fuzzy interval is defined with a fuzzy membership function. At step 2, clustering is conducted for data abstraction to extract inherent data distribution knowledge. At step 3, FARM-DS naturally transforms quantitative samples into “fuzzy discrete transactions” by projecting the center of each cluster extracted at step 2 on the fuzzy intervals generated at step 1. Finally, at step 4, simple “IF-THEN” Fuzzy Association Rules can be mined from the “fuzzy discrete transactions” by the traditional Apriori association rule mining algorithm. These FARs are thereafter used to predict unseen samples in the testing phase.

Fig. 3.1 shows a sketch of the FARM-DS algorithm. In the following, we assume that the classification problem at hand has n samples and m input features. Notice that step 1 and step 2 can be executed independently in parallel.

3.1 Step 1: Fuzzy Interval Partitioning

Step 1 builds a 1-in-1-out 0-order TSK fuzzy model [112, 114] for each feature:

$$\begin{aligned}
&\mathit{If}\ f_i\ \mathit{is}\ M_{i1}, \mathit{then}\ Y = \{-1, 1\},\\
&\mathit{If}\ f_i\ \mathit{is}\ M_{i2}, \mathit{then}\ Y = \{-1, 1\},\\
&\ \ \vdots \qquad\quad \vdots \qquad\quad \vdots\\
&\mathit{If}\ f_i\ \mathit{is}\ M_{ij}, \mathit{then}\ Y = \{-1, 1\}.
\end{aligned} \tag{3.1}$$

Here, $j \geq 2$ linguistic terms (M_{i1}, M_{i2},… M_{ij}) are defined for the ith input feature f_i, and the shape of the fuzzy membership function for each linguistic term will be selected in a data-dependant way from the following functions.

Triangular membership function specified by three parameters $\{a, b, c\}$ as follows:

$$\mathit{triangle}(x;a,b,c) = \begin{cases} 0, & x \leq a. \\ \dfrac{x-a}{b-a}, & a \leq x \leq b. \\ \dfrac{c-x}{c-b}, & b \leq x \leq c. \\ 0, & c \leq x. \end{cases}, \tag{3.2}$$

where $\{a, b, c\}$ determine the x coordinates of the three corners of the underlying triangular MF.

Trapezoidal membership function specified by four parameters $\{a, b, c, d\}$ as follows:

$$trapezoid(x;a,b,c,d) = \begin{cases} 0, & x \le a. \\ \frac{x-a}{b-a}, & a \le x \le b. \\ 1, & b \le x \le c. \\ \frac{d-x}{d-c}, & c \le x \le d. \\ 0, & d \le x. \end{cases}, \tag{3.3}$$

where {*a*, *b*, *c*, *d*} determine the *x* coordinates of the four corners of the underlying triangular MF.

Gaussian membership function specified by two parameters {*c*, *σ*} as follows:

$$gaussian(x;c,\sigma) = e^{-\frac{1}{2}(\frac{x-c}{\sigma})^2}, \tag{3.4}$$

where *c* represents the center and *σ* determines the width of the underlying Gaussian MF.

Generalized bell membership function specified by three parameters {*a*, *b*, *c*} as follows:

$$bell(x;a,b,c) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}}, \tag{3.5}$$

where *b* is usually positive.

Sigmoidal membership function specified by two parameters {*a*, *c*} as follows:

$$sig(x;a,c) = \frac{1}{1+\exp[-a(x-c)]}, \tag{3.6}$$

where *a* controls the slope at the crossover point *x* = *c*.

Left-Right membership function specified by three parameters {*α*, *β*, *c*} as follows:

$$LR(x;c,\alpha,\beta) = \begin{cases} F_L(\frac{c-x}{\alpha}), & x \le c. \\ F_R(\frac{x-c}{\beta}), & x \ge c. \end{cases}, \tag{3.7}$$

where $F_L(x)$ and $F_R(x)$ are monotonically decreasing functions defined on $[0,\infty)$ with $F_L(0) = F_R(0) = 1$ and $lim_{x\to\infty} F_L(x)= lim_{x\to\infty} F_R(x)=0$.

In Eq. 3.1, $Y = -1$ means a negative sample, and $Y = 1$ means a positive sample.

In its simplest form, only two linguistic terms (“low” and “high”) are defined for the *i*th input feature f_i, and the default membership function is a trapezoidal membership function (Eq. 3.3).

$$\begin{aligned} &If\ \ f_i\ is\ low,\ then\ Y = -1, \\ &If\ \ f_i\ is\ high,\ then\ Y = 1. \end{aligned} \tag{3.8}$$

Furthermore, parameters (defining MFs) in the 1-in-1-out TSK model are optimized by an ANFIS system to maximize the classification accuracy on the training dataset. The goal of this step is to achieve an approximate but suitable fuzzy partition for each feature efficiently (because here we consider each feature separately) and effectively (because we optimize the partition with a simple 1-in-1out ANFIS system).

Recently, cancer classification on microarray expression data is a hot bioinformatics research topic. A typical gene expression dataset is extremely high dimensional. The data usually comes with only dozens of samples but with thousands or even tens of thousands of gene features. As a result, the ability to extract a subset of informative genes while removing irrelevant or redundant genes is crucial for accurate classification. Furthermore, it is also helpful for biologists to find the inherent cancer-resulting mechanism and thus to develop better diagnostic methods or find better therapeutic treatments. From the data mining viewpoint, this gene selection problem is essentially a feature selection or dimensionality reduction problem. A good dimensionality reduction method should remove irrelevant or redundant features while keep informative or important features for classification. A classifier modeled in the resulted lower-dimensioned feature space is expected to capture the inherent data distribution better and thus has a better performance.

One more potential benefit of single dimension fuzzy partition described above is that features can be ranked according to classification accuracy of corresponding TSK models. For a high-dimensional classification problem such as cancer classification on microarray gene expression data, this feature ranking process may be useful for dimension reduction to make the following steps more efficient. This is an interesting future work.

3.2 Step 2: Data Abstracting

Step 2 groups training samples into several clusters by the K-means clustering algorithm.

1. Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set.
2. Assign each pattern to the closest cluster center.
3. Recompute the cluster centers using the current cluster memberships.
4. If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or inimal) reas-signment of patterns to new cluster centers, or minimal decrease in squared error.

Several variants of the k-means algorithm have been reported in the literature. Some of them attempt to select a good initial partition so that the algorithm is more likely to find the global minimum value.

Another variation is to permit splitting and merging of the resulting clusters. Typically, a cluster is split when its variance is above a prespecified threshold, and two clusters are merged when the distance between their centroids is below another pre-specified threshold. Using this variant, it is possible to obtain the optimal partition starting from any arbitrary initial partition, provided proper threshold values are specified. The well-known ISO-DATA algorithm employs this technique of merging and splitting clusters.

Another variation of the k-means algorithm involves selecting a different criterion function altogether. The dynamic clustering algorithm (which permits representations other than the centroid for each cluster) was proposed and describes a dynamic clustering approach obtained by formulating the clustering problem in the framework of maximum-likelihood estimation. The regularized Mahalanobis distance was used in Mao and to obtain hyperellipsoidal clusters.

K-means clustering can be viewed as a data abstraction method. That is, K-means partitions the samples into K mutually exclusive clusters, and returns a vector of indices indicating to which of the k clusters it has assigned each observation. Notice that K-means creates a single level of clusters. K-means is more suitable for clustering large amounts of data. It treats each sample as an object having a location in the feature space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible.

There are many different distance measurements.

- Squared Euclidean distance. Each centroid is the mean of the points in that cluster.
- Sum of absolute differences, i.e., L1. Each centroid is the component-wise median of the points in that cluster.

- One minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.
- One minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.
- Percentage of bits that differ (only suitable for binary data). Each centroid is the component-wise median of points in that cluster.

Which distance measurement is best depends on the kind of data being clustered. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. K-means computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure. K-means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. The details of the minimization can be controlled by using several optional input parameters to K-means, including ones for the initial values of the cluster centroids, and for the maximum number of iterations.

To decide the optimal/suboptimal number of clusters *K*, the whole FARM-DS algorithm runs several times with different *K* values. The *K* value with the largest training or cross-validation accuracy is selected as the optimal number of clusters. After *K* is fixed, the clustering with the largest overall silhouette value is selected to be the best clustering.

The silhouette value for a sample is a measure of how similar the sample is to samples in its own cluster compared with samples in other clusters, and ranges from -1 to +1. It is defined as

$$S(i) = \frac{\min(b(i,k)) - a(i)}{\max(a(i), \min(b(i,k)))} \quad , \qquad (3.9)$$

where *a(i)* is the average distance from the *i*th sample to other samples in its own cluster, and *b(i,k)* is the average distance from the *i*th sample to samples in another cluster *k*.

The larger silhouette values over all training samples mean that samples in the same cluster are more similar while samples between different clusters are more different, which in turns means a better clustering result.

We also tried fuzzy clustering algorithms for data abstraction.

1. Select an initial fuzzy partition of the N objects into K clusters by selecting the N 3 K membership matrix U. An element uij of this matrix represents the grade of membership of object xi in cluster cj.
2. Using U, find the value of a fuzzy criterion function, e.g., a weighted squared error criterion function, associated with the corresponding partition.
3. Repeat step 2 until entries in U do not change significantly.

In fuzzy clustering, each cluster is a fuzzy set of all the patterns. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be obtained from a fuzzy partition by thresholding the membership value. The most popular fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm. Even though it is better than the hard k-means algorithm at avoiding local minima, FCM can still converge to local minima of the squared error criterion. The design of membership functions is the most important problem in fuzzy clustering; different choices include those based on similarity decomposition and centroids of clusters. A generalization of the FCM algorithm was proposed through a family of objective functions. A fuzzy c-shell algorithm and an adaptive variant for detecting circular and elliptical boundaries was presented.

In FARM-DS, fuzzy C-means algorithm (FCM) is used to group samples into *K* clusters with centers $c_1,\cdots c_k,\cdots c_K$ in the feature space. FCM assigns a real-valued vector $U_i=\{\mu_{1i},\cdots \mu_{ki},\cdots,\mu_{Ki}\}$ to each sample. $\mu_{ki}\in[0,1]$ is the membership value of the *i*th gene in the *k*th cluster. The larger membership value indicates the stronger association of the sample to the cluster. Membership vector values μ_{ki} and cluster centers c_k can be obtained by minimizing

$$J(K,m)=\sum_{k=1}^{K}\sum_{i=1}^{N}(\mu_{ki})^m d^2(x_i,c_k)\,, \tag{3.10}$$

$$d^2(x_i,c_k)=(x_i-c_k)^T A_k(x_i-c_k)\,, \tag{3.11}$$

$$\sum_{k=1}^{K}\mu_{ki}=1\,,0<\sum_{i=1}^{N}\mu_{ki}<N\,, \tag{3.12}$$

where $1\le i\le N$ and $1\le k\le K$ [40, 14].

In Eq. 3.10, *K* and *N* are the number of clusters and the number of samples in the dataset, respectively. *m*>1 is a real-valued number which controls the ‘fuzziness’ of the resulting clusters, μ_{ki} is the degree of membership of the *i*th sample in the *k*th cluster, and $d^2(x_i,c_k)$ is the square of distance from *i*th sample to the center of the *k*th cluster. In Eq. 3.11, A_k is a symmetric and positive definite matrix. If A_k is the identity matrix,

$d^2(x_i, c_k)$ corresponds to the square of the Euclidian distance. Eq. 3.12 indicates that empty clusters are not allowed.

Notice in each step, the fuzzy membership values are defuzzified in such a way that a sample is always grouped into the cluster with the largest membership value and the cluster with the second largest membership value.

