Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques

Ashfaq Ahmed K and Sultan Aljahdali
College of Computers and Information Technology
Taif University,
Taif, Saudi Arabia
ashfaqme@gmail.com, aljahdali@tu.edu.sa

Nisar Hundewale and Ishthaq Ahmed K
Department of Computer Science
nisar@computer.org, ishthaq@gmail.com

Abstract—**The Concept of classification and learning will suit well to medical applications, especially those that need complex diagnostic measurements. Therefore classification technique can be used for cancer disease prediction. This approach is very much interesting as it is part of a growing demand towards predictive diagnosis. From the available studies it is evident that classification and learning methods can be used effectively to improve the accuracy of predicting a disease and its recurrence. In the present work classification techniques namely Support Vector Machine [SVM] and Random Forest [RF] are used to learn, classify and compare cancer disease data with varying kernels and kernel parameters. Results with Support Vector Machines and Random Forest are compared for different data sets. The results with different kernels are tuned with proper parameters selection. Results are analyzed with confusion matrix.**

Keywords- Support Vector Machine, Random Forest, Radial Basis Function, Sigmoid.

I. INTRODUCTION

Cancer disease prediction involves more than one physician from different specializations. This needs different biomedical markers and multiple clinical factors like the age, general health of the patient, its location, type of cancer, the grade and size of the tumor. Cell based, patient based and population based information all must be carefully considered by the attending medical practitioner to come out with a reasonable prediction. It is not so easy even for the most skilled technician to do. Both physicians and patients need to face same challenges when it comes to the matter of cancer prevention and cancer prediction. Family history, age, diet, weight, habits like smoking, heavy drinking, and exposure ultra violet radiations, radon, asbestos plays a major role in predicting an individual's risk for developing cancer. These conventional clinical, behavioral parameters and environment may not be sufficient to make better predictions. To predict the disease we need some specific molecular details about either the tumor or the patient's genetic status. With the speedy development of the proteomic, genomic and imaging technologies, this molecular scale information about patients or tumors is now can be readily acquired.

Many attempts to predict the cancer disease exist such as decision trees, expert systems, neural networks and genetic algorithms etc. However, little significant work has been performed to optimize the parameters of the svm model and compare the results with one of the other powerful learning techniques.

In [1] different novel algorithms are presented for cancer disease prediction. The paper establishes that the concept of support vector is good for better predictions.

In [2] micro array cancer data sets are used for predicting the cancer disease with random forest and support vector machine. It establishes that these techniques yield better results with smaller number of genes.

In [3] support vector machine are used to predict the different levels of cancer growth. It proposes the optimum size for training sets.

It is evident from the past studies that support vector machine and random forest are the better learning techniques for the cancer disease diagnosis. Also they yield much better results with proper parameters selection and size of training data sets.

Section II A discusses about the Support Vector Machine in detail, Section II B about the Random Forest, Section II C is about the data set used, Section III A discusses about experiment setup and Section III B about the actual experiments and Section III C is on discussions about results obtained.

II. CLASSIFICATION TECHNIQUES USED

A. Support Vector Machine

The technique of empirical data modeling is applicable to many engineering applications. In empirical data modeling an induction process is used to build up a model of the system, from which it can deduce responses of the system which are to be tested or observed. The observational nature data obtained is finite and taken as a sample. This sampling is non-uniform and due to the high dimensional nature of the problem data, the input space will be in a sparse distribution. As a result the problem is wrongly presented.

Neural network approaches have suffered problems with generalization producing models that get over fit with the data.

978-1-4673-0892-2/12/$31.00 ©2012 IEEE

This is a result of the optimization algorithms used for statistical method and parameter selection to select the best model. Other learning techniques like decision trees, expert systems and AII were used to predict. This problem of prediction and prognosis can be better solved with machine learning and classification support vector machine technique which implements classification. Machine Learning is a concept under Artificial Intelligence and it is concerned with the development of techniques and methods which enable the system to learn from the available data. This means the development of algorithms which enable the machine to learn from available data and perform tasks and activities of testing the new data. Machine learning works closely with statistics in many ways. There are many techniques and methodologies developed for machine learning tasks [6]. Support Vector Machine is one of the machine learning techniques. Support Vector Machine (SVM) was first introduced in 1992, by Boser, Guyon, and Vapnik in COLT-92. Support vector machines which are used for classification and regression are a set of related supervised learning methods [6]. These machines belong to a generalized family of linear classifiers. In another terms, Support Vector Machine is a classification and regression prediction tool that implements machine learning concepts to maximize predictive accuracy which avoids over fit to the data. A better learning technique must always avoid over fit of the data.

