Combining Nearest Neighborhood Classifiers using Genetic Programming

Abdul Majid, Asifullah Khan and Anwar M. Mirza
Faculty of Computer Science & Engineering, GIK Institute, Ghulam Ishaq Khan (GIK) Institute of Engineering Science & Technology, Topi-23460, Swabi, PAKISTAN { majid, akhan and mirza}@giki.edu.pk

Abstract

In this paper, GP based intelligent scheme has been used to develop an Optimal Composite Classifier (OCC) from individual nearest neighbor (NN) classifiers. In the combining scheme, first, the predicted information is extracted from the component classifiers. Then, GP is used to develop OCC having better performance than individual NN classifiers. The experimental results demonstrate that the combined decision space of OCC is more effective. Further, we observed that heterogeneous combination of classifiers has more promising results than their homogenous one. Another side advantage of our GP based intelligent combination scheme is that it automatically incorporates the issues of optimal model selection of NN classifiers to achieve a higher performance prediction model.

Keywords: kNN classifier; Receiver Operating Characteristics Curve (ROC); Area under the Convex Hull (AUCH); Genetic Programming (GP).

1. Introduction

There is a considerable research in engineering and medical applications to get more useful information from raw data [1]. Researchers are working on new emerging fields of Machine Learning, Data Mining and Knowledge Discovery to develop an intelligent information extraction model [3]. They are always in search of high performance prediction model.

In the current example of a thyroid disease diagnosis system, practitioners are interested in high performance classification model which precisely represent the status of patient of thyroid disease in term of adjustable tradeoff between true positive and false positive parameters of ROC curve. For this purpose, they require an improved ROC model, specifically, for a weak patient; the cost of misclassification may be harmful. Such patients may not afford even a small misclassification cost, that is, for healthy tissues to be classified as malignant during cancer therapy or other way round.

In a recent survey report in UK, it was estimated that 80% death rate can be controlled if fatal diseases may be detected and predicted precisely in its early stages. Therefore, a high performance classification model is inevitable a solution which can save thousands of lives [2].

In binary classification problem, to design an optimal class mapping functions that assign a data sample x to the correct class from a given set of training samples $s_i = (x_i, y_i)$, where $x_i \in S$ is considered to be in one of two classes $\{C_1, C_2\}$, and $y_i \in \{-1,1\}$. In kNN classifier, usually, the k nearest neighbors are determine from the training set, $x_i \in S_t$, and then classify x_i as the most common class among the k neighbors. A crucial issue in kNN classifier is the optimal choice of neighbor size k [1].

A lot of research has been done to optimize kNN classification models. Various researchers have offered application dependent solutions. Such techniques have been discussed in detail in our previous work [4]-[5]. Main point of the research is the selection of suitable models for kNN. Mostly, such approaches are based on the modification of the selection and voting scheme of kNN classifiers. However, these methods become intractable when the search space is large [4]. So, there is a need of an intelligent method to optimize such parametric models.

We are offering an alternative GP based combination of classifiers approach to optimize different NN classifiers. This combination idea have been successfully apply to optimize various classification models [3]-[9], [12]. The combination of classifiers has attained a considerable attention of the research community for higher classification performance [3]-[11]. The combination technique can improve the prediction accuracy by parameters tuning. Such a technique is expected to be more accurate than a single classifier [10]. But there are many challenges to combine the results provided by each classifier, due to the lack of general classifiers combination rules. In such situation GP offer an optimal search technique. GP has the capability to produce an optimal function

(numerical classifier) as well as there are good chances for GP based technique to combine classification models effectively.

Previously, GP has been used in combining different classification models like, artificial neural network [9], decision trees, SVM of different kernel functions [6]-[7], linear classifiers [4] and various statistical classifiers [8], to develop a composite classification model.

In this paper, we address issues related to the improvement of NN classifiers through two main contributions First, GP based OCC function is developed by combining the predictions of individually trained NN classifiers. This enables us to construct high performance composite decision space. Secondly, GP based combination of NN classifiers implicitly incorporates the desire of the selection of suitable k for high performance NN classifiers.

