IMPROVE THE EFFICIENCY OF CLASSIFICATION ALGORITHM IN DATA MINING

Dissertation proposal submitted in partial fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

By

POONAM RANI

11610308

Supervisor

MISS. KAMALDEEP KAUR

Assistant Professor, Lovely professional university

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

December 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0359

Supervisor Name : Kamaldeep Kaur **UID :** 21976 **Designation :** Assistant Professor (Contract Basis)

Qualification : __________________________ **Research Experience :** __________________________

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Poonam Rani	11610308	2016	K1637	8196826231

SPECIALIZATION AREA : Programming-II **Supervisor Signature:** ____________________

PROPOSED TOPIC : Improve the efficiency of classification algorithm in data mining

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.50
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.75
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	6.75
5	Social Applicability: Project work intends to solve a practical problem.	6.50
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.50

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NA
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Improve the efficiency of classification algorithm in data mining**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

TABLE OF CONTENTS

DECLARATION

I hereby declare that the dissertation proposal entitled, Improve the efficiency of classification algorithm in data mining submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: ______

Investigator

Regn. No. 11610308

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation proposal entitled "**IMPROVE THE EFFICIENCY OF CLASSIFICATION ALGORITHM"**, submitted by **Poonam Rani** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor
Miss. Kamaldeep Kaur
Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: ________________
   HoD Name: ___________________
   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: ________________

   Name: ________________

   Affiliation ________________

   Date: ________________

   Internal Examiner

   Signature ________________

   Name: ________________

   Date: ________________

ACKNOWLEDGEMENT

All praise in the name of almighty God, who give us in the darkness and help in difficulties. The dissertation is the result of full semester of work whereby I have been accompanied and supported by many people. It is a pleasant aspect to that I have the opportunity to express my gratitude for all of them.

I am also extremely indebted to my guide **Miss Kamaldeep Kaur** (Assistant professor, Department of Computer science, Lovely Professional University, Phagwara). I am very much thankful to **Miss Kamaldeep Kaur** for picking me as a student at the critical stage of my masters. I warmly thank her for her valuable advice, constructive criticism and her extensive discussions around my work.

I expand my thanks to my friends and family who always kept my spirits up with their extended love, affection and support at the time of my project work.

At last but not the least, I would like to pay high regards to the authors whose work I have consulted very often during my project work. And I would like to thank Lovely Professional University that provided me the road for the completion of my degree in this particular field.

Poonam Rani

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATION

ABSTRACT

Data mining is very important part of the computer science. It is a method through which we can extract the important patterns from the nudge data set. For doing this we are having a concept of KDD. Classification, clustering, regression, association all are different approaches to mine data. There are further so many types comes under these approaches. In classification one method for doing classification process is known as decision tree classifier. C4.5 and ID3 decision tree algorithms use the concept of entropy and information gain. There are different ways to use entropy. Using the concept of entropy we can do modifications in decision tree algorithms. As the consequences it would be more accurate as compared to previously available algorithms. There are many scholars who compare the different entropy techniques along with their limitations and advantages. But the combined entropy approach is not that much consider in these research papers. Using weka tool and net beans software we will implement the combined entropy in the field of data mining to improve the accuracy / efficiency of the classification algorithm (decision tree).

CHAPTER 1

INTRODUCTION

1.1 Data Mining:

Data mining is a method of finding and tacking out precious information gets from enormous 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- to analyze the market for mine collection and demanding products, fraud detection- to find some outlier if it enters in any ones business, science exploration- to do more research, can be in medical field to predict patient’s symptoms and treat them accordingly. 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 – it means to designing category and formation in the given data which implement some method beyond applying familiar formation of that data. 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 report we are going to discuss about classification and its different methods which all are comes under predictive type. Most of the researcher either used an individual algorithm to analysis the performance and in the last they just compare it.

1.2 KDD (Knowledge Discovery in Data Bases):

In other words we can say that data mining is also achieved through KDD process that is knowledge discovery in data base. There are following steps includes in KDD process:

1.2.1 Steps for KDD Process:

1. Selection
2. Preprocessing
3. Transformation

4. Data Mining
5. Interpretation/Evaluation

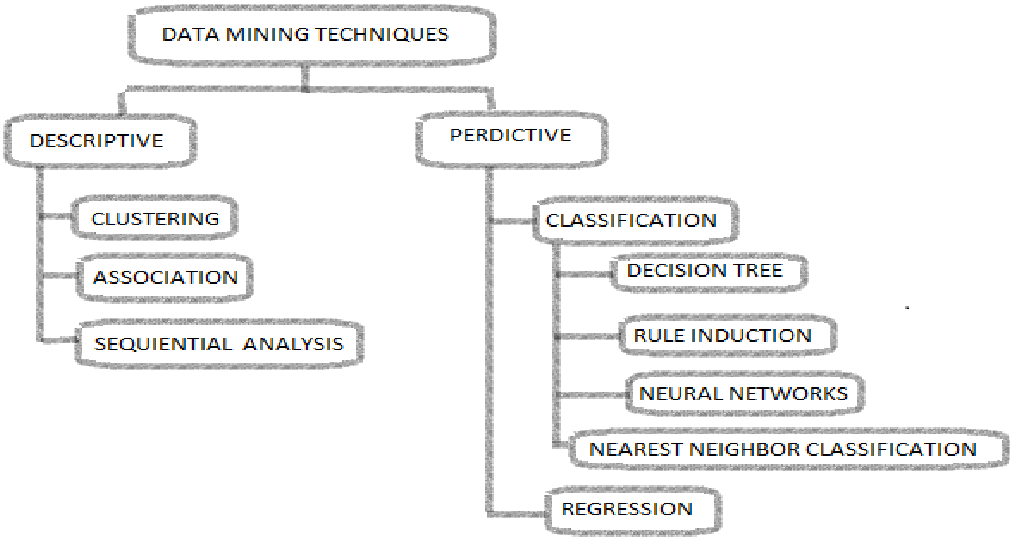


Figure 1.1: Data Mining Techniques

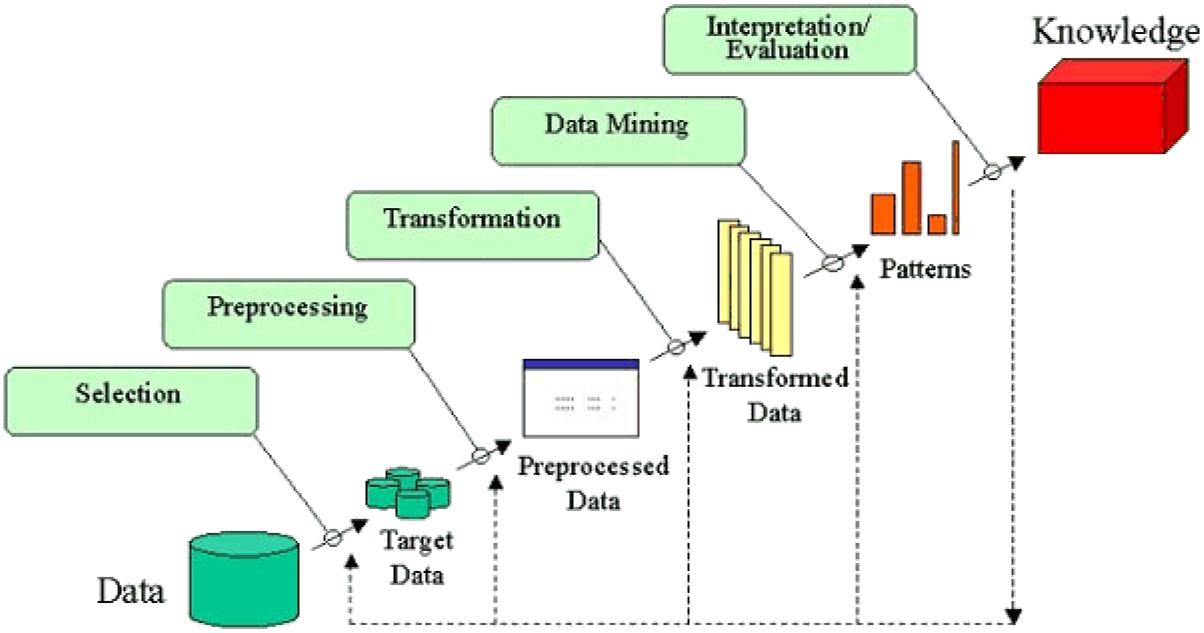


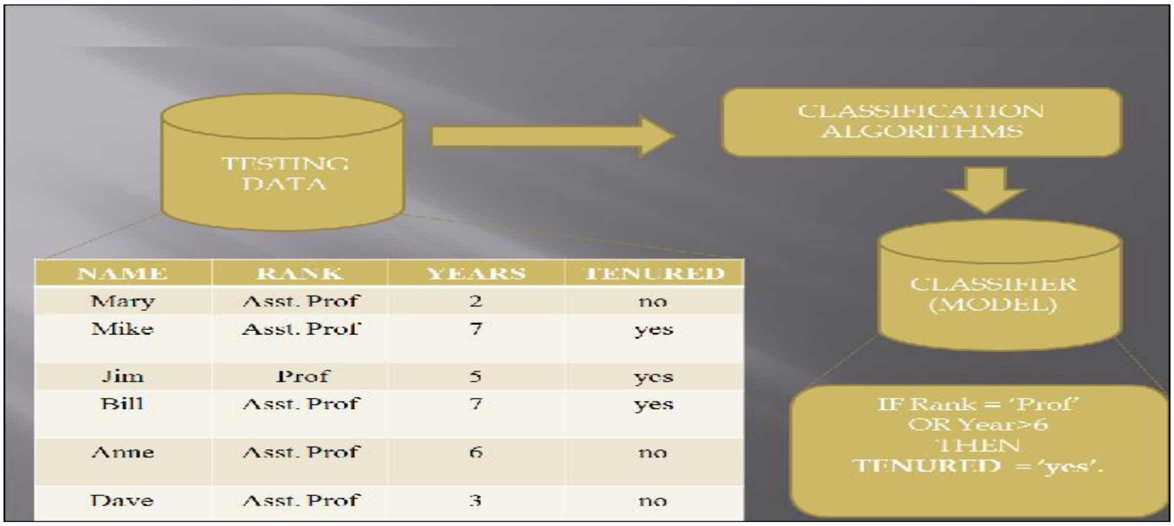
Figure 1.2: Process of KDD [41]

1.3 Classification:

It is a method of finding a target class for the given set of data. The major purpose of classification is to forecast a proper class label according to 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 example of classification where a bank organization needs to evaluate the data of the customer which wants to apply for loan and according to their given information we have to arrange them if they are safe or not for their business. In this example the loan officer actually wants to know the categorical class. There are steps to apply the classification.

