
Review Paper on Classification Algorithms in Data Mining

Poonam Rani

Student, School of Computer Science & Engineering,
Lovely Professional University, Phagwara,
Punjab, India. Pin Code: 144411
Email Id: ranipoonam405@gmail.com

Kamaldeep Kaur

Assistant Professor, School of Computer Science & Engineering,
Lovely Professional University, Phagwara,
Punjab, India. Pin Code: 144411
Email Id: kamaldeep.23523@lpu.co.in

Abstract-

Data mining is the emerging field in computer science. Data mining includes the identification of hidden predictive patterns from large databases. Data mining parameters includes association, classification, clustering, forecasting. There are different classification algorithms. This is a review paper regarding various classification algorithms of data mining and enhancement of those algorithms which has been applied to different datasets. This paper will give insights about various classification algorithms and as well as their enhancements. Numerous researcher's implemented different approaches to enhance the correctness and accuracy of the particular classification algorithm. They have different methods to implement a specific classification algorithm like to use a hybrid approach- combine two or more algorithms in a single one. To create a new algorithm by using the combination of clustering and classification in data mining or may be by adding a new formula construct a new algorithm for a given data set and produce another good solution for the previous available algorithm.

Keywords – Data Mining, Classification, naïve based classifier, decision tree classifier, SMO.

I. INRODUCTION

Data mining is the phenomenon of finding and taking out precious information from a mammoth data set; we can say that it is a method of mining useful knowledge from data. We have different applications of data mining those are: market analysis, fraud detection, customer retention, production control, science exploration etc. It is an interdisciplinary sub field of computer science. Data mining include six classes to mine useful information. Anomaly detection - to find some unusual data that might contain any error or some outlier detection comes under this. Association - market basket analysis comes under this category. Here we find the relationship between variables which associates their co-occurrence from the collection of the items.

Clustering - is the process of bringing out patterns and structures from the data that are in one way or another "similar", without using known structures. It can be homogeneous or heterogeneous Classification- to classify the data item according to the build classifier. Regression - attempt to identify which model is having less number of errors for the given datasets. Summarization - It is the last step where we have to provide a proper visualization and reporting of the data set. The following diagram shows us the various data mining techniques. In this paper we are going to discuss about classification and its different methods which all are comes under predictive type. Most of the researcher either used a individual algorithms to analysis the performance and in the last they just compare it.

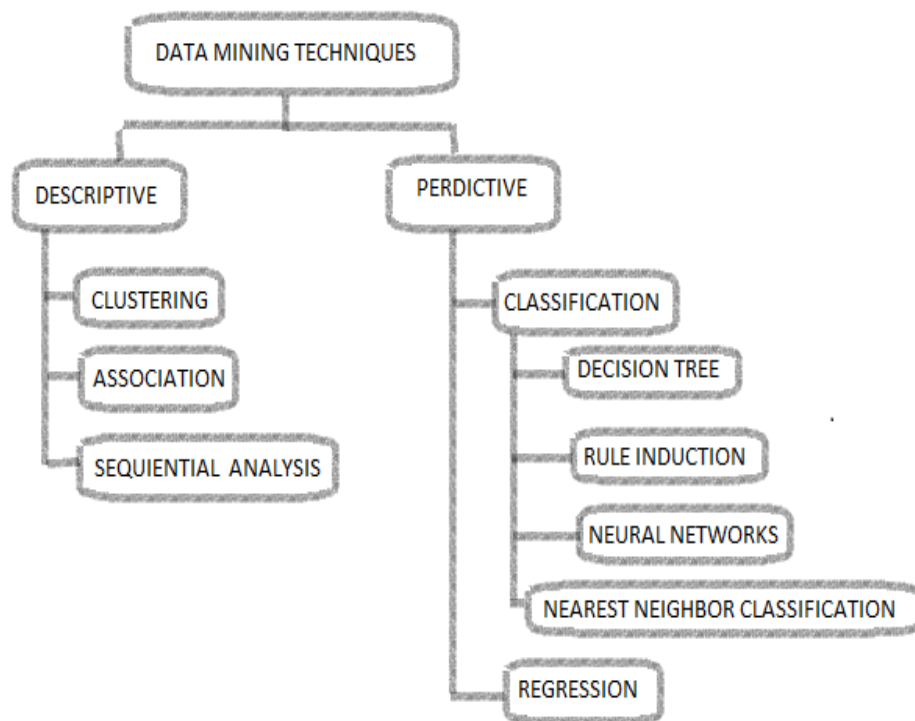


Figure 1: Data Mining Techniques

Classification is the method of ascertaining a target class for the given dataset. The prime objective of the classification is to predict an accurate class label for the dataset. We are having different classification methods in data mining and these all methods divide our dataset into a predictable form. There is an illustration of classification where a bank loan officer needs to examine the data in order to cognize the risky and safe loan applicant. In this example the loan officer actually wants to know the categorical class. There are steps to apply the classification.

