ISSN 2347 - 3983

Volume 8. No. 7, July 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter119872020.pdf

https://doi.org/10.30534/ijeter/2020/119872020

SVM Classification Model With The Fusion Of Multiple Kernels For Medical Dataset

Nandini S[1], Yathish D P[2]

[1]GITAM School of Technology, INDIA, nandinis1801@gmail.com

[2]GITAM School of Technology, INDIA, yathi0509@gmail.com

ABSTRACT

Data mining is a field of research whose real goal is to get learning from a lot of information. Information mining has acquired a lot of consideration in data industry and in the public arena, because of accessibility of tremendous measures of information and the quick requirement for tuning such data into helpful information and learning. In medical and healthcare areas, because of controls and because of the accessibility of PCs, a lot of information is getting to be accessible. Experts are required to utilize this information in their work at the same time, in the meantime. A major objective of this paper is to develop a SVM[9] classification model technique with the combination of multiple kernels[2] to classify the benchmark datasets. Classification of medical data is one of the most challenging problems and also medical system is very complex, since it deals with large data situation and many features. Medical data classification is a topic of interest for doctors, researchers, consultants and medical related industries. In this paper we have implemented SVM classification[3] techniques with the combination of multiple kernels to classify the benchmark of medical datasets. Comparison of SVM classifier is done with Naïve Bayes classifier[4]. The motivation behind the proposed paper work is to build the classification model with the fusion of multiple kernels to classify the benchmark medical datasets and compared with other classifier technique

Key words : Data Mining, Kernel, Naive Bayes Classifier, SVM Classifier.

.

1. INTRODUCTION

Information mining is a field of acquiring the knowledge from the patterns which is obtained from the data. It utilizes strategies at the crossing point of data frameworks, AI, machine learning and measurements. While data discovery typically refers to the method of discovering helpful data from knowledge. Data processing focuses on the appliance of algorithms for extracting patterns from knowledge. The general target of the information mining system is to concentrate data from learning set and modify it into an adequate architecture for more utilize. Information mining strategies is upheld rapidly on existing hardware stages and programming bundle to strengthen the value of current information resources, and might be joined with framework and new stock as they're brought on-line. It is an inspired technique that may be applied to extract helpful patterns. Additionally to grouping and managing of knowledge, data processing also includes analysis and prediction.

Classification techniques in information processing area unit those are fit for processing an outsized quantity of knowledge. Therefore it are often made public as associate inevitable a part of data processing and is gaining a lot of quality.

The most generally used frameworks of information mining are:

Decision Trees: The sets of decision that represents tree-shaped structures. These choices produce rules for the characterization of a dataset.

Rule Induction: The abstraction of fitting if-then standards from information in light of measurable implication.

Nearest Neighbour Method: A procedure that characterizes every last record in a dataset in view of a combination of the classes of the k records most identified with it in a chronicled dataset.

Artificial Neural Networks: Non-linear prescient models that learn through preparing and reproduce natural neural systems in architecture.

2. LITERATURE SURVEY

In [1], the authors propose an incipient hybrid kernel function to amend the performance of medical dataset of heart disease diagnoses. The hybrid kernel function built between two customary bits. The K-sort bit capacity which is an early portion amalgamates with the RBF kernel and polynomial kernel; linear combinations with the different kernel functions are constructed with hybrid bit capacity. The PSO calculation is used to enhance the coefficient of direct amalgamation, a punishment parameter C which is concerned. They tested the model with the heart disease dataset. They mainly concentrated on maximizing the hyperplane, by probing the optimum disunion of hyperplane is identically tantamount to understand the quadratic programming (QP). This quadratic programming quandary is solved by utilizing the Lagrange multipliers.

In [2], the authors completed the near investigation of various part works for breast disease discovery. The creators focus on transfer of SVM with various pieces and moreover made examination with the neural systems technique using MLP (Multilayer Perception). They look at the impacts of choice element subset before applying transfer with various parts. They use the Wisconsin bosom cancer database from machine learning repository of the University of California. This dataset contains 699 elements vectors out of 458 are considerate cases and 241 are harmful cases, 16 highlights vectors are fragmented. Every elements vector contains 9 highlights. They used every one of the 9 components to get test results with subsets of the element separated using hereditary calculation for assessing the execution of SVM with various bit capacity. They use the coarse and fine matrix seeks and 5×2 cross acceptance to locate the ideal estimations of these parameters. They consider the hereditary calculation with all the distinctive bits to get the diverse exactness and MLP uses the GA for highlight subset separate. From this near investigation of SVM with various pieces and MLP for bosom malignancy location and they utilized four portions: RBF, polynomial, mahalanobis and sigmoid parts, the outcome demonstrates that the execution of SVM with various bits is superior to the MLP bit.

