ANALYSIS OF METHODS AND STRATEGIES FOR DIVERSITY BASED GENETATION OF CLASSIFIER ENSEMBLE

M. S. Joshi¹, V. Y. Kulkarni²

¹ student, M.E. Computer Engineering, MIT, Pune, Maharashtra, India
² Associate Professor, Department of Computer Engineering, MIT, Pune, Maharashtra, India

Abstract

A classifier ensemble is a group of individual base classifiers. Each classifier is trained individually by modifying the given data set to achieve diversity. During the testing phase the results given by each classifier are collected to give the final result using a technique called as majority voting. Empirical results prove that diversity amongst the base classifiers improves the accuracy of the result. The diversity can be achieved either based on datasets used for training or based on attribute selection methods. In this paper we summarize the study of the ensemble generation methods. We also propose to test a different approach that will find the distance based diversity of randomly generated datasets before the ensemble generation which will not only optimize the construction process but will improve the efficiency also. The Random Forest (RF) is a widely used algorithm for generating ensemble of decision tree classifiers. During the training phase it uses bagging technique to create bootstrapped samples. The best way to test the model is using 10-fold cross validation. As the empirical results show that the diversity in the classifiers improve the overall accuracy of the ensemble, it is needed to verify how diverse the training samples are. This will verify the existence of the base classifier in the ensemble based on the nature of randomly sampled dataset. The base classifier we yield with this approach certainly maintains a higher diversity and the accuracy of the ensemble improves further with these diverse base classifiers. We aim to achieve higher diversity with less number of base classifiers retained in the model.

Key Words: base classifier, bagging, bootstrapping, classifier ensemble, diversity, prediction accuracy.

1. INTRODUCTION

The main concept of the ensemble methodology is to build several individual classifiers, and group them in order to obtain a classifier that is improved over every one of them. Ensemble learning is well established as a research area in the field of machine learning. Some widely used techniques in this context, such as bagging and boosting, had proved to be effective in solving classification tasks. Each classifier in the ensemble is referred as a base classifier. The success of a classifier ensemble is depends on the fact that the base classifiers perform diversely. Empirical results have shown that there exists positive correlation between accuracy of the ensemble and diversity among the base classifiers. Further, most of the existing ensemble learning algorithms can be found building diverse base classifiers implicitly. [8][9][10]

In an ensemble learning algorithm, if the term diversity is defined explicitly and optimized, it is said that the methods achieves diversity explicitly. In other case the algorithm achieves diversity implicitly. Since many ensemble algorithms have been successfully proposed by seeking diversity implicitly, it is important to know whether we can perform better by seeking the diversity explicitly. Many measures, of the relation between two classifiers output can be derived from the literature of statistics. There is less clarity on the subject when three or more classifiers are concerned. There are methods and formulas aiming at quantifying diversity but, because of the lack of a definition, little is put on a systematic basis. The general anticipation is that diversity measures will be helpful in designing the individual classifiers and the combination technology. [12][14]

The Different diversity measures used in the literature are grouped in two categories, pair-wise measures and non-pair-wise measures. As per the literature found so far, many researchers mainly focus on pair-wise measures, as they agree on the fact that it mainly helps in achieving the desired diversity than finding the diversity as a measure of whole ensemble directly. Similarly working on diversity of every individual base classifier helps in better generation of desired ensemble with improved accuracy. [14]

The diversity and ensemble generation methods those we reviewed are summarized below.
2. DIVERSITY GENERATION

Diversity is an important condition for obtaining more accurate ensembles. Diversified classifiers lead to uncorrelated classifications, which in turn improve classification accuracy. However, in the classification task, there is no complete theory to explain why and how diversity between individual models contributes toward overall ensemble accuracy. Here we have reviewed the methods and approaches conceptually, those we found useful in aiming at diversity of base classifiers in view of improvising the accuracy of the ensemble.

2.1 Manipulating the Training Samples

In this method, each classifier is trained on a different variation or subset of the original dataset. It is useful for inducers whose variance-error factor is relatively large i.e. small variations in the training set can cause a major change in the trained classifier. Most popular ensemble procedures (such as bagging and boosting) belong to this category. [12]

2.2 Sampling Methods

The distribution of records among the different classifier could be random as in the bagging algorithm or in the arbiter trees. Other methods distribute the records based on the class distribution such that the class distribution in each subset is approximately the same as that in the entire dataset. It has been shown that proportional distribution as used in combiner trees can achieve higher accuracy than random distribution. Instead of performing sampling with replacement, we can manipulate the weights that are assigned to each example in the training sample. In this case, the base induction algorithm should be able to take these weights into consideration. [12]

2.3 Starting point in the hypothesis space

Diversity can be obtained by starting the search in the Hypothesis Space from different starting points. For example in neural networks, we can assign different initial weights to the networks links. Experimental study indicate that the resulting networks differed in the number of cycles in which they took to converge upon a solution, and in whether they converged at all. While it is very simple way to gain diversity, it is now generally accepted that it is not sufficient for achieving good diversity.

