Hybrid Model Using Unsupervised Filtering Based On Ant Colony Optimization And Multiclass Svm By Considering Medical Data Set

Rashmi[1], Shaveta Saini[2]

[1] *M.Tech Scholar, Department of Computer Science & Engineering, Guru Nanak Dev University RC Jalandhar, Punjab*
[2] *M.Tech Scholar, Department of Computer Science & Engineering, Guru Nanak Dev University RC Jalandhar, Punjab*

***

Abstract - *Data mining is the computational procedure for discovering habits in big data pieces ("big data") regarding methods in the intersection of artificial thinking ability, machine learning, statistics, and also database programs. The total goal of the data exploration process is usually to extract information from the data set and transform it into a good understandable structure for further use. It has been found that that the Ant colony optimization outperforms over the J48 and random forest based machine learning algorithms. The overall objective of this research work is to propose a hybrid model which will use unsupervised filtering which will be followed by Ant colony optimization and multiclass SVM by considering the medical data set.*

***KeyWords*: Data mining, Ant colony optimization,Random forest, SVM, ACO.**

1.INTRODUCTION

Data Mining is defined as getting information and facts via substantial sets of data. To put it differently, most people can say which facts mining is the method with mining knowledge via data. The content or maybe knowledge removed consequently work extremely well in many programs like Market place Evaluation, Deception Diagnosis, Client Storage, Manufacturing Regulate, Scientific research Search etc. Normally, facts mining (sometimes identified as facts or maybe knowledge discovery) is the whole process of examining facts via various aspects along with summarizing the idea straight into valuable information and facts - information and facts which can be used to improve sales revenue, reduces prices, or maybe both. Details mining software is one of the systematic instruments intended for examining data. The idea permits consumers to research facts via several size or maybe angles, sort the idea, along with summarize this interactions identified. Technologically, facts mining is the whole process of obtaining correlations or maybe habits amongst a large number of domains around large relational databases.

Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

1.1 Ant Colony Optimization in Data Mining

Data mining (sometimes termed facts or expertise discovery) includes the usage of sophisticated facts evaluation equipment find out before mysterious, valid behavior and also romantic relationships inside significant datasets. These power tools may incorporate statistical types, statistical algorithms, and also appliance understanding solutions (algorithms of which better their overall performance routinely as a result of expertise, such as neurological communities or selection trees). For that reason, facts mining consists of greater than collecting and also managing facts – this also may include study and also prediction. Data mining can be on facts depicted inside quantitative, textual forms. Files mining apps can make use of various factors to check the data. Such as connections, collection or direction study, distinction, clustering and also forecasting. While facts mining solutions can be very powerful gear, they may not be self-sufficient applications. To overcome the objectives, facts mining needs qualified complex and also investigative Gurus that can design your study and also experience the outcome that's created. Frequently, inside facts mining projects virtually any of 4 sorts of romantic relationships usually are desired:
•Classes: stored data is used to locate data in predetermined groups.
•Clusters: data items are grouped according to logical relation-ships or user preferences
•Associations: data can be mined to identify associations.
•Sequential patterns: data is mined to anticipate behavior patterns and trends.

1.2 Random Forest

RF suits numerous category trees to the details fixed, and then fuses the actual predictions all the actual trees. The algorithm starts off with selecting a numerous (e.g., 500) bootstrap samples from the data. Inside a common bootstrap taste, roughly 63% connected with the very first observations happen at least once. Observations in the very first details fixed that will not happen inside a bootstrap taste are called out-of-bag observations. Your category sapling is match to each and every bootstrap taste, nevertheless at each and every node, only a few arbitrarily picked specifics (e.g., the actual sq cause of the number of variables) are available for the actual binary partitioning. The trees and shrubs usually are totally produced and they are all utilised to predict the actual out-of-bag observations.

The forecasted category of an statement is worked out by simply the vast majority political election in the out-of-bag predictions for your statement, with jewelry divide randomly.