In [3], the authors proposed a nascent element determination by a RBF piece space for the assignment of dermatology infections. The proposed technique ideally highlights by abstracting the incidental/excess elements in RBF bit space using piece mean. The components of the information space are changed into RBF space and corrected the F-score mean estimation of each of the element has been processed in part space. At that point the separated components have been used in transfer calculations. The creators use the SVM, RBFN and C4.5 calculations had been taken for assessing the proposed technique for transfer of dermatology ailments. Using ten-fold cross approval parcels the training information and test information up to mapping the kernel space. SVM, RBFN and C4.5 calculations are amalgamated with RBF portion highlight separate for assignment. Contrast the outcomes and different calculations, C4.5 with RBF piece space is inciting higher transfer execution than the others techniques used in the assignment of dermatology maladies.

In [4], authors proposed a proficient calculation support vector machines with various parts predicated on Isometric element mapping (Isomap) during the time spent bosom disease transfer. Firstly, they use Isomap is executed to decrease the higher measurement of the bosom disease information into a lower space and afterward utilize SVM model with various parts to consign the lower dimensional information. The primary implies of Isomap is to locate the characteristic geometry of the information, as caught in the geodesic complex separation between all sets of information focuses. The dataset has 569 specimens. Every specimen speaks to a genuine patient and has 30 restorative components. 357 examples are kind-hearted tumours, 212 specimens are dangerous tumours. They outline the SVM models predicated on RBF part work. At that point builds the SVM models with different parts with accumulation of RBF bit and polynomial portion work and use these models to consign the bosom growth dataset and outcomes are recorded. The execution of the calculation is appeared in ROC (Receiver Operating Characteristics) and AUC (Area under ROC) bend. The outcome demonstrates the ISOMAP-SVM with amalgamation pieces have higher transfer rate than the customary SVM with single kernel.

In [5], authors have worked to give a translation of a procedure that can be accustomed to raise arrangement flourishing rates. Region predicated support vector machine calculation partitions the preparation information into two subsets; in the main subset, emerge from the joining of rational information areas and second subset is involves the information parcels that is laborious to be bunched. Pima Indian Diabetes information acquired from the UCI storehouse of machine learning databases and this dataset incorporates data collected of female's patient those are no less than 21 years age of the Pima Indian legacy. The cases of dataset are the nature challenging to divide by a solitary summing up machine learning model. There calculation is trying to minimize the preparation blunder while developing an assignment limit.

In [6], authors have worked on effective hybridized classifier for breast malignancy finding. Build the classifier by mixing an unsupervised neural system (ANN) technique assigned self-sorting out maps (SOM) with a directed classifier called stochastic inclination drop (SGD). Bosom cancer Wisconsin dataset was amassed from UCI machine learning storehouse. They amass the 699 occurrences of dataset from January 1989 to November 1991. For check, another dataset kenned as Internet Advertisement dataset were are moreover used. Firstly the non-cross breed model of SGD was accustomed to consign the datasets. At that point SGD was combining with SOM model and the half breed framework was accustomed to consign the cases. The near examination is performed between the proposed strategy and three regulated best in class machine learning procedures Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). The correlation predicated up on assignment exactness that is incited by inducing a perplexity grid. The outcome demonstrates that transfer using hybridization of SOM with SGD is substantially better than SGD in disengagement.

In [7], the authors considers a hypothyroid identification and relegation using multi-class support vector machine classifier and its application. The primary target of this work is vigour of sundry sorts of parts for multi-class SVM classifier and an examination of various building strategies for Multi-class SVM, for example, One-Against-One and One-Against-All. The creators use the Multi-class Support Vector Machines (MCSVM) systems for recognition and transfer of hypothyroid scrape in people. They give exploratory results on hypothyroid infection dataset from UCI vault for MCSVM. The creators directed a test under Windows Vista running on a PC with framework design Intel P4 processor 2.6 GHz with 2 GB of RAM. Hypothyroid dataset had 4 classes with 30 qualities. The dataset contains 3772 specimens which are separated to 3167 examples as preparing information and 605 as test information. The vigour of MCSVM classifier is contrast with three distinct pieces with decides its accuracy and involution. He also leads explored different avenues regarding MCSVMs using the OAOSVM and the OAASVM with polynomial parts. The tested results on hypothyroid dataset assignment have demonstrated the fabulous execution and accuracy of MCSVM than decision tree and AdaBoost. It has withal demonstrated that the exactness of OAASVM with polynomial bit is superior to anything others.

