Stacked Regression Ensemble for Cancer Class Prediction

Muhammad Shoai B. Sehgal, Iqbal Gondal, Member, IEEE and Laurence Dooley, Senior Member, IEEE

Faculty of IT, Monash University, Churchill VIC 3842, Australia
e-mail: (Shoaiib.Sehgal ,Iqbal.Gondal, Laurence.Dooley) @infotech.monash.edu.au

Abstract—Design of a machine learning algorithm as a robust class predictor for various DNA microarray datasets is a challenging task, as the number of samples are very small as compared to the thousands of genes (feature set). For such datasets, a class prediction model could be very successful in classifying one type of dataset but may fail to perform in a similar fashion for other datasets. This paper presents a Stacked Regression Ensemble (SRE) model for cancer class prediction. Results indicate that SRE has provided performance stability for various microarray datasets. The performance of SRE has been cross validated using the k-fold cross validation method (leave one out) technique for BRCA1, BRCA2 and Sporadic classes for ovarian and breast cancer microarray datasets. The paper also presents comparative results of SRE with most commonly used SVM and GRNN. Empirical results confirmed that SRE has demonstrated better performance stability as compared to SVM and GRNN for the classification of assorted cancer data.

Index Terms—Support Vector Machines, Neural Networks, Stacked Generalization, Decision Based Fusion and Classifier Ensemble.

I. INTRODUCTION

Machine learning algorithms have been successfully applied to many bioinformatics applications [1] including diagnosis of breast [15], ovarian [1, 15], leukemia [16], lymphoma [17], brain cancer [18] and lung cancer [19]. Despite this success, there is still an awareness of the need for robust classification algorithms which exhibit performance stability for multiple types of data. Some classifiers have better classification results for one type of data, but fail to perform in a similar way for another data set. The reason of such variability in performance is due to lack of separability of data and much more number of features (thousands of genes) as compared to the number of samples in microarray data. This problem can be addressed if different classifier types are integrated to form an ensemble to identify a particular class [26, 27 and 28] so that best performance of the classifiers can be combined.

There are several techniques to combine multiple classifiers for example, sum, product, minimum, maximum, median and majority voting [21, 22, 23, 24 and 25]. The advantage of these techniques is that they are simple and do not require training [31], though this is counterevalied by the fact that they are not adaptive to the input data. Stacked generalization however, has emerged as a way to ensemble classifiers adaptively and requires further training for fusing the decision of different classifiers [8, 30]. This classifier fusion however requires careful selection of base classifiers [7] to achieve acceptable misclassification rates.

This paper presents Stacked Regression Ensemble (SRE) which used SVM with different kernels including Linear, Polynomial and Radial Basis Function (RBF) and Generalized Regression Neural Network (GRNN) as base classifiers. The motivation to use SVM and GRNN was their promising results in a variety of biological classification tasks, including gene expression microarrays. SVM has demonstrated better classification accuracy while classifying variety of microarray data than other commonly used classifiers [3, 14 and 32]. However, our research has demonstrated that if Generalized Regression Neural Networks (GRNN) is modified to add a distance layer it can perform better than SVM in the classification of genetic data [15]. Our experimental results will confirm that our proposed fusion models will perform consistently better than single classifiers including SVM and GRNN.

The well known ovarian cancer microarray data by Amir et al [36] and breast cancer data by Hedenfalk et al [37] is used for comparative purposes. The reason to address this classification problem present in the above data sets is due to the fact that cancer is one of the most appalling diseases for researchers due to its diagnostic difficulty and devastating effects on human kind. Some types of cancer are more dangerous than others for example, breast cancer is the second leading cause of cancer deaths in women today (after lung cancer) and is the most common cancer among women, excluding non-melanoma skin cancer and ovarian cancer [34]. Similarly ovarian cancer is the fourth most common cause of cancer related deaths in American women of all ages as well as being the most prevalent cause of death from gynaecologic malignancies in the United States [35]. Mutations in BRCA1, BRCA2 and Sporadic (without BRCA1 and BRCA2 mutation) are responsible for hereditary breast and ovarian cancer that can lead to carcinogenesis through different molecular pathways [36], so disease pathway mapping is helpful for the treatment of this disease in efficient way. If there is a family history of a particular gene mutation then this implies a possible mutation in the descendents, so we can combine the knowledge of gene mutation and family history for a more accurate and timely identification of ovarian and breast cancer.

This paper is organized as follows: section 2 outlines different class prediction algorithm used as base classifiers and for comparative purposes. Section 3 presents our proposed SRE model and methodology. Results and discussion on our experimental results are provided in section 4. Finally, section 5 concludes the paper.
II. CLASS PREDICTION MODELS

This section briefly reviews GRNN and SVM classification techniques that will be used in evaluating and building our proposed SRE model. A detailed review of these methods is provided by [2, 4 and 14].

