行政院國家科學委員會專題研究計畫　成果報告

應用資料探勘技術於醫院疾病資料庫之分析：以糖尿病為例

計畫類別：個別型計畫
計畫編號：NSC91-2320-B-040-028-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：中山醫學大學公共衛生系

計畫主持人：呂宗學

報告類型：精簡報告

處理方式：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中　華　民　國 93 年 2 月 11 日

行政院國家科學委員會專題研究計畫成果報告
應用資料探勘技術於醫院疾病資料庫之分析：以糖尿病為例

Use of data mining techniques in hospital-based disease dataset analysis: diabetes as an example

計畫編號：NSC91－2320－B－040－028－
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：呂宗學　　中山醫學大學公衛系

一、中英文摘要

背景：目前台灣大多數醫院的門診與住院相關的資料大多已經數位化，可惜這些例行累積的龐大訊息並沒有好好被利用來提供疾病管理有用的訊息。
目標：嘗試以資訊工程領域所開發的「線上分析處理」與「資料探勘」技術應用於醫院門診糖尿病相關資料庫，提供疾病管理快速有用的訊息。
方法：本研究資料來源為中山醫學大學附設醫院 1998 年至 2003 年所有門診相關資料庫。首先連結醫院例行儲存的不同資料庫來建立糖尿病患資料庫；接下來必須對原始資料進行清理的動作；第三是設計一個以主題為導向的資料倉儲，這時也必須將部份資料進行整合、轉換與運算；第四是建立線上分析處理模型；第五是選擇一個主題進行不同資料探勘方法的檢測，找尋最佳分類模型。
結論：由於疾病管理、線上分析處理與資料探勘在健康照護領域都還不是很成熟，本次研究也大多是不斷摸索嘗試與縮小範圍才比較能清楚掌握關鍵，但是本研究的經驗覺得這是相當有開發潛力的領域。

關鍵詞：疾病管理、糖尿病、線上分析處理、資料探勘、健康資訊

ABSTRACT

Background: Most of the information related to out-patients and in-patients are digitalized in most hospital in Taiwan. Nevertheless, these routinely collected huge data had not been fully used to provide useful information for disease management purpose.
Objective: To explore the feasibility of using OLAP (On-Line Analytical Processing) and Data Mining techniques developed by information engineering field on hospital diabetes-related disease dataset to provide useful information for disease management.
Method: All out-patients related data during the year 1998 through 2003 were collected for this study. We first linked different routine recorded data sets to establish a diabetic database as the basic workstation for this study. Second, the data cleaning procedure was processed. Third, we established a object-oriented data warehousing which involved several data integration, transformation, computing and summarization. Forth, we established a OLAP model. Fifth, we compared different data mining techniques to find the best classification model.
Conclusions: As disease management, OLAP and data mining were all very newly developed concept and techniques in healthcare fields. The research team spent a lot of time in try and errors and narrow down the questions and then caught some points. We highly believed that there will be a lot of potential in this combination.

Keywords: disease management, diabetes mellitus, OLAP, data mining, health informatics

二、緣由與目的

目前台灣大多數醫院的門診與住院相關的資料大多已經數位化，每天例行累積產生千萬筆交易資料。可惜這些例行累積的龐大訊息並沒有好好被利用來提供疾病

管理有用的訊息。傳統統計學資料分析是由上而下，也就是先有理論產生假說，再收集資料驗證假說。資料探勘則是由下而上，在龐大的資料庫中透過特殊的方法，找出有意義的模式或規則，發掘許多我們原本所不知的事實。

資料探勘的技術主要來自三個領域：一是傳統的統計學，二是資訊科學的人工智慧，三是決策支援系統[1]。資料探勘之所以受到重視是因為日常生活各種交易的全面電腦化，每筆交易都會紀錄在電腦資料庫中，如果能善加利用與分析這些資料庫，一定能產生許多有助於決策的相關訊息。資料探勘技術在商業界已經累積出相當的成果，醫療照護領域的資料也普遍電腦化，因此有必要好好應用此技術在疾病資料庫中挖掘出有用的訊息提供醫療照護決策參考。由於糖尿病是一個相當複雜且併發許多其他疾病的慢性病，花費的醫療費用也相當高，目前是衛生署及健保局非常重視的及病防治重點，因此特別需要有用的相關訊息協助照護過程的決策。

國外已經有許多以資料探勘技術分析醫療照護相關問題的例子，應用於大型資料建立的例子有世界衛生組織的國際藥物安全警訊計畫，目前正透過資料探勘技術來偵測藥物的可能副作用，這必須結合不同資料庫（就診資料庫、疾病資料庫、藥物使用資料庫）才可能達到此目的。過去以人工方式來達到此目的，除了非常耗費人力與時間外，也常常有錯誤與漏失[2,3]。

美國阿拉巴馬大學伯明罕醫學院的病理科也嘗試將不同檢驗資料庫連結[4]，同時也以此來監視抗藥性菌種之出現[5,6]。美國馬利蘭州也利用保險給付資料建立一個癌症監視系統，除了可以了解癌症疾病組合，也可評估治療花費與成本效益，同時也可以發現少數弱勢族群的特殊問題[7]。美國杜克大學（Duke University）利用臨床資料庫來作為早產兒發生的預測，進一步作為預防早產兒的發生[8]。

傳統資料庫的建立必須花很多金錢與人力定期收集，結果還是有很多缺失，資料探勘可以連結許多例行建立的資料庫形成資料倉儲，經過前置處理與整合，可以提高效益相輔相成。

傳統有關疾病預後的研究大多用迴歸分析，現在利用資料探勘技術來分析相同資料，常常能得到更準確的預測。譬如預測人工髖關節手術後的長期臨床功能[9]、預測脊椎損傷患者最後能否行走[10]、預測手術時間長短[11]、預測交通事故傷害嚴重度[12]、預測外傷手術後之存活率[13]、預測醫療使用率與費用[14]等。