1.3 Ant Colony Decision Trees (ACDT)

It is actually exciting to see the fact that proposed algorithm criteria connected with decision sapling structure is principally depending on the release connected with insect nest optimization. Several bit of a variations have already been presented the two while a whole new individually distinct optimization algorithm criteria pertaining to making decision woods as well as a whole new meta heuristics tactic in details exploration procedures. With ACDT each one insect prefers the appropriate feature pertaining to splitting in each one node with the produced decision sapling based on the heuristic operate as well as pheromone values.

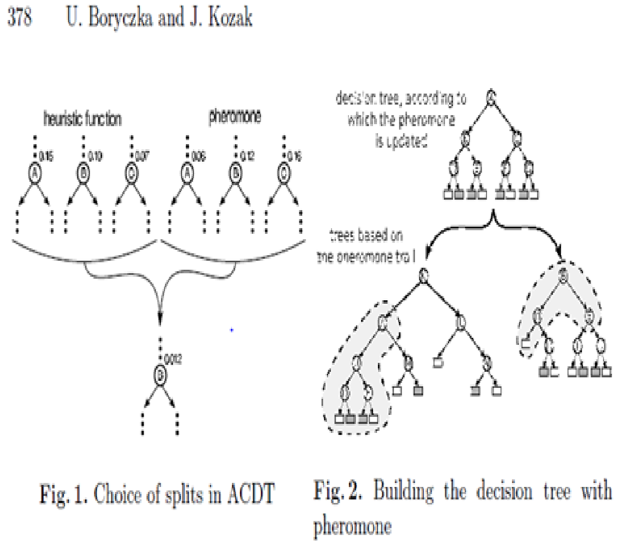
378 U. Boryczka and J. Kozak
heuristic function
pheromone
0.16
0.10
0.07
0.06
0.12
0.16
0.012
decision tree, according to which the pheromone is updated
trees based on the pheromone trail

Fig. 1. Choice of splits in ACDT

Fig. 2. Building the decision tree with pheromone

Fig1. Ant Colony Decision Trees

The particular displayed alterations are unveiled primarily move tip and perhaps they are treated seeing that a noticable difference on the expertise of the group mechanism. We certainly have utilised your conventional version with ACO easy alterations in regards to the main policies, dedicated to each agents–little bugs through the building this trips are contained in the scheme. We have got put on this conventional busting tip, first of all utilized in CART. Secondly, we're complied with the pheromone alterations which will are useful awareness intended for generating an acceptable division. With ACDT each ould like determines the perfect trait intended for busting within each node on the made determination tree according to the heuristic function along with pheromone beliefs (fig. 1). The particular heuristic function is definitely using the Twoing criterion, which supports little bugs divide this physical objects within a couple communities, of this analyzed trait values. In this way, this trait, which will nicely split this physical objects is definitely treated as being the finest situation for that analyzed node. The best busting is definitely noticed after we labeled exactly the same volume of physical objects in the right and left sub-contract woods using the ideal homogeneity within deciding classes. Pheromone beliefs signify the simplest way (connection) with the better than this subordinate nodes– most doable a combination in the analyzed subtrees. For each node most of us estimate the next beliefs according to the physical objects labeled utilizing the Twoing criterion with the highest node.The pseudo computer code on the offered criteria is definitely displayed below. At the start of it is deliver the results, each ould like forms a single determination tree . At the end of this hook, the top determination tree is definitely picked out and then the pheromone is definitely up-to-date according to the splits completed along the way with structure deciding tree, iteratively. Whilst constructing this tree, agents–little bugs are examining prior structures and many alterations are executed within the node. This technique is completed till the finest determination tree is definitely obtained. Is essential constructing deciding tree is definitely presented.

2. UNSUPERVISED FILTERING

Techniques to construct narrow techniques intended for raw, unclassified details are called without supervision mastering procedures with the theory with neurological networks. This sort of techniques will often be specified by their mastering rules, i.e., how they change their central dumbbells or even narrow coefficients. In this particular it will work with an alternative method to discover the narrow functions. Here we opt for very first many properties of the outcome vectors computed by the system. We layout a good (or energy) function that will procedures most of these properties. Eventually, most of us utilize an iterative search engine optimization method to discover the filters. It can be the benefit that will the massive current expertise with search engine optimization concept can certainly apply to discover powerful implementations with the training procedure.

