Optimization of Ensemble Classifier System based on Multiple Objectives Genetic Algorithm

Author
Nguyen, Tien Thanh, Liew, Alan Wee-Chung, Pham, Xuan Cuong, Nguyen, Mai Phuong

Published
2014

Conference Title
Machine Learning and Cybernetics (ICMLC), 2014 International Conference on (Volume:1 )

DOI
https://doi.org/10.1109/ICMLC.2014.7009090

Copyright Statement
Copyright 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Downloaded from
http://hdl.handle.net/10072/66706

Link to published version
OPTIMIZATION FOR ENSEMBLE SYSTEM BASED ON MULTIPLE OBJECTIVES GENETIC APPROACH

TIEN THAH NGUYEN¹, ALAN WEE CHUNG LIEW¹, XUAN CUONG PHAM², MAI PHUONG NGUYEN³

¹ School of Information Technology, Griffith University, Australia.
² School of Information Technology, Hanoi University of Science and Technology, Vietnam.
³ College of Business, Massey University, New Zealand.
Email: tienthanh.nguyen2@griffithuni.edu.au

Abstract:
This paper introduces a mechanism to learn optimal combining classifier algorithms associated with features and classifiers from an ensemble system. By using Genetic Algorithm approach that focuses on 3 objectives namely number of correct classified observations and number of selected features and classifiers, optimal solution can be achieved after several interactions of crossover and mutation. We also employ OWA operator in which a weight vector is built by Linear Decreasing (LD) function to average outputs from combining algorithms. Experiments on 2 well-known UCI Machine Learning Repository datasets demonstrate benefits of our approach compared with other state-of-art ensemble method like Decision Template and SCANN as well as all fixed combining algorithms in the ensemble system.

Keywords:
Combining Classifiers; Ensemble Method; Genetic Algorithm; multiple objectives; OWA operator

1. Introduction

1.1 Combining Classifiers based on Stacking Algorithm

In this paper, we concentrate on ensemble models with fixed base classifiers and their combined outputs to make predictions for unlabeled observations in test set. Wolpert [1] introduced Stacking Algorithm to generate meta-data as output of base classifiers by dividing the original dataset to B disjoined parts in which each part plays as test set while the others as training set. Label of observations in test set will be predicted by model formed by base classifiers and training set. Several state-of-art combining algorithms have been proposed which operate on meta-data to generate hypothesis about relationship between feature vector and its class label. A simple combining approach that just work with meta-data of test set was introduced by Kittler [6]. It consists of 6 combining rules, namely Sum, Product, Vote, Min, Max, and Average. According to our knowledge, Vote and Sum are the most common rules. Kuncheva [2] introduced the concept of Decision Template for each class generated from meta-data and the associated class label. She proposed 11 measurements between each Decision Template and meta-data of an observation to construct a classification framework. Meanwhile, Mezt [7] combined 3 algorithms namely Stacking, Correspondent Analysis and K Nearest Neighbor in a single algorithm called SCANN. His idea was to find relationship between rows (include all observations) and columns (include classification results of all classifiers). To implement it, he built an indicator matrix using {0, 1} labels and then used Singular Value Decomposition (SVD) transformation to make new scaled space. Ting [5] proposed Multiple Response Linear Regression algorithm (MLR) to combine posterior probabilities of each observation based on sum of weights calculated from K Linear Regression functions.

Recently, Szepannek [9] developed idea from pairwise combining by finding which classifier is best for a pair including Class i and Class j (i,j=1, M, i ≠ j) and used a pairwise coupling algorithm to combine outputs of all pairs to make posterior for each class. Zhang [10] used linear programming to find weight that each classifier puts on a particular class. Sen [8] introduced an approach inspired by MLR by applying hinge loss function to the combiner instead of using conventional least square loss. By using new function with regularization, he proposed three different combination, namely weighted sum, dependent weighted sum and linear stacked generalization based on different regularizations with group sparsity.

