行政院國家科學委員會專題研究計畫　期中進度報告

類神經網路特徵選取及其應用之研究(1/3)

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-009-072-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學工業工程與管理學系

計畫主持人：蘇朝墩

計畫參與人員：李得盛、許志華、薛友仁

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 29 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※

※

※

※　　　　　　類神經網路特徵選取及其應用之研究（**1/3**）

※

※

※

※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC 91-2213-E-009-072

執行期間：91 年 8 月 1 日至 92 年 7 月 31 日

計畫主持人：蘇朝墩

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學工業工程與管理學系

中　華　民　國 九 十二 年　五　月　二 十　日

行政院國家科學委員會專題研究計畫成果報告

類神經網路特徵選取及其應用之研究（1/3）

Feature Selection for Neural Classifiers with Application to Operation Management

計畫編號：NSC 91-2213-E-009-072

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：蘇朝墩　　國立交通大學工業工程與管理學系

計畫參與人員：李得盛、許志華、薛友仁

國立交通大學工業工程與管理學系

一、中文摘要

特徵選取是分類問題的主要工作之一，特徵選取的主要目的乃是要選取重要的特徵並獲得一可接受的分類精確度。類神經網路是處理分類問題的一個很受歡迎的方法，類神經網路的結構愈是簡化，其愈可改善網路之解釋和預測的能力。也就是說，降低特徵個數可以減少計算上的複雜度，並有可能可以提昇分類的精確度。

本計劃第一年提出一以基因演算法為基礎的類神經網路特徵選取方法，我們使用三個類神經網路（倒傳遞類神經網路、放射基準機能網路、學習向量量化網路）分別與基因演算法結合，並進行比較與討論。

關鍵詞：特徵選取；基因演算法；類神經網路

Abstract

Feature selection is one of the major tasks in classification problems. The main purpose of feature selection is to select the essential features used in the classification while maintaining an acceptable classification accuracy. Moreover, neural networks emerged as an attractive alternative to pattern classification. The simplified structure of neural networks can improve the interpretability and predictability of the network. That is to say, reducing the dimensions of the feature space can reduce the computational complexity and may increase estimated performance of the classifiers.

In this project (the first year), we propose a genetic algorithm (GA) based neural method for feature selection. Three neural network models, back-propagation network (BPN), radial basis function (RBF), and learning vector quantization (LVQ) are employed in the proposed approach for discussion.

Keywords: Feature selection, genetic algorithm, neural network

二、緣由與目的

The classification problem involves multi-dimensional information systems used to determine which item belongs to what class out of a set of possible classes. A number of variables stored in the multi-dimensional data sets are sometimes called features. Unfortunately, numerous of potential features considerably impact the efficiency of the classifiers, such as the *k*-nearest neighbor, C4.5 (Quinlan, 1993) and back-propagation classifier. Most of these features are either partially or completely irrelevant or redundant to the classified target. It is not known in advance which features will provide sufficient information to discriminate between the classes. It is also infeasible to classify the patterns or objects using all of the possible features. Feature selection is one of the major tasks in classification problems. The main purpose of feature selection is to select the features used in the classification while maintaining an acceptable classification accuracy.

Various algorithms have been used for feature selection in the past decades. Narendra and Fukunaga (1977) introduced the branch and bound algorithm to avoid the cost associated with searching through all of the feature subsets. Later, Foroutan and Sklansky (1987) introduced the concept of approximate monotonicity and used the branch and bound method to select features for piecewise linear classifiers. Siedlecki and Sklansky (1989) integrated the genetic algorithm (GA) with the *k*-nearest neighbor (KNN) classifier to solve the feature selection problem. The GA plays the role of selector to select a subset of features that can best describe the classification performance evaluated using the KNN classifier. In this work, we employed the idea from Siedlecki and Sklansky (1989) and compared the feature selection classification performance using neural network classifier.