由於台灣各層級醫院的不同部門的運作大多已經電腦化，其普遍的程度依序為門診掛號，批價系統，疾病診斷分類系統，健保申報系統，藥品及耗材管理系統，檢驗系統，門診診療系統，住院診療系統，影像系統等。每一部門會依據該單位的需求設計一套電腦系統，不同系統間常常也不一定能整合。更可惜的是這些系統的資料，在該部門統計完簡單的日週月報表後，就冷藏在大型電腦的磁帶內，最後終遭丟棄的命運。

更諷刺的是醫療人員為了研究疾病照護相關問題，還要另外申請研究計畫花錢請人將病歷資料摘要有用訊息，或設計問卷收集資料，這都是非常浪費不持久也不完整的做法。而且傳統資料分析做法有必須先有假說才去收集資料，許多潛藏的問題或特殊關聯現象也常常因為人類經驗的限制而忽略掉。

由於中山醫學大學與逢甲大學有建立策略聯盟，除了支援彼此師資教學資源外，也提供跨科技整合良好基礎。逢甲學資訊工程學系在資料探勘理論與技術的發展有相當的經驗，本研究將以資料探勘技術應用於中山醫學大學附設醫院的不同電腦資料庫，探討建立疾病相關資料庫的可行性，同時嘗試建立一些例行分析機制，希望能定期提供對疾病照護及醫務管理有用訊息。最後希望能將這些經驗推展到不同醫院，對國人疾病照護品質的提昇有所助益。由於本研究是一初步嘗試，所以先以糖尿病為例。因為糖尿病是一個相當複雜且併發許多其他疾病的慢性病，花費的醫療費用也相當高，目前也是衛生署及健保局非常重視的及病防治重點，因此特別需要有用的相關訊息協助照護過程的決

策。

三、材料與方法

本研究資料來源為中山醫學大學附設醫院 1998 年至 2003 年所有門診相關資料庫。首先連結醫院例行儲存的不同資料庫（譬如病患基本資料、門診、住院、藥物、檢驗、病理、耗材、健保申報、死亡等資料庫）來建立糖尿病患資料庫，作為計劃所需要的基本平台。

第二步必須對原始資料進行清理的動作，譬如刪除沒有出生年月日與性別等基本資料或是有矛盾的資料。第三是設計一個以主題為導向的資料倉儲，本研究區分了下列幾個倉儲：

1. 糖尿病患者就醫行為相關倉儲
2. 糖尿病患者檢驗結果相關倉儲
3. 糖尿病患者罹患其他疾病相關倉儲
4. 糖尿病患者治療方式相關倉儲
5. 糖尿病患者治療結果相關倉儲
6. 醫師處方檢驗行為相關倉儲
7. 醫師處方藥物行為相關倉儲
8. 醫師別與科別糖尿病患治療結果倉儲

為了提供有用訊息，必須將部份資料進行整合、轉換與運算，本次報告只以糖尿病患者就醫行為與空腹血糖檢驗結果為例說明。原本資料庫形式是以就診記錄唯一筆交易記錄來儲存，同一人一年可能有十幾筆就診記錄。如果我們要分析一位患者的就醫行為型態，就必須綜合一年所有就診記錄加以整合、轉換、運算與摘要。我們產生了下列幾種新的變項：

V1 就診期間：在研究期間出現第後一次就就診日期減去第一次就診日期；

V2 就診次數：在研究期間總共有幾次就診記錄；

V3 平均就診間隔：V2 除以 V1；

V4 平均每月就診次數：V1 除以 V2；

V5 每兩次就診間隔的變異係數，此數據是要反映規則性；

V6 平均空腹血糖值；

V7 平均糖化血色素值；

V8 是否有住院

第四是建立線上分析處理模型，儘量以圖形方式呈現疾病管理者想要了解的有用訊息：譬如高間隔患者的人口學特徵為何？不規則就診患者的平均空腹血糖值比規則就診患者高多少？哪些醫師的患者一年來都沒有處方糖化血色素檢查？

第五是選擇一個主題進行不同資料探勘方法的檢測，找尋最佳分類模型。本報告以前述例，找出依據人口學資料與就醫行為如何分類患者是屬於高空腹血糖值。

四、結果

圖一為中山醫學大學附設醫院門診相關資料庫的關聯圖。本研究花了相當多的時間來了解每一資料庫的變項定義與資料品質。表一為糖尿病患求醫行為的基本數據。因為篇幅關係，本次精簡報告沒有將線上分析處理的畫面展示。附錄是使用不同資料探勘技術的比較，已經寫成論文形式投稿。

五、討論

由於疾病管理、線上分析處理與資料探勘在健康照護領域都還不是很成熟，本次研究也大多是不斷摸索嘗試與縮小範圍才比較能清楚掌握關鍵，但是本研究的經驗覺得這是相當有開發潛力的領域。

不可否認門診診斷的準確度不高，但是本研究侷限就診期間一年且就診次數十二次以上，應該都是確定糖尿病診斷患者。至於空腹血糖值採取平均，有可能因為不同患者次數不同而造成統計不穩定。另外有些患者是初期控制，所以平均血糖偏高，應該是穩定患者進行比較比較合理。

本研究針對就醫行為進行分析有一大缺點就是患者可能去其他醫療院所就診，對於結果的評估可能有干擾。但是就疾病管理的立場而言，患者沒有在本醫院規則就醫就是管理不良的指標。當然了，如果還要更精確對治療結果進行預測，來要陸續加入是否罹患其他疾病以及其他危險因子之測量才合理。

六、計畫成果自評

本研究當除的確寫得太大了，以致延誤相當久才有初步成果產生，離當初的預期還相差非常遠。不過本研究團隊還會繼續努力，在這領域開花結果。

七、參考文獻

[1] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: for Marketing, Sales, and Customer Support*. New York: Joun Wiley & Sons, 1997.

