Study of Data Mining Algorithms in the Context of Performance Enhancement of Classification

Aditi Goel
M.Tech (CSE) Scholar
Department of Computer Science & Engineering
ABES Engineering College, Ghaziabad

Saurabh Kr. Srivastava
Sr. Assistant Professor
Department Of Information Technology
ABES Engineering College, Ghaziab

ABSTRACT
Data mining can help researchers to gain novel and deep insights for understanding of large datasets. Nowadays, people are using data mining algorithms in different contexts like banking, hospitals, marketing etc. Classification algorithm plays a vital role. In the study, we found that SVM is the best classifier amongst all the classifiers. Here we used learning algorithms with the historical dataset to train the classifier and the test samples are used to validate the correctness of the classifier. We might have structured semi-structured and unstructured datasets which are used for classification. We have performed the study of reputed literatures that belong to classification area to identify some new enhancements in the classifiers. A few most important classifiers are SVM, decision tree, neural network, Naive Bayes. We found most of the literatures were concentrated on SVM classifier so we targeted SVM classifier for the performance enhancement. SVM are important tool in data-mining to classify data. The aim of this review is to identify the effectiveness of kernel parameters for classification of data using Support Vector.

Keywords
SVM, data mining, classification algorithm, Naive Bayes, accuracy.

1. INTRODUCTION
Data mining helps in gathering important knowledge from the databases. The databases can be structured, semi-structured and unstructured. Researchers have done their work in all the directions. Role of data mining helps in knowledge discovery. Here is an example that proves the role of data mining in real scenario. Like, when you go to purchase the medicine the medical store owner saves the details of all the medicine purchased in the particular customer account. This helps the medical owner to keep a track on the medicines which are in stock or out-of-stock. With the help of this information the medical store owner will come to know about the sale of particular medicine. This indirectly helps the medicine manufacturers to know about the sale of their manufactured medicines. Data mining algorithms can help extract these kinds of important information.

Data mining has wide application. It is used in areas like Telecommunication, Fraud detection, Crime investigation, Businesses, Healthcare, Text mining, Web mining etc.

Data mining has following types of techniques:

1) Association : Association is a rule which means if we have two set of objects than the objects 1 should have something in common with object 2. Finding this similarity in the two object sets is the association rule. Multi-level association rule, Multi-dimensional association rule, Quantitative association rule are the types of association algorithm.
2) Classification : this technique contains two steps training step and the testing step. In the training step the model is constructed and in the testing step classifiers are tested to see the accuracy of the classifiers. Support Vector Machine (SVM), Neural Network, Decision tree, Naive Bayesian are the types of classification algorithm.
3) Clustering : Objects possessing similar characteristics are put together in form of a cluster. Two clusters do not have same characteristics or rarely have any same characteristics. Clustering algorithms are two types: Hierarchical and Partitional clustering algorithms.

2. RELATED WORK
In the context of classification following are the focused works reported by the researchers from the entire globe:

Iiihoi Yoo et al. [1] have demonstrated data mining and its technique. In their work they have discussed about the differences between Statistics and data mining. Some of the differences were that the statistics is a set of data that can be represented graphically, it is a mathematical term and it contains data in numeric values unlike the data mining which is a set of data which can be text, image, sound, categorical data. Data mining can even handle numeric data. Next they described about the data mining algorithms which are widely categorized into two types i.e. Predictive and Descriptive data mining algorithm. Predictive is a kind of supervised learning while the Descriptive is a type of unsupervised learning. Predictive data mining includes classification, regression and prediction algorithms. Descriptive data mining has types of clustering, association and summarization algorithms. Further, they described types of classification, clustering and association algorithms.

Murat Koklu et al. [2] have used a single dataset which is Pima Indian dataset of UCI repository. Three classifiers were used by the authors that are Multilayer Perceptron (MLP), Naive Bayes and J48 classifiers. They have explained MLP as a type of artificial neural network (ANN). ANN consists of neurons. It is a layered architecture which consists of input layer and the output layer. Every neuron is connected to every other neuron in a different layer by the weights assigned to it. The J48 algorithm is the java implementation of C4.5 algorithm in the Weka tool. C4.5 was introduced by Quinlan and it is improvement of basic ID3 algorithm. Naive Bayes classifier is used when input dimensionality is high. The dataset used has 768 instances and 8 attributes. The attributes are of numeric type. When the classifiers were applied MLP gave 75.130% accuracy, J48 gave73.828% accuracy and

Naive Bayes gave 76.302% accuracy. The Naive Bayes classifier had the best accuracy among the three classifiers.