1.3.1 Steps for Classification:

- Building the Classifier
- Using Classifier for Classification



TESTING DATA

CLASSIFICATION ALGORITHMS

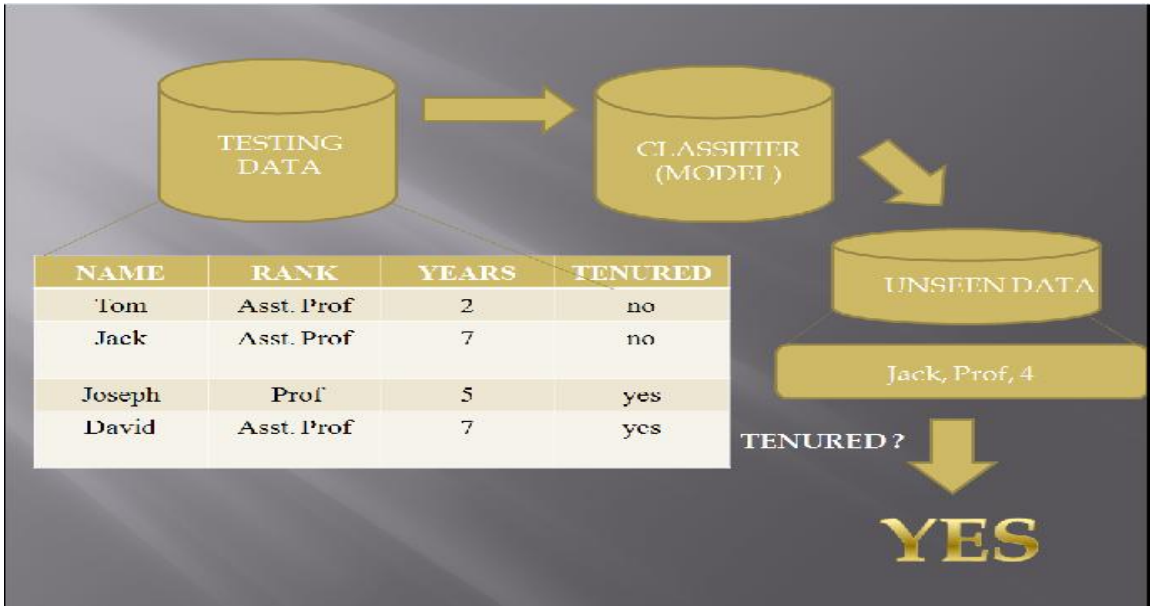
CLASSIFIER (MODEL)

NAME	RANK	YEARS	TENURED
Mary	Asst. Prof	2	no
Mike	Asst. Prof	7	yes
Jim	Prof	5	yes
Bill	Asst. Prof	7	yes
Anne	Asst. Prof	6	no
Dave	Asst. Prof	3	no

IF Rank = 'Prof'
OR Year>6
THEN
TENURED = 'yes'.

Figure 1.3: Building a Classifier Model Using Training Set

For building a classifier model we need to train our dataset. Here we need to build a new classifier with the help of classification algorithm. We need to prepare a training set from the given dataset. Each row that included in the training data set is point out particular allotted class. To use this classifier or model we have to apply the classifier 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 classifier, k-means classifier and SMO (sequential minimal optimization). As we know there are so many algorithms comes under classification, but the most common and effective approach is decision tree.

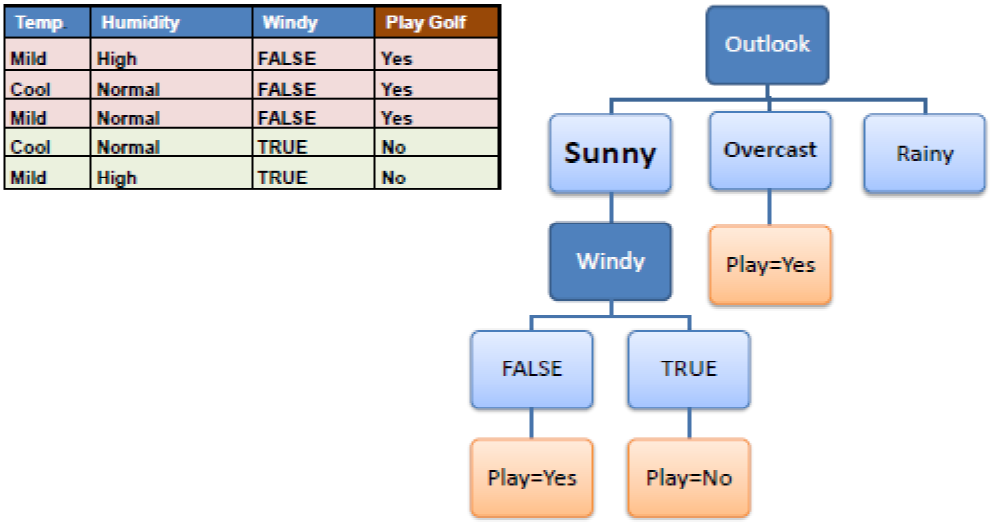


NAME	RANK	YEARS	TENURED
Tom	Asst. Prof	2	no
Jack	Asst. Prof	7	no
Joseph	Prof	5	yes
David	Asst. Prof	7	yes

Figure 1.4: Classifier Model Applying on Testing Data Set

1.4 Decision tree:

It is used for either classification or for regression. The leaf node describes a class for the given data which is previously not defined. There are different decision tree classifiers as well- ID3, C4.5, CART, CHAID, MARS [Wikipedia]. These algorithms have their own qualities and benefits.



Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No

Figure 1.5: decision tree derived from decision table [45]

Entropy and information gain are two main formulas used in decision tree. The entropy is the backbone of the decision tree ID3 and C4.5.

1.4.1 Entropy:

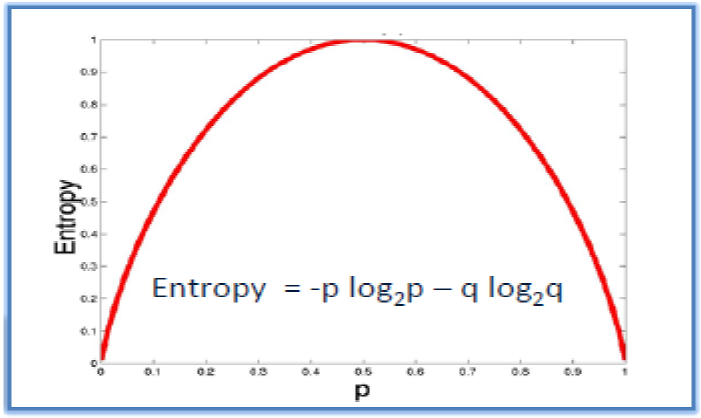
It is the known as sum of the probability of each class label times the log probability of that respective label [43]. In decision tree we need to calculate entropy two times. Firstly we have to calculate it for individual attribute, and then calculate it for the combined attributes. A simple formula to calculate Shannon's entropy is given below-

$E(S) = \sum_{i=1}^{c} -p\,i \log_2 pi$.................................**Equation 1.1**

To calculate the gain ratio the formula is-

Gain (T, X) = Entropy (T) – Entropy (T, X)........... **Equation 1.2**

There are three types of basic entropies we have in data mining. Those are Shannon's entropy, Renyi's entropy and Tsallis entropy. There are such researches which include the combination of these entropies to increase the accuracy of the traditional ID3 or C4.5 algorithm. In physics they also describes in some research paper that there are more combinations of the entropies which we can use for the further enhancements.



Entropy = -0.5 log$_2$0.5 – 0.5 log$_2$0.5 = 1

Figure 1.6: Entropy [42]

CHAPTER 2
REVIEW OF LITERATURE

Sneha Chandra et al. [1] present an adaptive classifier to enhance the classification accuracy in the field of medical data mining. They had been collected data from Nalanda Medical College Hospital (NMCH), Patna, Bihar, India. Data was collected regarding non-infectious disease Diabetes and the infectious disease Tuberculosis. In this research work she has been applied Laplacian Correction to Decision tree Classifier algorithm and Naïve Bayesian Classification algorithm. 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 RBC (Rule-Based Classifier), DTC (decision-Tree Classifier) and NBC (Naïve –Bayesian Classifier). They have been achieved 90% accuracy using adaptive algorithm.

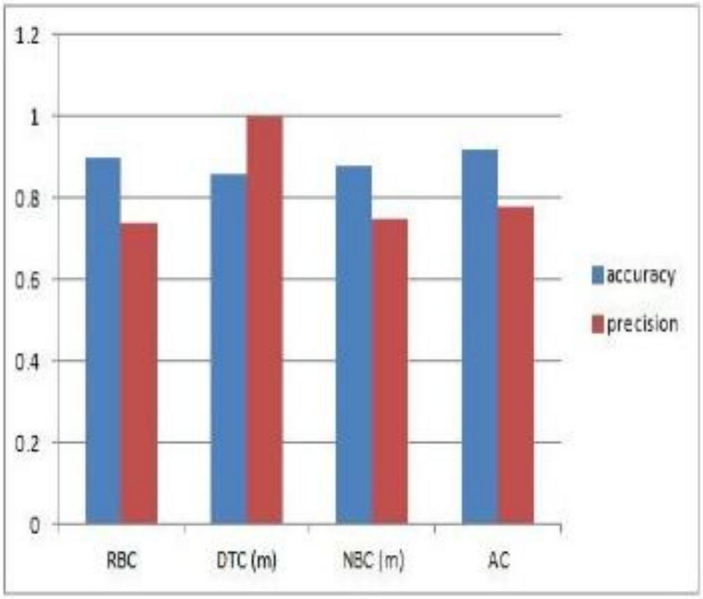


Figure 2.1: Accuracy and Precision Analysis [1]

Neelam Singhal et al. [2] worked on the performance enhancement of classification using hybrid algorithm in data mining. In this paper they merge Genetic Programming (GP) along with Decision tree classifier (DTC) to improve the accuracy, with the comprehensibility and respective timing of the classification. The algorithm has combination of Clustering and Feature Selection. For extracting data set the used software is weka. After that they have used 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 and Genetic programming individually.

