International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
419 – 422

A Survey on Improved Hybrid Classification Methods in Data Mining

Pulatsya R. Kanasagara
Computer Engineering Department
L.D. College Of Engineering
Ahmedabad , Gujarat, India
pkanasagara@gmail.com

Prof. Tushar J. Raval
Computer Engineering Department
L.D.College Of Engineering
Ahmedabad , Gujarat , India

Abstract— Data mining is powerful concept with great potential to predict future trends and behaviour. It refers to the extraction of hidden knowledge from large data set using various techniques. But as the amount of data generated is increasing exponentially, harnessing such voluminous data has become a major challenge. To address this problem there is proposed various improved classification methods into Data mining. All this methods use hybrid algorithm to improve classification in data mining. Here hybrid algorithm is nothing but logical combination of multiple pre-existing techniques to enhance performance and provide better results.

Keywords-*Knowledge discovery, Data Classification, Data mining , Hybrid Algorithm.*

*****

I. INTRODUCTION

In Today's world, people are overwhelmed with data which is now increasing exponentially. As the amount of information is increasing, its understanding is equally decreasing. Now there is much useful information is hidden in between the layers of useless information. Here Data Mining is computational process of discovering patterns and useful information using various methods. Data Classification is a data mining technique that is widely used for making group membership prediction of data instances in large database. A data classification model is trained with a set of training data then tested with another set of testing data to verify if its predictions that some data instances belong to some particular groups are sufficiently accurate before it is used to classify a new set of data. Here various classification techniques such as fuzzy decision trees, linear programming, neural network etc.

One of the short coming of decision tree algorithms is 'paralysis of analysis', where decision makers are burdened with information overload due to increasing the size of input dataset. To address this problem various hybrid algorithms are used. But this hybrid algorithms are developed using two or more algorithms which are already available.

II. PRELIMINARES

The algorithm proposed here uses the concepts of classification. These concepts are further explained below.

- ***Data Classification***
  - Data Classification has numerous applications in a wide variety of mining applications. It attempts to find the relationship between a set of feature variable and target variable of interest. Generally, classification algorithm consists of two parts:
    - Training part: this part is a learning step from training instances.
    - Testing part: this part evaluates the model constructed in the training part with another set of test instances. If the model is high performance, it then can be used to classify data.

- ***K means Clustering***

The real life datasets describe a data sample through a number of attributes. There arises a need of grouping these samples on the basis of similarity in their features for analysis of the datasets. A key element in data analysis procedures is grouping of data samples. Clustering is an unsupervised method of partitioning a large number of data objects into subsets called clusters. Clustering is different from supervised classification as the aim is to group a given collection of unlabeled patterns into meaningful clusters. Different clustering methods can generate different groupings for same set of data samples. Clustering can be broadly classified as partition based and hierarchical based. Some examples of the techniques used for partition based clustering are k-means and k-medeoids.

The algorithm proposed in this paper uses k-means algorithm. Initially k number of centroids are defined for clusters as the mean value of the points within the cluster. Then it assigns the data samples to these clusters based on distance parameter between the centroid of the cluster and the sample point. The value of centroids for the clusters is updated iteratively and data objects are reassigned to these clusters on the basis of updated values of centroids. This helps us to achieve a local optimum solution.

 ISSN: 2454-4248
419 – 422

- . ***Decision tree classification***

Decision tree classification is supervised learning technique that tries to divide dataset based on the attributes. A decision tree algorithm is greedy algorithm that uses top-down recursive way to determine the tree structure. The purpose of this algorithm is to make a decision tree from data set to show classification rules. Class labels are selected for classification on the basis on entropy or information values for each attribute.

The formulae used to calculate the value of information gain for each attribute test is that used in C 4.5 algorithm. If S is any set of samples, let freq(Ci, S) be the number of samples in S that belong to class Ci and ~S~ denotes the number of samples in the set S. Then the entropy of the set S is defined as:

$$Info(D) = \sum\left(\frac{freq(Ci,\ D)}{|D|}\right)\log\left(\frac{freq(Ci, D)}{|D|}\right))$$

After set S has been partitioned in accordance with n outcomes of one attribute tes

$$Infox(D) = \sum\frac{|Di|}{|D|}\ Info(Di)$$

Using (1) and (2) we can calculate gain for attribute as:

$$Gain(X) = Info(D) - Infox(D)$$

Attribute with highest gain value is selected as root node for the decision tree. The data samples are divided in decreasing order on entropy for each attribute X. The stopping criteria used to restrict the number of iterations for forming the tree is the minimum number of samples for the leaf nodes.
The decision tree formed is used to analyze the trends or patterns and formulate decisions to achieve desired objectives.