2.1 Multiclass SVMs

SVMs will be in the beginning designed for binary distinction problems. Extensions to be able to multiclass commonly contain possibly fixing a huge seo trouble immediately or perhaps contemplating the decomposition with the original trouble into smaller binary sub-contract troubles or perhaps next pairing their particular solutions. While either techniques, commonly, present no factor inside effectiveness in the event the super factors will be correctly updated [16], the decomposition one is much more computational attractive. There are two most important strategies with decomposition: One-Versus-One (OVO) and also One-Versus-All (OVA). And may generally utilized because of their ease, overall performance and also in the same way excellent distinction effectiveness [16]. This kind of document aims at on the OVO program, however the proposed approach may very well be well put on some other multiclass strategies since well. Your OVO process constructs N(N−1)/2 SVMs, acquiring note many binary combinations of classes. While an evaluation case in point is supplied, it truly is put on each of the SVMs and their produces will be for some reason combined. Your MaxWins voting program [9] utilized the following counts how frequently will you each and every course is definitely outputted through the binary SVMs and also the test case in point is associated with by far the most identified as class.

3.Related work

Duan, K. B., Rajapakse, J. C., Wang, H., & Azuaje, F. (2005)[1] has proposeda new aspect range procedure that runs on the backward elimination treatment just like that carried out throughout assist vector unit recursive aspect elimination (SVM-RFE). Contrary to a SVM-RFE procedure, at most move, a planned tactic computes a aspect rank credit score from your precise analysis involving weight vectors involving a number of straight line SVMs educated on subsamples involving the very first education data. Cios K.J and Moore G.W(2002)[2] has studiedEthical as well as legalised elements of health data mining as well as data property, concern with litigation, anticipated benefits plus unique administrator issues.Brameir.M and Banzhaf.W (2001) [3] has discussed a couple of strategies of acceleration and speed of innate encoding approach. First one is the use of a powerful protocol that will minimizes code. Next one is a demotic approach to virtually parallelize the device on a single processor.GP operation with professional medical explanation issues is usually in comparison coming from a standard repository along with outcomes acquired through nerve organs networks. Benefits reveal that GP functions equally inside explanation as well as generalization.Prather J.C and Lobach D.F (1997) [4] has used the tactics of web data mining (also called Understanding Uncovering inside databases) to search for interactions within a significant professional medical database. They summarize the particular functions interested in mining any professional medical database like files warehousing, files query& cleansing and files analysis.Paripinelli R.S,Lopes H.S and Freitas A.A92001)[5] has described a formula pertaining to concept breakthrough discovery inside data bank known as AntMiner.The intent of your algorithm criteria will be the removal of category procedures to help be relevant to undetectable details like a choice aid.AntMiner has been applied to health data bank to get category rules.Li J, Fu.A, W.c and He.H (2005) [6] has discussedthe trouble connected with acquiring probability habits inside professional medical results are discussed. Danger habits by way of a stats metric, distant relative probability which has been commonly used inside epidemiological exploration will be defined. A anti-monotone property to get mining exceptional probability pattern places is usually studied. The criteria features earned a few beneficial outcomes for professional medical researchers.Ghazavi S.N and Liao T.W (2008) [7] has presenteda details mining study associated with health care details using furred modelling procedures designed to use aspect subsets picked out by simply many methods. About three furred modelling procedures like the furred k-nearest neighbor algorithm criteria, a furred clustering based modelling and also the flexible multilevel based furred inference process are generally employed.Delan.D, Walker.G and Kadam.A (2005) [8] has stated thatimaginative biomedical systems, greater explanatory prognostic components will be calculated along with recorded in this paper. Available electronic advancements to build up idea versions with regard to cancer of the breast survivability are used.Pamulaparty, L., Rao, C. G., and Rao, M. S. (2016)[9] has discussedthat bunch evaluation isolates the information in groups that are important, practical as well as both. It is additionally utilized to be a place to start intended for some other reasons of knowledge summarization. These people reviewed a few very basic algorithms including K-means, Hairy C-means, Hierarchical clustering to think of groupings, and use Ur files mining tool. Your outcomes are subjected to testing within the datasets that is On line News flash Popularity, Eye Files Fixed plus coming from UCI files repository plus mi RNA dataset intended for health-related files analysis. All datasets appeared to be examined with different clustering algorithms. Every single formula have their own originality plus antithetical behavior.Sebag, M., Azé, J., and Lucas, N. (2016) [10] has taken from a NP finish marketing qualification to get supervised learning, the location underneath the ROC curve. This specific marketing qualification, handled together with progression approaches, will be experimentally when compared to basique probability qualification handled simply by quadratic marketing in Assistance Vector Machines. Very similar answers are attained with some standard difficulties from the Irvine repository, inside half this SVM computational cost.Yu, H., Vaidya, J., and Jiang, X. (2006) [11] has proposedvarious algorithms that handed out understanding breakthrough, even though furnishing makes certain within the non-disclosure connected with data. Group is a vital data mining difficulty pertinent in most varied domains. The aim of explanation should be to build a model which could forecast a characteristic (binary credit with this work) in line with the remainder of attributes. People propose a proficient plus protected privacy-preserving algorithm pertaining to assistance vector product (SVM) explanation through up and down partitioned data. Caruana, R., and Niculescu-Mizil, A.(2004)[12] has studiedvarious algorithms that handed out understanding breakthrough, even though furnishing makes certain within the non-disclosure connected with data. Group is a vital data mining difficulty pertinent in most varied domains. The aim of explanation should be to build a model which could forecast a characteristic (binary credit with this work) in line with the remainder of attributes. People propose a proficient plus protected privacy-preserving algorithm pertaining to assistance vector product (SVM) explanation through up and down partitioned data.