In the near future, other clustering algorithms such as Self-Organizing Maps or hierarchical clustering will be also tried for data abstraction.

3.3 Step 3: Generating Fuzzy Discrete Transactions

By grouping similar samples together in several clusters at step 2, a high-level data abstraction can be achieved. This way, the number of transactions and following rules is independent with the dimension of the input feature space. It is only decided by the number of clusters to generate a compact rule base, which in turn enhances the generalization capability and the interpretability to predict unknown new samples.

Step 3 transforms quantitative training samples into “fuzzy discrete transactions”. Firstly, the TSK models generated at step 1 are used to fuzzify the center of each cluster generated at step 2.

Currently, only two MFs for each feature at step 1 are considered. On each input feature f_i, two membership values μ_{low} and μ_{high} are calculated for a center by projecting the center on the feature. Fig. 3.2 shows an example of projecting a center with f_i =0.113 on the trapezoidal membership functions.

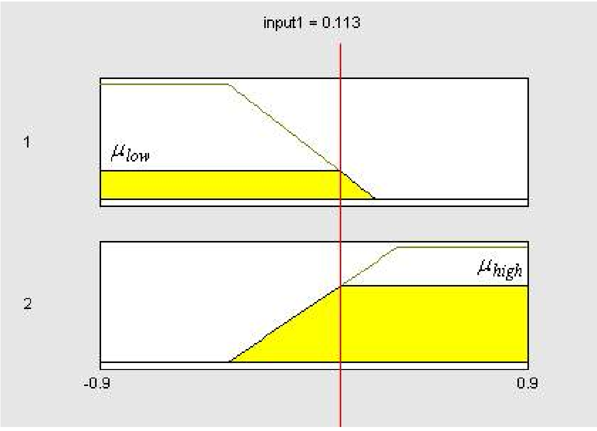

Figure. 3.2. an example to project a sample onto a feature

After that, for a cluster *k* with *sk*+ positive samples and *sk*- negative samples, | *sk*+ - *sk*-| same “fuzzy discrete transactions” are generated as follows:

If $s_{k+} \geq s_{k-}$, +1 is inserted into the transactions;

Else -1 is inserted into the transactions.

For each f_i,

$$\begin{array}{l} \text{if } \left|\mu_{high} - \mu_{low}\right| < \alpha, \\ \text{then } f_i \text{ is not inserted into the transactions.} \\ \text{if } \mu_{high} - \mu_{low} \geq \alpha, \\ \text{then } f_i \text{ is inserted into the transactions with the form of "i1".} \\ \text{if } \mu_{low} - \mu_{high} \geq \alpha, \\ \text{then } f_i \text{ is inserted into the transactions with the form of "i0".} \end{array}, \qquad (3.13)$$

Here $\alpha \in [0,1]$ is a threshold used to prune the resulted “fuzzy discrete transactions”. That is, if the difference between the “low” membership function value and the “high” membership function value of a feature is too small (less than α), this feature is treated as an unavailable feature on the resulted transactions. The pruning process improves the generalization capability of the clusters.

This projection method can also be extended to more than two MFs for some features at step 1.

3.4 Step 4: Mining Association Rules

The final step is mining association rules from the fuzzy discrete transactions generated at step 3 by the Apriori algorithm. It follows a rule-pruning process to eliminate the redundant and useless rules:

For a pair of rules A and B, if B is more specific than A (that means A is included by B), and B has the same support value as A, A is eliminated. A mined fuzzy association rule has the following format:

$$\textit{if } f_1 \textit{ is low}, f_2 \textit{ is high}, \cdots, f_h \textit{ is high}, \textit{then } y = \{-1, +1\}, \qquad (3.14)$$

where $0 < h \le m$. The rule is called a positive rule if y=1 or called a negative rule if y=-1. The length of a rule is defined to be the number of items in the antecedent part of this rule.

3.5 Testing Phase

In the testing phase, the performance of mined fuzzy association rules is evaluated on the testing dataset. Assume that there are *r*+ positive rules and *r*- negative rules. For each new sample, its positive weight *weight*+ and negative weight *weight*- are decided as follows:

$$weight_{+} = \sum_{i=1}^{r_{+}} strength_{i+}, \qquad (3.15)$$

$$strength_{i+} = \prod_{j=1}^{m} \max(0, \mu_{ij+} - \mu_{ij-}) \; . \qquad (3.16)$$

The *i*th positive rule is said to be fired if $strength_{i+} > 0$.

$$weight_{-} = \sum_{i=1}^{r_{-}} strength_{i-} , \tag{3.17}$$

$$strength_{i-} = \prod_{j=1}^{m} \max(0, \mu_{ij-} - \mu_{ij+}) \ . \tag{3.18}$$

The *i*th negative rule is said to be fired if $strength_{i-} > 0$.

Finally, a class label is calculated by the following equation:

$$y = sign(weight_{+} - weight_{-} + b) , \tag{3.19}$$

where $b \in R$ is a bias constant, which can be optimized by cross validation.

3.6 Parameter Selection

In the above process for FARM-DS modeling, many parameters need to be decided. At step 1, we need to decide the number of MFs for each feature; at step 2, the number of clusters need to be decided for data abstraction; at step 3, the threshold *α* need to be decided whether a feature should be inserted into a fuzzy discrete transaction; at step 4, bias *b* for final prediction also need to be decided. In general, some parameters can be decided based on prior knowledge for a specific problem, or at least limited into a field. On the other hand, cross-validation and bootstrapping are two common heuristics for parameter selection with the available training dataset.

For cross-validation, the dataset is randomly split into *k* equal-sized subsets. *k-1* subsets are combined as the dataset for modeling and another one is taken as the dataset for validation. The process is repeated *k* times such that each subset is used for validation once.

Another evaluation heuristic adopted is balanced .632 bootstrapping [20]: random sampling with replacement is repeated for *m* times (usually *m*=100 to 1000) on the training dataset. Each sample appears exactly *m* times in the computation to reduce variance [22]. Each time, on average 63.2% samples will appear for training and other samples for validation. The bootstrapping accuracy is defined to be the average accuracy on *m* times bootstrapping. The bootstrapping accuracy tends to be high-biased. The 0.632 bootstrapping accuracy

$$acc_{.632} = (1 - 0.632)acc_{training} + 0.632acc_{testing}\,, \tag{3.20}$$

tries to correct this bias via a weighted average of the training accuracy and the bootstrapping accuracy.

Chapter 4 FARM-DS from medical data

4.1 Experiments Design

The hardware we used is a desktop with P4-2.8MHz CPU and 256M memory. The software we developed is based on Matlab Fuzzy Logic Toolbox and Statistics Toolbox. The program of the Apriori association rule mining algorithm comes from http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html.

FARM-DS is compared with well-known SVMs and C4.5 classification algorithms. We run FARM-DS and SVMs in the experiments. We also compare our works with Bennett et al's works [13] on SVMs and C4.5 because the same experimental setup. For discrimination, the SVM built by us is called SVM1, and the SVM built in [13] is called SVM2. The Wisconsin breast cancer dataset and the Cleveland heart-disease dataset from UCI data mining repository [90] are used in the experiments. Table 41 lists the detailed characteristics of datasets.

TABLE 4.1
CHARACTERISTICS OF DATASETS USED FOR EXPERIMENTS

Dataset	Size	Attr	Ratio
Wisconsin Breast Cancer	683	9	239:444
Cleveland heart-disease	297	13	160:137

Note 1: Size = # of cases after removing cases with missing data, Attr = # of input features, Ratio = # of positive cases: # of negative cases.
Note 2: 16 cases in Wisconsin Breast Cancer and 6 cases in Cleveland heart-disease with missing values are removed.

5-fold cross validation is used for comparison. A dataset is randomly split into five equal-sized subsets, four of which are combined as the training dataset and another one is taken as the testing dataset. The training-testing process is repeated five times such that each subset is used as the testing dataset once. The input features are scaled and normalized to [-0.9, 0.9]. Note that the normalization process is based on the training dataset to avoid overfitting. For each fold:

$S(training):S(testing)=4:1$

$S(positive\_training):S(positive\_testing)=4:1$

$S(negative\_training):S(negative\_testing)=4:1$

$S(x)$ means the size of the dataset x.

According to [116], both SVMs with the linear kernel and the RBF kernel are used in our experiments. The best kernel and the parameters are optimized with grid search heuristic. For the linear kernel, the regulation parameter C is selected from

$$\begin{aligned} C \in \{&2^{-10},2^{-9.5},2^{-9},2^{-8.5},2^{-8},2^{-7.5},2^{-7},2^{-6.5},\\ &2^{-6},2^{-5.5},2^{-5},2^{-4.5},2^{-4},2^{-3.5},2^{-3},2^{-2.5},\\ &2^{-2},2^{-1.5},2^{-1},2^{-0.5},2^{0},2^{0.5},2^{1},2^{1.5},2^{2}\}.\end{aligned}$$

For the RBF kernel, the parameters γ, C are selected from

$\gamma \in \{2^{-16},2^{-14},2^{-12},2^{-10},2^{-8},2^{-6},2^{-4},2^{-2},2^{0},2^{2},2^{4}\}$,

$C \in \{2^{-6},2^{-4},2^{-2},2^{0},2^{2},2^{4},2^{6},2^{8},2^{10}\}$.

For FARM-DS, at step 1, trapezoidal membership functions are adopted for modeling a 1-in-1-out-0-order ANFIS system for each feature; the number of linguistic terms for each feature is fixed to be 2. At step 2, after some preliminary experiments, the optimal number of clusters is selected to be 11 for the Wisconsin breast cancer dataset and 21 for the Cleveland heart-disease dataset. For each fold, the clustering process is repeated 50 times and the one with the largest silhouette value is selected. The fuzzy discrete transactions pruning parameter $\alpha = 0.7$ is used at step 3. For mining association rules from the “fuzzy discrete transactions”, minimal support=0.1, and minimal confidence=0.8. At step 4, bias is 0 for the Wisconsin breast cancer dataset and is 3 for the Cleveland heart-disease dataset.

4.2 Results Analysis on Effectiveness

Firstly a clustering result on the Wisconsin breast cancer dataset is shown in Fig. 4.1. The clustering result is optimal in that it achieves the largest overall silhouette values 0.3899 [71]. From Fig. 4.1, we can see that except clusters 1, 7, and 11, the other clusters have good qualities.

Tables 4.2 reports the FARM-DS modeling results. For each fold, the largest overall silhouette value, and the numbers of mined positive rules and negative rules are reported. The validation accuracy is reported in Tables 4.3. Bennett et al also adopt 5-fold cross validation to evaluate the performance of C4.5 and SVM [13] on these two datasets. As a result, our simulation results can directly be compared with them. The experimental results demonstrate that FARM-DS with trapezoidal membership functions is competitive with the optimal SVM and better than C4.5 to achieve high prediction accuracy.