[2] M. Lindquist, M. Stahl, A. Bate, I. R. Edwards, and R. H. B Meyboom, “A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database,” Drug Safety, vol. 23, pp. 533-542, 2000.

[3] D. M. Coulter, A. Bate, R. H. B Meyboom, M. Lindquist, and I. R. Edwards, “Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study,” Br Med J, vol. 322, pp. 1207-1209, 2001

[4] J. M. Mcdonal, S. Brossette, and A. Moser, “Pathology information systems: data mining leads to knowledge discovery,” Arch Pathol Lab Med, vol. 122, pp. 409-411, 1998.

[5] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, “Association rules and data mining in hospital infection control and public health surveillance,” J Am Med Inform Asso, vol. 5, pp. 373-381, 1998.

[6] S. A. Moser, W. T. Jones, and S. E. Brossette, “Application of data mining to intensive care unit microbiologic data,” Emerging Inf Dis, vol. 5, pp. 454-457, 1999.

[7] G. A. Forgionne, A. Gangopadhyay, and M. Adya, “Cancer surveillance using data warehousing, data mining, and decision support systems,” Top Health Inform Mana, vol. 21, pp. 21-34, 2000.

[8] L. Goodwin and S. Maher, “Data mining for preterm birth prediction,” SAC 2000 (ACM Symposium on Applied Computing), pp. 46-51, 2000.

[9] B. Zupan, J. Demsar, D. Smrke, K. Bozikov, V. Stankovski, I. Bratko, and J. R. Beck, “Predicting patient’s long-term clinical status after hip arthroplasty using hierarchical decision modeling and data mining,” Method Inform Med, vol. 40, pp. 25-31, 2001.

[10] A. Ohrn and T. Rowland, “Rough sets: a knowledge discovery technique for multifactorial medical outcomes,” Am J Phys Med Rehabil, vol. 79, pp. 100-108, 2000.

[11] D. P. Strum, A. R. Sampson, J. G. May, and L. G. Vargas, “Surgeon and type of anesthesia predict variability in surgical procedure times,” Anesthesiology, vol. 92, pp. 1454-1466, 2000.

[12] S. Y. Sohn and H. Shin, “Pattern recognition for road traffic accident severity in Korea,” Ergonomics, vol. 44, pp. 107-117, 2001.

[13] N. Aoki, M. J. Wall, J. Demsar, et al., “Predictive model for survival at the conclusion of a damage control laparotomy,” Am J Surg, vol. 180, pp. 540-545, 2000.

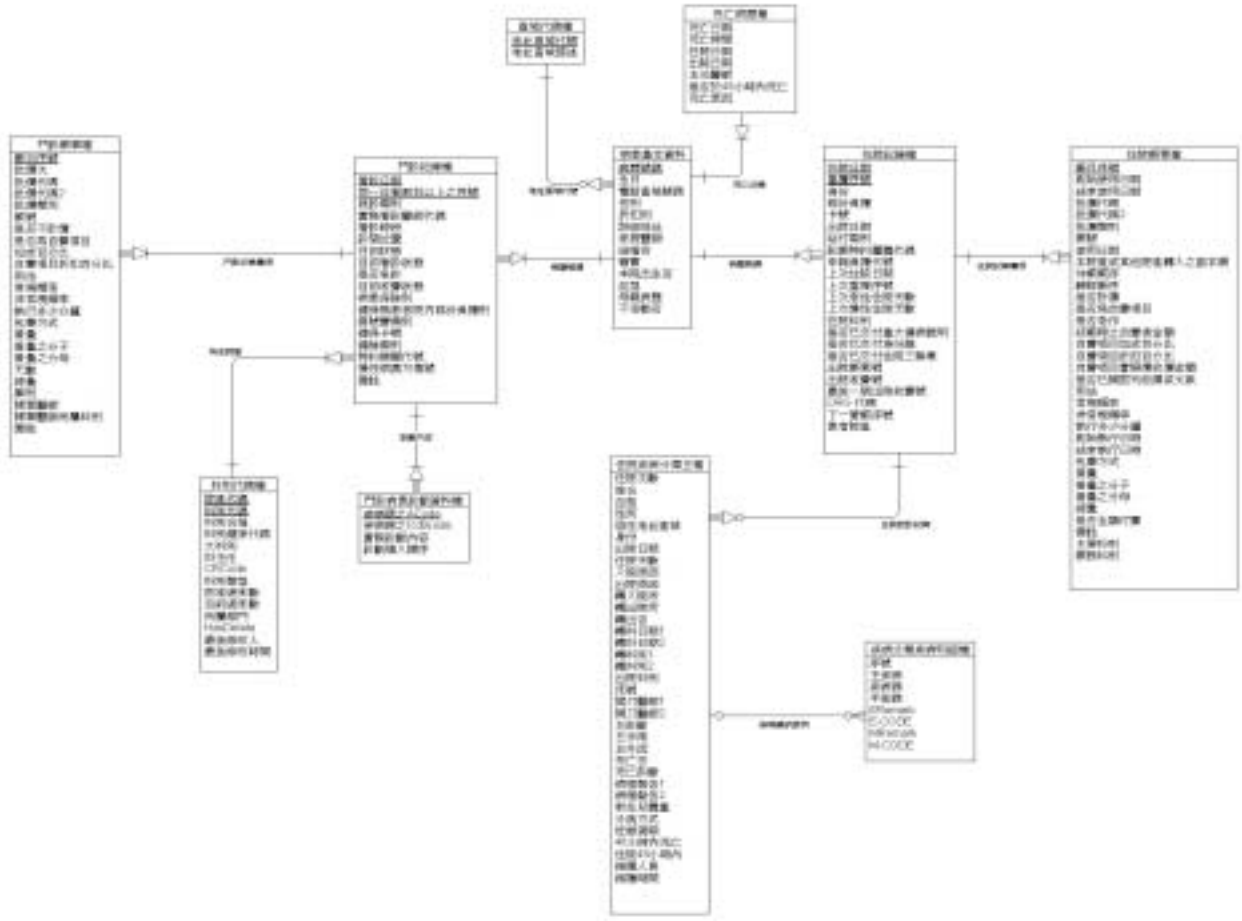
[14] F. R. Elevitch, A. Silvers, and J. D. Sahl, Projecting corporate health plan utilization and charges from annual ICD-9-CM diagnostic rates: a value-added opportunity