4. EXPERIMENTAL SETUP

MATLAB (matrix laboratory) is a multi-paradigm numerical processing environment and fourth-generation programming language. A proprietary programming language developed by Math Works, MATLAB allows matrix manipulations, plotting of data and functions, execution of algorithms, creation of end user interfaces, and interfacing with programs written in other languages such as C, C++, C#, Java, Fortran and Python.

Waikato Environment for Knowledge Analysis (**Weka**) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. It is a workbench which consists of collection of visualization tools and algorithms for data analysis and predictive modeling, along with graphical user interfaces for accessing the functions easily.

The Intel's DUAL CORE processor is used along with WINDOWS 7 with 2 GB RAM and 500 GB hard disk.

The data sets used for reference in order to evaluate the performance of proposed algorithm are:

Breast Cancer [26]-Breast cancer generally develops from breast tissue. Signs of breast cancer includes a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin.

Diabetes [26]-Diabetes is a metabolic disorder which is characterized by high blood sugar, insulin resistance, and relative lack of insulin. Common symptoms include increased thirst, frequent urination, and unexplained weight loss

E-coli [26]-Escherichia coli is a gram-negative, facultative anaerobic, rod-shaped bacterium of the genus Escherichia which is commonly found in the lower intestine of warm-blooded organisms. Most E. coli strains are harmless, but some serotypes can cause serious food poisoning in their hosts.

Heart Disease [26]-Cardiovascular disease (CVD) is a class of diseases which involves the heart or blood vessels. This is generally caused by high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol consumption.