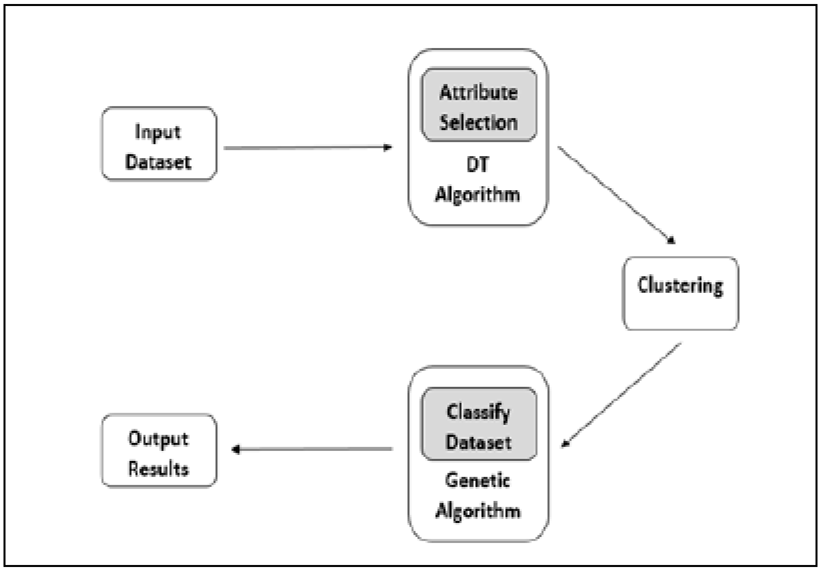
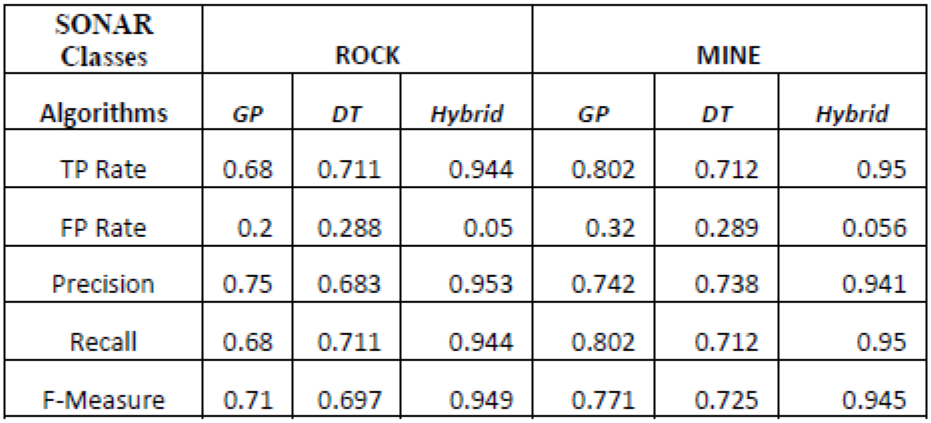


Figure 2.2: Pictorial Model for the Proposed Technique [2]



SONAR Classes	ROCK			MINE		
Algorithms	*GP*	*DT*	*Hybrid*	*GP*	*DT*	*Hybrid*
TP Rate	0.68	0.711	0.944	0.802	0.712	0.95
FP Rate	0.2	0.288	0.05	0.32	0.289	0.056
Precision	0.75	0.683	0.953	0.742	0.738	0.941
Recall	0.68	0.711	0.944	0.802	0.712	0.95
F-Measure	0.71	0.697	0.949	0.771	0.725	0.945

Figure 2.3: Results [2]

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 -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.

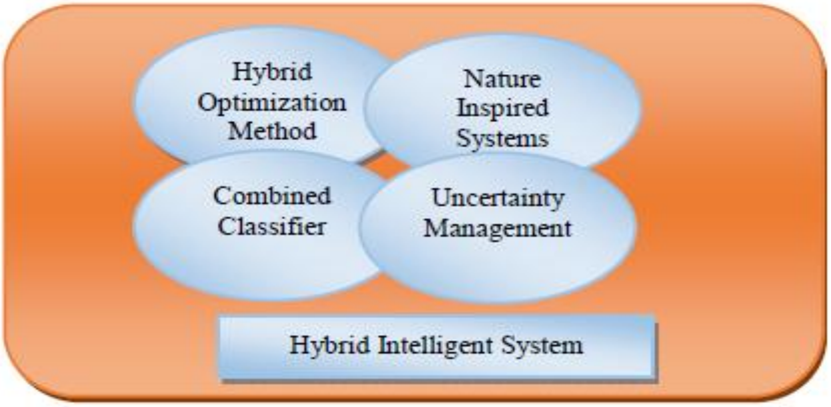


Figure 2.4: Domains of Hybrid Intelligent System [3]

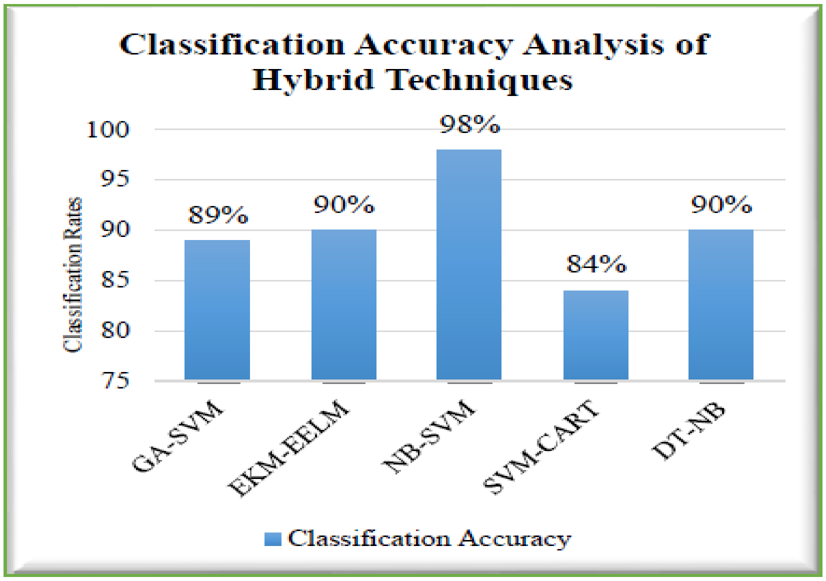


Figure 2.5: Results of Hybrid Approach [3]

This performance has been implemented on different image data sets. As the conclusion we get Naïve Bayes and Support Vector Machine hybrid approach having better accuracy in image classification that is 98%.

Mangesh M.Panchwagh et al. [4] present a music genre classification using data mining algorithm. 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. In this paper they have extracted three features Mel frequency Cepstral Coefficients (MFCC), Linear Productive coefficient and ZCR of different types of songs (pop, jazz, rock etc.). Dataset has been provided by Technical University Dortmund. The result shows that SMO classifier with PKIDiscretize pre-processing method provides accuracy of 100% for MFCC. Among the three MFCC provides the highest accuracy results.

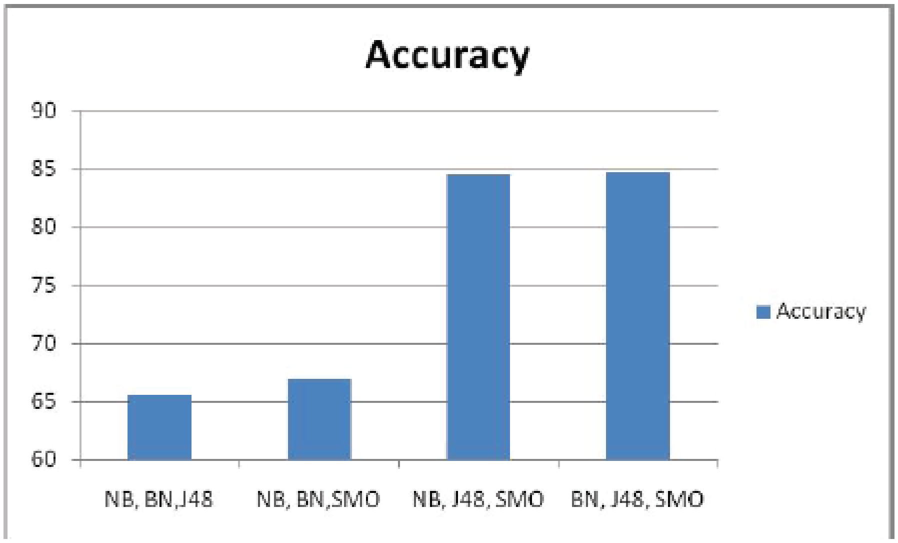


Figure 2.6: Accuracy Results of MFCC with Pkid Pre – Processing [4]

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. Data has been collected by NP and AV net protector. Then after performing various tests for virus using different parameters a data set has been 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 new 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 partition the data samples into pre-defined clusters. Then format the decision tree for each cluster. This algorithm has been implemented and 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.

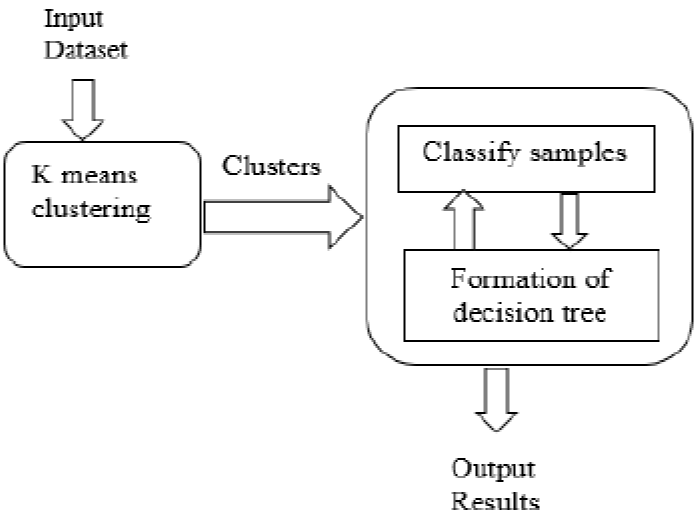


Figure 2.7: Diagram for Proposed Algorithm [6]

Rana Alaa El-Den Ahmed et al. [7] present a performance study of classification algorithms for online shopping attributes and behavior using data mining. 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. Data set has been collected from highly repudiated online shopping agency. Ten-fold cross validation is used for testing 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 data mining 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 the concluded that C4.5 is more accurate and consumes less execution time to mine data. They also provided the error rate difference that shown that C4.5 is best approach as compared to ID3.

Supreet kaur et al. [9] published a review paper on data mining classification techniques for detection 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 detection 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.

S.Vinodh Kumar et al. [10] described in their paper predicting fault – prone software modules using feature modules through mining algorithms that they analyzed the performance of different classification algorithms of data mining on defected software. This paper mainly focuses on the performance of the classifier algorithms on seven different datasets (PC1, PC2, PC3, CM1, Mw1 JM1, KC3, and PC4) mainly they have been categorized these datasets into two classes named as defective class and normal class. They have been performed 20 supervised algorithms on seven publically alreardy available datasets. They have taken data set from NASA MDP repository. The results show that Random Tree Classification algorithm provides 100% accuracy in classifying the given datasets.   Again here is important part they have been considered that is feature selection and they had verified their dataset according to the concluded results.