1.2. Genetic Approach in ensemble system

There are well-known optimization strategies applied to ensemble system to discover optimal subset of features and classifiers to improve the performance of classifier fusion. Here, we only discuss approaches based on Genetic Algorithm. Kuncheva [4] introduced two GA versions in which features are selected by join and disjoin mechanism. In
the former, she encoded feature by \( \{0,1,\ldots,K\} \) where \( k \leq k \leq K \) means that feature is only used by \( k \)th classifier and 0 means that feature is not used by any classifiers. In the latter, she added classifier encoding in the same chromosome with feature encoding but both of them work independently in crossover and mutation stage. She developed an encoding method based on Venn diagram for feature encoding and integer values for classifier encoding. Nanni [11] employed GA to improve SCANN [7] by building representations where each includes encoding of \( M \) classes. Gabrys [12] tried to put classifier, feature and rule encoding in a single chromosome as 3-dimensional cube. These approaches, however, are difficult to implement because of the complicated crossover stage. We address these issues with a new GA model for multi classifier system which achieves both effectiveness and ease of implementation.

1.3. OWA operator

The Ordered Weighted Averaging operator (OWA) [16] is one of the most well-known operations applied commonly in Decision Making Systems. This operator has been using to compute average value based on weight but instead of focusing on original data, it is linked with order of data; as a result elements on specific location like head, tail or middle can receive more attention than the others.

An OWA operator of dimension \( P \) is a mapping from \( \mathbb{R}^P \rightarrow \mathbb{R} \) that has an associated weight vector \( \{\omega_i\}_{i=1}^P \) such that \( \omega_i \in [0,1] \) and \( \sum_{i=1}^P \omega_i = 1 \). The function \( F(a_i | i = 1,P) \) determines aggregated value of the arguments \( A = \{a_i | i = 1,P\} : F = \sum_{i=1}^P b_i \omega_i \) where \( b_i \) is \( i \)th largest value in \( A \).

An interesting thing to work with OWA operators is the type of weights to be used. We introduce four popular types of function to build the weight vector for each \( \omega(k) \), which are Constant (eqn. 1), Gaussian (eqn. 2), Linear Increasing (eqn. 3) and Linear Decreasing (eqn. 4).

\[
\omega(k) = \frac{1}{K} \quad (1)
\]

\[
\omega(k) = e^{-\frac{(k-\mu)^2}{2\sigma^2}} \quad (2)
\]

\[
\mu = \frac{K+1}{2} \quad \text{and} \quad \sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K (k-\mu)^2}
\]

\[
\omega(k) = \frac{2k}{K(K+1)} \quad (3)
\]

\[
\omega(k) = \frac{2(K+1-k)}{K(K+1)} \quad (4)
\]

The novelty of our work is that in our ensemble, several fixed combining algorithms are present simultaneously. We propose a combining mechanism based on OWA operator to learning from training set to achieve:

- An optimal subset of combining algorithms
- An optimal subset of feature for each combination
- An optimal subset of classifier for each combination

Optimal solution resulted from multiple objective GA approach will be selected to predict label for all unlabeled observations.

2. Methodology

We denote \( N \) as the number of observations, \( K \) as the number of base classifiers, \( M \) as the number of classes \( \{W_j\}_j \), \( D \) as dimension of the original data, and \( T \) as the number of combining classifiers. For each observation \( X_i \), \( P_i(W_j | X_i) \) is the probability that \( X_i \) belongs to class \( W_j \) given by \( k \)th classifier. The meta-data (also called Level1 data) of all observations, a \( N \times M \times K \)-posterior probability matrix \( \{P_i(W_j | X_i)\} \) is defined by:

\[
\begin{bmatrix}
P_1(W_1 | X_1) & P_1(W_2 | X_1) & \ldots & P_1(W_M | X_1) \\
\vdots & \vdots & \ddots & \vdots \\
P_i(W_1 | X_i) & P_i(W_2 | X_i) & \ldots & P_i(W_M | X_i) \\
\vdots & \vdots & \ddots & \vdots \\
P_N(W_1 | X_N) & P_N(W_2 | X_N) & \ldots & P_N(W_M | X_N)
\end{bmatrix}
\]
Whereas Level1 data of an observation $X$ is defined by:

$$\text{Level1}(X) = \begin{bmatrix} P_1(W_1 | X) & \cdots & P_1(W_d | X) \\ \vdots & \ddots & \vdots \\ P_k(W_1 | X) & \cdots & P_k(W_d | X) \end{bmatrix}$$  \hfill (6)