5. METHODOLOGY AND RESULTS

5.1 Methodology

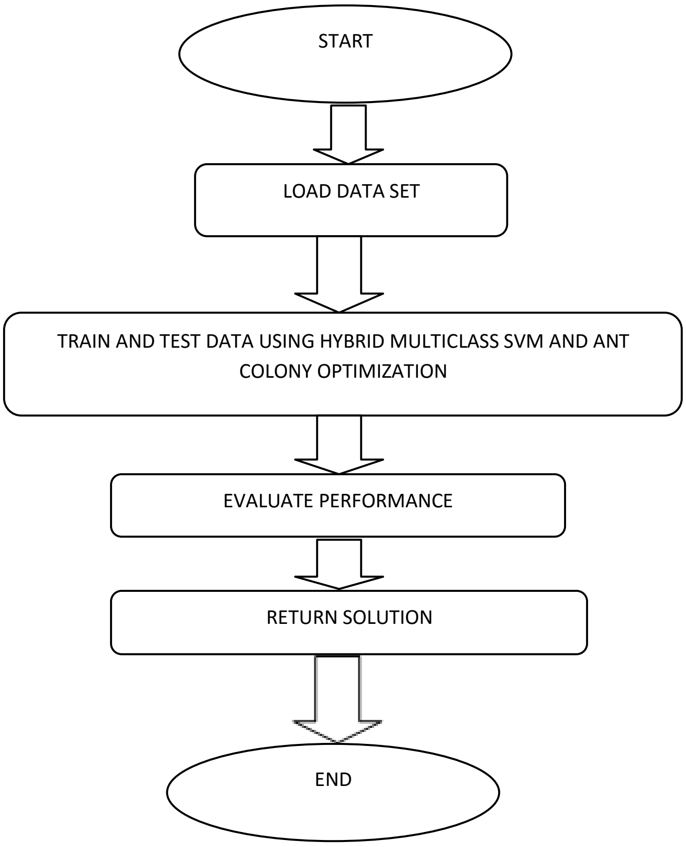
START
LOAD DATA SET
TRAIN AND TEST DATA USING HYBRID MULTICLASS SVM AND ANT COLONY OPTIMIZATION
EVALUATE PERFORMANCE
RETURN SOLUTION
END

Fig2.Proposed methodology

5.2 Performance Analysis

This paper has designed and implemented the proposed technique in MATLAB tool u2013a. The evaluation of proposed technique is done on the basis of following metrics i.e. Accuracy, F-measure, true positive rate and false positive raate. A comparison is drawn between all the parameters with proposed algorithm and figures shows all the results.

1. Correctly Classified Instances-It is defined as the number of instances which are classified as correct from the total number of instances used.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

2.Incorrectly Classified Instances-It is defined as the number of instances which are classified as incorrect from the total number of instances used.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

3. Kappa Statistics-Kappa statistics is the measure that determines inter rater agreement for qualitative items. Cohen's kappa measures the agreement between two raters who each classify *N* items into *C* mutually exclusive categories. It is calculated as-

$$K = \frac{p_o - p_e}{1 - p_e}$$

4. Accuracy-Accuracy refers to the ability of the model to correctly predict the class label of new or unseen data.It is calculated as-

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\frac{TP + TN}{TP + TN + FP + FN}$$

where,

True positives (TP) =No. of correct classifications predicted as yes(or positive).

True negatives (TN) =No. of correct classifications predicted as no(or negative).

False positive (FP) =No. of incorrect classifications predicted as yes(positive) when it is actually no(negative).

False negative (FN) =No. of incorrect classifications predicted as no(negative) when it is actually yes(positive).

5. F-Measure-It is the measure that combines precision and recall. It is the harmonic mean of precision and recall.It is calculated as-

$$F = 2 * \left[\frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}\right]$$

where, Precision and Recall are defined from the eqs below

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP + FN}$$

6.RESULTS AND PERFORMANCE EVALUATION

The results obtained from the proposed model are for Breast cancer, Diabetes, E-coli and Heart disease data sets. The detailed discussion of the results obtained for the data sets.

