International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

BRIDGING FEATURE SELECTION AND REDUCTION METHOD USING MSVM

Dr.C.Kalaiselvi,
Associate professor and Head,
Department of Computer Applications,
Tiruppur Kumaran College for Women,
Tirupur, Tamilnadu, India.

V.Mehalarajini,
M.Phil., Research Scholar,
Department of Computer Applications,
Tiruppur Kumaran College for Women,
Tirupur, Tamilnadu, India.

Abstract: A major part of current research in data mining is the ground of medical diagnosis. In the present learning with the Breast cancer Wisconsin data sets, a feature selection algorithm Modified Support Vector Machine Feature Selection (MSVMFS) predicts together diagnosis and prognosis by comparing several data mining classification algorithms. In the proposed approach, in level one of feature selection, features are selected based on rough set among different starting values of feature reduction. In level two features are selected since the reduced set based on the Correlation Feature Selection (CFS). Experiments show the proposed method is valuable by comparing through others in terms of number of preferred features and classification performance.

Keywords: Data mining, feature selection, rough set, correlation, breast cancer.

I. INTRODUCTION

Data mining is the procedure of extracting functional and related information form a database. Feature Selection (FS) is an essential concept in pattern recognition with data mining. It aims to select the unique features as of a set of features and eliminating pointless features. Rough set theory be able to used as a tool toward reduce unnecessary features and to deal with elusiveness and uncertainty in datasets. The main model in rough set theory is to define the necessity of features. The measures of necessity are calculated by the functions of approximation. Rough set have been used to improve the efficiency and usefulness of classification and applied for classification within various applications [2]. Breast cancer is considered like a major health issue for women. For the diagnosis of malignant ones in the Wisconsin Diagnostic Breast Cancer (WDBC) data set since for the recurrence of breast cancer in the Wisconsin Prognostic Breast Cancer (WPBC) data set, many techniques has been discussed [3], [4], [5] and [6].

In the present learning both Breast cancers Wisconsin Diagnostic with Prognostic datasets are used. Two levels of characteristic feature reductionions are proposed support for breast cancer detection and prognosis problems. In level one of feature feature reductionion, minimal feature subset is selected based on rough set and in level two features are selected starting the minimal feature subset obtained from level one based scheduled the correlation feature selection. Furthermore, special classification techniques are used to study the impact of the special features.

II. MATERIALS AND METHODS

2.1Rough Set Feature Selection

Rough set theory was introduced by Pawlak in 1982 [7]. In rough set theory, an information table is defined as I = (U,C∪ D,V,f) everyplace U is the universe of primitive objects and C a set of condition attributes, D a set of decision attributes, V a set of values of attributes in C as well as f:C→V a description function. For any P⊆C, there is an equality relation IND (P) as follows:

$$IND\ (P) = \{\ (x,y)\ \}\ \epsilon\ U\ |\ ʚ\ a\ \epsilon\ p\ a(x) = a(y)\ \}$$

If (x,y)∊ IND(P), x and y are indiscernible by attributes from P. Assuming *P* and *Q* are equivalence associations in *U*, the significant concept positive region is defined:

$$POSp(Q) = U\ x\epsilon u/Q\ Px$$

The positive region contains all items of *U* can be classified with certainty into classes of U/Q using attributes from P.

Modified Support Vector Machine algorithm [8] is frequently used in the direction of generate minimal feature subset. This algorithm uses the degree of dependency

$$\gamma_P(Q) = \frac{||POS_P(Q)||}{||U||}$$

It starts among an empty set of attributes. The greatest of novel attributes is determined and added to the set iteratively with the above dependency. The process is repeated until the dependency of feature reduction candidate equals to 1.

2.2Correlation Feature Selection (CFS)

It evaluates to worth a subset of attributes by considering the entity predictive ability of every feature along through the degree of redundancy between them. That is subsets of features so as to be highly correlated with the class while have low inter association are preferred. In this paper greedy search approach is used for WDBC data set and random search is used for WPBC data set.

