國立交通大學

工業工程與管理學系

博士論文

粒化計算處理不平衡資料之理論與應用

Granular Computing for Imbalanced Data: Theory and Applications

研 究 生：陳隆昇

指導教授：蘇朝墩、李榮貴　教授

中 華 民 國 九 十 五 年 三 月

粒化計算處理不平衡資料之理論與應用

Granular Computing for Imbalanced Data: Theory and Applications

研 究 生：陳隆昇　　Student：Chen, Long-Sheng

指導教授：蘇朝墩　　Advisor：Su, Chao-Ton

李榮貴　　Li, Rong-Kwei

國 立 交 通 大 學

工業工程與管理學系

博士論文

A Dissertation

Submitted to Department of Industrial Engineering and Management

College of Management

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in Industrial Engineering and Management

March 2006

Hsinchu, Taiwan, Republic of China

粒化計算處理不平衡資料之理論與應用

學生：陳隆昇　　指導教授：蘇朝墩
李榮貴

國立交通大學工業工程與管理學系博士班

摘　　要

近年來機器學習的發展為分類問題提供一項有效的工具。然而，當從不平衡資料(imbalanced data)學習時，傳統的方法在預測少數範例(minor examples)上，其能力是不足的。這類的問題相當重要，在許多環境、生命相關或商業重要領域中大量發生，譬如詐騙偵測、文字探勘、垃圾信件偵測、醫療診斷、錯誤監視及檢測等。在本論文中，我們提出稱為「粒化計算」(Granular Computing)的新穎方法來解決這種「類別不均問題」(Class Imbalance Problems)。

粒化計算以表示和處理資訊粒(Information Granule)為導向，是一種模仿人類資訊處理本能的計算模式，逐漸在資訊科學、邏輯、哲學等領域中成為一項重要的議題。當描述一個包含不完整、不確定或是模糊資訊的問題時，人類很難去考慮詳細的數值資料，而被迫考慮『資訊粒』—是由個別元素(individual elements)依據其相似性、功能接近性或是不可分辨度所構成的集合。粒化計算的模型不僅可以移除不必要的細節、使我們看清資料的本質，更能有效地用來解決『類別不均問題』。

本研究的目的在於發展出兩種粒化計算模型—「KAIG」與「IG based method」分別處理離散型(discrete)與連續型(continuous)資料。兩個模型中，兩種指標—H-index 與 U-ratio，被成功地導入以用來確定適當的顆粒性水準(level of granularity)，換言之，我們可以據此來確定適當的資訊粒數目。模糊適應共振理

論網路(Fuzzy ART neural network)被用來建構資訊粒。此外，在「KAIG」模型中，我們提出了「附屬屬性(sub-attributes)」的觀念來描述資訊粒並可解決資訊粒彼此重疊的現象。在「IG based method」方法中，我們則是以資料特性來表示資訊粒。本研究的主要目標詳述如次：

(1)發展 KAIG 模型來建構資訊粒，並從其中攫取知識。七個 UCI 資料銀行中的資料(包含一個不平衡診斷資料)，被用來評估 KAIG 模型的有效性，在使用不同的績效指標(如 *Overall Accuracy*, *G-mean* 和 *ROC curve*)評估下，相較於決策樹方法(decision tree, C4.5)與支持向量分類器(Support Vector Machine)，實驗結果說明了我們所提方法的優異性。

(2)應用 KAIG 模型解決工業工程相關領域中的「類別不均問題」。首先，在模擬的彈性製造系統(Flexible Manufacturing Systems)環境中，KAIG 模型被應用來改善動態排程系統的分類績效。其次，我們以一個手機檢測的實際案例來說明 KAIG 模型有極優異的能力偵測出極少數的不良品。此外，KAIG 模型可以減少多餘的測試項目並縮短檢驗時程。這兩個應用實例證實對於處理不平衡資料，KAIG 模型可以大幅提昇偵測少數範例的能力 (Negative Accuracy)，同時又不會減少整體的分類準確率(Overall Accuracy)。

(3)提出「IG based method」來處理連續型的不平衡資料。在這個方法中，不同的資料特性及其組合被用來表示建構好的資訊粒，然後再利用這些資訊粒的代表來建立分類器。一個糖尿病醫療診斷實例被用來評估所提方法的有效性。相較於傳統的方法，本研究所提的方法在不平衡資料的學習上表現出極佳的結果。

***關鍵字*: 粒化計算、資訊粒化、類別不均問題、模糊適應共振理論網路、知識攫取、機器學習**

Granular Computing for Imbalanced Data: Theory and Applications

student：Chen, Long-Sheng　　Advisors：Dr. Su, Chao-Ton
Dr. Li, Rong-Kwei

Department of Industrial Engineering and Management
National Chiao Tung University

ABSTRACT

In recent years, the development of machine learning techniques has provided an effective avenue for classification problems. However, when learning from imbalanced data, the traditional methods have poor predictive ability to identify minority instances. This problem is of crucial importance since it is encountered by a large number of domains of great environmental, vital or commercial importance such as fraud detection, text mining, spam detection, medical diagnosis and fault monitoring/inspection. In this study, we propose novel methods called “Granular Computing” models to tackle class imbalance problems.

Granular computing, which is oriented towards representing and processing Information Granules (IGs), is a computing paradigm that embraces a number of modeling frameworks. GrC imitates human instincts of processing information and is becoming a very important issue for computer science, logic, philosophy and others. When describing a problem which involves incomplete, uncertain, or vague information, we human beings tend to shy away from numbers and use aggregates to ponder the question instead. We are forced to consider IGs which are collections of entities arranged together due to their similarity, functional adjacency and indistinguishability. GrC model not only can remove unnecessary details and provide a better insight into the essence of data, but also effectively solve *class imbalance problems*.

This study aims to develop two kinds of GrC models, “Knowledge Acquisition via Information Granulation” (KAIG) model and “Information Granules based method” (IG based method), for dealing with discrete and continuous data, respectively. In both models, the homogeneity index (*H-index*) and the

undistinguishable ratio (*U-ratio*) are successfully introduced to determine a suitable level of granularity (i.e. determine suitable number of IGs). Fuzzy Adaptive Resonance Theory (Fuzzy ART) neural network is utilized to construct IGs. In addition, we propose the concept of "sub-attributes" to describe granules and tackle the overlapping among granules in KAIG model. In IG based method, data characteristics are employed to represent IGs. The main objectives of this study are:

1. Develop a KAIG model to construct IGs, and to discover knowledge from IGs. Seven data sets from UCI data bank (including one imbalanced diagnosis data), are provided to evaluate the effectiveness of KAIG model. By using different performance indexes, *Overall Accuracy*, *G-mean* and *ROC curve*, the experimental results comparing with *C4.5* and *Support Vector Machine (SVM)* demonstrate the superiority of our method.

2. Apply KAIG model to solve class imbalance problems in industrial engineering related areas. First, KAIG model is utilized to improve the classification performance of a dynamic scheduling system within a simulated Flexible Manufacturing System environment. Second, a real case of cellular phones inspection is provided to illustrate the excellent ability of KAIG model in identifying rare defective products. In addition, KAIG model can reduce redundant test items and shorten inspection time. For imbalanced data, these applications show KAIG model can dramatically increase *Negative Accuracy* (the capability of detecting minor instances) without losing *Overall Accuracy*.

3. Propose IG based method to deal with continuous imbalanced data. In this method, different data characteristics and their combinations are employed to denote constructed IGs. Then we build a classifier from these representatives of IGs. An actual medical diagnosis data of diabetes is used to evaluate the effectiveness of this method. Compared with traditional techniques, the proposed method is shown to be superior for learning on imbalanced data.

***Key words*:** Granular computing, Information granulation, Class imbalance problems, Fuzzy ART neural network, Knowledge acquisition, Machine learning.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

誌　　謝

首先要感謝的人，是我的指導教授－蘇朝墩老師。雖然他總是說：「我沒幫上甚麼忙」，但我心裡知道，從論文撰寫、投稿、國際會議參與、六標準差認證到千里馬計畫的申請；從寫作、報告技巧指導，生涯規劃的建議到人生經驗的傳承與分享，沒有他的指導，我無法順利取得學位。從他身上，我學到的是一份堅毅、努力不懈的研究態度，引領我邁向人生的康莊大道。口試期間，承蒙口試委員李榮貴教授、洪瑞雲教授、清華大學溫于平教授、大葉大學駱景堯教授及北科大陳穆臻教授，不吝給予指教，使得本論文更臻完整。同時我也要感謝李榮貴和洪瑞雲老師的幫忙，讓我能夠順利完成學業。此外，在普渡大學(Purdue University)訪問期間，工業工程學院的易玥文(Y. Yih)教授的細心指導，及朝龍、國浩、宗翰、Sandra、Steven 及其他台灣留學生在生活上的幫忙，在此也一併致上萬分謝意。

三年六個多月的時間是苦悶的，沒有學長許志華、許俊欽、劉正祥及同門楊健炘、黃榮輝、楊宗銘、林敬森、周家任、蕭宇翔、彭加景的陪伴，相互的服持與鼓勵，我很難撐到此刻。謝謝你們為我做的一切，希望你們都能平安順利。

最後，我要將這本論文獻給我的雙親及家人，沒有你們在背後默默地支持與鼓勵，就沒有今天的我。特別是我的妻子－林昕怡小姐，在我最艱難的時刻，他沒有對我失去信心，讓我有勇氣可以堅持下去。爾後，我會竭盡所能地貢獻一已微薄之力於社會，將這份感謝傳遞下去。

陳隆昇　Mar. 16, 2006

CHAPTER 1

INTRODUCTION

1.1 Research Motivations

When learning from imbalanced/skewed data, which almost all the instances are labeled as one class while far few instances are labeled as the other class, traditional machine learning algorithms such as *Neural Networks* (NN), *Decision Trees* (DT), and *Support Vector Machines* (SVM) tend to produce high accuracy over the majority class but poor predictive accuracy over the minority class. This minority class is usually the important one, like illness patients of medical diagnoses examples or abnormal products of finished-goods inspection data. This study tries to solve these *Class Imbalance* problems which caused by skewed data distribution.

There are two motivations why we propose the Granular Computing (GrC) to tackle class imbalance problems. The first one is human instinct (Zadeh, 2001). As human beings, we have developed a granular view of the world. When describing a problem or making decisions, we tend to shy away from numbers and use aggregates to ponder the question instead. This is especially true when a problem involves incomplete, uncertain, or vague information. It may be sometimes difficult to differentiate distinct elements, and so one is forced to consider “information granules” (IG) which are collections of entities arranged together due to their similarity, functional adjacency and indistinguishability (Bargiela and Pedrycz, 2003; Castellano and Fanelli, 2001; Yao and Yao, 2002; Zadeh, 1979). A typical example is the theory of rough sets (Walczak and Massart, 1999).

The process of constructing IGs is referred to as information granulation. This was first pointed out in the pioneering work of Zadeh (1979) who coined the term

‘information granulation’, and emphasized the fact that a plethora of details does not necessarily amount to knowledge. Granulation serves as an abstraction mechanism for reducing an entire conceptual burden. The essential factor driving the granulation of information is the need to comprehend the problem and have a better insight into its essence, rather than get buried in all the unnecessary details. By changing the size of the IGs, we can hide or reveal more or less details (Bargiela and Pedrycz, 2003). Granular Computing (GrC) is oriented towards the representation and processing of IGs.

The second motivation is about the behavior of data. In many practical datasets, such as medical/diagnosis, inspection, fault monitoring and fraud detecting data, the normal group and abnormal group are considered separate populations. Taguchi and Juoulum (2002) thought every abnormal condition (or a condition outside “healthy” group) is considered unique, since the occurrence of such a condition is different. Tolstoy’s quote in *Anna Karenina*: “*All happy families look alike. Every unhappy family is unhappy after its own fashion”* is also noted to illustrate their opinions (Taguchi and Juoulum, 2002). From the observations of practical data, we can clearly find the normal group (i.e. healthy patients, good products) look alike while the abnormal group (i.e. sick patients, defective products) are unique. If we construct IGs by the similarity of numerical data, the amount of IGs in normal group will be remarkably smaller than the size of normal numerical data. In other words, if we consider IGs instead of numerical data, it might increase the proportion of abnormal data and improve imbalanced/skewed situation of data.

1.2 Research Objectives

The purpose of this study is to develop two Granular Computing models to deal with imbalanced/skewed data. These two models can extract knowledge from IGs and are developed for discrete and continuous data, respectively. The main issue needed to tackle is how to measure and represent IGs if we want to acquire knowledge from IGs. In this study, we use Fuzzy Adaptive Resonance Theory (Fuzzy ART) neural network to construct IGs. The two indexes, the homogeneity index (H-index) and the undistinguishable ratio (U-ratio), are presented to measure IGs. In the first proposed model called "Knowledge Acquisition via Information Granulation" (KAIG), the concept of "sub-attributes" is presented to describe granules, and to tackle the overlapping among granules. In the second proposed approach called "Information Granules based method", we try to use different data characteristics such as mean, median, quartiles, minimum, maximum and combinations of them to represent IGs. Then, we extract knowledge from these IGs.

The KAIG model is designed for discrete imbalanced data. We will evaluate KAIG model by using UCI data and make a comparison between KAIG model and traditional knowledge acquisition algorithms which operate with numerical data. In addition, KAIG model is applied to deal with class imbalance problems in a dynamic scheduling problem within a simulated Flexible Manufacturing System. Besides, this study developed a feature selection procedure integrated the proposed KAIG model to find key test items and shorten inspection time. A real case of mobile phone inspection in Taiwan was used to evaluate effectiveness of the proposed procedure. We also show advantages and benefits of this procedure.

The IG based method is proposed to deal with continuous imbalanced data. The experimental results will be compared with cluster-based sampling method and

original machine learning techniques. Finally, an actual medical diagnosis data of diabetes is employed to illustrate the superiority of our method.

1.3 Framework and Organization

In practical application of machine learning (or data mining), class imbalance problems are emerging issues. According the report of available researches, sampling and moving decision threshold are widely used methods to tackle this problem. This study developed two kinds of GrC model, which is also new topic in information processing, to solve the class imbalance problems. The developed models will be compared with previous techniques. The research framework is shown in Figure 1.1.

This study is organized as follows. Chapter 1 presents the research motivations and objectives. Chapter 2 is the literature review of related researches toward granular computing, class imbalance problems, inductive learning and feature selection techniques. Chapter 3 proposes two GrC methodologies. In this chapter, we use Fuzzy ART neural network to construct IGs, present “H-index & U-ratio” to determine the suitable level of granularity, and develop the concept of “sub-attributes” and “data characteristics” to describe IGs. In chapter 4, several data sets from UCI machine learning group are provided to illustrate and evaluate the effectiveness of our methodologies. Chapter 5 describes the applications of KAIG model in dynamic scheduling system within a simulated FMS. In Chapter 6, we develop a KAIG model based feature selection procedure to reduce test items and shorten inspection time in mobile phone manufacturing. Chapter 7 provides a case study of diabetes diagnosis by using IG based method. Finally, conclusions and future works are described in Chapter 8.

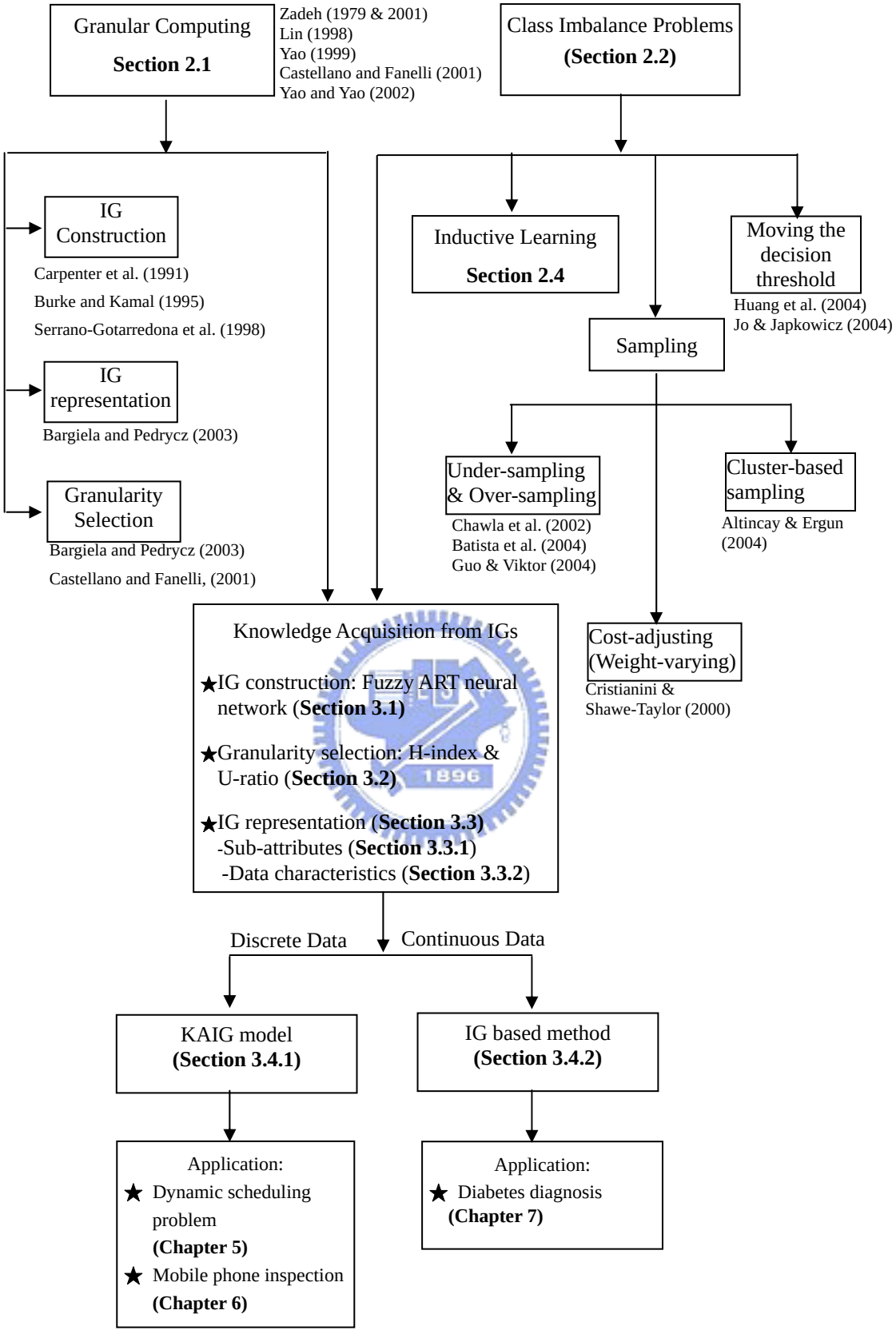


Figure 1.1 Research framework

CHAPTER 2

RELATED WORKS

2.1 Granular Computing

Humans have a remarkable capability to perform a wide variety of physical and mental tasks without any measurements and any computations, such as driving, parking, cooking, and playing computer game. We human beings use perceptions of direction, speed, time and other attributes of physical/mental objects, instead of numerical data. Basically speaking, reflecting the limited ability of human brains, perceptions are inaccurate. In more concrete terms, perceptions are granular. It means that the boundaries of perceived classes are unsharp; and the values of attributes are granulated (Zadeh, 2001). For example, the granules of temperature might be labeled very cold, cold, warm, hot, very hot, etc. The computation theory of perceptions (CTP) is inspired by the marvelous human ability. And, GrC belongs to related research areas of CTP.

GrC is quickly becoming an emerging conceptual and computing paradigm of information processing (Bargiela and Pedrycz, 2003). It is a superset of the theory of fuzzy information granulation, rough set theory and interval computations, and is a subset of granular mathematics. GrC as opposed to numeric computing is knowledge-oriented. Numeric computing is data oriented. The main issues (Castellano and Fanelli, 2001) of granular computing are how to construct the IGs, and to describe IGs. One particular question that arises is how to determine the level of granularity. If we want to acquire knowledge from IGs, we must try to solve these three questions which will be discussed in sections 3.1~3.3.

In the issue of constructing IGs, there are many approaches, such as the Self

Organizing Map (SOM) network (Bortolan and Pedrycz, 2002), Fuzzy C-means (FCM) (Castellano and Fanelli, 2001; Bargiela and Pedrycz, 2003b), rough sets, shadowed sets (Bargiela and Pedrycz, 2003a) used to do this. Because IGs exist at different levels of granularity, we usually group granules of similar “size” (that is granularity) in a single layer. If more detailed processing is required, smaller IGs are selected. Figure 2.1 illustrates this concept of granularity. At the lowest level, we are concerned with numeric processing. This is a domain completely taken over by numeric models, such as differential equations, regression models, neural networks, etc. At the intermediate level, we see larger IGs (viz. those embracing more individual elements). The top level is solely devoted to symbol-based processing, and as such invokes well-known concepts of Petri nets, qualitative simulation, etc (Bargiela and Pedrycz, 2003a).

In the issue of represntting IGs and determining the level of granularity, Bargiela and Pedrycz (2002) proposed the “hyperbox” and “inclusion & compatibility” to measure IGs. However, these researches focused on how to construct IG, how to describe IG and how to measure IG, individually. We need an advanced/integrated mechanism to imitate human ability of processing information, such as extracting knowledge from IGs and making decision based on them.

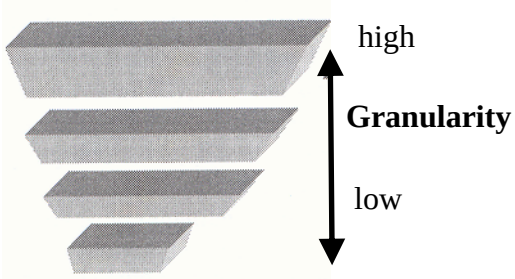


Figure 2.1 An information-processing pyramid (Bargiela & Pedrycz, 2003)

2.2 Class Imbalance Problems

Learning from imbalanced/skewed data is an important topic and rises very often in practice. In such kind of data, one class might be represented by a large number of examples while the other is represented by only a few. Many real world data have these characteristics, such as fraud detection, text classification (Chawla et al., 2002& 2004) telecommunications management, oil spill detection, risk management, medical diagnosis/monitoring**,** financial analysis of loan policy or bankruptcy (Batista et al., 2004; Chawla et al., 2004; Grzymala-Busse et al., 2004) and protein data (Provost and Fawcett, 2001)**.** Traditional classifiers seeking an accurate performance over a full range of instances are not suitable to deal with imbalanced learning tasks **(**Batista et al., 2004; Chawla et al., 2004; Guo and Viktor, 2004; Japkowicz and Stephen, 2002**)** since they tend to classify all data into the majority class, which is usually the less important class. Therefore, these traditional algorithms often produce high accuracy over the majority class, but poor predictive accuracy over the minority class.

To cope with imbalanced data sets, there are some methods proposed in literatures. There are two major groups of techniques designed to address class imbalance. The first group consists of supervised techniques that usually include five approaches: (1) *undersampling*, methods in which the minority population is kept intact, while the majority population is under-sampled, (2) *oversampling*, methods in which the minority examples are over-sampled so that the desired class distribution is obtained in the training set (Batista et al., 2004; Chawla et al., 2002; Guo and Viktor, 2004), (3) *cluster based sampling*, methods in which the representative examples are randomly sampled from clusters (Altincay and Ergun, 2004), (4) *moving the decision threshold*, methods in which researchers try to adapt the decision thresholds to impose bias on the minority class (Chawla et al., 2002; Huang et al., 2004; Jo and Japkowicz,

2004) and (5) *adjust costs matrices*, methods in which the prediction accuracy is improved by adjusting the cost (weight) for each class(Cristianini and Shawe-Taylor, 2000).