Thirumal P.C et al. [3] have used various data mining techniques for better prediction of disease and increase the accuracy of detection of disease. In this paper authors have used Pima Indian diabetes dataset. The database has 768 patients. The data has two classes which are represented by binary value '0' and '1'. '0' represents the negative test and '1' represents positive test for the diabetes. Range of the values differs widely therefore the authors applied normalization method. They used 'weka.filters. Discretize' method for normalization. In this paper Weka has been used for the results. Different classifiers used are Decision tree (C4.5), SVM, Naive Bayes, K-Means, K-NN classifiers. Cross-Validation is used after applying classifiers. The result shows that the accuracy of C4.5 algorithm is best among all the other classifier which is 78.2552% and Naive Bayes is the second best with accuracy of 77.8646%. Accuracy of K-NN is 77.7344% and that of SVM is 77.474%.

Rahul Samant et al. [4] have compared performance of SVM with different Kernel parameters to predict the risk of Hypertension disease. Hypertension is the main factor which can increase the risk of Heart diseases. But if Hypertension is detected in earlier stages it can decrease the risks. The dataset used was compiled under the study entitled Early Detection project (EDP) at IIT, Bombay, Mumbai, India. There were some values which were missing in the dataset. To impute the missing values KNN-imputations were used. To reduce the features and have only important feature, the Principal Component Analysis (PCA) feature reduction technique was used. Through which 13 important features were selected. Five different Kernel parameters were used to compare the accuracy namely Linear, quadratic, polyorder, MLP and RBF. Four datasets DS1 is a mixed dataset which contains samples of healthy patients and diabetic patients, DS2 dataset contains hypertensive and diabetic patients, DS3 contains diagnostic information about hypertensive and diabetics as well as diabetics and DS4 is KNN imputed dataset containing information about patients who have diabetes and are healthy. For DS1 SVM with linear/kernel gave best accuracy of 84.83% with sensitivity and specificity 87.32% and 83.22% respectively. For DS2 SVM with quadratic kernel gave best accuracy of 85.33% with sensitivity of 79.35% and specificity of 82.12%. For DS3 and DS4 all the kernel functions had satisfactory level of accuracy, but SVM with linear kernel was a better choice because of slightly better accuracy. The accuracy of DS1, DS2 and DS3 were higher than the DS4 as this dataset was KNN-imputed for missing values.

Nahla H. Barakat et al. [5] have used SVM for diagnosis of diabetes, where they have added an additional-rule based explanation component to provide comprehensibility. Diabetes mellitus is a chronic disease. Diabetes has some complications which can be prevented by early identification of people at risk. SVM operates by finding hyper-plane. Kernel functions are used to handle the non-linearly separable data. SVMs are black box models because they are incapable of providing a comprehensible justification for classification decisions they make and in medical diagnosis the explanation for classification decision is of great significance. Therefore, the authors have introduced the techniques for rule-extraction to enable the SVMs to be more intelligible. Authors have used two techniques for rule-extraction namely SQRex-SVM and Eclectic methods. Data from 4682 subjects of age 20 years and plus was collected using a questionnaire. Result shows that accuracy of rules extracted by SQRex-SVM and eclectic approaches have better accuracy than that of SVM. TP Rate of eclectic approach was highest i.e. 96% followed by TP rate of SQRex-SVM i.e. 96 % followed by TP rate of SVM i.e. 95%. In case of AUC, Eclectic attained the highest of 95%, second highest was of SQRex-SVM. The ROC curves for both the eclectic and SQRex-SVM approaches are almost identical to that of SVM. Furthermore, rulesets extracted from SVMs have smaller number of rules.

Ashfaq Ahmed K et al. [6] have proposed the work in which they have used Classification technique for prediction of Cancer disease. The classification techniques authors have used are SVM and Random Forest technique. The dataset used is duke breast cancer. The results are analyzed with confusion matrix. The implementation is done using SVM tool and RF tool on Matlab with Microsoft VC++ compiler installed over it. The training and testing data are formatted into SVM tool using read call. Then train feature takes this formatted data as its input and generate a model of classifier. Different training models are created using different SVM kernels i.e. linear, polynomial, sigmoid and RBF. For Random Forest technique, training and testing data are formatted into random tool format and then trained to generate a model of classifier which is a decision tree based model. Results are better with Radial basis function with SVM and in some cases results are comparable with Random Forest technique.

