ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

An Approach with SVM using Variable Features Selection on Breast Cancer Prognosis

Monika Lamba

PG Scholar, Dept. of CSE, USICT, GGSIP University, Delhi, India

ABSTRACT:-: The selection of parameters, with regards to accuracy is very important in support vector machines (SVM). In this paper using support vector machine classifier, i have used an approach to construct a model that is useful for classification. I have used 20 cross validation, 15 cross validation, 10 cross validation and 5 cross validation for variable selection on input feature vectors. Additionally, the accuracy (AUC), sensitivity, specificity and error rate of SVM is evaluated against well-known Breast Cancer medical dataset. SVM performance is much better than the other machine learning classifier. The empirical results demonstrate that the proposed variable feature selection SVM method can obtain much more appropriate model parameters that generates a high classification accuracy. Promisingly, the proposed SVM method can be regarded as a useful clinical tool for medical data classification and medical decision making.

KEYWORDS: Support Vector Machine, 20-Fold cross validation, 15-Fold cross validation, 10-Fold cross validation, 5-Fold cross validation, Machine Learning, feature selection.

I. INTRODUCTION

Machine Learning (ML) [1] plays a significant role in many applications of data mining [5] and pattern classification. Most of Machine Learning areas, where it can be successfully applied, are regression and classification problems, by improving the efficiency as well as the design of the systems. The features in dataset may be of different kind of dimensions, if features are given with known labels with corresponding correct outputs, is known as supervised learning. But in case of unsupervised learning, the attributes are unlabeled and outputs are not known. One more kind of Machine Learning exists, that is reinforcement learning where the training knowledge is given to the system by the external teacher that constitutes a measure of how well the system works is in the form of reinforcement learning. The learner is never instructed to make one of the desired actions, but rather discovering which actions lead to the best solution, by continuously trying every action to improve the efficiency and accuracy (AUC).

A. Supervised Learning Algorithms

Machine Learning is one of the method of learning a set of principles from instance of a training set, or more specific creating a classifier which may be used to generalize from new features or instances. The learning procedure is as follows: very primary step is to gather the dataset, if the collected dataset by any of the arbitrary process used is not directly suitable for induction, it may contain missing and noisy data values, hence requires significant pre-processing (Zhang, Schwartz, Wagner & Miller 2000). The next step is to prepare data, data pre-processing and the feature subset selection is identified as the process of identifying and removing as more redundant and irrelevant features as possible (10). This may lead for reducing the dimension of the data and allowing algorithms to perform efficiently and faster. As many features depend on one another and may affect the accuracy of supervised Machine Learning classification Models.

B. Algorithm Selection

The selection of learning algorithm [16] is one of the critical process. Once at primary stage when testing is judged and it comes out satisfactory, then that classifier is generalized. The accuracy of the classifier's is generalized based on prediction. There are multiple techniques mentioned to find the classifier's accuracy. One of the way is to split the training set by dividing the two-third for training and remaining for estimation performance. Another method used in this paper is known as cross validation in which training set is divided into equally-sized mutually exclusive subsets

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

and every subset of classifier is trained on the union of remaining subsets. The error rate in terms of average of each subset results an estimate of the rate of error for the classifier. If the error (%) is not tolerable then algorithm will go back to the previous stage of the Machine Learning supervised process.

There are two major motives to make use SVMs in the field of computational biology first, many problems have high dimensions as well as noisy data, for which SVM are called to perform well as compared to other statistical or ML methods. Second, in comparison to most ML methods, kernel methods like SVM can easily handle non-vector inputs, like graphs or variable length sequences. These types of data are most common in biology applications.

