Chapter-7

AN EVOLUTIONARY FEATURE SELECTION TECHNIQUE USING POLYNOMIAL NEURAL NET
This chapter presents a synergic approach of Polynomial Neural Network and Genetic Algorithm for classification tasks. The conventional PNN approaches, published in literature so far, the processing by PNN takes large computation time due to the expansion of the whole network at different levels and Genetic Algorithms are stochastic optimizers. In this synergic approach between PNN and GA, the feature subset selection is being done by the Polynomial Neural Network (PNN) using Genetic Algorithm (GA). A randomly selected subset of features of a dataset is passed to the PNN as input. The classification accuracy of PNN is taken as the fitness function of GA. In the proposed scheme, less number of features selected by the GA prevents PNN to grow at very early stages which reduces the computation cost as well as time. The proposed scheme is simulated on six benchmark databases and classification accuracies of PNN classifiers are compared with three other approaches. It is found that the classification accuracies obtained by proposed scheme is better in some cases and almost same in other cases. The strength of proposed scheme is (i) its high classification accuracy with much less computational cost and (ii) much less execution time taken by it as compared to other schemes.

7.1 Introduction

It is known that a set of features in a dataset that characterize a data point (class in most cases) for a classification problem may not be equally important; some features can be derogatory and may even have harmful effect on the task. Feature selection technique's aim is to discard the bad and irrelevant features from the available set of features. This reduction of features may improve the performance of classification, function approximation, and other pattern recognition systems in terms of speed, accuracy, and simplicity [48]. Another importance of feature
selection is in the task of mining large databases which is also known as
dimensionality reduction [186]. With an increase in dimensionality, the
hyper-volume increases exponentially and thus large dimensionality of
data demands a large number of training samples for an adequate
representation of the input space. Dimensionality reduction can be done
by selecting a small but important subset of features and generating
(extracting) a lower dimensional data preserving the distinguishing
characteristics of the original higher dimensional data [152]. Feature
Selection leads to savings measurement cost and time because some of the
features (redundant) get discarded. This concept can be utilized in pruning
those networks which otherwise would have taken a large time to compute
in presence of all features. In practice it is often found that additional
features actually degrade the performance of a classifier designed using
class-conditional density estimates when the training set is small with
respect to the dimensionality [167], [168]. This usually happens because
classifiers estimate the class-conditional densities from the available
training data. Thus, if the dimensionality is increased keeping the size of
the training set fixed, the number of unknown parameters automatically
increases and the reliability of the estimate decreases. As a consequence,
the performance of the classifier, constructed from a fixed number of
training instances, may degrade with the increase in dimensionality as was
illustrated by Trunk [210]. When feature selection methods use class
information, we call it supervised feature selection otherwise it is an
unsupervised feature selection [141] [44]. Feature selection is also
important in other areas such as finding cluster structures in data or in
other exploratory data analysis. Polynomial Neural Networks (PNN) [112]
have emerged recently as an extension of Artificial Neural Networks
(ANN) [185]. Some of the limitations of ANN have been claimed to be
countered in PNN. The PNN have also been used as classifiers like ANN.
The classification time depends on the number of features and the size of PNN, i.e. its architecture. The latter has been addressed in literature [136][137]. The PNN based classification using former approach has not been applied so far to our perception. In this paper we investigate a scheme to evaluate the performance of PNN classifier using reduced number of feature.