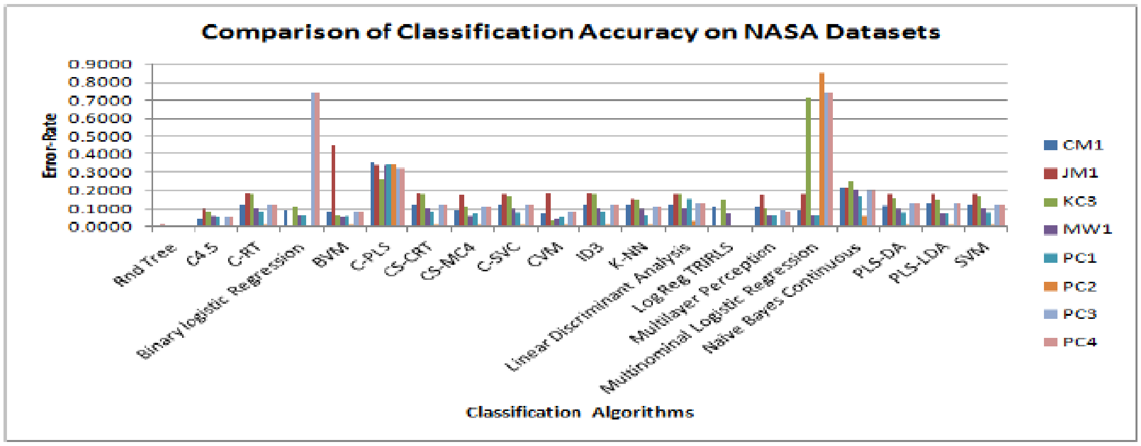


Figure 2.8: Comparison of the Algorithms on NASA Datasets [10]

Kewen Li , Peng Xie et al.[11] they have published a research paper on an improved algorithm for imbalanced data based on weighted KNN. In this paper they have been mentioned that the basic adaboost algorithm cannot provide a proper accuracy for balanced data for the minority class, they have been introduced a new k-adaboost algorithm for more accurate results. They have used here KNN for dividing the majority class into minority class and they consider more for the minority class. For avoiding the weight distortion they have used a threshold due to the classifying procedure. It also includes new error function to help us for neglecting the distortion. In standard classification methods have a high accuracy for the majority class and low accuracy for the minority class. That’s why we need to develop a new approach here. As the results they have been shown that the k-adaboot algorithm provides more accuracy as compared to the standard adaboost algorithm for the imbalanced data.

Jafar Tanha , Hamideh Afsarmanesh et al.[12] published a research paper on semi-supervised self training for decision tree classifiers. In this paper they have been described that the standard decision tree is not effective in self- training algorithm to semi-supervised learning. They also mentioned the reason for it that is the traditional decision tree cannot produce reliable probability estimation for specific given prediction. They have done various changes in the traditional decision tree for producing productive results. the various refinements done by them are naïve bayes tree combination of no-pruning and laplace correction grafting and distance- based measure. Further they also used ensembles approach for more accurate results. They have shown that the ensembles tree provides more accurate results as compares to the adapted decision tree for the learning process. They have been taken data sets from the UCI repository. They have compared all the previous approaches with the adopted decision tree self-training learner. According to the presented results it is clearly shows that the decision tree learner and the ensemble algorithm provide more accurate results. For the future work they have been suggested that to extend this work for the multiclass. In present state they worked on the binary class.

Rakesh Katuwal P.N. Suganthan Le Zang et al. [13] have published a new research paper an ensemble of decision trees with random vector functional link networks for multi-class classification. They have been proposed a new ensemble decision tree classifier combining it with RVFL. For partitioning the each individual classifier training set they have use RVFL algorithm further classification have done by using the decision tree. Developed algorithm has been tested on the 65 multi- class UCI datasets. They have shown that the proposed method providing more efficient results. They have compared RVFL with the RFL and RF. They named their algorithm as RFL which means RVFL with univariate trees and obRFL which means RVFL with oblique trees. As conclusion they described that RFL is better than obRFL, RFL and obRFL provides better accuracy than RF, obRF and RVFL.

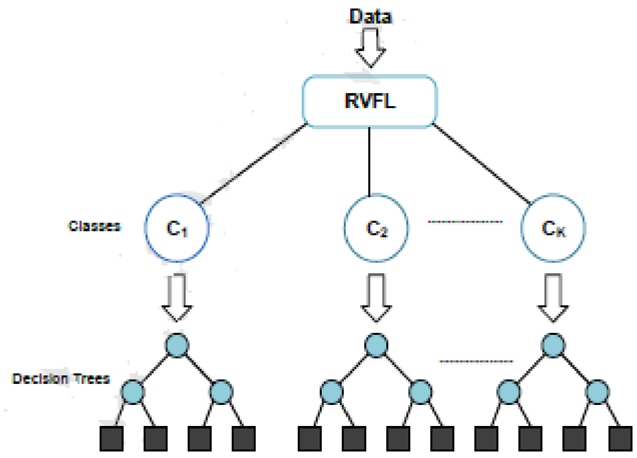


Figure 2.9: Proposed Ensemble Decision Tree with RVFL [13]

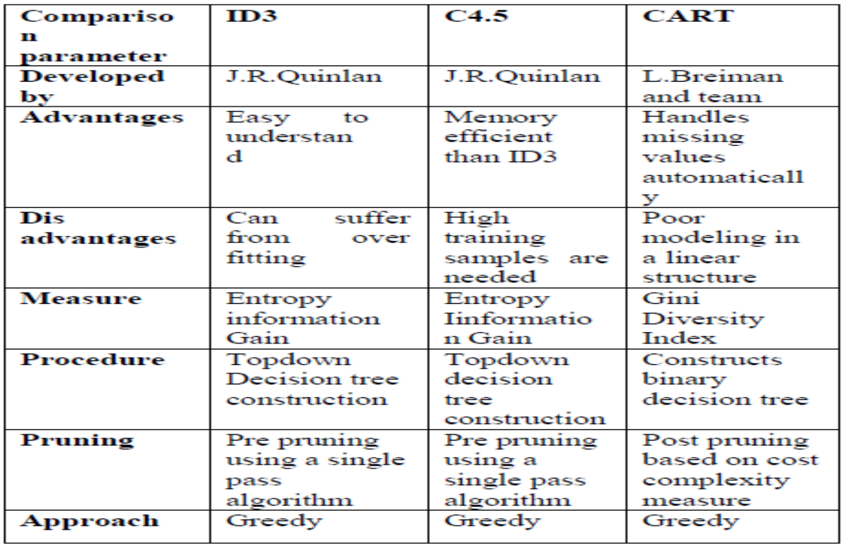
L. Truffet et al. [14] has been published a paper on Shannon entropy reinterpreted. This paper described the Shannon entropy and its limitation. To overcome the limitation of Shannon entropy they also proposed a new definition of the Shannon's entropy. They have been provided a systematic method to define a deformed algebra and deformed calculus provided. Multiplication is also called as frank copula. This paper is having domain in physics. These concepts are interrelated with each other. Fuzzy logic and calculus are also uses the same concept for the multiple working.

Krzysztof Gajowniczek et al. [15] have been published research paper on comparison of decision trees with renyi and tsallis entropy applied for imbalanced churn dataset. In this paper they have picked a concept of entropy from the decision tree C4.5. For the training set the performance of the both entropies can be balanced or imbalanced. But at the time of testing they showed that the renyi's entropy performs better. For overall the combination of the renyi's entropy and tsallis entropy performs the best. This combined approach is better than traditional Shannon's entropy.

Mehmet Niyazi Cankaya et al. [16] have been published research in statistical properties of Jizba- Arimitsu Hybrid Entropy. Here they have introduced the combined features of the renyi and tsllis entropy as well. Two persons developed this method described in this paper. Those are Jizb and Arimitsu. That's what this hybrid approach also known as Jizba- Arimitsu entropy. Here they have been discussed the relation between these two entropies. And on the behalf of their matching features they have proposed this new hybrid approach.

Snehal Bhogan et al. [17] has been published a paper on predicting student performance based on clustering and classification. Here they have compared different classification, clustering and regression approaches to predict the student performance based on such attributes. Using this prediction we can provide the necessary measurements to the student for increase his/ her performance in their exams. They have been proposed a hybrid method of enhanced k- strange points clustering algorithm and naïve bayes classification algorithm and they have also compared them with the existing k- means algorithm and decision tree algorithm. In last for predicting the student performance they have been used multiple-linear regression. Dataset has been collected from the collage Agnel Institute of technology and design. They have used 2012-2016 batch student data as the training set and used on k-means clustering algorithm and enhanced k-strange point clustering algorithm. Further for the testing purpose they have used data of the batch 2013-2017 for which they have applied decision tree, naïve bayes classification and multiple linear regressions. Results shows that the proposed algorithm having better accuracy as compared to the traditional algorithm.

S. Nagaparameshwara chary et al. [18] proposed a research paper a survey on comparative analysis of decision tree algorithms in data mining. In this paper they have described that the most used and appropriate method used for the analysis in DM is the decision tree. This follows two phases for the implementation. In the first phase we have to build a decision tree and in second phase pruning has been done. Here they have showed the comparison among the various decision trees for the analysis i.e. ID3, C4.5, J48 and CART.