2.3Wisconsin Diagnostic and Prognostic datasets

The Breast Cancer Wisconsin Diagnostic dataset (WDBC) and the Breast cancer Wisconsin Prognostic (WPBC) dataset were obtained to the UCI Machine Learning Repository [9]. Features are computed from digitized image of a Fine

International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

Needle Aspiration (FNA). The WDBC consists of 569 instances and 30 real valued input features where WPBC consists of 198 instances and 33 features. The attributes of two datasets are nearly the same so far the WPBC

Table 1. Attribute information of the Breast Cancer Wisconsin datasets

Diagnostic dataset	Prognostic dataset
1) ID Number 2) Diagnosis (M – Malignant, B-benign) 3) ----- 4-33) ten real valued features are computed for each cell nucleus: a. Radius (mean of distances from center to points on the perimeter) b. Texture (standard deviation of gray-scale values) c. Perimeter (perimeter of the cell nucleus) d. Area (area of the cell nucleus) e. Smoothness (local variance in radius lengths) f. Compactness (perimeter^2/area-1.0) g. Concavity (severity of concave portions of the contour) h. Concave points (number of concave portions of the contour) i. Symmetry (symmetry of the cell nuclei) j. Fractal dimension (coastline approximation-1)	
34) ---- -	Tumor Size (size of the tumor)
35) ---- -	Lymph node status

III. PROPOSED METHOD

MSVMFS is a two-level feature feature reductionion algorithm. The aim of this algorithm is achieve a feature subset with smallest number of features provided that efficient classification exactness.

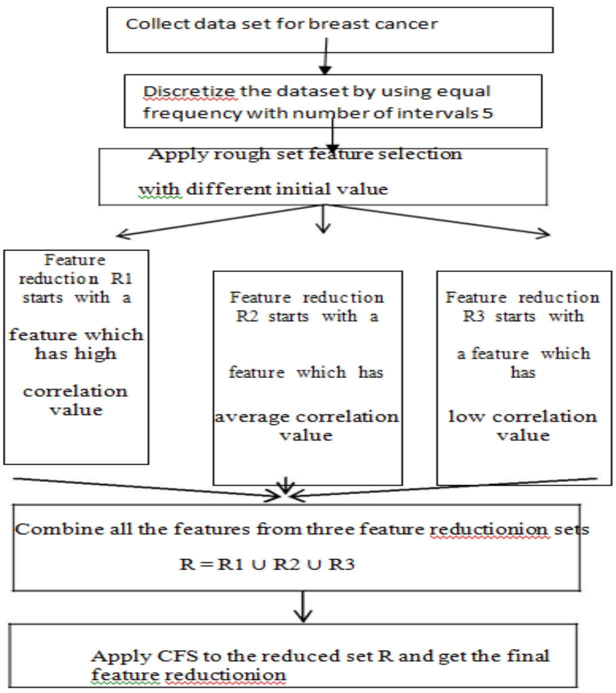


Figure 1. Proposed MSVMFS Method

In composed of two feature feature reductionion algorithms. Rough set Modified Support Vector Machine algorithm is useful initially to obtain minimal feature subset. Then second algorithm CFS is used to do further feature reductionion in minimal feature subset. Usually in Modified Support Vector Machine algorithm, the initial feature reduction set say R starts with an empty set but in our proposed method R is initialized with three different values in high correlation feature, regular correlation feature with low correlation feature and decision attribute (class label). Because to starting point in the attribute space influences the direction of exploration. Then CFS functional to the union of all the features of those three feature reduction sets. The data sets are discretized for the reason of rough set by using equal incidence with number of intervals 5 before applying feature selection algorithms

Algorithm:
Input: Data set D with all features.

Output: Optimum feature subset R

Step 1: Discretize the data set using equal frequency with number of intervals 5
Step 2: Apply rough set feature selection algorithm with an initial value of the feature reduction set R1= {attribute which has high correlation value}.
Step 3: Apply rough set feature selection algorithm with an initial value of the feature reduction set R2= {attribute which has average correlation value}.
Step 4: Apply rough set feature selection algorithm with an initial value of the feature reduction set R3= {attribute which has low correlation value}.
Step 5: Combine all the features from R1, R2 and R3 and obtain R. i.e., R=R1∪ R2∪ R3
Step 6: CFS is applied for the reduced set R to get further feature reductionion in the features.