7.2 Review of the work done

The problem of feature selection has been well addressed in literature [48] and it has been tried out in various paradigms. Previous studies on feature subset selection focused mainly around statistical approaches like principal component analysis (PCA) [86], linear discriminant analysis (LDA) [62], etc. These methods attempt to reduce the dimensionality of the feature space by creating new features which are combination of the original ones; known as feature extraction techniques. The main drawback of these methods is that the newly created features lose their original identity. Blum and Langley [25] have given an excellent survey on feature selection in machine learning. These approaches are different in evaluation of the feature subsets. There are many feature selection algorithms that use soft computing/computational intelligence tools. GA are used in [28] and [169] to select the relevant feature subsets. Neural Networks are used in [20], [56], [179], [189], [204] and [237] for feature selection. Feature selection has also been attempted using fuzzy and neurofuzzy techniques [47], [170]. There are also specialized methods to deal with feature selection for very large dimensional data sets that are typical in application areas such as bioinformatics [5], [123]. In [126], [127], [128], MacKay has considered neural network learning in a Bayesian framework. MacKay and Neal proposed a feature selection mechanism in the Bayesian learning framework called automatic
relevance detection (ARD) [145]. In the ARD model, each input variable is associated with a hyper parameter that controls the magnitude of the weights of connections out of that input unit. The significance of an input variable is determined according to the posterior distributions of these hyper parameters. In [153], Pal and Chintalpudi developed an integrated feature selection and classification scheme based on the multilayer perceptron (MLP) architecture. The feature selection phase in their method was integrated with the main learning task, and the MLP learned certain feature modulators along with the conventional weights and biases of a neural network. In [35], a neurofuzzy system was developed for simultaneous feature selection and system identification. The methodology developed in [35] was modified for a classifier in [36]. The feature selection methods described in [35], [36] and [153] are termed as online methods. In an online method, the feature selection phase is integrated with the task of learning other parameters of the system. The ARD [145] and its variant [205] also learn hyper parameters associated with the input features. Some evolutionary techniques [231] also learn the importance of input features along with other parameters of the classifier. For the classification problem, LIKNON [23] uses a linear programming formulation to learn feature weights along with other parameters of the separating hyperplane.

In [201], authors report on the plausible solution for ascertaining the composition of gases, during complex boiler flue gas data analysis, by taking a number classification problem as a model. For this purpose an indigenously developed arithmetic residue (AR) scheme has been devised as a feature selection technique. For the purpose of classification of data (number of classes of gases), a probabilistic neural network has been implemented and its classification capability has been
analyzed first for the data acquired from ORSAT analyzer and then for the data from KANE® analyzer.

Yumin et al. [233] recently proposes an Ant Colony Optimization based approach for feature selection using rough sets. Saxena et al. have used GA algorithms with Sammon's function for unsupervised feature selection [186].

It will also be worth addressing PNN and some of its applications in feature selection or classification. Classification is one of the core challenging tasks [136] in data mining [88], [132] pattern recognition [238], web mining [106], bioinformatics [15], and financial forecasting [26], [49]. The goal of classification [140] is to assign a new entity into a class from a pre-specified set of classes. One of the popular and widely used techniques is the method of feed-forward neural network (FNN) [211]. Although such FNN can classify a wide range of dataset properly/accurately yet the classification model cannot be comprehensible due to a large number of synaptic connections including weights and biases. In order to achieve high classification accuracy in FNN framework, one has to provide a well defined structure of FNN such as the number of input nodes, hidden and output neurons, and assumption of a proper set of relevant features. In this regard the trial and error method has been used to arrive at such kind of structures which are computationally very expensive. Similarly there are also other methods like rule extraction and decision tree [165], [166], which provides comprehensible rules that are based on the trade-off between the complexity and the classification accuracy of the rules. Misra et al. [136] presents a Particle Swarm Optimization (PSO) approaches to reduce optimized the architecture of PNN for classification task.
7.3 Proposed Scheme

Evolutionary approach based PNN generates populations or layers of neurons/simulated units/partial descriptions (PDs) and then trains and selects those neurons. It is known that the complexity of PNN model is directly proportionate to the feature set (for details see Chapter-5). Where feature set represents the set of all features in the dataset. A subset of this feature set is selected randomly which becomes a part of population (one chromosome) to be used in GA. The existence of a feature in the subset is represented by a 1 and its absence by a 0 in every chromosome of the population. This subset of features is given as input to PNN.

The classification accuracy of PNN is calculated using training and testing patterns from the dataset. This accuracy serves as the fitness value of the GA. Similarly the fitness of other subsets of features in the initial population is also calculated. The simple one point crossover and mutation operations are applied on initial population to produce a modified population. The fitness’ of the chromosomes of the modified population are compared with those of the initial population. The better chromosomes (Subsets) are retained in the next population. This completes one generation and the population with chromosomes of higher fitness values becomes the initial population for the next generation. This process continuous for a number of generations and at a satisfactory level the process is stopped. The feature subset with best classification accuracy of PNN is noted for comparison. Figure-7.1 describes how a PD is computed at a node of a PNN’s layer with reduced features using proposed scheme and the pictorial presentation of the flow of proposed scheme is given in Figure-5.7.
Figure 7.1: Computation of PD at a node of a PNN’s layer with some inputs treated absent and indicated by 0 using proposed scheme.