Rohit Arora et al. [7] have used J48 and MLP classification techniques to compare their performances and choose better algorithm on the basis of datasets. Weka tool was used for results. Authors have used five datasets which have been taken form UCI repository. The first dataset used was balance-scale dataset with 625 instances and 5 attributes. The second dataset they have taken is a diabetes dataset with 768 instances and 9 attributes. The third dataset is the glass dataset with 214 instances and 10 attributes. The fourth dataset was lymphography dataset with 148 instances and 19 attributes and the last dataset was vehicle dataset with 946 instances and 19 attributes. Authors have used some measures on each data to compare the performances of the two techniques on each data. The measures used are i.e. TPrate, FP rate, Precision, Recall, F-measure and ROC Area. For balance-scale dataset the values of MLP had higher values than that of J48. Therefore, in this case MLP is better technique. For diabetes dataset all measures have almost equal value except ROC Area in which MLP had higher accuracy than J48. So, MLP is better for diabetes dataset. For glass dataset, all measures have almost same accuracy except ROC Area which had higher accuracy for MLP. Therefore, MLP was better for this case also. For lymphography dataset, MLP had better accuracy except FP rate. So, MLP was better for this dataset. For vehicle dataset, MLP was a better technique. Further, for the final results the authors used only the accuracy measure of all the datasets and they concluded that MLP was a better technique.

Riccardo Bellazzi et al. [8] have discussed about predictive type of data mining algorithms in clinical medicine. Data mining has been used in different fields like banking, marketing, engineering and various areas of science. This method may be applied for prognosis, diagnosis and treatment planning. The authors have used an example to discuss about the basic concept of predictive data mining. For this example, authors have taken a dataset of 20 patients record, each record is described by three attributes. The attributes are health, timing and complication. The data includes the response variable called outcome that reports if the treatment was successful as evaluated after 2years of operation. To describe

this, the authors have used two techniques- Naive Bayesian and Decision trees. The result shows that both the classification techniques did not perform well but decision tree somewhat performed better. Further, authors have described the predictive data mining methods as Decision making, Decision rules, Logistic, Artificial Neural Network, SVM, Naive Bayesian classifier, Bayesian and K-Nearest Neighbors (K-NN). The *k* most similar training instances and classifies based on their prevailing class. Predictive data mining standards that are recently getting attention are CRISP-DM, SEMMA and PMML. Further, authors have discussed about the data mining and statistics. Both can address large set of datasets but the distinction is that data mining can handle sheer size of data. Data sets that are drawn from clinical medicine consists of predictive features like nominal, real-valued, etc and may contain missing feature values. Data mining and modern statistics can mostly handle these types of data but how they discover predictive models is the difference in these two techniques. In data mining one would depend on automatic techniques like constructive induction but in case of statistics most often it involves manual search. Authors have also discussed about data mining tasks and its guidelines.

R.Priya et al. [9] have proposed the work in which they have used two data mining techniques to diagnose Diabetic Retinopathy. Diabetic Retinopathy is an eye disease which is caused due to some complications in diabetes. This Diabetic Retinopathy leads to blindness if not detected early. There are two kinds of Diabetic Retinopathy i.e. Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). For classification of diabetic retinopathy authors have used fundus images. A fundus camera was used to get the internal surface of the eye. The image taken was of resolution 1280× 1024 in 24bit JPEG format which was then converted into gray scale image. After converting the image, adaptive histogram equalization was applied to improve the contrast of the image. After this, DWT was applied by which the size of image was reduced to half i.e. 640×512. After this, MFR (Matched Filter Response) and Fuzzy c-means clustering were applied to reduce the noise and segment the blood vessels in the image. Features of the image like radius, diameter, area, arc length, center angle, half area were calculated for each image after the pre-processing of the image was done. After this, modelling techniques like PNN, Bayes theory and SVM were used to compare their performances. Then, the images were classified into three groups- normal, **N**on-Proliferative and Proliferative Diabetic Retinopathy. For the pre-processing of images, the techniques used were Gray Scale Conversion, Adaptive Histogram Equalization, Discrete Wavelet Transform, Gaussian Matched Filter Response and Fuzzy c-means clustering. The result was implemented in Matlab and Microsoft Visual Basic 6.0. Two classification techniques used by authors were SVM and Probabilistic Neural Network (PNN). Authors were able to identify normal, NPDR and PDR Cases of Diabetic Retinopathy with accuracy of more than 80% and sensitivity of more than 90% in both models. The accuracy of SVM and PNN were 97.6 % and 89.6, the sensitivity of SVM and PNN were 98% and 90% and the specificity of SVM and PNN were 99% and 88% respectively. The results proved that SVM was much more efficient than PNN.

