Stacked regression ensemble for cancer class prediction

Conference or Workshop Item

Stacked Regression Ensemble for Cancer Class Prediction

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

Faculty of IT, Monash University, Churchill VIC 3842, Australia
e-mail: (Shoaib.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 countervailed 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.

0-7803-9094-6/05/$20.00 ©2005 IEEE

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_k of training cases x_i close to a given input case x, as given by:-

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

$$\text{where } W_i = \exp\left[\frac{-\|x - x_i\|^2}{2h^2}\right]$$

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:-

$$T = \{1, 2\} \qquad (2)$$

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 μ of T [14] so we used target values 1 and 2 because both have the same absolute distance from μ.

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:-

$$\text{W.Zi} + \text{b} = 0 \qquad (3)$$

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

$$Max \frac{1}{\|W\|^2} \qquad (4)$$

Subject to the condition

$$y_i\ (\text{W.Zi} + \text{b}) \geq 1 \qquad (5)$$

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

$$Max\left[\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l} a_i a_j y_i y_j K(x_i x_j)\right] \qquad (6)$$

where $0 \leq a_i \leq C$, l = number of inputs, i = 1.....l and $\sum_{i=1}^{l} a_i y_i = 0$. 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 $L_0 = \{(y_n, x_n) \forall\ n=1,2,....N\}$ 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_0 is first divided equally in k folds $k_1 ... k_n$. 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_v for a particular fold to be used as validation data $\mathcal{B}$ is :-

$$P_v = \frac{\eta}{N \times L} \qquad (7)$$

and the probability P_t to become part of training data $\acute{\Gamma}$ = L0- $\mathcal{B}$ is :-

$$P_t = \frac{(k-1) \times \eta}{N \times L} \qquad (8)$$

where N = total data items per class, k = number of folds, L = number of classes and η = number of samples in each fold. For each i^{th} iteration class prediction posteriori prob-

ability P_{pi} of class ω, for the given data x can be calculated as

$$P_{pi=}P(\omega_i \mid x) \qquad (9)$$

After entire cross validation process k prediction labels based on prediction probabilities P_p are assembled as $L_1 = \{(y_n, P_{pn}) \forall\ n=1,2,....k\}$ to get level-1 data. Then, GRNN, level-1 generalizer [11] is trained based on L_1 and the output is considered to be final output of SRE [12, 13] (See Fig. 1).

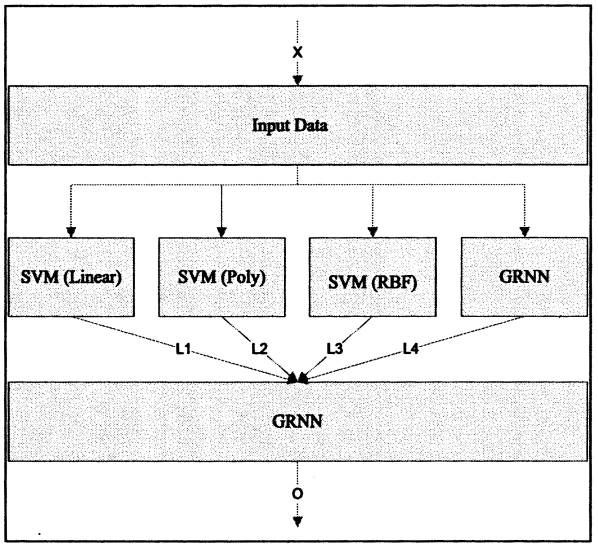

Fig. 1: Stacked Regression Ensemble

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 $\log_2$ 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-1 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)

Classification Models		B1	B2	S
Ensemble	SRE	**94**	**84**	**91**
Individual Classifiers	GRNN	94	75	81
	SVM-Linear	94	66	78
	SVM-RBF	88	78	75
	SVM-Poly	94	69	81

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.

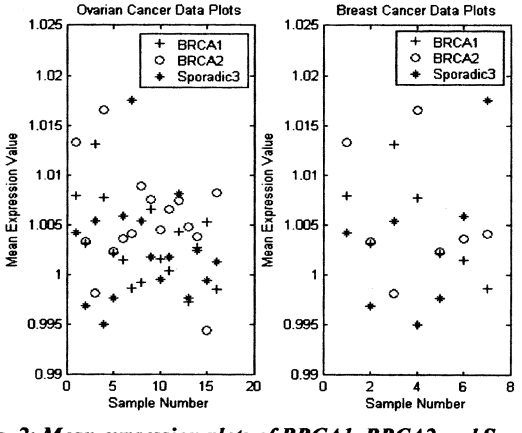

Fig. 2: Mean expression plots of BRCA1, BRCA2 and Sporadic mutations for Ovarian and Breast cancer data

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

Classification Models		B1	B2	S
Ensemble	SRE	**100**	**93**	**93**
Individual Classifiers	GRNN	86	93	93
	SVM-Linear	64	86	86
	SVM-RBF	64	64	64
	SVM-Poly	71	93	86

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 less 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

[1] E. Byvatov and G. Schneider. Support vector machine applications in bioinformatics. Applied Bioinformatics, vol. 2, 67—77, 2003.