7.4 Simulation Studies

The performance of proposed scheme is evaluated using the benchmark databases. A summary of these databases is given in Table-7.1 which is also available in the UCI machine repository [24]. Proposed scheme was simulated on a Pentium-III machine. For computing classification accuracy of PNN, cross validation was used. Each dataset was divided into two folds, one for training and other for testing. We have taken 50% patterns in fold 1 and remaining 50% in fold 2. The number of generations and sizes of populations used in the proposed scheme for different datasets are shown in Table-7.2. Table-7.3 presents the Times of execution, classification accuracies for different datasets using PNN, RCPNN with Gradient Descent, RCPNN with PSO and Proposed Scheme. Figure-7.2 shows a comparative bar chart for proposed scheme versus PNN, RCPNN with Gradient Descent, RCPNN with PSO schemes with respect to classification accuracies and Figure-7.3 shows a comparative line chart for proposed scheme versus RCPNN with Gradient Descent, RCPNN with PSO schemes with respect to time.
Table-7.1
Description of the Data Set Used

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Total</th>
<th>Attributes</th>
<th>Classes</th>
<th>Patterns in Class1</th>
<th>Patterns in Class2</th>
<th>Patterns in Class3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
<td>59</td>
<td>71</td>
<td>48</td>
</tr>
<tr>
<td>Pima</td>
<td>768</td>
<td>8</td>
<td>2</td>
<td>268</td>
<td>500</td>
<td>-</td>
</tr>
<tr>
<td>Bupa</td>
<td>345</td>
<td>6</td>
<td>2</td>
<td>145</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td>WBC</td>
<td>699</td>
<td>9</td>
<td>2</td>
<td>458</td>
<td>241</td>
<td>-</td>
</tr>
<tr>
<td>Thyroid</td>
<td>215</td>
<td>5</td>
<td>3</td>
<td>150</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>

Table-7.2
Parameters and population size used in GA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Population</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Wine</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>Pima</td>
<td>60</td>
<td>16</td>
</tr>
<tr>
<td>Bupa</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>WBC</td>
<td>60</td>
<td>18</td>
</tr>
<tr>
<td>Thyroid</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

Probability of Crossover = 0.5 and Probability of Mutation = 0.3 for all data sets
Table 7.3

A Comparison of Times of execution and respective Classification Accuracies obtained through proposed scheme with PNN, RCPNN with Gradient Descent and RCPNN with PSO.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Feature</th>
<th>Time (In Seconds)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P1*</td>
<td>P2*</td>
<td>P3*</td>
</tr>
<tr>
<td>Iris</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Wine</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Pima</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Bupa</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>WBC</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Thyroid</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*P1 - PNN
*P2 - RCPNN with Gradient Descent
*P3 - RCPNN with PSO
*P4 - Proposed Scheme

7.5 Results and Discussions

Table 7.3 presents a comparison between the performances evaluated using proposed scheme and those evaluated using other three schemes viz. PNN, RCPNN with Gradient Descent and RCPNN with PSO published in literature [136], [137]. The classification accuracy and the execution cost have been taken as the performance indexes. In other two schemes, efforts are made to reduce the architecture of the PNN. However, number of features has not been reduced in any of the three schemes. By observing Table 7.3, it is noted that in proposed scheme, a very less number of features are capable to produce a higher classification accuracy e.g. in iris data set only 2 features are required to produce 99.11% accuracy against other schemes which take all (4) features but produce a
lesser accuracy (86.22, 95.56, 98.67). Similar results are observed for Bupa, WBC, and Thyroid datasets. For Wine and Pima datasets the classification accuracies using proposed scheme are obtained fractionally smaller than that with other schemes but important to note that our scheme takes less number of features to acquire that much accuracy. This small difference in accuracy can be accepted looking to overall figures and number of features. In addition to classification accuracy; execution time is another parameter which has been compared with other three schemes. In proposed scheme, time of execution for iris dataset is 0.193 sec versus 54, 0.428, 0.830 sec respectively in other three schemes which is a noticeable difference. In each dataset, proposed scheme takes much less time compared to that in published schemes.

Figure-7.2 : Comparison of Classification Accuracies between proposed scheme and other schemes
Figure 5.32: FAB⁺-Mass Spectrum of Compound 9
Figure 7.3: Comparison of execution time between proposed and other schemes

*P1 being very large has not been shown in this figure.
Figure 5.33: $^1$H-NMR Spectrum of Compound 9