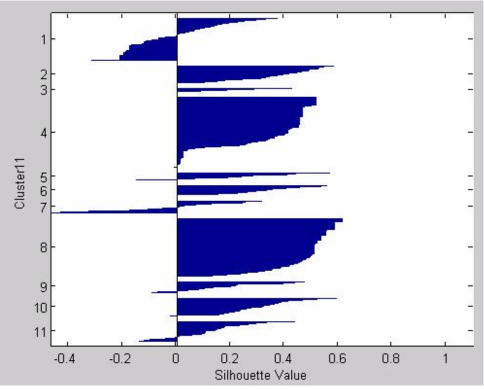

Figure.4.1. an example to decide the optimal

TABLE 4.2
FARM-DS MODELING RESULTS WITH TRAPEZOIDAL-SHAPED MEMBERSHIP FUNCTIONS BY 5-FOLD CROSS VALIDATION

Fold	Max sil value	# pos rules	# neg rules
Wisconsin breast cancer dataset			
1	0.2756	22	7
2	0.3829	26	7
3	0.2606	21	7
4	0.2993	20	9
5	0.3899	19	10
Cleveland heart-disease dataset			
1	0.4859	43	86
2	0.5002	67	75
3	0.4913	75	93
4	0.4823	63	98
5	0.4927	59	80

TABLE 4.3
VALIDATION ERROR COMPARISON BY 5-FOLD CROSS VALIDATION

Fold	FARM-DS	SVM1	SVM2 [27]	C4.5 [27]
Wisconsin breast cancer dataset				
1	97.81%	98.54%	**N/A**	**N/A**
2	97.81%	96.35%	**N/A**	**N/A**
3	97.81%	97.81%	**N/A**	**N/A**
4	97.08%	95.62%	**N/A**	**N/A**
5	95.56%	95.56%	**N/A**	**N/A**
Overall	**97.2%**	**96.8%**	**97.2%**	**93.4%**
Cleveland heart-disease dataset				
1	80%	81.67%	**N/A**	**N/A**
2	78.33%	83.33%	**N/A**	**N/A**
3	79.66%	81.36%	**N/A**	**N/A**
4	86.44%	83.05%	**N/A**	**N/A**
5	89.83%	88.14%	**N/A**	**N/A**
Overall	**82.8%**	**83.5%**	**81.5%**	**77.8%**

4.3 Result Analysis on Efficiency

Table 4.4 compares the running time of FARM-DS and that of SVM. The comparison shows that FARM-DS can finish in a reasonable period, although it is slower than SVM. Notice that the running time of FARM-DS is calculated under the assumption that the optimal number of clusters is known in advance. In the future, we plan to implement the parallel version of FARM-DS so that the same or similar efficiency can be achieved if the optimal number of clusters is unknown.

TABLE 4.4
RUNNING TIME COMPARISON WITH 5-FOLD CROSS VALIDATION

Dataset	FARM-DS	SVM1
Wisconsin	**46 seconds**	**45 seconds**
Cleveland	**61 seconds**	**27 seconds**

4.4 Result Analysis on Interpretability

As we know, a SVM only assigns a class label for a sample so that the classification exhibits little understandability, i.e., a diagnostic decision is essentially a black box, with no explanation on how it is reached. On the other hand, a decision tree built by C4.5 may be explained. Unfortunately, the classification accuracy of C4.5 is low on these two datasets.

In contrast, FARM-DS achieves high accuracy and also can return fired positive rules and fired negative rules for further analysis.

Due to relatively higher accuracy on the Wisconsin breast cancer dataset, we take it as the example to analyze the interpretability of mined FARS.

TABLE 4.5
THE FEATURE INFORMATION OF
THE WISCONSIN BREAST CANCER DATA SET

Feature	Medical meaning	Domain
1	clump thickness (the extent to which epithelial cell aggregates are mono or multilayered)	1 – 10
2	uniformity of cell size	1 – 10
3	uniformity of cell shape	1 – 10
4	marginal adhesion (cohesion of peripheral cells)	1 – 10
5	single epithelial cell size	1 – 10
6	number of bare nuclei	1 – 10
7	extent of bland chromatin	1 – 10
8	number of normal nucleoli	1 – 10
9	frequency of mitosis	1 – 10

For the Wisconsin breast cancer dataset, Table 4.5 describes the nine cellular features taken from fine needle aspirates (a fine needle aspiration is an outpatient procedure that involves using a small-gauge needle to extract fluid directly from a breast mass [30]) from human breast tissues. These nine features are believed to be useful to distinguish benign tumors from malignant ones.

Each of the nine features of the fine needle aspirates is graded one to ten at the time of sample collection so that a larger number signals a higher probability of malignancy. Thus, for the purposes of diagnosis, each tumor sample is represented as a 9-dimensional integer vector. Given such a 9-dimensional feature vector of an undiagnosed tumor, the problem is to determine whether the tumor is benign or malignant.

Extracted FARs enhance the interpretability of classification due to the following three benefits:

Firstly, FARs may help human experts to correct the wrongly classified samples. For example, 12 from 19 wrongly classified samples in the Wisconsin breast cancer dataset activate some correct rules. Table 4.6 lists the 12 samples. By analyzing these samples and corresponding rules, we can expect that the accuracy can be further improved. Consequently, more reliable decisions can be made.

TABLE 4.6
12 WRONGLY CLASSIFIED SAMPLES ON WISCONSIN BREAST CANCER DATASET

id	Real class	Predictive class	Positive weights	Negative weights
1	-1	+1	2.0000	0.9660
2	-1	+1	2.0000	0.9290
3	+1	-1	0.7300	0.9290
4	+1	-1	0.7085	0.9290
5	-1	+1	8.6263	0.1535
6	-1	+1	2.0000	0.9970
7	+1	-1	0.3756	0.9970
8	-1	+1	4.6971	0.9470
9	-1	+1	3.6049	0.9470
10	-1	+1	4.4657	0.8900
11	-1	+1	2.2007	0.1556
12	-1	+1	1.0000	0.8630

TABLE 4.7
THE MOST GENERAL AND THE MOST SPECIFIC FIRED RULES FOR THE 1ST SAMPLE IN FOLD 1 ON WISCONSIN BREAST CANCER DATASET

If bare nuclei (*f6*) is high, Then **y=1 (malignant)**. support=26.9%, confidence=100%, (most general)
If bare nuclei (*f6*) is high, mitosis (*f9*) is low, Then **y=1 (malignant)**. support=22.9%, confidence=100%, (most specific)
If normal nucleoli (*f8*) is low, Then **y=-1 (benign)**. support=77.6%, confidence=85.1%, (most general)
If normail nucleoli (*f8*) is low, marginal adhesion (*f4*) is low, single epithelial cell size (*f5*) is low, Then **y=-1 (benign)**. support=68.4%, confidence=96.6%, (most specific)
If normal nucleoli (*f8*) is low, marginal adhension (*f4*) is low, mitosis (*f9*) is low, Then **y=-1 (benign)**. support=71.4%, confidence=92.6%, (most specific)

For example, the first validation sample in fold 1 is classified to be positive but it is actually negative. (That is, it is false positive). Its positive weight weight+=2.0000, and its negative weight weight-=0.9660. For this sample, FARM-DS returns 2 fired positive rules and 5 fired negative rules, of which the most general ones and the most specific ones are shown in Table 4.7. The larger support of the negative rules may help human experts to make final correct decisions and find inherent disease-resulting mechanisms.

Secondly, FARs extracted by FARM-DS are short and compact. FARM-DS is executed again on the whole dataset. 22 positive rules and 8 negative rules are extracted. In average, the length of a positive rule is 2.6, the length of a negative rule is 4.3, and every sample activates 3.3 positive rules and 5.6 negative rules. We believe that both the short length and the small number of activated rules can make extracted FARs easy to understand for further study.

TABLE 4.8
ACTIVATION FREQUENCY OF FEATURES ON
THE WISCONSIN BREAST CANCER DATA

Feature	positive (malignant) count	negative (benign) count	activated frequency
1	6 high / 0 low / 22	0 high / 1 low / 8	0.3977
2	8 high / 0 low / 22	0 high / 3 low / 8	0.7386
3	8 high / 0 low / 22	0 high / 3 low / 8	0.7386
4	3 high / 0 low / 22	0 high / 6 low / 8	0.8864
5	0 high / 0 low / 22	0 high / 5 low / 8	0.6250
6	12 high / 0 low / 22	0 high / 4 low / 8	1.0455
7	1 high / 0 low / 22	0 high / 3 low / 8	0.4205
8	8 high / 1 low / 22	0 high / 4 low / 8	0.8182
9	0 high / 10 low / 22	0 high / 8 low / 8	0.5455

Thirdly, FARs are helpful to select important features. In Table 4.8, we count the activated numbers for each feature. As mentioned above, a larger number in a feature signals a higher probability of malignancy. So if a feature f is displayed in a positive rule in the format of “f is high”, it is correctly activated. If a feature is displayed in a positive rule in the format of “f is low”, it is wrongly activated. For negative rules, correct activation and wrong activation are defined reversely. The result demonstrates that the extracted FARs are reasonable because most of features are correctly activated. The activated frequency is calculated by decreasing the wrongly activated frequency from the correctly activated frequency. For example, the activation frequency of f8 is (8-1)/22 + 4/8 = 0.8122. The number of bare nuclei (f6), the degree of marginal adhesion (f4) and the number of normal nucleoli (f8) are most useful for classification because they are correctly activated most frequently. On the other hand, the degree of clump thickness (f1), the extent of bland chromatin (f7) and the frequency of mitosis (f9) are less useful. This kind of information is also helpful to human experts because they can pursue study on important features first.

There have been a lot of works to produce crisp or binary rule-typed knowledge on the Wisconsin breast cancer dataset [39, 94]. Compared with them, fuzzy rules with linguistic terms are more natural and hence easier to understand.

Peña-Reyes et al design the Fuzzy Cooperative Coevolution algorithm for breast cancer diagnosis to generate fuzzy rules [104]. FARM-DS combines Fuzzy Logic with Association Rule Mining, and hence provides an alternative rule mining method.

Chapter 5 FARM-DS from microarray expression data

5.1 Biological background

Every organism is composed of cell(s). In each cell, there is a nucleus, where the genetic material (DNA) is located. The coding segments of DNA, named “genes”, contain the sequence information for specific proteins, which are macro-molecules that play the key roles on biochemical and biological function, regulation and development of the organism. As a matter of fact, all cells in the same organism have exactly the same genome. However, due to different tissue types, different development stages, and different environmental conditions, genes from cells in the same organism can be expressed in different combinations and/or different quantities during the transcription process from DNA to messenger RNA (mRNA) and the translation process from mRNA to proteins. These different gene expression patterns, including both the combination and quantity, thus account for the huge variety of states and types of cells in the same organism [109]. Different organisms have different genomes and different gene expression patterns.

Very recently, DNA microarray (including cDNA microarray and GeneChip) has been developed as a powerful technology for molecular genetics studies, which simultaneously measures the mRNA expression levels of thousands to tens of thousands genes. A typical microarray expression experiment monitors expression level of each gene multiple times under different conditions or in different tissue types (for example, healthy tissue versus cancerous tissue, one kind of cancerous tissue versus another cancerous tissue). By recording such huge gene expression data sets, it opens the possibility to distinguish tissue types and to identify disease-related genes whose expression data are good diagnostic indicators [6, 10, 69, 92, 93, 96, 109].

From the viewpoint of data mining, it is a predictive data mining task [54] to distinguish different tissue types because the goal is to predict the unknown value of a variable (healthy or cancerous; if cancerous, which kind of cancer) of interest given known values of other variables (gene expression data). More specifically, it could be modeled as a classification problem. For example, one well-known problem by utilizing microarray gene expression data is to distinguish between two variants of leukemia, which are Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The AML/ALL problem could be modeled as a binary classification problem: if a sample is ALL, it is classified to be a negative case and -1 is output, otherwise it is AML and 1 is output.

5.2 Challenges for bioinformatics scientists

A typical gene expression dataset is extremely sparse compared to a traditional classification dataset: the data usually comes with only dozens of samples but with thousands or even tens of thousands of genes/features. This extreme sparseness is believed to significantly deteriorate the performance of a classifier. As a result, the ability to extract a subset of informative genes while removing irrelevant or redundant genes is crucial for accurate classification. Furthermore, it is also helpful for biologists to find the inherent cancer-resulting mechanism and thus to develop better diagnostic methods or find better therapeutic treatments. From the data mining viewpoint, this gene selection problem is essentially a feature selection or dimensionality reduction problem. A good dimensionality reduction method should remove irrelevant or redundant features while keep informative or important features for classification. A classifier modeled in the resulted lower-dimensioned feature space is expected to capture the inherent data distribution better and thus has a better performance.