III. LITERATURE REVIEW

Paper (i)

Paper Title: Improving Classification in Data mining using Hybrid Algorithm
Author(s): Akanksha Ahlawat , Bharti Suri
Journal(s): IEEE
Year of Published: 2016

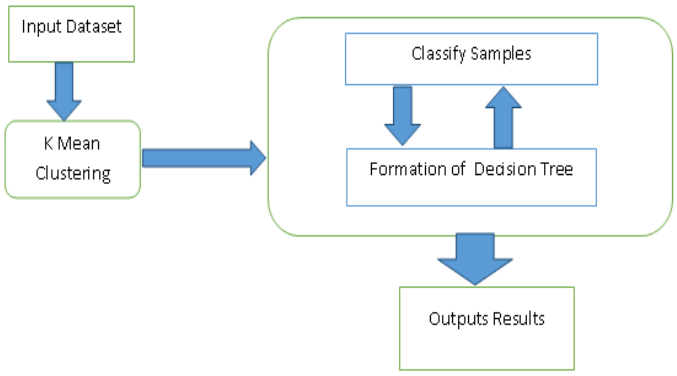
Case Study:

Data mining is a process of finding solution of future trends from set of data gathered from past and current trends. Through this, we can extract hidden knowledge from large datasets using various techniques such as clustering, machine learning, genetic algorithm, etc.

This survey uses the hybrid algorithm of data mining improving classification by make use of different data samples. It is only testing on static data on real life data sets, which shows improved accuracy in most cases.
This case study is not fit for the burdened information overload due to the large set of input dataset. Also into this there is a possibility of duplication of sub-trees on different parts
In this, they are going to input data in tabular form. It process the input data and output the decision tree with their respective accuracy.

Architectural Survey:

Figure to illustrate the architecture view of survey



The above figure describe the combination of two algorithms namely k-mean clustering and decision tree (hybrid algorithm - combination of more than one algorithm).
In this, we are taking input dataset on RDBMS server and applying K-mean clustering on it to get clusters as a output which results on applying decision tree algorithm on it to get output result. (The results are formed with the attribute with highest gain value is selected as root node for decision tree).

Algorithm Overview:-

- Take input datasets and store it's into tabular form.
- Compute information gain values for each attribute.
- Select attribute with heighest information gain value and apply K-Mean clustering on it.
- For each cluster (c) apply decision tree.
- Get output results.

Conclusion:-

This approach solve issues of burdening decision tree with large datasets by dividing into the clusters. But leads to bug in accuracy due to duplication of data and only considers structured data.

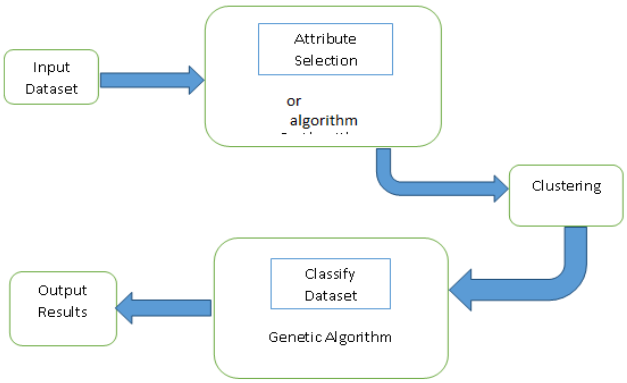
Paper (ii)

Paper Title: Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm

 ISSN: 2454-4248
419 – 422

Author(s): Neelam Singhal , Mohd.Ashraf
Journal(s): IEEE
Year of Published: 2015

Case Study:
Clustering basically allows users to make group of data to determine the patterns from it. the advantage over it is to make attributes to analyze data over it. By using the classification one ca extract the benefits of both supervised and unsupervised technique thus increasing the relevance of data. This includes the input dataset as an input which in turns going to apply a DT algorithm using attribute selection feature and parallely made classify dataset by applying genetic algorithm. Combined tother to form clustering and then it produces output as a result.



Algorithm:

- Take a dataset as input
- Divide dataset into training and testing dataset else,use n-fold cross validation.
- Select the AttributeSelection which is used for feature selection.
- Choose evaluator as ClassificationSubsetEval as it let us select features using a classifier.
- After that choose j48 under the Decision Tree as the classifier.
- Choose BestFirst searching method for search in feature selection.
- Apply the feature selection to the dataset.
- Choose AddCluster to perform clustering the dataset.
- Use SimpleKMeans cluster which is the simplest type of clustering technique.
- Choose number of cluster as the number of classes in the dataset.
- Put the index of the class attribute under the ignored AttributeIndices option so that unbiased cluster are created.
- Apply the clustering technique on the filtered dataset.
- Now classify the dataset using Genetic Programming.
- End.

Conclusion:-
It is applying hybrid approach(combination of decision trees and genetic algorithm) via feature selection and clustering which increases the accuracy but is only for limited input.

