Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine

Ravi Sanakal[#1], Smt. T Jayakumari[*2]

[1]M.Tech, Dept of Computer Science and Engineering, BTL Institute of Technology, Bangalore, India

[2]Assistant Professor, Dept of Computer Science and Engineering, BTL Institute of Technology, Bangalore, India

***Abstract*-- Clinical decision-making needs available information to be the guidance for physicians. Nowadays, data mining method is applied in medical research in order to analyze large volume of medical data. This study attempts to use data mining method to analyze the databank of Diabetes disease and diagnose the Diabetes disease. This study involves the implementation of FCM and SVM and testing it on a set of medical data related to diabetes diagnosis problem. The medical data is taken from UCI repository, consists of 9 input attributes related to clinical diagnosis of diabetes, and one output attribute which indicates whether the patient is diagnosed with the diabetes or not. The whole data set consists of 768 cases.**

Keywords: **Data Mining, Diabetes, Fuzzy C Means, Support Vector Machine, Sequential Minimal Optimization, Classification.**

1. INTRODUCTION

Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. One of the useful applications in the field of medicine is the incurable chronic disease diabetes. Data Mining algorithm is used for testing the accuracy in predicting diabetic status.

Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called training set, which contains the same set of attributes, except for the class label – not yet known. The algorithm analyses the input and produces a prediction.

2. REALTED WORK

Pardha Repalli [2], in their research work predicted how likely the people with diverse age groups are affected by diabetes based on their activities. They also found out factors responsible for the individual to be diabetic. Statistics given by the Centers for Disease Control states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modeling has 50784 records with 37 variables. They computed a new variable age_new as nominal variable, dividing in to three group's young age, middle age and old age and the target variable diabetes_diag_binary is a binary variable. They found 34% of the population whose age was below 20 years was not affected by diabetes. 33.9% of the population whose age was above 20 and below 45 years was not affected by diabetes. 26.8% of the population whose age was above 45 years was not diabetic.

Joseph L. Breault [3], in his research work used the publicly available Pima Indian diabetic database (PIDD) at the UCIrvine Machine Learning Lab. They tested data mining algorithms to predict their accuracy in predicting diabetic status from the 8 variables given. Out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. Rough sets as a data mining predictive tool applied rough sets to PIDD using ROSETTA software. The test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI .The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81%. Using a group of 10 random samples the mean accuracy was 73.2%.

G. Parthiban et al. [4]. The main objective of their research paper is to predict the chances of diabetic patient getting heart disease. In their study, they applied Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. They proposed a system which predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. It is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. The data set used in their work was clinical data set collected from one of the leading diabetic research institute in Chennai and contain records of about 500 patients. The clinical data set specification provides concise, unambiguous definition for items related to diabetes.

The WEKA tool was used for Data mining. They used 10 fold cross validation. They found most of the diabetic patients with high cholesterol values are in the age group of 45 – 55, have a body weight in the range of 60 – 71, have BP value of 148 or 230, have a Fasting value in the range of 102 – 135, have a PP value in the range of 88 – 107, and have a A1C value in the range of 7.7 – 9.6.

Padmaja et al. [5] in their research aimed at finding out the characteristics that determine the presence of diabetes and

to track the maximum number of women suffering from diabetes. They used Data mining functionalities like clustering and attribute oriented induction techniques to track the characteristics of the women suffering from diabetes. Information related to the study was obtained from National Institute of Diabetes, Digestive and Kidney Diseases. The results were presented in the form of clusters. Those clusters denote the concentrations of the various attributes and the percentage of women suffering from diabetes. The results were evaluated in five different clusters and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3. The study predicts the state of diabetes i.e., whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering from diabetes with specific characteristics. This is used to effectively in diagnosis and treatment.

3. DIABETES

Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes, your body either doesn't make enough insulin or cannot use its own insulin as well as it should. This causes sugar to build up in your blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death.

