An Improved Optimized Clustering Technique For Crime Detection

Nidhi Tomar (*Research Scholar*)
Dept. of CSE/IT
MITS
Gwalior, India
tomarnidhi4@gmail.com

Amit Kumar Manjhvar (*Asst.prof*)
Dept. of CSE/IT
MITS
Gwalior, India
amitkumar@mitsgwalior.in

***Abstract*—Data mining automates the finding predictive records procedure in big databases. Clustering is a most famous method in data mining and is an important methodology that is performed based on the similarity principle. The segregation of a big database is a stimulating and task of time consuming. It concludes two different stages: first, feature extraction maps all documents or record to a point in the space of high-dimensional, then algorithms for clustering automatically grouping the points into a cluster hierarchy. Clustering has various applications in different fields. Few of the fields include criminology, text mining, image resolution, machine learning. Crime detection has become one of the most attractive field as the crime rate in India and whole world is increasing at a greater pace. We as citizens of a country have to contribute towards its detection and removal. Thus, a comprehensive survey carried about the basics of clustering has given in this paper. Moreover, proposed work was given that gives the idea of the work going to be done in the upcoming time.**

Keywords—Clustering; partitioning; data mining; crime; k-means method optimization.

I. INTRODUCTION

This Data mining concern about to extracting and mining meaningful information from the largest dataset. It is the search for the relation and global pattern that exist in the huge database but are concealed among vast amounts of data, such as the connection between patient data and their medical diagnosis [1]. The main principle of data mining is to select information from a bulk data set and translate it into an intelligible structure for further use. This relationship represents expansive knowledge about the database and the object in the dataset. This mining technique purpose is to mine information from an enormous data set and make over it into a reasonable form for the supplementary purpose [2]. Data mining is a multistep process:

1. Selection:

The data that is useful and is related to the basic purpose is selected. The data which is selected is retrieved from the database for the further processing.

2. Preprocessing:

Preprocessing is the data cleaning stage where redundant information is removed.

3. Transformation:

Data is not merely transformed across mining, but transformation in order to be suitable for the task of data mining. In this stage, the data are made usable and navigable.

4. Data mining:

Pattern from the data is worried from this stage. Data mining is a multidisciplinary sub field of computer science.

5. Pattern evaluation and knowledge information:

The patterns obtained in data mining stage are converted into knowledge.

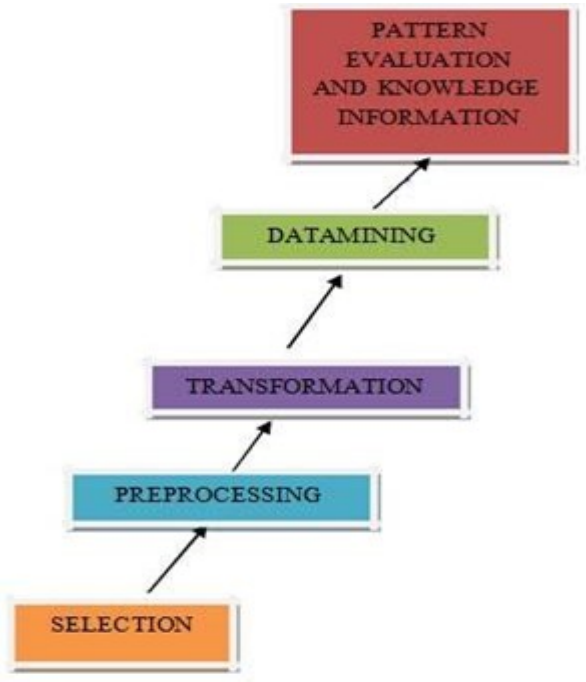

Fig. 1. KDD Process Steps

In the knowledge discovery process steps, data mining is the main step [4]. The main functions of data mining are:

978-1-5090-0669-4/16/$31.00 © 2016 IEEE

1. *Anomaly detection*
Anomaly detection is the recognition of odd data records that may be remarkable. It is the identification of items, events or observation, which does not conform to expected patterns or other items in a data set.

2. *Association rule*
Association rule training is the process to find the connection between the variables. These rules are if/then statements that help uncover a relationship between ostensibly unrelated data.

II. CRIME DETECTION USING CLUSTERING TECHNIQUES

Crime is one of the dangerous factors for any country, analysis of crime is the action in which analysis is done on crime activities. Today criminals have maximum use of all modern technologies and scientific methods in committing crimes [16]. The law enforcers have to effectively meet out challenges of crime control and maintenance of public order. One test to law implementation and insight organizations is the trouble of breaking down vast volumes of information included in criminal and terrorist exercises. Thus, making of information base for wrong doings and crooks is required. Information mining holds the guarantee of making it simple, advantageous and reasonable to investigate vast databases for associations and clients. we have detect the crime on the basis of some criteria-
1. Intelligence
2. Transference
3. Inquisition
4. Authoritative

III. CLUSTERING

Clustering is a major field of data analysis and data mining application. It is a set of methodologies for produce high superiority clusters and high intra-cluster similarity and low inter -class similarity. The types of data used for analysis of clustering are interval scatted variables binary, nominal, ordinal, ratio variables of mixed types.