II. RELATED WORK

In [1] a training algorithm that maximizes the margin between the training patterns and the decision boundary is presented. The technique is applicable to a wide variety of classification functions, including Perceptions, polynomials, and Radial Basis Functions. The number of parameters is adjusted automatically to match the complexity of the problem. The result is expressed as a linear combination of supporting patterns. Therefore, there are the subset of training patterns that are nearest to the decision boundary. Conditions on the generalization performance based on the leave-one-out method and the VC-dimension are given. In this paper describes the experimental results on optical character recognition problems that demonstrate as one of the good generalization obtained when compared with other learning algorithms. Generalization performance of various pattern classifiers is achieved when the capacity of the classification function is similar to the size of the training set. Classifiers, with the help of a large number of adjustable parameters and large capacity mostly learn the training set without bugs, but exhibit poor generalization. Conversely, a classifier with not sufficient capacity cannot be able to remember the task at all. In between, there is an optimal capacity of the classifier that minimizes the expected generalization bugs for a particular amount of training data. In this paper author describe a training algorithm which will automatically tunes the capacity of the classification function by maximizing as large as the margin between training examples and the class boundary, optionally after removing some meaningless examples from the training data. The classification resulting function depends on called supporting patterns.

In [2] the setting of parameters in the support vector machines (SVMs) [8], is very important with regard to its accuracy and efficiency. In this paper, author employed the firefly algorithm in order to train all parameters of the SVM simultaneously, including the smoothness parameter, penalty parameter, and Lagrangian multiplier. The method is called the firefly-based SVM. This tool is not considered for the feature selection, because SVM, together with feature selection, is not at all suitable for the application in a multiclass classification, mainly for the one-against-all multiclass SVM. In experiments, multiclass and binary classifications are explored. The classification performance of firefly-SVM is compared with the novel LIBSVM method associated with the grid search method and the (PSO) particle swarm optimization based SVM. The experimental result shows the use of firefly-SVM [11] for pattern classifications to have maximum accuracy.

Conclusion drawn

(1) The firefly-SVM attempts simultaneously to train three kinds of parameters: Lagrangian multiplier, smoothness parameter, and penalty parameter. Results demonstrate that firefly-SVM is able of dealing with the applications of classification of pattern.

(2) The firefly-SVM training algorithm has much good performance than the other two methods in the binary classification, so it is promising to apply firefly-SVM to many other practical problems.

(3)The firefly-SVM can converge with the optimal solution within a constraint time when it associates with the feature selection just because of its high complexity.

In [3] author talked about the application of microarray data, as how to select a small number of genes from thousands of genes that can contribute to the occurrence of cancers is very important issue. Many researchers use various intelligence methods to analyse gene expression data [2]. Based on statistical analysis, author proposed method that outperforms other classifiers for all test datasets, and is compatible to SVM for certain datasets. The housekeeping genes along with various expression patterns and tissue-specific genes are identified. These genes give a high discrimination power on cancer classification.

In [4] Vapnik defined Support vector machine as a new learning machine for two groups classification problems. The machine gives the following idea: input vector are non-linearly mapped to a very high-dimension feature space.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Essential properties of the decision surface guarantees high generalization ability of the machine learning [13]. The idea behind the support-vector network was implemented for the restricted case where the training data can be separately without errors. High generalization ability of support-vector networks [4] that is utilizing polynomial input transformations was demonstrated. Vapnik [4] compared and resulted the performance of the support-vector network to many classical learning algorithms that took part in mark study of Optimal Character Recognition (OCR).

In [5] Data mining and knowledge discovery in databases have become a significant amount of research, media attention and industry. The article gives an overview of the emerging field, clarifying how data mining and in databases knowledge discovery are related to each other and related fields, such as statistics, machine learning and databases. The article mentions particular world applications, challenges mainly involved in real-world applications of knowledge discovery, specific data-mining techniques, and future research directions in the field.

In [6] a novel method for accurate detection of (ROIs) regions of interest that contain circumscribed lesions in mammograms. The mammograms were segmented using a statistical threshold and a number of regions were extracted. Therefore, a set of qualification criteria is employed to filter regions retaining a radial-basis function neural network that makes the final decision marking them as ROIs that contain abnormal tissue. The proposed technique finds the exact location of the circumscribed lessons with an accuracy value of 90.9%, and a very low number of false positive regions.

