SVM Classification to Distinguish Parkinson Disease Patients

Ipsita Bhattacharya[1]
Dept. Of Computer Engg
Netaji Subhash Institute of technology
NSIT, University Of Delhi (DU)
New Delhi ,India
pinkiipsita@gmail.com

M.P.S Bhatia[2]
Dept. Of Computer Engg
Netaji Subhash Institute of technology
NSIT, University Of Delhi (DU)
New Delhi ,India
mpsbhatia@nsit.ac.in

ABSTRACT

In this paper we have discussed the importance of data mining in the field of bioinformatics and various subfields of bioinformatics in which data mining has shown its great impact. Using a data mining tool, Weka, we pre- process the dataset on which we have worked and then using one of the classification methods i.e. Support Vector Machine method (SVM), we distinguished people with Parkinson‗s disease from the healthy people. Appling libsvm we have tried to find the best possible accuracy on different kernel values for the given dataset. We study the ROC curve variation, and the way the value of true positive and false positive rates changes with increasing number of the cross validation folds.

Categories and Subject Descriptors

I.2.6 Learning (K.3.2): Features – *Analogies, Concept learning, Connectionism and Neural Nets, Induction, knowledge acquisition, Language acquisition, Parameter learning.*

General Terms

Experimentation

Keywords

Parkinson Disease (PD), Support Vector Machine (SVM) Classifier, Maximum Marginal Hyperplane (MMH), kernel, False Positive Rate (FPR), True Positive Rate (TPR), ROC Curve

1. INTRODUCTION

The huge amount of research work done in the field of biology resulted in a tremendous amount of raw data being generated. This fast growing data is extremely information rich, but without the use of some systematic tool it is impossible to extract those golden nuggets of knowledge. This is where data mining comes and is extensively used since then, giving rise to bioinformatics, which is the science of managing, mining, and interpreting information from biological sequences and structures. [7].

Here our aim will be to discriminate healthy people from those with PD based on the ‗status' attribute of the dataset, by using SVM method of classification [1]. Then we use Libsvm to classify on a random split of the dataset, and calculate accuracy for the different kernel values.

SVM is a supervised Machine learning algorithm. Thus, its goal is to build a concise model of distribution of the class labels, in terms of a predictor function. The resulting classifier obtained is then used to assign class labels to the testing instances, where the values of the predictor features are known, but the value of the class label is unknown.[13].

For this reason supervised machine learning algorithms are also called induction classification algorithm , i.e they learn a set of rules from the instances given in the training set.[13]

Parkinson‗s disease is a degenerative disorder of the central nervous system. The central Nervous system includes the Brain and the spinal chord. A very deep part of the brain called the Basal ganglia, which is a collection of the nerve cells, is responsible for the secretion of the Neurotransmitter called dopamine, which helps in sending message and coordinating movement in the body. Hence in PD the patient‗s movements, speech and other functions get impaired.[5] . Though its cause is related to the lack of cells of substantia nigra, which contain the neurotransmitter dopamine, but for any particular case, the cause is unknown. [14]. Some scientists believe that a change in a specific gene, may be the reason while other experts think that it could be something in the environment that causes the damage, such as pesticides or other chemicals.[5] Although the role that heredity plays isn't completely understood, about 1 million people in the United States have Parkinson's disease [5]. At present, there is no cure for PD, but a variety of medications like a combination of levodopa and carbidopa provide dramatic relief from the symptoms[4].

The diagnosis of the disease is difficult in some cases due to its overlapping symptoms, but vocal impairment is considered one of the earliest indications of the onset of this disease. [2, 1]. Amongst the various vocal tests conducted, sustained phonation is one. [1].

The National Institute of Neurological Disorders and Stroke (NINDS) conducts PD research in laboratories at the National Institutes of Health (NIH) and promotes the advancement of research directed to the understanding, treatment and eventual cure of PD [4].

Conference'10, Month 1–2, 2010, City, State, Country.