For example, the AML/ALL data has only 72 samples (tissues) with 7129 features (gene expression measurements). That means, without gene selection, we would need to discriminate and classify such a few samples in such a high dimensional space. It is unnecessary or even harmful for classification because it is believed that no more than 10% of these 7129 genes are relevant to Leukemia classification [48].

Moreover, we notice that most of current related works stop when a group of informative genes are selected. However, the behavior of the classifier modeled on the selected genes is difficult to understand by human experts. It is desirable to go one step further for knowledge discovery from the selected genes to ease further cancer study.

As a brief summary, there are three highly-correlated challenging tasks:

- Key Gene Selection: given some tissues, extract cancer-related genes while remove irrelevant or redundant genes.
- Cancer Classification: given a new tissue, predict if it is healthy or not; if not, predict which kind of cancer it has.
- Cancer-Gene Knowledge Discovery: After key genes are selected, extract knowledge from the classifier modeled on these key genes in the format of cases or rules.

FARM-DS can be applied to the 3rd task with mining fuzzy associations rules to uncover correlations between genes and cancers.

5.3 Simulation Environment and Datasets

The hardware used in the simulations is a laptop with centrino-1.6MHz CPU and 1024M memory. The software we developed is based on OSU SVM Classifier Matlab Toolbox [86], which implements a Matlab interface to LIBSVM [25].

TABLE 5.1
CHARACTERISTICS OF DATASETS

Dataset	#genes	#samples	#neg : #pos
AML/ALL	7129	72	47:25
colon cancer	2000	62	40:22
prostate cancer	12600	102	52:50

Table 5.1 lists characteristics of three datasets used in simulations for this work.

For the AML/ALL leukemia classification [48], there are 72 samples (47 ALL and 25 AML) from bone marrow and blood sample specimens. The 7129 features correspond to some normalized gene expression values extracted from the microarray image: 6817 of them come from human genes and the other 312 come from control genes.

The colon cancer dataset [6] is also used in simulations. For the colon cancer dataset, there are 22 normal tissues and 40 colon cancer tissues. Gene expression information of colon cancer on more than 6500 genes were measured using oligonucleotide microarray and 2000 of them with highest minimum intensity were extracted to form a matrix of 62 tissues × 2000 gene expression values. Similar to the AML/ALL dataset, some non-human genes are included for control.

The third dataset in our simulations is the prostate cancer dataset for tumor versus normal classification [110]. The dataset consists of 102 prostate samples (52 with tumors and 50 without tumors). The 12600 features correspond to some normalized gene expression values extracted from the microarray image.

5.4 Perfect gene subsets

GSVM-RFE can find multiple compact cancer-related gene subsets on each of which a SVM with 100% leave-one-out validation accuracy can be modeled [22]. In the following, such a gene subset is referred as a "perfect" gene subset. Table 5.2 lists a perfect subset of 8 genes for the AML/ALL dataset. Table 5.3 lists a perfect subset of 5 genes for the colon cancer dataset. Table 5.4 lists a perfect subset of 8 genes for the prostate cancer dataset.

TABLE 5.2
A PERFECT GENE SUBSET SELECTED ON THE AML/ALL DATASET

rank/ index	GAN	Description of Gene Function	References (PMID)
1/4847	X95735	Homo sapiens Zyxin	11433529
2/5039	Y12670	Leptin receptor gene-related protein	15337805
3/230	D14659	KIAA0103 gene	x
4/461	D49950	Interferon-gamma inducing factor (IL-18)	11261420 12860020 12804640 12513747 12363462
5/2242	M80254	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR	12846892
6/1834	M23197	Human differentiation antigen (CD33)	12939719 12899727 12162910 11590793
7/1796	M20902	Apolipoprotein C-I (VLDL)	7690733 15343346 2570021 8019964
8/1779	M19507	Myeloperoxidase	1336394 1751367 1650411 6304866

TABLE 5.3
A PERFECT GENE SUBSET SELECTED ON THE COLON CANCER DATASET

rank/ index	GAN	Description of Gene Function	References (PMID)
1/377	Z50753	GCAP-II/uroguanylin precursor	8519795
2/1353	M31303	Human oncoprotein 18 (Op18) gene, complete cds	x
3/1423	J02854	20-kDa myosin light chain (MLC-2)	1535481 1535480 3909097
4/353	T57882	Stratagene fetal spleen	x
5/1976	K03474	Human Mullerian inhibiting substance gene, complete cds	x

TABLE 5.4
A PERFECT GENE SUBSET SELECTED ON THE PROSTATE CANCER DATASET

rank/ index	GAN /GPL91	Description of Gene Function	References (PMID)
1/6185	X07732	hepatoma mRNA for serine protease hepsin	11518967
2/4649	M16942		12603425 11262202 9655265 7690428 6640262
3/5821	AF044311		x
4/5045	AL080150		7843088 2053044 745402 944985 5950231
5/10537	AF045229		x
6/6368	AB017363		x
7/11818	M21535	erg protein (ets-related gene)	x
8/5402	W27944	39g8 retina (?)	x

5.5 Gene-cancer knowledge discovery

TABLE 5.5
CLASSIFICATION ERRORS OF THE FOUR MODELS

Data (size)	SVM	DTs	FARM-DS	ANFIS
AML/ALL (72)	0	7	2	1
colon cancer (62)	0	9	13	1
prostate cancer (102)	0	13	7	8

TABLE 5.6
AUC OF THE FOUR MODELS

data	SVM	DTs	FARM-DS	ANFIS
AML/ALL	1.0000	0.8881	0.9600	0.9600
colon cancer	1.0000	0.8364	0.7966	1.0000
prostate cancer	1.0000	0.8731	0.9312	0.9858

TABLE 5.7
RULE NUMBERS OF THE FOUR MODELS

data	SVM	DTs	FARM-DS	ANFIS
AML/ALL	7	4	5	2
colon cancer	6	5	8	3
prostate cancer	7	8	15	4

TABLE 5.8
AVERAGE RULE LENGTHS OF THE FOUR MODELS

data	SVM	DTs	FARM-DS	ANFIS
AML/ALL	8.0	2.0	4.8	8.0
colon cancer	5.0	2.4	2.4	5.0
prostate cancer	8.0	4.1	3.1	8.0

In this section, FARM-DS is compared to other three classification models, including SVM, Decision Trees, and ANFIS on each of the three datasets with the corresponding perfect gene subset reported above. We evaluate a model's performance both in terms of accuracy and interpretability. Classification errors [54] (See Table 5.5) and area under the ROC curve (AUC) [19] (See Table 5.6) by the leave-one-out validation heuristic are used for accuracy comparison. A smaller error and a larger AUC mean a more accurate classifier.

On the other hand, number (See Table 5.7) and average length (See Table 5.8) of rules extracted on the whole dataset are reported for interpretability comparison. The length of a rule is defined to be the number of features appeared in the antecedent part of this rule. A classifier is easy to interpret if the extracted rules are few and short.

In the following, all results are reported and analyzed in the order of the AML/ALL dataset, the colon cancer dataset, and the prostate cancer dataset.

The extracted compact but highly informative gene subsets make it possible and meaningful to discover useful knowledge based on them. FARM-DS works on these gene subsets for fuzzy association rule mining to provide strong decision support for further cancer study. The consequent part of a FAR is limited to be the class label {-1, +1}.

5.6 Fuzzy association rules

TABLE 5.9
5 FUZZY ASSOCIATION RULES FOR AML/ALL DATASET

G1	G2	G3	G4	G5	G6	G7	G8	label
	-1	-1		-1				-1
	-1	-1	-1			-1	-1	-1
	-1		-1	-1		-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1
	-1	-1			+1			+1

TABLE 5.10
8 FUZZY ASSOCIATION RULES FOR COLON DATASET

G1	G2	G3	G4	G5	label
				+1	-1
			-1		-1
			-1	+1	-1
	-1		-1		-1
-1	-1	-1	-1		-1
		-1		-1	+1
-1		-1		-1	+1
-1	-1	-1		-1	+1

TABLE 5.11
15 FUZZY ASSOCIATION RULES FOR PROSTATE DATASET

G1	G2	G3	G4	G5	G6	G7	G8	label
				-1				-1
-1								-1
-1				-1		-1		-1
						+1		+1
+1								+1
	+1			+1				+1
+1						+1		+1
	+1			+1		+1		+1
+1	+1			+1				+1
+1		+1				+1		+1
+1		+1				+1	+1	+1
+1	+1		+1	+1				+1
+1	+1	+1		+1		+1		+1
+1	+1	+1	+1	+1		+1		+1
+1	+1	+1	+1	+1	+1	+1	+1	+1

FARM-DS has higher accuracy than DTs. On the other hand, compared with SVM, FARM-DS extracts much shorter rules and thus easier to interpret. 5, 8, 15 rules with average length 4.8, 2.4, 3.1 are extracted and reported in Tables 5.9-5.11, respectively. In the Tables, the empty cell means the “not available” condition of the corresponding gene in the corresponding rule. A low expressed gene is expressed as “-1”, which is actually a fuzzy membership function on the gene; while “+1” means a high expressed gene. Notice that the number of activated rules is even fewer for a special sample.

Chapter 6 Fuzzy-Granular Gene Selection from Microarray Expression Data

6.1 Introduction

Selecting informative and discriminative genes from huge microarray gene expression data is an important and challenging bioinformatics research topic. This chapter proposes a fuzzy-granular method for the gene selection task. Firstly, genes are grouped into different function granules with the Fuzzy C-Means algorithm (FCM). And then informative genes in each cluster are selected with the Signal to Noise metric (S2N). With fuzzy granulation, information loss in the process of gene selection is decreased. As a result, more informative genes for cancer classification are selected and more accurate classifiers can be modeled. The simulation results on two publicly available microarray expression datasets show that the proposed method is more accurate than traditional algorithms for cancer classification. And hence we expect that genes being selected can be more helpful for further biological studies.

The rest of the chapter is organized as follows. In Section 2, previous works on cancer classification and gene selection are briefly reviewed. After that, a new fuzzy-granular gene selection algorithm is proposed in Section 3. Section 4 evaluates the performance of this method on two microarray expression datasets. Finally, Section 5 summarizes the chapter.

6.2. Traditional algorithms for gene selection

6.2.1. SVM for cancer classification

Based on [50], Support Vector Machine (SVM) is believed to be a superior model for high-dimensional classification problems including cancer classification on microarray expression data. SVM is a new generation learning system based on recent advances in statistical learning theory [123].

Due to extreme sparseness of microarray gene expression data, the dimension of input space is already high enough so that the cancer classification is already as simple as a linear separable task [50]. It is unnecessary and even harmful to transfer it to a higher implicit feature space with a non-linear kernel. As a result, usually a SVM with a linear kernel (Eq. 6.1) [22] is adopted as the basic cancer classifier.

$$K(x_i, x_j) = (x_i \bullet x_j)\,. \qquad (6.1)$$

For a linear SVM, the margin width can be calculated by Equations 6.2-6.3.

$$w = \sum_{i=1}^{N_s} \alpha_i y_i x_i\,, \qquad (6.2)$$

$$\text{margin width} = 2/\|w\|\,. \qquad (6.3)$$

where N_s is the number of support vectors, which are defined to be the training samples with $0 < \alpha_i \le C$. Note that C is a "regulation parameter" used to trade-off the training accuracy and the model complexity so that a good generalization capability can be achieved. Interesting readers may refer [22, 33, 108, 123] for detailed knowledge about SVM.