In [7] this study investigates the efficacy of applying support vector machines (SVM) to bankruptcy prediction problem. Although it is one of the known fact that the back-propagation neural network (BPN) performs well in pattern recognition tasks, hence the method has some limitations that it is an art to find an appropriate model structure and optimal solution. Since SVM captures geometric characteristics of feature space without deriving weights of networks from the training data, it is able of extracting the optimal solution with the small training set size. In this study, authors showed that the proposed classifier of SVM approach outperforms BPN to the problem of corporate bankruptcy prediction. The results demonstrated that the accuracy and generalization performance of SVM is better than that of BPN as the training set size gets smaller. Authors also examined the effect of the variability in performance with respect to various values of parameters in SVM. In addition, authors investigate and summarize the several superior points of the SVM algorithm compared with BPN.

In [8] Feature selection was applied to reduce the number of features in many applications where data has thousands or hundreds of features. Already existing feature selection methods mainly focus on finding relevant features. In this paper, authors showed that feature relevance alone is insufficient for efficient feature selection of high-dimensional data. They define feature redundancy and propose to perform explicit redundancy analysis in feature selection. A new framework was introduced that decouples relevance analysis and redundancy analysis. They also develop a correlation-based method for redundancy and relevance analysis, and conduct an empirical study for its efficiency and effectiveness in order for comparing with representative methods. In this paper [8], authors have identified the need for redundancy analysis in feature selection, given with a formal definition of feature redundancy, and investigated the co-relationship between feature relevance and redundancy. They have proposed a new framework of efficient feature selection via with relevance and redundancy analysis, and a correlation-based technique that uses *C*-correlation for relevance analysis and both *C*- and *F*-correlations for analysis of redundancy. A new feature selection algorithm FCBF was implemented and evaluated by extensive experiments comparing with feature selection algorithms. The feature selection results were further verified by two different learning algorithms. Author Lie and Huan method demonstrates its efficiency along with effectiveness for feature selection, mainly in supervised learning where data contains many irrelevant and redundant features.

In [9] the correct diagnosis of breast cancer is one of the major problems in the medical field. From the literature it has been understood that different pattern recognition techniques can easily help them to improve in the medical domain. These techniques can very much help doctors to form a second opinion and can take a better diagnosis. In this paper author present a unique improvement in neural network training basically for pattern classification. The proposed training algorithm was inspired by the meta-plasticity property of Shannon's information theory and neurons. During the training phase of the Artificial meta-plasticity Multilayer Perceptron (AMMLP) algorithm gives priority for

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

updating the weights to the less frequent activations over the more frequent ones. In this way meta-plasticity is modelled. While maintaining MLP performance, AMMLP achieved a more efficient training. AMMLP performance was tested using classification specificity, accuracy and sensitivity analysis, and confusion matrix. The obtained AMMLP classification accuracy of 99.26%, result compared to the Back-propagation Algorithm (BPA) and various recent classification techniques applied to the same database.

In [10] this paper, a new support vector machines (SVM) parameter tuning scheme that uses the fruit-fly optimization algorithm was proposed. Termed as FOA-SVM (fruit-fly optimization algorithm), the scheme was successfully applied to the field of medical diagnosis. The proposed FOA-SVM, the FOA technique efficiently and effectively addresses the parameter set in SVM. Four competitive counter parts are employed for comparison purposes, including the particle swarm optimization genetic algorithm-based SVM (GA-SVM), algorithm-based SVM (PSO-SVM), bacterial forging optimization-based SVM (BFO-SVM), and grid search technique-based SVM (Grid-SVM). The results demonstrate that the proposed FOA-SVM process may be obtain much more appropriate model parameters reduced the computational time, which generates high classification accuracy. Promisingly, the proposed method may be regarded as a useful clinical tool for medical decision making.

In [11] Recent statistics result that breast cancer [9], [12], [15] is one of a major cause of death among women in all over the world. Early diagnostic with computer aided diagnosis systems is a very important tool. This task is not easy because of poor ultrasound resolution and huge amount of patient data size. Then, initial image segmentation was one of the most challenging and important task. Among various methods for medical image segmentation, the use of entropy for maximization the information between the background and foreground is a well known and applied technique. A new kind of entropy, hence called non-extensive entropy, has been proposed in the literature for generalizing the Shannon entropy. In this paper, author proposed the use of non-extensive entropy, called q-entropy, applied in a CAD system for classification of breast cancer in ultrasound of mammographic exams. In order to validate this proposal, author have tested protocol in a data base of 250 breast ultrasound images. With a cross-validation protocol, demonstrate system's as accuracy, sensitivity, specificity, positive predictive value with the negative predictive value as: 95%, 97%, 94%, 92% and 98%, respectively, as ROC curves and A_z areas.

