An Approach with Support Vector Machine using Variable Features Selection on Breast Cancer Prognosis

Sandeep Chaurasia
Department of Computer Science & Engineering,
Sir Padampat Singhania University,
Udaipur, India

Dr. P Chakrabarti
Department of Computer Science & Engineering,
Sir Padampat Singhania University,
Udaipur, India

***Abstract*—Cancer diagnosis and clinical outcome prediction are among the most important emerging applications of machine learning. In this paper we have used an approach by using support vector machine classifier to construct a model that is useful for the breast cancer survivability prediction. We have used both 5 cross and 10 cross validation of variable selection on input feature vectors and the performance measurement through bio-learning class performance while measuring AUC, specificity and sensitivity. The performance of the SVM is much better than the other machine learning classifier.**

Keywords—Breast cancer; feature selection; Support vectors; Support Vector Machine; Wisconsin Breast Cancer Dataset.

I. INTRODUCTION

A major category of problems in medical science deals with the diagnosis of disease, based upon various tests performed upon the patient. For this reason the use of classifier systems in medical diagnosis is gradually increasing. There is no doubt that evaluation of data taken from patients and decisions of experts are the most important factors in diagnosis. But the different artificial intelligence techniques for classification also help experts a great deal. Classification systems, minimizing possible errors that might be made because of fatigued or inexperienced experts, provide more detailed medical data for examination in a shorter time.

The importance of patterns classification of breast cancer is a major real world medical problem. Breast cancer has become one of the major causes of mortality around the world and research into cancer diagnosis and treatment has become an important issue for the scientific community. The etiologist of breast cancer remain unclear and no single dominant cause has emerged. [2][3]

Prevention is still a mystery and the only way to help patients survive is by early detection. If the cancerous cells are detected before spreading to other organs, the survival rate for patients is more than 97%. [4]

II. BACKGROUND MATERIAL

There are many applications for Machine Learning (ML) of which the most significant is data mining and pattern classification. Major areas of ML where it can often be successfully applied for classification and regression problems by improving the efficiency and design of the systems. Every instance or attribute in any of the dataset used by the machine learning algorithms is represented using the same set of features. The features may be of different dimension, if instances are given with known labels with corresponding correct outputs then this type of learning is called supervised learning, where as unsupervised learning, the instances are unlabeled or the outputs are unknown. Another kind of machine learning is reinforcement learning where the training information is provided to the learning system by the external teacher is in the form of a scalar reinforcement signal that constitutes a measure of how well the system operates. The learner is not instructed to take any desired actions, but rather discovering which actions yield the best solution, by continuously trying each action to improve the efficiency.

A. Supervised Learning Algorithms

Machine learning is the process of learning a set of rules from instance from a training set, or more generalizing, creating a classifier that can be used to generalize from new instances. The procedures or learning is as follow; the first step is to collect the dataset, if a dataset collected by any of the arbitrary method is not directly suitable for induction. It may contain noisy and missing data values, and therefore requires significant pre-processing [5]. The second step is the data preparation and data pre-processing and the feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible [6]. This reduces the dimension of the data and allowing algorithms to perform faster and more efficiently. But many features depend on one another and may influence the accuracy of supervised Machine Learning classification models.

B. Algorithm selection

Specifically the selection of learning algorithm is a critical procedure. Once at preliminary stage when testing is judged and it comes out satisfactory, then the classifier is generalized [11]. The accuracy of the classifier's evaluation is typically often based on prediction (the ratio of correct prediction over the total number of predictions). There are many techniques are available to calculate the classifier's accuracy. One way is to split the training set by dividing the two-thirds for training and rest for estimating performance. Another method is known as cross-validation in which the training set is divided into mutually exclusively equally-sized subsets and for every subset the classifier is trained on the union of remaining

subsets. The average error rate of each subset results an estimate of the error rate for the classifier. If the error (%) in not tolerable then the algorithm go back to the previous stage of the supervised ML process. There has been research on medical diagnosis of breast cancer with WBCD using Artificial Neural Networks (ANNs) in literature, and most has reported high classification accuracy.

