INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 3, Issue 1, January- June (2012), pp. 44-53
© IAEME: www.iaeme.com/ijcet.html
Journal Impact Factor (2011): 1.0425 (Calculated by GISI)
www.jifactor.com

IJCET
©IAEME

A NOVEL LNS SEMI SUPERVISED LEARNING ALGORITHM FOR DETECTING BREAST CANCER

S. Aruna
Research Scholar
Department of Computer Applications
arunalellapalli@yahoo.com

L.V. Nandakishore
Assistant Professor
Department of Mathematics
lellapalliarunakishore@gmail.com

Dr S.P. Rajagopalan
Professor Emeritus
Department of Computer Applications
sasirekharaj@yahoo.co.in

ABSTRACT

Semi supervised learning is a relatively new area in machine learning which represents the blend of supervised and unsupervised learning. It has the potential of reducing the need of expensive labeled data whenever only a small set of labeled examples are available. In this paper semi supervised learning algorithm combining Logical data analysis based on complete binary tree with Naïve Bayes and SVM with learning based on both labeled and unlabeled data is proposed for detecting breast cancer. Few labeled data are used as supportive set to build the diagnostic model which is used for classifying the unlabeled data. Wisconsin breast cancer dataset from the UCI machine learning depository is used for the experiment. This algorithm yielded an accuracy of 98.7% for unlabeled samples.

Keywords

Breast Cancer diagnosis, Logical data analysis, Naïve Bayes, Semi supervised learning, SVM.

1. INTRODUCTION

In machine learning, the classification task is commonly referred as supervised learning. Traditional supervised learning needs sufficient labeled data as training to get a strong generalization [1]. Obtaining a lot of labeled data is difficult in practice. Compared with labeled data, unlabeled data are sufficiently easier to obtain. If only a small amount of labeled data and large amount of unlabeled data are available semi supervised learning (SSL) strategy can provide a satisfactory classifier. SSL theory and algorithm developed quickly in recent years [2] because it has become a research focus in the field of machine learning attracting much more scholars to devote themselves to further study. SSL combines both labeled and unlabeled examples to generate an appropriate function or classifier. There has been increased interest in devising learning techniques that combine unlabeled data with labeled data. There are a number of medical areas to which machine learning systems have been applied. Breast cancer is one of them.

Breast cancer accounts for the second most cancer deaths in the world on which public awareness has been increasing in the last few decades. There has been a steady rise in the number of patients suffering from breast cancer. Although breast cancer is a potentially fatal condition early diagnosis of disease can lead to successful treatment [3]. One of the important steps to diagnose the breast cancer is classification of tumor. Tumors can be either benign or malignant. Only the malignant tumors are cancerous. Most breast cancers are detected by the patient as a lump in the breast. The majority of breast lumps are benign so it is the physician's responsibility to diagnose breast cancer, that is, to distinguish benign lumps from malignant ones. The long-term survival rate for women with breast cancer is improved by detecting the disease in its early stage [4]. Early diagnosis needs a precise and reliable diagnosis procedure that allows physicians to distinguish between benign breast tumors and malignant ones [5]. Unfortunately not all the physicians are experts in cross domain. Hence automation of diagnostic system is needed.

In this paper a novel LNS semi supervised learning algorithm is proposed for detecting breast cancer which is a mixture of 3 expert classifiers namely Logical data analysis model based on complete binary tree, Naïve Bayes and SVM. This paper is structured as follows. Section 2 gives a brief introduction to the expert classifiers used for the algorithm. The LNS algorithm is explained in section 3. Section 4 discusses the results obtained and concluding remarks are given in section 5 to address further research issues.

2. METHODOLOGIES

2.1 Classification Problem in SSL

Semi supervised learning is half way between supervised and unsupervised learning. The normal form of semi supervised learning set [2] is such that given set $X=\{\{X_t\},\{X_u\}$ a set of n points can be divided in to a training set $X_t=\{(x_i,c_i)|x_i \in R^m, c_i \in \{-1,1\}\}$ where $i=(1,..,m)$, c_i indicates the class to which the point x_i belongs and a test set $X_u=\{x_j| x_j \in R^{n-m}\}$ where $j=\{m+1,\ldots,n\}$ whose class labels are unknown. The goal of semi-supervised learning is to use existing labeled data in conjunction with unlabeled data to generate more accurate classifiers than using the labeled data alone. A good overview of semi-supervised learning is provided by [6].