2. RELATED WORK

Prominent areas of bioinformatics include (1) Gene expression analysis (2) Searching and understanding of protein mass spectroscopy of data (3) 3D structural and functional analysis and mining of DNA and protein sequences for structural and functional motifs (4) Text mining for biological knowledge discovery.[6].

The basic intention for many of the bioinformatics approaches is the evolution of organisms and the complexity of working with incomplete and noisy data.[3].BLAST and FASTE are designed by NCBI to deal with the challenge.[3].

Scientists have applied machine learning to information extraction in text. They developed a method that automatically extracts information from text in biomedical research articles, using various statistical methods. [6, 11]. Microarray technologies are used on the genotypic micro array data of patients and with the help of algorithms using support vector machine (SVM) various predictions are made.[6,8].

Scientists have applied machine learning to peptide identification through mass spectroscopy and also improved it‘s accuracy through kernel computation. They have also applied a block-based SVM method to perform protein homology prediction, with good results. [6, 9, 10].

In Neural network based approach; Probabilistic Neural Network (PNN) which is radial neural network, provides solution for classification problems using Baye‘s probabilistic rules. Here only one traversal of the training data is needed [2].

Scientists have used Support Vector Machine method and component-coupled method, also named as the covariant discrimination algorithm, to show that if they are complemented with each other, it can provide a powerful computational tool for predicting the structural classes of proteins. [13].

Support Vector machines have been used in many areas some of them includes drug design, image recognition and text classification, microarray gene expression data analysis and protein fold recognition [13].

3. DATASETS

Here we have downloaded the dataset for our work from http://archive.ics.uci.edu/ml/datasets/Parkinsons , where 197 voice recording instances are there of 31 people from which 23 are having the Parkinson disease. The various attributes are vocal fundamental frequency, absolute sound pressure level, extent of variation in speech i.e. jitter, shimmer indicating the speech amplitude and voice to harmonics ratio [1].

4. METHODS

Method that can be followed to achieve the goal is given below:

4.1 Support Vector Machines

SVM classifies, by finding a hyperplane also referred to as decision boundary. A hyperplane with large margin is expected to classify more accurately than that with smaller one so, SVM searches for MMH i.e. maximum marginal hyperplane. [7]. Data points that lie on the margin are known as support vector points and the solution is represented as a linear combination of only these points[13].

Support vectors are the most difficult instances to be classified, which provide us with a lot of information regarding the classification and also define the MMH. Hence once the support vectors and thus the MMH are defined, we get a trained SVM. [7].

Support vector machines use a linear separating hyperplane to create a classifier with a maximal margin. In order to do that, the learning problem for the SV machine will be cast as a constrained nonlinear optimization problem. In this setting the cost function is quadratic and the constraints linear[16].

In SVM a set of parameters **w** is the very subject of learning and generally these parameters are called weights, based on the hypothesis used these weights can be the hidden layer or output weights in multilayer perceptron or rules in fuzzy logic or the support vectors in SVM.The resultant model keeps a balance between overfitting and underfitting.[16].

The model complexity of an SVM is unaffected by the number of features encountered in the training. That is the reason why SVMs are well suited to deal with learning tasks, where the number of features is large with respect to the number of training instances. [13].

Real world problems can involve non-separable data, for which no hyperplane can be constituted. In those cases data can be mapped into a high dimensional space and can define a separate hyperplane called the transformed feature space [13].

The kernel functions are the special class of functions, which are used to map new points into the feature space for classification. [13].

4.2 Experiment Procedure

4.2.1 Pre-processing the datasets:

On pre- processing the dataset, it is found that the jitter and shimmer measure values are all very close to zero, with some rare examples of exceptionally high values. [1].Also it was found that there exists high correlation between various attribute values. [1]. Hence with the help of weka‘s filtering tool we remove those attribute.

4.2.2 Classification using SMO:

Next in order to classify we use support vector machine SVM. The basic advantage of using it is that it is much less prone to overfitting than other methods and secondly it gives a compact description of the learned model. [7].