In [12] Correct diagnosis is one of the major problems in medical field. This includes the drawbacks of human expertise mainly in diagnosing the disease manually. As seen from the literature it has been identified that pattern classification techniques such as (rbf) radial basis function neural network (RBFNN) and support vector machines can help them to improve in this domain. SVM and RBFNN with their remarkable ability to derive meaning from complicated data, may be used to detect trends and extract patterns that are too complicated to be noticed by either some humans or other computer techniques. This paper shows the use of polynomial kernel of RBFNN and RBFNN in ascertaining the diagnostic accuracy of cytological data. Well Known sets of cytologically proven tumor data was used to train the models to categorize cancer patients according to their diagnosis. This research has demonstrated that RBFNN outperformed the polynomial kernel of SVM for correctly classifying the tumors.

III. PROPOSED ALGORITHM

The support vector machine [17] , initially started with a binary classification method developed by Vapnik (Vapnik 1995) at Bell laboratories. Sometime for a binary problem, we have training data points: $\{x_i, y_i\}$, i=1 …I, $y_i \in \{-1,1\}$, $x_i \in R^d$. Suppose we have some hyper-planes that classify the positive label from the negative labels separating with the help of hyper-plane. The points x which is on the hyper-plane satisfy w.x + b = 0, where w is normal to the hyper-plane, $|b|/\|w\|$ is the perpendicular distance measured from the hyper-plane to the origin, and $\|w\|$ is the Euclidean norm of w. Let d+ d- be the shortest path from the dividing hyper-plane to the closest positive or negative points. Define the margin of a dividing hyper-plane to be d+ + d-. For the binary classes to be linearly separable, the support vector algorithm looks for the dividing hyper-plane with the highest margin. This can be mathematically stated as follows:

Assume that all the training data that fulfil the following constraints:

$$x_i . w + b \geq +1 \text{ for } y_i = +1, \quad (1)$$

$$x_i . w + b \leq -1 \text{ for } y_i = -1, \quad (2)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Combining (1) and (2) together as a set of in-equalities results:

$$Y_i (x_i . w + b) -1 \geq 0 \quad \forall i \qquad (3)$$

Considering, the equality in equation (1) holds that there exist a point which is equivalent to choosing a value for w and b. These points are on the hyper-plane H1: xi ·w + b = 1 with normal w and perpendicular distance from the origin |1-b|/||w||. Similarly, the points for the equality in equation (2) hold to lie on the hyper-plane H2: xi · w + b = -1, with normal once again w and perpendicular distance measured from the origin |-1-b|/||w||. Hence d+ =d- =1/||w|| and the margin 2/||w||.
So, to find the solution for a difficult two dimensional case to have the form as shown.

$$w = \sum_i . \alpha_i y_i x_i \qquad (4)$$
$$\sum_i . \alpha_i y_i = 0. \qquad (5)$$

Support vector training, hence amounts to maximizing LD with respect to the αi, subject to the constraints defined in equation (5) and positivity to the αi, with solution given by given in equation (4). Now we have Lagrange multiplier αi for each and every training point. Those points from solution set, where αi > 0, are known as support vectors and therefore are lying on any of the hyper-planes H1, H2. All other training points have αi = 0 and lie either on H1 or H2 as earlier defined in the equality mainly in equation (3) holds, or on other side of H1 or H2 such that it is well defined with inequality in equation (3) [20].

For these kind of models, namely the support vectors are major component of the training set. They are located closely to the decision boundary, as if we remove all the remaining training points or moved them subjected to a predefined condition that they do not intersect H1 or H2, and training has repeated and consequently the same hyper-plane is generated then the mentioned algorithm for linearly separable data when applied for the non-separable data which does not guarantee a feasible solution.
H2: xi · w + b = -1, with normal again w and perpendicular distance from the origin measured |-1-b|/||w||. Hence d+ =d- =1/||w|| and the margin 2/||w||.