- Building the Classifier
- Using Classifier for Classification



Figure 2: Classification Process - 1. Model Construction

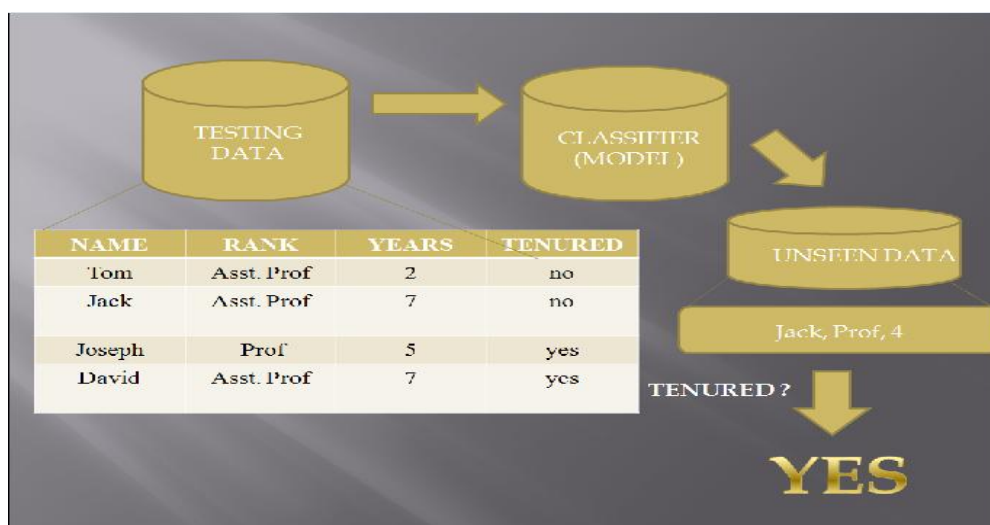


Figure 3: Classification Process - 2. Use of Model in Prediction

For building a classifier model we need to train our dataset. In this step the classification algorithm builds the classifier. We need to prepare a training set from the given dataset. Each row that includes in the training set is referred to as a label or class. To use this classifier or model we have to use the classification algorithm on the new dataset and use it for further prediction and to identify class labels. We are having different classification algorithms i.e. rule based classifier, naïve based classifier, decision tree, k-means and SMO (sequential minimal optimization). In this review paper three referenced papers contains methodology of Hybrid

Approach. Hybrid approach or Hybrid Intelligent System means to add two or more existing algorithms such a way that the resultant algorithm includes the features of both the algorithms and most of the time the Hybrid algorithm contains more accuracy and correctness than the individual algorithms. This is very vital method which is used in most of the researches. Even some of the researchers used the combination of both classification and clustering to develop the more accurate and efficient algorithm. We are having four domains to apply this approach as well.

1. Hybrid Optimization Method
2. Nature Inspired Systems
3. Combined Classifier
4. Uncertainty Management

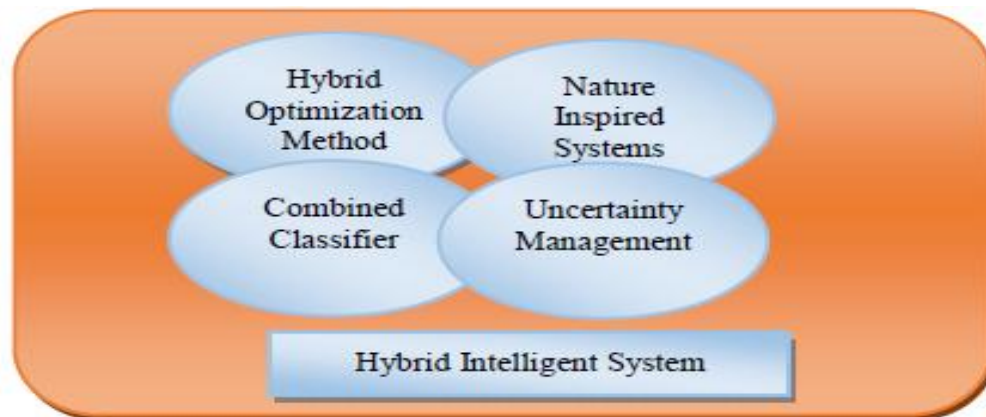


Figure 4 : Domains of Hybrid Intelligent System [3]