General Symptoms of Diabetes

- Increased thirst
- Increased urination - Weight loss
- Increased appetite - Fatigue
- Nausea and/or vomiting - Blurred vision
- Slow-healing infections - Impotence in men

Types of Diabetes

Type 1 - Diabetes also called as Insulin Dependent Diabetes Mellitus (IDDM), or Juvenile Onset Diabetes Mellitus is commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itself therefore; pancreas does not produce the hormone insulin. The body does not properly metabolize food resulting in high blood sugar (glucose) and the patient must rely on insulin shots. Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II - Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus.

Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the 40.India has the dubious distinction of being the diabetic capital of the world. Home to around 33 million people with diabetes, 19% of the world's diabetic population is from India. Nearly 12.5% of Indian's urban populations have diabetes. The number is expected to escalate to an alarming 80 million by the year 2030. Amongst the chronic diabetic complications, diabetic foot is the most devastating result. Over 50,000 leg amputations take place every year due to diabetes in India. Diabetes patients can often experience loss of sensation in their feet. Even the smallest injury can cause infection that can be various serious. 15% of patients with diabetes will develop foot ulcers due to nerve damage and reduced blood flow. Diabetes slowly steals the person's vision. It is the cause for common blindness and cataracts.

4. DATASET

The ability to diagnose diabetes early plays an important role for the patient's treatment process. The World Health Organization proposed the nine attributes, depicted below in Table 1, of physiological measurements and medical test results for the diabetes diagnosis.

TABLE 1

Sl.No	Attribute
1	Number of times pregnant
2	Plasma glucose concentration in an oral glucose tolerance test
3	Diastolic blood pressure (mm/Hg)
4	Triceps skin fold thickness (mm)
5	2-hour serum insulin (μU/ml)
6	Body mass index (kg/m^2)
7	Diabetes Pedigree function
8	Age (years)
9	Status (0-Healthy, 1-Diabetes)

The dataset [1], originally donated by Vincent Sigillito from the Applied Physics Laboratory at the Johns Hopkins University, is one of the most well-known datasets for testing classification algorithms. This dataset consists of records describing 768 female patients of Pima Indian heritage which are at least 21 years old living near Phoenix, Arizona, USA. From the 768 patients in the PID dataset, classification algorithms used a training set with 576 patients and a testing dataset with 192 patients.

5. ALGORITHMS & FLOW DIAGRAM

5.1 Fuzzy C-means clustering (FCM)

Fuzzy C-means clustering (FCM), relies on the basic idea of K-Means, with the difference that in FCM each data point belongs to a cluster to a degree of membership grade, while in K-Means every data point either belongs to a certain cluster or not. So FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. However, FCM still uses a cost function that is to be minimized while trying to partition the data set. The membership matrix U is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity:

$$\sum_{i=1}^{c} Uij = 1 \quad \text{J=1…..n} \qquad (1)$$

The objective function for FCM is

$$J(U, c1, c2 \ldots . cc) = \sum_{i=1}^{c} Ji = \sum_{i=1}^{c} \sum_{j=1}^{n} U_{ij}{}^{m} . dij_{m} \quad (2)$$

where U_{ij} is between 0 and 1; C_i is the cluster center of fuzzy group i $D_{ij}=\|C_i-X_j\|$ is the Euclidean distance between the ith cluster center and the *j*th data point; and m$\in$[1,∞] is a weighting exponent.

The necessary conditions for Equation (2) to reach its minimum are

$$C_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_{ki}}{\sum_{k=1}^{n} u_{ik}} \quad (3)$$

and

$$u_{ik} = 1 / \sum_{k=1}^{c} \left[\frac{d_{ij}}{d_{kj}}\right]^{2/(m-1)} \quad (4)$$

where,
U_{ij}=Membership Function
C_{ij}=Centroids
J=Objective Function
m=fuzziness exponent
n=No. of clusters

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode operation, FCM determines the cluster centers i c and the membership matrix U using the following steps:

Step 1: Initialize the membership matrix U with random values between 0 and 1 such that the constraint in Equation (1) is satisfied.

Step 2: Calculate c fuzzy cluster centers, C_i,i=(1……c), using Equation (3).

Step 3: Compute the cost function according to Equation (2). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new U using Equation (4). Go to step 2. As in K-means clustering, the performance of FCM depends on the initial membership matrix values; thereby it is advisable to run the algorithm for several times, each starting with different values of membership grades of data points.