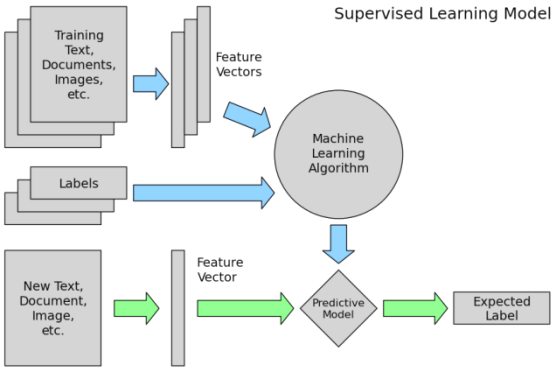

Fig. 1. Supervised learning model

The support vector machine (SVM) algorithm [8] is a classification algorithm that provides the best performance in various application domains such as object recognition, speaker identification, face detection and other classifications problems. Two main motivations to use SVMs in the field of computational biology first, many problems have high dimensional as well as noisy data, for which SVM are known to perform well as compared to other statistical or machine learning methods. Second, in contrast to most machine learning methods, kernel methods like the SVM can easily handle non-vector inputs, such as variable length sequences or graphs. These types of data are common in biology applications.

III. METHODOLOGY

A. Support vector machines

The support vector machine is originally a binary classification method developed by Vapnik et.al at Bell laboratories [9]. For a binary problem, we have training data points:$\{x_i y_i\}, i = 1 \ldots l$, $y_i \epsilon \{-1, 1\}, x_i \epsilon R^d$. Suppose we have some hyperplane that separates or classify the positive label from the negative labels with a separating hyperplane. The points x which is on the hyperplane satisfy $w \cdot x + b = 0$, where w is normal to the hyperplane, $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin, and $\|w\|$ is the Euclidean norm of w. Let d+ d- be the shortest distance from the separating hyperplane to the closest positive or negative points. Define the margin of a separating hyperplane to be d+ + d-. For the linearly separable classes, the support vector algorithm simply looks for the separating hyperplane with the biggest margin. This can be mathematically stated as follows: assume that all the training data satisfy the following constraints:

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

Combining (1) and (2) into one set of inequalities results:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall i \quad (3)$$

TABLE I. CONTRIBUTION IN MACHINE LEARNING

Researcher (Years)	Accuracy	Method
Quinlan (1996)	94.74%	C4.5 decision tree method
Hamiton, Shan, and Cercone (1996)	94.99%	RIAC method
Ster and Dobnikar (1996)	96.8%	linear discreet analysis method
Nauck and Kruse (1999)	95.06%	neuron-fuzzy techniques
Pena-Reyes and Sipper (1999)	97.36%	fuzzy-GA method
In Setiono (2000)	98.10%	Feed forward neural network rule extraction algorithm.
Albrecht, Lappas, Vinterbo, Wong, and Ohno-Machado (2002)	98.8%	Logarithmic simulated annealing with the perceptron algorithm
Abonyi and Szeifert (2003)	95.57%	supervised fuzzy clustering technique

Now considering the equality in equation (1) holds that require that there exist a point which is equivalent to choosing a value for w and b. These points are on the hyperplane H1: $xi \cdot w + b = 1$ with normal w and perpendicular distance from the origin $|1 - b|/\|w\|$ .Similarly the points for the equality in equation (2) holds to lie on the hyperplane H2: $xi \cdot w + b = -1$, with normal again w and perpendicular distance from the origin $|-1-b|/\|w\|$. Hence d+ = d- $= 1/\|w\|$ and the margin $2/\|w\|$

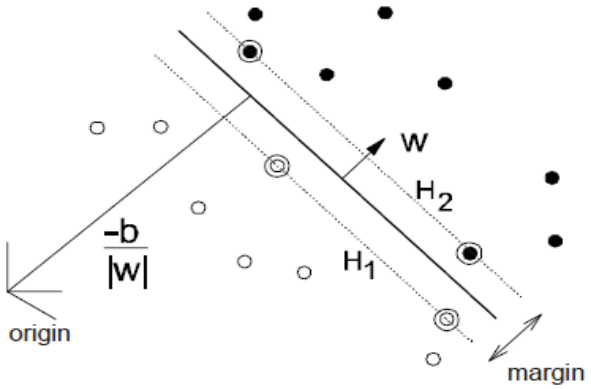

Fig. 2. Linear separating hyperplanes for the separable case. The support vectors are circled.