A. Generalized Regression Neural Network

Generalized regression neural networks are paradigms of the RBF used in functional approximation [20, 21]. To apply GRNN to classification, an input vector \( x \) (ovarian or breast cancer microarray data) is formed and weight vectors \( W \) are calculated using (2). The output \( y(x) \) is the weighted average of the target values \( t_i \) of training cases \( x_i \) close to a given input case \( x \), as given by:

\[
y(x) = \frac{\sum_{i=1}^{n} t_i W_i}{\sum_{i=1}^{n} W_i}
\]

(1)

where \( W_i = \exp \left[ -\frac{1}{2h^2} \right] \)

The only weights that need to be learned are the smoothing parameters, \( h \) of the RBF units in (2), which are set using a simple grid search method. The distance between the computed value \( y(x) \) and each value in the set of target values \( T \) is given by:

\[
T = \{1, 2\}
\]

(2)

The values 1 and 2 correspond to the training class and all other classes respectively. The class corresponding to the target value with the minimum distance is chosen. As, GRNN exhibits a strong bias towards the target value nearest to the mean value \( \mu \) of \( T \) [14] so we used target values 1 and 2 because both have the same absolute distance from \( \mu \).

B. Support Vector Machine

Support vector machines are systems based on regularization techniques which perform well in many classification problems [17, 22]. SVM converts Euclidean input vector space \( \mathbb{R}^n \) to higher dimensional space [18, 19] and attempts to insert a separating hyperplane [23, 24 and 25]. The gene expression data \( Z \) is transformed into higher dimensional space. The separating hyperplane in higher dimension space satisfies:

\[
W.Z_i + b = 0
\]

(3)

To maximize the margin between genetic mutation classes (5) and (6) are used.

\[
\text{Max} \frac{1}{\|W\|^2}
\]

(4)

Subject to the condition

\[
y_i (W.Z_i + b) \geq 1
\]

(5)

Using the Kuhn-Tucker condition [26] and LaGrange Multiplier methods in (7) is equivalent to solving the dual problem

\[
\text{Max} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right]
\]

(6)

where \( 0 \leq \alpha_i \leq C, ) \) = number of inputs, \( i = 1, \ldots, l \) and \( \sum_{i=1}^{n} \alpha_i y_i = 0 \). In (9) \( K(x_i, x_j) \) is the kernel function and in our experiments, a linear, polynomial (of degree 2 to 20) and RBF functions were used C is the only adjustable parameter in the above calculations and a grid search method is used to select the most appropriate value of the C.

III. STACKED REGRESSION ENSEMBLE

Classifier ensemble is used to increase the posteriori prediction performance. In the past several methods have been proposed to ensemble classifiers like Minimum, Maximum, Median, Majority Vote and Stacked Generalization. Comprehensive details for classifier fusion are provided by [7, 8 and 9]. However, due to adaptive behavior of stacked generalization it has proven to be a better technique than the techniques given in [7].

Fig. 1 demonstrates the working of SRE. For a given data \( L_2 = \{(y_n, x_n)\}_{n=1,2,\ldots,N} \) where \( y_n \) is the class label, \( x_n \) is a data sample and \( N \) is the total number of samples available in the data set. The input data \( L_2 \) is first divided equally in \( k \) folds \( k_1, \ldots, k_m \). The individual classifiers, (SVM with linear, polynomial and RBF kernel and GRNN) referred to as base classifiers or level-0 generalizer [10] are then cross validated by removing each fold randomly from \( k \) folds in such a way that selection probability \( P_s \) for a particular fold to be used as validation data \( T \) is :-

\[
P_s = \frac{n}{N \times L}
\]

(7)

and the probability \( P_t \) to become part of training data \( T = L_0 - T \) is :-

\[
P_t = \frac{(k-1) \times n}{N \times L}
\]

(8)

where \( N = \) total data items per class, \( k = \) number of folds, \( L = \) number of classes and \( n = \) number of samples in each fold. For each \( P^0 \) iteration class prediction posteriori prob-
ability \( P_p \) of class \( \omega_i \) for the given data \( x \) can be calculated as
\[
P_{p|x} P(\omega_i | x)
\] (9)

After entire cross validation process \( k \) prediction labels based on prediction probabilities \( P_p \) are assembled as \( L_k = \{P_{y_n|P_m}\} \) to get level-I data. Then, GRNN, level-I generalizer [11] is trained based on \( L_k \) and the output is considered to be final output of SRE [12, 13] (See Fig. 1).