Here Fig 1. shows the screen shot after removing the attributes identified as correlated ,in the pre-processing step. We have chosen _remove attributes‘ to eliminate those attributes and final list of attributes for classification is shown in fig 1.

In next step we have chosen classify icon in _Weka‘ and the algorithm chosen from the list is _SMO‘. With changing value of cross validation fold we have repeatedly classified the pre-prossesed dataset obtained in the Fig 1. The result for cross validation fold value _3_is given in Fig 2.

4.2.3 Classification using LIBSVM

Libsvm is an integrated software for support vector classification, regression and distribution estimation. It supports multiclass classification as well.

In order to find the best possible accuracy we have used the libsvm software. The preprocessed dataset obtained is split randomly and is transformed from the .CSV format to the libsvm format. We have used a perl script for the transformation from .CSV to libsvm format. On the training set, we have trained the classifier. And the test set instances have been classified.

For each kernel value :

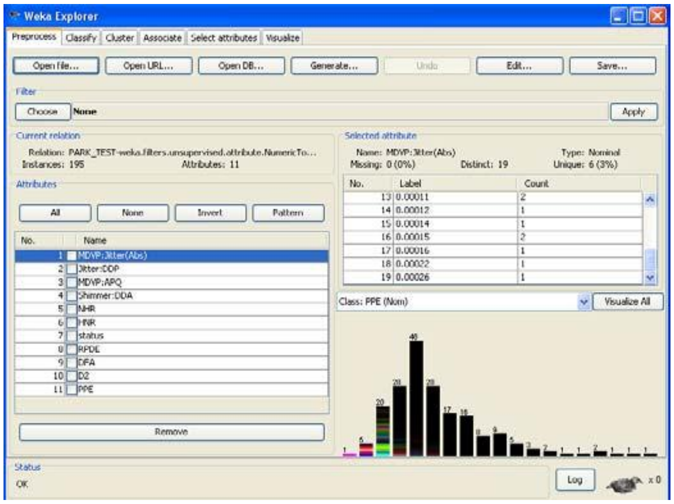
0—Linear

1---Polynomial

2-- radial basis function

3 -- sigmoid

We have changed the value of C which is the ‗Complexity' parameter. For each of the kernel values, we have increased C value, starting from 100 to 1000, and classify the test set in order to find the best possible accuracy . For a particular case i.e for RBF or Gaussian Kernel the result of libsvm classification has been shown in Fig 3.



Filter: None

Relation: PARK_TEST-weka.filters.unsupervised.attribute.NumericTo...
Instances: 195 Attributes: 11

No.	Name
1	MDVP:Jitter(Abs)
2	Jitter:DDP
3	MDVP:APQ
4	Shimmer:DDA
5	NHR
6	HNR
7	status
8	RPDE
9	DFA
10	D2
11	PPE

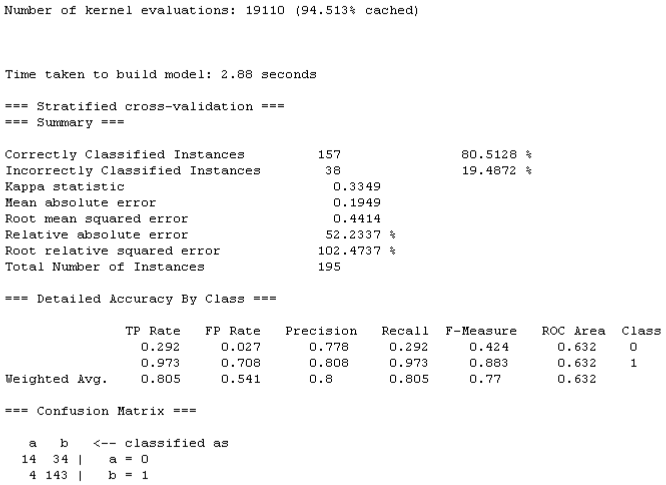
Name: MDVP:Jitter(Abs) Type: Nominal
Missing: 0 (0%) Distinct: 19 Unique: 6 (3%)