In [8], the authors have completed correlation among the diverse classifiers Naïve Bayes (NB), Instance Based for K-Nearest neighbour (IBK), Multi-Layer Perception (MLP), decision tree (J48) and Sequential Minimal Optimization (SMO). In this paper, they consider three distinct databases of bosom tumor are Wisconsin Diagnosis Breast Cancer (WDBC), Wisconsin Prognosis Breast Cancer (WPBC) and Wisconsin Breast Cancer (WBC). These datasets are by using assignment exactness and disarray network predicated on 10-fold cross approval technique. Combination of classifiers is amalgamating with different classifiers to show signs of improvement exactness. The test result demonstrates the exactness by transfer using the combination of MLP and J48 with the PCA are bosses to alternate classifiers using WBC dataset. The PCA is used in WBC dataset as a components decrease change technique in which cumulates an arrangement of associated elements. Cumulating all classifiers to the three diverse datasets to obtains preferred accuracy over different classifiers. In this paper, all trials are led in WEKA information mining instrument.

3. METHODOLOGY

To perform the information examination, it is critical to gather the datasets from various sources. At first datasets are in crude information structure; here and there the attributes of the information may not be ideal for classification. This information must change over into fitting structure before any mining can start. In this work, we consider the medicinal datasets for arrangement.

Dermatology Dataset: This dataset is taken from UCI machine learning repository and original owners are Niselilter, Gazi University, School of Medicine, Turkey and H. Altay Guvenir, Bilkent University, Dept. of Computer Engineering and Information Science.

Dermatology disease identification is troublesome issue for experienced specialists. The differential analysis of erythemato-squamous malady is genuine issue. They are all sharing the clinical components of erythema and scaling, with almost no qualifications. The inconvenience of the different conclusion that is an infection may show the elements of other ailment toward the beginning stage and may have the trademark highlights at the going with stages. The patients are at first surveyed clinically with 12 highlights. A short time later, skin tests were taken for the assessment of 22 histo pathological highlights.

In this work, we divide the dermatology dataset into two sets are training samples and test samples. It contains 358 instances out of which 321 are training samples and 27 are test samples, respectively, each samples involves 34 features with 6 different classes.

Attributes Selection[10]

In dermatology dataset we worked for this territory, the family history highlight has the value 1 if any of these infections has been found in the family and 0 for the most part. The age incorporate basically addresses the age of the patient. Every single other component (clinical and histo pathological) was given a degree in the extent of 0 to 3. Here, 0 exhibits that the element was not present, 3 demonstrates the greatest total possible, and 1, 2 shows that relative moderate qualities.

Following steps are used to develop a SVM model[9] based on multiple kernel functions for multiclass Dermatology Diagnosis Disease.

Step 1: Gather the dataset information from a recognized source which is used to obtain suitable results. Collect the dermatology diagnosis disease from the UCI repository and load the data to the SVM model.

Step 2: Separate the dataset into two sets are training set and test set. The training set consists 70% of samples and the test set consists 30% of samples of the dataset. The training samples have a great set of information about dataset. Before going to the step apply the 10-cross fold validation on training set.

Step 3: Construct the SVM classification model based on multiclass SVM and train the SVM using training samples with a parallel group vector.

Step 4: Classifies a given test set utilizing SVM[9] classifier as per one-vs-all relation. The i^{th} SVM is trained such that the samples in the i^{th} class are marked as positive samples and all the rest as negative samples.

Step 5: Develop a hybrid kernel function with combination of RBF kernel and MLP kernel functions to implement the multiclass SVM classification model.

Step 6: Evaluates the performance of a SVM classification model[9] by calculating the performance metrics are: precision, specificity, sensitivity, accuracy and F-measure.

These steps are depicted in a form of flow diagram in Figure 1.