5.2 Support Vector Machine (SVM)

Vladimir Vapnik invented Support Vector Machines in 1979 [14]. In its simplest, linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. In the linear case, the margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples. The output u of a linear SVM is

$$u = \vec{w}.\vec{x} - b \quad (1)$$

where w is the normal vector to the hyperplane and x is the input vector. The separating hyperplane is the plane u=0. The nearest points lie on the planes u = ±1. The margin m is thus

$$m = \frac{1}{||w||_2} \quad (2)$$

Maximizing margin can be expressed via the following optimization problem

$$\min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2 \text{ subject to } y_i(\vec{w}\cdot\vec{x}_i - b) \geq 1, \forall i, \quad (3)$$

where x_i is the *i*th training example, and y_i is the correct output of the SVM for the *i*th training example.

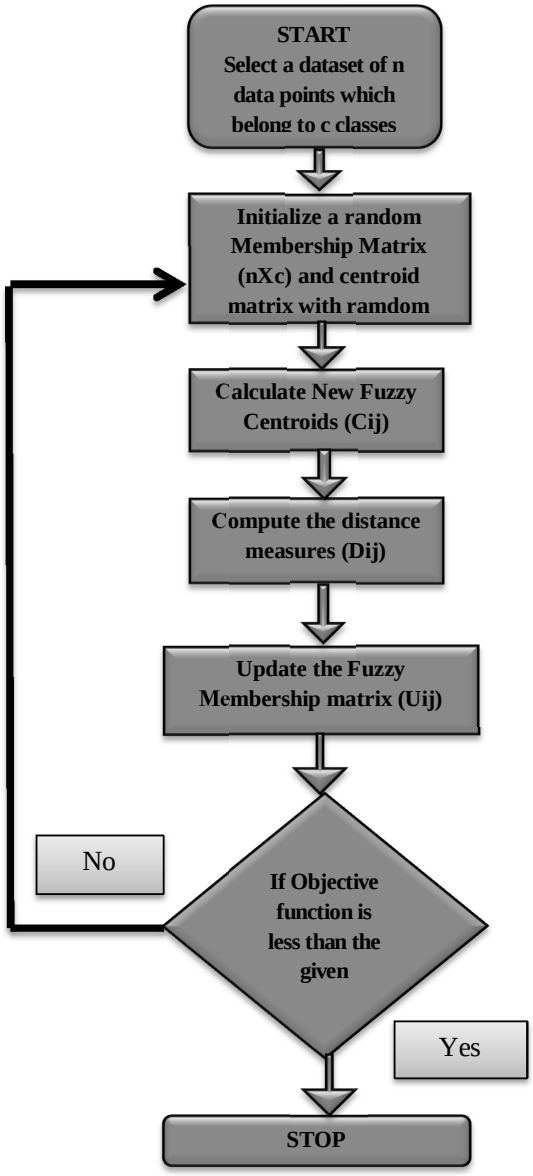

Figure 5.1 Flow of FCM

The value y_i is +1 for the positive examples in a class and –1 for the negative examples.
Using a Lagrangian, this optimization problem can be converted into a dual form which is a QP problem where the objective function ψ is solely dependent on a set of Lagrange multipliers α_i,

$$\min_{\vec{\alpha}} \Psi(\vec{\alpha}) = \min_{\vec{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i, \quad (4)$$

(where N is the number of training examples), subject to the inequality constraints,

$$\alpha_i \geq 0, \quad (5)$$

and one linear equality constraint,

$$\sum y_i \alpha_i = 0 \quad (6)$$

There is a one-to-one relationship between each Lagrange multiplier and each training example. Once the Lagrange

multipliers are determined, the normal vector $\vec{w}$ and the threshold b can be derived from the Lagrange multipliers:

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{x}_i, \quad b = \vec{w} \cdot \vec{x}_k - y_k \text{ for some } \alpha_k > 0. \qquad (7)$$

Because $\vec{w}$ can be computed via equation (7) from the training data before use, the amount of computation required to evaluate a linear SVM is constant in the number of non-zero support vectors.