2.2 Logical Data Analysis Model

LAD is a data mining method based on combinatorics, Boolean functions and optimization [7] that has been successfully applied to data analysis problems in different domains, including biology and medicine [8]. One of the underlying principles of LAD is to disregard the exact values of a variable, considers patient only whether the corresponding value of this variable is sufficiently high or low. In order to distinguish between measurements of benign and malignant tumors, only a fraction of the information contained in the dataset is needed. This set is called the supportive set.LAD was originally developed for analyzing binary data using the theory of partially defined Boolean functions. An extension of LAD for numerical data can be done by the process of binarization.

2.2.1 Tree based LAD model

First data discretization using equal width binning is done for the training set to reduce the number of values for a given continuous attributes by dividing into equal intervals. Five bins are used for discretization technique. Then the interval values are used to replace the original data values. Probability values for Benign P(B) and Malignant P(M) samples are calculated for each bin. In general binning techniques does not use class information. Binarization of the bin values are done such that if P(B)>P(M) then the bin values are changed to 0 else 1. Complete binary trees are created using the attribute values having 1. A binary tree T with n levels (root node at level 0) is complete if all levels except possibly the last are completely full, and the last level has all its nodes to the left side. A complete binary tree is very special tree. It provides the best possible ratio between the number of nodes and the height. The height h of a complete binary tree with N nodes is at most O (log N). From the binary tree various Boolean expressions are experimented to obtain a common pattern CP to distinguish the benign and malignant trees. Using P(B) and P(M) binarization of test set is done. Binary trees are created for the test set having nodes with attribute value as 1. Using CP the samples of the test set are labeled benign or malignant.

2.3 Naïve Bayes algorithm:

Naive Bayes classifier is a probabilistic classifier based on the Bayes theorem, considering a strong (Naive) independence assumption. Thus, a Naive Bayes classifier considers that all attributes (features) independently contribute to the probability of a certain decision. Taking into account the nature of the underlying probability model, the Naive Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect [9]. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

$$p(C|F_1,\dots,F_n) = \frac{p(C)\,p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}. \qquad (1)$$

where P is the probability, C is the class variable and F_1.......F_n are Feature variables F_1 through F_n The denominator is independent of C.

2.4 Support Vector Machines (SVM)

Support vector machines (SVM) are a class of learning algorithms which are based on the principle of structural risk minimization (SRM) [10]. SVMs have been successfully applied to a number of real world problems, such as handwritten character and digit recognition, face recognition, text categorization and object detection in machine vision

[11]. SVMs find applications in data mining, bioinformatics, computer vision, and pattern recognition. SVM has a number of advanced properties, including the ability to handle large feature space, effective avoidance of over fitting, and information condensing for the given data set.etc.[10]. Given training examples labeled either "yes" or "no", a maximum-margin hyper plane is identified which splits the "yes" from the "no" training examples, such that the distance between the hyper plane and the closest examples (the margin) is maximized. The use of the maximum-margin hyper plane is motivated by Vapnik Chervonenkis theory, which provides a probabilistic test error bound that is minimized when the margin is maximized. The parameters of the maximum-margin hyper plane are derived by solving a quadratic programming (QP) optimization problem. There exist several specialized algorithms for quickly solving the QP problem that arises from SVMs.

3. LNS SEMI SUPERVISED ALGORITHM

The LNS algorithm is the mixture of three expert classifiers namely tree based LAD, Naïve Bayes and SVM. This algorithm works on the principle of incremental learning. Unlabeled samples classified with high confidence are used to enlarge the pool of labeled samples. The primary goal of our research is to utilize the information from the unlabeled data effectively so that high accuracy of the classifier can be achieved.

Method: LNS algorithm

Input: F_{train} with n labeled samples and F_{test} with m unlabeled samples, n>0 and n<<m.

Output: Labels generated for F_{test} .

Step 1: Dsicretization of F_{train} using equal width binning.

Step 2: Calculating the probability values P(B) and P(M) for each bin.

Step 3: Binarization of attribute values using P(B) and P(M) obtained from step 2.

Step 4: Creation of CBT using values obtained in step 3.

Step 5: A common pattern of CP is obtained from CBT created in step 4 that can distinguish the state of the tumor as benign and malignant.

Step 6: Repeat step 3 and 4 for F_{test}.