The remaining paper is organized as follows: In section 2, we describe the method of combining classifiers. Results and discussion are presented in section 3. Finally, conclusions and future direction are given in section 4.

2. Methodology of combining classifiers

There are many problems for a general combination of NN classifiers, which are applicable to all type of input data. However, GP combination method has might effectively combine a small number of complementary components nearest neighbor classifiers. In our scheme, individual classifiers are tuned over a wide range of decision thresholds. GP is used to evolve an appropriate combination functions. GP fitness function, ROC curve, is based on a thorough measure of classifier performance.

The basic architecture of the construction of nearest neighbor based OCC is given in the Figure 1. This architecture is consisting of two layers. A set of NN component classifiers forms the first layer. This layer of the combination scheme is based on the concept of stacking the predicted information [11]. The concept of global scope is used to obtain the prediction of individual component classifiers [8]. In global scope scheme, each component classifier is applied to all the instances in a data set (instance space). GP based combining classifiers technique forms the second layer.

Suppose, there are n learning algorithms, $K_1, \cdots, K_n$. The thyroid disease dataset S is partitioned into three sets by using holdout method, such that: $S = \{ X_1 \cup X_2 \cup X_3 \}$ with $s_i = (x_i, y_i)$. A set of individual NN classifiers $C_1, \cdots, C_n$ are constructed during training on training dataset X_1 such that $C_j = K_j(x_{1i}),\ x_{1i} \in X_1$,

where, $j = 1, 2, \cdots, n$ and $i = 1, 2, \cdots, m$, and m is the number of training samples in X_1. A new feature set $(\hat{y}_i^1, \cdots, \hat{y}_i^n)$ is achieved by stacking the predictions of NN component classifiers as:

$$\hat{y}_i^j = C_j(x_{2i}), \quad x_{2i} \in X_2 \ , \ \forall \quad j = 1, 2, \cdots, n$$

GP meta-learning process is based on a new training data space of $((\hat{y}_i^1, \cdots, \hat{y}_i^n), \ \forall \ i = 1, \cdots, m$. The predictions $(\hat{y}_i^1, \cdots, \hat{y}_i^n)$ of individual classifiers are used as unary functions in GP tree. These unary functions are combined during GP crossover and mutation operations to obtain a high performance OCC.

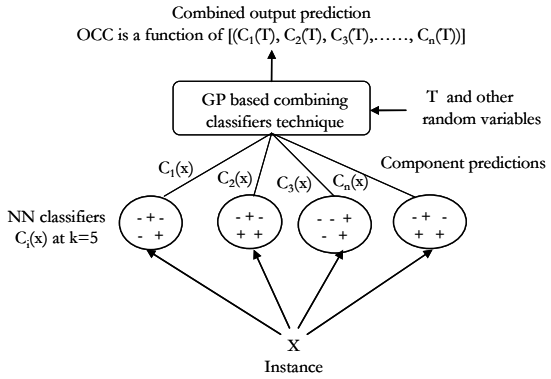

Figure 1: An example of architecture of developing a Homo-5 at k=5 for n classes

In developing homogeneous and heterogeneous OCC, separate GP simulations are carried out. In homogeneous combination, the OCC classifier is a function of any one of its component classifiers $f(kNN5)$, $f(kNN7)$, $f(kNN9)$. In heterogeneous combination, OCC might be a function of more than one component classifiers, As far as, the development of a homogenous combination of classifier is concerned, the best three of each of kNN5, kNN7 and kNN9, may be used separately in the combination set such that at a time in the combination set there are only three component classifiers.

Block diagram to develop OCC is shown in the Figure 2. GP combination technique can be divided into there are five main parts. In the first part, description of the thyroid data set is given. Then, construction and selection of component classifiers is explained, while the third part is based on the computation of fitness (AUCH of ROC curve) of each individual in the population. Fourth major part is the

GP mechanism to develop OCC. In the fifth part, GP based training pseudo code are given.