However, the sparseness of microarray data is so extreme that even a SVM classifier is unable to achieve a reliable performance for cancer classification. A preprocessing step for gene selection is necessary for SVM modeling to achieve more reliable classification.

6.2.2. Correlation-based feature ranking algorithms for gene selection

Correlation-based gene selection algorithms work in a forward selection way by ranking genes individually in terms of a correlation-based metric, and then the top ranked genes are selected to form the most informative gene subset [38, 46, 99].

Some commonly used ranking metrics are

Signal-to-Noise (S2N) [46]

$$w_i = \frac{|\mu_i(+) - \mu_i(-)|}{\sigma_i(+) + \sigma_i(-)}. \tag{6.4}$$

Fisher Criterion (FC) [99]

$$w_i = \frac{(\mu_i(+) - \mu_i(-))^2}{\sigma_i(+)^2 + \sigma_i(-)^2}. \tag{6.5}$$

T-Statistics (TS) [38]

$$w_i = \frac{|\mu_i(+) - \mu_i(-)|}{\sqrt{\frac{\sigma_i(+)^2}{n(+)} + \frac{\sigma_i(-)^2}{n(-)}}}. \tag{6.6}$$

In Equations 6.4-6.6, $\mu_i(+)$ and $\mu_i(-)$ are the mean values of the *i*th gene's expression data over positive and negative samples in the training dataset, respectively. $\sigma_i(+)$ and $\sigma_i(-)$ are the corresponding standard deviations. $n(+)$ and $n(-)$ denote the numbers of positive and negative training samples, respectively. A larger w_i means that the *i*th gene is more informative for cancer classification.

Correlation-based algorithms are straightforward to understand and work efficiently. If there are *d* genes originally, the ranking process takes *O(dlgd)* time. However, a common drawback is that these algorithms rank genes in one single group. Biologically, some genes may regulate cancers with a similar function and hence be similarly expressed.

With correlation-based algorithms, these genes may be ranked close enough. If they happen in the top of the ranking list, all of them may be selected as “informative” genes. As a result, the process of gene selection is biased to this single function and genes with other functions are removed. Because multiple different gene groups may regulate cancers in different ways, the biological analysis on genes selected by traditional correlation-based algorithms may lose other cancer-related information.

Biologically, different groups of genes may regulate cancers with different functions on one hand and one single gene may have more than one function to regulate cancers on the other hand. To select genes with more information from different function groups for reliable cancer classification and diagnosis, the novel Fuzzy-Granular based algorithm is presented in this work. The algorithm is based on the principles of granular computing.

6.3. A new fuzzy-granular based algorithm for gene selection

6.3.1. Granular computing

Granular computing represents information in the form of some aggregates (called “information granules”) such as subsets, classes, and clusters of a universe and then solves the targeted problem in each information granule [11, 83, 124-125]. On one hand, for a huge and complicated problem, it embodies Divide-and-Conquer principle to split the original task into a sequence of more manageable and smaller subtasks. On the other hand, for a sequence of similar little tasks, it comprehends the problem at hand without getting buried in all unnecessary details. As opposed to traditional data-oriented numeric computing, granular computing is knowledge-oriented [124]. From the data mining viewpoint, granular computing is knowledge-oriented. This means that data mining algorithms can be more effective by embedding the prior knowledge into the granulation process.

Lin does hybrid research in granular computing based on rough sets, fuzzy sets and topology and uses the granular computing theory in data mining applications [80-83]. Pedrycz applies interval mathematics, fuzzy sets, rough sets and random sets to granular computing research and relevant applications such as pattern recognition [15, 101-103].

In essential, our algorithm utilizes the Fuzzy C-Means clustering algorithm (FCM) to group genes into different function granules based on their expression patterns. Bezdek proposed FCM [14]. The advantage of FCM clustering is that it can assign a sample (a gene here) into multiple clusters with different membership values. Because a gene may regulate cancers with multiple functions, FCM matches the need to utilize this biological knowledge for granulation.

6.3.2. Relevance Index

"Relevance Index" (RI) was used to measure the relevance of a feature to a cluster in [35] to ease an unsupervised clustering process. Here the idea is extended as a preprocessing step. The goal here is to pre-filter some irrelevant genes to ease the following gene selection and supervised classification. Because a gene is possible to be negatively expressed or positively expressed, Equations 6.7-6.8 define the negative relevance index and the positive relevance index to measure the negative correlation and the positive correlation of a gene with the cancer being studied, respectively.

$$R_{i-} = 1 - \sigma_{i-}^2 / \sigma_i^2 , \tag{6.7}$$

$$R_{i+} = 1 - \sigma_{i+}^2 / \sigma_i^2 , \tag{6.8}$$

where σ_i^2, σ_{i-}^2, and σ_{i+}^2 are the variances of the projected values on the *i*th gene of the whole training samples, the negative training samples, and the positive training samples, respectively.

For example, gene X1 in Fig. 6.1 is positive-related because the local variance among positive samples is much smaller than the global variance on the whole samples. Similarly, gene X2 is negative-related, gene X3 is both negative-related and positive-related, and gene X4 can be viewed as an "irrelevant" gene in that it is neither negative-related nor positive-related.

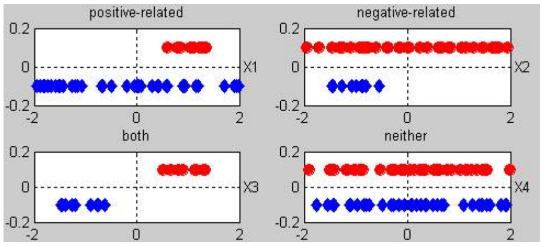

Figure. 6.1. positive-related gene, negative-related gene, both, neither

To apply RI metric for gene selection, a negative filtering threshold $\alpha_- \in [0,1)$ and a positive filtering threshold $\alpha_+ \in [0,1)$ need to be decided. The *i*th gene is "negative-related" if $R_{i-} \geq \alpha_-$. Similarly, it is "positive-related" if $R_{i+} \geq \alpha_+$. If $R_{i-} < \alpha_-$ and $R_{i+} < \alpha_+$, it is "irrelevant". A gene may be both negative-related and positive-related. These two filtering thresholds should be selected carefully: firstly, they can not be too large, otherwise the information loss may happen because some cancer-related genes are wrongly eliminated; secondly, they should be selected "in balance", which means negative-related genes and positive-related genes should be selected in balance, otherwise the minor genes are possible to be totally eliminated to result in performance degradation, especially when negative-related genes and positive-related genes are significantly imbalanced in the original dataset.

6.3.3. Fuzzy C-Means clustering

RI metric helps us to remove irrelevant genes. The next step is to pick up discriminative genes while removing redundant genes. By removing redundancy, genes with more regulation functions may be selected, assuming the number of genes is fixed.

Some genes may similarly regulate cancers and thus be similarly expressed. And hence these genes may play a similar role in cancer classification. As a result, if genes with similar expression patterns are grouped together into clusters, a few typical genes in a cluster may be selected and other genes in the cluster may be safely eliminated without significant information loss. On the other hand, an informative gene may contribute to cancer classification with complex correlations with multiple different clusters. Therefore, after the pre-filtering by RI metric, FCM is adopted to group genes into different function clusters.

FCM groups genes into K clusters with centers $c_1,\cdots c_k,\cdots c_K$ in the training samples space. (That is, each training sample is a dimension of the space). FCM assigns a real-valued vector $U_i = \{\mu_{1i},\cdots \mu_{ki},\cdots,\mu_{Ki}\}$ to each gene. $\mu_{ki} \in [0,1]$ is the membership value of the *i*th gene in the *k*th cluster. The larger membership value indicates the stronger association of the gene to the cluster. Membership vector values μ_{ki} and cluster centers c_k can be obtained by minimizing

$$J(K,m) = \sum_{k=1}^{K}\sum_{i=1}^{N}(\mu_{ki})^m d^2(x_i,c_k), \tag{6.9}$$

$$d^2(x_i,c_k) = (x_i - c_k)^T A_k (x_i - c_k), \tag{6.10}$$

$$\sum_{k=1}^{K}\mu_{ki} = 1, 0 < \sum_{i=1}^{N}\mu_{ki} < N, \tag{6.11}$$

where $1 \le i \le N$ and $1 \le k \le K$ [14].

In Eq. 6.9, K and N are the number of clusters and the number of genes in the dataset, respectively. $m>1$ is a real-valued number which controls the 'fuzziness' of the resulting clusters, μ_{ki} is the degree of membership of the *i*th gene in the *k*th cluster, and $d^2(x_i, c_k)$ is the square of distance from the *i*th gene to the center of the *k*th cluster. In Eq. 6.10, A_k is a symmetric and positive definite matrix. If A_k is the identity matrix, $d^2(x_i, c_k)$ corresponds to the square of the Euclidian distance. Eq. 6.11 indicates that empty clusters are not allowed.

6.3.4. Fuzzy-Granular based gene selection

We categorize genes into three classes:

- Informative genes, which are essential for cancer classification and diagnosis;
- Redundant genes, which are also cancer-related but there are some other informative genes regulating cancers similarly but more significantly;
- Irrelevant genes, which are not cancer-related and do not affect cancer classification;

A desirable algorithm should extract genes of the first category while eliminating genes of the last two categories. However, it is difficult to perfectly implement this goal. Firstly, inherent cancer-related factors are very possibly mixed with other non-cancer-related factors for classification. Secondly, some non-cancer-related factors may even have more significant effects on classifying the training dataset. It is actually the notorious "overfitting" problem. It is even worse when the training dataset is too small to embody the inherent real data distribution, which is common for microarray gene expression data analysis.

Correlation-based algorithms work by ranking genes in a same group. However, some really informative genes are possible to be wrongly eliminated. For example, an informative gene is ranked the highest in a function group. However, the genes in this function group are all ranked below another group of genes. As a result, all of genes including the informative gene in this function group are possibly eliminated.

The fuzzy-granular based algorithm is proposed in this work for more reliable gene selection. It works in two stages. Fig. 6.2 sketches the algorithm.

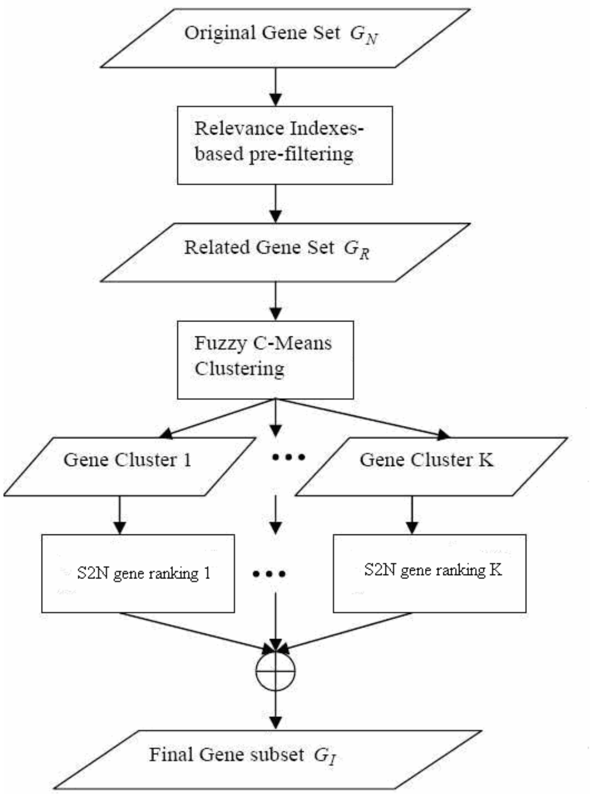

Figure. 6.2. Fuzzy-Granular gene selection

At the first stage, RI metrics are used to coarsely group genes into two granules: “relevant granule” and “irrelevant granule”. The relevant granule consists of negative-related genes (with $R_{i-} \geq \alpha_-$) and positive-related genes (with $R_{i+} \geq \alpha_+$), while the irrelevant granule is comprised of irrelevant genes (with $R_{i-} < \alpha_-$ and $R_{i+} < \alpha_+$). Notice $\alpha_- \in [0,1)$ is the negative filtering threshold and $\alpha_+ \in [0,1)$ is the positive filtering threshold. Only genes in the relevant granule survive for the following stages. The assumption is that irrelevant genes are not so useful for cancer classification or even possible to correlate other genes in some unknown complex way to confuse FCM to get good clusters/granules or confuse SVMs to get good classification. This pre-filtering process can dramatically decrease the number of candidate genes on which FCM works. Therefore, it can improve both the efficiency and the effectiveness of the following stages. Notice that the pre-filtering step by RI metrics is targeted at minimizing information loss by eliminating most of irrelevant genes.