Step 7: Using CP obtained in step 5 classify F_{test}.

Step 8: Move 2n pseudo labeled samples from F_{test} to F_{train}.

Step 9: Run the Naïve Bayes classifier to label F_{test} using the new F_{train}

Step 10: Repeat step 8.

Step 11: Run the SVM classifier to predict the labels for unlabeled samples using the new training set F_{train}.

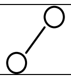
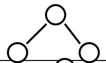
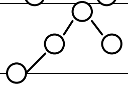
Highlights of the algorithm are 1) In general binning techniques does not use the class information. In our binning method we used the class information from the training test to calculate the probability values for benign and malignant samples in each bin. 2) CBT is used for data visualization because CBT can be expressed as various Boolean expressions than a table of data that is properly indexed and 3) Unsupervised discretization technique is combined with Naïve Bayes to improve the accuracy of the classifier.

The advantages of combining the three classifiers are LAD requires only a small amount of training information as supportive set to build the model, Naive Bayes performs often well even when the assumption is violated, it can be learned incrementally and it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification and SVM is based on the structural risk minimization principle (SRM). Comparing with other learning methods, its generalization is optimal. In training SVMs, the decision boundaries are determined directly from the training data so that the generalization ability is maximized. SVM require only few parameters for tuning the learning machine, Learning involves optimisation of a convex function and it scales relatively well to high dimensional data.

4. RESULTS

The bench mark well known Wisconsin breast cancer dataset from UCI machine learning depository is used for the experiment. It has 699 instances (Benign: 458 Malignant: 241) of which 16 instances has missing attribute values removing that we have 683 instances of which 444 benign and 239 are malignant The 683 samples are split randomly into a training set that consists of 25 malignant samples and 40 benign samples. The test set is simulated by removing the class labels from the remaining samples. Gist (http://svm.sdsc.edu) is used for all the training and testing of SVM. Gist is an implementation of SVM algorithm. Each attribute has domain values 1-10 and five bins are used for discretization. Binarization of training set is done using the P(B) and P(M) values. CBTs are created from the attribute values having 1. From the CBT various Boolean expressions are experimented to obtain a common pattern CP to distinguish the benign and malignant trees. Table 1 shows the attribute values, changed binary values , binary tree created and the Boolean expression to find the pattern CP for some samples in training set. From the table it can be seen that the Boolean expression if((btree$_i$) ^ (btree$_i$->rt)) is true for malignant trees and false for benign trees.

Table 1 Attribute values, changed binary values, binary tree created, Boolean expression and the class for training set.

Attribute values	Binarization using P matrix	Complete binary tree	Boolean expression if((btree)^(btree ->rt))	Class
1,1,1,1,2,1,2,3,1	0,0,0,0,0,0,0,0,0	Null tree	False	Benign
1,3,1,2,2,2,5,3,2	0,0,0,0,0,0,1,0,0		False	Benign
6,3,3,3,3,2,6,1,1	1,0,0,0,0,0,1,0,0		False	Benign
3,2,10,2,6,8,2,1,1	0,0,1,0,1,1,0,0,0		True	Malignant
7,2,2,1,6,10,9,3,1	1,0,0,0,1,1,1,0,0		True	Malignant

Results for the Naïve Bayes classifier with unsupervised discretization is given in table 3.

Table 2- Results for Naive Bayes Classifier

ID	Class
733823	Malignant
603148	Benign
183936	Benign
263538	Malignant
320675	Malignant
476903	Benign
566509	Benign
603148	Benign
659642	Malignant
255644	Malignant

Using gist SVM one can get the results as in Table 3 and Table 4. From the results, we see that the gist SVM is able to train a classifier with weight and discriminant Let D be the discriminant value. Then, according to the discriminant value D, we can classify the new data point into the positive class if D > 0, and classify it into the negative class when D < 0.