Comparison parameter	ID3	C4.5	CART
Developed by	J.R.Quinlan	J.R.Quinlan	L.Breiman and team
Advantages	Easy to understand	Memory efficient than ID3	Handles missing values automatically
Dis advantages	Can suffer from over fitting	High training samples are needed	Poor modeling in a linear structure
Measure	Entropy information Gain	Entropy Iinformation Gain	Gini Diversity Index
Procedure	Topdown Decision tree construction	Topdown decision tree construction	Constructs binary decision tree
Pruning	Pre pruning using a single pass algorithm	Pre pruning using a single pass algorithm	Post pruning based on cost complexity measure
Approach	Greedy	Greedy	Greedy

Figure 2.10: table represents the various analysis comparisons [18]

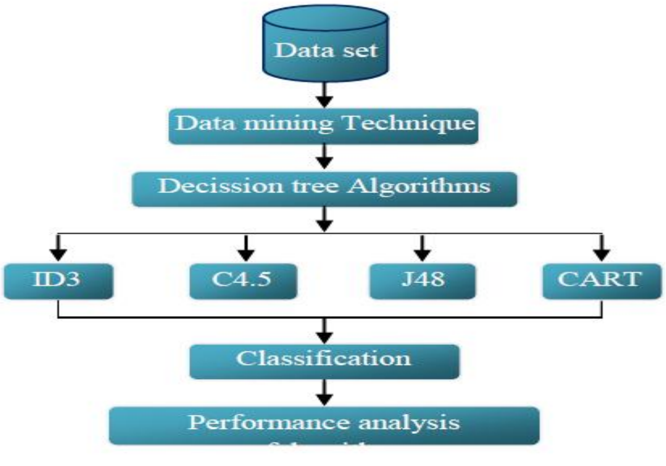
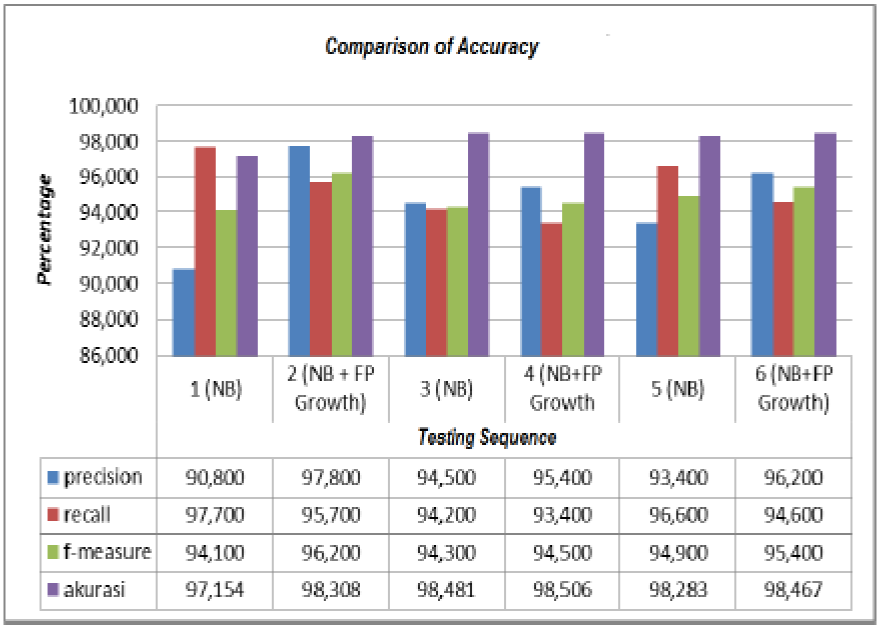


Figure 2.11: Proposed Methodology [18]

Dea Delva Arifin et al. [19] presents a paper on enhancing spam detection on mobile phone short message service(SMS) performance using FP-Growth and naïve bayes classifier. As we know that the use of SMS s very adoptable for advertisements and all for this reason the cases of SMS spam and fraud are also very common now a days. To detect these types of SMS frauds in this paper they have proposed a new combined approach. This includes the association and classification methods of the data mining. For the association purpose they have been used the FP growth and for the clustering they approach to the naïve bayes classifier, through this we can define a class of the fraud whether it is spam or ham. This enhanced approach gives the accuracy of 98%, 506% and 0,025%.



	1 (NB)	2 (NB + FP Growth)	3 (NB)	4 (NB+FP Growth	5 (NB)	6 (NB+FP Growth)
🟦 precision	90,800	97,800	94,500	95,400	93,400	96,200
🟥 recall	97,700	95,700	94,200	93,400	96,600	94,600
🟩 f-measure	94,100	96,200	94,300	94,500	94,900	95,400
🟪 akurasi	97,154	98,308	98,481	98,506	98,283	98,467

Figure 2.12: Accuracy Comparison [19]

Honjun Lu et al. [20] has been published a paper- neuro rule: a connectionist approach to data mining. In this paper they proposed a neuro rule connectionist approach for the data mining classification algorithm. Mainly they elaborate this approach in 3 phases. In the first step we have to train the neural network according to the given data set and with its required accuracy. In the second step pruning has been done as well as accuracy maintains. And in the last step we have to take the rules which are produced after pruning. They have mentioned that there are no such rules which are developed for the data mining approach and during the implementation when they have applied this approach the challenge was to reduce the execution time and total time to train our neural network. Further they have mentioned that in the future we can reduce the total execution time for this particular algorithm.

M Balamurugan et al. [21] published a research paper on the performance analysis of CART and C5.0 using sampling techniques. In this paper they have used the sampling approach on the balanced as well as unbalanced datasets. It's clearly mentioned that using the stratified sampling we can improve the decision tree accuracy. As we know that CART uses the gini index and C5.0 uses the information gain to find out the final results, but using the sampling we can increase the performance of these two algorithms but only for the unbalanced data. They have also mentioned that there is no such method to describe the particular sampling approach which increase the efficiency of the given decision tree. In future we can use this method for improving accuracy on the balanced data or we can design a better approach which defines a particular sampling approach suitable to the given data and the algorithm as well.

Tanupriya Choudhury et al. [22] present a paper on intelligent classification of lung and oral cancer through diverse data mining algorithms. In this paper they have used a simple logistic classifier in an intelligent way for analysis whether a patient have lung cancer or not. They have taken a dataset from a professional doctor and also from the internet repository. In this paper they are doing the image classification.

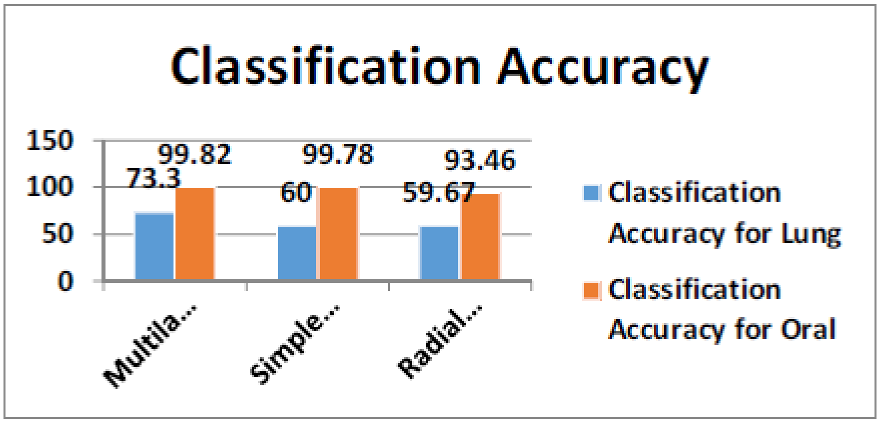


Figure 2.13: Accuracy Results [22]

If there is only single layer than it comes under preceptor and if there are multiple layers then it will be known as Multilayer Perceptron (MLP). There are three types of neural network classifier those are RBFN, MLP and SLA. The MLP is providing more accuracy as compare to the other two. Further we can implement this approach to real time data set and can also use this method for the bigger data.

Sachin Bhaskar et al. [23] have published a research paper on managing data in SVM supervised algorithm for data mining technology. This paper introduces a SVM classifier for the students just to understand this algorithm and various analyses like image classification, hand-writing recognition. This paper helps for the data mining students to understand the SVM algorithm and various functions of the kernels comes under the SVM. There are linear and non-linear kernels in the SVM algorithm. Here they have also described about the previous works done using the SVM algorithm. They also mentioned that there are same analysis areas can be work under the SVM. As I discussed earlier those are image reorganization, biosequence analysis and hand- writing reorganization.

K.Deeba et al. [24] have been published a paper on classification algorithms of data mining. In this paper they have been described the various classification algorithms. Mentioned algorithms are decision tree, Bayesian network model, naïve bayes, support vector machine and k nearest neighbor. All the advantages and disadvantages for the respective algorithms are also discussed here.

Sagar S. Nikam et al. [25] have been conductive a research on a comparative study of the techniques in data mining algorithms. C4.5, ID3, K-nearest neighbor, naïve bayes, SVM and ANN. Every algorithm has its own limitations and features. According to the desired features and the available data we can implement the suitable algorithm. These algorithms used for the pattern reorganization, spam filtering, artificial intelligence and also for the large volume of the dataset. This is totally depends on the situation that which classifier will be the best.

Diego Buenano Fernandez et al. [26] has been presents a paper on the comparison of the applications for educational data mining in engineering education. This paper discussed the comparison among the open source tools i.e. rapid miner, knime, weka used in educational data mining field. They have picked data set from the Ecuadorian university which includes the records of the tree years engineering programs. Main concentration they enforces on evaluating the best tool for analysis the student performance. They have been described that the number of newly available algorithms are mostly includes in weka as compared to two others. But the GUI is less friendly in weka rather than knime and rapid miner have. According to the produced results they have shown that it is very much important to know that which student needs training on which particular subject at what time, and to provide guidance accordingly.

Bharat S. Makhija et al. [27] has been published a paper on classification algorithms of data mining. This paper includes detail about the various classification algorithms available for data mining. The importance of these algorithms is also being described. Author mentioned that there is not a single classification algorithm by which we can conclude and say this is the best algorithm. This is totally depends on data set and various other factors as well. To use the algorithms in the best manner we can also go for ensemble approach through which we can combines two or more algorithm at the same time.

Haohang Li et al. [28] have been presents a paper on research on the high robustness data classification and the mining algorithm based on hierarchical clustering and KNN. The author has proposed a new approach named as prolonged supervised clustering and classification algorithm (S2CA). This new developed approach is used to enhance the capability of the existing algorithm. This is an adaptive approach which can be applicable on the data mining works as well.

K Prasanna Jyoti et al. [29] have been published research paper on a study of classification techniques of data mining techniques in health related research. We are having different classification algorithms for predicting the patient data sets. Some of them are C4.5, ID3, k-nearest neighbor, SVM, naïve bayes and ANN. They have been mentioned that how much it is important to classify patient data according to a particular disease. And all the classification methods are very useful to classify that data. The data has been collected from the health center. Various health related studies has been discussed here.

Saurabh Pal et al. [30] have been published research paper on performance analysis of students consuming alcohol using data mining techniques. This paper includes four dta mining algorithms – sequential minimal optimization (SMO), bagging, REP tree and decision table. Data has been collected from VBS Purvanchal University, Janupur, India and they had implemented it on the real based data set. They have been compared the discussed four algorithms. They have mentioned that bagging classification is providing accuracy 80.2532% and total time for execution taken is 0.14 seconds, which is the best among the four algorithms.