Figure 1 shows the learning mechanism on training set. Our purpose here is to find which combining algorithms associated with classifiers plus corresponding features are optimal for a particular training set. By applying GA on the training set, solution can be obtained through several interactions of crossover and mutation among individuals in population.

Specifically, the structure of chromosomes is given by two-part architecture (Figure 2):
- Part1: encoding for combining algorithms by using $\{0, 1\}$ elements, its length is $T$
- Part2: encoding for classifiers including features used by them which are employed for a particular combining algorithms by using $\{0, 1\}$ elements, its length is $(K+K\times D)\times T$

Traditionally, researchers have only focus on the accuracy of a classification system. In our approach, the other two objectives that are also considered are subset of features and classifiers used by the system. By reducing the number of selected features as well as classifiers, computation and storage cost will be saved. To achieve that, GA approach is employed with 3 objectives:
- Maximize number of correct classified observations in training set, denoted by $f_1$
- Minimize number of features used by system (denoted by $f_2$) so as to help reduce storage cost.
• Minimize number of classifiers used by system (denoted by $f_3$) so as to help reduce processing time as well as improve performance.

In this paper, we define an objective function formed by $f_1$, $f_2$ and $f_3$ as $F = f(f_1, f_2, f_3) = w(1)f(1) - w(2)f(2) - w(3)f(3)$, in which $w = (w(1), w(2), w(3))$ is a weight vector specified by users. To average the outputs from the selected combining algorithms, the fourth OWA function (eqn. 4) is chosen since we focus more on combining algorithm with higher accuracy. By using Linear Decreasing function, we put larger weights on higher accuracy algorithms so as to help increase converge speed in GA approach.

During testing, based on the encoding of optimal individual resulted from the training process we can determine:

• Which combining classifiers are being used or not used in ensemble system?
• Which classifiers plus their corresponding features are being used or not used for a particular selected classifier?

Features and classifiers are chosen corresponding with selected combining algorithms to classify each observation in test set to form Level1 data (eqn. 6). Next, this meta-data is employed as input to each selected combining algorithm to output class label as prediction of hypothesis associated with the algorithm. All outputs will be combined by a majority voting mechanism.

3. Experimental Results

In our experiment, we simply select the 6 simple rules namely Sum, Product, Max, Median, Min and Majority Vote as the fixed combining classifiers set in the model. As mentioned above, combining rules are very simple in computation compared with other approaches like Decision Template and SCANN. Here we want to show the benefit of our model, i.e. it has only simple combining algorithms but can achieve high accuracy. Three base classifiers are used: Linear Discriminant Analysis (LDA), Naïve Bayes, and KNN (with K set to 5, denoted as KNN(5)). We also initialize several parameters in GA, as detailed in TABLE 1. A 10-fold cross validation is used since we do not have separate training set and test set. We run the experiment 10 times so in total we have 100 test outcomes for each dataset. Evaluation is performed on Matlab2013a environment.

We choose 2 popular UCI datasets [19] namely Sonar and Vehicle in evaluation (TABLE 2). To show the advantage of our model, we compare its error rate with other single combining algorithms, including the 6 combining rules, Decision Template [2], MLR [5], and SCANN [7]. Experimental results of all algorithms are shown in TABLE 3 and 4.