From Fig 2 it has observed that H1 and H2 are parallel they have the same normal vector and that no training points fall between H1 and H2. So we can find the pair of hyperplanes which maximize the margin by minimizing $\|w\|2$, subject to constraints defined in equation (3). Thus to find the solution for a typical two dimensional case to have the form shown on Fig.2. We have to introduce non-negative Lagrange multipliers αi, where i = 1,…l for each one of the inequality

constraints in equation (3). As defined above the rule is that for constraints of the form ci ≥ 0, the constraint equations are multiply by the non-negative Lagrange multipliers and get subtracted by the objective function, to form the Lagrangian. For equality constraints, the Lagrange multipliers are unconstrained [12]. This gives Lagrangian:

$$L_p \equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i . w + b) + \sum_{i=1}^{l} \alpha_i, \quad (4)$$

We must now minimize Lp with respect to w and b, and maximize with respect to all αi simultaneously, all are subject to the constraints αi ≥ 0 as set of constraints named C1. We get a convex quadratic programming problem, as the objective function is also convex, and that points which are satisfying the constraints also generate a convex set. This concludes that we can also solve the following dual problem to maximize LP, subject to the constraints that the gradient of LP with respect to w and b vanish, and subject to the constraints that the αi ≥ 0 as a set of constraints named C2. This particular dual formulation of the problem is called the Wolfe dual [10]. It has the property that the maximize LP, subject to constraints C2, occurs for the same values of the w, b and α, as the minimize LP, subject to constraints C1. Requiring that the gradient of *LP* with respect to *w* and *b* vanish gives the conditions:

$$w = \sum_i \alpha_i y_i x_i , \quad (5)$$
$$\sum_i \alpha_i y_i = 0 . \quad (6)$$

As these are the equality constraints in the dual formulation, we can substitute them into equation (4) to give

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j . y_i y_j x_i . x_j \quad (7)$$

Now we have provided the Lagrangian with different labels P for primal and D for dual to emphasize that the two formulations are different: LP and LD generated from the same objective function but with different constraints, and the solution is obtained by minimizing LP or by maximizing LD respectively. Also note that if we formulate the problem with b = 0, which constitute that all hyperplanes passes through the origin, the constraint defined in equation 6 does not needed. This is a soft restriction for high dimensional spaces, and therefore it amounts to reduce the number of degrees of freedom by one.

Support vector training (linearly separable) therefore amounts to maximizing LD with respect to the αi, subject to the constraints defined in equation (6) and positivity to the αi, with solution given by given in equation (5). Now we have Lagrange multiplier αi for the every training point. Those points from solution set where αi > 0 are known as support vectors and therefore lying on any of the hyperplanes H1, H2. All other training points have αi = 0 and lie either on H1 or H2 as earlier defined 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 models the support vectors are major component of the training set. They are located nearest to the decision boundary, if we remove all the remaining training points or moved them around subjected to a condition that they do not cross H1 or H2, and training has repeated and consequently the same hyperplane is generated then the above algorithm for linearly separable data when applied for the non-separable data does not guarantee a feasible solution.

This will justify that using the objective function as dual Lagrangian that grows arbitrarily large. How we classify the non-separable data. To achieve this first we have to relax the constrained defined in equation (1) and equation (2) and for this we have to introduce positive slack variables e_i; i = 1,..., l, in the constraints, which then become:

$$x_i . w + b \geq +1 - e_i \text{ for } y_i = +1, \quad (8)$$
$$x_i . w + b \leq -1 + e_i \text{ for } y_i = -1, \quad (9)$$
$$e_i \geq 0 \forall i \quad (10)$$

If an error is occur then the corresponding e_i must exceed unity, so Σi e_i is an upper bound on the number of training errors. So to assign an extra cost for the errors is to change the objective function, it should be minimized from ||w||2/2 to ||w||2/2 + C(Σi e_i), where C is a parameter that has decided by the user for the large value of C correspond to high rate to errors. We have to generalized the above method to the case where sign (f(x)) represents the class (f(x) is a decision function) assigned to data point x is not a linear function of the data. The only approach is that we need to assure that the data appears in the training problem, is in the form of dot products of xi · xj. Now we first mapped the data to some other dimension such as Euclidean space H, using a mapping here we call as Φ:

$$\Phi: R^d \rightarrowtail H , \quad (11)$$

Then consequently the training algorithm would only depend on the data through dot products in H, i.e. on functions of the form $\Phi(x_i) \cdot \Phi(x_j)$. Now introducing the concept of the kernel function K such that $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, then only K is used in the training algorithm and we are not considering the value what Φ is. The kernel function has to satisfy Mercer's condition. One example for this function is Gaussian:

$$K(x_i , x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad (12)$$

In this particular example, H is infinite dimensional (Euclidean space), so it is not easy to work with Φ explicitly. However, if one replaces $x_i \cdot x_j$ by K (x_i, x_j) everywhere in the training algorithm, the algorithm will generate a support vector machine which lives in an infinite dimension space.

Now considering all of the previous section, since we are doing a linear separation, but in a different plane. To implement this model we need w, and that reside in H and in test phase an SVM is used by computing dot products of a given test point x with w, or more specifically by computing the sign of equation as stated below

$$f(x) \equiv \sum_{i=1}^{N_s} \alpha_i y_i \Phi(s_i). \Phi(x) + b = \sum_{i=1}^{N_s} \alpha_i y_i K(s_i, x) + b , \quad (13)$$

Where s_i are the support vectors and we can avoid computing Φ(x) by the use of $K(s_i, x) = \Phi(s_i) \cdot \Phi(x)$.

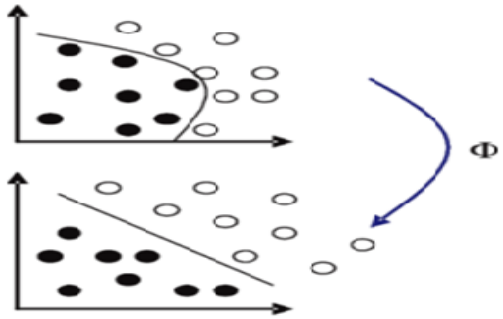

Fig. 3. General principle of SVM: projection of data in an optimal dimensional space.

IV. Experimental Setup

A. Breast Cancer Dataset

In this study, the Wisconsin Breast Cancer Database an UCI Machine Learning Repository was analysed. The WBCD dataset consists of 699 instances taken from Fine Needle Aspirates (FNA) of human breast tissue. Each record in the database has nine attribute.

TABLE II. Attributes Of The Simple Dataset

Attribute	Domain
1. Sample code number	id number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	2 - benign, 4- malignant

Attributes 2 through 10 have been used to represent instance. The nine attributes are detailed in Table 2. The measurements are assigned an integer value between 1 and 10, with 1 being the closest to benign and 10 to malignant. Associated with each sample is its class label, which is either benign or malignant. This dataset contains 16 instances with missing attribute values. Since many classification algorithms have discarded these data samples, for the ease of comparison, the same method is followed and the remaining 683 samples are taken for use. Therefore, the class has a distribution of 444 (65.0%) benign samples and 239 (35.0%) malignant samples.

B. Experimental Setup

The original data is present in the form of analogue values with values ranging from 0-10 [13]. The data are converted to their equivalent integer form. Scaling is required to map the dataset into desired range of variable ranging between minimum and maximum range of network input. Based on total number of attribute (assume N). N-1 will be numeric feature and 1 is class category. The numerical attributes are ranging in between 0 and 1 the new value obtained are converted into binary form by the following scaling grouping is done on the basis of range [0, x) ='0' and [x, 10] ='1'

These attributes are fed into the variable feature selection for training and testing to obtain the result for 10 and 5 cross fold validation to compute the performance of the support vector machine classifier. We have simulated the result using the kernel function as the radial basis function (rbf kernel).

TABLE III. Performance for the SVM classifier using variable feature selection with 10 and 5 cross validation