The second large class of techniques for detecting rare events involves an unsupervised framework, i.e. outlier detection or one-class classification (Manevitz and Yousef, 2001). Initially, minority examples are completely ignored and a model is trained by using all examples from the majority class (target class). Then, the outliers are detected as the data points with low probability of occurrence, small number of neighboring examples. In addition, SVM is usually used to tackle class imbalance problem (Wu and Chang, 2005).

These techniques have some disadvantages (Altincay and Ergun, 2004). For example, the computational load is increased and overtraining may occur due to the replicated samples in the case of over-sampling. Under-sampling does not take into account all available training data which corresponds to loss of available information. Huang et al. (2004) thought these supervised methods lack a rigorous and systematic treatment on imbalanced data. Moreover, the one-class classification methods only consider majority examples. It might miss some beneficial decision information of minority examples.

2.3 Fuzzy ART Neural Network

Fuzzy ART is one of clustering techniques and also the most recent adaptive resonance framework that provides a unified architecture for both binary and continuous valued inputs. Fuzzy ART clusters vectors based on two separate distance criteria, match and choice. For input vector I and category j, the match function is defined by

$$S_j(I) = \frac{|I \wedge w_j|}{|I|} \tag{2.1}$$

where w_j is an analog-valued weight vector associated with cluster *j*. $\wedge$ denotes the fuzzy AND operator, $(p \wedge q)_i = \min(p_i, q_i)$, and the norm $|....|$ is defined by $|p| = \sum_i |p_i|$.

The choice function is defined by

$$T_j(I) = \frac{|I \wedge w_j|}{\alpha + |w_j|} \tag{2.2}$$

where α is a small constant. Increasing α biases the search more towards clusters with large w_j. Each input vector is assigned to the category that maximizes $T_j(I)$ while satisfying $S_j(I) \geq \rho$, where the vigilance ρ, is a constant, $0 \leq \rho \leq 1$. The topological structure of the Fuzzy ART architecture is shown in Figure 2.2.

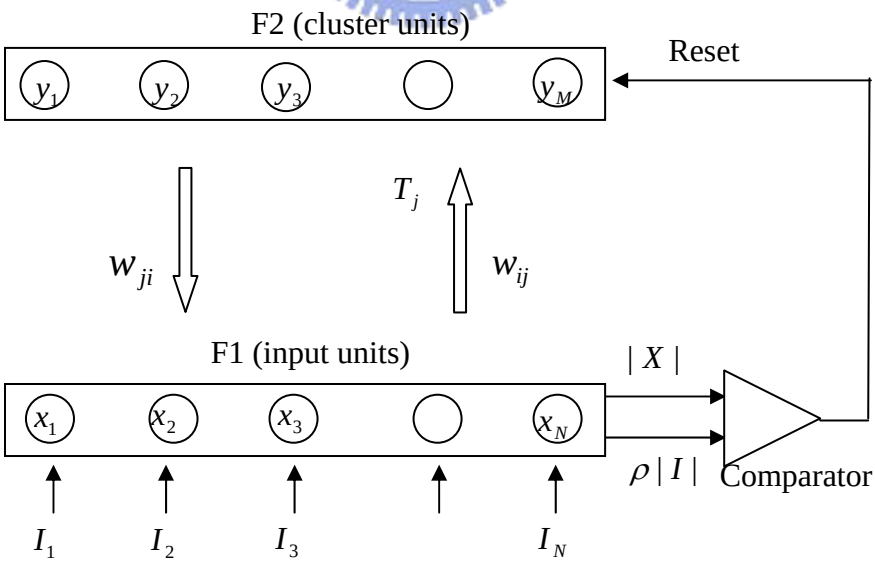


Figure 2.2 Topological structure of the Fuzzy-ART

2.4 Inductive Learning Methods

2.4.1 Decision Tree

The decision tree method is one of the most popular knowledge acquisition algorithms, and has been successfully applied in many areas. Decision tree algorithms, such as ID3 and C4.5, were originally intended for classification purposes. The core of C4.5 contains recursive partitioning of the training examples. Whenever a node is added to a tree, some subsets of the input features are used to pick the logical test at that node. The feature that results in the maximum information gain is selected for testing at that node. In other words, the algorithm chooses the “best” attribute to partition the data into individual classes at each node. After the test has been determined it is used to partition the examples, and the process is continued recursively until each subset contains examples of one class or satisfies some statistical criteria (Su and Shiue, 2003).

In this study, See5 (C4.5 commercial version) software was utilized to construct a decision tree. In See5 there are two parameters that can be tuned during the pruning phase: the minimal number of examples represented at any branch of any feature-value test; and the confidence level of pruning. In order to avoid the occurrence of overfitting and generating a simple tree, 2 was set as the minimum number of instances at each leaf, and the confidence level for pruning was set at 25%.

2.4.2 Back-propagation Neural Network

Neural nets have been used widely in pattern recognition, function approximation, optimization, and clustering. Generally speaking, neural nets can be classified into two categories, feed-forward and feedback networks. In this study, the

feed-forward network, shown as Figure 2.3, was employed because of their superior ability of classification.

The back-propagation learning algorithm (Rumelhart & McClelland, 1986) is the best known training algorithm for neural networks and still one of the most useful. This iterative gradient algorithm is designed to minimize the mean square error between the actual output of a multilayer feed-forward perceptron and the desired output. According to the rule of thumb and reports of available published papers, the number of hidden layers should be one or two. The back-propagation algorithm includes a forward pass and a backward pass. The purpose of the forward pass is to obtain the activation value and the backward pass is to adjust weights and biases according to the difference between the desired and actual network outputs. These two passes will go through iteratively until the network converges. The feed-forward network training by back-propagation can be summarized as the following steps:

Step 1: Select an architecture

Step 2: Randomly initialize weights

Step 3: While error is too large

For each training pattern (presented in random order)

Step 3.1: Select training pattern and feedforward to find actual network output

Step 3.1.1: Apply the inputs to the network

Step 3.1.2: Calculate the output for every neuron from the input layer, through the hidden layer(s), to the output layer

The output from neuron j for pattern p is O_{pj} where

$$O_{pj}(net_j) = \frac{1}{1+e^{-net_j}} \tag{2.3}$$

and

$$net_j = bias + \sum_k O_{pk} W_{jk} \tag{2.4}$$

k ranges over the input indices and W_{jk} is the

weight on the connection from input *k* to neuron *j*.

Step 3.2: Calculate errors and backpropagate error signals

Step 3.2.1: Calculate the error at the outputs

The output neuron error signal δ_{pj} is given by

$$\delta_{pj} = (T_{pj} - O_{pj}) \times O_{pj} \times (1 - O_{pj}) \tag{2.5}$$

where T_{pj} is the target value of output neuron *j* for pattern *p* and O_{pj} is the actual output value of output neuron *j* for pattern *p*.

Step 3.2.2: Use the output error to compute error signals for pre-output layers

The hidden neuron error signal δ_{pj} is given by

$$\delta_{pj} = O_{pj}(1 - O_{pj}) \sum_{k} \delta_{pk} W_{kj}) \tag{2.6}$$

where δ_{pk} is the error signal of a post-synaptic neuron *k* and W_{kj} is the weight of the connection from hidden neuron *j* to the post-synaptic neuron *k*.

Step 3.3: Adjust weights

Step 3.3.1: Use the error signals to compute weight adjustments

Compute weight adjustments ΔW_{ji} at time t by

$$\Delta W_{ji}(t) = \eta \times \delta_{pj} \times O_{pi} + \alpha \times \Delta W_{ji}(t - 1) \tag{2.7}$$

where η is the learning rate and α is the momentum coefficient ($\alpha \in [0,1]$).

Step 3.3.2: Apply the weight adjustments

Apply weight adjustments according to

$$W_{ji}(t + 1) = W_{ji}(t) + \Delta W_{ji}(t) \tag{2.8}$$

Step 4: Evaluate performance using the test data set

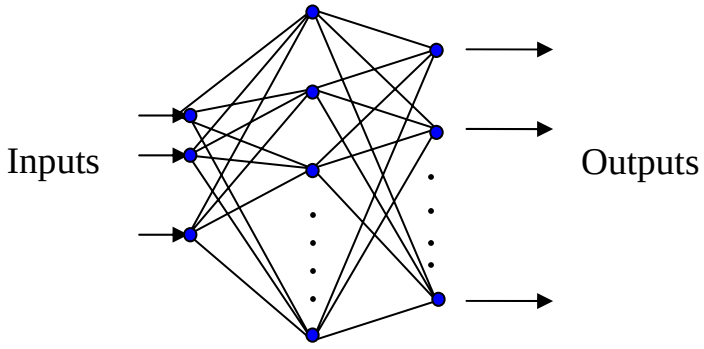


Figure 2.3 The back-propagation neural network structure

2.4.3 Rough Sets

The rough sets theory was introduced by Pawlak (1985) to deal with imprecise or vague concepts (Swiniarski and Skowron, 2003; Walczak and Massart, 1999). Rough sets deal with information represented by a table called the *information system* which contains objects and attributes. An information system is composed of a 4-tuple as follows:

$$S = \langle U, Q, V, f \rangle \tag{2.9}$$

where U is the universe, a finite set of N objects $\{x_1, x_2, \ldots x_N\}$, Q is a finite set of attributes, $V = \cup_{q \in Q} V_q$, where V_q is a value of attribute q, and $f : U \times Q \to V$ is the total decision function called the *information function* such that $f(x, q) \in V_q$ for every $q \in Q$, $x \in U$. For a given subset of attributes $A \subseteq Q$ the IND(A)

$$IND(A) = \{(x, y) \in U : \text{for all a} \in \text{A}, f(\text{x}, \text{a}) = f(\text{y}, \text{a})\} \tag{2.10}$$

is an *equivalence relation* on universe U (called an *indiscernibility relation*).

Some of the information systems can be designed as a *decision table*

$$Decision\ table = \langle \mathrm{U}, \mathrm{C} \cup \mathrm{D}, \mathrm{V}, f \rangle \tag{2.11}$$

where C is a set of condition attributes, D is a set of decision attributes, $V = \cup_{q \in C \cup D} V_q$, where V_q is the set of values of attribute $q \in Q$, and $f : U \times (C \cup D) \to V$ is a total *decision function* (decision rule in a decision table) such that $f(x, q) \in V_q$ for every $q \in Q$ and $x \in V$.

For a given information system *S*, a given subset of attributes $A \subseteq Q$ determines the approximation space $AS = (U, IND(A))$ in *S*. For a given $A \subseteq Q$ and $X \subseteq U$ (a concept of X), the A-*lower approximation* $\underline{A}X$ of set X in *AS* and A-*upper approximation* $\overline{A}X$ of set X in *AS* are defined as follows:

$$\underline{A}X = \{x \in U : [x]_A \subset X\} = \cup\{Y \in A^* : Y \subseteq X\}, \tag{2.12}$$

$$\overline{A}X = \{x \in U : [x]_A \cap X \neq \varnothing\} = \cup\{Y \in A^* : Y \cap X \neq \varnothing\} \tag{2.13}$$

where A^* denotes the set of all equivalence classes of IND(A). The process of finding a set of attributes smaller than the original one with the same classificatory power as the original set is called *attribute reduction*. A *reduct* is the essential part of an information system (subset of attributes) which can discern all objects discernible by the original information system. By means of the dependent properties of the attributes we can find a reduced set of attributes, providing that by removing the superfluous attributes there is no loss in classification accuracy.

2.4.4 Support Vector Machines

SVM is a powerful learning method and often employed to tackle class imbalance problems (Wu and Chang, 2005). SVM learns a decision boundary between two classes by mapping the training data (through kernel functions) onto a

higher dimensional space, and then finding the maximal margin hyperplane within that space. Finally, this hyperplane can be viewed as a classifier. Figure 2.4 illustrates the concept of feature mapping and two-class separation.

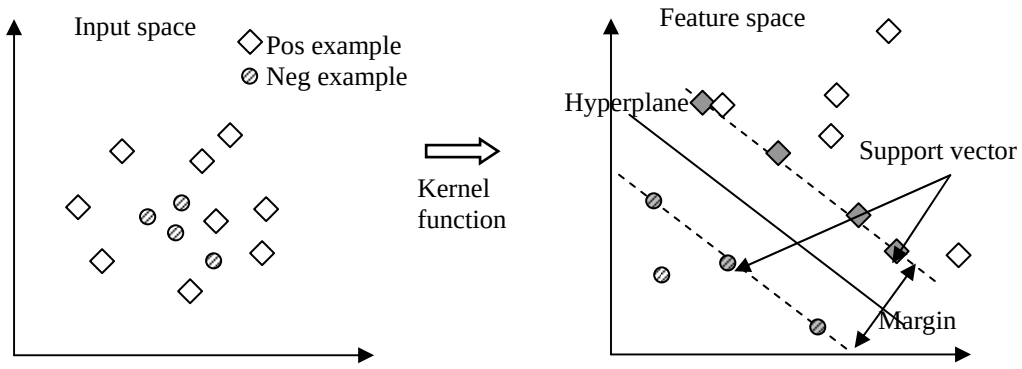


Fig. 2.4 The operations of Support Vector Machine

Consider a classifier, which uses a hyperplane to separate two class of patterns based on given examples $S = \{x_i, y_i\}_{i=1}^{n}$, $y_i \in \{-1,+1\}$. The hyperplane is defined by (w,b), where w is a weight vector and b a bias. Let w_0 and b_0 denote the optimal values of the weight vector and bias. Correspondingly, the optimal hyperplane can be written as

$$w_0^T x + b_0 = 0 \tag{2.14}$$

To find the optimum values of w and b, it requires to solve the following optimization problem.

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ \text{Subject to} \quad & y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.15}$$

where ξ is the slack variables, C is the user-specified penalty parameter of the error term ($C > 0$), and ϕ is the kernel function.

In this research, we used the LIBSVM (version 2.8), which is an integrated tool

for support vector classification and regression, and is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. We used the standard parameters of the algorithm. All optimal parameters can be automatically generated in this program and the default kernel function is Radial Basis Function (RBF).

2.5 Feature Selection From Imbalanced Data

Reduction of pattern dimensionality via feature selection belongs to the most fundamental steps in data processing (Swiniarski and Hargis, 2001). A large feature set often contains redundant and irrelevant information, and can actually degrade the performance of the classifier (Oyeleye and Lehtihet, 1998). The main purpose of feature selection is to remove irrelevant or redundant attributes and improve the performance of classification.

Feature selection is often applied in pattern classification, data mining, as well as in machine learning. Among many feature selection methods, GA, rough sets and neural networks have attracted much attention, and have become popular techniques for feature selection. However, when these methods are applied to imbalanced data, it usually suffers from some drawbacks, such as ignoring the minority examples and viewing them as outliers. It was reported (Batista et al., 2004; Chawla et al., 2004) that these methods seeking an accurate performance over a full range of instances are not suitable to deal with imbalanced learning tasks since they tend to classify all data into the majority class, which is usually the less important class. This is because typical classifiers that are designed to optimize overall accuracy without taking into account the relative distribution of each class.

Rough sets emerged as a major mathematical tool for discovering knowledge and feature selection (Walczak, B. and D. L. Massart, 1999). One of the fundamental

principles of a rough set-based learning system is discovering redundancies and dependencies between the given features of a problem to be classified. A reduct generated by the rough sets approach is defined as the minimal subset of attributes that enables the same classification of objects with full attributes. When applying rough sets in practice, its computational complexity increases dramatically with the growth of the data. In addition, the deterministic mechanism for the description of error is very simple in rough sets. Therefore, the rules generated by rough sets are often unstable and have a low classification accuracy (Li and Wang, 2004).

Feature selection with neural networks can be thought of as a special case of architectural pruning (Reed, 1993), where the input features are pruned rather than the hidden neurons. Su et al. (2002) attempted to determine the important input nodes of a neural network based on the sum of absolute multiplication values of the weights between the layers. They (Su et al., 2002) proposed an algorithm to remove unimportant input nodes from a trained back-propagation neural network (BPNN). The essence of this method is to compare the multiplication values of the weights between the input-hidden layer and the hidden-output layer. Only the multiplication weights with large absolute values are kept and the rests are removed. The equation for calculating the sum of absolute multiplication values is defined as follows.

$$Node_i = \sum_j \left| W_{ij} \times V_{jk} \right| \tag{2.14}$$

where W_{ij} is the weight between the *i*th input node and the *j*th hidden node, and V_{jk} is the weight between the *j*th hidden node and the *k*th output node. Then, we must set a threshold to remove the irrelevant input nodes. The threshold should be determined by the user to obtain a suitable number of input nodes. Unfortunately, the training of neural networks when using imbalanced data is very slow (Bruzzone and Serpico, 1997).

Another common understanding is that some learning algorithms have built-in feature selection, for example, ID3 (Quinlan, 1986), FRINGE and C4.5 (Quinlan, 1993). Almuallim and Dietterich (1994) suggested that one should not rely on ID3 or FRINGE to filter out irrelevant features. There are some cases in which ID3 and FRINGE miss extremely simple hypotheses. In addition, the negative examples of imbalanced data might be removed in the pruning phase of the tree construction.

In other words, when faced with imbalanced data, the performance of feature selection tools drops significantly (Akbani et al., 2004). Pendharkar et al. (1999) mentioned that the ratio of the number of objects belonging to positive and negative examples impacts upon effective learning. If the data set contains many positive examples and very few negative examples, there is a bias in the discriminant function that the technique will identify, and it therefore follows that this bias results in a lower reliability of the technique. An and Wang (2001) suggested to balance the data by sampling. However, this is sometimes not feasible due to there being so few negative examples.

CHAPTER 3

PROPOSED GRANULAR COMPUTING MODELS

In this chapter, we propose two kinds of GrC model, "*Knowledge Acquisition via Information Granulation*" (KAIG) model and IG based method, to tackle class imbalance problems. The KAIG model is suitable for dealing with discrete data and the IG based method is designed for continuous data. These two approaches can improve classification performance by controlling the reduction of unnecessary details.

In both of proposed models, Fuzzy ART (Adaptive resonance theory) neural network is utilized to construct IGs. The two indexes, the homogeneity index (H-index) and the undistinguishable ratio (U-ratio), are developed to determine a suitable level of granularity. In KAIG model, the concept of "sub-attributes" is presented to describe IGs and tackle the overlapping among granules. In IG based method, we propose three strategies which utilize different data characteristics and their combinations to represent IGs.

3.1 Construction of Information Granules

In this study, the Fuzzy ART is utilized to construct IGs. ART is a well established neural network theory developed by Carpenter et al. (1991). The ART network is also a famous method of clustering. Instead of clustering by a given number of clusters, it assigns patterns onto the same cluster by comparing their similarity. The detailed algorithm of Fuzzy ART can be found in (Serrano-Gotarredona et al., 1998).

The major difference between ART and other unsupervised neural networks is the so called vigilance parameter (ρ) which is viewed as a granularity and can be adjusted by the users to control the degree of similarity of patterns placed on the same cluster. In an ART, the degree of similarity between a new pattern and a stored pattern is defined. This similarity, compared to ρ, is a measure to ensure whether the new pattern is properly classified or not. The other unsupervised learning neural networks which do not implement vigilance may cause a significantly different input pattern to be forced into an inappropriate cluster. In contrast to some other cluster methods, an ART network will not automatically force all input vectors onto a cluster if they are not sufficiently similar. This is the reason why the ART network is employed in this study to construct the IGs.

There are three similar ART architectures, namely ART 1, ART 2, and Fuzzy ART. ART 1 is designed for binary-valued input patterns, and ART 2 is for continuous-valued patterns. Fuzzy ART is the most recent adaptive resonance framework that provides a unified architecture for both binary and continuous valued inputs. There are several factors that motivated us to use Fuzzy ART, and they are as follows (Burke and Kamal, 1995):

(1)Unlike ART1, Fuzzy ART does not require a completely binary representation of the parts to be grouped. In addition, Fuzzy ART possesses the same desirable stability properties as ART1 and a simpler architecture than that of ART2.

(2)ART2 can experience difficulty in achieving good categorizations if the input patterns are not all normalized to a constant length. However, such normalization can possibly destroy valuable information. Besides, there is a serious dependency of classification results in the case of ART1 on the sequence of input presentation.

As a result, the Fuzzy ART network is employed to construct IGs in this study.

3.2 Selection of Granularity

Selecting an appropriate size of IGs is a difficult task. Enough background knowledge is required to determine how similar objects should be gathered together to form one IG. An objective index is needed to select the appropriate similarity of granules. We propose H-index and U-ratio to solve this problem.

The basic assumption of the H-index is that the classes of objects should be equal if their values of attributes are sufficiently similar. This implies that we always make the same decision under a similar condition. Because we form granules by the similarity of objects, the objects in the same granule should have the same class. The H-index is used to measure the consistency of the class of the objects in one IG. The H-index is defined as

$$H-index=\sum_{m}\frac{i}{n}\Big/m \tag{3.1}$$

where n represents the number of all objects in one granule, m is the number of all IGs and i is the amount of objects possessing the majority class.

For example, Table 3.1 shows one IG involving five objects (n=5). There are 4 condition attributes (namely A, B, C and D) in the iris data. The decision attribute (class) of the first 4 objects is “versicolor”, but the last one has a different decision attribute, “setosa”. In this example, “versicolor” is the majority class and $i=4$. The H-index of this IG is $\frac{4}{5}$.

Table 3.1 The information granule- iris example

Condition attributes				Decision attribute (Classes)
A	B	C	D	
5.8	2.7	4.1	1	versicolor
6.2	2.2	4.5	1.5	versicolor
5.6	2.5	3.9	1.1	versicolor
5.9	3.2	4.8	1.8	versicolor
5	3.3	1.4	0.2	setosa

Table 3.2 The undistinguishable information granule

Condition attributes				Decision attribute
A	B	C	D	
5.4	2.2	3.9	1.2	versicolor
6.8	3.4	5.6	2.4	virginica

Another index for selecting similarity is the U-ratio. In the preceding example, “versicolor” is the majority of the classes. Therefore, it is assigned to be the class of this IG. If there is another granule described as Table 3.2, and we are unable to distinguish the class of the IG, then we call that granule an “undistinguishable granule.” The U-ratio is defined as

$$U - ratio = \frac{u}{m} \tag{3.2}$$

where u represents the number of undistinguishable granules

and m represents the quantity of all granules.

This index is to calculate the proportion of undistinguishable granules to all granules. If there are ten granules and two of them are undistinguishable granules, which means u is equal to 2 and m is equal to 10, then the U-ratio is equal to 0.2.

By using these two indexes, we also need a “granularity selection criteria” to determine the similarity of the IGs. In the present study, the larger the H-index the better it is, because it means that more objects in one IG possess the same class. There is no need to set up the index to a fixed value. The size of the index depends on the domain knowledge or how large an error you can tolerate. On the other hand, the U-ratio is the opposite. As far as the U-ratio is concerned, the smaller the better. It’s difficult to process an undistinguishable granule, so we need to view them carefully. However, we try to avoid this situation by setting the U-ratio as small as possible. In other words, if we select a specific similarity where the H-index is larger and the

U-ratio is smaller, then this similarity is the best solution.