Table 3 Results for training set

Example	Class	Weight	Train_Classification	Train_Discriminant
790287	**1**	0.5976	**1**	0.5008
63375	1	0.6362	1	0.4667
837480	1	0.6384	1	0.4648
142932	1	0.6692	1	0.4395
95719	1	0.7243	1	0.3931
558538	-1	-3.186	-1	-0.1526
183913	-1	-2.309	-1	-0.3917
167528	-1	-1.886	-1	-0.5015
566346	-1	-1.628	-1	-0.5668
636437	-1	-1.543	-1	-0.5939

Table 4 Results for test set

Example	Classification	Discriminant
508234	**1**	1.31174
488173	1	1.31113
601265	1	1.23666
608157	1	0.636931
555977	1	0.538721
635844	-1	-0.114497
378275	-1	-0.200625
521441	-1	-0.225032
303213	-1	-0.226484
324427	-1	-0.254418

Example in table 3 and 4 is the name provided for the samples, Class (training results only) is the class membership provided for the samples, Weight (training results only) is the importance of the example in setting the location of the decision boundary (which is the maximum margin hyperplane). Examples with non-zero weights are support vectors, train_classification (training results only) or classification is the predicted class of the example, or, for training, the location of the example with respect to the decision boundary. In training, if it differs from the Class, a training error is counted and train_discriminant (training results only) or discriminant (test results) is how far the example is from the decision boundary. Larger values correspond to greater certainty that the sample belongs to the predicted class.

Table 5 shows the results for the 3 passes of the LNS algorithm. Empirical comparison of the results show that the information from unlabeled data have improved the accuracy of the classifier.

Table 5 Results of the LNS algorithm for WBC dataset

Algorith m	Training-Test partition(%)	Training set accuracy(%)	Test set accuracy (%)
Tree based LAD	10 - 90	97.5	92.6
Naïve Bayes	30 - 70	93.3	95.8
Gist-SVM	50 - 50	99.1	98.7

There have been several studies reported in literature focused on medical diagnosis of breast cancer with Wisconsin breast cancer (WBC) dataset. Classification accuracies obtained with our algorithm and other classifiers (Both supervised and Semi supervised) are given in table 6.

Table 5 Classification accuracies obtained with our method and other classifiers (supervised and semi supervised) from literature for WBC dataset (*SSL methods)

Author(year, Ref)	**Method**	**Classification accuracy (%)**
Quinlan (1996)[12]	C4.5	94.74
Hamiton et al. (1996)[13]	RAIC	95.00
Ster and Dobnikar (1996)[14]	LDA	96.80
Nauck and Kruse (1999)[15]	NEFCLASS	95.06
Pena-Reyes and Sipper (1999)[16]	Fuzzy-GA1	97.36
Setiono (2000)[17]	Neuro-rule 2a	98.10
Albrecht et al. (2002)[18]	LSA machine	98.80
Abonyi and Szeifert (2003)[19]	SFC	95.57
Übeyli (2007)[20]	SVM	99.54
Polat and Günes_ (2007)[21]	LS-SVM	98.53
Guijarro-Berdias et al. (2007)[22]	LLS	96.00
Akay (2009)[23]	SVM-CFS	99.51
Karabatak and Cevdet-Ince (2009)[24]	AR + NN	97.40
Peng et al. (2009)[25]	CFW	99.50
Marcano-Cedeno et al. (2011)[26]	AMMLP	99.26
Aruna et al (2011) [27]*	SVM-Naïve Bayes	93.3 (training set) 86.6 (test set) 50-50 partition

Our Method LNS algorithm*		99.1(training set)
		98.7(test set)
		50-50 partition

From the results it can be observed that even though this is a semi supervised algorithm the classification accuracies are almost equal to the supervised learning methods.

5. CONCLUSION

Machine learning and knowledge discovery from databases (KDD) are increasingly being applied in health care to build models for better medical decision making. Medical decision making can be seen as classification problem. Physician classifies the symptoms of a patient to a certain disease group on the basis of knowledge. For many real world applications, such as medical diagnosis, forensic science, fraud detection, etc labeled examples are very minimal whereas unlabeled examples are abundant. In such situations semi supervised learning can provide a satisfactory classifier. In this paper we propose a LNS semi supervised algorithm combining Tree based LAD, Naïve Bayes and SVM for detecting breast cancer. A diagnostic model is built by the Tree based LAD using the small quantity of expert classified labeled data as a supportive set. This algorithm is based on incremental learning. Naïve bayes classifier with unsupervised discretization is used to classify the test set by adding some pseudo labeled samples to the training set. SVM is used finally to classify all the unlabeled samples. Wisconsin breast cancer data set from UCI machine learning depository is used for the experiment. The algorithm yielded an accuracy of 98.7% for unlabeled samples. In this paper this algorithm is used for breast cancer domain. Further research in future using different domains and combining feature selection techniques will provide a broader experimental evaluation in improving the algorithm.