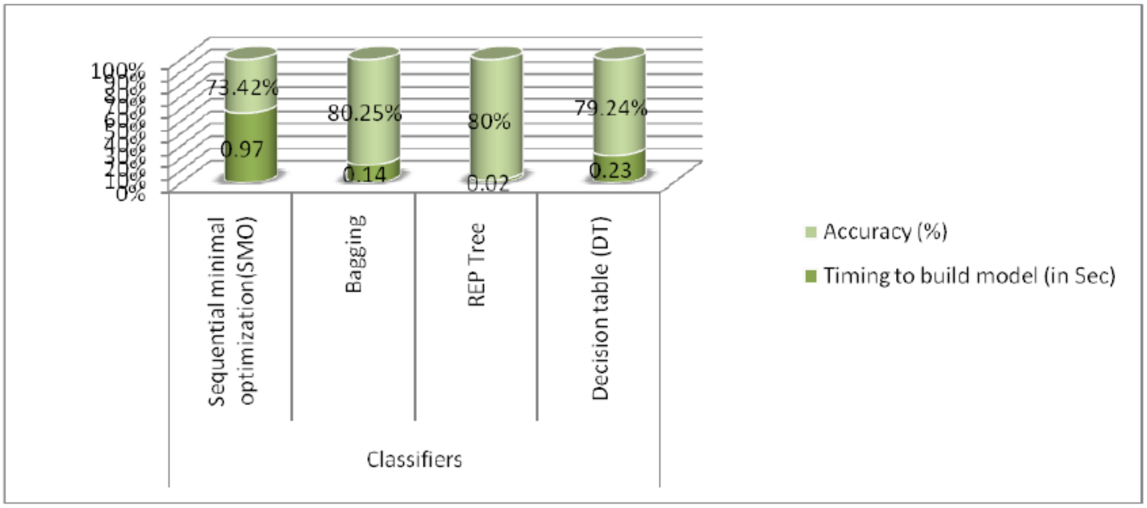


Figure 2.14: Accuracy Results [30]

Rui Han et al. [31] has been published a research paper on elastic algorithms for guaranteeing quality monotonicity in big data mining. According to the published paper author described that whenever we need to mine data from the available big data there are two things an algorithm has to full fill. First one is to provide the best accuracy according to the given data. Second thing is to provide a solution in minimum time constrains and less resources. They have used the concept of Shannon's entropy to increase the accuracy of the proposed algorithm and increase the quality monotonicity in a minimum time with limited given resources. They have used the KNN and CF (Collaborative Filtering) with R-tree data structure, this enhance the quality of the proposed algorithm. In future work they have been mentioned that we can implement this approach by using other clustering methods. Proposed algorithm is using naïve KNN concept. As the results they have shown that the elastic algorithm provides the better accuracy for the mining process in big data within limited resources and efficient time constrains.

Muhammad Yousefnezhad et al. [32] have being published paper on weighted spectral cluster ensemble. Cluster ensemble selection (CES) is a new approach through which we can combine the result of the different clustering algorithms. Ensemble approach is provide better results as compare to any individual algorithm. Here they have proposed a new WSCE approach as an ensemble method. There are two problems occurs whenever a data set have to deal with clustering algorithm. The first one is these algorithms are mostly depends on Shannon's based, and send one is to estimating the threshold values. To overcome these two problems the authors have proposed a new ensemble based clustering algorithm without threshold values. For implementing this proposed algorithm they have used the concept of graph based clustering algorithm and community detection arena. To generate the graph based results they have used two kernel spectral clustering method. The consequences describe that the proposed algorithm having highest accuracy rate.

Gang Sun et al. [33] have published a paper on a coal mine safety evaluation method based on concept drifting data stream classification. Monitoring data in coal mining uses the concept of drifting data stream classification.

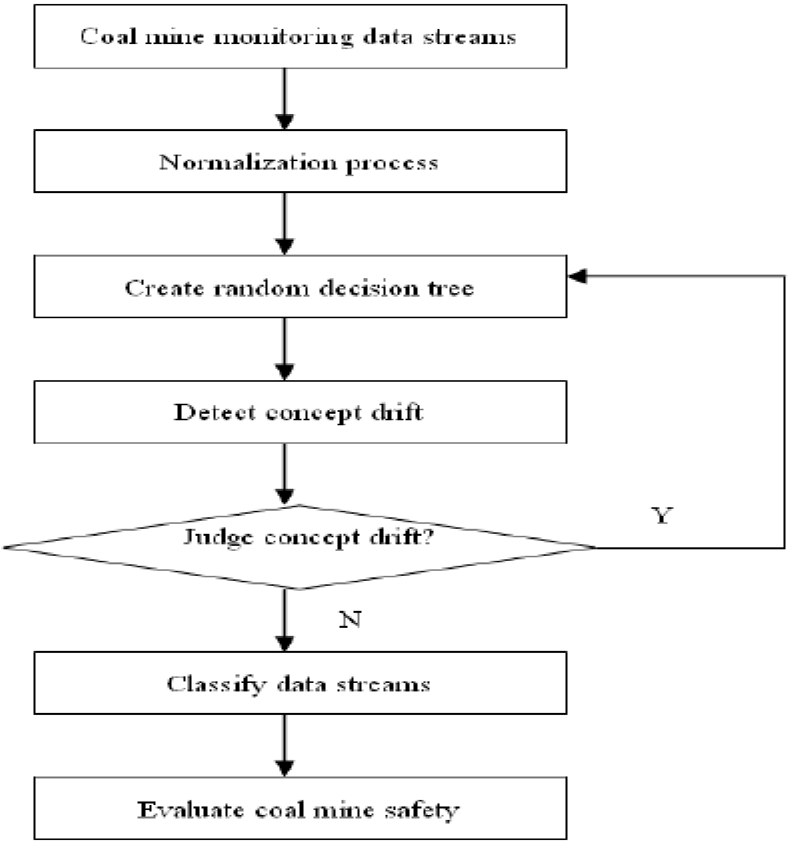

Coal mine monitoring data streams
Normalization process
Create random decision tree
Detect concept drift
Judge concept drift?
Y
N
Classify data streams
Evaluate coal mine safety

Figure 2.15: Proposed Algorithm's Work Flow [33]

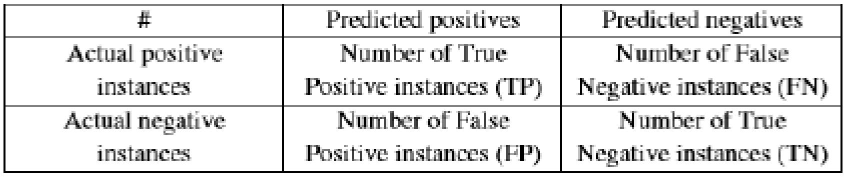
It is very important to ensure the safety of this task. To ensure these authors have proposed a random decision tree model, this uses the concept of Hoeffding Bounds inequality and information entropy, which normally not be included in this process. For standard execution to calculate the split data point normally available approach is random selection. For determining the threshold values they have been used Hoeffding Bounds inequality, to detect the drifting concept. The results determine clearly that to detect the concept drift the proposed method is a good approach. It also evaluated the data in more accuracy as compare to the standard approach. This is a new research in case of classifying data streams under the coal safety evaluation.

Petrônio Lopes et al. [34] have published a paper on GPU-oriented stream data mining traffic classification. In a networking word it is very hard to keeping records of each and every traffic bit. No doubt, there are AI (machine learning) techniques and methodology to list the network traffic. But still we need an algorithm which supports this type of architecture, and provides accurate results with good speed as well. This paper has proposed architecture to do stream data classification with in good speed along with accuracy. GPU (graphical processing unit) combines with data mining streaming characteristics to enhance the speed and accuracy of the architecture. Consequently, the proposed approach gives accuracy up to 95% along with speed 62x. After performing this experiment they have compared this approach with the previous implementations. But this method shows the best results in cases, accuracy and speed. This implementation has been done on real time data set.

Ranganatha S. et al. [35] have published the paper on medical data mining and analysis for heart disease dataset using classification techniques. Data mining plays a vital role in medical field. This era is full of new diseases, to keep record of each and every medicine record and to use this information for the research purpose or for patient’s treatment, it is very difficult task. To make this task easy this paper introduces an algorithm which gets the new hospitalized patient’s records and generates report in simple understandable language. This work has been done for heart disease patient. They have combined the ID3 decision tree and naïve Bayesian algorithm. ID3 decision tree generates report in a tree form which is easy to understand. Naïve Bayesian algorithm forecast the chances of a heart disease on the bases of providing details.

Ondrej Kazik et al. [36] have presented the paper on clustering based classification in data mining method recommendation. The enlargement of data base collection is a major issue for present scenario; the problem is to classification of data for the analysis purpose. This paper introduces the Meta learning based algorithm. By using the Meta data concept we can also determine a particular class or cluster for the unpredictable data set. Here they have done analysis on classification as well as clustering based algorithms. They have presented comparison among these algorithms. Using the concept of Meta learning they have recommended a mining approach for doing classification or clustering. In future they have mentioned the concept of association is not used in this paper, if anybody wants to pursue their work on the bases of this mining recommendation they can use the concept of association as well.

Manoj Kumar Das et al. [37] have been published review paper on opinion mining and sentiment classification: a review. This is the era of web technology, people don not only use this web information for their works but also provides opinions for the collected web information. These opinions are negative or positive. Using these feedback information sentiment analysis can to be implements using data mining concept. Supervised learning algorithms like naïve bayes, maximum entropy and SVM includes in this study. SVM algorithm also provides accuracy for the given data. This paper concludes that sentiment analysis is very challenging thing to do.



#	Predicted positives	Predicted negatives
Actual positive instances	Number of True Positive instances (TP)	Number of False Negative instances (FN)
Actual negative instances	Number of False Positive instances (FP)	Number of True Negative instances (TN)

Figure 2.16: Sentiment Classification Evaluation [37]

Ms.A.M.Abirami et al. [38] have been published a research paper on a survey on sentiment analysis methods and approach. Before purchasing anything a customer goes for views at least once, these all views comes under sentiment analysis as the business prospective. This paper present a survey on sentiment analysis methods and their use approaches. Customers provide opinions for the collected web information. These opinions are negative or positive. Using this feedback information sentiment analysis can be done. This paper includes SVM, naïve bayes, maximum entropy and various machine learning algorithms. The authors have presented the accuracy for the respective algorithms as well. As we know sentiment analysis is not easy to do at all, there are some problems exist in binary classification, polarity ratio. As the shown comparison they have shown that the machine learning algorithms provides better accuracy as compare to mining algorithms.

Tarun B. Mirani et al. [39] have been published a paper on sentiment analysis of ISIS related tweets using absolute location. ISIS terrorist attacks got famous using the social media. Tweeter allows every person to post anything like text, videos, and audios as publically. ISIS uses hashtag for their communication. This paper presents a method for the sentiment analysis on ISIS using tweets. This approach also includes the geolocations information regarding various comment on ISIS. The jeffrey breen algorithm is used by the authors for the sentiment analysis. Polarity based classification in ISIS related tweets have been done using different data mining algorithms – SVM, random forest, bagging, decision trees and maximum entropy. They have been collected data of tree days related to these tweets. They have also mentioned the most frequent words used in these tweets. The average accuracy shown is 90%. Among all the algorithms the maximum entropy algorithm gives the best accuracy that is 99%, after 10 folds validation.