for pathologists," Arch Pathol Lab Med, vol. 121, pp. 1187-1191, 1997.

表一：經過資料前處理及基本運算後之資料筆數

	人數	就診次數	有檢驗空腹血糖筆數	有檢驗糖化血色素筆數	有住院筆數
1998-2003 年門診總計	86,542	1,715,510	44,364	5,231	141,646
有糖尿病診斷者	3,635	41,396	14,634	3,739	4,129
就診期間一年或以上者	1,287	33,563	5,790	1,614	763
就診次數 12 次或以上者	947	31,488	5,037	1,390	582

圖一：中山醫學大學附設醫院門診相關資料庫關聯圖



Design of Accurate and Compact Fuzzy Classifiers with Linguistic Rules Using Intelligent Genetic Algorithms

Shinn-Ying Ho (何信瑩)*[a], Hung-Ming Chen (陳宏銘) [a], Chih-Jen Kao (高志仁) [a], and Tsung-Hsueh Lu (呂宗學)[b]

[a]Department of Information Engineering and Computer Science
Feng Chia University, Taichung, Taiwan 407, R.O.C.
[b]Department of Public Health
Chung Shan Medical University, Taichung, Taiwan 402, R.O.C.

*TEL: 886-4-24517250 ext. 3753, FAX: 886-4-24516101, e-mail: syho@fcu.edu.tw

Abstract

For a fuzzy rule-based classification system, it is critical to establishing fuzzy rules and to determining fuzzy partition area in order to accurately classify the training pattern. In this paper, an evolutionary design method for fuzzy rule-based classifier is proposed. Flexible trapezoid membership functions are used to generate accurate grid partition of feature space. The membership functions and rule base of the fuzzy classifier are simultaneously optimized by the intelligent genetic algorithm (IGA). We also incorporated heuristics to enhance the search capability of the IGA. Experimental results revel that the proposed IGA-based method is efficient in terms of the classification accuracy and the compactness of fuzzy classifiers.

Keywords: Fuzzy classifiers, membership functions, intelligent genetic algorithms**.**

1. Introduction

The concept of "*fuzzy sets*" was proposed to deal with vagueness, uncertainty, and imprecision intrinsic to many problems. The main characteristic feature of fuzzy sets is to apply the fuzzy concept to precise sets, then to describe and simplify the input space of a given problem. In general, daily-life practical problems involve a large number of uncertainty and complexity, therefore fuzzy theory has been widely used.

The classification problems have been playing an important and crucial role in many daily-life problems and engineering area. The approach to establish fuzzy rules to describe the complex and uncertain system is perfectly adapted to the classification problems.

The design process for a fuzzy classifier is first to generate the partition of input space, then to develop the fuzzy rule for each partition region. In general there are three types of partition methods, grid, scatter, and tree partitions. First, the grid partition defines each region as a square area and it is the most widely used method, especially in control system. Second, the scatter partition determines each region by covering a subset of the whole input space. Third, the tree partition specifies each region based on a corresponding decision tree. The choice of

partition region could affect the performance of classification. If a partition is too coarse, the performance may be low. If a partition is too fine, it will generate too many rules and cause cumbersome operations.

In order to full utilize the ability of fuzzy classifiers, further optimization for the membership functions or the rule base is required. Various literatures have successfully applied genetic algorithms (GAs) [1] to fuzzy classifier optimization. Ishibuchi *et al.* proposed a series of evolutionary approaches for fuzzy rule selection with fixed membership functions [4-6]. Murata *et al.* used GAs to adjust triangular membership functions [9]. However, many of them mainly focused on the simple problems with low input dimensions, which would not be feasible to the classification problems.

To design an efficient fuzzy classifier, the design method have to satisfied the following requirements:

1) The membership functions should be flexible enough to generate accurate fuzzy partition.
2) The optimization algorithm should have the ability for solving large parameters optimization problems in fuzzy classifier design.

In this paper, we proposed an evolutionary approach for design accurate and compact fuzzy classifiers with grid partition of feature space:

1) Adopt the trapezoidal membership function to ensure the flexibility of the fuzzy classification system;
2) Simultaneously adjust the membership functions and the fuzzy rule base;
3) Use the intelligent genetic algorithm (IGA) [13] to solve high-dimensional classification problems; and
4) Incorporate heuristics to enhance the search capability of IGA.

