Improving the Performance of Artificial Neural Networks via Instance Selection and Feature Dimensionality Reduction

Ali Abroudi, Mohammad Shokouhifar, and Fardad Farokhi

Abstract—This paper presents a hybrid approach with two phases for improving the performance of training artificial neural networks (ANNs) by selection of the most important instances for training, and then reduction the dimensionality of features. The ANNs which are applied in this paper for validation, are included Multi-Layer Perceptron (MLP) and Neuro-Fuzzy Network (NFN). In the first phase, the Modified Fast Condensed Nearest Neighbor (MFCNN) algorithm is used to construct the subset with instances very close to the decision boundary. It leads to achieve the instances more useful for training the network. And in the second phase, an Ant-based approach to the supervised reduction of feature dimensionality is introduced, aims to reduce the complexity, and improve the accuracy of learning the ANN. The main purpose of this method is to enhance the classification performance by improving the quality of the training set. Experimental results illustrated the efficiency of the proposed approach.

Index Terms—Artificial neural networks, instance selection, feature selection, ant colony optimization, EDA, MFCNN.

I. INTRODUCTION

The learning procedure has different steps including, constructing a Training Set (TS), training the network, testing the behavior of the network on the new instances. The first step consists of editing and condensing the data. The most important objective of any abstraction technique is to obtain a TS with the best performance in classification process. To build an optimal TS, we have to select the instances and features in an appropriate way to guarantee attaining to the best result.

The whole data is located in a matrix which its rows show the instances and the columns are the features. The instance selection schemes generally try to reduce the number of rows in entire dataset with no loss in the classification performance [1]. Non-parametric K-Nearest Neighbor (KNN) rule [2], predicts the label of a new instance by computing a similarity scale between the selected one and all other instances in TS. Condensed Nearest Neighbor rule (CNN) [3], was introduced for the nearest neighbor classification method and tried to build a condensed subset from the whole set of examples. The basic idea of this algorithm is that if an example is not correctly classified, this may occur because it is a boundary instance, and it should be considered in the selected condense set.

Fast Condensed Nearest Neighbor (FCNN) [4], starts by set S that is initialized with centroid of each class in TS. Then, until the TS is not completely classified with condensed set S, the algorithm selects for insertion a representative of the misclassified points of each Voronoi (Voronoi of point P ∈ S is the set of all instances in TS that are closer to P than to any other instance in S) cell induced by the current subset. Edited Nearest Neighbor (ENN) removes noisy instances, as well as close border points, leaving smoother decision boundaries [5]. It edits out those prototypes that misclassified using KNN rule. Allknn [6] is based on the ENN method. This algorithm, for k=1 to m, marks as removable any prototype misclassified by its k NNs. When the algorithm iterates m times, it eliminates the prototypes flagged as removable.

The Patterns with Ordered Projections (POP) algorithm [7] consists of removing the instances that are not boundary ones. That is, the behavior of this algorithm is in opposite direction from that of Allknn and ENN. The Reduced Nearest Neighbor rule (RNN) produces the smallest subset which classifies all the learning prototypes correctly [8]. In Drop3 algorithm [9], we have the associate concept, so that the associate of the instance $x_i$, is the instance $x_a$ which has $x_i$ as nearest neighbor. This algorithm eliminates $x_i$ if at least majority of its associates in TS would be classified correctly without $x_i$. Unlike most of instance selection methods that concentrate on improving the performance of KNN algorithm, in our previous work [10], we used the concept of FCNN rule aims to select the most efficient instances for training the ANNs.

Since most of the practical problems are written in high dimensional spaces, their features have to be either selected or combined into a fewer number. Feature selection approaches mostly are employed for reduction of the features which must be extracted. However, the feature combination methods are applied to improve the performance of training the ANN, which is the main concern in this paper. All feature selection algorithms according to subset evaluation strategies, can be categorized in wrapper and filter approaches. In filters, a classifier independent index is used for evaluation the gathered subsets; while the classification error is used in wrappers [11].