IV. RESULT

The proposed methods have been implemented using MATLAB (Version 7.12). It is used eliminate the insignificant in addition to redundant features. The reduced attribute set obtained for WDBC dataset is: {11, 17, 22, 24, 25, 26, 28, and 31}. The reduced attribute set obtained for WPBC is {3, 27, and 35}. In this paper WEKA toolkit is implemented and to analyze the datasets with the data mining algorithms [10]. Table 2 shows three different feature reduction sets obtained using rough set and table 3 shows the final feature reduction sets of WDBC and WPBC datasets after applying CFS

Table 2. Features sequentially selected against different initial values

Data set	Feature reduction R1	Feature reducti on R2	Feature reduction R3
WDBC	31,17,25,8,22	25,26,28,22,14	18,24,29,25,14,11

WPBC	27,9,32,15,3	28,6,3,3 5,31	3,19,35,16 ,13

Table3. Feature subsets selected on the breast cancer datasets

Data set	Combine all three feature reductions R1,R2 and R3	Apply CFS to get final feature reduction R
WDBC	8,11,14,17,18,22,24,25, 26,28,29,31	11,17,22,24,25,26 ,28,31
WPBC	3,6,9,13,15,16,1 9,27,28,31,32,3 5	3,27, 35

The data mining algorithms such as Bayes net, Naïve bayes, Multilayer perceptron, RBF network, IBK, J48 and Simple cart are used to classify WDBC and WPBC datasets through all the features with optimum features selected via our projected method. The results are shown in Table 4 and Table 5. To obtain high accuracy of a prediction model, optimal parameter setting play an essential role. In this paper to evaluate the appropriate algorithmic parameters of all the mentioned eight data mining algorithms and use 80-20 training-testing detachment of the data. In our results demonstrate the projected method improves to the classification accuracy almost all the data mining algorithms. The graphical demonstration of the performance to the classification algorithms WDBC and WPBC are portrayed in Fig. 2 and Fig. 3 respectively.

Table 4. Classification Accuracy on WDBC dataset

S. No	Data Mining Algorithm	Considering all the features. (Accuracy %)	MSVMFS feature subset (Accuracy %)
1	Bayes Net		94.7368
2	Naïve Bayes	90.3509	94.7368
3	MLP	96.4912	100
4	RBF	92.9825	99.1228
5	SMO	97.3684	97.3684
6	IBK	95.6140	98.2456
7	J48	92.9825	96.4912
8	Simple Cart	92.1053	94.7368

Table 5. Classification Accuracy on WPBC dataset

S. No	Data Mining Algorithm	Considering all the features. (Accuracy %)	MSVMFS feature subset (Accuracy %)
1	Bayes Net	77.5	82.5
2	Naïve Bayes	72.5	82.5
3	MLP	72.5	82.5
4	RBF	75	82.5
5	SMO	77.5	85
6	IBK	72.5	80
7	J48	75	82.5
8	Simple Cart	77.5	80

Figure 2.Classifier performance before and after feature selection on WDBC data set

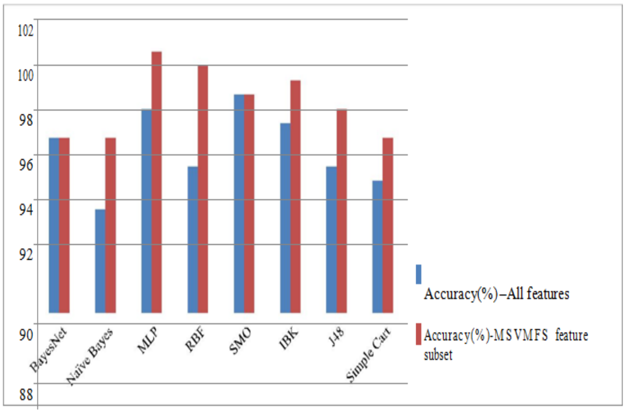
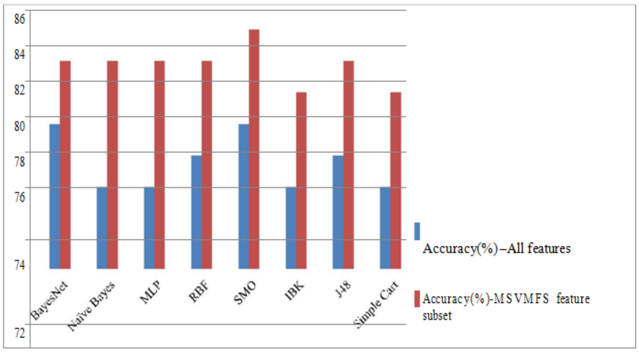