REFERENCES

[1] D.J.Miller and H.Uyar's, "A mixture of experts classifier with learning based on both labelled and unlabelled data" *Advance in NIPS* 9, pp 577, 1997

[2] Oliver Chapelle, Bernhard Scholkopf, Alexande Zien, "*Semi supervised Learning [M]*," The MIT Press, 2006

[3] I. Harirchi, et al., "Breast cancer in Iran: a review of 903 case records," Public Health, 114(2): p. 143-145, 2000.

[4] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002

[5] T. Subashini, V. Ramalingam, and S. Palanivel, "Breast mass classification based on cytological patterns using RBFNN and SVM" Expert Systems with Applications, 36(3): p. 5284-5290, 2009.

[6] X.J.Zhu, "Semi-supervised learning literature survey [R]" *Technical Report* 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, December, 2007.

[7] E. Boros ,PL. Hammer ,T. Ibaraki ,A. Kogan ,E. Mayoraz ,I. Muchnik ," An Implementation of Logical Analysis of Data", *IEEE Trans on Knowl and Data Eng* , 12: 292-306, 2000.

[8] PL. Hammer , TO. Bonates," Logical Analysis of Data: From Combinatorial Optimization to Medical Applications", *Ann Operations Res* , 148**:**203-225, 2006.

[9] S. Belciug, "Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection", *Proc. 8th International conference on Artificial Intelligence and Digital Communications,* 2008.

[10] Vladimir N. Vapnik. (1998) *Statistical Learning Theory*. New York: Wiley.

[11] You et al, (2010) "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network**"** *BMC Bioinformatics,* 11:343.

[12] R. Quinlan, Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, 779-809. 1996.

[13] H.J. Hamiton, N. Shan, & N. Cercone, RIAC: A rule induction algorithm based on approximate classification, In International conference on engineering applications of neural networks, University of Regina, 1996.

[14] Ster, B., & Dobnikar, A. Neural networks in medical diagnosis: Comparison with other methods. In Proceedings of the international conference on engineering applications of neural networks, pp. 427–430, 1996.

[15] D. Nauck, & R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data", *Artificial Intelligence in Medicine*, 16, 149–169, 1999.

[16] C.A. Pena-Reyes, & M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis", *Artificial Intelligence in Medicine*, 17, 131–155, 1999.

[17] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 18(3), 205–217, 2000.

[18] A.A. Albrecht, G. Lappas, S.A. Vinterbo, C.K. Wong & L. Ohno-Machado, Two applications of the LSA machine, In Proceedings of the 9th international conference on neural information processing, 184–189, 2002.

[19] J. Abonyi, & F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers", *Pattern Recognition Letters*, 14(24), 2195–2207, 2003.

[20] E.D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection", *Expert Systems with Applications*, 33(4), 10541062, 2007. Doi: 10.1016/ j.eswa.2008.02.064.

[21] K. Polat & S. Günes, "Breast cancer diagnosis using least square support vector machine", *Digital Signal Processing*, 17(4), 694–701, 2007.

[22] B. Guijarro-Berdias, O. Fontenla-Romero, B. Perez-Sanchez, & P. Fraguela, "A linear learning method for multilayer perceptrons using least squares", *Lecture Notes in Computer Science*, 365–374, 2007.

[23] Akay, M. F, "Support vector machines combined with feature selection for breast cancer diagnosis", Expert Systems with Applications, 36(2), 3240–3247, 2009.

[24] M. Karabatak & M. Cevdet-Ince "An expert system for detection of breast cancer based on association rules and neural network", *Expert Systems with Applications*, 36, 3465–3469, 2009.

[25] Peng, L, Yang, B., & Jiang, J, "A novel feature selection approach for biomedical data classification", Journal of Biomedical Informatics, 179(1), pp 809–819, 2009.

[26] A. Marcano-Cedeno, J. Quintanilla-Domínguez, D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network", Expert Systems with Applications 38 , 9573–9579, 2011.

[27] Aruna. S, Nandakishore. L. V. and Dr Rajagopalan. S. P., "An Algorithm Proposed for Semi- Supervised Learning in Cancer Detection", In Conference Proceedings Second International conference on Seiscon , ISBN-978-93-80430-00-3. Vol III, pp 860-864