Atrribute	Sensitivity	Specificity	ErrorRate	AUC
[2,3]	0.945945	0.945606	0.054172	0.956400
[2,3,4]	0.959459	0.949791	0.043924	0.988900
[2,3,4,5]	0.952703	0.966527	0.042460	0.955500
[2,3,4,5,6]	0.950451	0.966527	0.043924	0.955600
[2,3,4,5,6,7]	0.954954	0.987447	0.033675	0.988600
[2,3,4,5,6,7,8]	0.941441	0.987948	0.042460	0.955600
[2,3,4,5,6,7,8,9]	0.936937	0.991631	0.043400	0.966700
[2,3,4,5,6,7,8,9,10]	0.930180	0.987447	0.049700	0.945833
10 cross validation				
Atrribute	Sensitivity	Specificity	ErrorRate	AUC
[2,3]	0.959641	0.941176	0.046783	0.966700
[2,3,4]	0.954954	0.932773	0.052780	0.977300
[2,3,4,5]	0.954751	0.983193	0.035294	0.965900
[2,3,4,5,6]	0.936937	0.983333	0.046780	0.977800
[2,3,4,5,6,7]	0.959641	0.991667	0.029155	0.988900
[2,3,4,5,6,7,8]	0.932126	0.975000	0.052786	0.955600
[2,3,4,5,6,7,8,9]	0.940909	1.000000	0.038340	0.965900
[2,3,4,5,6,7,8,9,10]	0.919280	1.000000	0.052631	0.977800
5 cross validation				

C. Result

The result obtained using the support vector machine classifier by selecting variable attribute selection. As shown in table 3 the classifier gives the best sensitivity with 0.9886 0.9889 with attribute A [2, 3, 4, 5, 6, 7] are selected for training and testing the machine for each 10 and 5 cross validation respectively. The best specificity is achieved when attribute A [2, 3, 4, 5, 6, 7, 8, 9] are selected for training and testing with 0.99163 and 1.00 respectively. The lowest error rate and the best AUC are obtained with A [2, 3, 4, 5, 6, 7]. The accuracy is the proportion of the total number of predictions that were correct. The best accuracy is evaluated when we considered the attribute A [2, 3, 4, 5, 6, 7] is 96.4%

and for the remaining selection of attributes the accuracy lies between in a range of 95. 09 to 95. 7.

V. CONCLUSION

This paper describes the potency of SVMs in the field of computational biology for which SVM are known to perform well as compared to other statistical or machine learning methods. After a better understating of the strengths of each method it has been observed that the results are generated on the basis of AUC, sensitivity and specificity. The accuracy of support vector machine is far better as compared with other machine learning classifier. The result may be much better for the larger set of real data.

REFERENCES

[1] T. S. Subashini, Vennila Ramalingam, S. Palanivel Breast mass classification based on cytological patterns using RBFNN and SVM Expert Syst. Appl. 36(3): 5284-5290 (2009).

[2] Ioanna Christoyianni, Evangelos Dermatas, George K. Kokkinakis Automatic detection of abnormal tissue in mammography ICIP (2) 2001: 877-880

[3] P.S. Rodrigues, R. Chang, and J.S. Suri, "Non-Extensive Entropy for CAD Systems of Breast Cancer Images", ;in Proc. SIBGRAPI, 2006, pp.121-128.

[4] American Cancer Society Hompage, http://www.cancer.org/ 2008

[5] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences J Comput Biol. 2000 Feb-Apr;7(1-2):203-14.

[6] Lei Yu, Huan Liu Efficient Feature Selection via Analysis of Relevance and Redundancy The Journal of Machine Learning Research Volume 5, 12/1/2004 Pages 1205-1224

[7] AstroML: Machine Learning and Data Mining for Astronomy http://www.astroml.org/

[8] Boser, B. E., Guyon, I. M., and Vapnik, V. A training algorithm for optimum margin classifiers. In Fifth Annual Workshop on Computational Learning Theory, Pittsburgh. ACM. 1992

[9] V. Vapnik. The nature of statistical learning Theory, 2nd Ed. Springer, NewYork, 1999.][C. J .C. Burges. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Disc. 2 (1998) 121.

[10] R. Fletcher. Practical methods of optimization. 2nd Ed. Wiley & Sons. Chichester (1987)

[11] Marcano-Cedeño J. Quintanilla-Domínguez, D. Andina WBCD breast cancer database classification applying artificial metaplasticity neural network Expert Systems with Applications 38 (2011) 9573–9579

[12] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev A Multi-class SVM Classifier Utilizing Binary Decision Tree. Informatica 33 (2009) 233-241 233

[13] Sandeep Chaurasia, Prasun Chakrabarti, Neha Chourasia An Application of Classification Techniques on Breast Cancer Prognosis"International Journal of Computer Applications (0975 – 8887) Volume 59– No.3, page 6-10, December 2012.