Support Vector machines are defined as systems that use hypothesis space of a linear function in a bigger dimensional feature space. These systems are trained with a learning algorithm from optimization theory that uses a learning bias taken from the theory of statistical learning. Support vector machine was earlier famous with other communities but now it is playing an important role in machine learning research. This technique also used in many other critical domains. SVM becomes more important while using pixel maps as input; the accuracy of SVM is comparable to neural networks with extended features in a handwriting recognition task [7]. SVM is also being used for many applications, such as face analysis, hand writing analysis, engineering, business, management and many more. SVMs are also being used for pattern classification and regression based applications.

The Support Vector Machines SVM have been developed by Vapnik [8] and are popular due to many challenging features and better empirical performance. The methodology of SVM uses the Structural Risk Minimization (SRM) principle, this is superior [9] to traditional Empirical Risk Minimization (ERM) principle, being used by conventional neural networks. ERM technique tries to minimize the error on the training data but SRM tries to minimize an upper bound on the expected risk. This difference makes SVM to work with a better ability to generalize. This is always the goal of statistical learning. SVMs were developed basically to solve the classification problem, but currently they are also being used to solve regression problems [10].

Fig. 1 shows an over fitting classifier where data is overlapping with training data. *Fig. 2* shows a better classifier where there is almost no overlapping.

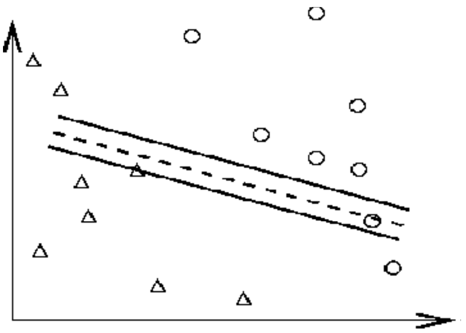


Figure 1. Over Fitting Classifier

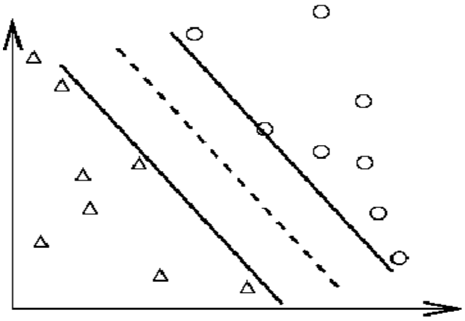


Figure 2. Better Fitting Classifier

Methodology used in SVM:

Classification technique usually separates data into two different data sets one training and the other testing sets. Every instance in the training set contains one target value and several attributes or features. SVM produces a model using the training data and their features which in turn will predicts the target values of the test data given only the attributes of the input test data.

In machine learning, classifying data is the main job. With some given input training data points each belong to one of two classes, the job is to decide which class a new data point will fit without overlapping. In support vector machines, every data point is considered as a p dimensional vector where a vector means a list of p values. We must know whether we can separate such points with a $(p - 1)$ dimensional hyper plane. This approach is called as a linear classifier. There may be several hyper planes that might classify the data. We need to choose one reasonable choice as the best hyper plane. This hyper plane represents the largest separation or margin between the two classes. So we select the hyper plane one for which the distance from it to the nearest data point on each side is maximum. Such a hyper plane is known as the maximum margin hyper plane and the linear classifier it defines is called as a maximum margin classifier or otherwise, the preceptor of optimal stability.

With a training set of instance label pairs of sample training data (x_i, y_i),i = 1…l where x_i ε R^n and y ε $\{1,-1\}^1$, the support vector machines SVM (Boser et al., 1992; Cortes and Vapnik, 1995) applies optimization technique. The function ǿ maps the training vectors x_i into a higher dimensional space.

SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. The function used K (x_i, x_j) = ό $(x_i)^T$ ό (x_j) is called the kernel function. There are many kernels have been proposed but following are four basic kernels that are in use:

The sequence of steps for Classification with SVM includes data preprocessing with categorization of features and proper scaling. The next steps include selection of kernel function and cross validation.

B. *Random Forest*

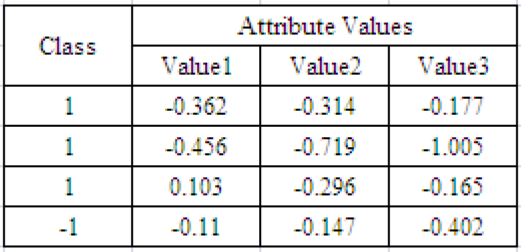
Random Forest is another classification technique; it is a collection of a group of tree predictors. Here each tree depends on the values of a vector independently with the same distribution over all trees in the forest. Error with generalization converges as the number of trees in the forest becomes more. The error associated with generalization of this classifier primarily depends on the strength of the individual trees in the forest and the correlation between the trees. With a random selection of features to split each node results error rates that can be compared. This Random Forest [16] is good even with greater noise. The internal working of this technique make better internal estimates monitor error, strength, and correlation. These are then used to show the response to increasing the number of features used in the splitting. Internal estimates can also be used to find variable importance. These concepts are also applied to regression functionality of the technique.

C. *Datasets Used*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

A duke breast cancer data set is chosen for experiments. Training data set with few records is as shown in the following Table I.

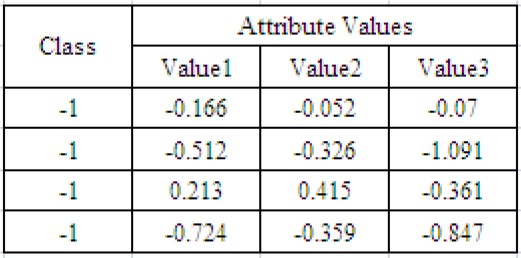
TABLE I. TRAINING DATA



Class	Attribute Values		
	Value1	Value2	Value3
1	-0.362	-0.314	-0.177
1	-0.456	-0.719	-1.005
1	0.103	-0.296	-0.165
-1	-0.11	-0.147	-0.402

Testing data set with four records is as shown in the following Table II.

TABLE II. TESTING DATA



Class	Attribute Values		
	Value1	Value2	Value3
-1	-0.166	-0.052	-0.07
-1	-0.512	-0.326	-1.091
-1	0.213	0.415	-0.361
-1	-0.724	-0.359	-0.847

III. EXPERIMENTAL SETUP AND RESULTS

A duke breast cancer data set is chosen for experiments. This data set is classified with support vector machine and Random Forest. The results are analyzed with a prediction analysis technique called confusion matrix. The results with different parameters are tuned and parameters selections for optimal classification results are automated.

A. *Setups*

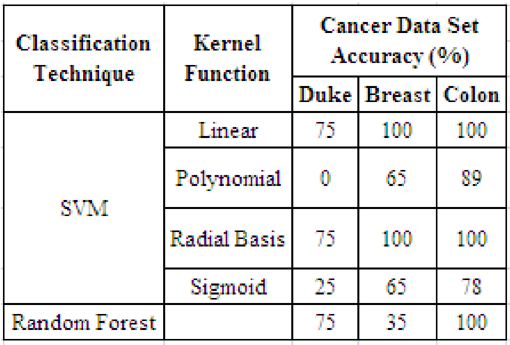
Implementation is done with SVM tool on Mat lab with Microsoft VC++ compiler installed over it. Training data and testing data are formatted into svm tool format using read call then train feature takes formatted data as an input and generates a model of classifier. This model is a statistical model. The varying types of input parameters like kernel functions. Different training models are created using different kernel functions like Linear, Polynomial, RBF and Sigmoid.

Implementation is also done with RF tool on Mat lab with Microsoft VC++ compiler installed over it. Same duke breast cancer data set is used to carry out experiments. Training data and testing data are formatted into random tool format using then trained to generate a model of classifier. This model is a decision tree based model. Given as an input data set for prediction the tool classifies the given test data using the input model. The results of classification are presented as percentage of accuracy.