Typically feature selection algorithms which are based on Evolutionary Algorithms (EAs), perform better than the conventional techniques especially for high-dimensional data sets. In [12], a rough set approach based on Ant Colony Optimization (ACO) was proposed, which adopts mutual information based feature significance as heuristic information. In [13], an ACO was proposed, which in the each iteration, the classification accuracy is used for evaluation of feature subsets represented by ants. In the
previous works, we proposed both filter and wrapper feature selection approaches based on ACO [14]-[17], and artificial bee colony (ABC) algorithm [18], [19].

Feature combination algorithms can be classified into two categories, depending on whether the problem is supervised or not. In unsupervised situations, Principal Component Analysis (PCA) [20] is widely used as feature dimensionality reduction algorithm. And when supervision is available, Fisher’s Discriminant Analysis (FDA) [21] is mostly employed. In FDA, a projected class is located far from the others as possible, and the patterns are grouped around their means. This is a smart measure for class separation. FDA is a dimensionality reduction technique, not a classification method. It is mostly used as a preprocessing step before classification. Evolutionary Discriminant Analysis (EDA) [22], was introduced by Sierra and Echeverria as a modified version of FDA, which use a genetic algorithm to optimize the classification error.

In this paper, a modified FCNN (named MFCNN) is introduced for effective selection of most important instances for train the ANN. And then an Ant-based EDA (named AntEDA) is proposed in order to achieve the reduced feature space, and it has been shown that this approach is more powerful to correctly classify the patterns compare to the others as possible, and the patterns are grouped around their means.

III. METHODOLOGY

A. Instance Selection Using MFCNN Algorithm

In this sub-section, we describe the instance selection
method that we use for our hybrid model. Random selection of instances may not be the most adequate way for training the ANNs. The FCNN is one of the most efficient techniques for selecting instances, thus we perform some modifications on this method and name it: Modified FCNN (MFCNN).

Our proposal of MFCNN has two levels that are shown in Fig. 2. In level 1 of MFCNN, the centroids of the all classes in TS are determined and then transferred to condensed set S as initial instances. Then, during each iteration, for each instance in S, the Voronoi set is obtained. Then, those samples of Voronoi that have a different label from corresponding instance (we called this set as Voren) are picked and all class centroids of Vorenare calculated and eventually inserted to S.

The stop criterion of this level is similar to previous level. is selected and inserted to P sample to SS.

This level stops when no more instances can be added to S, in other words, when TS is correctly classified using condensed-set S. In level 2 of MFCNN, we try to find the smaller condensed set (double-condensed set) from the output of level 1. In this level, we apply the FCNN base rule on S set and finally the SS set is produced.

In level 2, first of all, the centroids of all classes in resultant S are determined and placed into the new condensed set SS (During each iteration, we used S and SS as temporary sets for S and SS sets respectively). And after that, the Voren set of each instance P ∈ SS is constructed, then the closest sample to P is selected and inserted to SS as representative. The stop criterion of this level is similar to previous level.

B. Feature Dimensionality Reduction Using Evolutionary Discriminant Analysis and ACO Algorithm

ACO has been introduced based on the foraging behavior of real ants [24]. Ants deposit some pheromone on ground when they walking, in order to mark some favorably paths that should be followed by the other ants. Pheromone will be evaporated soon, but in the short time it remained on the ground as trail of ants. The ants in option to choose between two paths, probability choose that path which has more amount of pheromone. We have been previously proposed some ACO-based feature selection and ranking approaches [14-17]. In this paper, we propose a new feature combination algorithm using ACO (named AntEDA), to achieve the most useful projected features for classification.