3.3 Representation of Information Granules

3.3.1 The Concept of "Sub-attributes"

In KAIG model, we propose the concept of "sub-attributes" to represent IGs. First, we utilize hyperboxes to represent IGs (Pedrycz and Bargiela, 2002). For example, a hyperbox $[b]$ defined in R^n is fully described by its lower (b^-) and upper corner (b^+), where b^- and b^+ are vectors in R^n. An important and frequently used universal set is the set of all points in the *n*-dimensional space. This set is denoted as R^n. Using b^- and b^+ we can express the hyperbox as $[b]=[b^-,b^+]$. Consider two IGs (hyperboxes) $A=[a]$ and $B=[b]$ defined in R^2. More explicitly, we follow a full notation $[a]=[a^-,a^+]$ and $[b]=[b^-,b^+]$. These two granules are described as Table 3.3.

Table 3.3 Two IGs represented by hyperbox form

IGs \ Attributes	X_1	X_2
A	$\{a_1^-,a_1^+\}$	$\{a_2^-,a_2^+\}$
B	$\{b_1^-,b_1^+\}$	$\{b_2^-,b_2^+\}$

As Figure 3.1 shows, there are overlaps between two granules A and B. This makes it difficult to handle by knowledge acquisition tools. This is because most of knowledge acquisition algorithms are not designed to deal with IGs, especially when overlapping occurs between granules. Unfortunately, the overlapping situation always happens in real world. In this study, we introduce the concept of "sub-attributes" to tackle the problem of overlaps between granules.

We can explain this idea of "sub-attributes" by using Figure 3.1. In axis X1

(attribute 1), the overlapping part of two granules are separated into overlapping part ($[b_1^-, a_1^+]$) and non-overlapping parts ($[a_1^-, b_1^-]$ & $[a_1^+, b_1^+]$). These sub-intervals, $[a_1^-, b_1^-]$, $[b_1^-, a_1^+]$ & $[a_1^+, b_1^+]$, are named as X_{11}, X_{12}, X_{13} which are so called "sub-attributes." The binary variable which is employed to be the values of sub-attributes represents whether an IG contains these sub-intervals or not. The results of rewriting the IGs by using sub-attributes can be found in Table 3.4. We divide the original attribute X_1 into sub-attributes X_{11}, X_{12}, X_{13}; and attribute X_2 into X_{21}, X_{22}, X_{23}. Then, these two granules are rewritten by replacing the original attributes with sub-attributes. By introducing the concept of sub-attributes, we can easily extract knowledge from the granules even if the overlapping situation always exists.

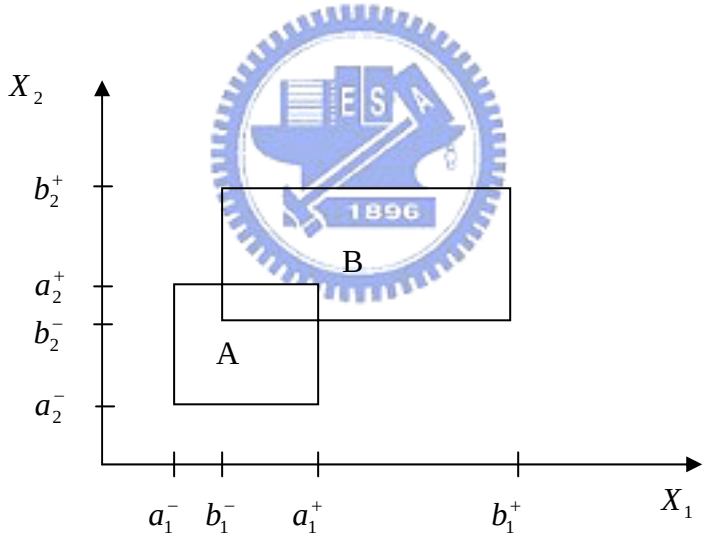


Figure 3.1 The overlap between IGs

Table 3.4 The IGs with sub-attributes

Original attributes	X_1			X_2		
Sub-attributes / IGs	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	X_{23}
	$[a_1^-, b_1^-]$	$[b_1^-, a_1^+]$	$[a_1^+, b_1^+]$	$[a_2^-, b_2^-]$	$[b_2^-, a_2^+]$	$[a_2^+, b_2^+]$
A	1	1	0	1	1	0
B	0	1	1	0	1	1

The concept of “sub-attributes” can maintain the complete characteristics of data. The IGs with addition of sub-attributes are suitable for all knowledge acquisition algorithms. It is not required to adjust the computational architecture of these algorithms. However, too many sub-attributes may be generated in the situation of natural overlapping which the values of the condition attributes are continuous and diverse. Therefore, as we often do in data preparation phase of data mining, we suggest descretizing data before implementing KAIG model to control the number of sub-attributes.

3.3.2 Using Data Characteristics to represent IGs

As mentioned above, too many sub-attributes will increase computational complexity. In order to avoid this situation, we propose another idea which uses data characteristics to describe IGs. Unlike “sub-attributes” which use intervals to represent IG, we utilize different data points such as mean, median, maximum, minimum, and quartiles to describe IGs in IG based method. Three IG representation strategies are provided. In strategy 1, we utilize single value, mean and median, to describe IGs. The strategy 2 uses double-value combinations of data characteristics, Q1+Q3 and Maximum+Minimum. In strategy 3, we employ triple-values combinations, Q1+Median+Q3 and Maximum+Mean+Minimum.

3.4 Proposed Methodologies

This section summarizes the procedure of two proposed GrC models. First, we address how the IGs are formed from numerical data. Secondly, H-index and U-ratio are introduced to determine the level of granularity which can be used to construct IGs in Fuzzy ART. Then, we try to describe IGs and extract knowledge from them.

3.4.1 The KAIG Model

Figure 3.2 shows the proposed KAIG model. We summarized KAIG model by the following steps:

Step 1: Information Granulation

In step 1, we use Fuzzy ART to construct IGs. But, first thing we need to determine is to select the suitable level of granularity (vigilance). The IGs are formed by the selected granularity. The initial value of granularity is set 1 and then decrease gradually until find one satisfying criteria of H-index and U-ratio. The found suitable granularity is employed to construct IGs.

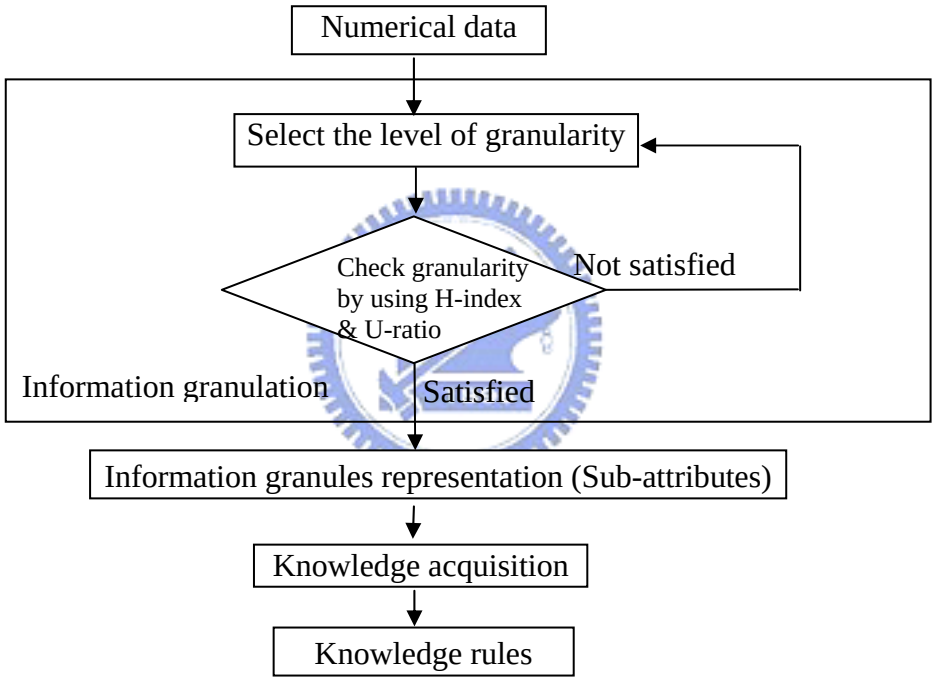


Figure 3.2 Knowledge Acquisition via Information Granulation (KAIG) model

Step 2: Information Granules Representation

IGs are represented in a suitable form that can be handled by knowledge acquisition tools. As mentioned in section 3.2.3, these formed IGs are described in hyperboxes. Then, the sub-attributes are applied in these IGs to solve the problem and finally we can extract knowledge from these IGs.

Step 3: Knowledge Acquisition

After describing IGs appropriately and tackling the overlapping situation, we can

use knowledge acquisition tools to extract knowledge rules from the granules. In this study, we will compare three famous data mining algorithms, C4.5, Rough sets and back-propagation neural network, to evaluate their effectiveness in KAIG model.

3.4.2 The IG based Method

In KAIG model, we use “sub-attributes” to describe IGs and solve the overlapping situation effectively. However, when dealing with continuous data, KAIG may generate so many sub-attributes that increase the computational complexity of knowledge acquisition algorithms. The same situation may occur while the discretization algorithms dividing the continuous attribute’s value into too many discrete intervals. Therefore, we propose the IG based method in this section.

In this method, the “information granulation” process is the same with KAIG model. Only one difference is the description of IGs. This method utilizes data characteristics to denote IGs without using sub-attributes. This IG based method follows the three steps described as bellow. We adopt three strategies which are listed in Step 2 to describe IGs. They are different combinations of data characteristics (mean, median, quartiles, maximum & minimum), single-value, double-value, and triple-value strategies. Then we can build a classifier from these data characteristics.

Step 1: Information Granulation

Step 2: IG Representation: Data Characteristics

Strategy 1- Single value: Mean, Median.

Strategy 2-Double values: Max+Min, Q1+Q3

Strategy 3-Triple values: Max+Mean+Min, Q1+Median+Q3

Step 3: Knowledge Acquisition

CHAPTER 4

NUMERICAL EXAMPLES

In this chapter, several data sets from UCI data bank are employed to illustrate our models and evaluate the effectiveness. Besides, some imbalanced data sets are provided to demonstrate the superiority of our methods in solving class imbalance class problem by using the indexes, *Overall Accuracy*, *G-mean* and *Receiver Operation Characteristic (ROC) curve*.

4.1 Performance Measures

Before implementing, we should discuss the effectiveness of performance index in class imbalance situation. The easiest way to evaluate the performance of classifiers is based on the confusion matrix described as Table 4.1. *TP*, *FP*, *TN* and *FN* are defined as bellows.

TP: the number of True Positive examples

FP: the number of False Positive examples

TN: the number of True Negative examples

FN: the number of False Negative examples

Traditionally, the performance of a classifier is evaluated by considering the overall accuracy against test cases. However, when learning from imbalanced data sets, the measure is often not sufficient. For example, it is straightforward to create a classifier having an accuracy of 95% in a domain where the majority class proportion corresponds to 95% of the examples, by simply forecasting every new example as belonging to the majority class. Another fact is the metric considers different classification errors to be equally important. But as we know, a highly imbalanced class problem does not have equal error costs that favor the minority class, which is often the class of primary interest. Therefore, following the available studies (Batista

et al., 2004; Estabrooks et al., 2004; Guo and Viktor, 2004; Provost and Fawcett, 2001; Radivojac et al., 2004), we use *Overall Accuracy* (including *Positive Accuracy* and *Negative Accuracy*), *G-Mean* and *Receiver Operation Characteristic (ROC) curve* to evaluate our KAIG model. The *G-mean* is defined as

$$\sqrt{\text{Positive Accuracy} \times \text{Negative Accuracy}} \tag{4.1}$$

where *Positive Accuracy* and *Negative Accuracy* are calculated as *TP/(FN+TP)* and *TN/(TN+FP)*. This measure is to maximize the accuracy on each of two classes while keeping these accuracies balanced. For instance, a high *Positive Accuracy* by a low *Negative Accuracy* will result in poor *G-mean.*

Table 4.1 Confusion matrix for binary class problem

	Predicted Positive	Predicted Negative
Actual Positive	*TP* (the number of True Positive)	*FN* (the number of False Negative)
Actual Negative	*FP* (the number of False Positive)	*TN* (the number of True Negative)

Another index is *ROC curve*, which is a technique for summarizing a classifier's performance over a range by considering the tradeoffs between *TP rate* and *FP rate*. The *TP rate* and *FP rate* are calculated as *TP/(FN+TP)* and *FP/(FP +TN)*. We use the term *ROC* space to denote the coordinate system used for visualizing classifier's performance. In *ROC* space, *TP rate* is represented on the Y axis and *FP rate* is represented on the X axis. Each classifier is represented by the point in *ROC* space corresponding to its (*FP rate, TP rate*) pair. A ROC analysis also allows the performance of multiple classification functions to be visualized and compared simultaneously. The area under ROC curve (AUC) represents the expected performance as a single scalar. The AUC has a known statistical meaning: it equals to the Wilconxon test of ranks, and is equivalent to several other statistical measures for

evaluating classification and rank models (Hand, 1997).

4.2 Implementation of KAIG Model

4.2.1 Illustrative Example

We apply the KAIG model to the well-known data set, iris. It is comprised of 150 examples. We rearrange it randomly and divide it into two subsets, training set (100 objects) and test set (50 examples). We will illustrate the process of KAIG step by step.

Step 1: Information Granulation

We input the 100 training examples to the Fuzzy ART to form IGs. We set the parameters of Fuzzy ART $\alpha = 0.01$ and $\beta = 1$. The number of IGs varies with the different level of similarity (vigilance). In this study, similarity value varies gradually from 1 to 0. The similarity 1 represents the numerical data. Next, we need to determine which similarity is suitable by the H-index and the U-ratio. The H-index is 'the larger the better' and the U-ratio is 'the smaller the better'. In Figure 4.1, we can find more than one similarity that satisfies this criterion. These similarities are 0.95-0.8 and 0.7-0.55, where H-index = 1 and U-ratio = 0. Their performances of classification, as described in Figure 4.2, are equal to each other. All classification accuracies are equal to 100%.

When the performances are equally good, the amount of granules becomes another criterion for selecting the similarity. In this study, we use IGs instead of numerical data to acquire knowledge and make decisions. If the smaller similarity is selected, the lesser the amount of granules will be dealt with. This smaller amount of granules may save some training time during the building of the model. Therefore, we select a similarity of 0.55 and the amount of granules is 3.

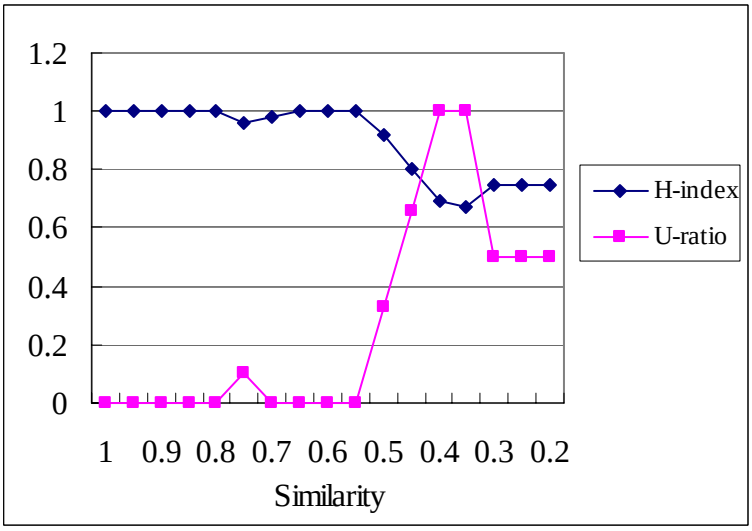


Figure 4.1 The H-index and U-ratio of the iris data

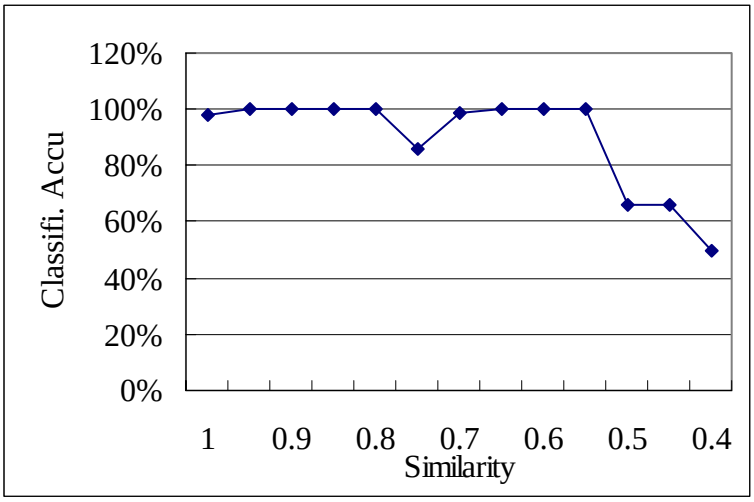


Figure 4.2 The performance of classification (Iris data)

Step 2: Representing the IGs

We describe these 3 granules in hyperboxes form and they are shown in Table 4.2. L_i represents the lower bound of attribute values, and U_i represents the upper limit of attribute values in the *i*-th granule. Take granule #1 for example, it contains 33 objects. In condition attribute A, the minimum is 4.4 and the maximum is 5.7. We utilize the low limit and upper limit to describe all examples in the same one granule. Granule 1 possesses the same class, setosa. Granule 2 contains 33 examples which are of the same class, versicolor. Granule 3 is comprised of 34 examples which have the

same class, virginica.

Table 4.2 The IGs with the similarity of 0.55.

No. of granules		Condition attribute				Classes (No. of examples)		
		A	B	C	D	setosa	versicolor	virginica
#1	L_1	4.4	2.3	1	0.1	33	0	0
	U_1	5.7	4.2	1.9	0.6			
#2	L_2	5	2.2	3	1	0	33	0
	U_2	6.8	3.4	5.1	1.8			
#3	L_3	5.6	2.2	4.8	1.4	0	0	34
	U_3	7.9	3.8	6.9	2.5			

Next, the original attributes are divided into several sub-attributes. Table 4.3 shows the IGs and their sub-attributes. The four original condition attributes (A, B, C, D) are divided into 17 sub-attributes (A1, …, D4). These 17 sub-attributes are used as the inputs for the operation of knowledge acquisition algorithms.

Step 3: Knowledge Acquisition

The rough sets method can be utilized to remove superfluous sub-attributes and to acquire knowledge. The theory of rough sets emerged as a major mathematical tool for discovering knowledge. A fundamental principle of a rough set-based learning system is to discover redundancies and dependencies between the given features of a problem to be classified (Mitra et al., 2002). In the rough set method, a reduct is the minimal subset of attributes that enable the same classification of objects with full attributes. All results of rough sets are operated by Rosseta software. Readers can find additional information on the theory of rough sets in the references (Hu et al., 2002; Walczak and Massart, 1999). The knowledge rules extracted by rough set method are listed as follows:

Rule 1: IF B4=1 THEN Class= setosa;

Rule 2: IF D2=1 THEN Class= versicolor;

Rule 3: IF B4=0 AND D2=0 THEN Class= virginica;

These knowledge rules can be translated as follows:

Rule 1: ATTRIBUTE B $\in$ (3.8,4.2] THEN Class= setosa;
Rule 2: ATTRIBUTE D $\in$ (1.0,1.4] THEN Class= versicolor;
Rule 3: ATTRIBUTE B $\notin$ (3.8,4.2] AND ATTRIBUTE D $\notin$ (1.0,1.4]
THEN Class= virginica;

These knowledge rules are applied to test the remaining 50 examples. Table 4.4 is the minimal reduct of the testing granules. The sub-attributes of testing granules, B4 and D2, are put into these extracted knowledge rules. The predicted decisions are fully equal to the true ones. Therefore, the classification accuracy is 100%.

In this illustrative example, we reduce some unnecessary detailed information by acquiring knowledge from IGs, but the classification accuracy remains high. Also, the knowledge rules for decision-making are fewer than those extracted from numerical data, which may save the response time of a decision. Table 4.5 shows the comparison of classification performances.

Table 4.3 The IGs with sub-attributes

Original attributes	A					B				C				D				Classes
Sub-attributes	A1	A2	A3	A4	A5	B1	B2	B3	B4	C1	C2	C3	C4	D1	D2	D3	D4	
	4.4-5.0	5-5.6	5.6-5.7	5.7-6.8	6.8-7.9	2.2-2.3	2.3-3.4	3.4-3.8	3.8-4.2	1-1.9	3-4.8	4.8-5.1	5.1-6.9	0.1-0.6	1-1.4	1.4-1.8	1.8-2.5	
Granule No.1	1	1	1	0	0	0	1	1	1	1	0	0	0	1	0	0	0	setosa
Granule No.2	0	1	1	1	0	1	1	0	0	0	1	1	0	0	1	1	0	versicolor
Granule No.3	0	0	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	virginica

Table 4.4 The minimal reduct of IGs for testing

IGs No.	B4	D2	Classes	
	3.8-4.2	1-1.4	Predicted	True
#1	1	0	setosa	setosa
#2	0	1	versicolor	versicolor
#3	0	0	virginica	virginica

Table 4.5 The comparison of processing with information granules and numerical data

Methods	Rough Sets		KAIG	
Data type	Numerical data (Similarity=1.0)		Information granules (Similarity=0.55)	
Classification Accuracy	100%	98%	100%	100%
No. of rules	16		3	

Table 4.6 The background of five data sets

Data set	Title	No. of instances	No of attributes	Value of attributes	Class distribution
WDBC	Wisconsin Diagnostic Breast Cancer	683 (699 minus 16 missing data)	9 (remove first attribute-"ID")	All discrete	Benign (65.5%) Malignant (34.5%)
CE	Car Evaluation Database	1728	6	All discrete	Unacceptable (70.023 %) Acceptable (22.222 %) Good (3.993 %) Very good (3.762 %)
TAE	Teaching Assistant Evaluation	151	5	1-continuous 4-discrete	Low (32.45%) Medium (33.11%) High (34.44%)
BUPA	BUPA liver disorders	345	6	All continuous	Class 1 (42.03%) Class 2 (57.97%)
WINE	Wine recognition data	178	12	All continuous	Class 1 (33.15%) Class 2 (39.89%) Class 3 (26.96%)
PIMA	Pima Indians Diabetes	768	8	All continuous	Healthy (65%) Diabetic (35%)

4.2.2 Evaluation of KAIG Model

To evaluate the effectiveness of the KAIG model, five data sets which come from databank of UCI machine learning group (http://www.ics.uci.edu/~mlearn/) are considered in this section. Table 4.6 provides brief explanation about the data background, including data size, number of features, data characteristics (binary/continuous), and defined classes. Before implementing, we divide all data sets into training set and testing set with the proportion of 3:1.

With the help of the H-index and the U-ratio shown in Figure 4.3, we can find the suitable similarity of these data sets. According to these determined similarities, numerical data is transformed into IGs by Fuzzy ART. Then, three famous knowledge acquisition algorithms, neural network (BP), decision tree (C 4.5 algorithm) and the rough set method, are utilized. Professional II PLUS is employed to build neural network in this study. The optimal neural network (BP) parameter settings, structure and learning iterations shown in Table 4.7 are obtained by trial and error. See5 (C4.5 commercial version) software was utilized to construct a decision tree in this study. The inputs and outputs of decision tree and the rough set method are condition attributes and defined classes respectively.