2.1. Thyroid disease data set

This data set [13] is used to train and test various classifier. The ann.train data set consists of 3772 data samples and ann.test contains 3475 data samples. Originally, it is a three class problem; two classes for abnormal thyroid are combined into one class to form a binary class. The data is divided into three equal sets training data 1, testing data1 and testing data2 using holdout method. For the experimental study, the most distinguishable last five and three features are used to train and test the classifiers.

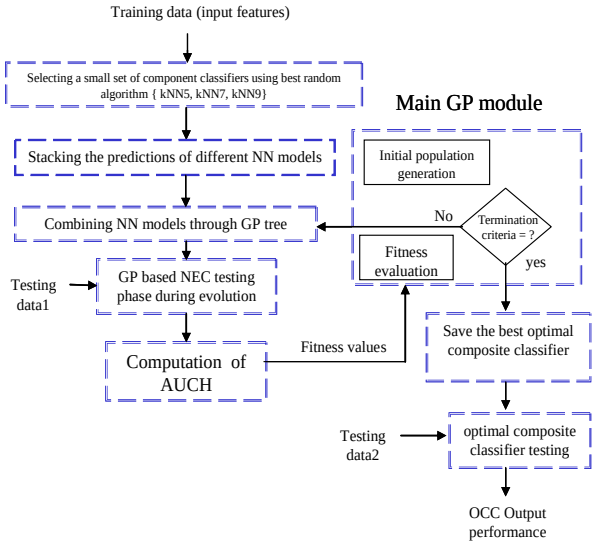

Figure 2: Block diagram to develop GP based OCC.

2.2. Selection of component classifiers

First step in the selection of suitable component classifiers is the construction of individual NN classifiers by using different k. several higher performing complementary nearest neighbor classifiers are constructed by using best random sampling algorithm [15]. In this algorithm, best three out of ten NN classifiers are selected by using different k = 5, 7 and 9. For example, in case of kNN9 classifier, k = 9 and s = 10 are chosen. And only the best three component (of kNN9) classifiers are selected for further experiments. Similar procedure is repeated to construct kNN3, kNN5 and etc. The selection algorithm might have chosen complementary component classifiers.

2.3. Computation of AUCH of ROC curve

ROC curve are important parameters to analyze the performance of classifiers. When there is no prior knowledge of the true ratio of misclassification costs for a classification system, ROC curve is a reasonable performance measure [4]. In order to evaluate general performance of classifier under a certain tradeoff, ROC curve is selected. We characterize the classification system over its entire operating range. Figure 3 shows the status of tissues of a patient with and without a disease arranged according to the value of a diagnostic test. The test results are given in a scale [0, 1] where 0 represents a normal case (negative result) and 1 an abnormal case (positive result). A cut-off value determines the number of true positives, true negatives, false positives and false negatives.

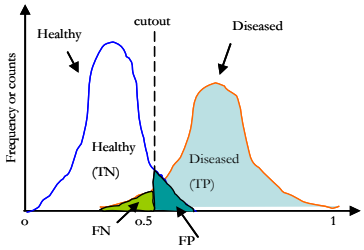

Figure 3: Status of healthy and diseased tissues in term of TP and FP.

Convex hull of a classifier's ROC is taken as a measure of the performance of various classification models [3]-[9]. In order to calculate AUCH of each individual, in the GP evolution at the specified feature set, testing data 1 is used. The decision values of a classifier are obtained. Then, TPR and FPR of the entire test samples in the testing data 1 are computed at varying threshold T in the range of [0-1]. These values of TPR and FPR are plotted to produce ROC curve and then AUCH of ROC curve is determined.

2.4. GP module

There are three main steps in GP module to develop an OCC, as follows: First step, in GP module is the suitable selection of arithmetic, logic and trigonometric operators according to application, e.g. +, -, *, protected division, LOG, EXP, SIN and COS. Three NN classifiers kNN5, kNN7 and kNN9 are used as unary functions in GP tree as shown in Figure 4. Some other specific logical statements are also used, e.g. greater than (gt) and less than (lt). Threshold T is taken as a variable terminal in the range of [0, 1]. Randomly generated numbers in the range of [-1, 1] are used as constant terminals in GP tree.