The used structure for performing AntEDA can be seen in Fig. 3. In this structure $f_i$ is the $i^{th}$ original feature, and $\bar{f}_j$ is the $j^{th}$ projected feature, which is calculated by “Eq. (10)”. The $m$ and $n$ ($n=m$) are respectively the number of all original features and the number of projected features. The mean of projected feature $\bar{f}_j$ for the instances belonging to class $k$ is calculated as “Eq. (11)”.

$$\bar{f}_j = \sum_{i=1}^{m} f_i w_{ij} \quad (10)$$

$$\bar{M}^k_j = \sum_{i=1}^{m} M^k_i w_{ij} \quad (11)$$

The AntEDA classifies patterns as belonging to the class of the closest projected mean as “Eq. (12)”. In this manner, the instance $x$ belongs to the class that it closest to its mean.

$$\text{class}(x) = \arg\min_{k} \left( \frac{1}{n} \sum_{i=1}^{n} (\bar{f}_j - \bar{M}^k_j)^2 \right) \quad (12)$$

We are searching for the optimal linear projections. To do that, we propose an ACO algorithm to optimize the weights of the structure. It leads to minimize the number of misclassified patterns, in order to achieve the projected features more useful than original ones. Note that, the total weights of the AntEDA structure are equal to $mxn$. Typically ACO has performed to discrete problems. The allowable values for each weight is considered from the set of $\{-1, -0.99, -0.98, ..., 0.98, 0.99, 1\}$. The optimization process using ACO for obtaining the optimal weights is performed by generation a colony consists of $k$ artificial ants to search through the possible weights of AntEDA structure. To construct a solution, each ant selects values for each weight among allowable values, with respect to probability rule as “Eq. (13)”. The probability of $k^{th}$ ant to select the $i^{th}$ value for the $j^{th}$ weight at time $t$ is given as:

$$P^i_j(t) = \frac{\rho^i_j(t)}{\sum_{i=0}^{T} \rho^i_j(t)} \quad (13)$$

where $\rho^i_j$ is the pheromone amount of $i^{th}$ value for the $j^{th}$ weight. $T$ is the number of discreteness of the weights which is 201. After all ants construct their projections, they are evaluated by “Eq. (14)”. The projection’s cost is proportional to the number of misclassified patterns. In the other words, the
cost of a given projection is equal to the percentage of all misclassified patterns for that projection.

\[ J = \frac{1}{N} \sum_{x} L(t(x), class(x)) \]  \hspace{1cm} (14)

Which \( t(x) \) is target output of pattern \( x \). If \( t(x) = class(x) \) then \( L = 0 \), and otherwise \( L = 1 \). And \( N \) is the number of all patterns within the dataset. After all ants traverse among the problem space and have constructed their projections, pheromone trails on each node (each value of each weight) will be updated. The change in the pheromone of each node is according to:

\[ \Delta \rho_{ij}(t) = \rho + \left\{ \begin{array}{ll}
Q & \text{for edges of the best ants} \\
0 & \text{otherwise}
\end{array} \right. \]  \hspace{1cm} (15)

where \( \rho_{ij}(t) \) is the pheromone amount on node \((i,j)\) at this iteration. Note that, the node \((i,j)\) means the \(i^{th}\) value of \(j^{th}\) weight. And \(\lambda(0<\lambda<1)\) is the evaporation of pheromone trails for all nodes. \(Q\) is a user-defined constant that adjust the amount of pheromone deposition on the associated nodes of the best ants. During each iteration, the pheromone amount of all nodes must be bounded between two values. The overall steps to achieve the optimal weights using ACO algorithm can be summarized as follow:

- An initial random population of ants is generated. Each ant randomly selects a possible value for each weight, and constructs a projection.
- The following steps are repeated until the error rate is less than \textit{max-iterations}, or a prescribed number of generations is reached.
  1) The gathered projections are evaluated by the number of misclassified patterns according to “Eq. (14)”.
  2) The best ants are selected.
  3) The pheromone trails are updated by “Eq. (15)”: evaporation for all values, and deposition just for the values of best ants.
  4) A new population is generated according to pheromone trails by the selection rule as “Eq. (13)”.
  5) A new projection becomes the current best projection when the error is reduced.
- The best projection found so far, is chosen as output.

Note that our approach classifies patterns by proximity to means in the projected space. Since this classification rule is not necessarily optimal, after performing AntEDA, a neural network must be used to take full advantage of the projected features. So after performing the proposed MFCNN instance selection algorithm and the AntEDA feature selection approach, MLP and NF networks are used for classification on achieved TS with useful features. The overall flowchart of proposed hybrid approach can be seen in Fig. 4.