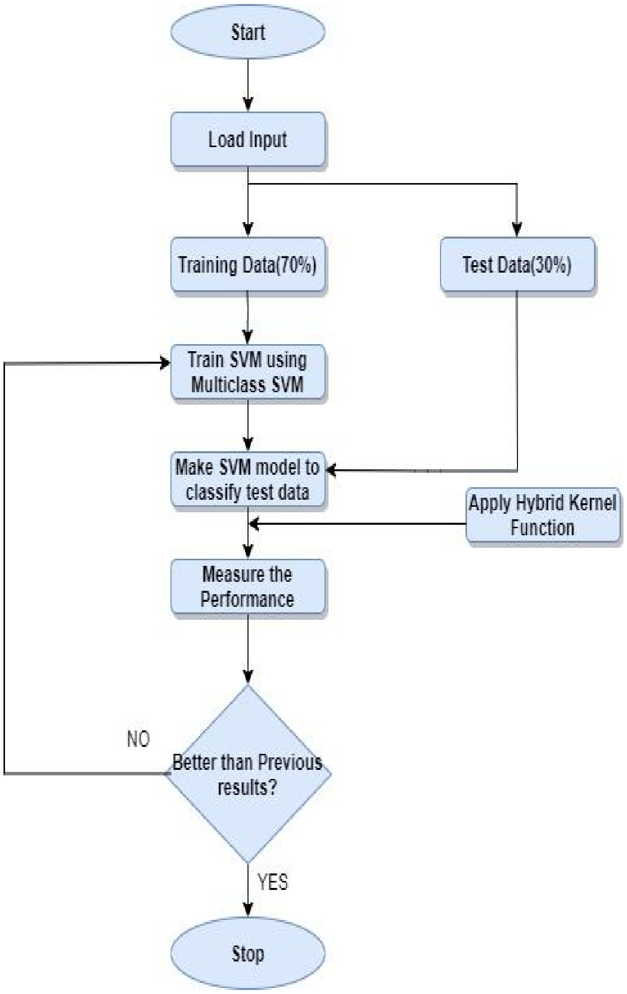
Start
Load Input
Training Data(70%)
Test Data(30%)
Train SVM using Multiclass SVM
Make SVM model to classify test data
Apply Hybrid Kernel Function
Measure the Performance
NO
Better than Previous results?
YES
Stop

Figure 1: Work Flow

4. RESULTS

Performance Comparison b/w Traditional kernels

Below figure 2 shows the performance comparison between traditional kernels and by using proposed kernel . The figure 2 depicts that the performance is more in a proposed system.

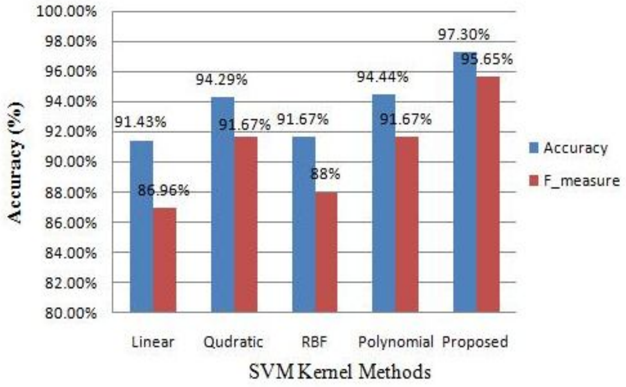

Figure 2: Performance comparison

Compare Performance b/w Naive Bayes Classifier

When the performance of the proposed system is compared with Naive Bayes classifier, it is seen in figure 3 that the proposed system has more accuracy and F_measure.

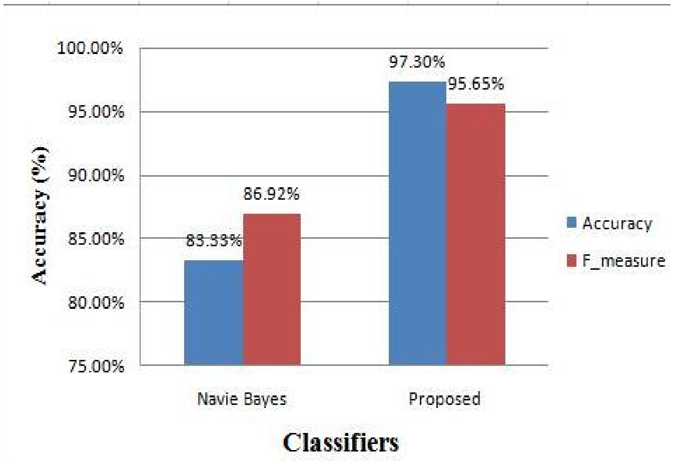

Figure 3: Performance Comparison between Naive Bayes Classifier and proposed system

5. CONCLUSION

In medical areas lot of diagnosis information is hidden in the form of raw data, before using these raw data pre-processing is required which means the data should be in understandable format. Diagnosis of healthcare conditions is really a challenging task. Medical data includes a number of tests necessary to diagnosis a particular disease and these diagnosis is conducted based on the experience of physicians. Hence, health industries use the data mining techniques to develop an efficient model.