1. Discusssion for Wisconsin Breast Cancer data set

The accuracy of proposed Hybrid multiclass SVM and LAD tree model(shown in "bold" text in table 3) is computed as 96.5 for Breast Cancer data set.The accuracies of other classification models have also been computed in the table. It is clearly evident from the results obtained that the accuracy of proposed model is computed as the highest among all other classification models. In addition to accuracy, another important metric i.e. Kappa Statistics is also evaluated and it also shows the highest reading of 0.9157 from all other existing classifiers.

Another important metric ROC is also calculated. The ROC area of proposed model i.e. 0.994 is also highest among all other existing classifiers.

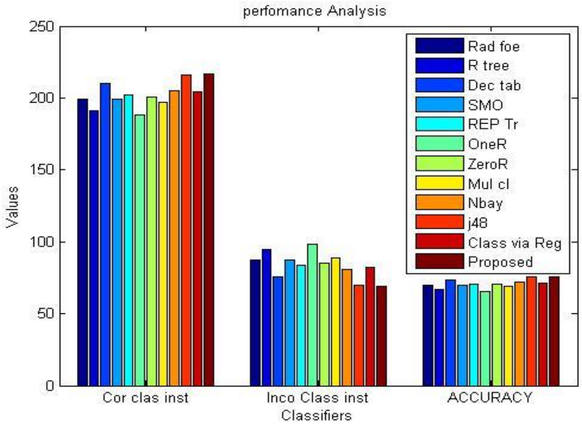

Fig 3.Analysis of corr class istances, Inc class ins,Accuracy

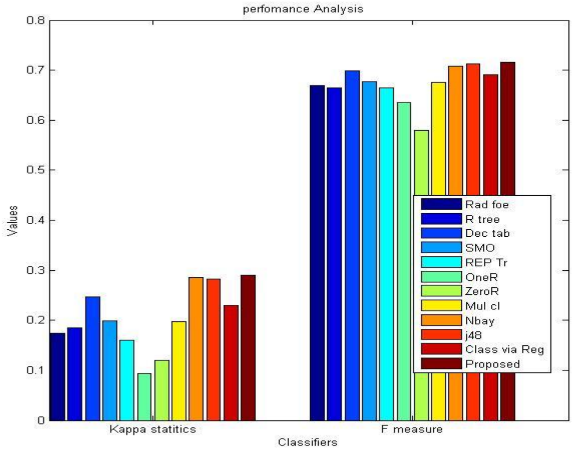

Fig 4.Analysis of kappa stat and F-measure

2. Discussion for Diabetes data set

The accuracy of proposed Hybrid multiclass SVM and LAD tree model (shown in "bold" text in table 4) is computed as 99.2 for Diabetes data set. The accuracies of other classification models have also been computed in the table. It is clearly evident from the results obtained that the accuracy of proposed model is computed as the highest among all other classification models. In addition to accuracy, another important metric i.e. Kappa Statistics is also evaluated and it also shows the highest reading of 0.9827 from all other existing classifiers.

Another important metric ROC is also calculated. The ROC area of proposed model i.e.1 is also highest among all other existing classifiers.