Figure3. Classifier performance before and after feature selection on WPBC data set



Comparing the accuracies in Table 4 and 5, it is establish the MSVMFS improves to accuracy of all the data mining algorithms. Multilayer perceptron has been produced 100 percent accuracy in classifying the WDBC dataset and SMO is the greatest performing classification algorithm on the WPBC dataset which provides 85 percent classification accuracy Classification accuracy of MSVMFS algorithm and other methods for WDBC and WPBC from literature are summarized in Table 6

Table 6. Accuracy rate comparison of MSVMFS with other approaches from existing researches on WDBC and WPBC datasets.

Classifier	Data set	Accuracy (%)
CART with feature selection (Chi-square) [11]	WDBC	92.61%
Hybrid Approach [12]	WDBC	95.96%

Jordan Elman neural network[13]	WDBC WPBC	98.25 70.725%
Rough set K-Means Clustering[14]	WDBC	99.12%
Proposed method(MSVMFS)	WDBC WPBC	100 85

V. CONCLUSION

In this paper MSVMFS feature selection algorithm is proposed for breast cancer datasets. In a two level attribute feature reductionion algorithm with combination of rough set and CFS. In the goal of this algorithm is select minimum number of features given that high classification accuracy. It is observed that our proposed model achieved highest classification accuracy compared to additional feature selection methods. MLP classification algorithm created 100 percent accuracy classifying the WDBC dataset. It also affirms that the SMO algorithm is the best performing algorithm on the WPBC dataset which provides85 percent classification accuracy.

VI. REFERENCES

[1]. Liu H. and Motoda H., "Feature Selection for knowledge Discovery and Data Mining ", Kluwer Academic Publisher, 1999.

[2]. Jensen R. and Shen Q.," A Rough Set-Aided System for Sorting WWW Bookmarks", In Zhong N *et al.* (Eds.), Web Intelligence: Research and Development, pp. 95-105, 2001.

[3]. Setiono R., "Generating concise and accurate classification rules for breast cancer diagnosis", Artificial Intelligence in Medicine, 18:205–219, 2000.

[4]. Chen D., Chang R.F., Huang Y.L., "Breast cancer diagnosis using self-organizing map for sonography", Ultrasound in Medical Biology 2000, Vol. 26, pp. 405– 11.

[5]. Giger M., Huo Z., Kupinski M., Vyborny C., "Computer-aided diagnosis in mammography. In Handbook of Medical Imaging", (Eds.) Sonka, M., Fitzpatrick, J., Medical Image Processing and Analysis, Vol. 2. SPIE Press, pp. 917–986, 2000.

[6]. Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E.," A neural network approach to breast cancer diagnosis as a constraint satisfaction problem", Med. Phys. Vol.28, pp. 804–811, 2001.

[7]. Zdzislaw Pawlak,"Rough Sets-Theoretical Aspects and Reasoning about Data", Klower Academic Publication. 1991..

[8]. A.E.Hassanien, Z.Suraj, D.Slezak, and P.Lingras, "Rough Computing: Theories, Technologies, and Applications,"

[9]. NewYork: Information Science Reference, 2008.

[10]. http://archive.ics.uci.edu/m1/machine-learning-databases/breast-cancer-wisconsin/breast-cancer.

[11]. Weka: Data Mining Software in java http://www.cs.waikato.ac.nz/ml/weka/

[12]. D.Lavanya, Dr.K.Usha Rani.," Analysis of feature selection with classification: Breast cancer datasets", Indian Journal of Computer Science and Engineering (IJCSE), October 2011. D. Lavanya, "*Ensemble Decision Tree Classifier for Breast Cancer Data*," International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.

[13]. Chunekar, V.N.; Ambulgekar, H.P. (2009). "Approach of Neural Network to Diagnose Breast Cancer on Three

[14]. Different Data Set," Proceedings Advances in Recent Technologies in Communication and Computing 2009 ARTcom-2009), 27th-28th Oct., IEEE, Kottayam. pp:893-895.

[15]. T.Sridevii and A. Murugan, "An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set" International Journal of Computer Applications (IJCA), Vol. 85, No.11, pp 38-42, Jan 2014