At the second stage, genes which survive after the first stage are grouped by FCM into several "function granules". In each function granule, some correlation-based metric is used to rank genes in the descending order. The lower-ranked genes are removed. And then all remaining genes in these function granules are combined disjunctively to form the final gene subset. By using FCM, our algorithm explicitly groups genes with similar expression patterns into clusters and then the lower-ranked genes in each cluster could be safely removed as redundant genes because the more significant genes with similar functions will survive. Furthermore, due to complex correlation between genes, the similarity is by no means a "crisp" concept. FCM deals with complex correlation between genes by assigning a gene into several clusters with different membership values. Therefore, a really informative gene achieves more than one opportunity to survive.

6.4. Simulation

In our simulation, the new fuzzy-granular based algorithm is compared with three correlation-based algorithms, S2N, FC and TS. The hardware we used is a desktop with

P4-2.8MHz CPU and 256M memory. The software we developed is based on OSU SVM Classifier Matlab Toolbox [86] which implements a Matlab interface to LIBSVM [25].

6.4.1. Evaluation metrics

Three metrics, accuracy (Eq. 6.12), sensitivity (Eq. 6.13) and specificity (Eq. 6.14), are used to evaluate classification performance.

Here, sensitivity is defined to be the fraction of the real negatives that actually are correctly predicted as negatives. Specificity is defined to be the fraction of the tissues predicted as negatives that really are negatives.

$$accuracy = \frac{TN + TP}{TN + FN + FP + TP}. \tag{6.12}$$

$$sensitivity = TN/(TN + FP). \tag{6.13}$$

$$specificity = TN/(TN + FN). \tag{6.14}$$

By the definitions, the combination of sensitivity and specificity can be used to evaluate a model's balance ability so that we know if a model is biased to a special class.

We also report the area under the ROC curve (AUC) [19] for each algorithm. The AUC value can indicate a model's generalization capability as a function of varying a classification threshold. An area of 1 represents a perfect classification, while an area of 0.5 represents a worthless model.

6.4.2. Data description

The prostate cancer dataset for tumor versus normal classification [79] is used in our simulation. It consists of 136 prostate samples (77 with tumors and 59 without tumors). The 12600 features correspond to some normalized gene expression values extracted from the microarray image. Here negatives are defined to be the normal prostate samples without tumor, while positives are the tumor samples.

The colon cancer dataset is also used for comparison [79]. There are 22 normal tissues and 40 colon cancer tissues. Gene expression information of colon cancer on more than 6500 genes were measured using oligonucleotide microarray and 2000 of them with highest minimum intensity were extracted to form a matrix of 62 tissues × 2000 gene expression values.

6.4.3. Data modeling

The same as [48], the original dataset is simply normalized so that each gene vector has 0 for mean and 1 for standard deviation. To avoid overfitting, for leave-one-out or bootstrapping validation accuracy evaluation, validation samples are kept out from calculating these two values.

The regulation parameter $C \equiv 1$ for the linear SVMs. For FCM, the "fuzziness degree" $m = 1.15$, the maximal iteration number is 100, and the minimal improvement $\varepsilon = 10^{-5}$. For fuzzy-granular gene selection, genes are grouped into 10 clusters, in each of which S2N/FC/TS is used for gene ranking, and then 2 highest ranked genes in each of the 10 clusters are combined disjunctively to form the final gene set with the size (at most) 20. For comparison, top 20 ranked genes are also selected based on S2N, FC and TS, respectively.

Notice that fuzzy membership values are defuzzified in such a way that a gene is always grouped into the cluster with the largest membership value and the cluster with the second largest membership value. The assumption is that different gene function groups are clustered based on their expression strengths. Some genes whose expression strengths are between two groups may be more suitable to be clustered into the two groups at the same time. This way, each gene achieves two opportunities to be selected.

The genes distribution in the prostate cancer dataset is highly imbalanced between negative-related genes and positive-related genes. If $\alpha_+ = \alpha_- = 0.5$, 4761 positive-related genes and only 110 negative-related genes are survived. To alleviate the imbalance, $\alpha_+ = 0.75$ and $\alpha_- = 0.5$ are used to select 721 positive-related genes and 110 negative-related genes. There is no overlapping between positive-related genes and negative-related genes. Similarly, for the colon cancer dataset, $\alpha_+ = 0.5$ and $\alpha_- = 0.1$.

The leave-one-out validation is used [50]: in each fold, one sample is left for validation and the other samples are used for training. Another evaluation heuristic adopted is balanced .632 bootstrapping [20]: random sampling with replacement is repeated for 100 times on each of the two datasets. Each tissue sample appears exactly 100 times in the computation to reduce variance [29].

6.4.4. Result analysis

Table 6.1 reports the leave-one out validation performance of six gene selection algorithms, named S2N, Fuzzy-Granular with S2N, FC, Fuzzy-Granular with FC, TS and Fuzzy-Granular with TS. Table 6.2 reports .632 bootstrapping performance. The results show that Fuzzy-Granular gene selection improves prediction performance compared to correlation-based algorithms both in terms of accuracy and AUC. Specifically, Fuzzy-Granular with S2N has the best performance under both leave-one-out validation and .632 bootstrapping validation. This means fuzzy-granular algorithm can select more informative genes.

TABLE 6.1
LEAVE-ONE OUT VALIDATION PERFORMANCE ON THE PROSTATE CANCER DATASET

model	accuracy	AUC	sensitivity	specificity
S2N [46]	0.8309	0.8388	0.7792	0.9091
FG+S2N	**0.9191**	**0.9226**	**0.8961**	**0.9583**
FC [99]	0.8824	0.8803	0.8961	0.8961
FG+ FC	0.9118	0.9102	0.9221	0.9221
TS [38]	0.8603	0.8588	0.8701	0.8816
FG+TS	0.9191	0.9206	0.9091	0.9459

TABLE 6.2
.632 BOOTSTRAPPING PERFORMANCE ON THE PROSTATE CANCER DATASET

model	accuracy	AUC	sensitivity	specificity
S2N [46]	0.8323	0.8484	0.8047	0.9073
FG+S2N	**0.8684**	**0.9125**	**0.9045**	**0.9357**
FC [99]	0.8489	0.8688	0.8938	0.8829
FG+ FC	0.8621	0.9054	0.9002	0.9280
TS [38]	0.8556	0.8734	0.8953	0.8880
FG+TS	0.8530	0.8864	0.8379	0.9436

TABLE 6.3
LEAVE-ONE OUT VALIDATION PERFORMANCE ON THE COLON CANCER DATASET

model	accuracy	AUC	sensitivity	specificity
S2N [46]	0.8710	0.8591	0.9000	0.9000
FG+S2N	**0.9516**	**0.9523**	**0.9500**	**0.9744**
FC [99]	0.8710	0.8693	0.8750	0.9211
FG+ FC	0.8710	0.8591	0.9000	0.9000
TS [38]	0.7903	0.7761	0.8250	0.8462
FG+TS	0.8710	0.8591	0.9000	0.9000

TABLE 6.4
.632 BOOTSTRAPPING PERFORMANCE ON THE COLON CANCER DATASET

model	accuracy	AUC	sensitivity	specificity
S2N [46]	0.8419	0.8701	0.9092	0.9116
FG+S2N	**0.8428**	**0.8881**	**0.9285**	**0.9212**
FC [99]	0.8314	0.8690	0.9087	0.9128
FG+ FC	0.8323	0.8806	0.9236	0.9160
TS [38]	0.8150	0.8437	0.8741	0.9010
FG+TS	**0.8428**	**0.8881**	**0.9285**	**0.9212**

There are two reasons for the good performance of fuzzy-granular gene selection. Firstly, RI-based pre-filtering eliminates most of irrelevant genes and hence decreases correlation-induced noise. Secondly, FCM explicitly groups genes into different clusters with different expression patterns so that informative genes from different function granules (clusters) are selected in balance.

Similar performance gain of fuzzy-granular gene selection is observed for the colon cancer dataset (Table 6.3 and Table 6.4).

6.5. Summary

To select a more informative gene set for reliable cancer classification, the fuzzy-granular based algorithm is proposed in this chapter. Firstly, it utilizes Relevance Index metrics to remove most of irrelevant genes to improve the efficiency and decrease the noise effect at the same time. Secondly, it explicitly groups genes with similar expression patterns into “function granules” with the Fuzzy C-Means clustering algorithm. Therefore, the lower-ranked genes in each “function granule” can be safely removed as redundant genes because more significant genes with similar functions will survive. Finally, it deals with complex correlation between genes by assigning a gene into several clusters with different membership values so that a really informative gene is more possible to survive.

Our fuzzy-granular based algorithm is more reliable for cancer classification, as the experiment results on the prostate cancer dataset and the colon cancer dataset demonstrated. The gene set selected by our algorithm is expected to be more helpful for biologists to uncover the inherent cancer-resulting mechanism.

Because of the inherent advantage to eliminate irrelevant or redundant genes while selecting really informative genes, we expect that this superior performance can also be true in processing other microarray datasets. This work is currently in processing.

Chapter 7 Conclusions and future works

In this dissertation, two fuzzy granular based algorithms have been proposed. The first one is a general Fuzzy Association Rule Mining for Decision Support algorithm (FARM-DS). By combining data clustering techniques with fuzzy interval partitions on input features, the high-level data abstraction can be extracted and the quantitative data can be efficiently transformed into fuzzy discrete transactions, on which a traditional Apriori algorithm works on mine association rules that can be utilized for classification and decision support.

The FARM-DS algorithm is compared with state-of-the-art classification algorithms on medical or biological datasets. The empirical study demonstrates that FARM-DS is accurate for classification. More importantly, besides a class label, FARM-DS also returns the fired rules for an unseen sample to human experts, and thus can provide strong decision support to assist human experts to make correct decisions.

The second algorithm is applying fuzzy granulation on the microarray expression dataset for gene selection. This algorithm utilizes Relevance Index metrics to remove most of irrelevant genes, groups genes with similar expression patterns into granules then, ranks them with correlation-based methods in each granule, finally, lower-ranked genes are removed as redundant genes.

The experiment results show that this algorithm is more reliable for cancer classification. The gene set selected by our algorithm is expected to be more helpful for biologists to uncover the inherent cancer-resulting mechanism.

As a long term research plan, our goal is to build a hybrid intelligent knowledge discovery and data mining system based on granular computing, soft computing and statistical learning to provide effective and efficient decision support for diseases diagnosis and drug design, and many other applications. The algorithms proposed in this dissertation can be viewed as a preliminary step toward the goal.

Bibliography

[1] M. Anvari and G.F. Rose, "Fuzzy relational databases," in Bezdek, Ed., analysis of Fuzzy Information, Vol. II, CRC Press, Boca Raton, FL, 1987.