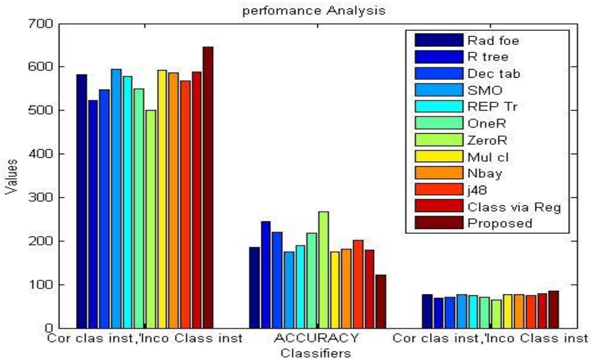

Fig 5.Analysis of corr class istances, Inc class ins,Accuracy

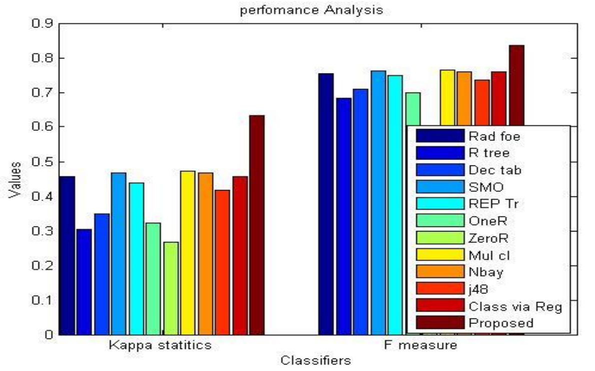

Fig 6.Analysis of kappa stat and F-measure

3. Discussion for E-coli data set

The accuracy of proposed Hybrid multiclass SVM and LAD tree model(shown in "bold" text in table 5) is computed as 99.4 for E-coli data set. The accuracies of other classification models have also been computed in the table. It is clearly evident from the results obtained that the accuracy of proposed model is computed as the highest among all other classification models. In addition to accuracy, another important metric i.e. Kappa Statistics is also evaluated and it also shows the highest reading of 0.9918 from all other existing classifiers.

Another important metric ROC is also calculated. The ROC area of proposed model i.e.1 is also highest among all other existing classifiers.

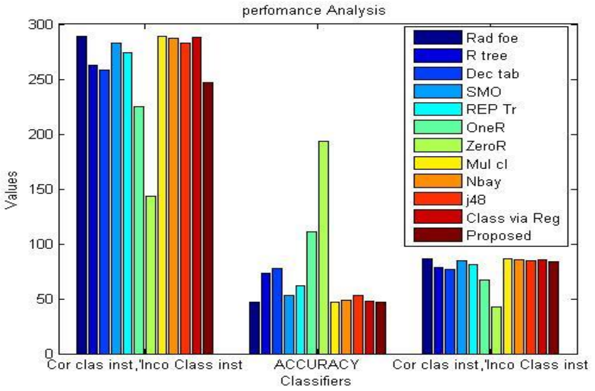

Fig 7.Analysis of corr class istances, Inc class ins,Accuracy

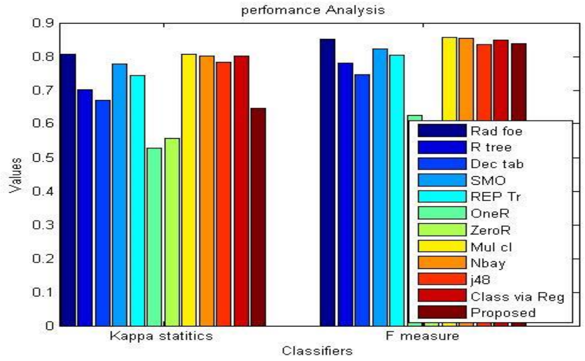

Fig 8.Analysis of kappa stat and F-measure

4. Discussion for Heart Disease data set

The accuracy of proposed Hybrid multiclass SVM and LAD tree model (shown in "bold" text in table 6) is computed as 98.3 for Heart disease data set. The accuracies of other classification models have also been computed in the table. It is clearly evident from the results obtained that the accuracy of proposed model is computed as the highest among all other classification models. In addition to accuracy, another important metric i.e. Kappa Statistics is also evaluated and it also shows the highest reading of 0.9632 from all other existing classifiers.

Another important metric ROC is also calculated. The ROC area of proposed model i.e.0.998 is also highest among all other