In the paper, the dermatology diagnosis disease dataset is used. Dermatology is a study of skin cancer that is very difficult and complicated to diagnose. The attribute or feature[10] type is classification; the six different classes of these diseases share the identical features of erythema. The general objective of the proposed work is to build a new classification model with combination of multiple kernels. Dataset is divide into two sets which contains 70% of training samples and 30% of test samples. Apply the 10-cross fold validation on training set to train the each subset and SVM classification model. Then construct the SVM model based on multiclass SVM and train the SVM using training samples with corresponding support vectors. Here SVM classifier of one-vs-all relation classifies the given test set. Now we build a new hybrid kernel by combining good characteristics of RBF kernel and MLP kernel functions in order to classify the dermatology dataset. The performance of SVM classification model is evaluated by calculating the performance measures and the results are compared with Naïve Bayes classifier and traditional kernels functions.

REFERENCES

1. Yuanbin Mo, Shuihua Xu., **Application of SVM based on hybrid kernel function in heart diseases diagnosis**, *2010 IEEE International Conference of Intelligent Computing and Cognitive Informatics.*
2. Muhammad Hussain, Mohammed Berbar, Ali Elzaart, Summrina KanwalWajid., **A Comparison of SVM Kernel Functions for Breast Cancer Detection,** *2011 IEEE Eighth International Conference Computer Graphics, Imaging and Visualisation.* https://doi.org/10.1109/CGIV.2011.31
3. Rajkumar .N, Jaganathan P., **A new RBF kernel based learning method applied to multi-class dermatology diseases classification**, *Conference on Information and Communication Technologies, IEEE, 2013.*
4. Xuifeng Yang, HuiPeng, Mingruhi Shi, **SVM with Multiple Kernels based on Manifold Learning for Breast Cancer Diagnosis,** *August 2013, proceeding of the IEEE International Conference on Information Automation, Yinchuan, China.*
5. Savvas Karatsiolis, Christos.N.Schizas, **Region based Support Vector Machine Algorithm for Medical Diagnosis on Pima Indian Diabetes Dataset,** *proceeding of the 2012 IEEE 12th International Conference on Bioinformatics and Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November.* https://doi.org/10.1109/BIBE.2012.6399663
6. Dishant Mittal, Dev Gaurav, Sanjiban Sekhar Roy, **An Effective Hybridized Classifier for Breast Cancer Diagnosis**, *2015, IEE, International Conference on Advanced Intelligent Mechatronics (AIM), July 7-11, Busan, Korea.* https://doi.org/10.1109/AIM.2015.7222674
7. Fereshteh Falah Chamasemani, Yashwant Prasad Singh, **Multi-class Support Vector Machine Classifiers – An Application in Hypothyroid detection and Classification,** *2011 IEEE Sixth International Conference on Bio-Inspired Computing: Theories and Application.*
8. Gouda I. Salama, M.B. Abdelhalim, MagdyAbd-elghnay Zeid, **Breast Cancer Diagnosis on Three Different Datasets using Multi-Classifiers,** *September-2012 International Journal of Computer and Information Technology (2277 - 0764), Volume 01 – Issue 01.*
9. Dr. D. Nagajyothi , Rakshith Addagudi , Tejaswini Gunda , Sindhu santhoshi Logitla, **Detection of Lung Cancer using SVM Classifier,** *International Journal of Emerging Trends in Engineering Research,* Volume 8. No. 5, May 2020 https://doi.org/10.30534/ijeter/2020/113852020
10. Amal Fouad , Hossam M. Moftah , Hesham A. Hefny , **MRI Brain Cancer Diagnosis Approach Using Gabor Filter and Support Vector Machine**, *International Journal of Emerging Trends in Engineering Research, Volume 7, No. 12 December 2019* https://doi.org/10.30534/ijeter/2019/297122019