II. LITERATURE SURVEY

Sneha Chandra et al. [1.] present an adaptive classification method to ameliorate the classification accuracy for medical data mining. Data provide details of non-infectious disease Diabetes and the other infectious disease Tuberculosis. In this author has used Laplacian Correction to Decision tree algorithm and Naïve Bayesian. They have been applied individual algorithms as well as the adaptive algorithms for same data set. She actually compared the results of adaptive algorithm to [1] RBC (Rule-Based Classifier), DTC (decision-Tree Classifier) and NBC (Naïve –Bayesian Classifier). They have been achieved 90% accuracy using adaptive algorithm.

NeelamSinghal et al. [2.] worked on the performance improvement of classification using hybrid algorithm in data mining. Author has merged Genetic Programming along with Decision tree algorithm to enhance the accuracy, comprehensibility as well as timing of the classification process. [2] This research suggests an algorithm that has benefits of both Clustering and Feature Selection. They extract dataset in the WEKA Software. Then use attribute selection which is used for feature selection. They have used the J48 classifier and then apply the best feature selection on the dataset. For combining clustering simple KMeans cluster has been used. In last they used Genetic Programming to classify data. The algorithm has been implemented on the 5 different download datasets from UCI repository. As a result the hybrid approach contains better performance as compared to Decision Tree as well as Genetic programming individually.

Thamilselvanv et al. [3.] published a review paper on image classification in data mining using hybrid data mining algorithm. In this review they explained the comparison of various hybrid classification algorithms. They have been chosen various datasets to find the technique and accuracy of the different algorithms. In this paper proposed hybrid approaches are –[3] Genetic Algorithm and Support Vector Machine, Extreme K-Means Algorithm and Effective Extreme Learning Machine, Naïve Bayes and Support Vector Machine, Support Vector Machine and Classification regression tree. Accuracy are 89%, 90%, 98%, 84% and 90% respectively. This performance has been implemented on numerous image data sets. As the conclusion we get Naïve Bayes and Support Vector Machine hybrid approach having improved accuracy in image classification that is 98%.

MangeshM.Panchwagh et al. [4.] present a music genre classification using data mining algorithm. In this paper they compared the single classification algorithm for music categorization and for ensemble classification model for music categorization. Feature selection and classification are the two important steps to classify any music data. Author has extracted three features [4] Mel frequency Cepstral Coefficients (MFCC), Linear Productive Coefficient (LPC) and Zero Crossing rate (ZCR) of various types of songs (pop, jazz, rock etc.). Data has been offered by TU Technical University Dortmund. The result shows that SMO classifier with PKIDiscretize pre-processing method gives the highest accuracy of 100% for MFCC. Among the three features MFCC, LPC and ZCR the MFCC gives the most accurate results.

Jaitendra Jain et al. [5.] present data mining classification approach to detect worms. Naïve Bayesian classifier has been used to identify a particular class for unknown worm. In this paper three unknown worm classes are used – type of scheduler used, type of protection used and scan frequency for the worm detection. [5] Data was collected by NP and AV net protector. Then after performing various tests for virus using multiple parameters a data set was created. Then Naïve Bayesian classifier has been applied on this data set. Finally the category of the worm comes under either executable or script type. Simple calculation has been performed. Further they

described that using more parameters and more number of classes we can extend this work as well.

Akanksha Ahlawat et al. [6.] proposed a novel hybrid algorithm for improving classification in data mining. For this algorithm they have two ideas first to make cluster and second one is to use classification. They have been used K-Means clustering for dividing the data samples into pre-defined clusters. Then format the decision tree for each cluster. This algorithm was tested on various real life data sets available from UCI repository. For the result evaluation they have used two parameters - Percentage accuracy and Cohen's kappa where value range from 0 to 1. Using this approach the accuracy of Decision Tree has been increased. The issue of burdening decision tree with large data set has been abolished. Overall percentage accuracy is 68.41% and Cohen's kappa is 0.31 for diabetes. For iris percentage accuracy is 92.41% and Cohen's kappa is 0.91.

