MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 3, Issue 2, July-December 2018

A Comparative Analysis of Cancer Disease Diagnosis Using Machine Learning Algorithms

T. Manikyala rao[1], P. Meena[2], K. Mounika[3], M.Prasanna[4]
Sri Sivani College of Engineering, Chilakapalem
Email Id: *manik.me.au@gmail.com[1]*

Abstract

Breast cancer is the second largest cause of cancer deaths among women. At the same time, it is also among the most curable cancer types if it can be diagnosed early. Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability. In this paper, breast cancer diagnosis is based on a SVM-based method combined with Random Forest Classifier has been proposed. Experiments have been conducted on different training-test partitions of the Wisconsin breast cancer dataset (WBCD), which is commonly used among researchers who use machine learning methods for breast cancer diagnosis. The performance of the method is evaluated using classification accuracy, correlation matrix, scattering matrix. The result shows that the highest classification accuracy (100%) is obtained for the random forest classifier model that contains five features, and this is very promising compared to the previously reported results

Keywords:- *Support vector machine, Random Forest Classifier, Wisconsin breast cancer (WBCD), machine learning, correlation matrix, scattering matrix*

INTRODUCTION

The high incidence of breast cancer in women has increased significantly in the recent years. It the cause of most common cancer death in women

This study aims to build a computer-aid diagnosis system to distinguish benign breast tumor from malignant one. This method involves two stages in a machine learning algorithm random forest is

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

hybridized. The first stage of the whole system conducts a data reduction process for learning algorithm random forest of the second stage. This provides less training data for random forest and so prediction time of the algorithm can be reduced in a great deal. With a selected features set, the explanation of rationale for the system can be more readily realized.

Our proposed method averagely obtained 97.6% and 95.90% classification accuracy in test phase and these results are the highest one among the studies applied for Wisconsin Breast Cancer Diagnosis Dataset (WBC-DD) and Wisconsin Breast Cancer Prognostic Dataset (WBCPD) classification problem so far. It also indicates that the proposed method can be applied confidently to other breast cancer problems with different data sets especially with ones that have a higher number of training data.

RELATED WORK

There has been a lot of research on medical diagnosis of breast cancer Hui-Ling Chen et al. used a rough set (RS) based supporting vector machine classifier (RS_SVM) and the reported accuracy was 96.87% in average. Based on neuro-fuzzy rules Ali Keles et al. presented a decision support system for diagnosis breast cancer. As be reported, the system has high positive predictive rate (96%) and specificity (97%). In A. Marcano et al. proposed a method named AMMLP based on the bio-logical meta-plasticity of neurons and Shan- non's information theory.

PROPOSED METHOD

In a proposed method we are using two algorithms for cancer disease diagnosis namely supports vector machine and random forest classifier.

Support vector machine

A support vector machine (SVM) performs classification by constructing an N dimensional hyper plane that optimally separates the data into two categories. SVM models are closely related to neural networks. In fact, a SVM model using a sigmoid kernel function is equivalent to a two layer, perceptron neural network.

Support vector machine (SVM) models are a close cousin to classical multilayer perceptron neural networks. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifier in which the weights of the network are found by solving a quadratic programming problem with linear

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training.

In the parlance of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyper plane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyper plane that separates clusters of vectors in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyper plane are the support vectors. The figure below presents an overview of the SVM process.

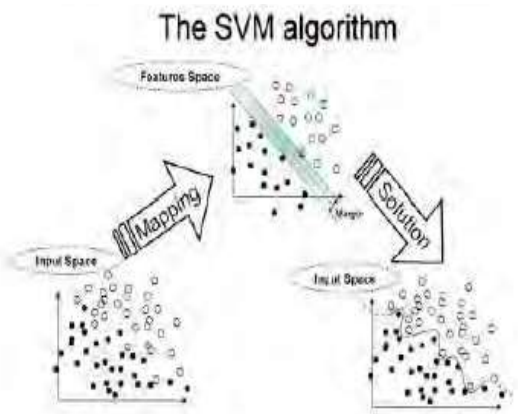
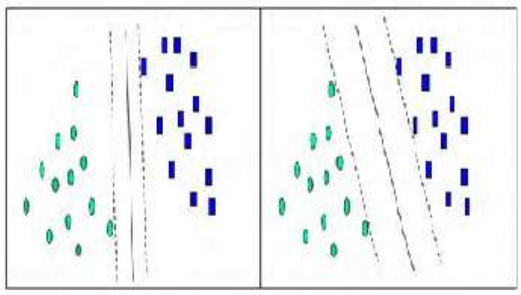