Support vector training (linearly separable) so it will amounts to maximizing LD with respect to the αi, subject to the constraints defined in equation (5) and positivity to the αi, with solution given by given in equation (4). Now we have Lagrange multiplier αi for all training point. Those points from solution set, where αi > 0, are called as support vectors (SVs) and therefore are lying on any of the hyper-planes H1, H2. All other training points have αi = 0 and will lie either on H1 or H2 as defined earlier also in the equality in equation (3) holds, or on other side of H1 or H2 such that it is defined inequality in equation (3) holds.

For these kind of models, it is mentioned that the support vectors are one of the major component of the training set. They are located closest to the decision boundary, if we remove rest all the remaining training points or moved them subjected to a condition that they do not intersect H1 or H2, and training has repeated and consequently the same hyper-plane is generated then the given algorithm for linearly separable data when applied for the non-separable data that always does not guarantee a feasible solution [22].

IV. EXPERIMENTAL SETUP

In order to evaluate the multi-variant Support Vector Machine, medical dataset was used from the UCI machine learning data repository, including Breast Cancer Wisconsin (Original) dataset. Table 1 contains detailed descriptions of the dataset. Breast Cancer Wisconsin dataset contains missing values, and the missing categorical attributes are replace with the mode of the attributes.

The breast Cancer dataset consists of 699 instances, each record has ten attributes. So, the class has a distribution of 458(68.46%) benign samples and 241(36.02%) malignant samples.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 1. Description of the five medical datasets used in the experiments.

No.	Datasets	# of classes	# of Instances	# of features	Missing Values.
1.	Breast Cancer Wisconsin(original)	2	699	10	Yes

Detailed description of Medical Datasets.

Table 2 Using Five cross validation on different attributes of Breast Cancer Dataset

Attribute	Sensitivity	Specificity	Error Rate	AUC
[2,3]	0.95196506	0.94605809	0.050071530	0.9499
[2,3,4]	0.95633187	0.95020746	0.045779685	0.9542
[2,3,4,5]	0.94759825	0.96680497	0.045779685	0.9542
[2,3,4,5,6]	0.94759825	0.97510373	0.042918454	0.9571
[2,3,4,5,6,7]	0.94323144	0.97510373	0.045779685	0.9542
[2,3,4,5,6,7,8]	0.93668122	0.97925311	0.048640915	0.9514
[2,3,4,5,6,7,8,9]	0.95196506	0.95435684	0.047210300	0.9528
[2,3,4,5,6,7,8,9,10]	0.92139737	0.99170124	0.054363376	0.9456

Breast Cancer Dataset with attributes on five cross validation

Table 3 Using Ten cross validation on different attributes of Breast Cancer Dataset

Attribute	Sensitivity	Specificity	Error Rate	AUC
[2,3]	0.941048034	0.96265560	0.051502145	0.9485
[2,3,4]	0.94541484	0.97095435	0.045779685	0.9542
[2,3,4,5]	0.94759825	0.96680497	0.045779685	0.9542
[2,3,4,5,6]	0.94759825	0.97095435	0.044349070	0.9557
[2,3,4,5,6,7]	0.94541484	0.97510373	0.044349070	0.9557
[2,3,4,5,6,7,8]	0.93886462	0.97510373	0.048640915	0.9514
[2,3,4,5,6,7,8,9]	0.93231441	0.98755186	0.048640915	0.9514
[2,3,4,5,6,7,8,9,10]	0.92576419	0.99170124	0.051502145	0.9485

Breast Cancer with attributes on ten cross validation

Table 4 Using Fifteen cross validation on different attributes of Breast Cancer Dataset