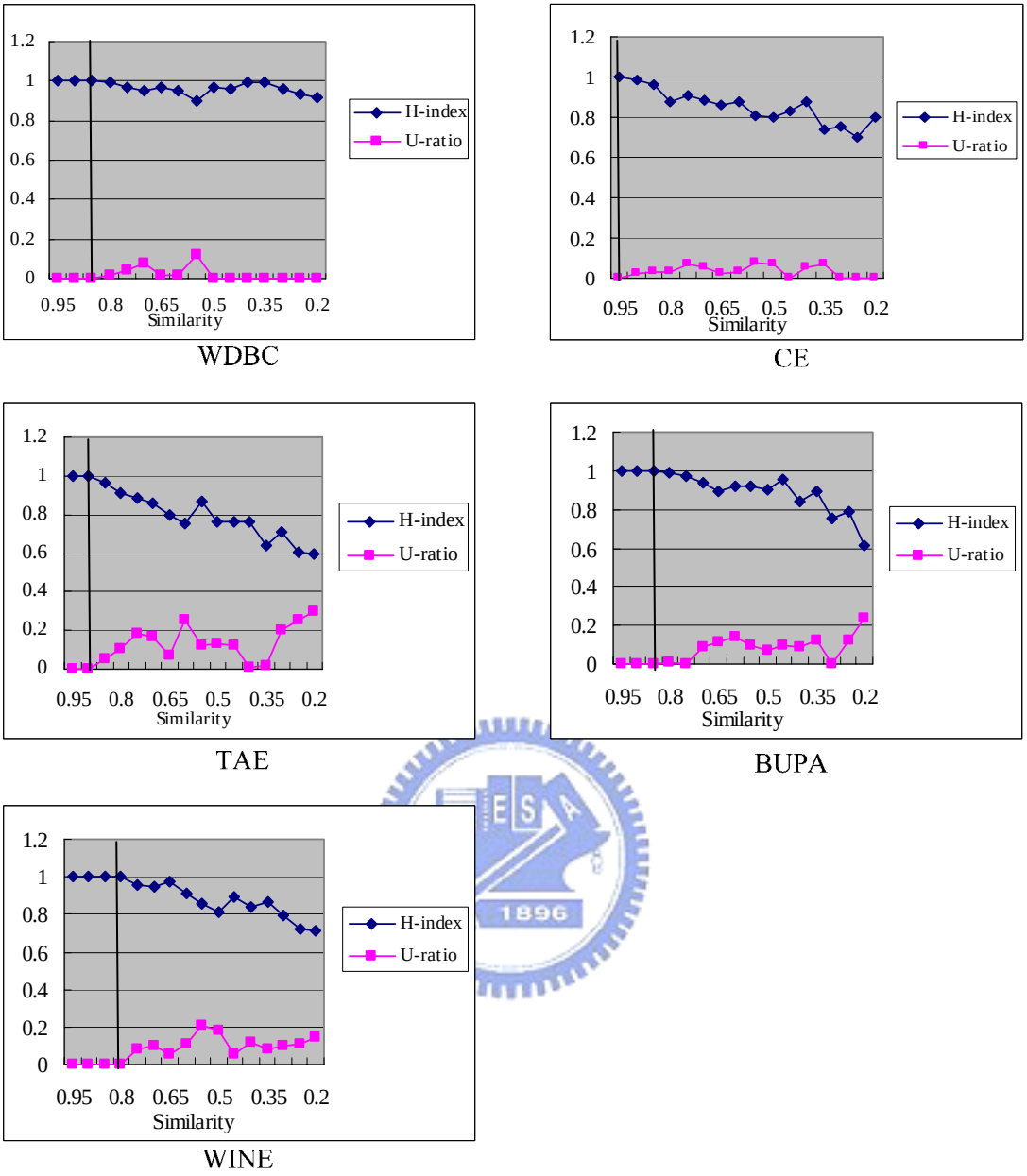


Figure 4.3 The H-indexes and U-ratios of five data sets

The comparisons of implementation results are provided in Table 4.8. Except WDBC, KAIG model has better classification performances in the other five data sets than those of traditional methods which use numerical data. In average, the classification accuracy increases 2.33% and the number of rules is reduced by 48.67% compared with traditional methods. In KAIG model, we can use different kind of knowledge acquisition tools and the results will be different. The classification

accuracy averagely increases 0.86%, 2.238%, 1.182% by applying Rough sets, C4.5 and BP, respectively. In addition, C4.5 has fewer number of knowledge rules (12 rules in average) than those of Rough sets (84.8 rules in average). Therefore, C4.5 is more suitable to be employed in KAIG model than the other two methods.

Table 4.7 The setting of parameters in neural network (BP)

Data Set	Data type	Structure	Learning rate	Momentum	Iterations
WDBC	Numerical	9-11-1	0.2	0.9	20000
	Granule(0.85)	90-160-1	0.2	0.9	20000
CE	Numerical	6-9-1	0.2	0.9	10000
	Granule(0.95)	21-35-1	0.3	0.9	20000
TAE	Numerical	5-6-1	0.2	0.9	20000
	Granule(0.85)	69-120-1	0.2	0.9	20000
BUPA	Numerical	6-5-1	0.3	0.8	30000
	Granule(0.85)	26-31-1	0.2	0.9	15000
WINE	Numerical	13-5-1	0.3	0.7	10000
	Granule(0.8)	35-7-1	0.2	0.8	15000

4.2.3 Implementation in Imbalanced Data

This section will apply KAIG method to overcome the class imbalance problems. C4.5 and SVM are usually utilized as benchmarks or basic learners in related works (Batista et al., 2004; Guo and Viktor, 2004; Huang et al., 2004; Jo and Japkowicz, 2004; Provost and Fawcett, 2001; Radivojac et al., 2004). Therefore, the experimental results of KAIG will be compared with these two methods. A brief introduction about SVM can be found in (Wu and Chang, 2005; Cristianini and Shawe-Taylor, 2000).

The imbalance class problems often happen in medical diagnosis data. Therefore, pima-indians-diabetes whose information shows in Table 4.6 is employed to verify effectiveness of our model. Results for this data set, shown in Table 4.9, were averaged over 4-fold cross validation (CV) experiments, which the data set was

partitioned into 4 equal sized sets and each set was then in turn used as the test set. Besides, in order to test the robustness of KAIG model, we reduce the proportion of minority class from 35% to 10% and 5% by removing the number of minor examples randomly.

Table 4.8 The comparison of classification performance

Data type / Classification Accuracy / Methods		Numerical data (Similarity =1.0) Traditional methods			Granules (Similarity =0.85) KAIG		
	Methods	Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
WDBC	Rough sets	100	92.23	212	100	89.47	58
	Decision tree (C 4.5)	97.5	97.06	10	93.4	94.74	4
	Neural Network (BP)	96.66	100	-	100	89.64	-
CE	Phase	Similarity =1.0			Similarity =0.95		
		Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
	Rough sets	100	89.58	385	100	88.96	207
	Decision tree (C 4.5)	97.4	92.8	75	98.4	95.58	36
	Neural Network (BP)	91.18	91.09	-	94.04	92.80	-
TAE	Phase	Similarity =1.0			Similarity =0.9 0		
		Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
	Rough sets	84.96	84.21	90	95.95	87.37	68
	Decision tree (C 4.5)	60.2	47.36	13	64.9	48.39	11
	Neural Network (BP)	68.23	69.05	-	78.61	84.21	-
BUPA	Phase	Similarity =1.0			Similarity =0.85		
		Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
	Rough sets	100	63.95	165	100	66	80
	Decision tree (C 4.5)	76.4	65.1	15	78.2	70	5
	Neural Network (BP)	69.35	64.47	-	100	66.15	-
WINE	Phase	Similarity =1.0			Similarity =0.8		
		Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
	Rough sets	100	93.18	31	100	95.65	11
	Decision tree (C 4.5)	95.6	90.9	6	96.7	95.7	4
	Neural Network (BP)	87.63	86.49	-	85.87	84.21	-

In the experiments of 35%, 10% and 5%, the results indicate that KAIG model

has better performance than those of SVM and C4.5 against highly imbalanced data sets, in term of the *Negative Accuracy*. In average, KAIG owns 58.08% of *Negative Accuracy* far better than 14.55% of SVM and 27.49% of C4.5. It means KAIG has excellent capability of detecting minor examples (diabetic patients). Meanwhile, KAIG doesn't lose *Overall Accuracy* and *Positive Accuracy*. They are even better than those of SVM and C4.5 in experiment of 35%.

Table 4.9 The results in different proportion of minor class examples

Methods	KAIG				SVM				Decision tree (C 4.5)			
(35%)	Training		Test		Training		Test		Training		Test	
	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
Overall Accuracy	91.97%	2.5%	78.78%	2.5%	76.82%	1.4%	75.52%	2.8%	81.50%	4.28%	74.22%	3.1%
Pos. Acc.	93.07%	2.3%	84.00%	4.5%	93.07%	0.5%	92.60%	2.7%	87.94%	7.50%	83.20%	2.8%
Neg. Acc.	85.24%	2.1%	70.52%	8.0%	46.52%	4.7%	43.66%	3.3%	71.40%	8.73%	57.46%	8.3%
G-mean	90.67%	2.59%	76.46%	3.85%	65.73%	3.18%	63.56%	3.29%	78.95%	3.24%	68.99%	4.80%
(10%)	Training		Test		Training		Test		Training		Test	
	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
Overall Accuracy	95.01%	1.1%	87.05%	2.3%	89.93%	0%	89.93%	0%	91.55%	1.7%	88.49%	1.6%
Pos. Acc.	99.33%	0.8%	92.20%	1.2%	100%	0%	100%	0%	98.73%	1.9%	96.80%	3.1%
Neg. Acc.	52.98%	11.9%	41.08%	20.5%	0%	0%	0%	0%	27.38%	20.4%	14.29%	24.0%
G-mean	72.16%	7.91%	59.73%	17.0%	0%	0%	0%	0%	44.43%	30.6%	23.63%	32.1%
(5%)	Training		Test		Training		Test		Training		Test	
	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
Overall Accuracy	97.48%	0.7%	94.89%	1.6%	94.94%	0%	94.70%	0%	96.52%	0.8%	93.56%	1.0%
Pos. Acc.	98.54%	0.9%	98.60%	1.7%	100%	0%	100%	0%	99.47%	0.7%	98.20%	0.8%
Neg. Acc.	72.50%	13.2%	28.57%	26.1%	0%	0%	0%	0%	41.25%	14.9%	10.72%	13.7%
G-mean	84%	7.69%	44%	33.3%	0%	0%	0%	0%	63%	12.7%	23%	26.9%

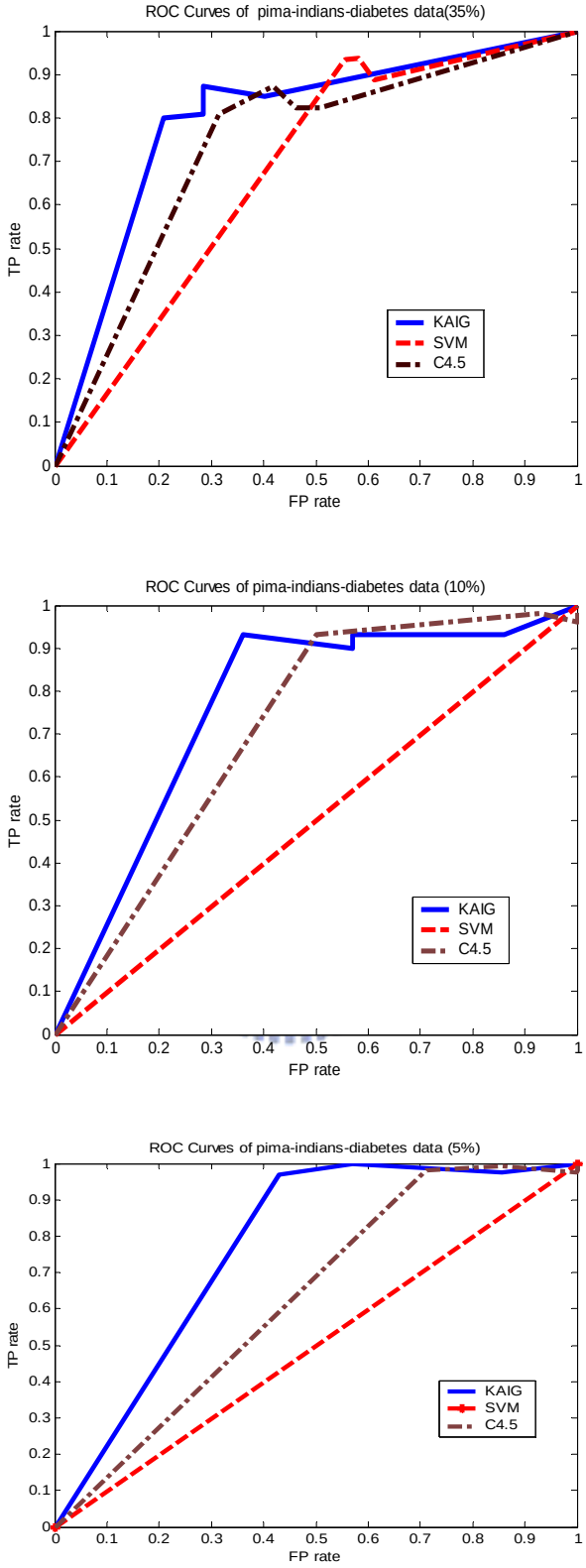


Figure 4.4 ROC curves of pima-indians-diabetes data

Both *G-mean* and *ROC* curves shown in Figure 4.4 also demonstrate the superiority of our method. In extreme skewed data (10% and 5%), *G-mean* is more

sensitive than *Overall Accuracy*. When *Negative Accuracy* decreases dramatically, *G-mean* can indicate these changes but *Overall Accuracy* can not. *ROC* curves provide visual results which can easily compare these three methods and find KAIG has best performances (*AUC*) in different experiments.

4.2.4 Discussion and Concluding Remarks

A novel method called KAIG model is presented to solve class imbalance problems. In this model, we propose two indexes to determine the level of granularity and the "sub-attributes" concept to describe IGs. The experimental results show that the KAIG model can improve classification performance by reducing unnecessary details of information. We also demonstrate that the proposed method has excellent ability of identifying the minority examples in imbalanced learning tasks. In medical diagnosis data, our method can dramatically increase *Negative Accuracy* without losing *Positive Accuracy* and *Overall Accuracy*. *ROC curves* and *G-mean* also illustrate the superiority of KAIG model compared with C4.5 and SVM.

Construction of IGs is one of many interesting and important issues in granular computing. IGs are aimed at building efficient and user-centered views of the external world and supporting/facilitating our perception of the surrounding physical and virtual world. In our research, we construct IGs by objects' "similarity", the parameter (vigilance) of Fuzzy ART. It can define the "indistinghishable, similar, coherency and alike" relations of objects. However, other relations whose definitions are not specific/ concrete, such as "functional adjacency", also can employ to construct IGs. But, it is hard to define these "not specific" relations. Therefore, more efforts of studying different relations are necessary in the future researches.

4.3 Implementation of IG based Method

4.3.1 Illustrative Example

In this section, the Haberman's survial data is used to be an illustrative example of IG based method. This data set contains 306 objects. 225 of them are training examples and the rests are for testing. The operation of IG based method will be illustrated step by step.

Step 1: Information Granulation

The information granulation process is the same with KAIG model. We input the training data to the Fuzzy ART. The parameters of Fuzzy ART, α, β, are set as 0.01, 1, respectively. Then, we vary gradually the parameter "vigilance" (level of similarity) from 1 to 0. According to 'granularity selection criteria' mentioned in section 3.2, we can find the suitable similarity is 0.85. The results of H-index and U-ration can be found in Figure 4.5. Forty-two IGs are constructed during information granulation process.

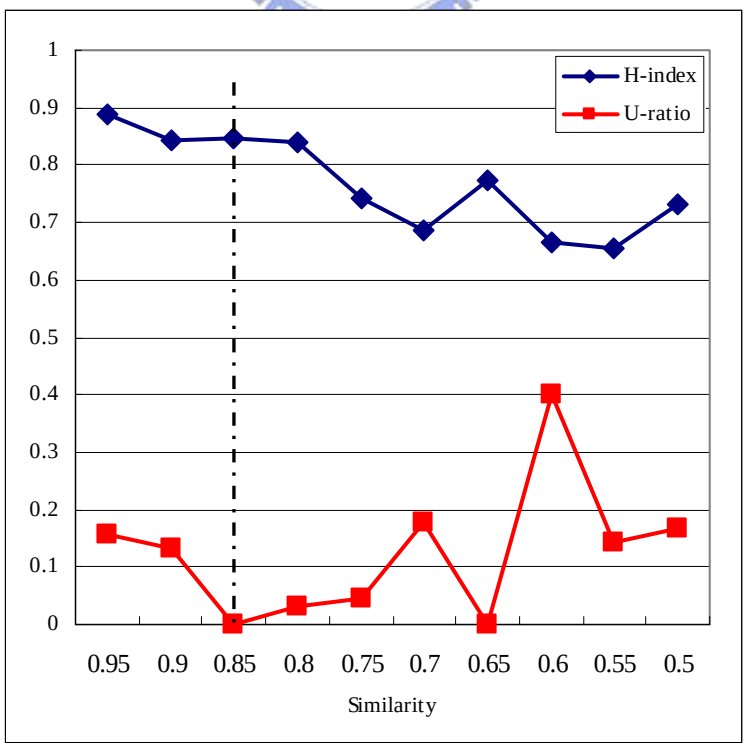


Figure 4.5 H-index and U-ratio of Haberman's survial data

Table 4.10 An illustrative example of IG based method

IG No	Attributes			Class	Data Characteristics				
	x1	x2	x3			x1	x2	x3	class
1	59	64	0	+1	**Mean**	59.6	64.4	0.22	+1
1	62	66	0	+1	**Median**	60	64	0	+1
1	60	64	0	+1	**Q1**	59	64	0	+1
1	59	64	1	+1	**Q3**	61	65	0	+1
1	57	64	0	+1	**Min**	57	64	0	+1
1	61	64	0	+1	**Max**	62	66	1	+1
2	43	60	0	+1	**Mean**	44	62.1	0.57	+1
2	46	63	0	+1	**Median**	44	63	0	+1
2	43	63	2	+1	**Q1**	43	61.5	0	+1
2	46	62	0	+1	**Q3**	45	63	1	+1
2	44	61	0	+1	**Min**	42	60	0	+1
2	42	63	1	+1	**Max**	46	63	2	+1
2	44	63	1	+1					
3	39	67	0	+1	**Mean**	42.4	65.4	0.55	+1
3	41	65	0	+1	**Median**	43	65	0	+1
3	45	67	1	+1	**Q1**	41	64	0	+1
3	43	64	3	+1	**Q3**	43	67	0.5	+1
3	45	67	0	+1	**Min**	39	64	0	+1
3	42	65	0	+1	**Max**	45	67	3	+1
3	43	64	2	+1					
3	41	65	0	+1					
..	..	..	..	..	..	..	..	..	..
34	47	63	23	-1	**Mean**	50	64	22.5	-1
34	46	65	20	-1	**Median**	50	64	23	-1
34	54	65	23	-1	**Q1**	46.8	63	22.3	-1
34	53	63	24	-1	**Q3**	53.3	65	23.3	-1
					Min	46	63	20	-1
					Max	54	65	24	-1
36	66	61	13	-1	**Mean**	62.7	59.7	14.3	-1
36	62	59	13	-1	**Median**	62	59	13	-1
36	60	59	17	-1	**Q1**	61	59	13	-1
					Q3	64	60	15	-1
					Min	60	59	13	-1
					Max	66	61	17	-1
37	48	58	11	-1	**Mean**	47.7	58.3	11	-1
37	51	59	13	-1	**Median**	48	58	11	-1
37	44	58	9	-1	**Q1**	46	58	10	-1
					Q3	49.5	58.5	12	-1
					Min	44	58	9	-1
					Max	51	59	13	-1

Step 2: IG Representation by using Data Characteristics

In this step, we use different data characteristics and their combination to represent the constructed IGs. Table 4.10 provides parts of the IGs and their data characteristics. We utilize Mean, Median, Q1+Q3, Min+Max, Q1+Median+Q3, and Min+Mean+Max, to describe those IGs.

Step 3: Knowledge Acquisition

Those data characteristics mentioned above serve as the training data of a learner. Decision tree is employed as basic classifier in this method. The results of using Mean to represent IGs are listed as bellow. During training phase, the classification performance, Overall Accuracy, Positive Accuracy, and Negative Accuracy are 85.71%, 84.62%, 87.5%, respectively. The evaluation of three IG representing strategies is provided in next section.

Rule 1: IF x1 <= 43 THEN Class= +1 [Support: 0.833]
Rule 2: IF x3 <= 4 THEN Class= +1 [Support: 0.808]
Rule 3: IF x1 > 43 AND x3 > 4 THEN Class= -1 [Support: 0.689]

4.3.2 Experimental Results

In order to evaluate the effectiveness of the IG based method, three data sets from UCI are considered in this chapter. The brief illustration about these data sets is provided in Table 4.11. Before implementing, the data sets are divided into training and testing set (4:1). In other word, we use 80% data to build model and the rest 20% is employed to criticize the constructed classifier. The basic learner will be C 4.5 and SVM is the benchmark.

Table 4.11 Data background (UCI)

Data set	No of examples	No of attributes	Value of attributes	Class distribution
Diabetes	759	9	Continuous	Healthy: 496 (66%) Diabetic: 263 (34%)
Haberman's survial data	306	4	Continuous	Survived (74%) Died (26%)
Post-operative patient data	90	10	All discrete	I (sent to Intensive Care Unit) (2%) S (prepared to go home) (27%) A (sent to general hospital floor) (71%)

The experimental results of IG based method are summarized in Table 4.12. We have tried three strategies, single, double, triple value strategy. Considering average Overall Accuracy (G-mean), the performances of single, double, triple-value strategy are 73.93% (44.98%), 71.44% (39.04%), 70.59% (43.06%), respectively. Obviously, triple-value strategy has better performance than those of single and double strategy. From Figure 4.6 & 4.7, the combination of Q1+Median+Q3 has better Overall Accuracy, G-mean, and Negative Accuracy than others.

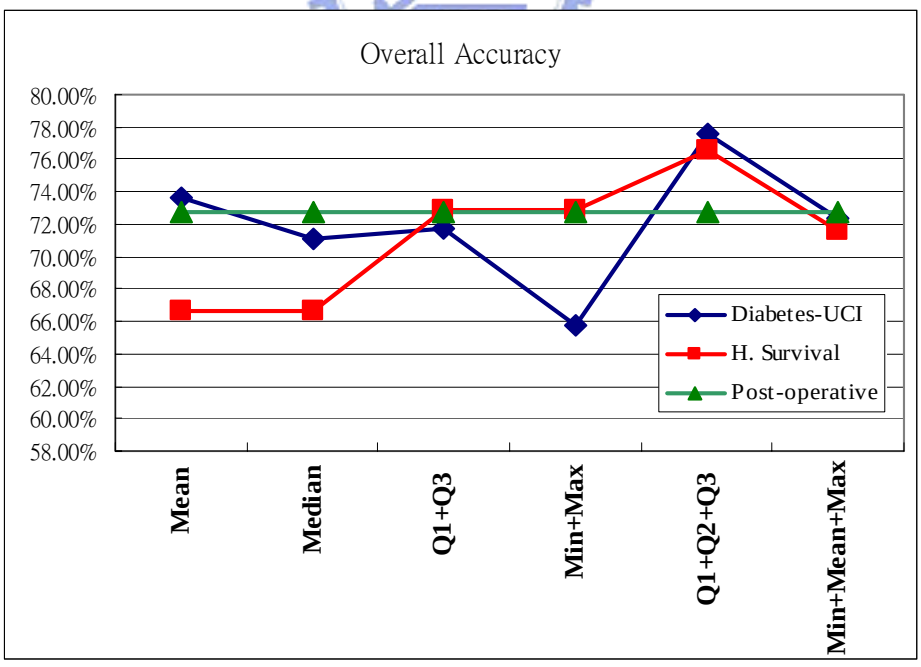


Figure 4.6 Overall accuracies of different strategies (UCI)

Table 4.13 provides the comparison of IG based method and other methods, like DT, SVM, under-sampling, and cluster-based sampling. It's easy to find our method has excellent performance in Overall Accuracy, G-mean, and Negative Accuracy. As seen in Figure 4.8, the proposed IG based method can dramatically increase Negative Accuracy. It means IG based method has marvelous ability to classify minority examples without losing Overall Accuracy. Compared with cluster-based sampling, under-sampling, SVM, and DT, the experimental results also show our proposed methods have the superiority in G-mean.

