A Survey on Hard Computing Techniques based on Classification Problem

Ankita Mishra[1] Bed Prakash[2] Abha Choubey[3]
[1]M.E. Scholar [2]PT Lecturer [3]Associate Professor
[1,3]Department of Computer Science & Engineering [2]Department of Information Technology
[1,3]SSTC Bhilai (C.G.) [2]UPU Govt. Poly. Durg (C.G.)

Abstract— In this survey paper we survey about various techniques related to Hard computing techniques useful for improve efficiency and time complexity to solve complexity occur in classification techniques in data mining. In this survey we analyzed that multiple hard computing techniques provide different feature such as K-medoid, FCM, PAM, CLARA, CLARANS, BIRCH, CURE, ROCK and CHAELEON etc. In above mentioned techniques very useful and reliable techniques that to be introducing in this survey paper is K-means. K-means techniques based on hard computing and k number of cluster partitions concept. We also analysis of UCI library dataset that purely based on classification i.e. Iris, Wine, PimaIndians, Shuttle, Magic etc.

Key words: Cluster, Analytical Model, Crisp Analysis Numerical analysis

I. INTRODUCTION

These Now a days there is enormous measure of Data being accumulated and set away in databases everywhere over the globe. The slant is to keep growing truly quite a while. It is not hard to find databases with Terabytes of Data in endeavors and investigation workplaces. That is more than 1,099,511, 627,776 bytes of Data. There is critical information and Data "concealed" in such databases; and without customized methodologies for removing this Data it is in every way that really matters hard to burrow for them. Amid the time various counts were made to uproot what is called bits of gaining from broad plans of Data. There are a couple of unmistakable methods of insight to approach this issue: Classification, Associations standard, Clustering, thus on[19].

Information Classification is an alternate system that incorporates diverse techniques and criteria for sorting Data within a database or vault. This is generally done through a database or business learning programming that gives the ability to clear, recognize and disengage data. A couple tests and utilization of Data Classification include:

- Distinguishing and keeping as frequently as could reasonably be expected used data as a piece of plate/memory store[12].
- Data sorting in perspective of substance/record sort, size and time of information[10].
- Classifying in order to sort for security reasons data into restricted, open or private data sorts[11].

Characterization is utilized as a part of various regions like bio-informatics, science, hereditary qualities and Healthcare etc.

Data Classification and Data identification have information all about store lots of facts that's you are tagging in backup. So it can be retrieve easily and work with efficient way. But in organization can also save their original information as well as duplicating information every day. Which cuts capacity and reinforcement costs, whilst accelerating information seeks[19]. The two major problem-solving technologies include:

- Hard computing
- Soft computing

In this survey paper we introduce hard computing techniques for solving data classification problem. Generally two techniques have already introduced first is tradition techniques which is called Hard Computing and second is modern techniques, which is called Soft Computing[14].

Hard Computing deals with precise models where exact arrangements are accomplished rapidly. Traditional computing techniques taking into account standards of exactness, instability and thoroughness. The issues taking into account diagnostic model can be effortlessly tackled utilizing such strategies. True issues which manage changing of data and loose conduct can-not be taken care of by hard computing techniques. Generic Algorithm[15] and Particle Swarm Optimization Techniques have developed as potential and vigorous improvement apparatuses as of late. The course substance will be taught by famous specialists in the field, having satisfactory showing and research experience. This course will be valuable to staff from every building order as a potential registering device in their examination exercises. This course is gone for recently enlisted educators (Less than five years Experience), composed from the fundamental ideas to application and surveys[1].

II. CLASSIFICATION TECHNIQUES

Based on our analysis two basic techniques are used for classification Problem[13]:

- Hard Computing
- Soft Computing

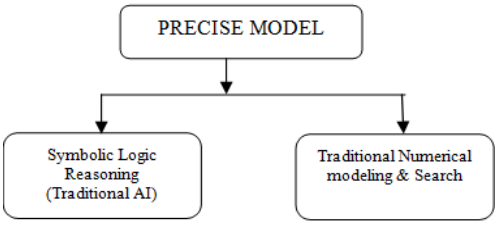

Fig. 1: Model

A. *Various Extracted features of Hard Computing:*

Hard techniques are purely based on a Traditional method. Hard Computing deals with precise models where exact arrangements are accomplished rapidly. Traditional computing techniques taking into account standards of exactness, instability and thoroughness. The issues taking into account diagnostic model can be effortlessly tackled utilizing such strategies[10].Hard Computing have various features[2]:

- Hard computing schemes, which strive for exactness and full truth[1].
- Hard computing based on binary logic, crisp systems, numerical analysis and crisp software.
- Hard computing has the characteristics of precision and categoricity[4].
- Hard computing requires programs to be written.
- Hard computing uses two-valued logic.
- Hard computing is deterministic.
- Hard computing requires exact input data.
- Hard computing is strictly sequential.
- Hard computing produces precise answers.