Attribute	Sensitivity	Specificity	Error Rate	AUC
[2,3]	0.94104803	0.96265560	0.05150214	0.9485
[2,3,4]	0.94759825	0.96680497	0.04577968	0.9542
[2,3,4,5]	0.94759825	0.96265560	0.04721030	0.9528
[2,3,4,5,6]	0.94541484	0.97925311	0.04291845	0.9571
[2,3,4,5,6,7]	0.94323144	0.97510373	0.04577968	0.9542
[2,3,4,5,6,7,8]	0.93668122	0.97510373	0.05007153	0.9499
[2,3,4,5,6,7,8,9]	0.93231441	0.98755186	0.04864091	0.9514
[2,3,4,5,6,7,8,9,10]	0.92576419	0.99179124	0.05150214	0.9485

Breast Cancer with attributes on fifteen cross validation

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 5 Using Twenty cross validation on different attributes of Breast Cancer Dataset

Attribute	Sensitivity	Specificity	Error Rate	AUC
[2,3]	0.94104803	0.96265560	0.05150214	0.9485
[2,3,4]	0.94759825	0.97095435	0.04434907	0.9557
[2,3,4,5]	0.94759825	0.96265560	0.04721030	0.9528
[2,3,4,5,6]	0.94541484	0.97095435	0.04577968	0.9542
[2,3,4,5,6,7]	0.94323144	0.97925311	0.04434907	0.9557
[2,3,4,5,6,7,8]	0.93668122	0.97510373	0.05007153	0.9499
[2,3,4,5,6,7,8,9]	0.95196506	0.94605809	0.05007153	0.9499
[2,3,4,5,6,7,8,9,10]	0.92576419	0.99170124	0.05150214	0.9485

Breast Cancer with attributes on twenty cross validation

V. SIMULATION RESULTS

As in **Breast Cancer** dataset, the classifier gives the best sensitivity 0.95759825 with attribute [2,3,4,5], [2,3,4,5,6] are selected for training and testing the machine for 5 cross validation as shown in Table 2, for 10 cross validation best sensitivity achieved is 0.94759825 with attribute [2,3,4,5], [2,3,4,5,6] as shown in Table 3, for 15 cross validation best sensitivity achieved is 0.94759825 with attribute [2,3,4], [2,3,4,5] as shown in Table 4, for 20 cross validation best sensitivity achieved is 0.95196506 with attribute [2,3,4,5,6,7,8,9] as shown in Table 5. The best specificity is achieved 0.99170124 with attribute [2,3,4,5,6,7,8,9,10] for 5 cross validation as shown in Table 2, 0.99170124 with attribute [2,3,4,5,6,7,8,9,10] for 10 cross validation as shown in Table 3, 0.99170124 with attribute [2,3,4,5,6,7,8,9,10] for 15 cross validation as shown in Table 4, 0.99170124 with attribute [2,3,4,5,6,7,8,9,10] for 10 cross validation as shown in Table 5. The best accuracy is achieved 0.9571 with attribute [2,3,4,5,6] for 5 cross validation as shown in Table 2 and 0.9557 with attributes [2,3,4,5,6] and [2,3,4,5,6,7] for 10 cross validation as shown in Table 3, 0.9571 with attributes [2,3,4,5,6] for 15 cross validation as shown in Table 4, 0.9557 with attributes [2,3,4] and [2,3,4,5,6,7] for 20 cross validation as shown in Table 5.

Table 6. Summarized result of 5, 10, 15 and 20 cross validation in Breast Cancer Dataset

Types of Cross Validation	5-Cross Validation	10-Cross Validation	15-Cross Validation	20-Cross Validation
Accuracy	0.9571	0.9557	0.9571	0.9557
Error Rate	0.042918454	0.044349070	0.042918454	0.044349070
Sensitivity	0.95633187	0.94759825	0.94759825	0.95196506
Specificity	0.99170124	0.99170124	0.99179124	0.99179124

VI. CONCLUSION AND FUTURE WORK

This paper describes the potency of SVM in the field of computational biology for which SVM are known to perform well as compared to other machine learning or statistical methods. 20-fold, 15-fold, 10-fold and 5-fold cross validations are approaches for solving medical data classification problems. In this paper, I explore the use of 20-fold, 15-fold, 10-fold and 5-fold cross validations for classification. Based on the results of the current datasets of the UCI database, I conclude that various selection of features results different accuracy. This indicates that the proposed 20-

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

fold, 15-fold, 10-fold and 5-fold cross validation methods help in selecting those variables that will result in highest accuracy. The result may be much better for larger set of real data.