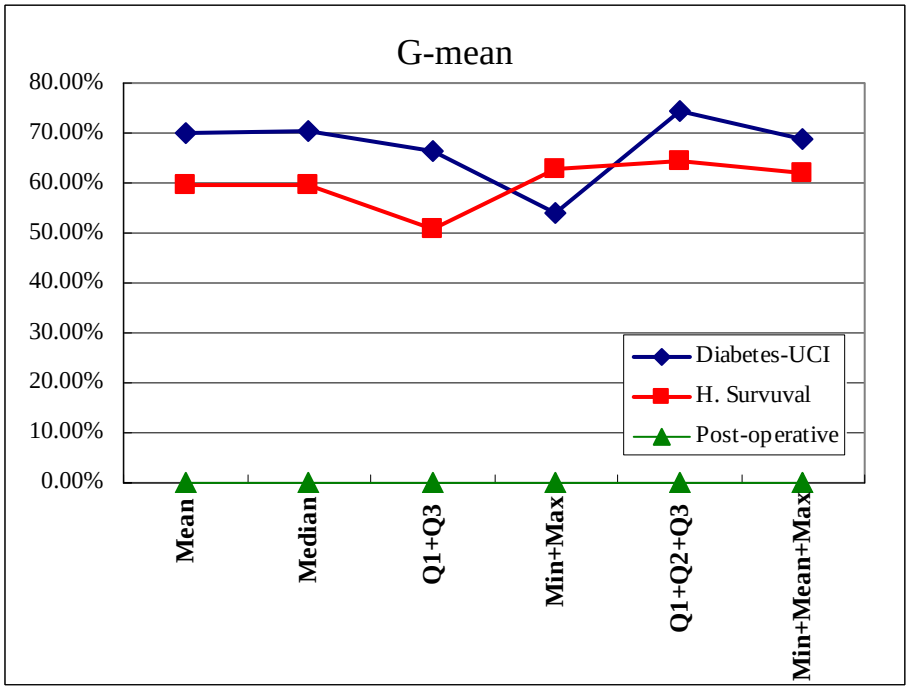


Figure 4.7 G-mean of different strategies (UCI)

Table 4.12 The experimental results of IG based method

Methods / Performance	IG based method (Data characteristics)					
	Mean	Median	Q1+Q3	Min+Max	**Q1+Q2+Q3**	Min+Mean+Max
Overall Accuracy	71.03%	70.15%	72.43%	70.45%	**75.63%**	72.23%
Pos. Accu.	83.86%	81.17%	87.96%	86.76%	**88.97%**	85.24%
Neg. Accu.	36.79%	39.31%	28.66%	28.62%	**38.68%**	36.79%
G-mean	43.11%	43.26%	39.15%	38.93%	**46.29%**	43.66%

Table 4.13 The experimental results of sampling methods

Methods / Performance	IG based method	Cluster based sampling	Under- sampling	SVM	DT
Overall Accuracy	75.63%	71.99%	73.30%	74.98%	72.65%
Pos. Accu.	88.97%	85.27%	86.62%	94.25%	88.64%
Neg. Accu.	38.68%	32.43%	33.69%	19.06%	28.03%
G-mean	46.29%	40.08%	41.39%	30.50%	39.03%

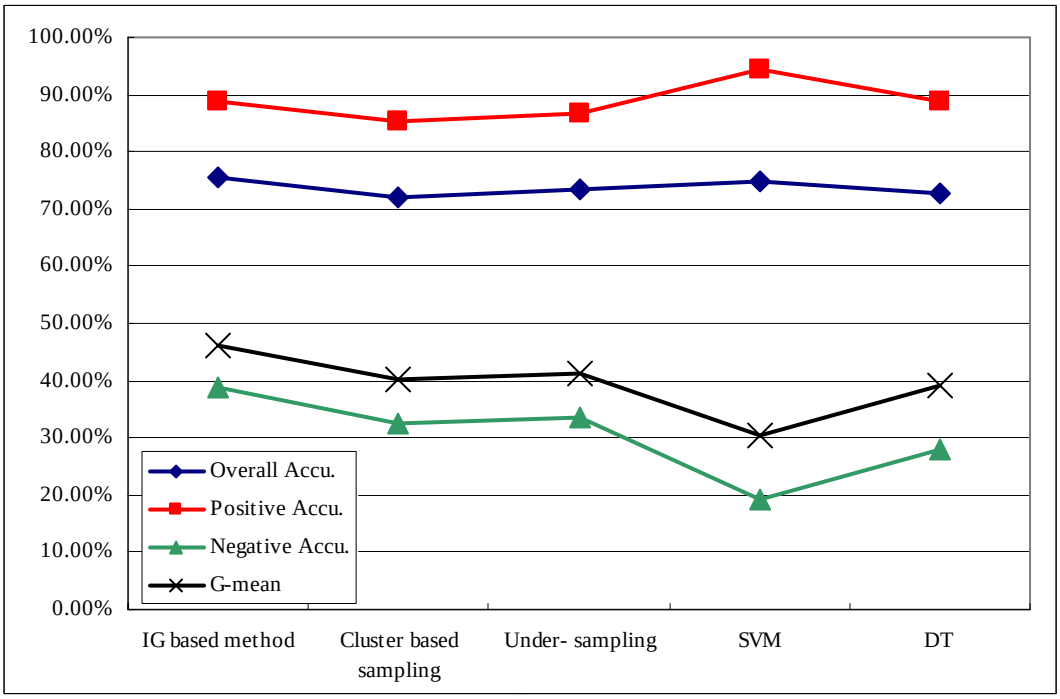


Figure 4.8 Comparison of the proposed IG based, cluster-based sampling, under-sampling, DT and SVM

4.3.3 Discussion and Concluding Remarks

In this section, we proposed IG based method. The experimental results show our method can remarkably improve the class imbalance problems. In other words, IG based method can dramatically raise the ability of identifying minority examples without losing Overall Accuracy. Among three IG representing strategies, triple-value strategy (Q1+Q2+Q3) outperforms double and single-value strategies.

In addition, the sampling methods can improve the class imbalance situation, but they cannot guarantee to find the optimal solutions. This is because sampling methods

lack a systematic method to find representative examples and determine what the proportion of majority to minority is. The IG based method can be viewed as another kind of sampling method. However, the problems mentioned above can be easily resolved by using our approach. IG based method has a clear procedure to find suitable number of clusters. Once the number of clusters is determined, the representative examples and the proportion of majority to minority can be easily found. In other words, IG based method enhances sampling methods without sacrificing the advantages. Besides, compared with KAIG model, IG based method can avoid increasing the number of input variables.

CHAPTER 5

APPLY KAIG MODEL TO BUILD A GRANULAR COMPUTING BASED SCHEDULING SYSTEM WITHIN DYNAMIC MANUFACTURING ENVIRONMENT

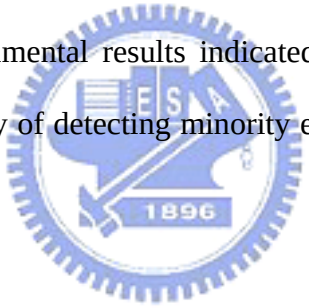
5.1 Problem Description

Scheduling must efficiently reconcile conflicts in the assignment of various resources and also the constraints between them, in order to keep the production system operating smoothly. Traditionally, researchers use analytical tools such as mathematical modeling technology, dynamic programming, branch-and-bound methods or other heuristic algorithms (Li et al. 2003) to solve static scheduling problem. Unfortunately, scheduling environments are usually dynamic. Therefore, researchers try to tackle the issues of dynamic scheduling problems by employing machine learning approaches (Aytug et al., 1994) such as Artificial Neural Networks (Li and She, 1994; Sim et al., 1994; Li et al., 2003; Min and Yih, 2003), Decision Trees (Su and Shiue 2003), Support Vector Machines (SVM) (Gersmann and Hammer, 2005), Genetic Algorithm (Wang and Uzsoy, 2002) and etc. According to related works, applying inductive learning techniques is a useful way in acquiring dynamic scheduling knowledge and can effectively solve dynamic scheduling problems.

However, when inducing knowledge by generalizing from environment-provided examples, there are some issues needed to be overcome. One of them is class imbalance problem (Japkowicz and Stephen, 2002; Wu and Chang, 2005). This problem is of crucial importance since it is encountered by a large number of domains of great environmental, vital or commercial importance, and was shown, in certain cases, to cause a significant bottleneck in the performance attainable by standard

learning methods, which assume a balanced class distribution. For example, it is relatively cheap and simple to obtain training examples from a normally working machine in manufacturing environments. But, the sampling from a faulty machine will require that the machine has to be damaged in several ways to obtain defective examples. The creating of a balanced training set will therefore be very expensive or impractical.

This chapter develops a new scheduler which integrates GrC model to conduct class imbalance problem in dynamic scheduling environment. We implement this new scheduler within a simulated Flexible Manufacturing System (FMS) environment. A highly imbalanced simulation data is generated to evaluate the effectiveness of this scheduler. Compared with traditional techniques, “cluster based sampling” and “costs adjusting” method, the experimental results indicated that the proposed model can remarkably improve the ability of detecting minority examples while reforming other classification performance.



5.2 A Granular Computing Based Scheduler

In this section, we’ll discuss the proposed model within a dynamic scheduling environment. The scheduler contains two major mechanism, information granulation and inductive learning mechanism. The first mechanism is to construct IGs. The second one is to extract scheduling knowledge from IGs.

5.2.1 Information Granulation Mechanism

The main purpose of information granulation is to transfer numerical data into IGs. This mechanism has three phases described as Figure 5.1. Phase 1 and 2 try to find a suitable level of granularity and construct IG within Fuzzy ART neural network. Phase 3 is to describe these constructed IGs by using sub-attributes. Then the

inductive learning mechanism can extract knowledge from these IGs with addition of sub-attributes. A more concise algorithm for the information granulation mechanism follows.

Step 1: Initialize parameters of Fuzzy ART.

Select values for vigilance parameter ($\rho = 1$), choice parameter ($\alpha \to 0$) and learning rate ($\beta = 1$).

Step 2: Construct IGs by Fuzzy ART according to the selected vigilance parameter (similarity).

Step 3: Compute H-index & U-ratio.

Step 4: Check "granularity selection criteria".

IF H-index & U-ratio don't satisfy "granularity selection criteria"

$\rho = \rho - \Delta$ AND go to step 2.

Where Δ represents a constant increment defined by users

ELSE

Go to step 5.

Step 5: Describe IGs by hyperbox form.

Step 6: Divide original attributes into sub-attributes.

Step 7: Re-formulate IGs with sub-attributes.

5.2.2 Inductive Learning Mechanism

In this study, we consider two learning methods, Support Vector Machine (SVM) and Decision Tree (C 4.5). SVM is a powerful learning method and often employed to tackle class imbalance problems (Wu and Chang, 2005). Decision tree is the benchmark of comparison and it's also the basic learner in many literatures.

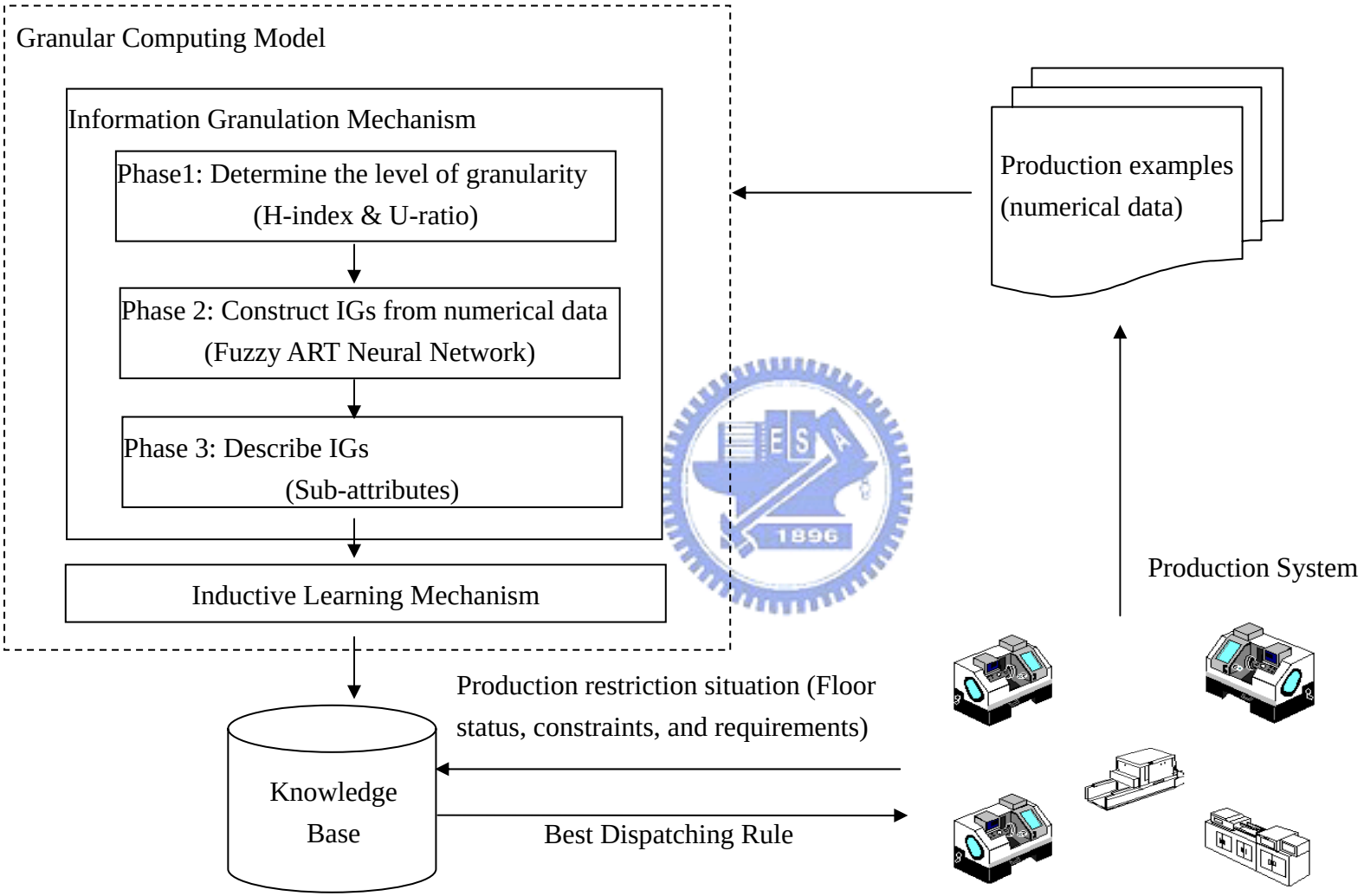


Fig. 5.1 An Granular Computing based scheduler

5.3 Comparative Techniques

We use two techniques, *Cost Adjusting* and *Cluster Based Sampling*, to be the comparative algorithms. In order to compare the classification performance of standard SVM and these two techniques, we use SVM to be the learner in different schemes.

5.3.1 Costs Adjusting Method

This method is to improve the classification performance by increasing the misclassification cost for minority class. Traditional performance indexes consider the misclassification costs of majority and minority instances are equal. Under the assumption of maximizing overall classification accuracy, the minority examples will be neglected. If we give penalty (cost) to minority class, the class imbalance problem will be improved. In this method, different misclassification costs can be incorporated into classes, which avoid direct artificial manipulation on the training set.

5.3.2 Clusters Based Sampling Method

Under-sampling might remove some important examples and over-sampling introduces noises into the training data set. Therefore, we consider cluster based sampling method in this study. The purpose of cluster based sampling method is to find representative examples and balance class distribution. This technique can be illustrated in Figure 5.2. First, we separate majority and minority examples into two groups. The minority population is kept intact. Second, the majority examples are clustered depending on their similarity. Third, we sample representative majority examples from those clusters. Finally, those samples and minority examples are joined together and we can acquire knowledge from both class examples. In order to improve class imbalanced situation, we only vary the number of majority examples by randomly sampling and do nothing for minority examples.

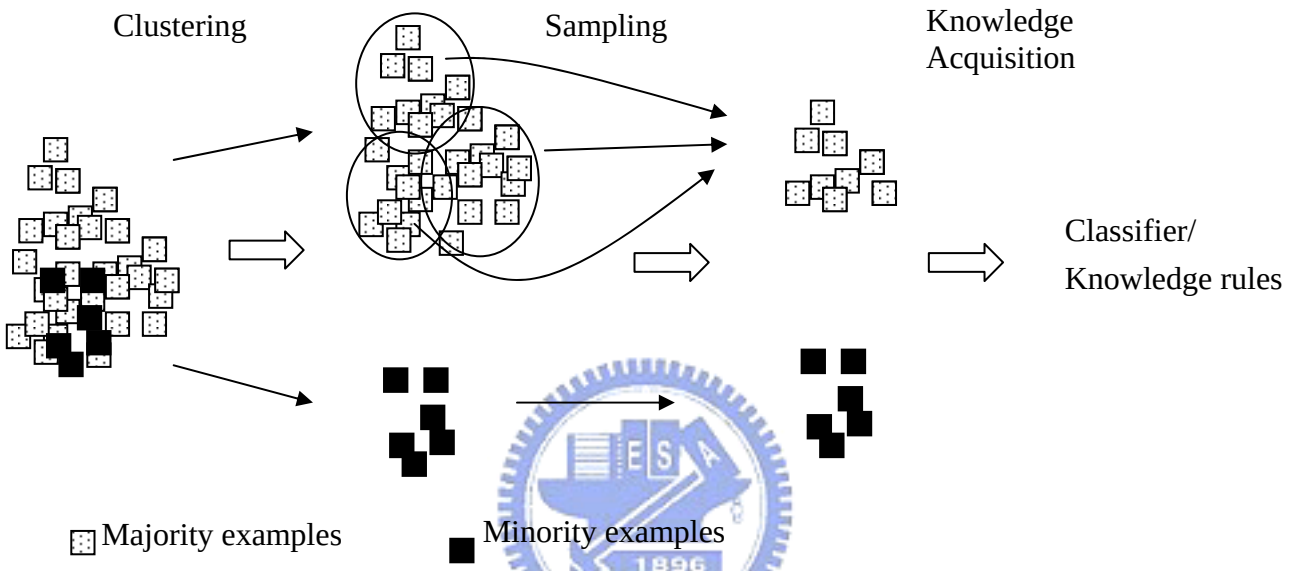


Fig. 5.2 Illustration of cluster based sampling method

5.4 Implementation

In this section, a *Flexible Manufacturing System* (FMS) schedule decision problem is provided to illustrate the class imbalance problem. PROMODEL II software is employed to build this simulation model. And the GrC model is programmed with the use of the software of Matlab 6.1.

5.4.1 Description of Simulated System

We slightly modify the FMS simulation model built in Li's paper (Li et al., 2003) by adding one processing part. Basically, we follow most assumptions like physical layout, simulation time and others. This modified FMS simulation model involves 5 different types of parts, 4 numerical control (NC) machines, 4 Work-In-Process (WIP) buffers located in front of each machine, and 1 load/unload station. To evaluate the performance of the FMS, "mean machine utility" was selected to evaluate production performance. For simplicity and clarity, three basic dispatching rules were set for deciding which part should be processed when there is more than one part in the WIP buffer waiting to be processed. They are *FCFS* (First come, first served), *SPT* (Shortest processing time) and *EDD* (Earliest due date). Besides, four system control attributes were defined to describe the operating conditions of this FMS model. They are as follows:

-Buffer size: capacity of WIP buffers (unit).

-Arrival rate of parts: frequency of parts incoming to the FMS (batch/min).

-Batch size: the volume of parts per batch (unit/batch).

-Speed of AGV (m/min).

There are 1500 examples collected and 4 folds CV method is applied for training and testing. The distribution of each class is 42.8% (*FCFS*), 49.6% (*SPT*) and 7.6% (*EDD*). All attributes' values are continuous. In such data, the examples labeled EDD obviously are viewed as the minority compared with those of FCFS and SPT.

5.4.2 Using the Costs Adjusting Method

By increasing the misclassification cost of minority examples from 1, 2, 5 to 10, we train a SVM classifier with RBF kernel function and the parameter setting of γ and tolerance are 0.5, 0.00001, respectively. The experimental results are shown in Figure 5.3. C=1 represents equal misclassification cost of classes. From Figure 5.3, with the increase of cost, we can find the *Negative Accuracy and G-mean* rises remarkably while the *Positive Accuracy* decreasing slowly. But, the Overall Accuracy has slight improvement compared with the origin (C=1). Considering the tradeoff of these indexes, we choose the C=5 as the optimal solution in this method.

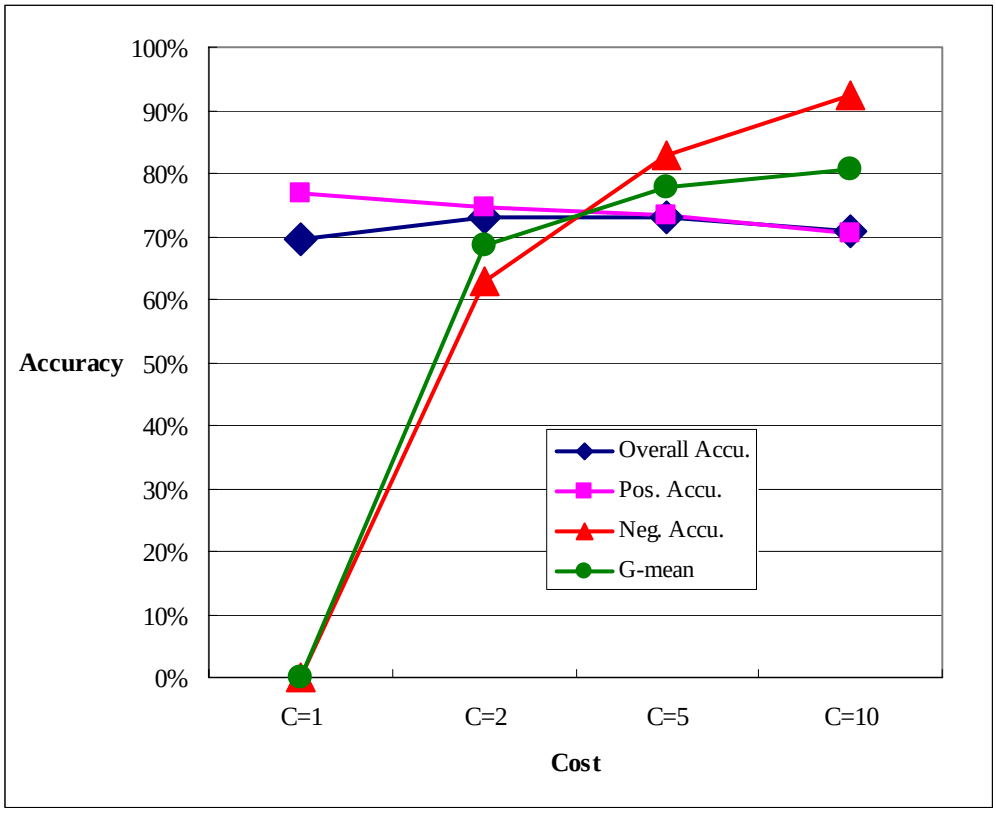


Fig 5.3 Overall accuracy, Positive Accuracy, Negative Accuracy & G-mean of SVM with different costs

5.4.3 Using the Cluster Based Sampling Method

In this method, we vary the proportion of majority and minority by only changing the number of majority examples. Four different combinations of proportion

of majority to minority (1:1), (2:1), (3:1), (4:1) and (9:1) are employed to implement this method. In order to compete with cost adjusting method, we use the same inductive learning algorithm (SVM) and same test data. Figure 5.4 shows the experimental results of different proportions. In 2:1 proportion, the cluster based sampling method has best G-mean and good performances in Accuracies. In 1:1 situation, the *Negative Accuracy* reaches to perfect 100%, but other indexes drops. This is because the data size (86+86=172) of training dataset smaller than the number of test dataset (375).