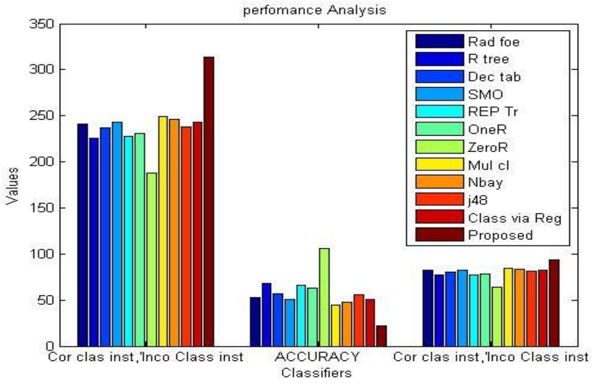

Fig 9.Analysis of corr class istances, Inc class ins,Accuracy

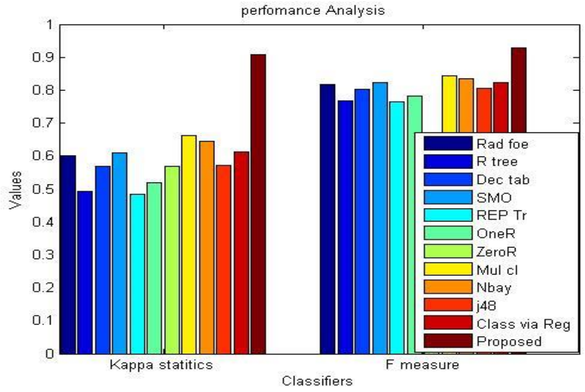

Fig 10.Analysis of kappa stat and F-measure

6.CONCLUSION

In this paper, we have analyzed existing ANT colony optimization and random forest tree based data mining. The proposed unsupervised filtering by ACO and multiclass SVM based data mining gives better results. This paper has shown comparison between exiting and proposed data mining techniques on the basis of parameters like Correctly classified instances,Incorrectly classified instances,Kappa statistics,Accuracy and F-measure.This proposed technique of data mining shows better results as compared to the existing technique.

REFERENCES

[1]Duan, K. B., Rajapakse, J. C., Wang, H., & Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience, 4(3), 228-234.

[2]Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. Artificial intelligence in medicine, 26(1), 1-24.

[3]Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. IEEE Transactions on Evolutionary Computation, 5(1), 17-26.

[4]Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse. In Proceedings of the AMIA annual fall symposium (p. 101). American Medical Informatics Association.

[5] Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2001, July). An ant colony based system for data mining: applications to medical data. In Proceedings of the genetic and evolutionary computation conference (GECCO-2001) (pp. 791-797).

[6] Li, J., Fu, A. W. C., He, H., Chen, J., Jin, H., McAullay, D., & Kelman, C. (2005, August). Mining risk patterns in medical data. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 770-775). ACM.

[7] Ghazavi, S. N., & Liao, T. W. (2008). Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine, 43(3), 195-206.

[8] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

[9] Pamulaparty, L., Rao, C. G., & Rao, M. S. (2016). Cluster Analysis of Medical Research Data using R. Global Journal of Computer Science and Technology, 16(1).

[10] Sebag, M., Azé, J., & Lucas, N. (2003, October). ROC-based evolutionary learning: Application to medical data mining. In International Conference on Artificial Evolution (Evolution Artificielle) (pp. 384-396). Springer Berlin Heidelberg.

[11]Yu, H., Vaidya, J., & Jiang, X. (2006, April). Privacy-preserving svm classification on vertically partitioned data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 647-656). Springer Berlin Heidelberg.

[12] Caruana, R., & Niculescu-Mizil, A. (2004, August). Data mining in metric space: an empirical analysis of supervised learning performance criteria. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 69-78). ACM.

[13] Delen, D. (2009). Analysis of cancer data: a data mining approach. Expert Systems, 26(1), 100-112.

[14] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

[15] Raikwal, J. S., s& Saxena, K. (2012). Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. International Journal of Computer Applications, 50(14).

[16] Moses, D. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, 42(2).

[17] Nichat, A. M., & Ladhake, S. A. (2016). Brain Tumor Segmentation and Classification Using Modified FCM and SVM Classifier. Brain, 5(4).

[18] Verma, L., Srivastava, S., & Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. Journal of Medical Systems, 40(7), 1-7.

[19] Li, D. C., Liu, C. W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. Computers in biology and medicine, 40(5), 509-518.

[20]Kazemzadeh, R. S., & Sartipi, K. (2005, September). Interoperability of data and knowledge in distributed health care systems. In 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05) (pp. 230-240). IEEE.