The main advantage of the proposed method is that the accuracy of fuzzy classifier is maximized while the great interpretation ability of grid partition is also preserved. The high classification rate with fewer fuzzy rules can be obtained through the flexibility of trapezoidal membership and the search ability of the intelligent genetic algorithm.

The experimental results show that the proposed method is efficient in designing high-performance fuzzy classifiers in terms of classification accuracy and the number of fuzzy rules, compared with existed grid partition approaches.

2. Related Work

2.1 Genetic Algorithms

The theoretical foundation of genetic algorithm was originally proposed by John Holland [2]. The concept came from the evolution process that operates on chromosomes. The natural selection process reveals that the chromosomes that encode successfully structures reproduce more often than those that do not. As of today, genetic algorithms have been successfully applied to a variety of research areas, for instance optimizations, machine learning, control systems, and pattern recognition.

One of the successful applications that genetic algorithms have been employed is the fuzzy if-then rule for classification problems. The real-life classification problems often involve many input attributes.

2.2 Genetic Rule Selection

For instance, if there is a ten-dimensional pattern classification

problem with six fuzzy sets for each attribute, the total number of possible fuzzy if-then rules is 6^{10} (more than 2 billion). Thus how to select a small number of rules in order to construct a compact fuzzy rule-based system is crucial and important. Ishibuchi *et al.* [2][4] proposed the idea of rule selection using genetic algorithms that can generate smaller numbers of rule sets.

K fuzzy sets are defined for each feature axis with fixed shape of triangular membership functions. They also proposed the "*don't care*" concept if some features are not used in the classifier. Thus the *n*-dimensional feature space is then partitioned into regions and number of possible fuzzy rules is $(K+1)^n$. Since various regions do not contain any patterns, the corresponding fuzzy rules can be eliminated without losing accuracy.

The remained candidate fuzzy rules are further optimized using a simple genetic algorithm. The solution is encoded as a bit string. The bit '1' represents the corresponding rule is selected into the rule set. Otherwise, the rule is not used.

However, the rule selection approach is not suitable for solving high-dimensional pattern classification problems. The number of rules exponentially increases with the number of input features. Even after the rule elimination, the simple genetic algorithm will suffer to the huge number of candidate fuzzy rules.

2.3 Fuzzy Genetic-Based Machine Learning

For solving high-dimensional pattern classification problems, Ishbuchi *et al*. proposed a fuzzy genetic-based machine-learning (GBML) algorithm [17]. They use the same membership functions as the early rule selection method, but the antecedent conditions of the fuzzy rules are encoded into chromosomes and directly evolved by the genetic algorithm instead. The variable-length representation of chromosomes allows genetic algorithm to minimize the number of fuzzy rules. Compared with the rule selection approach, the fuzzy GBML has the ability for solving high-dimensional classification problems.

However, the shapes of the triangular membership functions of the fuzzy GBML proposed by Ishibuchi *et al*. are fixed. In real-world applications, the membership functions should be flexible enough to generate accurate fuzzy partition. Homaifar and McCormick [12] showed that simultaneous design of membership functions and fuzzy rules can enhance the performance of fuzzy systems. It also increases the number of parameters which will be optimized. Thus the final classification performance (i.e., accuracy and rule base complexity) of the fuzzy classifier will depend on the search ability of the genetic algorithm.

3. Evolutionary Fuzzy Classifier Design

3.1 Membership Functions and Fuzzy Partition

In this paper, we used flexible trapezoid membership functions (Fig. 1) in the fuzzy classifier design. Each membership function is represented by five parameters. Without losing the generalization ability, we assume that all the attribute value is normalized in the unit interval [0, 1]. The membership function $\mu(x)$ of a fuzzy set is defined as follows:

$$\begin{cases} 0 & \text{if } x < a \text{ or } x > d \\ 1 & \text{if } b \le x \le d \\ \dfrac{x-a}{b-a} & \text{if } a < x < \text{b} \\ \dfrac{d-x}{d-c} & \text{if } c < x < d, \end{cases} \quad (1)$$

where $x \in [0,1]$ and $a \le b \le c \le d$. The variables a, b, c, and d determining the shape of a trapezoidal fuzzy set are to be optimized. The parameters of the membership functions are optimized by the intelligent genetic algorithm later. The parameter L can decrease the interaction of encoded parameters, which can further extend the optimization performance of intelligent genetic algorithm.

For each feature axis, a specified number of membership functions are defined. Each membership function represents a fuzzy set with a linguistic label. Fig. 2 is an example of a feature axis with three trapezoidal membership functions.

3.2 Fuzzy Rules and Fuzzy Reasoning Method

The following fuzzy if-then rules for n-dimensional pattern classification problems are used in our design of fuzzy classifier systems:

R_j: If x_1 is $A_{1,j}$ and … and x_n is $A_{n,j}$ then Class C_j with CF_j, $j = 1, \ldots, N$,

where R_j is a rule label, x_i denotes a feature variable, $A_{n,j}$ is an antecedent fuzzy set, $C_j \in \{1, \ldots, C\}$ denotes a consequent class, C is a number of classes, CF_j is a certainty grade of this rule in the unit interval [0,1], and N is a number of fuzzy rules in the initial fuzzy-rule base. The antecedent conditions of each fuzzy rule are generated by the intelligent genetic algorithm. Once the membership functions and the antecedent conditions are given, the variable C_j and CF_j are determined by the following equations in the training phase [18].