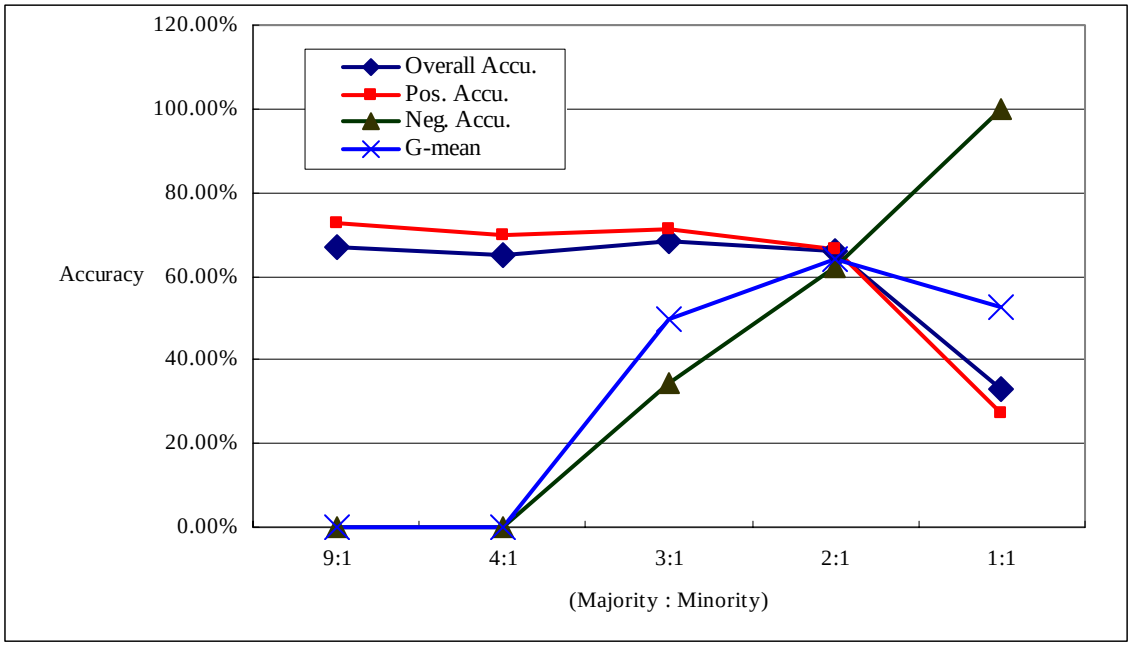


Fig. 5.4 Overall Accuracy, Positive Accuracy, Negative Accuracy & G-mean of Cluster based method with different sample size

5.5 Experimental Results

The experimental results are summarized in Table 5.1 and Figure 5.5. In *Overall Accuracy*, our proposed method is slightly better than C4.5 and obviously superior to Costs Adjusting, Cluster-based Sampling and SVM, which can not identify any minor examples in this case. In average, the proposed approach remains the ability (89.53%) of identifying class FCFS and SPT, comparing with Costs Adjusting (73.45%),

Cluster based sampling (61.74%) and C4.5 (89.54%). Moreover, our method has better capability (91.95%) in identifying minor instances (EDD) than 82.76% of Costs Adjusting, 64.37% of Cluster-based sampling, and 67.81% of C4.5. Our proposed method also outperforms the other methods in *G-mean*. The same conclusion is also shown in Figure 5.6. It's easy to see that AUC of the proposed method is bigger than those of other methods.

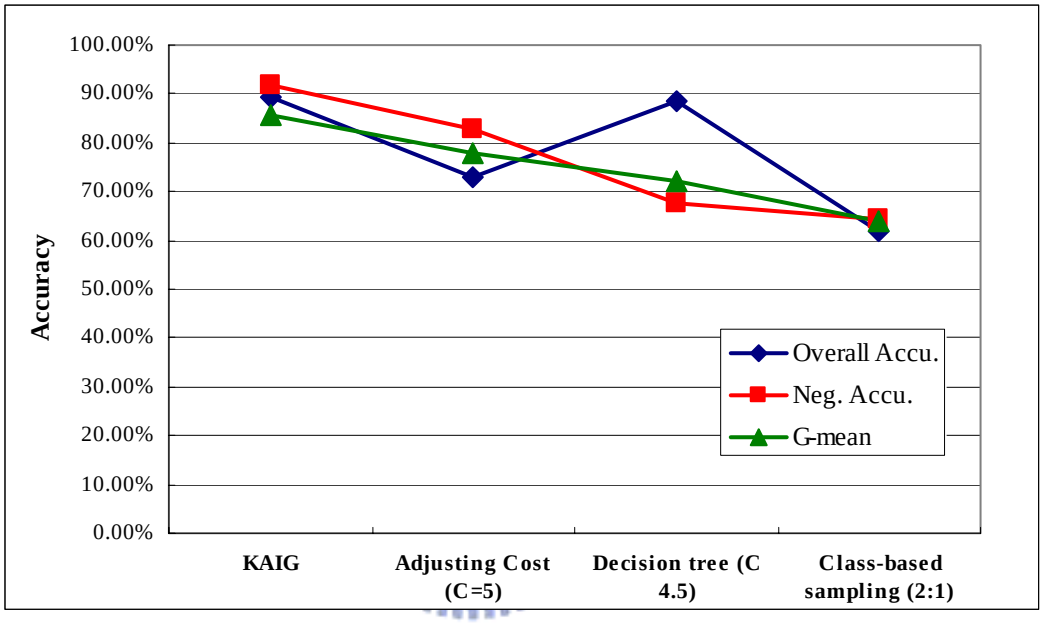


Fig. 5.5 The comparison of classification performances

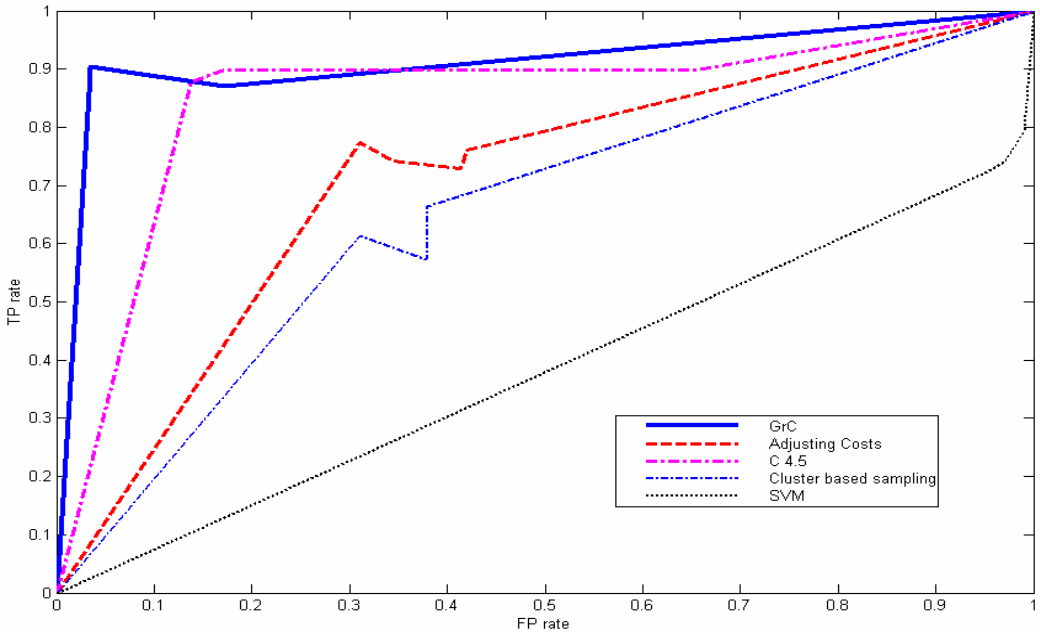


Figure 5.6 ROC curves of FMS data

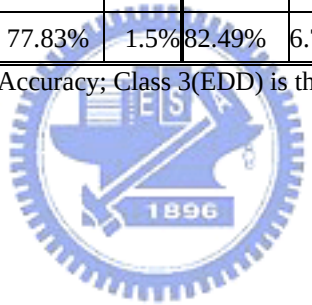
5.6 Discussion and Concluding Remarks

Most conventional classifiers assume more or less equally balanced data classes and do not work well when one class is severely undersampled. Actually, when introducing inductive learning to real world applications, class imbalance problems are necessary to be considered especially in situations where the minority examples are crucial, such as fault monitoring or finished products inspection. In this chapter, we applied KAIG model to induce scheduling knowledge within an FMS. Compared with the experimental results of Cost Adjusting, Cluster-based Sampling, DT (C4.5) and SVM, our proposed model can significantly improve the ability of a scheduling system in detecting minority examples while increasing the overall classification performance. The other two comparative approaches, Costs Adjusting method and Cluster-based Sampling approach, can effectively raise the ability of detecting minority examples in highly imbalanced data. Unfortunately, both techniques might result in decreases of *Positive* and *Overall Accuracy*.

Table 5.1 The experimental results of FMS simulated data

Method	GrC model				Adjusting Costs (C=5)				Decision tree (C 4.5)				Class-based sampling (2:1)			
Phases	Training		Test		Training		Test		Training		Test		Training		Test	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Overall Accuracy	90.64%	1.9%	89.16%	2.2%	74.49%	1.6%	73.00%	2.0%	92.63%	2.6%	88.35%	0.9%	61.94%	2.80%	61.87%	4.27%
Class 1(FCFS)	91.56%	1.9%	92.71%	1.9%	82.11%	0.7%	81.25%	3.5%	92.65%	6.3%	93.96%	3.1%	62.74%	13.01%	62.29%	13.20%
Class 2(SPT)	90.30%	4.7%	85.84%	4.5%	65.62%	3.4%	64.38%	4.3%	82.59%	8.1%	85.12%	4.8%	61.37%	5.68%	61.11%	2.76%
Class 3(EDD)	87.60%	4.8%	91.95%	8.0%	89.24%	3.5%	82.76%	4.9%	88.96%	0.6%	67.81%	28.9%	63.45%	5.29%	64.37%	3.98%
G-mean	85.04%	2.12%	85.50%	5.33%	81.50%	1.5%	77.83%	1.5%	82.49%	6.71%	72.18%	16.7%	62.23%	3.10%	64.03%	1.74%

Note: Class 1(FCFS) and Class 2(SPT) can viewed as Positive Accuracy; Class 3(EDD) is the Negative Accuracy.



CHAPTER 6

APPLY KAIG MODEL TO SHORTEN THE CELLULAR PHONE TEST PROCESS

6.1 Problem Description

Personal wireless communication services have been available to the general public for only about 10 years, since the breakthrough of cellular phones (Hannikainen et al., 2002). At the same time the technology employed by mobile telecommunications is evolving rapidly. New designs in cellular phones and novel functions are being introduced at an ever increasing pace. This is leading to fierce competition and short product life cycles. Consequently, one of the major concerns of OEM (Original Equipment Manufacture) and EMS (Electronic Manufacturing Service) phone manufacturers is how to decrease testing costs, especially in the low profit environment in which they operate. This is because testing equipments for mobile phones are expensive, and the testing times long. In one estimate, it costs around 1 US dollar and 1-3 minutes per phone (VI service Network). However, these testing cost and time will increase dramatically because more and more newly developed modules like digital camera, mp3 player, personal digital assistant (PDA), and blue-tooth transmitter are added to cellular phones. We have to spend extra time and money to inspect these new functions. These factors often hinder the enhancement of the overall output of cellular phones (VI service Network).

Another key issue affecting handset vendor success is time-to-market (Agilent Technologies). If we can shorten manufacturing time, the time-to-market will be reduced. Cellular manufacturers can quickly response customers' demand. In the manufacturing process of cellular phones shown as Figure 6.1, the Radio Frequency (RF) function is a crucial test and needs more operation time than any of the other

inspection processes. In order to save inspection costs and shorten production time, manufacturers need an effective method to reduce the RF function test items. A number of soft computing approaches, such as neural networks (Verikas and Bacauskiene, 2002), genetic algorithms (GA) (Zhu and Guan, 2004), decision tree and rough sets (Swiniarski and Hargis, 2001; Swiniarski and Skowron, 2003) have been widely used to remove irrelevant, unnecessary, and redundant attributes (test items). However, when these methods are applied to real world problems, there are many issues that need to be addressed. One of them is the "class imbalance" problem.

In modern production systems, the defective rate of products is becoming quite low. In the six sigma quality management system for example, we should use "ppm"(parts per million) instead of "%" to calculate the defective rate. In a mature manufacturing industry the amount of good products far exceeds the defective products. When feature selection approaches encounter imbalanced data such as this, it becomes difficult to acquire knowledge from the few negative examples (defective products). Fewer abnormal products will be viewed as outliers or bias by feature selection methods (Pendharkar et al., 1999). This leads to a high level of type II errors (customer risks, the probability that customers accept defective products) which are critical to OEM/EMS companies. A low level of Negative Accuracy will cause great losses, requires compensation and may result in the loss of orders from important customers.

In this chapter, we use KAIG model to effectively reduce RF function test items. A real case with imbalanced data is studied, and the implementation results show that our method can find relevant test items without losing Overall Accuracy.

6.2 Proposed Feature Selection Procedure

Figure 6.2 shows the basic idea of the proposed methodology. A large amount of

similar objects are gathered together to form fewer granules. When the information granulation approach is employed, numeric data will transfer to IGs and the number of positive and negative granules will be decreased compared with numeric data. The ratio of negative to positive examples will be increased. It may improve imbalanced data situation. Next, these IGs are described with appropriate form and then we can use feature selection method to extract knowledge rules or key attributes from these granules. The detailed procedure of the neural network based information granulation approach is described as follows.

Step 1: Identify condition attributes and class attributes
Step 2: Data Preprocessing
    Step 2.1: Data cleaning (Fill in missing data and remove noisy or inconsistent data)
    Step 2.2: Data transformation (Normalize or discretize the data)
Step 3: Measure the information granules
    Step 3.1: Select the degree of similarity
    Step 3.2: Check the suitability
    Step 3.3: Determine the suitable similarity
Step 4: Construct the information granules
Step 5: Define the information granules
    Step 5.1: Describe the information granules
    Step 5.2: Tackle the overlaps among the information granules
Step 6: Acquire key attributes and extract knowledge rules

Step 1 and Step 2 are data preparing phases. In these phases, we should identify the condition attributes (inputs) and the decision attributes (outputs) first. Then, data should be prepared for the process, like removing noisy data, filling missing data, and discretizing data. In step 3, the users need to determine suitable level of granularity. After that, Fuzzy ART neural network can be utilized to construct the IG, depending on the selected similarity (granularity). Next, we describe these IGs using the appropriate form. Finally, the relevant attributes can be found by feature selection methods.

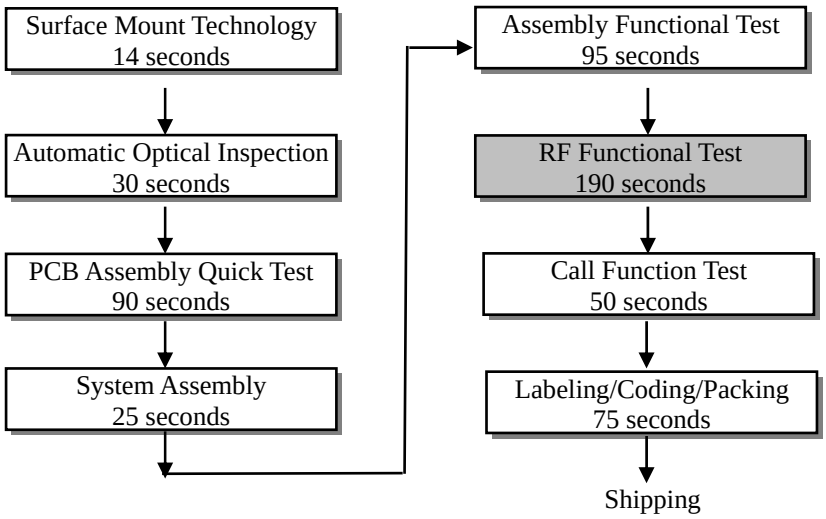


Figure 6.1 A manufacturing process of a cellular phone

In addition, real-world data tend to incomplete, noisy, and inconsistent. Data cleaning (step 2.1) routines attempt to fill in miss values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Moreover, discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals. In this study, “equal frequency bining” algorithm is utilized to discretize data. This unsupervised method is to divide the range into *b* bins of equal frequency. This method is less susceptible to outliers, and the intervals would be closer to each other in regions where there are more elements and farther apart in sparsely-populated regions, which represents the distribution of each variable better than the equal-width method. In summary, data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of data mining process.

.

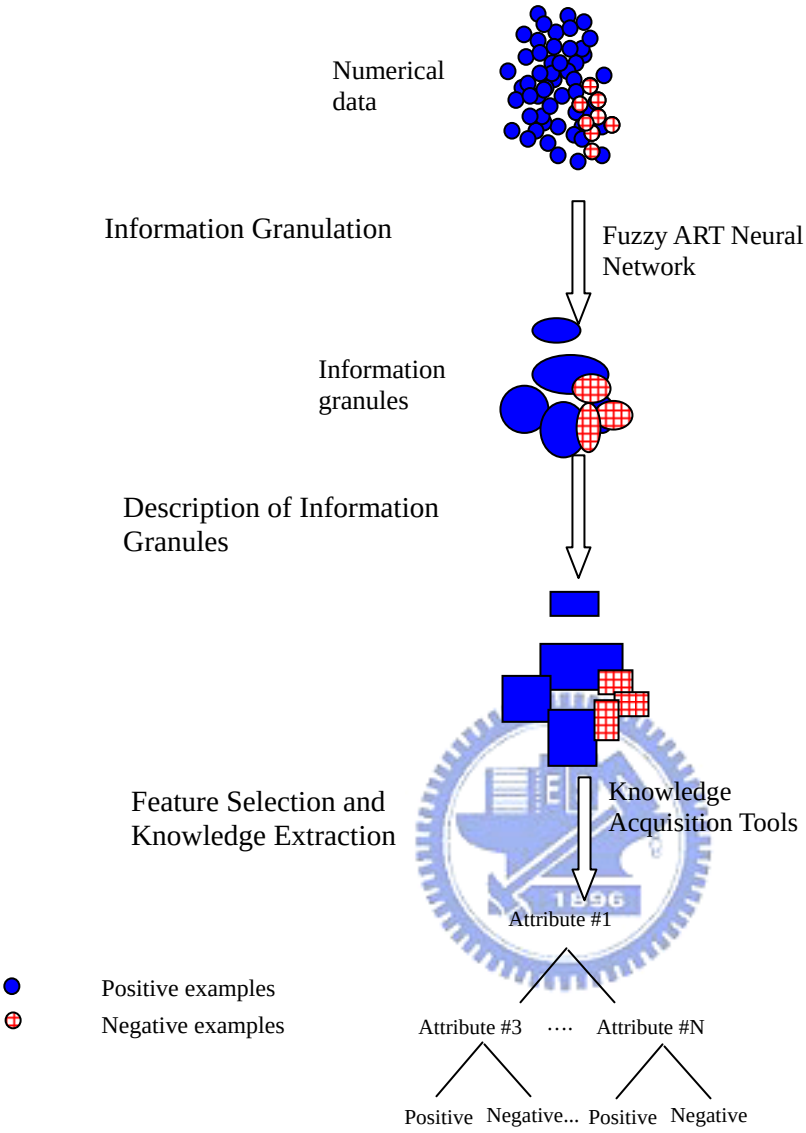


Figure 6.2 Basic idea of the proposed methodology

6.3 Case Study

The actual case comes from a cellular phone OEM/ODM company which was established in 1984. It is located in Taiwan and the company owns several factories in mainland China. In 2003, its total annual revenue reached US 4.713 billion dollars, and it has a worldwide workforce of over 10 000. The production volume of cellular phones in 2004 was about 7.5 million units.

6.3.1 The Problem

In this case, the objectives of the cellular phone manufacturer are to reduce the time-to-market and reduce test time and consequently cost. Figure 6.1 provides the manufacturing process of the cellular phone including the operation time of each process. We find that the RF functional test is the bottleneck of entire process. The RF test is aimed at inspecting whether or not the mobile phone receive/transmit signal satisfies the enabled transmission interval (ETI) protocol on different channels and different power levels. In order to ensure the quality of communication of mobile phones, the manufacturers usually add extra inspection items, such as several different frequency channels and power levels, resulting in the inspection time being increased and as a result the test procedure becomes a bottleneck.

If we can reduce the numbers of items tested in the RF function test, without losing inspection accuracy, then the inspection time will be shortened. At the same time this reduction of test items will help lower the cost of testing and the 'time-to -market'.

6.3.2 Data Collection

The 1006 RF function test data containing 62 test items (27 are continuous attributes and 35 are discrete attributes) as described in Table 6.1 are collected. There are eight major RF functional tests: the power versus time (PVT; symbol: A), the power level (TXP; symbol: B), the phase error and the frequency error (PEFR; symbol: C), the bit error rate (BER -20; symbol: D and BER -102; symbol: E), the ORFS-spectrum due to the switching transient (ORFS_SW; symbol: F), the ORFS-spectrum due to modulation (ORFS_MO; symbol: G), the Rx level report accuracy (RXP_Lev_Err; symbol: H), and the Rx level report quality (RXP_QUALITY; symbol: I). According to different channels and power levels, each

test item has several separate test attributes. Each form of the test attributes is to be represented as: test item-channel-power level. In the 1006 collected objects, there are only 44 negative examples (defective products) and the rests are positive examples (normal products). The defective rate is about 4%. We separate the 1006 examples into a training set which includes 756 objects (722 objects are normal, and 34 objects are defective) and a test set that includes 250 objects (240 objects are normal, and 10 objects are defective).