Fig 1: overview of the svm process

A Two-dimensional example

Before considering N-dimensional hyper planes, let's look at a simple 2-dimensional example. Assume we wish to perform a classification, and our data has a categorical target variable with two categories. Also assume that there are two predictor variables with continuous values. If we plot the data pints using the value of one predictor on the X-axis and the other on the Y-axis we might end up with an image such as shown below. One category of the target variable is represented by the rectangles while the other category is represented by ovals.



In this idealized example,, the cases with one category are in the lower left corner and the cases with the other category are in the upper right corner; the cases are completely separated. The SVM analysis attempts to find a 1-dimensional hyper plane (i.e., a line) that separates the cases based on their target categories. There are an infinite number of possible lines; two candidate lines are shown above. The

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

question is which is better, and how do we define the optimal line.

The dashed lines drawn parallel to the separating line mark the distance between the dividing line and the closet vectors to the line. The distance between the dashed line is called the margin. The vectors (points) that constrain the width of the margin are the support vectors. The following figure illustrates this.

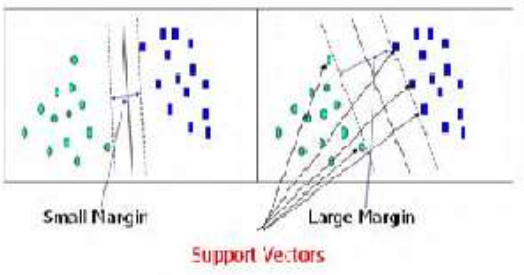


Fig 2: Margin representation

SVM analysis finds the line (or, in general, hyper plane) that is oriented so that the margin between the support vectors is maximized. In the figure above, the line in the right panel is superior to the line in the left panel.

If all analysis consisted of two-category target variables with two predictor variables, and the cluster of points could be divided by a straight line, life would be easy. Unfortunately, this is not generally the case, so SVM must deal with (a) more than two predictor variables, (b) separating the points with non-linear curves, (c)handling the cases where clusters cannot be completely separated and (d) handling classifications with more than two categories.

Random Forest

The random forest (RF) algorithms form a family of classification methods that relay on the combination of several decision tress (Figure 1). The particularly of such Ensembles of Classifiers (EoC) is that their tree-based components are grown from a certain amount of randomness. Based on this idea, RF is defined as a generic principle of randomized ensembles of decision trees. The basic unit of RF (the so-called base learner) is a binary tree constructed using recursive partitioning (RPART). The RF tree base learner is typically grown using the methodology of CART, a method in which binary splits recursively partition the tree

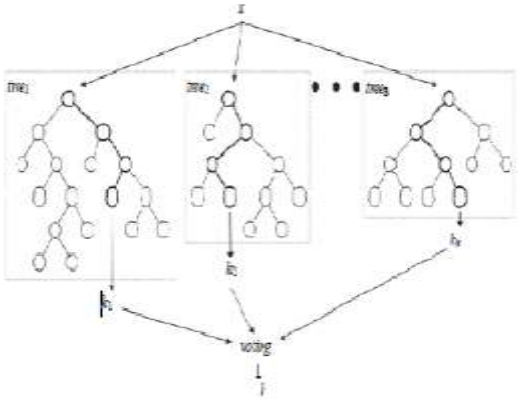


Fig 3: Decision tree representation

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

Into homogenous or near homogenous terminal nodes (the ends of the tree). A good binary split pushes data from a parent tree-node to its two daughter nodes so that the ensuring homogeneity in the daughter nodes is improved from the parent node. RF is often a collection of hundreds to thousands of trees, where each tree is grown using a bootstrap sample of the original data. RF trees differ from CART as they are grown non-deterministically using a two-stage randomization procedure.