$$C_j = \arg\max_k \beta_{\text{Class}\,k}(R_j), \quad (2)$$

$$\beta_{Class_k}(R_j) = \sum_{x_p \in Class_k} \mu(x_p), \quad (3)$$

$$\mu(x_p) = \mu_{j1}(x_{p1}) \times \ldots \times \mu_{jn}(x_{pn}), \quad (4)$$

where $x_p = \{x_{p1}, \ldots x_{pn}\}$ is the training pattern and $\mu_{ji}(.)$ is the membership function of fuzzy set A_{ji}, and

$$CF_j = \{\beta_{ClassCj}(R_j) - \overline{\beta}\} / \sum_k \beta_{Classk}(R_j) \quad (5)$$

where

$$\overline{\beta} = \sum_{k \ne C_j} \beta_{Classk}(R_j) / (C-1). \quad (6)$$

In the test phase, the class label of a test pattern is determined by the Single-Winner-Rule strategy [16]. Given a test pattern x_p, the value of $\mu_j(x_p) \cdot CF_j$ of each fuzzy rule is calculate, and the winner rule R_{j^*} is the rule with maximal value $\mu_{j^*}(x_p) \cdot CF_{j^*}$ and the class of the test pattern is determined by the consequent class of R_{j^*}.

3.3 Fitness Function and Chromosome Representation

Our three goals for design of fuzzy classifiers are: 1) maximize the classification accuracy, 2) minimize the numbers of fuzzy rules, and 3) minimize the total number of antecedent conditions of the fuzzy rules. We formulate these criteria as the fitness function $F(S)$:

$$\text{Maximize } F(S) = NCP(S) - w_{\text{r}} \cdot N_{\text{r}} - w_{\text{a}} \cdot N_{\text{a}}, \quad (7)$$

where S is the fuzzy classifier generated by IGA, $NCP(S)$ is the number of correctly classified training patterns by S, N_{r} is the number of fuzzy rules, N_{a} is the total number of antecedent conditions. w_{r} and w_{a} are the weighted values to control the importance of N_{r} and N_{a}, respectively.

The parameters of membership functions (parametric genes) and the antecedent part of fuzzy rules (rule base genes) are encoded into chromosomes. Each parametric gene is a binary coded real value in the interval [0,1] and each rule base gene can be one of the conditions 0, 1, …, *K*-1, or '#', where *K* is the number of fuzzy sets per feature, and '#' is the "*don't care*" condition.

In order to minimize the number of used fuzzy rules, the control genes are also encoded in the chromosomes. Each fuzzy rule is controlled by a gene with one bit length. If the bit is 1, the corresponding fuzzy rule will selected into the rule base; otherwise, the fuzzy rule will not be used. Fig. 3 shows the chromosome representation for $K = 3$. It is noticed that each membership function located at the boundary only required two parameters.

In our experiments, we set the value of *k* to be 3. The total number of encoded parameters is $(9n + m + m{\cdot}n)$ where *n* is the number of features and *m* is the maximal number of rules. In our experiments, the maximal number of fuzzy rules is $3C$ where *C* is the number of classes of the classification problem.

3.4 Intelligent Genetic Algorithm

In this section, we proposed a modified version of intelligent genetic algorithm for design of fuzzy classifiers.

Population initialization. We applied the well-known fuzzy *c*-means (FCM) clustering algorithm to generate the initial membership functions. For each feature, we calculate the projection of training patterns and the means obtained by FCM are the initial values of parameters L_i. The other parametric genes and the control genes are randomly initialized. Finally, the following direct rule generation heuristic [10] is used to initialize the rule base genes.

Step 1: Randomly select a training pattern.

Step 2: Calculate the membership values of fuzzy sets for each feature.

Step 3: The antecedent conditions of the fuzzy rule are the fuzzy sets with maximal membership values for each feature.

Step 4: Randomly select half of the antecedent conditions and set them to be '#' ("*don't care*").

Step 5: Perform step 1 to 4 until all fuzzy rules are generated.

Selection. The binary tournament selection without replacement is used in the proposed algorithm. The selection strategy ensures that the best individual of each generation can be selected into the mating poor.

Intelligent crossover operator. A modification of intelligent crossover using orthogonal experimental design (OED) is proposed. We use OED to generate high-quality offspring in the crossover operator. In order to minimize the number of parameters involving in the orthogonal experiments, any corresponding pair of genes with the same value from two chromosomes can be simply ignored in the crossover operator. Furthermore, if any corresponding pair of control genes is '0', the corresponding rule base genes can be also ignored. The remained genes are randomly divided into several segments and the orthogonal experimental design is applied.

Mutation operators. Two types of mutation operators are used in the proposed method. For parametric and control genes, the simple bit inverse mutation operator is used. For rule base genes, the direct rule generation heuristic is also used here. Let m_{err} is the number misclassified training patterns and N_{r} is the number of decoded fuzzy rules. If $m_{\text{err}} < \lfloor N_{\text{r}}/2 \rfloor$, then m_{err} new fuzzy rules are generated from the misclassified training

patterns; otherwise, $\lfloor N_r/2 \rfloor$ new fuzzy rules are generated.

The framework of the proposed algorithm is described as follows:

Step 1: Initialize population $P(1)$, and let $t = 1$.

Step 2: Evaluate $P(t)$.

Step 3: Select individuals from $P(t)$ into mating poor.

Step 4: Perform intelligent crossover and mutation and the newly generated individuals form $C(t)$.

Step 5: Select $P(t+1)$ from $P(t)$ and $C(t)$.

Step 6: Let $t = t + 1$.

Step 7: If the termination condition is satisfied, stop the evolution; otherwise, go to step 2.

The termination condition is defined as user's preference, such as maximal number of generations or maximal number of function evaluations. The detail information and theoretical analysis of the intelligent genetic algorithm can be found in [13].