No.	Label	Count
13	0.00011	2
14	0.00012	1
15	0.00014	1
16	0.00015	2
17	0.00016	1
18	0.00022	1
19	0.00026	1

Class: PPE (Nom)

Status: OK

Figure 1: Screen Shot of Preprocessed Data in Weka



```
Number of kernel evaluations: 19110 (94.513% cached)



Time taken to build model: 2.88 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         157               80.5128 %
Incorrectly Classified Instances        38               19.4872 %
Kappa statistic                          0.3349
Mean absolute error                      0.1949
Root mean squared error                  0.4414
Relative absolute error                 52.2337 %
Root relative squared error            102.4737 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.292     0.027     0.778     0.292     0.424      0.632    0
                 0.973     0.708     0.808     0.973     0.883      0.632    1
Weighted Avg.    0.805     0.541     0.8       0.805     0.77       0.632

=== Confusion Matrix ===

   a   b   <-- classified as
  14  34 |   a = 0
   4 143 |   b = 1
```

Figure 2: Result Of Classification in Weka

Value of C : 100
Training set accuracy : 95.302%
(142/149) Test set accuracy : 60.8696%
(28/46)
.....*.*optimization finished, #iter =
956 nu = 0.229053
obj = -2673.587854, rho = -
1.080998 nSV = 42, nBSV = 30

Figure 3: Result Of Classification using Libsvm

5. RESULTS

In the result of classification using Weka in Fig 2 , we can see the details of accuracy measure by class. Here true positive rate and false negative rate for cross validation fold value 3 is shown.

Similarly by changing the value of cross validation fold from 2 to 10 we obtain the values shown in Table I. and Table II We can further see the confusion matrix , where each row of matrix represents instances in a predicted class while each column represents the instances in an actual class.

From the result we can infer that with the increasing value of cross validation fold the true positive rate value keeps on increasing in class 0 i.e in Table I. Also the value of false positive rate decreases with the increasing value of cross validation fold for class 1 i.e in Table II.

These results were expected because with the increasing value of cross validation fold we are increasing the training set instances and decreasing the test set instances, hence accuracy keeps on increasing.

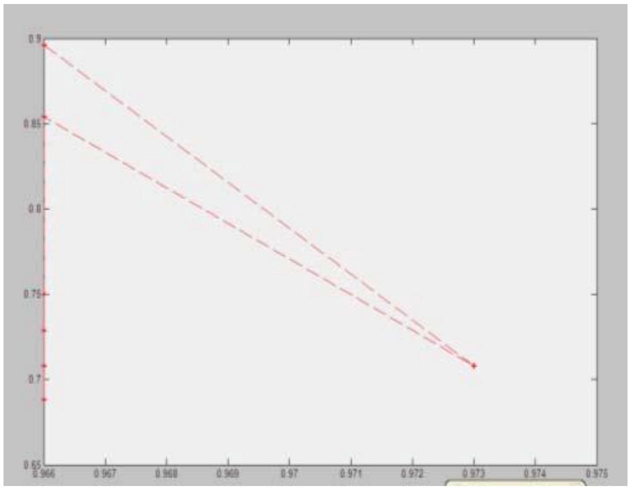
TABLE I: for class 0, HEALTHY PEOPLE

Cross Validation Fold Value	True Positive Rate(TPR)	False Positive Rate(FPR)
2	0.014	0.034
3	0.292	0.027
4	0.146	0.034
5	0.313	0.034
6	0.250	0.034
7	0.271	0.034
8	0.271	0.034
9	0.292	0.034
10	0.313	0.034