Rana Alaa El-Den Ahmed et al. [7.] present a performance analysis of classification methods for online shopping attributes and behavior by applying data mining. Here they compared eleven different algorithms to search the best fit classifier according to consumer. Those algorithms are- Bayes Net, naïve Bayes, K star, classification and clustering, filtering classifier, END, JRIP, RIDOR, Decision Table, J48, Simple Cart. Data set has been collected from highly repudiated online shopping agency. Ten-fold cross validation is applied for checking the accuracy. WEKA tool kit has been used for implementation. Results contain 87.13% accuracy for Decision Table and time taken is 0.24 seconds. This is the most accurate classifier for online shopping.

Davinder Kaur et al. [8.] wrote a review paper of decision tree algorithms that are ID3 and C4.5. In this paper they have been described the terminology to create a decision tree. They concentrate on elements of the decision tree. They compared ID3 and C4.5. In the end they concluded that C4.5 is more precise and takes lesser execution time for mining data. They also provided the error rate difference that shown that C4.5 is best approach as compared to ID3.

Supreetkaur et al. [9.] published a review paper for classification techniques for spotting of lung cancer. She provides reviews of various researchers who had published research in the field of lung cancer. The main propose was to identify a model for early spotting and correct diagnosis for providing help to doctor to prevent the patient in the better way. The main thing they concluded that is to apply hybrid classification scheme and create data mining tool well suited according to the patient disease and its treatment.

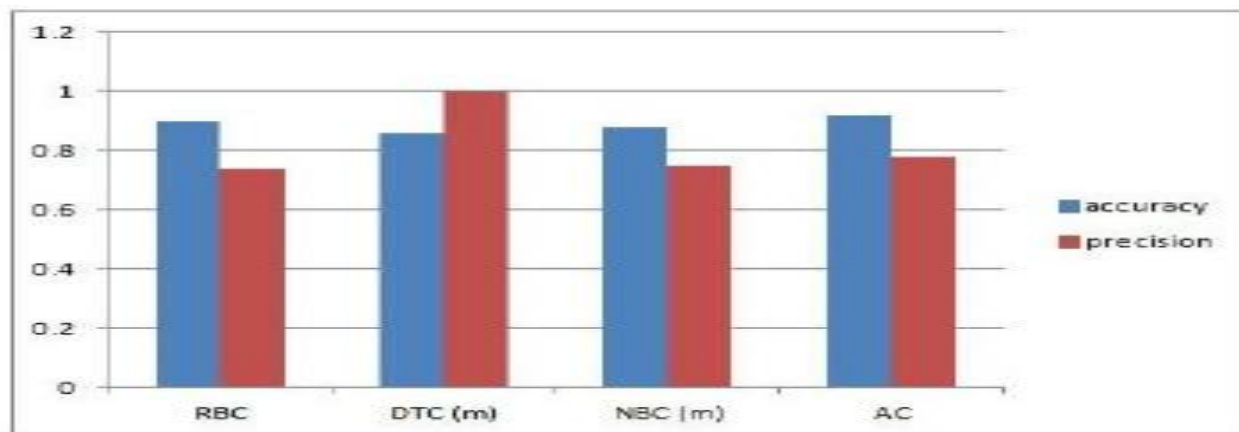


Figure 5: [1] Analysis of accuracy, precision of different classifiers (Diabetes Dataset)

S. No.	Author Name	Method Used	Results and Future Work
1.	Sneha Chandra, Maneet Kaur [1]	Compared the results of adaptive algorithm to RBC (Rule-Based Classifier), DTC (decision-Tree Classifier) and NBC (Naïve –Bayesian Classifier)	Achieved 90% accuracy using adaptive algorithm In future work author claimed to achieve higher classification accuracy for the sample medical datasets also working on other medical disease.
2.	NeelamSinghla, Mohd. Ashraf [2]	For combining clustering simple KMeans cluster has been used. In last they used Genetic Programming to classify data.	As a result the hybrid approach contains better performance as compared to Decision Tree and Genetic programming individually. In future work we can compare this hybrid algorithm with existing one and modify them to achieve better performance
3.	Thamilselvan, Dr.J.G.R.Sathiaselan [3]	Review paper on image classification using hybrid data mining algorithm. They explained the comparison of various hybrid classification algorithms.	Getting Naïve Bayes and Support Vector Machine hybrid approach has improved accuracy in image classification that is 98%.
4.	MangeshM.panchwagh, VijayD.Katkar [4]	In this paper they compared the single classification algorithm for music categorization and for ensemble	The result shows that SMO classifier with PKIDiscretize pre-processing method gives the highest accuracy of 100% for MFCC(Mel frequency Cepstral