Although not all data sets are linearly separable. There may be no hyperplane that splits the positive examples from the negative examples. In the formulation above, the non-separable case would correspond to an infinite solution. However, in 1995, Cortes & Vapnik [15] suggested a modification to the original optimization statement (3) which allows, but penalizes, the failure of an example to reach the correct margin. That modification is:

$$\min_{\vec{w},b,\vec{\xi}} \tfrac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \xi_i \text{ subject to } y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i, \forall i, \qquad (8)$$

where ξ_i are slack variables that permit margin failure and C is a parameter which trades off wide margin with a small number of margin failures.

When this new optimization problem is transformed into the dual form, it simply changes the constraint (5) into a box constraint:

$$0 \leq \alpha_i \leq C, \forall i. \qquad (9)$$

SVMs can be even further generalized to non-linear classifiers. The output of a non-linear SVM is explicitly computed from the Lagrange multipliers:

$$u = \sum_{j=1}^{N} y_j \alpha_j K(\vec{x}_j, \vec{x}) - b, \qquad (10)$$

where K is a kernel function that measures the similarity r or distance between the input vector $\vec{x}$ and the stored training vector $\vec{xj}$. Examples of K include Gaussians, polynomials, and neural network non-linearities. If K is linear, then the equation for the linear SVM (1) is recovered.

The Lagrange multipliers α_i is still computed via a quadratic program. The non-linearities alter the quadratic form, but the dual objective function ψ is still quadratic in α:

$$\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(\vec{x}_i, \vec{x}_j)\alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i,$$
$$0 \leq \alpha_i \leq C, \forall i, \qquad (11)$$
$$\sum_{i=1}^{N} y_i \alpha_i = 0.$$

The QP problem in equation (11), above, is the QP problem that the SMO algorithm will solve. The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions for an optimal point of a positive definite QP problem. The KKT conditions for the QP problem (11) are particularly simple. The QP problem is solved when, for all *i*:

$$\alpha_i = 0 \Leftrightarrow y_i u_i \geq 1,$$
$$0 < \alpha_i < C \Leftrightarrow y_i u_i = 1, \qquad (12)$$
$$\alpha_i = C \Leftrightarrow y_i u_i \leq 1.$$

where u_i is the output of the SVM for the *i*th training example. Notice that the KKT conditions can be evaluated on one example at a time, which will be useful in the construction of the SMO algorithm.

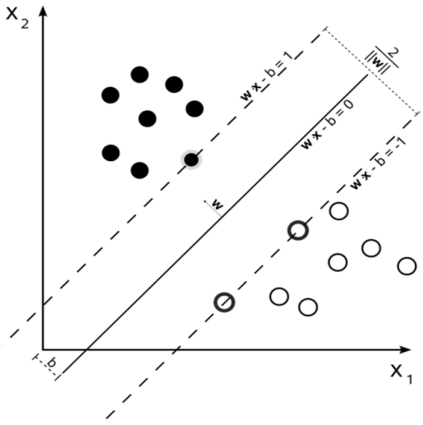

Figure 5.2 SVM Classification

6. RESULTS AND DISCUSSION

As proposed, the FCM clustering is implemented in MATLAB R2012a. To evaluate this clustering Pima Indian Diabetes Dataset [1] was used. The dataset has 9 attributes and 768 instances. Attributes are exacting, all patients now are females at least 21 years old of Pima Indian heritage. From the 768 patients in the PID dataset, classification algorithms used a training set with 576 patients and a testing dataset with 192 patients.

6.1 Performance Measures

To find the performance metrics such as sensitivity, specificity and accuracy, a distinguished confusion matrix is obtained based on the classification results from these algorithms. Confusion matrix is a matrix representation of the classification results as shown in Table 2.

TABLE 2

	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted.

Sensitivity is the percentage of positive labeled instances that were predicted as positive.

These performance criterion for the classifiers in disease detection are evaluated as follows from the confusion matrix.