Yurong Zhong et al. [40] has been published a paper on the analysis of cases based on decision tree. In the concept of data mining dissrent steps are included, but classification is an important factyor that we have to include in process. Under classification as we know that there are several algorithms, but according to this paper decision tree is very usefull approch because of some features like easy to understand and provides good accuracy results. They have presented different cases on the basis of ID3 decision tree. The authors have been proposed an improved version of the decision tree classifier on the bases of tylor method and entropy. The entropy is works as backbone of the decision tree. So, they have picked this entropy concept and enhace the performance of the ID3 decision tree. They have also reduced the complexity of the same algorithm as well. Consiquently, by using the concept of entropy modification we can say ID3 decision tree contains better accuaracy and less complex approch as well.

CHAPTER 3

SCOPE OF THE STUDY

As per proposed study improve the efficiency of classification algorithm in data mining, we can extend the accuracy of the standard algorithm decision tree ID3 and C4.5. Another option we have to adopt the hybrid approach for more accurate mining in less and effective time. In proposed study the concept of entropy is used. Entropy is a backbone of decision tree i.e. ID3 and C4.5. The present work includes the standard entropy formulas in decision tree for the classification purpose. In future there are still some combination comes under the domain of physics, from where we get an idea to use those concepts in data mining. For more experiments we can also do analysis according to the different data sets and using those analysis predict the best approach for the particular data set, because there are some authors who has mentioned this problem.

CHAPTER 4

OBJECTIVES OF THE STUDY

The objective of the proposed study is to increase the efficiency of a classification algorithm (decision tree) in data mining. To increase this accuracy the concept of entropy is involving. Information gain and entropy are two standard things, without which this is totally difficult to calculate the values of the decision tree. The standard entropy uses in the decision tree ID3 is Shannon's entropy. There are other entropies like Renyi's entropy and Tsallis entropy which are also very useful for the data mining prospects. There are some researchers who have published there research in comparing these entropies and also described that Renyi's entropy, Tsallis entropy and some other's entropies having more accurate consequences as compared to Shannon's entropy. But there are some more concepts of entropy which can be used in data mining, other than physics only. In short the proposed objective is just want to conclude that, using these entropy properties we will increase the accuracy of the decision tree in data mining.

CHAPTER 5

RESEARCH METHODOLOGY

Proposed study topic is “improve the efficiency of classification algorithm in data mining”. As we know that there are so many classification algorithms comes under data mining, proposed methodology is on decision tree. Decision tree is mostly used for the classification purpose in data mining. It creates a hierarchy on the bases of two things – entropy and information gain. Entropy is the sum of the probability of each label times the log probability of that same label [44].

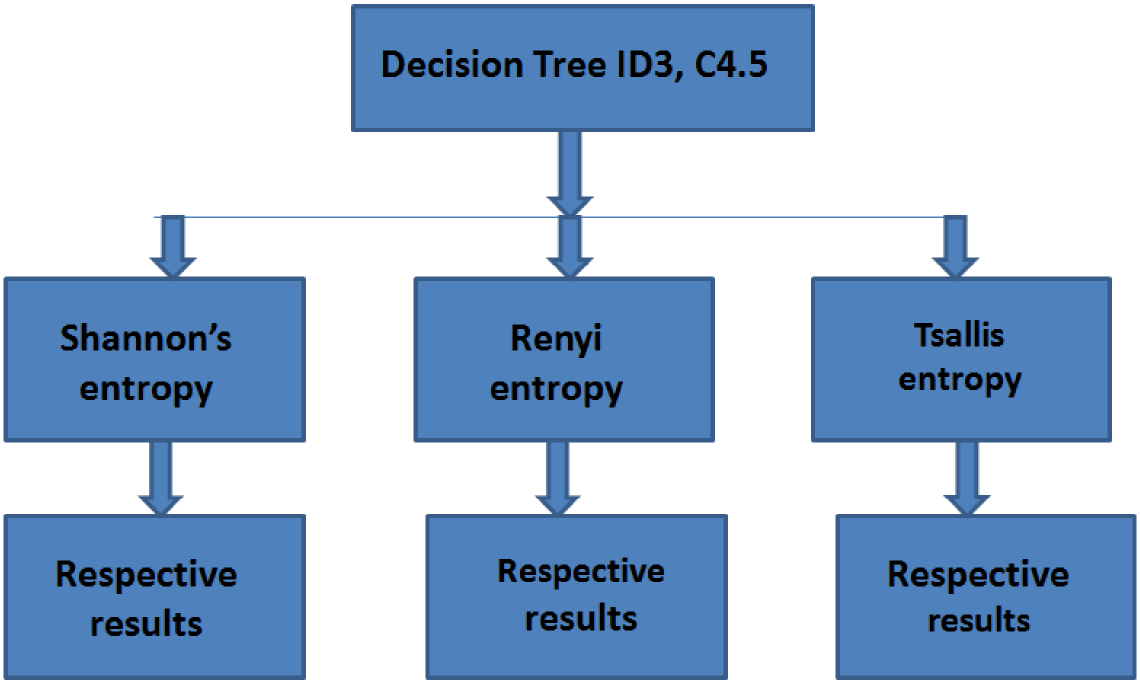


Figure 5.1: Previous Experiments on Decision Tree Using Different Entropy Concept

The last leaf provides us the appropriate class. Here it is clear that why entropy is so much important in decision tree ID3 or C4.5, because this is backbone of this algorithm. There are tsome types of basic entropies in the field decision tree [15]. Those are Shannon’s entropy, Renyi’s entropy and Tsallis’s entropy and some more entropies have been introduced. As you can see from the figure 23 this provides the result of implemented till the date on ID3 and C4.5 using the concept of entropy. In [15] author has also compared the results of the respective entropies. In [16] they have mentioned that there are some other entropy concepts like there asymptotic statics through which we can combine the two or more entropies in a single function and then we can calculate the consequences.

A simple formula to calculate Shannon's entropy is given below-

$$E(S) = \sum_{i=1}^{c} -p\,i \log_{2} \text{pi} \ldots\ldots\ldots\ldots\ldots\ldots \textbf{Equation 5.1}$$

To calculate the gain ratio the formula is-

$$\text{Gain (T, X)} = \text{Entropy (T)} - \text{Entropy (T, X)} \ldots \textbf{Equation 5.2}$$

These two are Shannon –Boltzmann-Gibbs entropy based formulas. But as we can see from the [16] there are other methods to use these entropies in the efficient manner.

Formula to calculate ranyi entropy-

$$\text{H}\alpha(\text{X}) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p_{i}^{\alpha}\right) \ldots\ldots\ldots\ldots\ldots\ldots \textbf{Equation 5.3}$$

Formula to calculate tsallis entropy-

$$\text{S}_{\text{q}}\ (\text{p}_{\text{i}}) = \frac{k}{q-1} . \left(1 - \sum_{i=1}^{.} p_{i}^{q}\right) \ldots\ldots\ldots\ldots\ldots\ldots \textbf{Equation 5.4}$$

On these bases we are proposing a modified entropy based algorithm through which we can enhance the accuracy of the decision tree. It can be implementing it on ID3 as well as on C4.5. In figure 24 proposed methodologies is showing that there are different factors on which bases we will increase the accuracy of the proposed algorithm. Those factors are AUC (area under ROC curve), confusion matrix, cross fold validation and pruning. After that proposed study will compare the modified entropy based algorithm with the previous decision tree which also had applied different entropy methods. For all this implementation we will use Weka tool for the analysis and comparison purpose. Weka is free and open source software which is easily available. For algorithm modification we will use net beans software, which is again a free software this supports java language. We are just using the concept of entropy as mentioned in [16], this is concept of physics. It's very clear that data mining is an interdisciplinary approach. This modified entropy concept is never used by any research scholar in data mining till the date. This is only defined in physics under the thermodynamics concept. But according to the described concept in [16] we can say that it is totally applicable for the data mining domain as well. To implement this proposed algorithm we will get data set from the UCI repository. This is university of California repository, a free repository which contains number of authorized and verified data sets. So, for the training as well as for testing the data set will be get from the UCI repository.

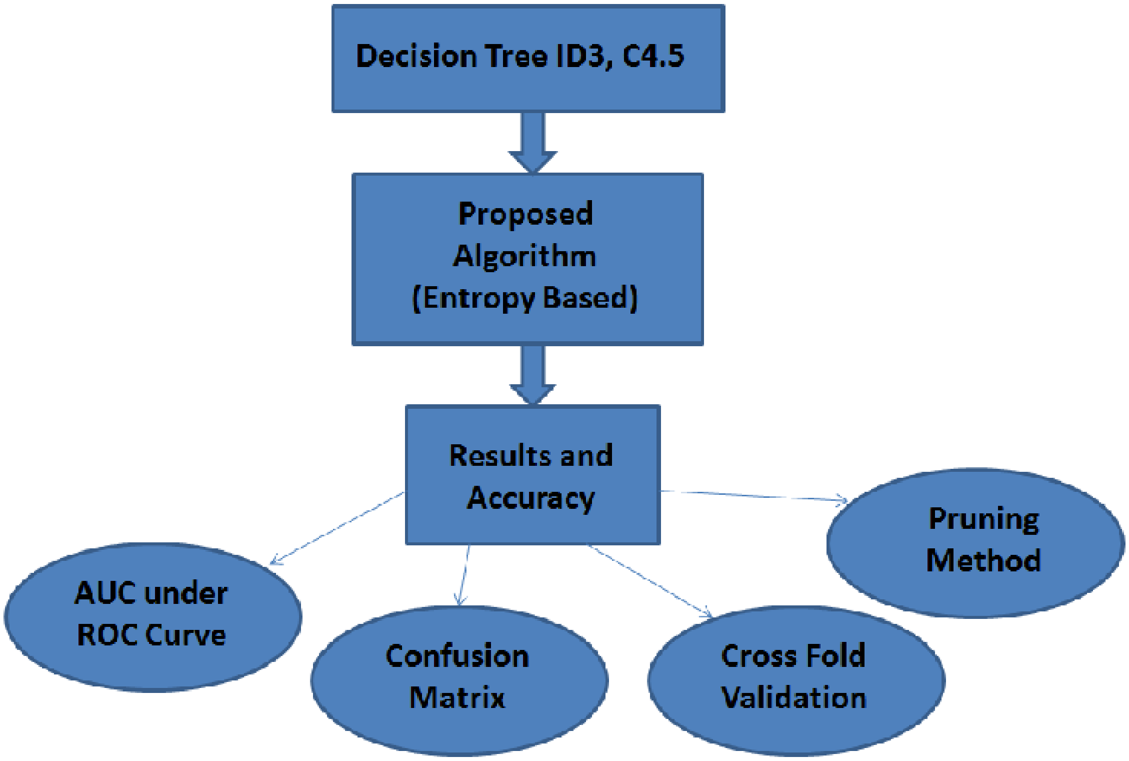


Figure 5.2: Methodology for the Proposed Modified Algorithm

CHAPTER 6

EXPECTED OUTCOMES

The expected results would be enhanced version of the decision tree algorithm ID3 and C4.5.As you can see from the research methodology the consequences will contain more accurate result from the previous algorithm. Till the time the concept of the hybrid entropy, and on the bases of similar features the combined entropy has been introduced only in the physics domain. If some scholars has introduces this site then they have only mentioned the comparison among the different entropies and the respective results only. [15] As the author introduces the some similar features through which we can combines Renyi’s entropy and Tsallis’s entropy and other entropies to increase the efficiency of the new approach using theses combinations. So, we will work on this entropy concept and provide the better performance in terms of the accuracy.