2.4 Changing the target attribute representation

In the methods that manipulate the target attribute, instead of inducing a single complicated classifier, several classifiers with different and usually simpler representations of the target attribute are induced. This manipulation can be based on an aggregation of the original targets values (known as Concept Aggregation) or more complicated functions (known as Function Decomposition). Classical concept aggregation replaces the original target attribute with a function, such that the domain of the new target attribute is smaller than the original one. [2]

2.5 Partitioning the search space

The idea is that each member in the ensemble explores a different part of the search space. Thus, the original instance space is divided into several sub-spaces. Each sub-space is considered independently and the total model is a (possibly soft) union of such simpler models. When using this approach, one should decide if the subspaces will overlap. At one extreme, the original problem is decomposed into several mutually exclusive sub-problems, such that each sub problem is solved using a dedicated classifier. In such cases, the classifiers may have significant variations in their overall performance over different parts of the input space. At the other extreme, each classifier solves the same original task. In such cases, If the individual classifiers are then appropriately chosen and trained properly, their performances will be (relatively) comparable in any region of the problem space. However, usually the sub-spaces may have soft boundaries, namely subspaces are allowed to overlap. This leads to get diversity in the base classifiers. [5][7][12]

3. ENSEMBLE GENERATION METHODS

The research work on ensemble generation methods has been considering different approaches to focus on the diversity and accuracy of the ensemble. These approaches and their strategy are discussed in this section.

3.1 An Information Theoretic approach

The IT-based approach measures the accuracy and diversity of a given ensemble initially. Then, given a pool of M available classifiers, a classifier ensemble can be created so as to maximize the ITS (Information Theory Score i.e. aiming at obtaining the highest accuracy and diversity). An iterative method was proposed for the creation of an ensemble, where nodes are added incrementally. This approach allows for the creation of the classifier ensembles without requiring an exhaustive search among all possible classifier combinations. Q-statistic that is based on the probability of coincident errors for two classifiers and the overall Q-statistic of an ensemble is computed as the average Q-value for all pair wise combination of classifiers. [8]
3.2 Dynamic classifier selection

In the case of fusion-based methods, also known as combination-based methods, one assumes that all base classifiers are equally expert in the whole feature space. Thus, the decisions of all classifiers are taken into account for any input pattern. There are many different fusion-based methods reported in the literature. Those can be classified as linear or non-linear. For example, the simplest linear way to combine multiple classifiers is the sum and average of the outputs. For the non-linear case, some examples are combination using majority voting strategies. [1]

For selection-based methods, unlike for the fusion-based methods, only one classifier is used to classify the input pattern. In order to do so, a procedure to choose a member of the ensemble that will be employed to make the decision has to be defined. Such a procedure often receives as input the testing pattern to be classified. That is, the choice of a classifier that will be used to produce the final output is made during the operation phase. Preference is given to more certain classifiers. In order to provide the most suitable output for the input pattern, hybrid methods use characteristics of both selection and fusion techniques. Indeed, in this context, there is often a procedure to decide whether to use the selection or fusion method. Also, the main idea is to use selection only and if only the best classifier is good, according to some criterion, at classifying the testing pattern. Otherwise, a fusion method is used.

There are three main choices in the design of a MCS: the organization of its components, the system components and the combination methods that will be used. In terms of the organization of its components, a MCS can be defined as modular and ensemble. In the modular approach, each classifier becomes responsible for a part of the whole system and they are usually linked in a serial way. In contrast, in the ensemble approach, all classifiers are able to answer to the same task in a parallel or redundant way. Moreover, there exists a combination module that is responsible for providing the overall system output. In this paper, the kind of MCS analyzed is of the ensemble type.