IV. EXPERIMENTAL RESULTS

To further verify the effectiveness of our method, 13 UCI classification data sets were used as listed in Table I. To validate the proposed approach, we applied our method on MLP and NFN. The training results by the randomly selected TS using all original features and instances (we called it RND), and the selection via proposed approach (HYBRID), and also MFCNN are reported in Table. II. Furthermore the number of features and instances extracted from each datasets (represented by F and I respectively) is included. pe.

As shown in this table, the performance of the Hybrid and MFCNN methods is always better than conventional random selection method. Furthermore, because of selection of effective features in Hybrid method, this method guarantees the same and even better classification accuracy than MFCNN approach. The results approve the power of the presented Hybrid method for training artificial intelligent networks with the most effective features and instances.

<table>
<thead>
<tr>
<th>Database</th>
<th>RND</th>
<th>MFCNN</th>
<th>HYBRID</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>NFN</td>
<td>MLP</td>
<td>NFN</td>
</tr>
<tr>
<td>Heart</td>
<td>90.2</td>
<td>83.25</td>
<td>95.34</td>
</tr>
<tr>
<td>Letter</td>
<td>98.36</td>
<td>98.12</td>
<td>95.99</td>
</tr>
<tr>
<td>Breast</td>
<td>97.21</td>
<td>96.43</td>
<td>95.99</td>
</tr>
<tr>
<td>Twonorm</td>
<td>95.78</td>
<td>93.84</td>
<td>98.31</td>
</tr>
<tr>
<td>Sonar</td>
<td>77.12</td>
<td>74.43</td>
<td>80.21</td>
</tr>
<tr>
<td>Nursery</td>
<td>96.32</td>
<td>97.92</td>
<td>98.45</td>
</tr>
<tr>
<td>Pima</td>
<td>73.14</td>
<td>71.36</td>
<td>84.72</td>
</tr>
<tr>
<td>Spambase</td>
<td>90.41</td>
<td>89.52</td>
<td>94.87</td>
</tr>
<tr>
<td>Magic</td>
<td>84.74</td>
<td>81.68</td>
<td>91.35</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>88.73</td>
<td>87.01</td>
<td>88.99</td>
</tr>
<tr>
<td>Bupa</td>
<td>60.68</td>
<td>73.22</td>
<td>77.74</td>
</tr>
<tr>
<td>Chess</td>
<td>92.39</td>
<td>91.88</td>
<td>94.97</td>
</tr>
</tbody>
</table>

As shown in this table, the performance of the Hybrid and MFCNN methods is always better than conventional random selection method. Furthermore, because of selection of effective features in Hybrid method, this method guarantees the same and even better classification accuracy than MFCNN approach. The results approve the power of the presented Hybrid method for training artificial intelligent networks with the most effective features and instances.
V. CONCLUSION

In this paper a new Hybrid approach based on instance selection and feature selection, was proposed for improving the performance of training ANNs. In FDA, two different criteria should be optimized: the Fisher’s measure, and the classification error in the projected space. But in proposed AntEDA, the error is minimized directly without using any intermediate criteria by aid of ACO. The proposed Hybrid algorithm efficiently reduced the number of features and instances. As a result, it achieves better results than random selection method and even MFCNN approach that uses all of instances. As a result, it achieves better results than random selection algorithm will be applied for very large datasets.

REFERENCES


M. Shokouhifar was born in Dezful, Iran, in 1987. He received the MS degree in electronic engineering from Islamic Azad University of Central Tehran Branch, Tehran, Iran, in 2011. He currently is a MS student in electronic engineering in the field of analog circuit design, Shahid Beheshti University, Tehran, Iran. His research interests include: pattern recognition, image processing, fuzzy sets theory, analog and digital circuit design, wireless sensor networks, feature selection, and solving NP-complete problems by using evolutionary algorithms. He has more than 20 international journals and conference papers in the field of evolutionary algorithms. In 2010, he received the award of superior researcher from the Islamic Azad University of Central Tehran Branch. He is a student member of IEEE.