[2] V. N. Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998.

[3] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov and T. Poggio. Support vector machine classification of microarray data. Technical Report, A.I Lab. MIT, 2000.

[4] S. Haykin. Neural Networks, Prentice Hall, 1999.

[5] Land, W., Wong, L., McKee, D., Masters, T., and Anderson, F., "Breast Cancer Computer Aided Diagnosis (CAD) Using a Recently Developed SVM/GRNN Oracle Hybrid. 2003 IEEE International Conference on Systems, Man, and Cybernetics" October 2-8, 2003, Washington, DC.

[6] W.S. Sarle, "Stopped training and other remedies for overfitting", Proc. 27th Symposium on the Interface of Computing Science and Statistics, pp. 352-360, 1995.

[7] Ludmila I. Kuncheva, A Theoretical Study on Six Classifier Fusion Strategies, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.24 n.2, p.281-286, February 2002.

[8] S. Dzeroski and B. Zenko. Is combining classifiers with stacking better than selecting the best one? ACM-*Machine Learning,* Volume 54, Issue 3, 2004.

[9] Wolpert, D.H., 1992. Stacked generalisation. Neural Networks 5 (2), 241–260.

[10] Ting, K.M. and I.H. Witten. Stacked Generalization: When Does It Work? , Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, 1997, pp. 866-871.

[11] Breiman, L. (1993). Stacked regression. Technical report, Department of Statistics, University of California at Berkeley.

[12] M. LeBlanc and R. Tibshirani, Combining estimates in regression and classification, J. Amer. Statist. Assoc.91 (1996), 1641-1650.

[13] D. H. Wolpert. Stacked generalization. Technical Report LA-UR-90-3460, Los Alamos, NM, 1990.

[14] C. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1996.

[15] Statistical Neural Networks and Support Vector Machine for the Classification of Genetic Mutations in Ovarian Cancer

[16] T. R Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439):531—537, 1999.

[17] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C.Aster and T. R. Golub. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nat Med, 8(1):68—74, 2002.

[18] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E.McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 415(24):436—442, 2002.

[19] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. F. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA 98:13790–13795, 2001.

[20] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P.Mesirov, T. Poggio, W. Gerald, M. Loda., E. S. Lander and T. R. Golub, Multiclass cancer diagnosis using tumour gene expression signatures, Proc. Natl. Acad. Sci., USA, 98(26):15149—15154, 2001.

[21] Fairhurst, M.C., Rahman, A.F.R., 1997. Generalised approach to the recognition of structurally similar handwritten characters using multiple expert classifiers. IEE Proc. Vision, Image Signal Process. 144 (1), 15–22.

[22] Breiman, L., 1996. Bagging predictors. Machine Learn. 24, 123–140.

[23] Dietterich, T., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. J. Artificial Intell. Res., 263–286.

[24] Ho, T.K., Hull, J.J., Srihari, J.N., 1994. Decision combination in multiple classifier systems. IEEE Trans. Pattern Anal. Machine Intell. 16 (1), 66–75.

[25] M.I Jordan, R.A. Jacobs, Hierarchical mixture of experts and the em algorithm. Neural Comput. 6, 181–214, 1994.

[26] Y. Freund, R. Schapire, Experiments with a new boosting algorithm. In: 13th Internat. Conf. on Machine Learn., pp. 148–156, 1996.

[27] Jacobs, R.A., 1995. Methods for combining experts' probability assessments. Neural Comput. 7, 867–888.

[28] Kittler, J., 1998. Combining classifiers: a theoretical framework. Pattern Anal. Appl. 1, 18–27.

[29] Xu, L., Krzyzak, A., Suen, C.Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Systems Man Cybernet. 22 (3), 418–435.

[30] Woods, K., Kegelmeyer, W.P., Bowyer, K., 1997. Combination of multiple experts using local accuracy estimates. IEEE Trans. Pattern Anal. Machine Intell. 19, 405–410.

[31] F.M.Alkot "Modified product fusion"

[32] M .P. S Brown., W. N. Grundy, D. Lin, N. Cristianini., C. Sugnet., T. S. Furey, M. Ares and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. Proc. Natl. Acad. Sci., 262—267, 1997.

[33] I. Inza, P. Larranaga, B. Sierra and M. Nino, "Combination of classifiers, A case study on oncology", Technical Report EHU-KZZA-IK-1-98, 1998.

[34] J. Laurier, WHO report: alarming increase in cancer rates, 2003.

[35] T. S. Furey, N. Cristianini, N.Duffy, D. W.Bednarski, M. Schummer, D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, vol. 16(10):906—914, 2000.

[36] A. J. Amir , C. J. Yee, C. Sotiriou, K. R. Brantley, J. Boyd, E. T. Liu. Gene expression profiles of brca1-linked, brca2-linked, and sporadic ovarian cancers. Journal of the National Cancer Institute, vol. 94 (13), 2002.

[37] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon and P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, J. Trent. Gene-expression profiles in hereditary breast cancer, N. Engl. J. Med., 22; 344(8):539—548, 2001.