Table 6.1 Test items of the RF function

No.	Test items	Code	No.	Test items	Code	No.	Test items	Code
1	TXP	B105	22	BER(-20)	D1145	43	RXP_QUALITY	I522-102
2	PEFR	C105	23	BER(-102)	E1145	44	TXP	B6880
3	BER(-20)	D105	24	ORFS_SW	F1145	45	PFER	C6880
4	BER(-102)	E105	25	ORFS_MO	G1145	46	BER(-20)	D6880
5	ORFS_SW	F105	26	RXP_Lev_Err	H114-102	47	BER(102)	E6880
6	ORFS_MO	G105	27	RXP_QUALITY	I114-102	48	ORFS_SW	F6880
7	RXP_Lev_Err	H10-102	28	TXP	B9655	49	ORFS_MO	G6880
8	RXP_QUALITY	I10-102	29	PFER	C9655	50	TXP	B6883
9	TXP	B725	30	BER(-20)	D9655	51	TXP	B6887
10	PFER	C725	31	BER(-102)	E9655	52	TXP	B68815
11	BER(-20)	D725	32	ORFS_SW	F9655	53	RXP_Lev_Err	H688-102
12	BER(-120)	E725	33	ORFS_MO	G9655	54	RXP_QUALITY	I688-102
13	ORFS_SW	F725	34	RXP_Lev_Err	H965-102	55	TXP	B8750
14	ORFS_MO	G725	35	RXP_QUALITY	I965-102	56	PEFR	C8750
15	TXP	B727	36	TXP	B5220	57	BER(-20)	D8750
16	TXP	B7211	37	PEFR	C5220	58	BER(-102)	E8750
17	TXP	B7219	38	BER(-20)	D5220	59	ORFS_SW	F8750
18	RXP_Lev_Err	H72-102	39	BER(-102)	E5220	60	ORFS_MO	G8750
19	RXP_QUALITY	I72-102	40	ORFS_SW	F5220	61	RXP_Lev_Err	H875-102
20	TXP	B1145	41	ORFS_MO	G5220	62	RXP_QUALITY	I875-102
21	PFER	C1145	42	RXP_Lev_Err	H522-102			

6.3.3 Data Preparation

In this case, the inspection data are collected automatically by computers, and there are no missing values. In the data preparation phase we remove 11 attributes (D105,I10-102,D725,I72-102,D1145,I114-102,D9655,I965-102,D6880,I688-102,D8750) which have the same value. These 11 attributes have no classification ability.

Consequently, only 51 attributes labeled X1~X51 are left to be analyzed further. Before implementation, these collected data need to be normalized due to different scale of attributes' value, which may affect the performance of Fuzzy ART. All values of attributes were normalized to the interval [0, 1] by employing a min-max normalization equation, shown as equation (6.1). In this equation, $\max_i$ is the maximum and $\min_i$ is the minimum of the *i*-th attribute values, and v_{ij} is the value of *i*-th attribute of *j*-th objects and $v_{ij}^{'}$ is the normalized value.

$$v_{ij}^{'} = \frac{v_{ij} - \min_i}{\max_i - \min_i} \tag{6.1}$$

6.3.4 Information Granulation

Next, we utilize the Fuzzy ART to construct IGs. The proposed procedure is programmed with the use of the software of Matlab 6.1. Depending on H-index and U-ratio, the suitable similarity was determined as 0.8.

Once the similarity is determined, Fuzzy ART is again utilized to construct IGs. We set the Fuzzy ART parameters α, β, ρ to be 0.01, 1, 0.8, respectively. Thirty-three IGs are constructed. Twenty four of them are IGs of good products and the rests belong to the defective products. Each IG is described by using the lower limit and upper boundary (hyperbox form) as shown in Table 6.2. In addition, the overlapping parts among granules are separated from the original attribute by designating them as new attributes or so-called "sub-attributes." We divide the original attribute X_1 into sub-attributes X_{11}, X_{12}, X_{13}, X_{14}, X_{15}; and the same happens for the other attributes. These 33 granules are rewritten as Table 6.3.

Table 6.2 The information granules described as hyperbox form

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	Y
L1	4	3	1	1	1	1	1	1	1	1	1	1	1	2	1	2	2	1	1	1	1	5	2	1	1	1	1	3	3	1	1	1	1	1	1	3	3	1	1	1	2	3	2	1	5	2	1	1	1	1	1	1
U1	4	4	1	1	1	1	2	1	1	1	1	2	2	4	1	3	2	1	1	1	1	5	2	1	1	1	1	3	5	2	1	2	1	2	1	3	7	1	1	1	2	3	3	1	5	4	1	2	2	1	1	1
L2	3	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	4	1	1	1	1	1	3	2	1	1	2	1	1	1	3	1	1	1	1	2	3	3	1	4	1	1	1	1	1	1	1
U2	4	3	1	1	1	1	2	1	1	1	1	3	3	4	1	3	2	1	2	1	1	5	1	1	2	1	1	3	4	1	3	2	1	2	1	3	5	1	1	1	2	3	3	1	5	4	2	2	2	1	1	1
L3	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	3	2	1	1	1	1	2	1	1	1	1	1	3	1	1	1	2	1	1	1	2	1	1	1	1	2	3	3	1	3	1	1	1	1	1	1	1
U3	4	3	1	1	1	1	2	1	1	1	1	2	4	4	1	4	2	1	1	1	1	5	2	1	2	1	1	3	4	3	2	2	1	2	1	3	5	1	1	1	2	3	3	1	5	4	2	2	2	1	1	1
L4	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	2	1	1	1	3	1	1	1	1	2	3	3	1	5	2	1	1	1	1	1	1
U4	4	3	1	1	1	1	2	2	1	1	1	2	3	3	1	4	2	1	1	1	1	5	2	1	2	1	1	3	4	1	2	2	1	2	1	3	7	1	1	1	2	3	3	1	5	4	2	2	2	1	1	1
L5	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	5	1	1	1	1	1	3	1	1	1	2	1	1	1	3	1	1	1	1	2	3	2	1	5	1	1	1	1	1	1	1
U5	4	4	1	1	1	1	2	1	1	1	1	2	3	4	1	3	2	1	1	1	1	5	2	1	2	1	1	3	5	1	2	2	1	2	1	3	6	1	1	1	2	3	3	1	5	4	3	2	2	1	1	1
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
L31	3	1	1	1	1	1	1	1	1	1	1	1	1	3	1	3	1	1	1	1	1	3	2	1	1	1	1	3	1	1	1	2	1	3	1	3	1	1	1	1	2	3	3	1	5	2	1	1	1	2	1	2
U31	4	3	1	1	1	2	1	1	1	1	1	2	2	4	2	3	2	1	1	1	2	5	2	1	1	1	2	3	4	1	1	2	1	3	1	3	1	1	1	1	2	3	3	2	5	3	2	2	1	2	1	2
L32	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	3	1	1	1	1	1	1	4	1	1	1	1	1	2	2	1	1	1	1	1	1	1	2	4	1	2	1	2	2
U32	3	3	2	1	1	1	1	1	5	1	1	2	3	2	1	3	2	2	1	1	1	1	1	5	1	1	1	1	2	1	4	1	1	2	2	1	4	2	1	1	1	1	1	1	1	4	4	1	2	1	2	2
L33	3	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1	1	1	1	1	1	4	1	1	1	1	1	3	1	1	1	2	1	1	1	3	1	1	1	1	2	3	3	1	1	1	1	1	1	1	1	2
U33	4	2	1	1	1	2	2	1	1	1	1	2	2	4	1	3	2	1	1	1	2	5	2	1	2	1	2	3	4	1	1	2	1	2	1	3	1	1	1	1	3	3	3	2	5	2	2	2	1	1	1	2

Notes: (1) L1 and U1 represent the lower limit and upper limit of the 1st IG.

(2) X represents the condition attributes, and Y is the decision attribute.

(3) The data shown in the table are discretized.

Table 6.3 The IGs with the addition of sub-attributes

Original Attributes	X1					X2					X3		X4		X5		X6		..	..	..	X51		Y
Sub-attributes	X11 (X1=1)	X12 (X1=2)	X13 (X1=3)	X14 (X1=4)	X15 (X1=5)	X21 (X2=1)	X22 (X2=2)	X23 (X2=3)	X24 (X2=4)	X25 (X2=5)	X31 (X3=1)	X32 (X3=2)	X41 (X4=1)	X42 (X4=2)	X51 (X5=1)	X52 (X5=2)	X61 (X6=1)	X62 (X6=2)	..	..	..	X175 (X51=1)	X176 (X51=2)	
IG #1	0	0	0	1	0	0	0	1	1	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #2	0	0	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #3	0	1	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #4	1	1	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #5	0	0	0	1	0	1	1	1	1	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #6	0	1	1	1	0	1	1	1	1	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
IG #7	0	0	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	0	..	..	..	1	0	1
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..			
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..			
IG #27	1	0	0	0	0	0	1	1	1	1	0	1	1	0	1	0	1	0	..	..	..	1	1	2
IG #28	0	0	0	1	0	1	1	1	0	0	1	0	1	0	1	0	1	1	..	..	..	1	0	2
IG #29	0	1	1	1	0	0	0	0	0	1	1	0	1	1	1	0	1	0	..	..	..	1	0	2
IG #30	0	0	1	1	0	1	1	0	0	0	0	1	1	0	1	0	1	0	..	..	..	0	1	2
IG #31	0	0	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	1	..	..	..	1	0	2
IG #32	1	1	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	..	..	..	0	1	2
IG #33	0	0	1	1	0	1	1	0	0	0	1	0	1	0	1	0	1	1	..	..	..	1	0	2

Table 6.4 The implementation results by rough sets

Method	KAIG (rough sets)		Rough sets	
Phase	Training	Test	Training	Test
Data size (good:bad)	33 (24:9)	14 (9:5)	756 (722:34)	250 (240:10)
Pos. Accu.	100%	100%	99.3%	99.6%
Neg. Accu.	100%	90%	1000%	10%
Overall Accu.	100%	99.6%	99.34%	96%
No. of rules	4		433	
Extracted features	**B725, H114-102**		C105, B727, B1145, H114-102, C9655, H522-102, B8750, E8750	

Note: (24:9) is the proportion of good products to bad products.

Table 6.5 The implementation results by decision tree (C 4.5)

Method	KAIG (DT)		DT	
Phase	Training	Test	Training	Test
Data size (good:bad)	33 (24:9)	14 (9:5)	756 (722:34)	250 (240:10)
Pos. Accu.	100%	100%	100%	100%
Neg. Accu.	100%	90%	76.47%	60%
Overall Accu.	100%	99.6%	98.9%	98.4%
No. of rules	3		7	
Extracted features	**B725, H114-102**		E105, C725, G725, H72-102, H965-102, H688-102	

Table 6.6 The implementation results by BPNN (full attributes)

Method	KAIG (BPNN)		BPNN	
Phase	Training	Test	Training	Test
Data size (good:bad)	33 (24:9)	14 (9:5)	756 (720:34)	250 (240:10)
Pos. Accu.	99.5%	100%	99.86%	100%
Neg. Accu.	88.24%	100%	70.59%	50%
Overall Accu.	98.9%	100%	98.54%	98%
Structure	16-15-1		17-4-1	
Parameters	Learning rate: 0.2 Momentum: 0.9 50000 iterations		Learning rate: 0.2 Momentum: 0.8 2000 iterations	
Extracted features	B7211, **H114-102**, E8750, B8750, C8750, **B725,** H965-102, H688-102, H10-102, B727, E5220, B7219, C105, C6880, C9655, B68815		C9655, B725, C725 ,B8750, B105, B727, C8750, F1145, B5220, B7211, H114-102, B6880, B68815, F725, B6887, E1145, I522-102	

6.3.5 Feature Selection and Knowledge Acquisition

Now three feature selection algorithms, rough sets method, decision tree (C 4.5 algorithm) and neural network, are implemented. The inputs and outputs of the decision tree and rough sets are 176 sub-attributes and defined classes respectively. In the neural network based method, the back-propagation neural network with one hidden layer is adopted and implemented using Professional II PLUS software. All parameters of the BPNN are obtained by trial and error, including the number of training iterations and the structure of the network.

Implementation results are shown in Tables 6.4~6.6. In Tables 6.4 and 6.5, our proposed approach obviously outperforms the traditional approach without granulation, in both classification accuracy and type II error. In addition, fewer knowledge rules and attributes are obtained. In Table 6.6, the classification accuracy and type II error of our approach are still better than those by the original BPNN. All the attributes, kept and ranked by priority, are listed in Table 6.6. By comparing the implementation results of these three methods, six attributes {B7211, H114-102, B725, B8750, C8750 and E8750} are reserved as final test items for the RF functional test. The knowledge rules listed in Figures 6.7(a) & (b) are generated by using rough sets and decision tree methods. These rules may not only help engineers to predict the yield rate of products, but may also enhance the performance of knowledge management.

6.4 The Benefits

By implementing the proposed method, test items are reduced from 62 to 6 items. The test time is reduced from 190 seconds to 95 seconds. The amount of employed test equipment is reduced from 8 machines to 4 machines. As a result the company will save about $ US 200 000 per year. In addition we should not forget the resulting

rise in customer satisfaction and the reduction in risk for the customers. The potential benefits of implementation are substantial.

Rule 1:
$B725 \in [31.7440,31.7446)$ AND $H114-102 \notin (1,*]$ TEHN Class = Good Product
[Accuracy: 1.0; Supports: 24]
Rule 2:
$B725 \in [31.7440,31.7446)$ AND $H114-102 \in (1,*]$ TEHN Class = Bad Product
[Accuracy: 1.0; Supports: 2]
Rule 3:
$B725 \notin [31.7440,31.7446)$ AND $H114-102 \notin (1,*]$ TEHN Class = Bad Product
[Accuracy: 1.0; Supports: 4]
Rule 4:
$B725 \notin [31.7440,31.7446)$ AND $H114-102 \in (1,*]$ TEHN Class = Bad Product
[Accuracy: 1.0; Supports: 3]

Figure 6.7 (a). Knowledge rules extracted by rough sets

Rule 1:
$B725 \in [31.7440,31.7446)$ AND $H114-102 \notin (1,*]$ TEHN Class = Good Product
[Accuracy: 0.962; Supports: 24]
Rule 2:
$B725 \notin [31.7440,31.7446)$ TEHN Class = Bad Product
[Accuracy: 0.889; Supports: 7]
Rule 3:
$H114-102 \in (1,*]$ TEHN Class = Bad Product
[Accuracy: 0.857; Supports: 5]

Figure 6.7 (b). Knowledge rules extracted by decision tree (C 4.5)

6.5 Discussion and Concluding Remarks

In most cases of inspection data, the amount of good products is far greater than the amount of defective products. The few defective products are usually viewed as outliers and are removed in the generalization phase of the classification tools. Actually, all normal products look alike, and the abnormal products have individual styles. That phenomenon is also noted by Taguchi and Jugulum [22]. We should pay

more attention to this, and consider the categories of instances instead of the data size when developing feature selection algorithms.

Traditional machine learning techniques tend to generate a huge amount of knowledge rules and lead to a low level of Negative Accuracy when dealing with imbalanced data. This chapter applied KAIG model to select key test items in mobile phone inspection. KAIG model not only extracts fewer knowledge rules, but also outperforms the traditional methods regarding the Negative Accuracy and Overall Accuracy.

A real case study of cellular phone test process was employed to demonstrate the effectiveness of KAIG model. When encountering imbalanced data, KAIG model is effective in removing unnecessary RF function test items, saving testing costs and shortening the time to market. It is suitable for reducing the inspection process in the high technology industry, especially now that we are facing the six-sigma age, i.e. the defective rate of products is becoming extremely low.

The experimental results also show that there is a trade-off relationship between Positive Accuracy and Negative Accuracy. The KAIG model can raise the level of Negative Accuracy without dropping Positive Accuracy. This is very important to OEM/ODM manufacturers because a low level of Negative Accuracy will inevitably lead to orders being lost. The inconsistence of the extracted attributes when using different feature selection methods is an important issue for future research, because it might confuse users (engineers) when applying these feature selection techniques in practice. To solve the inconsistence, a robust approach is needed to be developed in the future.

CHAPTER 7

APPLY IG BASED METHOD TO ENCHANCE THE DIABETES DIAGNOSIS ABILITY

7.1 Problem Description

Diabetes, which can result in a variety of complications, including heart disease, kidney disease, eye disease, erectile dysfunction, and nerve damage, has become a serious problem in human society. In 2000, the World Health Organization (WHO) estimated that over 177 million people had diabetes. By 2030, it is estimated that there will be 366 million diabetes patients world-wide. The WHO is calling diabetes an epidemic, and recently it is having a huge economic impact on countries in Africa, India, and China. Diabetes is a bigger killer than AIDS, and the cost of supporting a person who has lost a foot due to diabetes may drain three-quarters of the income of a poor family. In Taiwan, diabetes has been the 5th leading cause of death since 1987 and it became number 4 in 2004.

In recent years, researchers have tried to use artificial intelligence (AI) methods to build diagnostic classifiers (Srikanth *et al*., 1997) in order to identify diseases quickly and economically, and therefore help diagnose patients in those developing countries that lack sufficient medical resources. The AI methods acquire knowledge from examples of existing diagnoses and apply the extracted knowledge to diagnose an illness. However, the data obtained from examples of diagnoses are often imbalanced or skewed, with almost all the instances being labeled as one class, while only a scant few instances are labeled as the other class, usually the important class. When building a classifier from such imbalanced/skewed diagnosis data, class imbalance problems are necessary to be considered. In this chapter, we apply IG based method to increase the ability of identifying diabetic patients.

7.2 Data Collection

The experimental data comes from the health examination database of a regional hospital in Taiwan. We obtained 2000 raw data. After removing missing value examples, 1829 objects remained, which contain 1729 positive instances (healthy patients) and 100 negative instances (diabetic patients). These examples are divided into training (1464: 1384+80) and test objects (365: 345+20). The proportion of negative examples is about 5%.

Table 7.1 shows 23 attributes of this data, they are biochemical or physical test items and all their values are continuous except for the first one "Gender." Although there are different types of diabetes like Type 1, Type 2 and gestational diabetes, they are combined and considered as diabetes. Therefore we have 2 classes, positive (healthy patients) and negative (diabetic patients).

Table 7.1 Attributes

#1	Gender	#5	FEV1	#9	SGOT	#13	BUN	#17	Thyroxine	#21	HDL
#2	Age	#6	PFR	#10	SGPT	#14	Creatinine	#18	Uric acid	#22	ELDL
#3	Vital Capacity	#7	albumin	#11	APAE	#15	Glucose AC	#19	Cholesterol	#23	LDL
#4	Predicted VC	#8	Total protein	#12	Total bilirubin	#16	Glucose PC	#20	Triglyceride		

7.3 Implementation of IG based method

In this section, we try to build a classifier to identify diabetic patients automatically by using IG based method. The comparison of our method, SVM, DT (C 4.5), cluster based sampling, and under-sampling will be provided to show the superiority of IG based method in highly class imbalance situation.

Table 7.1 shows the results of H-index and U-ratio. Seventy-nine IGs are constructed depending on the selected similarity (vigilance), 0.7. Then, three quartiles,

Q1, Q2, and Q3 are utilized to describe those IGs which are shown in Table 7.2.

Table 7.2 Using Q1+Q2+Q3 to describe IGs

Data char.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	class	IG No.
Q3	1	57	4.16	128.8	86	143	5	8	30	40	71	1.1	16	1.1	100	96.75	1.4	6.3	238	109	57	22	173	+1	1
Q2	1	50	3.22	110	82	101	4.7	8	17	23	62.5	0.9	14	1.1	91	92.5	1.1	5.9	219	99	52	20	142	+1	1
Q1	0	37	2.725	77.25	80	80.5	4.6	7	17	15	57.3	0.7	13	0.9	85	84.5	0.7	5.1	198	78.3	43	16	123	+1	1
Q3	1	68	3.785	116.4	88	137	4.8	8	25	22	152	1.1	20	1.2	90.8	120.8	1.2	8.7	187	163	58	33	114	+1	2
Q2	1	60	2.98	109.5	84	127	4.5	7.5	20	20	93	1	19	1.2	89.5	92	0.9	6.8	174	128	44	26	111	+1	2
Q1	0.3	44	2.61	101	79	103	4.1	7	17	15	70.8	0.8	16	1.1	86	88	0.7	6.6	171	80.5	31	16	108	+1	2
Q3	1	60	3.04	108	85	82	4.6	8	19	16	87	1.6	19	1.1	101	120	4.8	6	184	112	57	22	110	+1	3
Q2	0	54	2.47	96	80	78	4.5	7	14	13	53	1	19	0.9	98	88	1.4	5.8	168	94	55	19	81	+1	3
Q1	0	51	1.76	71	79	75	4.4	7	14	11	45	0.7	16	0.8	91	74	0.9	5.3	146	74	52	15	70	+1	3
Q3	1	62	3.51	107	87	142	4.7	7	25	48	57.5	1.1	18	1.2	103	106.5	2.9	6.9	252	175	67	35	169	+1	4
Q2	1	58	2.985	94.6	84	126	4.5	7	23	27	54	0.9	16	1.1	91.5	95	1.5	6.4	223	115	58	23	120	+1	4
Q1	0.8	50	2.62	89	83	101	4.4	7	19	16	41.8	0.7	15	1	87	91.75	0.8	6.1	193	88.8	48	18	106	+1	4
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Q3	1	71	2.64	99.5	83	89	4.5	7	30	37	85	0.9	17	1.2	135	224.5	4.3	7.4	210	171	46	34	133	-1	75
Q2	0.5	66	2.27	79.45	75	68	4.1	7	22	25	74	0.7	15	1	127	206.5	2.2	6.2	189	144	42	29	112	-1	75
Q1	0	60	2.0175	65	69	47.9	4	7	20	20	63	0.7	13	0.9	94.5	189.5	0.9	5.4	154	116	38	23	88	-1	75
Q3	1	56	3.3825	105.5	88	139	4.9	8	25	37	84	1.1	19	1.2	221	256	1.7	7.4	245	200	52	40	142	-1	76
Q2	1	51	2.87	98	85	124	4.8	8	22	29	79.5	0.9	18	0.9	137	218	1.1	6.5	189	188	50	38	104	-1	76
Q1	0	47	2.585	78.25	83	102	4.7	7	20	21	62.3	0.7	15	0.8	118	137.5	1	5.8	152	99.5	48	20	74.3	-1	76
Q3	1	64	3.57	125	87	184	4.9	8	32	44	78	0.9	20	1.2	258	278	1.1	6.2	199	145	61	29	115	-1	77
Q2	1	64	3.53	108	82	134	4.9	8	19	25	71	0.9	20	1.2	146	224	1	5.9	188	115	52	23	112	-1	77
Q1	1	56	2.95	108	76	79	4.5	8	18	20	63	0.8	16	1	113	222	0.9	5.7	159	79	46	16	91	-1	77
Q3	1	60	3.3525	101.5	77	94.3	4.3	8	34	53	94.5	1	17	1.2	264	427.3	1.6	5.5	274	362	56	69	164	-1	78
Q2	1	54	3.075	100.5	75	68.5	3.9	7.5	24	29	83.5	0.8	14	1	222	338	1.3	5	256	228	45	46	153	-1	78
Q1	1	49	2.9025	94.75	71	51.5	3.7	6.8	20	25	76	0.7	11	0.9	198	285	1	4.5	226	152	37	31	132	-1	78
Q3	0	67	2.48	118	98	177	5	7	30	17	71.5	0.9	18	1.1	135	166	1.9	7	246	230	70	46	145	-1	79
Q2	0	66	2.38	114	96	175	4.9	7	23	14	60	0.8	16	1	134	161	1.6	6.3	242	195	63	39	140	-1	79
Q1	0	66	2.28	110	95	174	4.7	7	17	11	48.5	0.8	14	1	132	156	1.3	5.5	237	159	55	32	135	-1	79

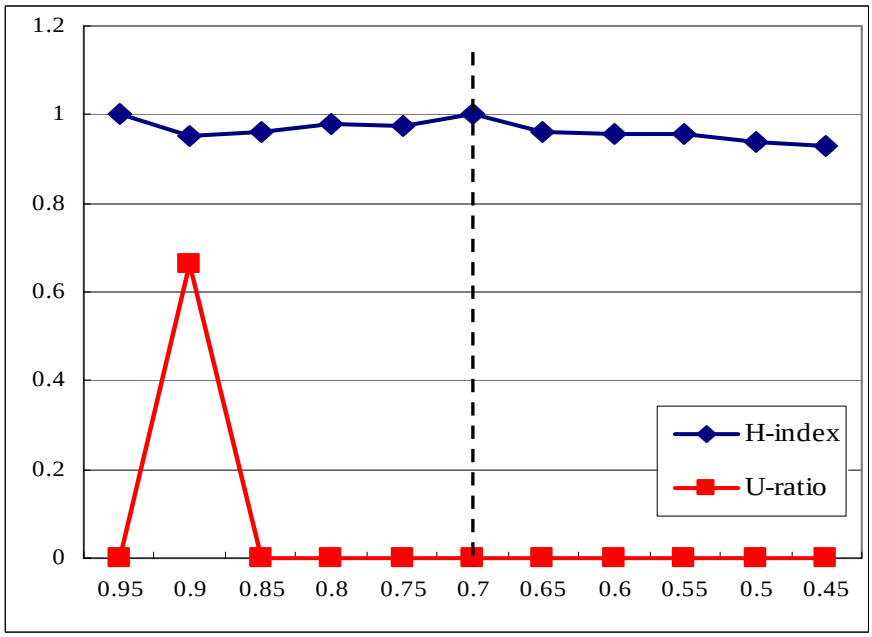


Figure 7.1 H-index & U-ratio of diabetes diagnosis data

Table 7.3 summarizes the results of experiments. Although Overall Accuracy and Positive Accuracy drop slightly, IG based method has excellent performance in Negative Accuracy (75%), which is better than cluster based sampling (65%), under-sampling (45%), SVM (10%), and DT (15%). It means IG based method can remarkably increase the ability of detecting diabetic patients. Considering G-mean, IG based method also outperforms other methods. The comparison is also demonstrated in Figure 7.2.