Paper (iii)
Paper Title: An Extended ID3 Decision Tree Algorithm for Spatial Data
Author(s): Imas Sukaesih Sitanggang , Razali Yaakob , Norwati Mustapha , Ahmad Ainuddin B Nuruddin
Journal(s): IEEE
Year of Published: 2011

Case Study:
This case study aims of utilizing data mining tasks such as classification on spatial data is more complex than non-spatial data. spatial information gain to choose the best splitting layer from a set of explanatory sets. one layer relates to other layers to create objects in a spatial datasets by applying spatial relations such as topological relations and metric relation. Spatial decision trees refers to the model expresing classification rules induced from spatial data. The algorithm considers not only attributes of the object to be classified but to consider also attributes of neibouring objects.Instead of using number of tuples in a partition, spatial information gain is calculated using sptial measures namely area.

Conclusion:
This concludes to fix data to a spatial data only containing discrete features. In this layers are using to seperates the dataset into smaller partitions that belong to same class only. It treads to high response time and complexity.

Paper (iv)
Paper Title: A New Hybrid Model of PSO and DE Algorithm for Data Classification
Author(s): Wannaporn Teekeng , Pornkid Unkaw
Journal(s): IEEE
Year of Published: 2017

Case Study:
Data classification is a data mining technique that is widely used for making group membership predictions of data instances in large datasets. Data classification has myraid of applications is a various applications of mining. It usually consists of two parts: training part and testing part. It begins with the input data which in turns generates initial particles and then apply objective function on it so as to acheiveoptimized solution. It then helps in improving fight and particles strength. selection criteia on particals is then processed.Finally a training on testing took place to end the whole process.

 ISSN: 2454-4248
419 – 422

Conclusion:

It concludes to combine classification algorithm of PSO and DE algorithm. This leads to help in generating better solution but has lower response time due to the diversify the particles.

IV. CONCLUSION

After all the study we can conclude that in traditional algorithm use hybrid algorithm (combination of two existing algorithm) to improve the classification method in data mining. There all hybrid algorithms takes input data in tabular form and all testing dataset is static. Also it reduce the accuracy as per dataset is increased because it creates duplications. There response time also slow because it use more than on algorithm to improve classification.

REFERENCES

[1] Akansha Ahlawat, Bharti Suri.”Improving Classification In Data Mining Using Hybrid Algorithm” IEEE 2016.

[2] Neelam Singhal, Mohd Ashraf.”Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm”

[3] Imas Sukaesih Sitanggang, Razali Yaakob, Norwati Mustapha, Ahmad Ainuddin B Nuruddin. “An Extended ID3 Decision Tree Algorithm for Spatial Data”. IEEE 2011.

[4] Wannaporn Teekeng, Pornkid Unkaw. “A New Hybrid Model of PSO and DE Algorithm for Data Classification”. IEEE June 26 2017.

[5] Masaki Kurematsk, Hamido Fujita. “ A Framework for Integrating a Decision Tree Learning Algorithm and Cluster Analysis”. 12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques September 22-24,2013, Budapest, Hungary.

[6] Mrutyunjaya Panda, Arijit Abraham. “Hybrid evolutionary algorithm for classification datamining”Springer. 10 August 2014

[7] Jjiawei Han, Micheline Kamber, Jian Pei “Data Mining: Concepts and Techniques”3rd edition published by Morgan Kaufman.

[8] Yao Yu, Fu Zhong-linang, Zhao Xiang-hui, Cheng Wen-fang.” Combining classification based on decision tree” IEEE International Conference on Information Engineering Vol 2. July 2009.

[9] K.C. Tan, E. J. Teoh, Q. Yu, k. C. Goh. “A Hybrid evolutionary algorithm for attribute selection in data mining”. Expert system with application 2008, Elsevier.

[10] Harvinder Chauhan, Anu Chauhan, “ Evaluating performance of Decision Tree Algorithms,” International Journal of Scientific and Reserch Publication, Volume 4, Issue 4, April2014.

[11] Linna Li, Xuemin Zhang.”Study of Data Mining Algorithm Based on Decision Tree” 2010 International Conference On Computer Design And Application(ICCDA 2010) 2010 IEEE.

[12] Chinnpat Kaewchinporn, Nattakan Vongsuchoto, Anantaporn Srisawat.”A Combination of Decision Tree Learning and Clustering for Data Classification” 2011 8th International Joint Conference on Computer Science and Software Engineering(IJCSSE) 2011 IEEE.

[13] Bhaskar N. Patel, Satish G. Prajapati, Dr. Kamaljit I. Lakhtaria” Efficient Classification of Data using Decision Tree. Bonfring International Journal of Data Mining, Vol.2 No. 1 March 2012.

[14] Dharm Singh, Naveen Chodhary, Jully Samota “Analysis of Data Mining Classification with Decision tree Technique” Global Journal Of Computer Science And Technology Software And Data Engineering vol13 year 2013.