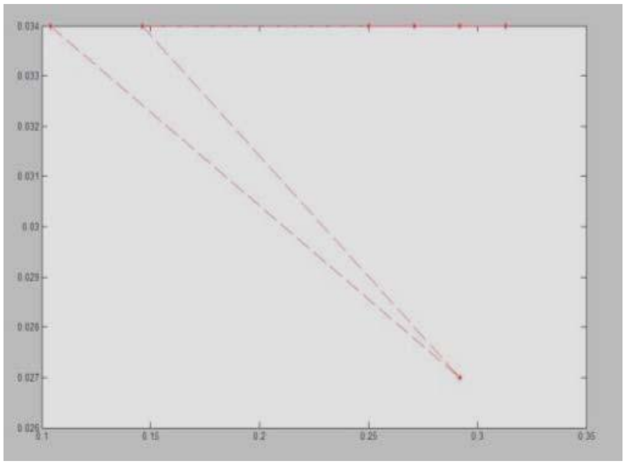
TABLE II: For Class 1 PDP

Cross Validation Fold	True Positive Rate(TPR)	False Positive Rate(FPR)
2	0.966	0.896
3	0.973	0.708
4	0.966	0.854
5	0.966	0.688
6	0.966	0.75
7	0.966	0.729
8	0.966	0.729
9	0.966	0.708
10	0.966	0.688

The result we have plotted using matlab , i.e the ROC curve. ROC stands for Receiver operating characteristic which is represented by plotting the fraction of true positives (TPR = true positive rate) vs. the fraction of false positives (FPR = false positive rate). The graph obtained after plotting the points, with TPR on the X axis and FPR on the Y axis.



For class 1 ROC Curve



For class 0 ROC Curve

In the result of classification using Libsvm , we have used different values of kernel as well as changed the value of C from 100 to 1000 , the result is shown in Table III , IV ,V.

TABLE III For RBF Kernel

Value of C	Training Set Accuracy	Test Set Accuracy
100	95.302% (142/149	60.8696% (28/46)
200	96.6443%(144/149)	54.3478%(25/46)
300	96.6443%(144/149)	54.3478% (25/46)
400	98.6577%(147/149)	52.1739% (24/46)
500	98.6577%(147/149)	50% (23/46)
600	98.6577%(147/149)	50% (23/46)
700	98.6577%(147/149)	50% (23/46)
800	98.6577% (147/149	52.1739% (24/46)
900	98.6577%(147/149)	52.1739% (24/46)
1000	98.6577% (147/149	52.1739% (24/46)

TABLE IV For linear kernel

Value of C	Training Set Accuracy	Test Set Accuracy
100	94.6309%(141/149)	65.2174%
200	97.3154% (145/149	65.2174%
300	96.6443% (144/149	65.2174%

400	96.6443% (144/149	63.0435%
500	96.6443% (144/149	65.2174%
600	96.6443% (144/149	65.2174%
700	96.6443% (144/149	65.2174%
800	96.6443% (144/149	65.2174%
900	96.6443% (144/149	65.2174%
1000	96.6443% (144/149	65.2174%

TABLE V For Polykernel

Value of C	Training Set Accuracy	Test Set Accuracy
100	95.302% (142/149	60.8696% (28/46)
200	96.6443%(144/149)	54.3478%(25/46)
300	96.6443%(144/149)	54.3478% (25/46)
400	98.6577%(147/149)	52.1739% (24/46)
500	98.6577%(147/149)	50% (23/46)
600	98.6577%(147/149)	50% (23/46)
700	98.6577%(147/149)	50% (23/46)
800	98.6577% (147/149	52.1739% (24/46)
900	98.6577%(147/149)	52.1739% (24/46)
1000	98.6577% (147/149	52.1739% (24/46)

From Table III we infer that RBF Kernel is not at all suitable as the test set accuracy keeps on decreasing with the increasing value of C and Training set accuracy keeps on increasing, which is not our aim and which indicates overfitting.

In case of Sigmoid Kernel For C value ranging from 100 to 1000 the training set accuracy is 83.8926% (125/149) and the test set accuracy is 47.8261% (22/46), which is unaffected by changing value of C. For Polykernel as well we see the test set accuracy decreases.