REFERENCES

1. B.E, I.M, V.N. A Training Algorithm for Optimal Margin Classifiers, ACM, New York, NY, USA, 1992.
2. Chen, Wang, Tsai, Wang, Adrian, Cheng, Yang, Teng, Tan, & Chang. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics, 2014.
3. Chao, Horng. The Construction of support vector machine classifier using the firefly algorithm. Comput Intell Neurosci, 2005.
4. Cortes, Vapnik. Support -Vector Networks. Mach. Learn. 20 (3) Springer. 273–297, 1995.
5. Fayyad, Piatetsky-Shapiro, Smyth. From Data Mining to knowledge Discovery in Database - AI magazine. aaai.org, 1996.
6. Ioanna, Evangelos , George. Automatic detection of abnormal tissue in mammography ICIP (2) 877-880, 2001.
7. K.S, T.S, H. An application of support vector machines in bankruptcy prediction model, Expert Syst. Appl. 127–135, 2005.
8. Lei, Huan. Efficient Feature Selection via Analysis of Relevance and Redundancy The Journal of Machine Learning Research Volume 5, Pages 1205-1224, 2004.
9. Marcano-Cedeño, Quintanilla-Domínguez, D. WBCD breast cancer database classification P.S,applying artificial metaplasticity neural Network Expert Systems with Applications 9573-9579, 2011.
10. Shen, Chen, Yu, Kang, Zhang, Li, Yang, Liu. Evolving support vector machines using fruit fly optimization for medical data classification. Science Direct, 2016.
11. R, & J.S. Non-Extensive Entropy for CAD Systems of Breast Cancer Images. Pp 121-128, 2006.
12. T. S, Vennila, & S. Breast mass classification based on cytological patterns using RBFNN and SVM Expert Syst. Appl. 36(3): 5284-5290, 2009.
13. Tzanis, Berberidis, Vlahavas. Machine Learning and Data Mining in Bioinformatics, 2012.
14. Vapnik. The Nature of Statistical Learning Theory, Springer, New York, 1995.
15. Zhang, Schwartz, Wagner, Miller. A greedy algorithm for aligning DNA sequences J Comput Biol. 203-14, 2000.
16. Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory – COLT '92.
17. Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" Journal of Machine Learning Research, 2: 125–137, 2001.
18. Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P."Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing. New York: Cambridge University Press, 2007.
19. Aizerman, Mark A.; Braverman, Emmanuel M.; and Rozonoer, Lev I. "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control, 1964.
20. Meyer, D.; Leisch, F.; Hornik, K. "The support vector machine under test". Neurocomputing 55: 169, 2003.
21. Hsu, Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen . A Practical Guide to Support Vector Classification (Technical report). Department of Computer Science and Information Engineering, National Taiwan University, 2003.
22. Hsu, Chih-Wei; and Lin, Chih-Jen . "A Comparison of Methods for Multiclass Support Vector Machines". IEEE Transactions on Neural Networks, 2002.
23. R.-E. Fan; K.-W. Chang; C.-J. Hsieh; X.-R. Wang; C.-J. Lin. "LIBLINEAR: A library for large linear classification". Journal of Machine Learning Research, 2008.
24. Monika Lamba; "Variable Features Selection for classification of Medical Data Using SVM". International Journal of Engineering Technology, Management and Applied Sciences, Volume 4, Issue 5, ISSN 2349-4476, May 2016.

BIOGRAPHY

Monika Lamba has received her B.Tech degree in Computer Science and Engineering from Ansal Institute of Technology (AIT), Gurgaon affiliated to Guru Gobind Singh Indraprastha University (GGSIPU), Delhi in 2014 and pursuing M.Tech degree in Computer Science and Engineering in College University School of Information and Communication Technology (USICT), Dwarka affiliated to GGSIPU, Delhi in 2014-2016.