III. Different Techniques of Hard Computing Methodologies

Hard Computing techniques generally use established method of arithmetic, Science and computing to solve Classification Problem[1].It has various methods related to hard computing techniques:

- Binary Logic,
- Boolean Logic,
- Analytical Model,
- Deterministic Search,
- Crisp Analysis,
- Numerical analysis.

Hard Computing is the thing that we have been used to, what has been going on, right from the earliest starting point of computational science[1]. All the traditional thinking and displaying approaches that depend on Boolean rationale, investigative models and fresh orders fall in this classification. Hard computing drops the hammer on accuracy ruling out approximations. This can be computationally costly, time devouring and once in a while even unthinkable for application to complex genuine issues since numerous such issues are normally sick characterized frameworks, hard to demonstrate and with vast arrangement spaces. In such cases, delicate computing acts the hero however with a cost of bargaining on a percentage of the standards of hard figuring. It is tolerant of imprecision, vulnerability, incomplete truth, and approximation[5]. The managing rule of delicate figuring is: adventure the resistance for imprecision, vulnerability, incomplete truth, and estimation to accomplish tractability, vigor and low arrangement cost[11].

IV. K-cluster Methodologies

Most Organization produces unlimited bulk of data daily and store them in Database. This large amount of data has valuable hidden fact[1].

Many research proposed various clustering algorithm for solving classification problem (because Classification is unsupervised whereas clustering is supervised method).The two main techniques are proposed in clustering is Partitioning clustering and hierarchical clustering. Most of the clustering algorithm in the survey are K-means, K-medoid, FCM, PAM, CLARA, CLARANS, BIRCH, CURE, ROCK and CHAELEON[1]. Above mentioned among the techniques mostly K-means and FCM algorithm are used portioning techniques by the many surveys.

A. Algorithm:

K-means[3] is the most popular hard clustering algorithm. Each data point belongs to only one cluster. This method required previous knowledge about number of cluster. In this techniques, k indicates number of cluster

K-cluster calculation has three stages including:

- Step 1) k cluster center are indicated, arbitrarily i.e. one center for every cluster,
- Step 2) for every input, separation from each cluster center is calculated. The input data fits in with the cluster which has the nearest distance from the middle. This step is repeated for all input, and
- Step 3) the barycenters of cluster (which are produced in step 2) are figured and considered as new cluster center and after that the calculation goes to step 2.

These steps are repeated until centers do not change for the two consecutive iterations[9][1].

- The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, where $k < n$.
- It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data.
- It assumes that the object attributes form a vector space.

An algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing data points so as to minimize the sum-of-squares criterion:

$$j = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2$$

Where x_n is a vector representing the the n^{th} data point and u_j is the geometric centroid of the data points in S_j. Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid[1].

B. Flow Diagram:

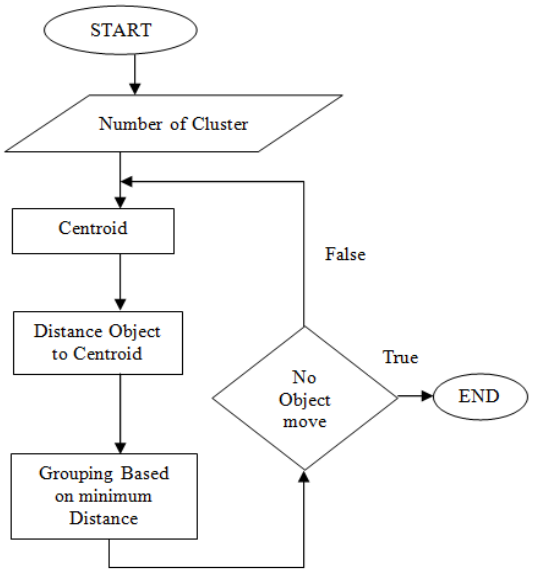

Fig 1: K-means Clustering Algorithm[9]

C. DataSet with Cluster:

Data Set	Size	Number of Attribute	Number of Cluster
Iris	150	4	3
Wine	178	13	3
Pima Indians	768	8	2
Shuttle	43,500	9	2
Magic	19,020	10	2

Table 1: Various dataset extracted from uci library

V. RESULT

K-means Clustering have been used in many real world area. This techniques applied for many real areas like image segmentation, patterns in biological sequences, face recognition, cellular manufacturing, network intrusion detection system, prediction of students academic performance, meteorological data application, application in university libraries, evaluation of success of software reuse and so on. In this survey paper data taken from the UCI library which is fast as compare to other clustering methods.