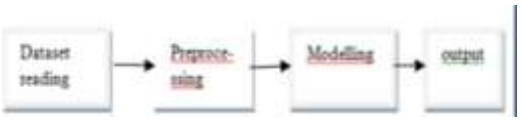
In addition to the randomization introduced by growing the tree using a bootstrap sample of the original data, a second layer of randomization is introduced by growing the tree using a bootstrap sample of the original data, a second layer of randomization is introduced at the node level when growing the tree. Rather than splitting a tree node using all variables, RF selects at each node of each tree, a random subset of variables and only those variables are used as candidates to find the best split for the node. The purpose of this two-step randomization is to de-correlate trees so that the forest ensemble will have low variance, a bagging phenomenon. RF trees are typically grown deeply.

N-Fold Cross Validation

In n-fold cross validation approach, we randomly partition into N sets of equal size and run the learning algorithm N times. Each time, one of the N sets is the test set, and the model is trained on the remaining N-1 sets. The value of K is scored by averaging the error across the N test errors. We can then pick the value of K that has the lowest score, and then learn model parameters for this K. A good choice for N is N=M-1, where M is the number of data points. This is called Leave-one-out cross validation formally reflects what we have learned about the validity of the hypothesis from consideration of the data.

Procedure

To build the random forest classifier and SVM we are going to use the breast cancer dataset by using two algorithms to predict breast cancer type (Benign or malignant).



Dataset reading:

Input is the breast cancer dataset which contains 569 samples and 30 attributes. Each instance has one of 2 possibilities namely Benign or malignant one of the other numeric value columns is ID column of instances.

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

Preprocessing:

Drop the unnamed column which is related to patient ID. In a preprocessing we can split the data into two parts namely training set and testing set.

Modelling:

In modeling we are applying algorithm on the given dataset namely support vector machine and random forest classifier.

Output:

We can obtain the performance metrices that is accuracy. Accuracy in classification problems is the number of correct predictions made by the model over all kinds prediction made.

EXPERIMENTAL RESULTS

In this section we present experimental results of the proposed method for the breast cancer diagnosis were implemented by using python 3.5 versions with pandas library package.

Table 1: Represents the accuracy

S. No	Algorithm	Output (accuracy)
1.	Random forest classifier	97.6%
2.	Support vector machine	95.9%

Output1:

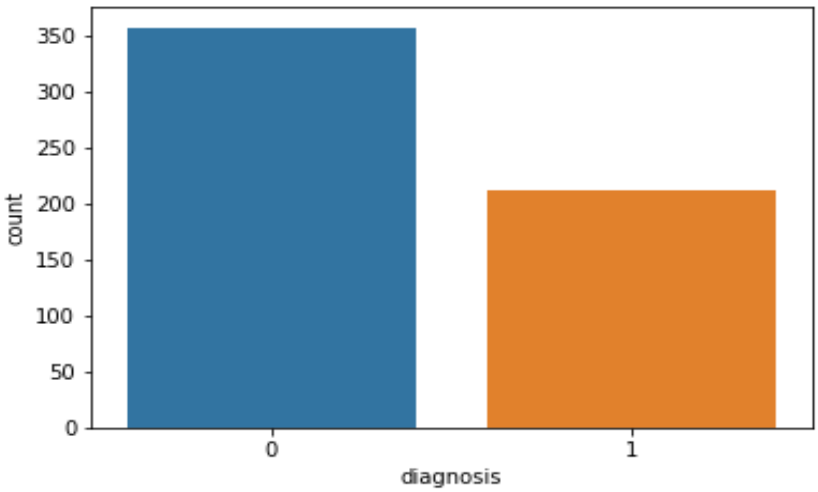


Fig 4: Graph between diagnosis and count

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

Output 2:

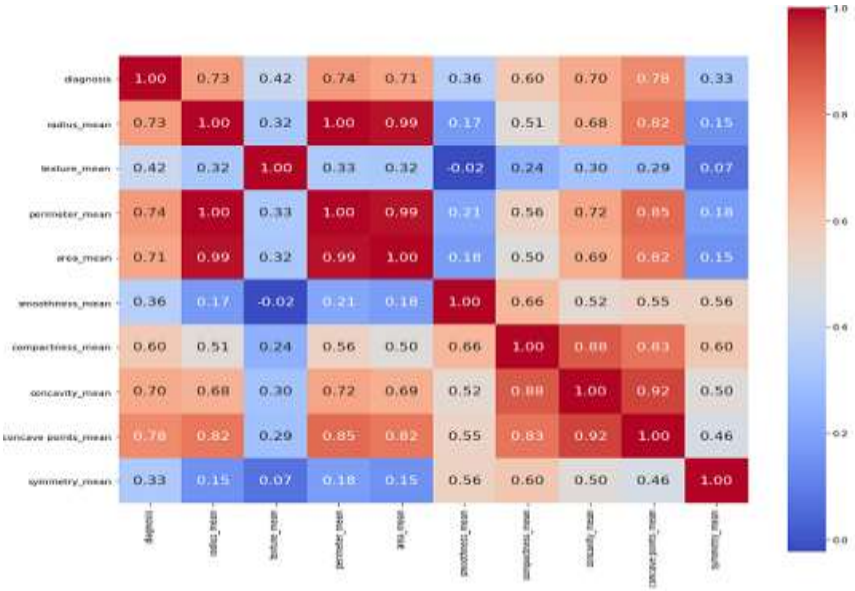


Fig 2: Representation of correlation matrix

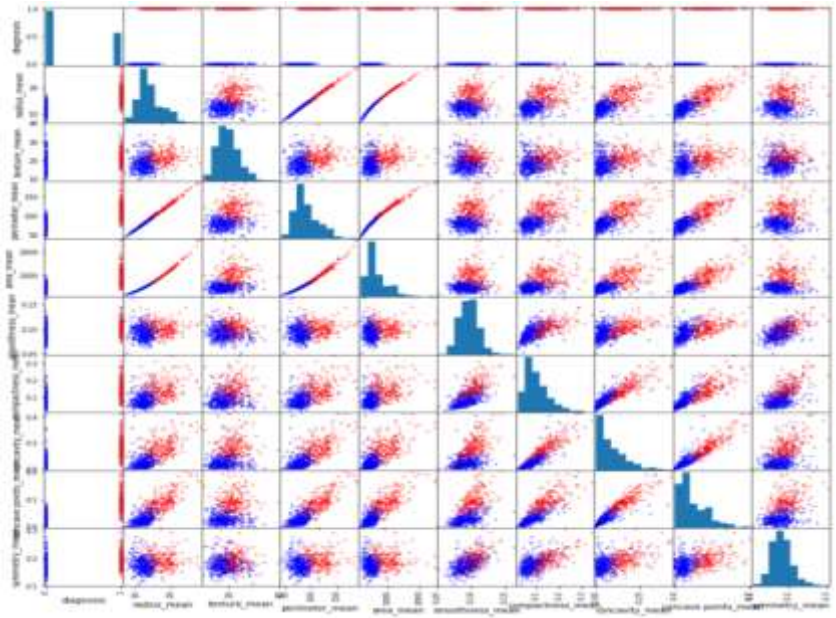


Fig 5: Representation of scatter plot matrix

CONCLUSION

In order to significantly improve predictive accuracy for breast cancer diagnosis, we developed a classifier model by combining the random forest classifier and support vector machine. The classification accuracies of the proposed models are by taking 70%as a training set and 30% as testing set by implementing these two algorithm with in accuracy 97.6% and 95.9%.