Hence we see that the best possible result is obtained for linear kernel which is 65.217%.

6. CONCLUSION

Hence we conclude that the SVM method can be used successfully to classify the instances into two classes, and hence can differentiate healthy people from those with Parkinson‘s disease. On the random split of dataset, we conclude that the best accuracy achieved is 65.2174%. On changing the split ratio and repeating the test we can achieve better result. The ROC curve study shows that with the increasing number of cross validation fold starting from 2 and increasing till 10, False positive and True positive rates changes. On plotting the graph for each class we get the above shown graphs. Hence by accurate and early detection, of the disease efficient treatment can be provided to the patients.

7. FUTURE SCOPE

We can extend our work by testing the same dataset on different tools like in Matlab and compare the efficiency of the two. With proper partitioning of the dataset better accuracy can be achieved. This process of analyzing the voice sample, of patients can help detect the disease in a wide range of people, staying far away from the diagnostic clinics, as was suggested by the scientists.[1].

8. ACKNOWLEDGEMENT

From UCI Machine Learning dataset repository dataset was collected site: http://archive.ics.uci.edu/ml/datasets/Parkinsons We worked on open source data mining tool: weka downloaded from site: http://www.cs.waikato.ac.nz/~ml/weka/.

We have classified using Libsvm-2.9 downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/

9. REFERENCES

[1] Little M. A., McSharry P.E., Hunter E.J., Ramig L.O. (2008), Suitability of dysphonia measurements for telemonitoring of Parkinson's disease ser IEEE Transactions on Biomedical Engineering

[2] Marius Ene Neural network-based approach to discriminate healthy from those with Parkinson's disease, Annals of the University of Craiova, Math. Comp. Sci. Ser. Volume 35, 2008, Pages 112{116 ISSN: 1223-6934}.

[3] JACQUES COHEN, Bioinformatics—An Introduction for Computer Scientists, Brandeis University. ACM Computing Surveys, Vol. 36, No. 2, June 2004, pp. 122–158.

[4] http://www.ninds.nih.gov/disorders/parkinsons_disease/parkinsons_disease.htm

[5] http://kidshealth.org/kid/grownup/conditions/parkinson

[6] Jinyan Li, Limsoon Wong, Qiang Yang, Data mining in bioinformatics, Institute for Infocomm Research, National University of Singapore Hong Kong University of Science and Technology.

[7] Han and Kamber, data mining concepts and techniquesy, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 2006.

[8] H. Liu, J. Li, and L. Wong, ―Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data," Bioinformatics, vol. 21, no. 16, 2005, pp. 3377–3384.

[9] Yan Fu et al., ―Exploiting the Kernel Trick to Correlate Fragment Ions for Peptide Identification via Tandem Mass Spectrometry," Bioinformatics, vol. 20, no. 12, 2004, pp. 1948–1954

[10] Yan Fu et al., ―A Block-Based Support Vector Machine Approach to the Protein Homology Prediction Task in KDD Cup 2004," ACM SIGKDD Explorations, vol. 6, no. 2, 2004, pp. 120–124

[11] S. Ray and M. Craven, ―Learning Statistical Models for Annotating Proteins with Function Information Using Biomedical Text," BMC Bioinformatics, vol. 6, suppl. 1, 2005, p. S18.

[12] Support Vector Machines for predicting protein structural class Yu-Dong Cai, Xiao-Jun Liu, Xue- biao Xu and Guo-Ping Zhou

[13] Supervised Machine Learning: A Review of Classification Technique: .S B Kotsiantis Department of Computer Science and Technology, University of Peloponnese, Greece End of Karaiskaki, 22100, Tripolis GR. Informatica 31 (2007)

[14] http://www.freelibrary.com/Parkinson+disease+current+evidence+for+acute+care+management

[15] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.

[16] Huang, Keeman and Kopriva, Kernel Based Algorithms For Mining Huge Datasets, Vol 17, Springer-Verlag Berlin Heidelberg 2006.