VI. CONCLUSION

Recently, lots of data extract from different large organization faced classification problem and several works have used clustering and classification in sequential structure to increase the performance mean that time and space complexity of classification algorithm. It is represented the efficiency of classification learning is enhanced if the input data is first clustered and then used for classification. According to analysis of Table1 various data set that applied k-means clustering algorithm that is partition various cluster and data takes at their particular cluster..

The general approach can be applied for improve the quality of data complexity problem that can be caused during create classes.

REFERENCES

[1] A Comparative Study of Hard and Fuzzy Data Clustering Algorithms with Cluster Validity Indices By O. A. Mohamed Jafar1, and R. Sivakumar.
[2] Okyay Kaynak, Senior Member, IEEE, Kemalettin Erbatur, and Meliksah Ertugrul," The Fusion of Computationally Intelligent Methodologies and Sliding-Mode Control—A Survey", © IEEE Transactions On Industrial Electronics, Vol. 48, No. 1, February 2001.
[3] Gehard J.Woginger,"Extract Algorithm for NP-hard Problem ", International Journal on Soft Computing (IJSC), Vol.2, No.3, August 2011.
[4] Mark J. Embrechts Rensselaer Polytechnic Institute, Troy, New York, USA ,Boleslaw Szymanski Rensselaer Polytechnic Institute, Troy, New York, USA ,Karsten Sternickel Cardiomag Imaging Inc., Schenectady, New York, USA," Introduction to Scientific Data Mining: Direct Kernel Methods & Applications ",The fusion of hrd and Soft computing,Weily NewYork 2005.
[5] Francisco Herrera," Introduction: Genetic Fuzzy Systems",INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 13, 887 890 1998.
[6] Mr. Ankit R. Deshmukh1, Prof. Sunil R. Gupta," DATA MINING BASED SOFT COMPUTING METHODS FOR WEB INTELLIGENCE", International Journal of Application or Innovation in Engineering & Management (IJAIEM).
[7] Dharmendra Kelde, Deepak Nagde, Raviraj Patel Pavan Pawar," Information Forensic Application using Soft Computing Techniques", International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 69 - 72.
[8] Choo Jun TAN1, Chee Peng LIM2*, Yu-N CHEAH1, Shing Chiang TAN," Classification and Optimization of Product Review Information Using Soft Computing Models", International Symposium on Affective Engineering 2013 (ISAE 2013).
[9] K-means algorithm with their flow diagram in slideshare.net/parryprabhu/k-meanclustering-algorithm
[10] Rudolf Kruse, Christian Borgelt, Detlef D. Nauck, Nees Jan van Eck and Matthias Steinbrecher," The Role of Soft Computing in Intelligent Data Analysis ".
[11] Anna Thomas and Karthik Pattabiraman," An Intermediate Code Level Fault Injector For Soft Computing Applications".
[12] P. Anil babu, K.Koteswara Rao," Software Testing Using Soft computing Technique ",International Journal of Advanced Research in Volume 3, Issue 8, August 2013 .
[13] Min Pei, Erik D. Goodman, William F. Punch III and Ying Ding ,"Genetic Algorithms For Classification and Feature Extraction".
[14] "Feature Extraction, Construction and Selection: A Data Mining Perspective" edited by Huan Liu, Hiroshi Motoda.
[15] Bruce Vanstone ,"A Survey of the Application of Soft Computing to Investment and Financial Trading"
[16] S. Manoharan,"A Comparison and Analysis of Soft Computing Techniques for Content based Image Retrieval System", International Journal of Computer Applications (0975 – 8887) Volume 59– No.13, December 2012.
[17] Bezdek J C. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology, 1(1), pp. 57–71, 1974.
[18] BezdekJ C. Mathematical models for systematic and taxonomy. In: Proceedings of 8th international conference on numerical taxonomy, San Francisco, pp. 143–166, 1975.
[19] Mustafa Karabulut and Turgay Ibrikci. Fuzzy c-means based DNA Motif Discovery. Lecture Notes in Computer Science, Advanced intelligent computing theories and applications with aspects of theoretical and methodological issues, vol. 5226, pp. 189–195, 2008. R. Xu, D. Wunsch II," Survey of Clustering Algorithms". IEEE Transactions On Neural Networks, Vol. 16, No. 3, MAY 2005.
[20] Cifarelli C, Manfredi G and Nieddu L. Statistical face recognition via a k-Means iterative algorithm. Seventh international conference on machine learning and applications (ICMLA'08), pp. 888–891, 2008