Accuracy = (TP+TN) / (TP+FP+TN+FN)
Sensitivity = TP / (TP+FN)
Specificity = TN / (FP+TN)
Positive Prediction=FP / (TP+FP)
Negative Prediction = FN / (TN+FP)

6.2 Results

The Fuzzy C Means classification is implemented here with tolerance of 10^{-10} fuzziness co-efficient (m) =1.25 and No. of clusters=2. The True/False counts i.e. outcome of FCM classification are tabulated in TABLE 3. The implementation of FCM in MATLAB generated the results as shown in Table 4.

TABLE 3

	Classified as Healthy	Classified as not Healthy
Actual Healthy	63	3
Actual not Healthy	8	120

TABLE 4

Accuracy	94.300518%
Sensitivity	95.384615%
Specificity	93.750000%
Positive prediction	88.571429%
Negative prediction	97.658537%

FCM clustering applied on Diabetes dataset yields relatively better classification result of 94.3% accuracy. After two classes have been identified, the Euclidean distance measures between mean of centers of these two classes and new testing data are applied to pattern compute the confusion matrix.

SMO is an improved training algorithm for SVMs. Like other SVM training algorithms, SMO breaks down a large QP problem into a series of smaller QP problems. Unlike other algorithms, SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly.

The test results using the PIMA Indian Diabetes data set shows that the implementation of a SVM through SMO algorithm has been carried out correctly. The results are specified in TABLE 5 and TABLE 6 below.

TABLE 5

	Classified as Healthy	Classified as not Healthy
Actual Healthy	387	113
Actual not Healthy	198	70

TABLE 6

Accuracy	59.5052%
Sensitivity	77.4%
Specificity	26.1194%

CONCLUSION

Early detection of any kind of disease is an essential factor. This helps in treating the patient well ahead. In this, research paper is aimed to design a system that would assist doctors in medical diagnosis. This paper presents a diagnostic FCM as well as SVM using SMO and decides which technique helps in diagnosis of Diabetes disease.

The best result is by FCM with an accuracy of 94.3% and positive predictive value which is 88.57%. SVM has an accuracy of 59.5% which is quite low. These results are quite satisfactory, due to the fact that detecting the Diabetes is a very complex problem. Perhaps the most important result of this study was the understanding gained through the implementation and the results obtained here are also very encouraging and open the doors of the future research towards the detection of Diabetes disease.

This study can be further extended to deal datasets with multiple classes.

REFERENCES

[1] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu.

[2]Pardha Repalli, "*Prediction on Diabetes Using Data mining Approach*".

[3] Joseph L. Breault., "*Data Mining Diabetic Databases:Are Rough Sets a Useful Addition*".

[4] G. Parthiban, A. Rajesh, S.K.Srivatsa, "*Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method* ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[5] P. Padmaja, "*Characteristic evaluation of diabetes data using clustering techniques*", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[6] Lin, C., Lee, C., "*Neural Fuzzy Systems,*" PrenticeHall, NJ, 1996.

[7] Tsoukalas, L., Uhrig, R., "*Fuzzy and Neural Approaches in Engineering,*" John Wiley & Sons, Inc., NY, 1997.

[8] De Oliveira, J. V., & Pedrycz, W. (2007). *Advances in fuzzy clustering and its applications.* (1 ed., pp. 4-69). London: Wiley.

[9] Everitt, S., Landau, S., Leese, M. (2011). Cluster Analysis. (5 ed., pp. 76-80). London: Wiley.

[10] T. J. Ross, *Fuzzy Logic with Engineering Applications*, Third Edition, ISBN: 978-0-470-74376-8, John Wiley & Sons, 2010

[11] http://en.wikipedia.org/wiki/Diabetes_mellitus

[12].Yamaguchi M, Kaseda C, Yamazaki K, Kobayashi M. "*Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining"*. Med Biol Eng Comput.2006; 44(6):451–7.

[13] Centers for Disease Control and Prevention. National diabetes factsheet: national estimates and general information on Diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011. http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf

[14] Vapnik, V., *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, (1982).

[15] Cortes, C., Vapnik, V., "*Support Vector Networks*," Machine Learning, 20:273-297, (1995).

[16] Burges, C. J. C., "*A Tutorial on Support Vector Machines for Pattern Recognition,*" submitted to Data Mining and Knowledge Discovery,http://svm.research.bell-labs.com/SVMdoc.html, (1998).