Muhamad Hariz Muhamad Adnan et al. [10] have reviewed data mining for medical systems. They have discussed data mining as the process including formulating a hypothesis, collecting data, performing pre-processing, estimating the model and interpreting the model and draw the conclusions. Authors have discussed various data mining techniques- Artificial Neural Network (ANN), Decision Tree, Bayesian Classifiers and SVM. Data mining in medical domain is used to classify and predict the diseases like in Cornory Heart Diseases (CHD), Breast Cancer, Diabetes, In-vitro Fertilization and Childhood Obesity.

Lung-Cheng Huang et al. [11] have compared different classification techniques with and without feature selection. They have used the classification techniques to predict Chronic Fatigue Syndrome. To do this they have used genetic factors such as Single nucleotide polymorphisms (SNPs). The dataset used by authors have been taken from CDC Chronic Fatigue Syndrome Research Group. The dataset was total of 109 subjects in which 55subjects were of Chronic Fatigue Syndrome and 54 subjects were of non-fatigue controls. Authors focused on only 42 SNPs. In the data, 1.08% of SNPs call was missing which were imputed by the authors. Classification techniques used for comparison were C4.5 decision tree, naive bayes and SVM with four kernel- linear, polynomial Gaussian radial basis function and sigmoid. Two feature selection approaches were used to find a subset of SNPs that could maximize the performance of prediction model. First technique combined the information-gain method and chi-squared method. Second technique used was the wrapper-based feature selection approach, in this the feature selection algorithm acted as a wrapper around the classification algorithm. Measures used for performance comparison were AUC, sensitivity, specificity. Results were implemented in Weka. 10-fold Cross-Validation method was used. Firstly, the result of 10-fold cross-validation by naive bayes, SVM and C4.5 decision tree without using two feature selection techniques were implemented. The results shows that SVM with Gaussian radial basis function kernel performed the best with AUC=0.62. Secondly, three classifiers for hybrid feature selection approach that combined the chi-square and information-gain methods were implemented which showed that the naive bayes was best with AUC=0.70. Finally, the classification techniques for the other feature selection approach i,e. For the wrapper-based feature selection approach were implemented which showed that naive bayes outperformed the rest with AUC=0.70. Overall naive bayes classifier with hybrid and wrapper-based feature achieved the highest performance when compared with other techniques**.**

3. MOTIVATION TO WORK

(i) The research as now predicted some health issues are quite popular nowadays. People are using data mining to identify the health risks related to diabetes, flue, chronic fatigue syndrome and cancer.

(ii) The motivation towards this work came from the health risk that has increased nowadays. Like, diabetes is one of the top 10 most common diseases. According to World Health organization (WHO) in 2014, 9% of people of age 18 and above had diabetes. In 2012 diabetes was cause of 1.5 million deaths. Following are the deaths caused due to disease depicted in Table 1.

Table 1. Death rate corresponding to diseases

S.NO	CAUSES OF DEATH	DEATH RATE
1)	Heart disease	611,105
2)	Cancer	584,881
3)	Chronic lower respiratory diseases	149,205
4)	Stroke (Cerebrovascular diseases).	128,978
5)	Alzheimer's disease	84,767
6)	Diabetes	75,578
7)	Influenza and Pneumonia	56,979
8)	Nephritis, nephrotic syndrome, and nephrosis.	47,112

These above are two points of my motivation to work with data mining algorithms. If these diseases are detected in earlier stage then the death rates can be minimized. Data-mining in Healthcare using any of the classifier can help to decrease the death rate. This death rate minimization due to diseases motivated towards this work. Following Table 2. describes the paper wise algorithms and datasets used.