[2] R. Agrawal, J. Gehrke, D. Gunopoulos, P. Raghavan. Automatic subspace clustering of high dimensional data for datamining applications. In Proc. 1998 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'98), pages 94-105. June 1998.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," Proc. Of ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'93), pp45-51, Washington, DC, July 1993.

[4] R. Agrawal, J.C. Shafer, "Parallel Mining of Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.

[5] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," Proc. of Int. Conf. Very Large Data Bases (VLDB'94), pp 487-499, Santiago, Chile, Sept. 1994.

[6] U. Alon, N. Barkai,D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. of Natl Acad Sci,USA 96: 6745-6750, 1999.

[7] W. H. Au and K. C. C. Chan. An effective algorithm for discovering fuzzy rules in relational databases. In Proc. 7th IEEE Int. Conf. Fuzzy Systems, Anchorage, AK, 1998, pages 1314-1319.

[8] W. H. Au and K. C. C. Chan. FARM: a data mining system for discovering fuzzy association rules. In Proc. 8th IEEE Int. Conf. Fuzzy Systems, Seoul, Korea, 1999, pages 1217-1222.

[9] W.-H. Au and K. C. C. Chan, "Mining Fuzzy Association Rules in a Bank-Account Database," IEEE Transactions on Fuzzy Systems, Volume 11, Issue. 2, pp. 238-248, April 2003.

[10] E. Bair, R. Tibshirani, “Machine learning methods applied to DNA microarray data can improve the diagnosis of cancer,” SIGKDD Explorations, vol. 5(2), pp. 48-55, 2003.

[11] A. Bargiela, Granular Computing: An Introduction, Kluwer Academic Pub, Kluwer, 2002.

[12] E.Baralis, P. Gazza, “A lazy approach to pruning classification rules,” Proc. IEEE Int. Conf. on Data Mining (ICDM’04), pages 35-42, 2002.

[13] K. P. Bennett and J. Blue, “A Support Vector Machine Approach to Decision Trees,” R.P.I Math Report No. 97-100, Rensselaer Polytechnic Institute, Troy, NY, 1997.

[14] J.C. Bezdek, (1981) Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum Press, New York.

[15] G. Bortolan and W. Pedrycz, “Reconstruction problem and information granularity,” IEEE Transactions on Fuzzy Systems, 2, 1997, 234-248.

[16] P. Bosc, O. Dubois, O. Pivert, H. Prade, “On fuzzy association rules based on fuzzy cardinalities,” Proc. of IEEE Int. Conf. on fuzzy system, pages 461-464, 2001.

[17] P.Bosc, L. Lietard and O. Pivert, “Functional dependencies revised under graduality and imprecision,” NAFIPS, pp. 57-62, 1997.

[18] P.Bosc, O. Pivert, “On some fuzzy extensions of association rules,” Proc. of joint 9th IFSA World Congress and 20th NAFIPS Int. Conf., pages 1104-1109, 2001.

[19] A. P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[20] U. Braga-Neto and E. R. Dougherty, “Is cross-validation valid for small-sample microarray classification?,” Bioinformatics, vol. 20, pp. 374-380, 2004.

[21] B.P. Buckles and F.E. Petry, “A fuzzy representation of data for relational databases,” Fuzzy Sets and Systems, Vol.7, pages 213-226, 1982.

[22] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, 2(2), pp. 121-167, 1998.

[23] K. C. C. Chan and W.-H. Au, “Mining fuzzy association rules,” Proc. Of 6th Int. Conf. Information Knowledge Management, pp. 209–215, Las Vegas, NV, 1997.

[24] K. C. C. Chan and W. H. Au. Mining fuzzy Association rules in database containing relational and transactional data. In Data Mining and Computational Intelligence, A. Kandel, M. Last, and H. Bunke, Eds. New Yorks: Physica-Verlag, 2001, pages 95-114.

[25] C. -C. Chang, C. -J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[26] A. Chatterjee and A. Rakshit, “Influential Rule Search Scheme (IRSS)-A New Fuzzy Pattern Classifier,” IEEE Transactions on Knowledge and Data Engineering, Volume 16, Issue. 8, pp. 881-893, August 2004.

[27] G. Chen, Q. Wei, and E. Kerre. Fuzzy data mining: Discovery of fuzzy generalized association rules. In Recent Issues on Fuzzy Database, G. Bordogna and G. Pasi, Eds. Physica-Verlag, 2000. “Studies in Fuzziness and Soft Computing” Series.

[28] G. Chen and Q. Wei. Fuzzy association rules and the extended mining algorithms. Information Sciences, vol 147, pages 201-228, 2002.

[29] M. Chernick, Bootstrap Methods: A Practitioner’s Guide, Wiley, New York, NY, 1999.

[30] B. C. Chien, Z.L. Lin and T. P. Hong. An effiecient clustering algorithm for mining fuzzy quantitative association rules. In Proc. of IFSA World Congress and NAFIPS Conference, Vol 3, pages 1306-1311, 2001.

[31] K. K. Chin, “Support vector machines applied to speech pattern classification,” Master's thesis, Engineering Department, Cambridge University, 1999.

[32] P. Clark and S. Matwin. Using qualitative models to guide induction learning. In Proc. 1993 Int. Conf. on Machine Learning (ICML'93), pages 49--56, Amherst, MA, 1993.

[33] N. Cristianini, J,Shawe-Taylor, An introduction to support Vector Machines and other kernel-based learning methods, Cambridge University Press, New York, NY, 1999.

[34] M. Delgado and A. Gonzalez, “An inductive learning procedure to identify fuzzy systems,” Fuzzy Sets and Systems, Vol. 55, pp.121-132, 1993.

[35] M. Delgado, N. Marin, D. Sanchez and M. A. Vila, “Fuzzy Association Rules: General Model and Applications,” IEEE Transactions on Fuzzy Systems, Volume 11, Issue. 2, pp. 214-225, April 2003.

[36] M. Delgado, N. Marin, M. J. Martin-Bautista, D. Sanchez and M. A. Vila, “Mining Fuzzy Association Rules: An Overview,” 2003 BISC International Workshop on Soft Computing for Internet and Bioinformatics, 2003.

[37] S. Dick and A. Kandel, “Combinatorial rule explosion eliminated by a fuzzy rule configuration,” IEEE Transactions on Fuzzy Systems, vol. 7, pp. 475-477, Aug. 1999.

[38] K. Duan, J. C. Rajapakse, “A Variant of SVM-RFE for Gene Selection in Cancer Classification with Expression Data,” Proc. of IEEE CIBIB 2004, pp. 49-55, San Diego, 2004.

[39] W. Duch, R. Adamczak and K. Grabczewski, “A new methodology of extraction, optimization and application of crisp and fuzzy logical rules,” IEEE Transactions on Neural Networks, Volume 12, pp. 277-306, 2001.

[40] J.C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” Journal of Cybernetics, Vol. 3, pp. 32-57, 1973.

[41] G. Dong and J. Li, Efficient mining of emerging patterns: Discovering trends and differences. In Proc. 1999 Int. Conf. on Knowledge Discovery and Data Mining (KDD’99), pages 43-52, San Diego, CA, Aug. 1999.

[42] M. Ester, H.P. Kriegel, and X. Xu, "Knowledge discovery in large spatial databases: focusing techniques for efficient class identification," In Proc. of 4th Int. Symp. Large Spatial Databases (SSD'95), pp. 67-82, Portland, ME, Aug. 1995.

[43] A.W.C. Fu, M.H. Wong, S.C. Sze, W.C. Wong, W.L. Wong, and W.K. Yu, Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes. In Proc. Int. Symp. on Intelligent Data Engineering And Learning (Ideal'98), Hong Kong, 1998, pages 263-268.

[44] T. Fukuda, Y. Morimoto, S, Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms and visualization. In Proc. 1996 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'96), page 13-23, Montreal, Canada, June 1996.

[45] T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama. Mining optimized association rules for numeric attributes. In Proc. 1996 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'96), pages 182-191, Montreal, Canada, June 1996.

[46] T. Furey, N. Cristianini, N.Duffy, D. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," Bioinformatics, vol. 16 pp. 906-914, 2000.

[47] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," In Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'98), pp. 73-84, Seattle, WA, June 1998.

[48] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, 286, pp. 531-537, 1999.

[49] S. Gunn, "Support vector machines for classification and regression," ISIS technical report, Image Speech & Intelligent Systems Group, University of Southampton, 1998.

[50] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, vol. 46, pp. 389-422, 2002.

[51] J. Han and Y. Fu, “Discovery of multiple-level association rules from large databases,” In Proc. of 1995 Int. Conf. on Very Large Databases (VLDB’95), pages 420-431, Zurich, Switzerland, Sept. 1995.

[52] J. Han and M. Kamber. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[53] J. Han, J. Pei and Y. Yin. “Mining frequent patterns without candidate generation,” In Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD’ 00), pages 1-12, Dallas, TX, May 2000.

[54] D. Hand, H. Mannila and P. Smyth, Principle of Data Mining, The MIT Press, Cambridge, London, 2001.

[55] R. Haux and U. Eckert, “Nondeterministic dependencies in relations: an extension to the concept of functional dependency,” Information Systems 10, Vol. 2, pp. 139-148, 1985.

[56] Y. C. He, Y.C. Tang, Y.-Q. Zhang and R. Sunderraman, “Mining Fuzzy Association Rules from Microarray Gene Expression Data for Leukemia Classification,” Proc. of International Conference on Granular Computing (GrC-IEEE 2006), Atlanta, pp. 461-465, May 10-12, 2006.

[57] Y.C. He and Y.C. Tang, Y.-Q. Zhang and R. Sunderraman, “Adaptive Fuzzy Association Rule Mining for Effective Decision Support in Biomedical Applications,” International Journal of Data Mining and Bioinformatics, Vol. 1, No. 1, pp. 3-18, 2006.

[58] Y.C. He, Y.C. Tang, Y.-Q. Zhang and R. Sunderraman, “Fuzzy-Granular Gene Selection from Microarray Expression Data,” Proc. of DMB2006 in conjunction with IEEE-ICDM2006, Hong Kong, Dec. 18, 2006, (accepted).

[59] Y.C. He, Y.C. Tang, Y.-Q. Zhang and R. Sunderraman, “Fuzzy-Granular Methods for Identifying Marker Genes from Microarray Expression Data,” Computational Intelligence for Bioinformatics, Gary B. Fogel, David Corne, and Yi Pan (eds.), IEEE Press, 2007.

[60] T.P.Hong, C.S.Kuo, and S.C. Chi, “Mining association rules from quantitative data,” Intelligent Data Analysis, vol. 3, pp. 363-376,1999.

[61] T.P. Hong, C. S. Kuo, S.C. Chi, and S.L. Wang, “Mining fuzzy rules from quantitative data based on the AprioriTid algorithm,” ACM SAC 2000 Como, pp 534-536, Italy, March 2000.

[62] T.P. Hong and S.S. Tseng, “A generalized version space learning algorithm for noisy and uncertain data,” IEEE Transactions on Knowledge and Data Engineering, Vol 9, No 2, pp. 336-340,1997.

[63] Y.-C. Hu, R.-S. Chen and G.-H. Tzeng, “ Mining fuzzy association rules for classification problem,” Computers and Industrial Engineering, Volume 43, Issue 4, pp. 735-750, 2002.

[64] Yi-Chung Hu, Ruey-Shun Chen, Gwo-Hshiung Tzeng, “Discovering fuzzy association rules using fuzzy partition methods,” Knowledge Based Systems, vol. 16, pp. 137-147. 2003.

[65] Yi-Chuang Hu, Gwo-Hshiung Tzeng, “Elicitation of classification rules by fuzzy data mining,” Engineering Applications of Artificial Intelligence, Vol. 16,pp. 709-716,2003.