		classification model for music categorization.	Coefficients)
5.	Jitendra Jain, Dr. Parashu Ram Pal [5]	Presents data mining classification approach to detect worms. Naïve Bayesian classifier has been used to identify a particular class for unknown worm.	Finally the category of the worm comes under either executable or script type. Further they described that using more parameters and more number of classes we can extend this work as well.
6.	Akanksha Ahlawat , Bharti Suri [6]	For this algorithm they have two ideas first to make cluster and second one is to use classification. They have been used K-Means clustering for partition the data samples into pre-defined clusters. Then format the decision tree for each cluster.	Overall percentage accuracy is 68.41% and Cohen's kappa is 0.31 for diabetes. For iris percentage accuracy is 92.41% and Cohen's kappa is 0.91. In future we can use this approach for the big data for the use of map reduce to handle the huge dataset.
7.	Rana Alaa El-Deen Ahmed, M.Eleman. Shehab, Shereen Morsy, Nermeen Mekawie [7]	Here they compared eleven different algorithms to find the best fit classifier according to consumer. Those algorithms are- Bayes Net, naïve Bayes, K star, classification via clustering, filtering classifier, END, JRIP, RIDOR, Decision Table, J48, Simple Cart.	Results contain 87.13% accuracy for Decision Table and time taken is 0.24 seconds. This is the most accurate classifier for online shopping.
8.	Davinder Kaur, Rajeev Bedi, Dr, Sunil Kumar Gupta [8]	They have been described the terminology to create a decision tree. They compared ID3 and C4.5.	C4.5 came out to be accurate and consumed less execution time to mine data. The error rate difference that shown that C4.5 is best approach as compared to ID3.
9.	Supreet Kaur , Amanjot Kaur Grewal [9]	Provides reviews of various researchers who had published research in the field of lung cancer. The main propose was to identify a model for early detection and correct diagnosis for providing help to doctor.	to apply hybrid classification scheme and create data mining tool well suited according to the patient disease and its treatment.

Figure 6: Summary of all the papers.

III. CONCLUSION

This is review paper on various classification algorithms in data mining. As we can see from the figure 6 which represents the overall summary for all the given research papers in this review paper. All researchers have different approaches to enhance the accuracy and correctness of the algorithms. Most of them used the hybrid approach that means to collect the two or more algorithms in single one and then implement it on the data set. Some of them have changed the calculations and provides the better approach to us. In some recherche they just provide the comparative study of different algorithms for the same data set. Few of the researcher paper contain the analysis and predictive analysis information just to find the value of some attributes. Other than this we can also say that we can use these classifications algorithms on different data sets like medical data to predict the disease and provide diagnosis according to that, worms detection propose, music categorization, feature selection. We can also combine the clustering and classification to produce the better approach for the classification in data mining. As given in the above table we can check for the given future works and further enhance the performance of the specific algorithm.

REFERENCES

1. Sneha Chandra, Maneet Kaur “Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining” 2nd International Conference on Computing for Sustainable Global Development(INDIACom 2015) IEEE.
2. NeelamSinghla, Mohd. Ashraf “Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm” International Conference on Computing, Communication and Automation (ICCCA2015) IEEE.
3. Thamilselvan, Dr.J.G.R.Sathiaseelan “Image Classification using Hybrid Data Mining Algorithms – A Review ” IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS’15. IEEE 2015.
4. MangeshM.panchwagh,VijayD.Katkar “Music Genre Classification Using Data Mining Algorithm” 2016 conference on advances in signal processing(CASP) Cumminm College of engineering for Women ,Pune.Jun 9-11,2016, IEEE.
5. Jitendra Jain, Dr. Parashu Ram Pal “Detecting Worms Based on Data Mining Classification Technique” International Journal of Engineering Science and Computing, May 2017, Volume 7 Issue No.5IJESC.

6. Akanksha Ahlawat , Bharti Suri “Improving Classification in Data mining using Hybrid algorithm” 2016 IEEE.
7. Rana Alaa El-Deen Ahmed, M. Eleman. Shehab, Shereen Morsy, Nermeen Mekawie “Performance study of classification algorithms for online shopping attitudes and behavior using data mining” 2015 Fifth International Conference on Communication Systems and Network Technologies. 2015 IEEE.
8. Davinder Kaur, Rajeev Bedi, Dr, Sunil Kumar Gupta “Review of Decision Tree Data Mining Algorithms: ID3 and C4.5” Proceeding of International Conference on Information Technology and Computer Science July 11-12, 2015, ISBN: 9788193137307 .2015 IEEE.
9. Supreet Kaur , Amanjot Kaur Grewal “A Review paper on Data Mining Classification Techniques for Detection of Lung Cancer” international Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue:11-Nov-2016.