The GA is a powerful feature selection tool, especially when the dimensions of the original feature set are large (Siedlecki et al., 1989). Reducing the dimensions of the feature space not only reduces the computational complexity, but it also increases estimated performance of the classifiers. Kudo (2000) presented three versions of the feature selection. These three problem types result in specific objectives and different types of optimization. The first problem version involves determining a subset that yields the lowest classifier error rate. This version of the problem leads to unconstrained combinatorial optimization in which the error rate is the search criterion. The second problem version involves seeking the smallest feature subset that has an error rate below a given threshold. This version leads to a constrained combinatorial optimization task, in which the error rate serves as a constraint and the number of features is the search criterion. The third problem version involves finding a compromise objective between version one and version two by minimizing the penalty function. In this work, we will focus the proposed method in the second problem version.

三、結果與討論

The proposed GA-based feature selection approach is similar to the KNN-GA approach developed by Siedlecki and Sklansky (1989) in the literature. In the KNN-GA approach, given a set of feature vectors of the form $X=\{x_1, x_2, \ldots, x_n\}$, the GA produces a transformed set of vectors of the form $X'=\{w_1x_1, w_2x_2, \ldots, w_nx_n\}$ where w_i is a weight associated with feature i. A KNN classifier is used to evaluate each set of feature weights. This algorithm introduces a binary masking vector along with feature weight vector on the chromosome. Using the GA optimization technique, this algorithm can efficiently search for the optimal solution, the maximal classification accuracy or minimal classification error rate.

The proposed GA-based feature selection method is summarized as follows:

Phase I: Training the neural networks

Step 1: Collect a set of observed data.

Step 2: Divide the data into training and test data sets.

Step 3: Set the training parameters (such as learning rate, momentum, etc).

Step 4: Train the different neural network structures.
Step 5: Choose a trained network with the highest accuracy rate and remember the weights between the layers.
Phase II: GA optimization process
Step 1: Initialize the GA chromosome (assigning 1 to each binary node in mask vector along with the weights obtained from step 5 on phase I)
Step 2: Set the GA operating conditions (e.g. generation size, population size, crossover rate and mutation rate).
Step 3: Use the input data to obtain the initial solution.
Step 4: Repeat steps 5-8 until a stopping condition is reached.
Step 5: Calculate the output value by entering the input data sets and mask vector into the trained network (obtained from step 5, phase I).
Step 6: Transfer the output value to a class label.
Step 7: Calculate the classification accuracy rate by comparing the target with the class label.
Step 8: Select, crossover and mutate the chromosome according to the fitness function (equation 3.1).
Step 9: Obtain an optimal subset of input variables based on the binary mask vector (denoted by "1") and weights between the layers
Phase III: Testing process
Step 1: Find the test data.
Step 2: Apply the data to the trained GA-based neural classifier from step 9 in phase II.
Step 3: Obtain the classification results.

四、計劃成果自評

This project proposed a novel GA-based algorithm that integrated the GA selector and neural network classifiers for feature selection. In this work, the GA searched for near-optimal solutions for the subsets in the feature space. The proposed method can preserve the accuracy using the best subset of features instead of using all of the available features. This algorithm can improve the performance because it can eliminate noisy and irrelevant features that may mislead the learning process. The results demonstrated that the LVQ-GA outperformed the BP-GA and RBF-GA algorithms in both examples because of its learning algorithm and neural network structure. The classification performance also shows that the proposed method is robust and effective in a multi-dimensional data system.

The above research results have been submitted for publication in *International Journal of Systems Science*.

五、參考文獻

1. Foroutan, I. and Sklansky, J., 1977, Feature selection for automatic classification of non-Gaussian data. *IEEE Transactions on System, Man and Cybernet,* **17**, 187-198.
2. Kudo, M. and Sklansky, J., 2000, Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition*, **33**, pp.25-41.
3. Quinlan, J. R., 1993, C4.5: programs for machine learning, Morgan Kaufmann, San Mateo, CA.
4. Siedlecki, W. and Sklansky, J., 1988, ON automatic feature selection. *International Journal of Pattern Recognition and artificial Intelligence*, **2**, 197-220.
5. Siedlecki, W. and Sklansky, J., 1989, A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, **10**, 335-347.