Optimizing the Member Selection for Ensembles of Classifiers: An Application of Rainfall Forecasting in Sri Lanka

Nagahamulla, H.R.K.

Dept of Computing & Information Systems, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka

Email: harshaninag@yahoo.com

Abstract

A collection of classifiers trained to do the same task is called an ensemble of classifiers. Ensembles can be created using a set of classifiers of the same type or using a set of classifiers in different types (Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, ect.). The generalization ability of an ensemble is significantly increased than that of a single classifier. To achieve increased generalization ability the members of an ensemble has to be accurate (able to produce correct forecast) and diverse (errors in different regions of the error space). However accuracy and diversity are two conflicting conditions that have to be balanced carefully to achieve good performance. Thus members for an ensemble need to be selected carefully in order for them to have the perfect balance between accuracy and diversity. This study aims to optimize the member selection for the ensembles using Genetic Algorithms (GA) to increase the ensemble performance in the context of time series forecasting. The selected application is rainfall forecasting in Sri Lanka. Rainfall is very difficult to forecast accurately because it is a very complex hydrological process. Forecasting rainfall requires manipulating huge datasets with large number of variables. But accurate rainfall forecasts are in high demand because of the close relationship rainfall has with human life.

There are three steps in creating an ensemble; creating the pool of classifiers, selecting the members for the ensemble from the pool and combining the selected members using a combiner method. The performance of the ensemble depends on the techniques used in each of these steps. First a pool of classifiers, including different types of classifiers such as SVM, Back Propagation Network (BPN), Radial Basis Function Network (RBFN) and Generalized Regression Neural Network (GRNN) was created by training the classifiers using different training data. Then a number of ensembles were created by selecting different combinations of classifiers from the pool randomly and combining them using a separate GRNN. These ensembles were the initial population of the GA. A simple binary genetic algorithm was then used to create new generations of ensembles and find the ensemble that gave the best result. The fitness of the ensembles were calculated to balance the accuracy and the diversity of the ensemble. The chromosomes were ranked and sorted according to their fitness. Then, the mating pool was prepared by selecting the chromosomes with highest fitness and the pairs were selected using roulette wheel rank weighting. Mating took place using one point crossover with 0.6 crossover probability and the new generation was mutated with 0.1 mutation probability. To train and test the models rainfall data from 1961 to 2001 (41 years) of Colombo, Sri

Lanka is used. Input data set consisted of 26 variables obtained from the NCEP_1961-2001 dataset and the output data was daily rainfall of Colombo. The dataset was partitioned to training data (first 60%), validation data (next 20%) and testing data (the remaining, more recent 20%). To create different training datasets from the available training data moving block bootstrap method was used. The dataset containing 10958 records was split into 9863 overlapping blocks of length 1096 and out of these 9863 blocks 10 blocks were selected to train each classifier. To validate the proposed method another two ensembles were created using two well known ensemble creation methods bagging and boosting. The performance of the best ensemble (ENN-GA) was compared with the performance of a single SVM, BPN, RBFN, GRNN, the best performing ensemble in the initial population (ENN), bagging model and the boosting model. Forecasting accuracy of each model was measured for the test dataset using Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R2).

The best performing ensemble comprised of two SVM, three BPN, two RBFN and five GRNN. The number of generations for convergence was 287. The following table summarizes the results for individual classifiers, ENN, bagging, boosting and ENN-GA.

Classifier Type	**SVM**	**BPN**	**RBFN**	**GRNN**	**ENN**	**Bagging**	**Boosting**	**ENN-GA**
RMSE	9.21	9.44	8.69	8.22	8.16	8.11	8.04	7.99
MAE	5.44	5.36	5.04	4.76	4.98	5.14	4.77	4.67
R2	0.50	0.47	0.55	0.60	0.55	0.54	0.58	0.61
T-Critical two-tail	1.960554664	1.960554811	1.960554811	1.960554811	1.960558216	1.960558365	1.960557621	1.960562414
run time (s)	1	1	1	<1	2	2	2	2

The proposed model ENN-GA gave more accurate results than the single classifiers used in the study with smaller RMSE and MAE values and larger R2 and the time and space requirements were very small. The proposed model managed to predict the overall rainfall with reasonable accuracy; zero rainfall accurately, smaller rainfall with slight differences and some higher rainfall with considerable differences. These higher differences were obtained for very high rainfall that occurred suddenly. Although the number of these occurrences were very few the difference between the actual and forecasted rainfall was high. The RMSE values were larger compared to MAE values because the errors in high rainfall were magnified in RMSE.

The proposed method outperform the single classifiers, ENN model and bagging and boosting models in forecasting rainfall for Colombo, Sri Lanka.

Keywords: *Artificial Neural Network, Ensembles, Genetic Algorithms, Time Series Forecasting*