There are a vast number of combination-based methods reported in the literature. They could be classified according to their characteristics as Linear or Non-linear. In Linear combination methods, currently, the simplest ways to combine multiple neural networks are the sum and average of the neural networks outputs. In Non-linear methods, the class includes rank-based combiners, such as Borda Count, majority voting strategies, the Dempster-Shafer technique, fuzzy integral, neural networks and genetic algorithms. In Selection-based Methods, unlike the combination-based methods, only one classifier is needed to correctly classify the input pattern in selection-based methods. In order to do so, it is important to define a process to choose a member of the ensemble to make the decision, which is usually based on the input pattern to be classified. The choice of a classifier to label is made during the operation phase. This choice is typically based on the certainty of the current decision. Preference is given to more certain classifiers. [1][3]

3.3 Clustering and selection

In this context, there is an initial pool of base classifiers. In this method a set of classifiers from the initial pool is chosen based on both their accuracy and diversity, whereas in a single classifier is elected from the pool based only on its accuracy. In fact, a classifier is only chosen if it is statistically better than other classifiers in the pool. Otherwise, a fusion method is used with all the base classifiers. It is important to understand that, in all experiments, researchers use some fusion-based method to combine the output of the base classifiers forming the ensemble. [1][10][11]

3.3 Diversity based selection of components

Classifiers may be fused in different ways, for example, by combining decisions or by combining discriminate function outputs. We can consider combining classifiers through their decisions. The decisions of the set of classifiers, selected according to their rankings, are combined into a decision fusion vector and the final decision is then made by classifying the decision fusion vector into one of the signal classes. Method to fuse data sets of an ensemble into a data fusion vector includes summing/averaging, pooling, and concatenation. Out of the three data fusion methods, data concatenation is the most promising, therefore, the concatenation model is used to fuse the selected data sets. However, other classifiers such as the nearest neighbour, k-nearest neighbour, and the nearest mean classifier can be implemented. [3]

Ensemble selective strategy based on Bagging is composed of two major phases, namely construction of the initial ensemble by randomly selecting base model and iterative refinement of the ensemble members. [2][3]

3.4 ASDM algorithm: Attribute Selection and Diversity Measure

ASDM algorithm selects randomly the attributes from the original attribute sets, and constructs the training subset based on the selected attributes, and then the base classifier is learned on the training subsets. Because the base classifier was trained on different subsets, there may be diversity between different classifiers. The entire diversity between the learned classifier and ensemble are calculated. If the measure of diversity is higher than a given threshold, the learned classifier is added into the ensemble, otherwise this learned classifier is discard. In the final, majority voting is adopted. [6][7]
3.6 Selective Ensemble Learning Method

Selective Ensemble learning is to select ensemble models with selective strategies after generating many different base models. Different selective strategies can get different ensemble learning algorithms. We can utilize feature-based or data-based method to select training data which are used for generating base models. In this paper, we mainly study data-based method to select training data and introduce four different selective ensemble learning approaches which include Hill Climbing, Ensemble Backward Sequential Selection, Ensemble Forward Selection and Clustering Selection. We use decision tree and neural network algorithms to train base models. [11]

Hill-Climbing(HC) - Hill Climbing ensemble selective strategy based on Bagging is composed of two major phases, namely construction of the initial ensemble by randomly selecting base model and iterative refinement of the ensemble members. Initial ensemble members are formed using the randomly selective method. The second phase is aimed to improve the accuracy and diversity of the ensemble classifiers. For all the learning models, an attempt which includes adding or deleting one model is made to switch ensemble members, and keep better changes. This process is continued until no further improvements are possible.

Ensemble Backward Sequential Selection – It begins with all learning models and repeatedly removes a model whose remove yields the maximal performance improvement. The cycle repeats until no improvement is obtained.

Ensemble Forward Selection – It begins with zero attributes, evaluates all base models with exactly one model, and selects the one with the best performance. It then adds to the Result Set that yields the best performance for models of the next larger size. The cycle repeats until no improvement is obtained.

3.7 Classifier Ensemble based on Feature Transformation

A basic classification system consists of three levels. In the first level the raw data of input patterns are collected. After that, in next level, some features are extracted from these data. Finally a classifier is trained based on extracted features. If one is interested to create a collection of diverse classifiers, he must create classifiers which differ from each others in one of these levels. In some ensemble based methods such as bagging and boosting learning is done on different patterns sets. Some other methods, e.g. Random Subspace Method (RSM), use different feature sets to make base classifiers different. Using different learning method or the same learning method with different parameters can also be used as a diversity creation method. Here they propose a diversity creation method based on feature transformation. Let C be the number of target classes of input patterns. Train M base classifiers where M is equal 1 to G*C. These classifiers are divided to C groups, where each group corresponds to a specific class. The features given to the classifiers of each group are the ones that discriminate the corresponding class of this group from other classes in a proper manner. It should be noted that although the input features to each classifier are appropriate for discriminating one class from others, all classifiers are asked to learn all classes. [3][9][13]