Second step, is the selection of initial population generation method. The most commonly used Ramped half and half method is used. In order to generate next population three GP operators, namely replication, mutation and crossover, are used. They help in converging to optimal/near optimal solution. GPLAB software [12] adjusts the ratio of crossover and mutation.

In the last step, selection and evaluation of each individual in the population is carried out. Fitness in term of AUCH of ROC is used as a feedback to the GP module providing fitness of each individual. Figure *2* shows the usage of fitness function as a feedback. Higher AUCHs of individuals indicate higher performance. At the end of GP evolution, the best-evolved OCC expression is obtained.

Similarly, homogenous combinations of classifiers (Homo-5, Homo-7 and Homo-9) are developed by using different combinations of kNN5, kNN7 and kNN9 classifiers such that only one type of NN classifiers are selected as a unary function in the GP tree as shown in figure 4. All the necessary parameter settings for the GPLAB software are shown in table 1

In order to analyze and compare the performance of OCC, their AUCH is computed on a novel test data2. The experimental results are repeated 10 times by randomly permuting the test data2 and their average performances are reported.

2.5. GP based training pseudo code

Suppose S_o: represents the data in the original training set, S_t: representation the new training meta-data, x: data set instance, $C(x)$: class of x, OCC: a composite classifier in the form of homogenous or heterogeneous, C_i: component classifiers, $c_i(x)$: prediction of C_i on input x, n: number of component classifiers each represent only one class. Here, we are explaining the general training procedure for n classes, n = 2, for our binary class problem.

2.5.1 Train-Composite classifier (S_o, OCC)

Step 1: All the input instances x from S_o are obtained to form a prediction vector corresponding to each classifier C_i form the combination set.

Step 2: Collect $[c_0(x), c_1(x), \cdots, c_n(x)]$ in S_t for each component classifier.

Step 3: Start GP combining method while using prediction set as unary function in GP tree. A threshold T is used as a variable to compute (AUCH of ROC curve) a fitness function.

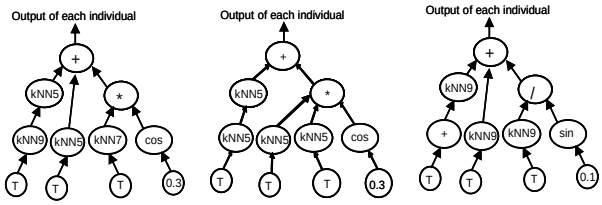

Figure 4: GP trees representation of heterogeneous and homogeneous combination of different NN component classifiers.

Table 1 GP Parameters Selection

Objective:	**To develop an OCC**
Function set	+, -, *, protected division, GT, LT, EXP. (for hetero. combination of kNN5, kNN7 and kNN9 are used as Unary Functions) and for homo. combination only kNN5, kNN7 or kNN9 are used as unary functions
Terminal Set:	Terminals set consist of variable T and random constants in between -1 to +1
Fitness :	AUCH at 11 ROC points.
Selection:	Generational
Wrapper:	Positive if >= 0, else Negative.
Pop. Size:	300
Initial pop.	Ramped half and half
Sampling	Tournament
Survival mechanism	Keep best
Termination:	Generation 55

2.5.2 Pseudo code for the classification

Step 1: Apply (OCC, x) to composite classifier and instance from S_o.

Step 2: $x = [c_0(x), c_1(x), \cdots, c_n(x)]$, stack the predictions to form meta-data.

Step3: Compute $OCC(x)$.

During GP learning process, training data taken from S_o are used as input to the component classifiers to form a new (meat-data) training vector. In order to classify a new test x', first, classify it using component classifiers C_i, $i = 1, \cdots, n$ and then form a vector of predicted labels $[c_0(x'), c_1(x'), \cdots, c_n(x')]$. OCC is then applied to this vector of predictions to obtain the final output result.