| TABLE 1: INITIALIZATION VALUE OF PARAMETERS IN GA |
|-----------|-----------|
| Parameters                | Value    |
| Population Size           | 20        |
| Mutation Rate (Part 1)    | 0.02      |
| Mutation Rate (Part 2)    | 0.03      |
| Weight vector for multiple objectives | $w(1) = 1, w(2) = 0.1$ and $w(3) = 0.1$ (Sonar) |
|                          | $w(1) = 0.5, w(2) = 0.1$ and $w(3) = 0.1$ (Vehicle) |

| TABLE 2: INFORMATION OF 2 DATASETS IN EXPERIMENT |
|-----------|-----------|
| Name of Dataset | Sonar | Vehicle |
| Number of Observations | 208 | 864 |
| Number of Attributes | 60 | 18 |
| Feature Type | Real | Integer |
| Number of Classes | 2 | 4 |

TABLE 3. ERROR RATES OF 6 COMBINING RULES ON 2 DATASETS
### Table 4. Error Rates of Other Combining Algorithms and Proposed Model

<table>
<thead>
<tr>
<th>File name</th>
<th>Decision Template</th>
<th>SCANN</th>
<th>MLR</th>
<th>Proposed model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>Sonar</td>
<td>0.2129</td>
<td>8.80E-03</td>
<td>0.2128</td>
<td>8.01E-03</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.2161</td>
<td>1.50E-03</td>
<td>0.2224</td>
<td>1.54E-03</td>
</tr>
</tbody>
</table>

### Table 5. Average Number of Selected Features and Classifiers in 100 Tests

<table>
<thead>
<tr>
<th></th>
<th># of selected features</th>
<th># of selected classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonar</td>
<td>37.18</td>
<td>1.62</td>
</tr>
<tr>
<td>Vehicle</td>
<td>16.66</td>
<td>1.57</td>
</tr>
</tbody>
</table>

Figure 3. Trend of combining algorithms selected among all experiments (Left) Vehicle (Right) Sonar

Remarkably, the proposed model is better than all combining algorithms in fixed set for both datasets. Comparing with other state-of-art algorithms like Decision Template, SCANN and MLR, our model is highly competitive. This is an excellent outcome since our model is formed by 6 simple combining rules (TABLE 3 and 4).

Another interesting point to note is the average number of features and classifiers chosen. Only 37.18 and 16.66 features are used by our model in the final solution among the entire 60 and 18 features on Sonar and Vehicle, respectively. Besides, our model only uses on average 1.62 and 1.57 classifiers from the 3 input classifiers on the 2 data files (TABLE 5).

In addition, the percentage of different combining algorithms used in 200 tests is illustrated in Figure 3. It is worth noticing that up to 198 cases, only single combining algorithm is chosen as optimal solution. Clearly, by cutting the smallest values from the sequence \( a_1 \leq a_2 \leq \ldots \leq a_T \), we will have a new sequence with higher average. Hence, if a chromosome has \( T \) algorithms ordered by correct number of classified observations \( a_1 \leq a_2 \leq \ldots \leq a_T \), it may be replaced by chromosomes in the next generation with fewer algorithms by cutting the smallest values in the sequence. As a result, in the end, the final chromosome with optimal solution only has several algorithms with same accuracy. Hence, we use a special OWA operator to focus more attention on outstanding algorithms. This boosts the system’s performance by increasing the speed of convergence in GA.

### 4. Conclusion and Future Development

In this paper, we have introduced a novel approach to learn optimal subset of combining classifier algorithms corresponding with subset of classifiers and features to construct a powerful and flexible ensemble system. We used multiple objectives GA to discover optimal solution of system architecture by proposing novel structure for chromosome and combined outputs from selected combination by OWA operator. Experiments conducted on two well-known datasets demonstrates the advantage of our
approach that optimal combining algorithms with optimal features and classifiers can be found.

Our work can be extended in several directions. For example, in scenarios where more than one combining algorithm is chosen, voting strategy plays an important role to gather hypotheses to form single one. Here we simply used majority vote strategy on the predictions of selected combining algorithms. To improve effectiveness and efficiency of final prediction, other methods such as Majority Teach Minority by Zhou [18] can be considered.

Another interesting point to be addressed is averaging outputs from combining classifiers to form an objective function in the fitness function. In our experiment, we used a particular OWA operator to compute average value for the correctly classified observations. An open question here is exploring new strategies to combine outputs from the entire fixed input combining algorithms.

References