4. IDENTIFYING THE DIVERSITY

As a case study, we are using Random Forest algorithm, RF classifier is a collection of individual decision tree classifiers; therefore it can improve classification accuracy. In our system, each tree is constructed using a different bootstrap sample of the training dataset. For each observation, each individual tree votes for one class and the RF classifier classifies the class that has the plurality of votes. The detailed steps of the random forest algorithm are as follows: Let N trees be the number of trees to build for each of N trees iterations.[4][7]

Random forest belongs to the family of classifier ensemble methods that use randomization to produce a diverse pool of individual classifiers. It can be defined as a classifier consisting of a collection of L tree-structured classifier \( \{ h(x, \Theta_k) \}, k = 1 \ldots L \) where the \( \{ \Theta_k \} \) are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input \( x \). A random forest can be built by randomly sampling a feature set for each decision tree (as in Random Subspaces), and/or by random sampling a training subset for each decision tree (as in Bagging). [2][4]

The overall process of construction of RF can be explained by following steps.

1. Select a new bootstrap sample from training set.
2. Grow an un-pruned tree on this bootstrap.
3. At each internal node, randomly select ‘m’ predictors and determine the best split based on selected measures as attribute selection measure using only these predictors.
4. Output overall prediction as the majority vote from all individually trained trees.
Empirical results have proved that in a group of classifiers if the diversity is increased, it improves the accuracy of the whole ensemble as such. [14]

<table>
<thead>
<tr>
<th>$D_k$ correct (1)</th>
<th>$D_k$ wrong (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N^{11}$</td>
<td>$N^{10}$</td>
</tr>
<tr>
<td>$N^{01}$</td>
<td>$N^{00}$</td>
</tr>
</tbody>
</table>

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

Figure 1: Architecture of classifier ensemble (RF)

Figure 2: Relationship between pair of classifiers

Each base classifier from the ensemble is compared with all other base classifiers to find the diversity between them. The measures which evaluate the diversity in this fashion are called as pair-wise measures. The Q statistics pair-wise measure is calculated in this experimentation. [7][14]

The main notations used are summarized as follows:
- $L$: total number of base classifiers.
- $N$: total number of training samples.
- $m_i$: margin of an ensemble on the training sample $x_i$.
- $P$: average classification accuracy of the base classifiers on the training data.
- $p_j$: classification accuracy of the base classifier $h_j$.
- $O_{ij}$: output of the classifier $h_j$ on the training sample $x_i$.
- $div$: diversity among the base classifiers in the ensemble.
- $D_1, D_2, D_3, \ldots, D_n$: Bootstrapped Datasets

For all datasets, for each attribute $a_1, a_2, a_3, \ldots, a_m$
Find the distance as a diversity measure and evaluate the accuracy of ensemble using 10-fold cross validation.

5. EXPERIMENTATION

As an initial experimentation we developed a java program within WEKA that will find out the parameters like accuracy, mean absolute error, RMS error and Kappa statistic. We tested the program using five UCI datasets namely, car, hypothyroid, kr-vs-kp, sick and spambase. We create a pool of 100 samples and build the RF with number of trees in the step of 5, ranging from 5 to 100. The selection of samples is based on the ranks given to all 100 samples.[7][8] We also compare the accuracy given by RF with selective samples based on ranks with that of RF build in standard WEKA.

**Time Complexity:** The process requires one to iterate through Set of size $N$ of bootstrapped samples to find the vector of attributes statistics. So it requires $N$ iterations initially. To find the distance and rank of each dataset, it runs $N^2$ times. So the overall time complexity of approach is $O(N^3)$

**Space Complexity:** The subsets are evaluated simultaneously and those with favourable results are stored back onto the disk in the form of text-file. Space requirements vary linearly with the size of subset due to the number of trees contained in the set. Maximum space requirement while running turns out to be $O(N)$. Hence the space complexity is $O(N)$.

6. RESULTS AND ANALYSIS

The results obtained show the correlation between the accuracy and the error measures taken on the dataset.
These statistics are helpful in finding the diversity within the dataset used to construct the base tree. For statistically independent classifiers, the expectation of $Q_{i;k}$ is 0. $Q$ varies between -1 and 1. Classifiers that tend to recognize the same objects correctly will have positive values of $Q$, and those which commit errors on different objects will render $Q$ negative.

The disagreement measure is to characterize the diversity between a base classifier and a complementary classifier and for measuring diversity in decision forests. It is the ratio between the number of observations on which one classifier is correct and the other is incorrect to the total number of observations. The double-fault measure is used to form a pairwise diversity matrix for a classifier pool and subsequently to select classifiers that are least related. It is defined as the proportion of the cases that have been misclassified by both classifiers.

### References