4. Experimental Results

In this section, the proposed IGA-based method is compared with the existed fuzzy classifiers with grid partition and the decision tree C4.5 release 8 [15]. Several databases for performance evaluation are selected from UCI Machine Learning Repository [14]. Table 1 is the summarized information of the databases.

The parameter settings of the proposed method are as follows:

Number of fuzzy sets per feature: 3;
Maximal number of fuzzy rules: 3*C*;
Population size N_{pop}: 20;
Number of fuzzy sets for each feature: 3;
Weighted values of the fitness function: $w_r = 0.1$, $w_a = 0.001$;
Crossover probability P_c: 0.8;
Mutation probability P_m: 0.01;
Stopping condition: 20000 fitness evaluations.

4.1 Experiment 1

In the following experiments, all the patterns are used as the training patterns for fuzzy classifier design. The results of the proposed IGA-based method are obtained from 20 independent runs. Tables 2 and 3 are the best performance of the IGA-based method and other fuzzy classifiers for iris and wine classification problems, respectively. Table 4 is the average classification performance of IGA-based method. Figs. 4 is the membership functions and rule base for the iris classification problem obtained by the proposed method.

In the experiments, the IGA-based method can obtained very high classification rates and the used number of fuzzy rules is fewer than the existed fuzzy classifiers.

4.2 Experiment 2

In order to show the generalization ability of the proposed method, we applied the ten-fold cross validation method (10-CV) to the databases. For *m*-fold cross validation, the database is randomly divided into *m* subsets with equal size. Then the classifier is trained *m* times, each time with a different set held out as a validation (or test) set. The estimated performance is the mean of these *m* results.

Table 5 is the average performance of the IGA-based method obtained from 30 independent runs. Table 6 is the performance of C4.5 release 8 with pruned tree using certainty level $CF = 0.25$ (the default value of the program).

We observed that the average number of fuzzy rules for each class is less than 2. It is show that the proposed method is effective to minimize the number of fuzzy rules. The results shown that the performance of the proposed IGA-based is comparable to or better than that of C4.5. In half of the databases, the classification accuracy of test data of the IGA-based method is better than that of C4.5, while the number of used rules of the IGA-based method is also less than that of C4.5. That is, the proposed method can obtain more compact classifiers with high classification accuracy. The fuzzy rules with linguistic expression are more comprehensible for human.

5. Concluding Remarks

In this paper, we proposed an evolutionary approach for design of accurate and compact fuzzy classifiers. The flexible trapezoid membership functions can archive the accurate grid partition of feature space. The proposed evolutionary design method simultaneously optimizes the shapes of membership functions and the fuzzy rule base, which improves the classification performance. Finally, several useful heuristics are incorporated into the proposed method, which can further extend the optimization performance of the intelligent genetic algorithm.

The performance of the proposed method is evaluated using various databases which are frequently used in the research area of pattern recognition. It is shown empirically that the performance of the proposed method is superior to the existed rule-based methods in terms of classification accuracy and the compactness of classifiers.

References

[1] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, 1989.

[2] J.H. Holland, *Adaption in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan, 1975.

[3] H. Ishibuchi, K. Nozaki, N. Yamamtot, and H. Tankara, “Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms,” *Fuzzy Sets and Systems*, vol. 65, no. 2-3, pp.237-253 1994.

[4] H. Ishibuchi and T. Murata, “A genetic-algorithm-based fuzzy partition method for pattern classification problems,” in *Proc. Genetic Algorithm and Soft Computing* (*Studies in Fuzziness, vol. 1.8*), pp.555-578, 1996.

[5] H. Ishibuchi, K. Nozaki, N. Yamamtot, and H. Tankara, “Selecting fuzzy if-then rules for classification problems using genetic algorithms,” *IEEE Trans. Fuzzy Systems*, vol. 3, no.3, pp.260-270, 1995.

[6] H. Ishibuchi, T. Murata, and I.B. Turksen, “Single-objective and two-objective genetic algorithm for selecting linguistic rules for pattern classification problems,” *Fuzzy Sets and Systems*, vol. 89, no.2, pp.135-150, 1997.

[7] H. Ishibuchi and T. Murata, “Minimizing the fuzzy rule base and maximizing its performance by a multi-objective genetic algorithm,” in *Proc. FUZZ-IEEE’97*, pp.259-264, 1997.

[8] H. Ishibuchi, and T. Murata, “Multi-objective genetic local search for minimizing the number of fuzzy rules for pattern classification problems,” *IEEE Trans. Syst., Man, Cybern. B*, pp.1100-1105, 1998.

[9] T. Nurata, H. Ishibuchi, and M. Gen, “Adjusting fuzzy partitions by genetic algorithms and histograms for pattern classification problems,” in *Proc. IEEE Conf. Computational Intelligence*, pp. 9-14, 1998.

[10] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Information Sciences*, vol. 136, pp. 109-133, 2001.

[11] K. Nozaki, H. Ishibuchi, and H. Tanaka, "Adaptive fuzzy rule-based classification system," *IEEE Trans. Fuzzy System*, vol. 4, no. 3, pp. 238-250, 1996.

[12] A. Homaifar and E. McCormick, "Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 2, pp. 129-139, 1995.

[13] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases, 1998. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[14] S.-Y. Ho *et al.*, "Intelligent genetic algorithm with a new intelligent crossover using orthogonal arrays," in *Proc. Genetic and Evolutionary Computation Conference*, 1999, pp. 289-296.

[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, Morgan Kauffman, Loas Altos, CA, 1993.

[16] H. Ishbuchi and T. Nakashima, "Effect of rule weights in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Systems*, vol. 9, no. 4, pp. 506-515, 2001.