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

REFERENCES

1. The Women's Health Resource (2013) http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-2
2. WHO (2012) Breast cancer: Prevention and controlhttp://www.who.int/cancer/detection/bre-astcancer/en/index1.html
3. UCI Machine Learning Repository (2012) Wisconsin breast cancer dataset.
4. Chen, H.-L., Yang, B., Liu, J. and Liu, D.-Y. (2011) A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications
5. Keles, A., Keles, A. and Yavuz, U. (2011) Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Systems with Applications,38,.doi:10.1016/j.eswa.2010.10.061
6. Marcano-Cedeño,A., Quintanilla-Dominguez, J. and Andina, D. (2011) WBCD breast cancer database classification applying artificial meta plasticity neural network. Expert Systems with Applications,38,.doi:10.1016/j.eswa.2011.01.167
7. Ince, M.C. and Karabatak, M. (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36, 3465-3469. doi:10.1016/j.eswa.2008.02.064
8. Polat, K. and Günes, S. (2007) Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 11, 694-701. doi:10.1016/j.dsp.2006.10.008
9. Ubeyli, E.D. (2007) Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33, 1054-1062. doi:10.1016/j.eswa.2006.08.005
10. Sahana, S., Polat, K., Kodaz, H. and Günes, S. (2007) A new hybridmethod based on fuzzy-artificial immune system and k-nn algorithmfor breast cancer diagnosis. Com- puters in Biology and Medicine, 377, 415-423. doi:10.1016/j.compbiomed.2006.05.003
11. Abonyi, J. and Szeifert, F. (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognition Letters

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

12. Quinlan, J.R. (1996) Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research,
13. Krishnan, M.M.R., Banerjee, S., Chakraborty, C., Cha-560 C. Nguyen et al. / J. Biomedical Science and Engineering 6 kraborty, C. and Ray, A.K. (2010) Statistical analysis of mammographic features and its classification using support vector machine. Expert Systems with Applications
14. Stoean, R. and Stoean, C. (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Systems with Applications
15. Mu, T.T. and Nandi, A.K. (2007) Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier. Journal of the Franklin Institute
16. Zhang, Z.W., Shi, Y. and Gao, G.X. (2009) A rough set- based multiple criteria linear programming approach for the medical diagnosis and prognosis. Expert Systems with Applications
17. Li, D.-C., Liu, C.-W. and Hu, S.C. (2011) A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artificial Intelligence in Medicine
18. Ghazavi, S.N. and Liao, T.W. (2008) Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine.
19. Breiman, L. (2001) Random forests. Machine Learning Journal Paper
20. Wu, X.D. and Kumar, V. (2009) The top ten algorithm in data mining. Chapman & Hall/CRC, London.
21. Biau, G., Devroye, L. and Lugosi, G. (2008) Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9, 2015-2033.
22. Verikas, A., Gelzinis, A. and Bacauskiene, M.(2011) Mining data with random forests: A survey and results of new tests. Pattern Recognition, doi:10.1016/j.patcog.2010.08.011
23. Liaw, A. and Watthew, M. (2002) Classification and regression by random forest. R News, 3, 18-22.

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved

24. Breiman,L. (2004). RFtools—For predicting and understanding data.Technical Report.
25. Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1993) Classification and regression tree. Chapman & Hall, London.
26. Su, X. (2007) Bagging and random forests.
27. Efron, B. (1994) The jackknife, the bootstrap and other resampling plans. 6th Edition, Capital City Press, Baton Rouge, 1994.
28. Dupret, G. and Koda, M. (2001) Theory and methodology: Boostrap resampling for unbalanced data in super- vised learning. Eropean Journal of Operational Research, 134, 141-156. doi:10.1016/S0377-2217(00)00244-7
29. Good, P.I. (2006) Resampling methods: A practical guide to data analysis. 3rd Edition, Birkhauser.
30. Hsu, C.-C., Wang, K.-S. and Chang, S.-H. (2011) Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization. Expert Systems with Applications,38,46984704. doi:10.1016/j.eswa.2010.08.150
31. Koch, K.-R. (2007) Introduction to Bayesian statistics. Springer, New York, 2007.
32. Brase, C.H. and Brase, C.P. (2012) Understanable statistics. 10th Edition, Cengage Learning, Stamford.

Page 108-117 © MANTECH PUBLICATIONS 2018. All Rights Reserved