Table 7.3 Summary of experimental results

Methods / Performance	IG based method	Cluster based sampling	Under- sampling	SVM	DT
Overall Accuracy	93.70%	92.88%	95.07%	95.07%	95.34%
Pos. Accu.	94.78%	94.49%	97.97%	100%	100%
Neg. Accu.	75%	65%	45%	10%	15%
G-mean	84.31%	78.37%	66.40%	31.62%	38.73%

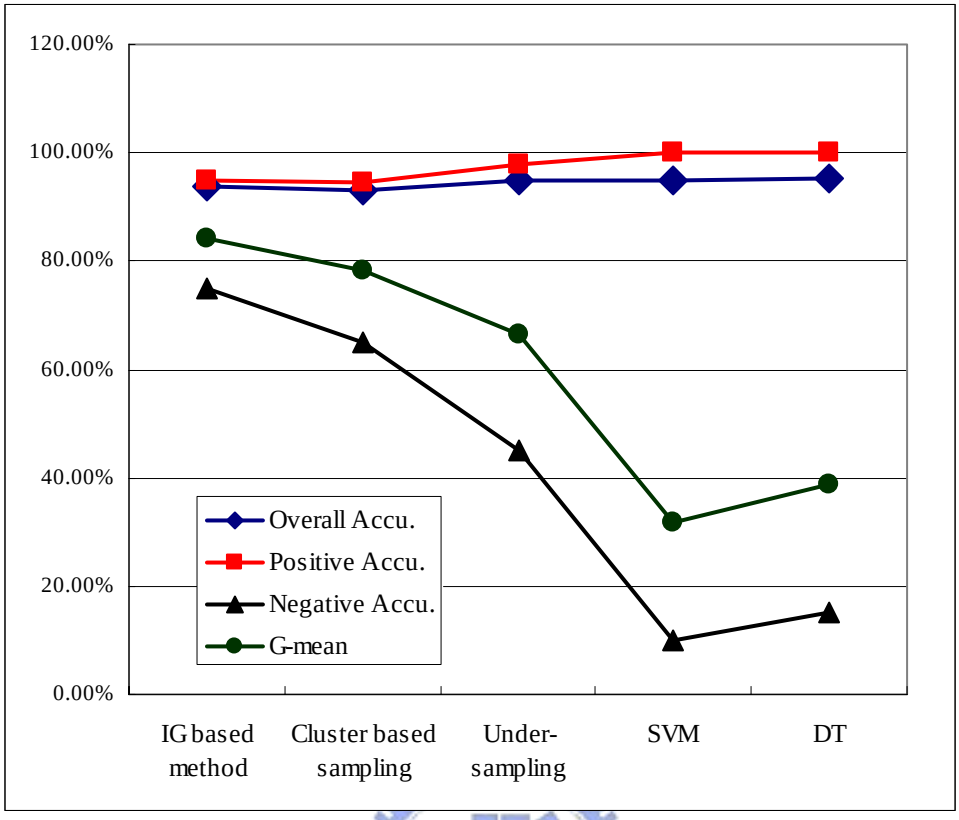


Figure 7.2 The comparison of experimental results

7.4 Discussion and Concluding Remarks

Medical diagnosis data is usually highly imbalanced. It's especially true when diseases are rare. When applying machine learning techniques to build classifiers to diagnose an illness, it's necessary to consider class imbalance problems. The chapter applied IG based method to diabetes diagnosis. An actual case was provided to illustrate the effectiveness of our method. The experimental results show our proposed method can significantly increase the ability of detecting diabetic patients (minority) in highly skewed data situation.

CHAPTER 8

CONCLUSIONS

8.1 Summary

Granular computing, which imitates human capability of performing tasks or processing information, is a new direction of Artificial Intelligence. Albert Einstein (1879-1955) said “As far as the laws of mathematics refer to reality, they are not certain; as far as they are certain, they do not refer to reality.” His words can be used to explain why researchers paid lots of attentions on uncertainty/vagueness in human decision making, such as fuzzy sets, rough sets, granular computing, and etc. These researches are not intended to replace traditional measurement-based methods which operate numerical data. Their purpose is to let the developed computational theories refer to reality.

The main contributions of this dissertation are to propose two practical GrC models, KAIG and IG based method, for dealing with discrete and continuous data, respectively. The first proposed KAIG indeed has impressive classification ability for imbalanced data, but it also has a drawback of generating so many “sub-attributes” that may increase computational complexity when dealing with continuous data. We can control the number of “sub-attributes” by adopting a discretization algorithm. However, in some situations, the discretization algorithm might still generate lots of discrete intervals. Therefore, we proposed the second IG based method to aim at continuous data. Unlike traditional data mining approaches which acquire knowledge from numerical data, our methods can extract knowledge from IGs while controlling the reduction of information (i.e. removing unnecessary details). In both GrC models, the procedures to find appropriate number of IGs and to represent the constructed IGs

were provided. Integrating Fuzzy ART neural network and two objective indexes (H-index and U-ratio), an information granulation procedure was presented to construct IGs. This study also developed two kinds of IG representation methods, the concept of "sub-attributes" and "data characteristics." Experimental results show their efficacy to represent IGs.

In addition, this study also discussed class imbalance problems which are becoming serious issues when applying data mining techniques to practical areas, such as spam detection, defective products inspection, and diseases/illness diagnosis. Recently, we observe an increase of research activity in data mining from imbalanced data sets. This increase in interest gave rise to two workshops held in 2000 (the American Association for Artificial Intelligence; AAAI) and 2003 (International Conference on Machine Learning; ICML) on learning from imbalanced data sets. The Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, *SIGKDD Explorations*, also published a special issue to discuss the same issue (June 2004).

In related researches, some presented techniques like re-sampling and weight-adjusting techniques can merely slightly improve the imbalanced situation, but they cannot guarantee optimal solutions. Compared with IG based method, cluster-based sampling methods lack a systematic mechanism to determine a suitable number of clusters and to find representative examples to denote clusters. Nevertheless, our methods provided an effective avenue to solve these problems. Considering ROC curve analysis and other performance indexes, KAIG and IG based method have been shown their superiority of identifying minority examples without losing overall accuracy.

In practical applications, we found proposed GrC models actually can increase classification performance by reducing detailed information. Our methods are

especially effective for imbalanced inspection, diagnosis, faults monitoring, and fraud detecting data. The case studies showed KAIG model improved predictive ability of a scheduling system within an FMS environment. KAIG model also removed redundant test items and shortened inspection time of cellular phone manufacturing. In addition, the IG based model was proposed to deal with continuous data and it effectively increased the ability of a classifier to detect rare diabetic patients.

8.2 Further Research

Some potential directions for improvement and future work are clear. The first issue is about reduction of information. How many unnecessary details should be removed? Although we employed two indexes to tackle this issue, it's necessary to pay more efforts to do advanced researches.

The second issue is about class imbalance problems. Usually, data from two classes are available in conventional classification; the decision boundary is supported from both sides of example objects. When the class distribution is extremely skewed or the minority examples are completely absent, the traditional algorithms might fail. One-class classification method which is assumed that only one class information available might be the possible solution.

REFERENCES

[1] Agilent Technologies, Agilent's TS-5550 cellular phone functional test platform, http://www.home.agilent.com.

[2] Akbani, R., Kwek, S. and Japkowicz, N., 2004, Applying support vector machines to imbalanced datasets, *ECML 2004: 15th European Conference on Machine Learning* (J.-F. Boulicaut et al. Eds.), LNAI 3201, pp. 39-50.

[3] Almuallim H. and Dietterich, T., 1994, Learning boolean concepts in the presence of many irrelevant features, *Artificial Intelligence*, **69**, pp. 279–305.

[4] Altincay, H. and Ergun, C., 2004, Clustering based under-sampling for improving speaker verification decisions using AdaBoost, *Lecture Notes in Computer Science*, **3138**, pp. 698-706.

[5] An, A. and Wang, Y., 2001, Comparisons of classification methods for screening potential compounds, *Proceedings of the IEEE International Conference on Data Mining (ICDM.01)*, San Jose, CA, pp. 11-18.

[6] Aytug, H., Bhattacharyya, S., Koehler, G. J., and Snowdon, J. L., 1994, A review of machine learning in scheduling, *IEEE Transactions on Engineering Management*, **41** (2), pp. 165-171.

[7] Bargiela, A. and Pedrycz, W., 2003a, *Granular Computing: an introduction*, Kluwer Academic Publishers, Boston.

[8] Bargiela, A. and Pedrycz, W., 2003b, Recursive information granulation: aggregation and interpretation issues, *IEEE Transactions on Systems, Man, and Cybernetics*, **33** (1), pp. 96–112.

[9] Batista, G., Prati, R.C. and Monard, M.C., 2004, A study of the behavior of several methods for balancing machine learning training data, *SIGKDD Explorations*, **6** (1), pp. 20-29.

[10] Bortolan, G. and Pedrycz, W., 2002, Fuzzy descriptive models: an interactive framework of information granulation, *IEEE Transactions on Fuzzy Systems*, **10** (6), pp. 743-755.

[11] Bruzzone, L. and Serpico, S.B., 1997, Classification of imbalanced remote-sensing data by neural networks, *Pattern Recognition Letters*, **18**, pp. 1323–1328.

[12] Burke, L. and Kamal, S., 1995, Neural networks and the part family/machine group formation problem in cellular manufacturing: a framework using fuzzy ART, *Journal of Manufacturing Systems*, **14** (3), pp. 148-159.

[13] Carpenter, G.A., Grossberg, S. and Rosen, D.B., 1991, Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Networks*, **4**, pp. 759-771.

[14] Castellano, G. and A. M. Fanelli, 2001, Information granulation via neural network-based learning," *IFSA World Congress and 20th NAFIPS International Conference*, **5**, pp. 3059-3064.

[15] Chawla, N. V., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002, SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, pp. 231-357.

[16] Chawla, N. V., Japkowicz, N. and Kolcz, A., 2004, Editorial: special issue on learning from imbalanced data sets, *SIGKDD Explorations*, **6** (1), pp. 1-6.

[17] Cristianini, N. and Shawe-Taylor, J., 2000, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge.

[18] Estabrooks, A, Jo, T. and Japkowicz, N., 2004, A multiple resampling methods for learning from imbalanced data sets, *Computational Intelligence*, **20** (1), pp. 18-36.

[19] Gersmann, K. and Hammer, B., 2005, Improving iterative repair strategies for scheduling with the SVM, *Neurocomputing*, **63**, pp. 271-292.

[20] Grzymala-Busse, J. W., Stefanowski, J. and Wilk, S., 2004, A comparison of two approaches to data mining from imbalanced data, *Lecture Notes in Computer Science*, **3213**, pp. 757-763.

[21] Guo, H. and Viktor, H. L., 2004, Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, *SIGKDD Explorations*, **6** (1), pp. 30-39.

[22] Han, J. and Kamber, M., 2001, *Data mining: concepts and techniques*, Morgan Kaufmann publishers, San Francisco.

[23] Hand, D. J., 1997, *Construction and assessment of classification rules*, John Wiley & Sons, New York.

[24] Hannikainen, M., Hamalainen, T. D., Niemi, M. and Sarinen, J., 2002, Trends in personal wireless data communications, *Computer Communications*, **25**, pp. 84-99.

[25] Hu, X., Cercone, N., Han, J. and Ziarko, W., 2002, GRS: a generalized rough sets model, *Data Mining, Rough sets and Granular Computing*, in: T. Y. Lin, Y. Y. Yao and L. A. Zadeh (Eds.), Physica-Verlag, New York, pp. 447-460.

[26] Huang, K., Yang, H., King, I. and Lyu, M., 2004, Learning classifiers from imbalanced data based on biased minimax probability machine, *Proceedings of the 04' IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 558-563.

[27] Japkowicz, N. and Stephen, S., 2002, The class imbalance problem: a systematic study, *Intelligent Data Analysis*, IOS Press, **6**, pp. 429–449.

[28] Jo, T. and Japkowicz, N., 2004, Class imbalances versus small disjuncts, *SIGKDD Explorations*, **6** (1), pp. 40-49.

[29] Li, D.-C. and She, I.-S., 1994, Using unsupervised learning technologies and simulation analysis to induce scheduling knowledge for flexible manufacturing systems, *International Journal of Production Research*, **32**, pp. 2187–2199.

[30] Li, D.-C., Chen, L.-S., and Lin, Y.-S., 2003, Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments, *International Journal of Production Research*, **41**, pp. 4011–4024.

[31] Li, R. and Wang, Z., 2004, Mining classification rules using rough sets and neural networks, *European Journal of Operational Research*, **157**, pp. 439-448.

[32] Lin, T.Y., 1998, Granular computing on binary relations: Part I and II, in: Skoworn, A. and Polkowski, L. (Eds), *Rough sets in Knowledge Discovery*, pp. 107-140, Springer-Verlag: Berlin, Germany.

[33] Manevitz, L. M. and Yousef, M., 2001, One-class SVMs fro document classification, J*ournal of Machine Learning Research*, **2**, pp. 139-154.

[34] Min, H. and Yih, Y., 2003, Development of a real-time multi-objective scheduler for a semiconductor fabrication system, *International Journal of Production Research*, **41**, pp. 2345–2364.

[35] Mitra, S., Pal, S. K. and Mitra, P., 2002, Data mining in soft computing framework: a survey, *IEEE Transactions on Neural Networks*, **13** (1), 3-14.

[36] Montgomery, D. C., 2005, *Introduction to statistical quality control*, John Wiley & Sons, USA.

[37] Oyeleye, O. and Lehtihet, E. A., 1998, A classification algorithm and optimal feature selection methodology for automated solder joint defect inspection, *Journal of Manufacturing Systems*, **17**, pp. 251-262.

[38] Pawlak, Z., 1985, Rough sets and fuzzy sets, *Fuzzy Sets and Systems*, **17**, pp. 99-102.

[39] Pedrycz, W. and Bargiela, A., 2002, Granular clustering: a granular signature of data, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, **32** (2), pp. 212-224.

[40] Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N. and Benner, M., 1999, Association, statistical, mathematical and neural approaches for mining breast cancer patterns, *Expert Systems with Applications*, **17**, pp. 223-232.

[41] Provost, F. and Fawcett, T., 2001, Robust classification for imprecise environments, *Machine Learning*, **42**, pp. 203-231.

[42] Quinlan, J., 1986, Induction of decision trees, *Machine Learning*, **1**, pp. 81–106.

[43] Quinlan, J., 1993, *C4.5: Programs for machine learning*, Bari: Morgan Kaufmann.

[44] Radivojac, P., N. C. Chawla, A. K. Dunker and Z. Obradovic, 2004,

Classification and knowledge discovery in protein databases, *Journal of Biomedical Informatics*, **37**, 224-239.

[45] Reed, R., 1993, Pruning algorithms- a survey, *IEEE Transactions on Neural Networks*, **5**, pp. 740-747.

[46] Serrano-Gotarredona, T., Linares-Barranco, B. and Andreou, A. G., 1998, *Adaptive Resonance Theory Microchips-Circuit Design Techniques*, Kluwer Academic Publishers, Boston.

[47] Sim, S. K., Yeo, K. T. and Lee, W. H., 1994, An expert neural network system for dynamic job shop scheduling, *International Journal of Production Research*, **32**, pp. 1759-1773.

[48] Su, C.-T., Hsu, J.-H. and Tsai, C.-H., 2002, Knowledge mining from trained neural network, *Journal of Computer Information Systems*, pp. 61-70.

[49] Su, C.-T. and Shiue, Y.-R., 2003, Intelligent scheduling controller for shop floor control systems: a hybrid genetic algorithm/decision tree learning approach, *International Journal of Production Research*, **12**, pp. 2619–2641.

[50] Swiniarski, R. W. and Hargis, L., 2001, Rough sets as a front end of neural-networks texture classifiers, *Neurocomputing*, **36**, pp. 85-102.

[51] Swiniarski, R. W. and Skowron, A., 2003, Rough set methods in feature selection and recognition, *Pattern Recognition Letters*, **24**, pp. 833-849.

[52] Taguchi, G. and Juoulum, R., 2002, *The Mahalanobis-Taguchi strategy- a pattern technology system*, John Wiley & Sons, New York.

[53] Verikas, A. and Bacauskiene, M., 2002, Feature selection with neural networks, *Pattern Recognition Letters*, **23**, pp. 1323-1335.

[54] VI Services Network, Mobile phone online test application, www.VI-china.com.cn

[55] Walczak, B. and Massart, D. L., 1999, Tutorial: rough sets theory, *Chemometrics and Intelligent Laboratory Systems*, **47**, pp. 1-16.

[56] Wang, C.-S. and Uzsoy, R., 2002, A genetic algorithm to minimize maximum lateness on a batch processing machine, *Computers and Operations Research*, **29** (12), pp. 1621-1640.

[57] Wu, G. and Chang, E. Y., 2005, KBA: kernel boundary alignment considering imbalanced data distribution, *IEEE Transactions on Knowledge and Data Engineering*, **17** (6), pp. 786- 795.

[58] Xiu, B.-X., Zhang, W.-M, Wang, S. and Tang, J.-Y., 2003, Generalized multilayer granulation and approximations, *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pp. 1419-1423.

[59] Yao, Y.Y., 1999, Granular computing using neighborhood systems, in: Roy, R., Furuhashi, T., and Chawdhry, P.K. (Eds.), *Advances in Soft Computing: Engineering Design and Manufacturing*, Springer- Verlag, London, U.K., pp.

539-553.

[60] Yao, Y. Y. and Yao, J. T., 2002, Granular computing as a basis for consistent classification problems, *Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining*, pp. 101-106.

[61] Zadeh, L. A., 1979, Fuzzy sets and information granularity, in: Gupta, M. M., Ragade, R. K., and Yager, R. R. (Eds.), *Advances in Fuzzy Set Theory and Applications*, North Holland, Amsterdam, pp. 3-18.

[62] Zadeh, L. A., 2001, A new direction in AI: Toward a computational theory of perceptions, *AI Magazine*, **22** (1), pp. 73-84.

[63] Zhu, F. and Guan, S., 2004, Feature selection for modular GA-based classification, *Applied Soft Computing*, **4**, pp. 381-393.

陳隆昇

學歷

1. 國立交通大學工業工程與管理系博士（91.09 ~ Now）
2. 國立成功大學工業管理研究所碩士 （87.09 ~ 89.07）
3. 國立成功大學工業管理科學系學士 （83.09 ~ 87.07）

經歷

1. 國科會九十四年千里馬計畫--普渡大學訪問學生 (94.01~94.08)
2. 成功大學工管系系學會會長 （85.09 ~ 86.07）
3. 陸軍步兵 168 旅少尉行政官（89.12～91.03）

榮譽與認證

1. 八十七年度救國團大專優秀青年
2. 摩托羅拉大學六標準差綠帶認證(Motorola University Green Belt Certification)
3. 九十一年陸軍總司令部獎狀

論文發表

(A)期刊論文

[1] Li, D.-C., **Chen, L.-S**. and Lin, Y.-S. (2003), "Using Functional Virtual Population as assistance to learn scheduling knowledge in dynamic manufacturing environments," *International Journal of Production Research*, 17, 4011–4024 (**SCI**, **EI**)

[2] Su, C.-T., **Chen, L.-S.** and Yih, Y. (2006), "Knowledge acquisition through information granulation for imbalanced data," *Expert System with Applications*, 31(4). (**SCI**, **EI**) (In press)

[3] Su, C.-T., **Chen, L.-S.**, Chiang, T.-L. (2006), "A neural network based information granulation approach to shorten the cellular phone test process," *Computers In Industry* (**SCI, EI**). (In press)

[4] **Chen, L.-S**. and Su, C.-T. (submitted), "Using granular computing model to induce scheduling knowledge in dynamic manufacturing environments," *International Journal of Computer Integrated Manufacturing* (**SCI, EI**)

(B)研討會論文

[1] Su, C.-T., Yang, T., **Chen, L.-S** and Chiang, T.-L., A novel method to shorten inspection process: a case study, INCOM 2006: 12th IFAC Symposium on Information Control Problems in Manufacturing, Saint-Etienne, France.

[2] **Chen, L.-S.** and Su, C.-T., 2006, Multiple techniques for learning from imbalanced/skewed data, 36th International Conference on Computers and Industrial Engineering, Taipei, Taiwan (submitted Ref. No. CIE00107).

[3] **Chen, L.-S.** and Su, C.-T., 2005, A study of improving the classification accuracy in diabetes diagnosis, *the 3rd ANQ Congress & the 19th Asia Quality Symposium*, Taipei, Taiwan

[4]. **Chen, L.-S.** and Su, C.-T., 2004, Information granulation to reduce the amount of massive data, *IIE Annual Conference*, Houston, USA

[5]. **Chen, L.-S.**, Siao, Y.-S. and Su, C.-T., 2004, MTS for Mobile Phone Test Process, *the 2nd ANQ Congress & the 18th Asia Quality Symposium*, New Delhi, India

[6]. Li, D.-C., Heish, K. Y., and **Chen, L.-S.**, 1998, Two Model to Reduce Size of Decision Tree in Inductive Learning, in *Proceedings of the Quantitative Management Techniques and Applications in Taiwan Conference*, pp. 356-363, Tainan, Taiwan