3. Results and Discussion

During GP training phase, different composite classifiers are trained and then test using the last three and five features of thyroid data. Then the performance of these classifiers is estimated in term of AUCH of ROC curves, by calculating TPR and FPR of each classifier under a threshold T of the rang of [0,1]. The results obtained are shown in Figure 5 and Figure 6 collectively.

A useful finding is observed from Figure 5 that OCC developed by heterogeneous combinations performs better than their homogenous combinations (Homo-5, Homo-7 and Komo-9) with higher AUCH values. This might be the fact that in the heterogeneous combination, we have used different (complementary) component classifiers that might have a better chance to learn with different aspect of data during GP evolution process. Therefore, combined decision space of heterogeneous combination of component classifiers is more effective than the space generated by their homogenous combination.

Bar chart in Figure 6 shows the comparison of experimental results in more summarized form at two different feature sets of thyroid data. In the comparative analysis, the performances of various classification models are compared with reference to the performance of 1-NN classifier. In 1-NN classifier, decision is made on the basis of Euclidian distance using all the training instances from a given test sample, which is computationally expensive. While, using (complementary) components classifier, we have achieved comparable performance with 1-NN classifier. Each NN component classifier store only a small number of k prototypes, taken from training instances, for decision i.e. k=5, 7, 9. These component classifiers were, initially, generated to create diversity in the combination set.

The experimental results in Figure 6 shows that the overall performance of GP based developed composite classifiers are superior to all individual component classifiers and 1-NN classifiers. Therefore, OCC using GP based combination method might have a better chance to extract useful information from its component classifiers to construct a more optimal decision space. The general order of the performance of classifiers is:

Hetro > Homo > kNN classifiers

From Figure 6, it is observed that all the classifiers; in general, improve their performance with the increase of the size of features set. This might be the fact that, in general, with the provision of more useful information, better decisions are expected.

This improvement in the performance of ROC curve due to GP based composite classifiers causes low FPR and high TPR values on ROC curve, that is, points on the curve are shifting towards upper left corner. They are the most desirable points. As a result, OCC provides better decision at the current classification problem. This type of situation is highly desirable in those applications where cost of FPR is too important, such as, detection/diagnosis of malignant and benign tissues due to various fatal diseases, lungs/ liver cancer, tumor detection, thyroid disease before cancer therapy.

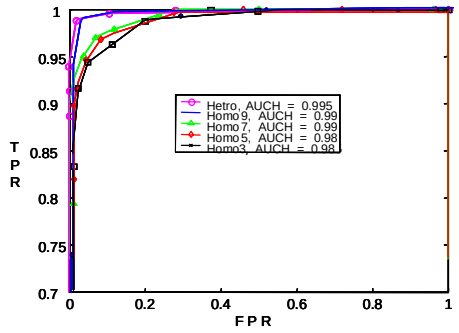

Figure 5: AUCH of ROC curve of heterogeneous and homogenous classifiers at last three features set of thyroid data.

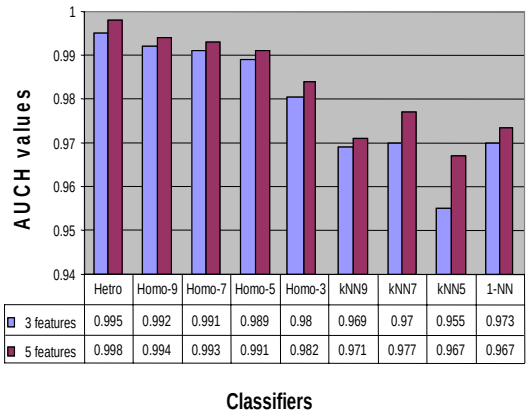

Figure 6: Bar Chart of AUCHs of different Classifiers at three and five features of thyroid data

Expression of the best evolved OCC classifier, in prefix form, using 5 features is:

OCC(kNN5,kNN7,kNN9)=plus(le(sin(kNN7),le(plus(sin(kNN5),sin(sin(kNN9))),abs(minus(kNN5,abs(minus(sin(sin(kNN5)),0.12826))))))),plus(minus(kNN7,.15),plus(plus(sin(plus(le(0.047),kNN7), ((abs(le(plus (kNN7,(kNN5,plus (times(divide

(kNN5,plus(sin(ab (kNN9)),plus(cos(kNN7) ,sin(sin(kNN5))))) ,0.51) le(sin(sin(abs (minus(sin(kNN7)

This expression shows the dependency of OCC function on its constituent (kNN5, kNN7 and kNN9) classifiers, random constants, arithmetic operators and other special operators.

4. Conclusions and future directions

The work of this paper has investigated that GP combination technique can intelligently combine different NN classifiers for better decision. With GP meta-search strategy, we have automatically reduced outliers which cause to degrade the performance of kNN classifiers by finding optimal numeric function from GP solution space. Moreover, the decision space generated by heterogeneous combination is more informative than homogenous combination.

This method is flexible and general under certain environment. It may develop optimal composite classifiers for a binary classification problem, specifically for imbalance medical data, where it is very difficult for conventional classification models to train on such kind of data. In future, we intend to apply this technique on other medical data.

Our current and previous investigations [4]-[8] have explored the GP potential to optimally combine the decision information from its constituent classifiers.

References

[1] R.O. Duda, P.E. Hart, and D. G. Stork, "*Pattern Classification"*, John Wiley & Son s, Inc., New York, second edition, 2001.

[2] K. Rajpoot, and N. Rajpoot, "SVM Optimization for Hyperspectral Colon Tissue Cell Classification", *LNCS3217, Springer-Verlag* Sep. 2004, pp. 829-837.

[3] W.B. Langdon and S.J. Barrett, "Genetic Programming in Data Mining for Drug Discovery", in Evolutionary Computing in Data Mining, *Physica Verlag*, 2004, pp. 211-235.

[4] A. Majid, "Optimizing and Combination of classifiers Using Genetic Programming", *Ph.D. thesis*, FCS&E, GIK institute of Engineering Science and Technology, 2005.

[5] A. Majid, A. Khan, and A.M. Mirza, "Improving Performance of Nearest Neighborhood Classifier Using Genetic Programming", in *the proc. of International conference on machine learning and its applications ICMLA'04*, Louisville, KY, USA, Dec. 2004.

[6] A. Majid, A. Khan and A.M. Mirza,, "Combination of Support Vector Machines Using Genetic Programming", submitted *in the International Journal of Hybrid Intelligent Systems (IJHIS)*, 2005, pp.1-11.

[7] A. Majid, A. Khan and A.M. Mirza, "Intelligent combination of Kernels information for improved classification", accepted *in the International conference on machine learning and its applications ICMLA'05*, Los Angeles, California, USA, 2005.

[8] A. Khan, A. Majid, and A.M. Mirza, "Combination and Optimization of Classifiers in Gender Classification Using Genetic Programming", *International Journal of Knowledge-Based Intelligent Engineering Systems*, vol. 8, 2004.

[9] W.B. Langdon, S.J. Barrett, and B.F. Buxton, "Combining Decision Trees and Neural Networks for Drug Discovery in Genetic Programming", *Proc. of the 5th European Conference, EuroGP2002*, Springer-Verlag, 2002.

[10] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorization", *Information Fusion*, Elsevier, 6(1), March 2005, pp. 5–20.

[11] S. Dzeroski and B. Zenko, "Is combining classifiers with stacking better than selecting the best one?", *Machine Learning*, 54(3), 2004, pp. 255–273.

[12] http://gplab.sourceforge.net

[13] http://www.ics.uci.edu/~mlearn/MLRepository

[15] D.B. Skalak, "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms", *In Proceedings of the Eleventh International Conference on Machine Learning*, Morgan Kaufmann, 1994.