B. *Results*

The results obtained with both the techniques with breast cancer data set are tabulated shown in Table III.

TABLE III. CLASSIFICATION RESULTS



Classification Technique	Kernel Function	Cancer Data Set Accuracy (%)		
		Duke	Breast	Colon
SVM	Linear	75	100	100
	Polynomial	0	65	89
	Radial Basis	75	100	100
	Sigmoid	25	65	78
Random Forest		75	35	100

Table IV presents the analysis of the results in the form a confusion matrix. Confusion matrix presents test data as versus predicted data as shown in Table 3.2.2**.** This matrix represents how many times true is being predicted as true, true as false, false as true and false as false. This is helpful in analyzing the results obtained with a dataset.

Results obtained for duke data set with SVM Classification and the computed confusion matrix for the above results is

Confusion Matrix (Cm):

TABLE IV. CONFUSION MATRIX

	T	**F**
T	0	0
F	1	3

C. Discussions

Results obtained for different cancer disease data sets with SVM and Random Forest using different kernel functions like linear, polynomial, radial basis and sigmoid are tabulated. It can be observed that there is a varying accuracy of classification with different probabilistic estimate with different kernel function.

Results are observed much better with Radial basis function with SVM and in some cases results are comparable with Random Forest technique.

IV. CONCLUSION AND FUTURE WORK

It is concluded that varying results are obtained with svm classification technique with different kernel functions. Each kernel has its own parameters. A function call is incorporated to tune kernel parameters for best accuracy possible with that kernel. For data sets like duke and colon, random forest technique is yielding results comparable with parameter tuned svm results. The results are better analyzed with confusion matrix. This work can further be extended with other new kernel functions and other classification techniques.

REFERENCES

[1] Chen AH. Exploring novel algorithms for the prediction of cancer classification. Software Engineering and Data Mining(SEDM), 2010 2nd International Conference on Software Engineering and Data Mining, June 2010 pages: 378-383, Tzu-chi Univ., Hualien, Taiwan.

[2] M.Klasssen. Learning Microarray Cancer Datasets by Random Forests and Support Vector Machine. Future Information Technology (FutureTech) 2010, 5th International Conference California University, Thousand Oaks, CA, USA, page(s): 1-6

[3] Furuta K, Aoki Kinoshita, K.F Wai-Ki ching,, Support Vector Machine Methods for the prediction of Cancer growth, 2010, 3rd international joint conference on computational science and optimization(CSO); volume:1; Pages:229-232

[4] Burke HB, Goodman PH, Rosen DB, et al. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79:857-62.

[5] Leenhouts HP. 1999. Radon-induced lung cancer in smokers and nonsmokers: risk implications using a two-mutation carcinogenesis model. *Radiat Environ Biophys*, 1999 38:57-71.

[6] E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144{152. ACM Press, 1992.

[7] C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] Cortes and V. Vapnik. Support-vector network. Machine Learning, 20:273-297, 1995.

[9] E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classi_cation. Journal of Machine Learning Research, 9:1871 1874, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/papers/liblinear. pdf.

[10] L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. Brinkman. PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. Nucleic Acids Research, 31(13):3613-3617, 2003.

[11] S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation, 15(7):1667{1689, 2003}.

[12] T. Lin and C.-J. Lin. A study on sigmoid kernels for SVM and the training of non- PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University, 2003. URL http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.

[13] Michie, D. J. Spiegelhalter, C. C. Taylor, and J. Campbell, editors. Machine learning, neural and statistical classi_cation. Ellis Horwood, Upper Saddle River, NJ, USA, 1994. ISBN 0-13-106360-X. Data available at http://archive.ics. uci.edu/ml/machine-learning-databases/statlog/.

[14] W. S. Sarle. Neural Network FAQ, 1997. URL ftp://ftp.sas.com/pub/neural/ FAQ.html. Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

[15] Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 1995.

[16] Breiman, Leo (2001). "Random F orests". *Machine Learning* 45 (1): 5–32. *doi:10.1023/A:1010933404324.*

[17] Ho, Tin (1995). *"Random Decision Forest"*. 3rd Int'l Conf. on Document Analysis and Recognition. pp. 278–282.