[66] H. Ishibuchi, T. Nakashima and T. Murata, “Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems,” IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume 29, Issue 5, pp. 353-361, Oct. 1999.

[67] J. –S. R. Jang, “ANFIS: Adaptive-Network-Based Fuzzy Inference System,” IEEE Transactions on Systems, Man and Cybernetics, Volume 23, Issue 3, pp. 665-685, 1995.

[68] C.Z. Janikow, “Fuzzy decision trees: issues and methods,” IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume 28, Issue 1, pp. 1-14, Feb. 1998.

[69] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: A survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370-1386, 2004.

[70] M. Kamber, J.Han, and J. Y. Chiang. Metarule-guided mining of multidimensional association rules using data cubes. In Proc. 1997 Int. Conf. on Knowledge Discovery and Data Mining (KDD’97), pp. 207-210, Newport Beach, CA, Aug. 1997.

[71] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.

[72] G. Karypis, E.H. Han, and V. Kumar, "CHAMELEON: a hierarchical clustering algorithm using dynamic modeling," COMPUTER, Vol. 32, pp. 68-75, 1999.

[73] M.Kaya, R. Alhajj. Integrating mutlti-objective genetic algorithms into clustering for fuzzy association rules mining. In Proc. of the Fourth IEEE Int. Conf. on Data Mining (ICDM'04), pp. 431-434, Brighton, UK, Nov. 2004.

[74] R. Kruse and D. Nauck, "NEFCLASS - A Neuro-Fuzzy Approach for the Classification of Data," Proc. of the 1995 ACM Symposium on Applied Computing, 1995.

[75] C.-M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," SIGMOD Record, Vol. 27, No.1, pp.41-46,1998.

[76] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One sided selection," Proc. of the 14th International Conference on Machine Learning (ICML1997), pp.179-186, 1997.

[77] J. H. Lee and H. L. Kwang, "An extension of association rules using fuzzy sets," Proc of IFSA'97, 1997

[78] B. Lent, A. Swami, and J. Widom. "Clustering association rules," Proc. of 1997 International Conference on Data Engineering (ICDE'97), pp. 220-231, Birmingham, England, Apr, 1997.

[79] W. Li, J. Han and J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," Proc. of ICDM 2001, pp369-376, 2001.

[80] J. Li and H. Liu, "Kent ridge bio-medical data set repository," Available at http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

[81] T. Y. Lin, "Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems." Rough Sets in Knowledge Discovery, A. Skowron and L. Polkowski (eds), Physica-Verlag, 1998, 107-121.

[82] T. Y. Lin, "Granular Computing on Binary Relations II: Rough Set Representations and Belief Functions," Rough Sets In Knowledge Discovery, A. Skowron and L. Polkowski (eds), Physica -Verlag, pp. 121-140,1998.

[83] T. Y. Lin, "Granular Computing: Fuzzy Logic and Rough Sets," Computing with words in information/intelligent systems, L.A. Zadeh and J. Kacprzyk (eds), Physica-Verlag (A Springer-Verlag Company), pp.183-200, 1999;

[84] T. Y. Lin, "Data Mining and Machine Oriented Modeling: A Granular Computing Approach," Journal of Applied Intelligence, Kluwer, Vol 13, No 2, pp. 113-124, 2000.

[85] B.Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," Proc. of 4th International conference on knowledge discovering and data mining KDD'98, pp 80-86, 1998.

[86] J. Luo and S. M. Bridges, "Mining Fuzzy Association Rules and Fuzzy Frequency Episodes for Intrusion Detection," International Journal of Intelligent Systems, Vol. 15, No. 8, pp.687-704, 2000.

[87] J. Ma, Y. Zhao, and S. Ahalt, OSU SVM Classifier Matlab Toolbox, Available at http://www.ece.osu.edu/~maj/osu_svm/.

[88] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symposium on mathematics, Statistics and Probability, pp. 281-298, 1967.

[89] O. L. Mangasarian, W, N. Street and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," Mathematical Programming Technical Report 94-10, University of Wisconsin, 1994.

[90] J.M. Medina, M.A. Vila, J.C. Cubero, and O. Pons, "Towards the implementation of a generalized fuzzy relational database model," Fuzzy Sets and Systems, Vol. 75, pp 273-289, 1995.

[91] C.J. Merz and P.M. Murphy. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[92] R. J. Miller and Y.Yang. Association rules over interval data. In Proc. 1997 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'97), pages 452-462, Tucson, AZ, May 1997.

[93] F. Model, P. Adorjan, A. Olek, and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification," Bioinformatics, vol. 17 Suppl. 1, pp. S157-S164, 2001.

[94] E. J. Moler, M. L. Chow, and I. S. Mian, "Analysis of molecular profile data using generative and discriminative methods," Physiol. Genomics, vol. 4, pp. 109-126, 2000.

[95] M. Muselli and D. Liberati, "Binary rule generation via Hamming Clustering," IEEE Transactions on Knowledge and Data Engineering, Volume 14, pp.1258-1268, 2002.

[96] R. Ng and J. Han, "Efficient and effective clustering method for spatial data mining," In Proc. of 1994 Int. Conf. on Very Large Databases (VLDB'94), pp. 144-155, Santiago, Chile, Sept. 1994.

[97] W. S. Noble, "Support vector machine applications in computational biology," Kernel Methods in Computational Biology. B. Schoelkopf, K. Tsuda and J.-P. Vert, ed. MIT Press, pp. 71-92, 2004.

[98] J. S. Park, M. S. Chen, and P. S. Yu, "An effective hash-based algorithm for mining association rules," In Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'95), pages 175-186, San Jose, CA, May 1995.

[99] P. Pavlidis, J. Weston, J. Cai, et al., "Gene functional analysis from heterogeneous data," Proc. RECOMB, New York: ACM Press, pp. 249-255, 2001.

[100] M. J. Pazzani, S. Mani, and W. R. Shankle. Beyond concise and colorful: Learning intelligible rules. In Knowledge Discovery and Data Mining, pages 235--238, 1997.

[101] W. Pedrycz, Granular Computing: An Emerging Paradigm, Physica-Verlag, 2001.

[102] W. Pedrycz, "Granular computing in Data Mining," M.Last and A. Kandel (eds.), Data Mining & Computational Intelligence, Springer-Verlag, 2001.

[103] W. Pedrycz and G. Vukovich, "Granular computing in pattern recognition," Neuro-Fuzzy Pattern Recognition, (H. Bunke and A. Kandel, eds.), World Scientific, 2002.

[104] C.-A. Peña-Reyes and M. Sipper, "Applying Fuzzy CoCo to breast cancer diagnosis," Proc. of the 2000 Congress on Evolutionary Computation (CEC00), vol. 2, pp. 1168-1175, 2000.

[105] J.R. Quinlan. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[106] K. V. S. V. N. Raju and A. K. Majumdar: "Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems", ACM Transactions on Database Systems, Vol. 13, NO. 2, June 1988, 129-166.

[107] A. Savasere, E. Omiecinski, and S. Navathe. "An efficient algorithm for mining association rules in large databases," In Proc. of 1995 Int. Conf. on Very Large Databases (VLDB'95), pages 420-431, Zurich, Switzerland, Sept. 1995.

[108] B. Schölkopf, I. Guyon, and J. Weston, "Statistical Learning and Kernel Methods in Bioinformatics," Artificial Intelligence and Heuristic Methods in Bioinformatics 183, (Eds.) P. Frasconi und R. Shamir, IOS Press, Amsterdam, The Netherlands, pp. 1-21, 2003.

[109] G. P.-Shapiro, P. Tamayo, "Microarray data mining: facing the challenges," SIGKDD Explorations, vol. 5(2), pp. 1-5, 2003.

[110] D. Singh, P. G. Febbo, K. Ross, et al, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell 1:203-209, 2002

[111] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In Proc. 1996 ACM-SIGMOD Int. Conf. on Manage of Data (SIGMOD'96), pages 1-12, Montreal, Canada, June 1996.

[112] M. Sugeno and G. T. Kang, "Structure identification of fuzzy model," Fuzzy Sets System, Volume 28, pp 15-33, 1988.

[113] K.A. Swets, "Measuring the accuracy of diagnostic systems," Science 240, 1988, pp. 1285-1293..

[114] T.Takagi and M.Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” IEEE Transactions on Systems, Man and Cybernetics, Volume 15, pp. 116-132, Jan. 1985.

[115] Y.C. Tang, Y.C. He, Y.-Q. Zhang, Z. Huang, X. Tony Hu and R. Sunderraman, “A Hybrid CI-Based Knowledge Discovery System on Microarray Gene Expression Data,” Proc. of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2005), San Diego, pp.25-30, Nov. 14-15, 2005.

[116] Y.C. Tang, B. Jin, Y. Sun, Y.-Q. Zhang, "Granular Support Vector Machines for Medical Binary Classification Problems," Proc. of IEEE CIBIB 2004, San Diego, Oct. 7-8, 2004.

[117] H. Toivonen, “Sampling large databases for association rules,” In Proc. of 1996 Int. Conf. on Very Large Databases (VLDB’96), pages 134-145, Bombay, India, Sept. 1996.

[118] S.L. Wang, J.S. Tsai and T.P. Hong, “Mining Functional Dependencies from fuzzy relational databases,” ACM SAC 2000, Como, pp. 490-493, Italy, March 2000.

[119] X.-Z. Wang, D.S. Yeung and E.C.C. Tsang, “A comparative study on heuristic algorithms for generating fuzzy decision trees,” IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume 31, Issue 2, pp. 215-226, Apr. 2001.

[120] K.Wang, S. Zhou, and Y. He, “Growing decision trees on support-less association rules,” In Proc. of the sixth ACM-SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’00), pages 265-269, Boston, MA, Aug. 2000.

[121] S.S.C. Wong and S. Pal, “Mining fuzzy association rules for web access case adaptation,” Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case-Based Reasoning (ICCBR'01), 2001.

[122] V. Uebele, S. Abe and M.-S. Lan, “A neural-network-based fuzzy classifier,” IEEE Transactions on Systems, Man and Cybernetics, Volume 25, Issue 2, pp. 353-361, Feb. 1995.

[123] V. Vapnik, Statistical Learning Theory, New York, John Wiley and Sons, 1998.

[124] Y.Y. Yao, “On Modeling data mining with granular computing,” Proceedings of COMPSAC 2001, pp.638-643, 2001.

[125] J.T. Yao, Y.Y. Yao, “A granular computing approach to machine learning,” Proceedings of FSKD’02, Singapore, pp732-736, 2002.

[126] X. Yin and J. Han, “CPAR: Classification based on Predictive Association Rules,” Proc. Of SIAM Int. Conf. on Data Mining (SDM'03), pp. 331-335, San Francisco, CA, 2003.

[127] K. Y. Yip, D. W. Cheung and M. K. Ng, “HARP: A Practical Projected Clustering Algorithm,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1387-1397, 2004.

[128] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Computing optimized rectilinear regions for association rules. In Proc. 1997 Int. Conf. on Knowledge Discovery and Data Mining (KDD’97), pages 96-103, Newport Beach, CA, Aug. 1997.

[129] Y. Yuan and M.J. Shaw: “Induction of Fuzzy Decision Trees”, Fuzzy Sets and Systems, 69, 1995, 125-139.

[130] L. Zadeh, “Fuzzy Sets,” Journal of Information and Control, Volume 8, pp 338--353, 1965.

[131] L. Zadeh, “Outline of a New Approach to the Analysis of Complex Systems and Decisions Processes,” IEEE Transactions on Systems, Man and Cybernetics, SMC-3(1), Jan. 1973.

[132] L.A. Zadeh and J. Kacprzyk, “Fuzzy logic for the management of uncertainty,” John Wiley & Sons Inc., 1992