![Stacked Regression Ensemble](image)

**Fig. 1: Stacked Regression Ensemble**

**IV. RESULTS AND DISCUSSION**

To compare the different classification models, ovarian cancer data by Amir et al [36] and breast cancer data by Hedenfalk et al [37] is used for the experimental purpose. The ovarian cancer data set contains 18, 16 and 27 samples of BRCA1 mutations, BRCA2 and sporadic mutations (neither BRCA1 nor BRCA2) respectively. Each data sample contains logarithmic microarray data of 6445 genes. There are 21 samples for the breast cancer data that contains 7, 7 and 8 samples of BRCA1 mutations, BRCA2 and sporadic mutations (neither BRCA1 nor BRCA2) respectively and each data sample contains microarray data of 3226 genes. The microarray data is asymmetric, so log2 of the input data is used to make data symmetric. This logarithmic data is then used as the input to the classifier to identify the mutation of genetic data.

Results in Table 1 show that our proposed SRE model has demonstrated higher classification accuracy as compared to single models, including SVM (with Linear, Polynomial and RBF kernel) and GRNN for the class prediction of BRCA1 and Sporadic mutations in ovarian cancer. The reason of these better results is the adaptive behavior of GRNN as level-I generalizer. For example, SVM with RBF kernel performed better than rest of single classifiers and GRNN adapted to give it more weight while classifying BRCA2 mutations as compared to the rest of base classifiers. The other reason is that the ensemble can only be accurate than its base classifiers when base classifiers disagree with each other [33] and in this case the classifiers in ensembles performance is better than the best accuracies of individual base classifiers.

**Table 1: Classification accuracies of ovarian cancer (B1 = BRCA1, B2 = BRCA2 and S = Sporadic)**

<table>
<thead>
<tr>
<th>Classification Models</th>
<th>B1</th>
<th>B2</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>94</td>
<td>84</td>
<td>91</td>
</tr>
<tr>
<td>Individual Classifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRNN</td>
<td>94</td>
<td>75</td>
<td>81</td>
</tr>
<tr>
<td>SVM-Linear</td>
<td>94</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>SVM-RBF</td>
<td>88</td>
<td>78</td>
<td>75</td>
</tr>
<tr>
<td>SVM-Poly</td>
<td>94</td>
<td>69</td>
<td>81</td>
</tr>
</tbody>
</table>

For example, the classification accuracy of SRE for BRCA2 mutation is better than the best accuracies of the rest of models due to the reason that these models showed different class results while classifying the same sample. Similarly, the better classification results for Sporadic genetic mutation of SRE is not due to class separability rather it is hard to classify as compared to the other mutations (see Fig. 2). The reason for better results by SRE for Sporadic mutation classification is due to the presence of SVM with polynomial kernel and GRNN as base classifiers (whose individual accuracy is maximum for the classification of Sporadic mutations) and their different behavior while classifying the same sample.

**Table 2: Classification accuracy of breast cancer (B1 = BRCA1, B2 = BRCA2 and S = Sporadic)**

<table>
<thead>
<tr>
<th>Classification Models</th>
<th>B1</th>
<th>B2</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>100</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>Individual Classifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRNN</td>
<td>86</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>SVM-Linear</td>
<td>64</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>SVM-RBF</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>SVM-Poly</td>
<td>71</td>
<td>93</td>
<td>86</td>
</tr>
</tbody>
</table>

Table 2 shows that for breast cancer data SRE again outperformed rest of the models for the classification of BRCA1 and Sporadic mutations. However, performance accuracy of SRE for BRCA2 and Sporadic mutation (see table 2) is not
better than the best accuracies of base classifiers due to the reason that all the classifiers misclassified the same sample. So, assembling classifiers using proposed Staked regression technique can perform better than the best accuracies of individual base classifiers if the classifiers differ in decisions otherwise it is never better than the best accuracies of individual classifiers.

V. CONCLUSIONS

This paper has presented staked regression ensemble which has shown consistent performance when used for ovarian and breast cancer microarray data sets as compared to single machine learning algorithm, including SVM and GRNN. Our proposed models have shown improved classification accuracies of 94%, 84% and 91% for ovarian cancer data for BRCA1, BRCA2 and Sporadic mutations respectively for ovarian cancer data. The respective accuracies of SRE for breast cancer data are 100%, 93% and 93% for the aforementioned mutations. The SRE model demonstrated better classification accuracies than the best accuracies of individual base classifiers when they differ in the class prediction results. However, the proposed model has accuracy equal to the best accuracy of base classifier when the base classifiers don’t differ in class prediction results. So, combining classifiers in such a way gives either better accuracy than the best accuracy of base classifiers or at least it provides the same classification performance as can be demonstrated by the best base classifiers. Therefore, the better classification results for multiple datasets demonstrate the ability of our innovative SRE model to classify various types of data more accurately.

VI. REFERENCES


[15] Statistical Neural Networks and Support Vector Machine for the Classification of Genetic Mutations in Ovarian Cancer
[31] F.M Alkoft “Modified product fusion”.