CHAPTER 7
SUMMARY AND CONCLUSION

Data mining is vital research area in the field of computers. For every business we need to predict the customer demands so that we can plan accordingly. To mine such data and to extract some meaningful knowledge we have KDD process. For this process we need some methods like association, classification, clustering and regression also. Proposed study chosen classification part in data mining. Classification means on the bases of similar properties of the different objects we have to define a class label. To narrow down this topic, studied about various classification algorithms and finally picks a decision tree algorithm i.e. ID3 and C4.5. In decision tree there is concept of entropy and information gain. To calculating entropy we are having basic entropy formula called Shannon-Boltzmann-Gibbs entropy. But it need to be redefine because some of its limitations. To overcome these limitations later another entropies has been introduced those are Renyi's entropy and Tsallis's entropy. Some scholars have done some work on comparing these entropies. Here is concept to combining entropies on the basis of the similar properties. So we bring this idea [16] from a physics domain paper. And will work on the entropy. As per purposed topic mentioned in the research methodology using the concept of entropy we will enhance the ID3, C4.5 algorithms accuracy.

LIST OF REFERENCES

[1] Chandra Sneha M K 2015 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* (IEEE)

[2] Singhal N and Ashraf M 2015 Performance enhancement of classification scheme in data mining using hybrid algorithm *International Conference on Computing, Communication & Automation* (IEEE) pp 138–41

[3] Thamilselvan P and Sathiaseelan J G R 2015 Image Classification using Hybrid Data Mining Algorithms – A Review *IEEE International Conference on Innovations in Information Embedded and Communication Systems* (IEEE) pp 71–6

[4] Panchwagh M M 2016 Music Genre Classification Using Data Mining Algorithm *conference on advances in signal processing(CASP) Cumminm College of engineering for Women* (Pun: IEEE) pp 49–53

[5] Jain J and Pal P R 2017 Detecting Worms Based on Data Mining Classification Technique *Int. J. Eng. Sci. Comput.* **7** 11388–91

[6] Ahlawat A and Suri B 2016 Improving classification in data mining using hybrid algorithm *2016 1st India International Conference on Information Processing (IICIP)* (IEEE) pp 1–4

[7] Ahmeda R A E-D, Shehaba M E, Morsya S and Mekawiea N 2015 Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining *Proceedings - 2015 5th International Conference on Communication Systems and Network Technologies, CSNT 2015* (IEEE) pp 1344–9

[8] Kaur S and Kaur H 2017 Review of Decision Tree Data mining Algorithms: CART and C4.5 *Proceeding of International Conference on Information Technology and Computer Science* vol 8(IEEE)pp 4–8

[9] Dogra A K 2015 A Review Paper on Data Mining Techniques and Algorithms *Int. Res. J. Eng. Technol.* **4** 1976–9

[10] Kumar S V and Jacob S G 2012 Predicting Fault –Prone Software Modules Using Feature Selection and Classification through Data Mining Algorithms 1 *2012 IEEE International conference on computational Intelligence and Computing Research* (IEEE) pp 1–4

[11] Li K, Peng X, Jiannan Z and Wenying L 2017 An Improved Adaboost Algorithm for Imbalanced Data Based on Weighted KNN *2017 IEEE 2nd Internation Conference on Big Data Anaysis* (IEEE) pp 30–4

[12] Tanha J, van Someren M and Afsarmanesh H 2017 Semi-supervised self-training for decision tree classifiers *Int. J. Mach. Learn. Cybern.* **8** 355–70

[13] Katuwal R, Suganthan P N and Zhang L 2017 An ensemble of decision trees with random vector functional link networks for multi-class classification *Appl. Soft Comput. J.*

[14] Truffet L 2017 Shannon Entropy Reinterpreted

[15] Gajowniczek K, Ząbkowski T and Orłowski A 2015 Comparison of Decision Trees with Rényi and Tsallis Entropy Applied for Imbalanced Churn Dataset *Comput. Sci. Inf. Syst.* 39–44

[16] C M N 2016 Physica A *2017 Elsevier* **0** 1–11

[17] Bhogan S, Sawant K, Naik P, Shaikh R, Diukar O and Dessai S 2017 PREDICTING STUDENT PERFORMANCE BASED ON CLUSTERING AND CLASSIFICATION *IOSR J. Comput. Eng.* **19** 49–52

[18] Nagaparameshwara S and Rama B 2017 A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining *International Conference on Innovation Applications in Engineering and Information Technology(ICIAEIT-2017)* pp 91–5

[19] Arifin D D and Bijaksana M A 2016 Enhancing Spam Detection on Mobile Phone

Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier *2016 IEEE Asis Pacific Conf. Wirel. Mobile(APWiMob)* 80–4

[20] Lu H 2017 NeuroRule : A Connectionist Approach to Data Mining *Proceedings of the 21st VLDB Conference Zurich* (Swizerland)

[21] Balamurugan M and Kannan S 2016 Performance Analysis of Cart and C5 . 0 using Sampling Techniques *2016 IEEE International Conference on Advances in Computer Application (ICACA)* (IEEE) pp 72–5

[22] Engineering T 2016 Intelligent Classification of Lung & Oral Cancer through diverse data mining algorithms *2016 International Conference on Micro-Electronics and Telecommunication Engineering* (IEEE) pp 135–40

[23] Road B 2014 Managing Data in SVM Supervised Algorithm for Data Mining Technology *IT in Business, Industry and Government (CSIBIG), 2014 Conference* (Indore, India: IEEE)

[24] Deeba K and Amutha B 2016 Classification Algorithms of Data Mining *Indian J. Sci. Technol.* **9**

[25] Nikam S S 2015 ORIENTAL JOURNAL OF A Comparative Study of Classification Techniques in Data Mining Algorithms *An Int. Open Free Access, Peer Rev. Res. J.* 13–9

[26] Fernández D B and Luján-mora S 2017 Comparison of applications for educational data mining in Engineering Education *IEEE* 0–4

[27] Makhija B S and Raut A B 2016 Classification Algorithms of Data Mining *Int. J. Sci. Eng. Technol. Res.* **5** 1639–42

[28] Li H and Wang S 2017 Research on the high robustness data classification and the mining algorithm based on hierarchical clustering and KNN *Communication and Electronics Systems (ICCES), International Conference* (Coimbatore, India: IEEE)

[29] Engineering C 2017 A Study of Classification Techniques of Data *Int. J. Innov. Res.*

Comput. Commun. Eng. 13779–86

[30] Pal S and Chaurasia V 2017 PERFORMANCE ANALYSIS OF STUDENTS CONSUMING ALCOHOL USING DATA MINING *3rd International Conference on "Latest Innovation in Science, Engineering and Management"* (The International Centre Goa, Panjim, Goa(India)) pp 24–36

[31] Han R, Nie L, Ghanem M M and Guo Y 2013 Elastic algorithms for guaranteeing quality monotonicity in big data mining *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013* 45–50

[32] Yousefnezhad M and Zhang D 2016 Weighted spectral cluster ensemble *Proc. - IEEE Int. Conf. Data Mining, ICDM* **2016**–**January** 549–58

[33] Sun G, Wang Z, Zhao J, Wang H, Zhou H and Sun K 2016 A coal mine safety evaluation method based on concept drifting data stream classification *2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.* 1125–9

[34] Lopes P, Fernandes S, Melo W and Sadok D H 2014 GPU-oriented stream data mining traffic classification *2014 IEEE Symp. Comput. Commun.* 1–7

[35] Anusha C, Vinay S K, Pooja Raj H J and Ranganatha S 2013 Medical data mining and analysis for heart disease dataset using classification techniques *Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013)* 1.09-1.09

[36] Kazik O, Peskova K, Smid J and Neruda R 2013 Clustering Based Classification in Data Mining Method Recommendation *2013 12th Int. Conf. Mach. Learn. Appl.* 356–61

[37] Das M K, Padhy B and Mishra B K 2017 Review *Inventive Systems and Control (ICISC), 2017 International Conference* (Coimbatore, India: IEEE) pp 4–6

[38] Abirami M A M and Gayathri M V 2016 a Survey on Sentiment Analysis Methods and Approach *2016 Eighth International Conference on Advanced Computing (ICoAC)* (Chennai, India: IEEE) pp 72–6

[39] Mirani T B and Sasi S 2016 Absolute location *Computational Science and Computational Intelligence (CSCI), 2016 International Conference* (Las Vegas, NV, USA: IEEE) pp 1140–5

[40] Yurong Zhong 2016 The analysis of cases based on decision tree *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (Beijing, China: IEEE) pp 142–7

[41]https://www.google.co.in/search?q=knowledge+discovery+in+database&rlz=1C1CHBF_enIN761IN761&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj6lY_J-sDXAhVJNY8KHaC0By4Q_AUICigB&biw=1366&bih=662#imgrc=Oeyjiu_FfEkI6M

[42] http://www.saedsayad.com/decision_tree.htm

[43] https://stackoverflow.com/questions/1859554/what-is-entropy-and-information-gain

[44] https://stackoverflow.com/questions/1859554/what-is-entropy-and-information-gain

[45] http://www.saedsayad.com/images/Entropy_Sunny.png

APPENDIX

DM – Data Mining

KDD- Knowledge Discovery in Databases

ID3 - Iterative Dichotomiser 3

CART- Classification And Regression Tree

CHAID- CHi-squared Automatic Interaction Detector

WEKA- Waikato Environment for Knowledge Analysis

UCI- University of California

SMO- Sequential Minimal Optimization

AI- Artificial Intelligence

SVM- Support Vector Machine

KNN- K-Nearest Neighbor

TP- True Positive

TN- True Negative

FP- False Negative

FN- False Positive

NMCH -Nalanda Medical College Hospital

RBC - Rule-Based Classifier

DTC- decision-Tree Classifier

NBC- Naïve –Bayesian Classifier

GP-Genetic programming

SVM-Support Vector Machine

MFCC- Mel frequency Cepstral Coefficients

LPC- Laplacian Productive Coefficient

ZCR- Zero Crossing rate

TU - Technical University

RVFL- Random Vector Functional Link

SMS- Short Message Service

MLP- Multilayer Perceptron

S2CA- supervised clustering and class