[17] H. Ishibuchi, T. Nakashima, and T. Murata, "Performance evaluation of fuzzy classifier systems for multi-dimensional pattern classification problems," *IEEE Trans. Syst., Man, Cybern. B*, vol. 29, pp. 601–618, Oct. 1999.

[18] H. Ishibuchi, K. Nozaki, and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," *Fuzzy Sets Syst.*, vol. 52, no. 1, pp. 21–32, 1992.

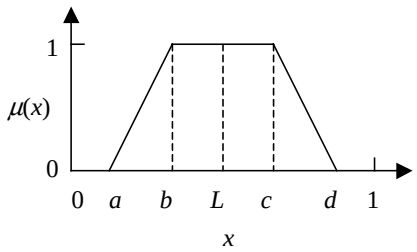


Fig. 1. Flexible trapezoid membership function.

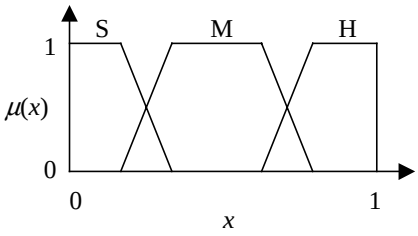


Fig. 2. An example of a feature axis with three membership functions.

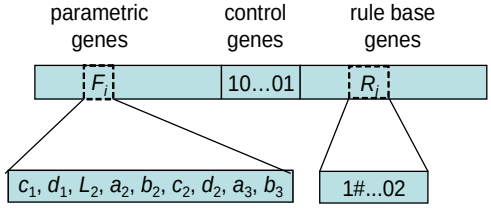
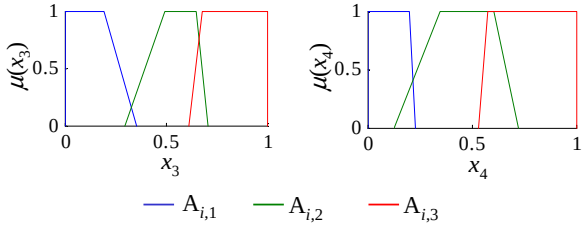


Fig. 3. Chromosome representation.



R_1: Class 0 with $CF = 0$

R_2: If x_3 is A_{32} and x_4 is A_{42} then Class 2 with $CF = 0.94$

R_3: If x_4 is A_{43} then Class 3 with $CF = 0.62$

Fig. 4. The membership functions and rule base for iris classification using all patterns as training data.

Table 1. Databases with numerical attribute values. N_p is the number of encoding parameters in the chromosomes of IGA.

Database	Pattern number	Dimension n	Class number C	N_p
cmc	1473	9	3	171
glass	214	9	6	261
haberman	306	3	2	51
heart-c[†]	297	13	5	327
iris	150	4	3	81
liver-disorder	345	6	2	96
new-thyroid	215	5	3	99
pima-diabetes	768	8	2	126
wdbc	569	30	2	456
wine	178	13	3	243

Table 2. Best performance of fuzzy classifiers for iris classification problem.

	IGA-based	Rule Selection [10]	Ishibuchi (1996) [11]	Ishibuchi (1995) [5]	Ishibuchi (1996) [4]	Ishibuchi (1998) [9]
Classification rate	98%	98%	100%	100%	100%	100%
Number of rules	3	5	11	13	56	58
Number of antecedent conditions	3	7	44	52	NA	NA

Table 3. Best performance of fuzzy classifiers for wine classification problem.

	IGA-based	Rule Selection [10]	Fuzzy GBML [10]	Ishibuchi (1995) [5]	Ishibuchi (1996) [4]
Classification rate	99.4%	96.1%	99.4%	100%	100%
Number of rules	5	4	8	16	180
Number of antecedent conditions	10	6	15	208	NA

Table 4. Average performance of IGA-based method

	iris	wine
Classification rate	97.20%	97.83%
Number of rules	3.7	6.1
Number of antecedent conditions	4.2	11.9

Table 5. The average performance of IGA according to the fitness value using 10-CV.

Database	Fitness	*TrCR* (%)	*TeCR* (%)	N_r	N_a	N_r/C	Avg. Length
cmc	726.553	54.85	52.54	5.2	11.8	1.7	2.3
glass	129.408	67.92	59.52	13.6	46.5	2.3	3.4
haberman	203.509	73.98	73.21	2.3	2.2	1.1	1.0
heart-c	166.279	62.56	54.92	9.2	31.2	1.8	3.4
iris	131.307	97.54	95.29	3.6	4.0	1.2	1.1
liver-disorder	212.002	68.39	63.56	3.3	5.4	1.7	1.6
new-thyroid	175.306	90.98	88.89	7.3	14.7	2.4	2.0
pima-diabetes	510.269	73.87	72.04	3.1	4.7	1.6	1.5
wdbc	469.815	91.83	89.13	4.1	37.3	2.1	9.1
wine	154.867	97.11	88.83	6.8	21.9	2.3	3.2
Average	287.931	77.90	73.79	5.85	17.98	1.81	2.86

Table 6. The performance of C4.5 with pruned tree using 10-CV.

Database	*TrCR* (%)	*TeCR* (%)	N_r	N_r/C
cmc	54.85	52.54	5.2	1.7
glass	67.92	59.52	13.6	2.3
haberman	73.98	73.21	2.3	1.1
heart-c	62.56	54.92	9.2	1.8
iris	97.54	95.29	3.6	1.2
liver-disorder	68.39	63.56	3.3	1.7
new-thyroid	90.98	88.89	7.3	2.4
pima-diabetes	73.87	72.04	3.1	1.6
wdbc	91.83	89.13	4.1	2.1
wine	97.11	88.83	6.8	2.3
Average	77.90	73.79	5.85	1.81