Table 2. Algorithms and datasets used in the reference papers

Paper Reference	Algorithm Used	Dataset Used
Ref [1]	Reviwed classification, regression and prediction, clustering, association and summarization algorithms.	Reviewed only algorithms
Ref [2]	Multilayer Perceptron (MLP) [Ref 4,7], Naive Bayes [Ref 8,11] and J48.	Pima Indian Dataset of UCI repository.
Ref [3]	Decision tree (C4.5), SVM [Ref 4,5,6,8,9,10,11], Naive Bayes, K-Means, K-NN classifiers [Ref 8].	Pima Indian diabetes dataset
Ref [4]	SVM with Linear, quadratic and polyorder, RBF. [Ref 3,6,8,9,10,11] and MLP [Ref 2,7]	Dataset used was compiled under the study entitled Early Detection project (EDP) at IIT, Bombay, Mumbai, India
Ref [5]	Support Vector Machine (SVM) [Ref3,4,6,8,9,10,11], SQRex-SVM and eclectic approaches	Data from 4682 subjects of age 20 years and plus was collected using a questionnaire.
Ref [6]	Support Vector Machine (SVM) [Ref 3,4,5,8,9,10,11] and Random Forest technique	Duke breast cancer dataset.
Ref [7]	J48 and Multilayer Perceptron (MLP) [Ref 2,4]	Five datasets have been used which were taken from UCI repository. The five datasets were balance-scale dataset, diabetes dataset, glass dataset, glass dataset, vehicle dataset.
Ref [8]	Decision tree [Ref 10] decision rules, logistic regression, Artificial neural network, Support Vector Machine (SVM)[Ref 3,4, 5,6,9,10,11], the naive Bayesian[Ref 2,11], Bayesian networks and K-nearest neighbours algorithms(KNN) [Ref 3].	Reviewed only algorithms
Ref [9]	Support Vector Machine (SVM) [Ref 3, 4,5,6,8,10,11] and Probabilistic Neural Network (PNN)	The algorithms were applied on the fundus image
Ref [10]	Artificial Neural Network (ANN), Decision Tree [Ref 8], Bayesian Classifiers and Support Vector Machine[Ref 3,4, 5,6,8,9,11]	Reviewed only algorithms.
Ref [11]	C4.5 decision tree, naive bayes [Ref 2,8] and SVM with four kernel- linear, polynomial Gaussian radial basis function and sigmoid[Ref 3,4,5,6,8,9,10,]	Dataset has been taken from CDC Chronic Fatigue Syndrome Research Group.

4. CONCLUSION

In the paper different data mining classification algorithms are covered. Every classification algorithm has its own role and importance. After this review of all the papers it was found that Support Vector Machines (SVM) plays an important role in classification. Here tuning parameters can be used that can improve the classification accuracy in prediction. Further, it is a requirement to validate the hypothesis generated "how kernel parameter plays an important role in SVM classification".

5. FUTURE WORK

Two main tasks that comes out in the above conclusion:

1) To test and validate the hypothesis generated.

2) How SVM performs best among other classification algorithms.

6. REFERENCES

[1] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang and Lei Hua, 2012, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature" *Springer*, J Med Syst,36, p.p-:2431–2448.

[2] Murat Koklu and Yavuz Unal, 2013," Analysis of a Population of Diabetic Patients Databases with Classifiers*", International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, Vol:7, No:8.

[3] Thirumal P. C. and Nagarajan N., January 2015, "Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus- A Case Study", *ARPN Journal of Engineering and Applied Science*, VOL. 10, NO. 1, ISSN 1819-6608.

[4] Rahul Samant and Srikantha Rao et al, 2013, "A study on Comparative Performance of SVM Classifier Models with Kernel Functions in Prediction of Hypertension", *International Journal of Computer Science and Information Technologies*, Vol. 4 (6), p.p. 818-821.

[5] Nahla H. Barakat, Andrew P. Bradley and Mohamed Nabil H. Barakat," July 2010, Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", *IEEE,*, VOL. 14, NO. 4.

[6] Ashfaq Ahmed K and Sultan Aljahdali, Nisar Hundewale and Ishthaq Ahmeed K, 2012, "Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques", *IEEE*.

[7] Rohit Arora and Suman, September 2012, "Comparitive Analysis of Classification Algorithms on Different Datasets using WEKA", *International journal of Computer Applications*, Vol. 54- No.13.

[8] Ricacardo Bellazzi and Blaz Zupan, 2008, "Predictive data mining in clinical medicine : Current issues and guidelines", *International Journal Of Medical Informatics*, p.p.-81-97.

[9] R.Priya and P. Aruna, March 2012, "SVM and Neural Network based Diagnosis of Diabetic Retinopathy", *International Journal of Computer Applications (0975 – 8887)",* Volume 41– No.1.

[10] Muhamad Hariz Muhamad Adnan, Wahidah Husain and Nur'Aini Abdul Rashid, 2012, "Data Mining for Medical Systems: A Review*", International Conference on Advances in Computer and Information Technology*, ACIT.

[11] Lung-Cheng Huang, Sen-Yen Hsu and Eugene Lin, 2009, "A Comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data", *Journal of Translational Medicine*.