A. Hierarchical

A hierarchical method creates a disintegration of the given dataset of the objects hierarchically. Here the tree of clusters called as dendrogram is built. Each cluster node contains child clusters siblings. In hierarchical clustering, we assign.

a) Agglomerative

In agglomerative hierarchical clustering is a bottom up approach each scrutiny starts in its seize cluster, and pairs of cluster are merged as one moves up and about to hierarchy.

b) Divisive

Divisive is a top down approach clustering method all observation starts from one cluster and splits are performed on iterative as one step down the hierarchy unavoidable.

B. Partitioning Based

In the partitioning algorithm split data points in k cluster.The partition where each point into a cluster and partition is done based on certain objective function.

C. Density based

If cluster grows, according to density of neighboring objects and is based on the concept of density reach ability and density connectivity both of which are depends on the input parameter size of the epsilon e neighborhood and minimum terms of local distribution of the nearest neighbor. DBSCAN focusing on low dimensional spatial data used DENCLUE algorithm.

D. Grid based

This grid based methods are object space quantizes into a finite number of cells that form a lattice structure. All clustering operations are performed on grid structure, main advantage of this approach is fact process time, which classically independent of the number of data objects and dependent only on the cells in each dimension in Quantize space.

E. High dimensional data clustering

It is a most important tack in cluster analysis because a lot of applications require the analysis of an object. Contain a large number of features or dimension for ex-a text document may contain thousand of keyword as a feature. This high dimensional data method of clustering is challenging due to curse of dimensionality. Many dimensions may not be relevant, the number of dimension gradually more sparse, so that the distance measurement is likely to be low CLIQUE and PROCLUS two subspace clustering method.

F. Constraint based

This constraint based clustering method performs clustering by including user specific or application oriented constraint. A constraint describes user express and provides an effective means for communication with clustering process [7].

IV. CRIME DETECTION IN INDIA

Data mining is to model crime detection problems. Crime is a social dearly in several ways. Any research that can help in solving crimes faster will pay for itself. About 20% of the crime in this paper we use a clustering algorithm for a data mining method to help detect the crimes and its patterns and speed up the process of solving crime .we look at k-means clustering with some enhancement to aid in the process of identification in of crime patterns and optimize the type of crime in India and other country.

V. CLASSIFICATION OF CLUSTERING TECHNIQUES

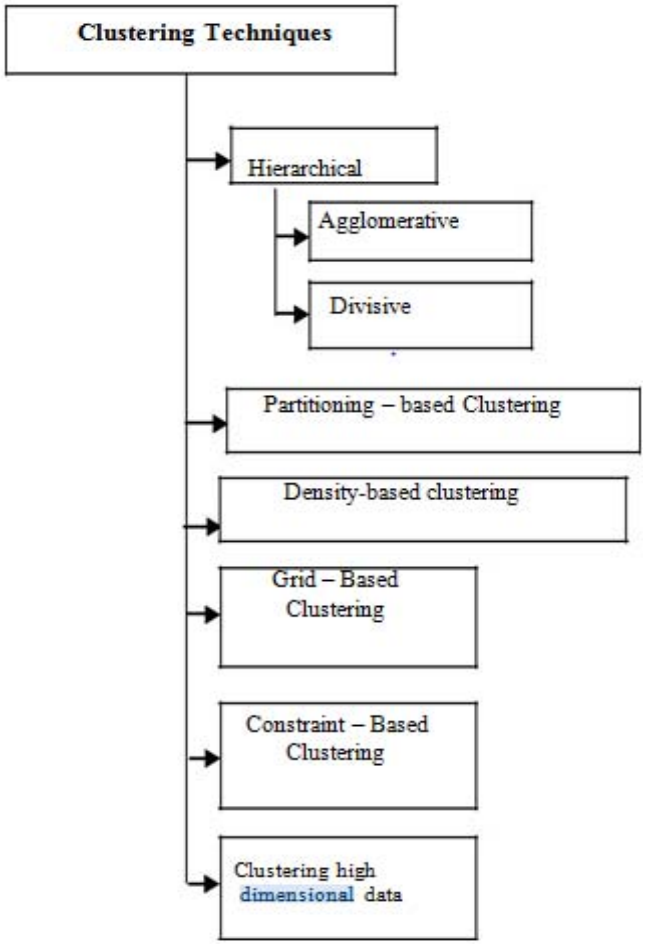

Fig. 2. Clustering Techniques

A. Techniques used in clustering analysis

a) Feature Selection Methods: The feature selection phase is an significant preprocessing step which is required in order to increase the superiority of the essential clustering. Not all features are similarly appropriate to finding clusters, while some may be further rowdy than another. consequently, it is often cooperative to use a preprocessing phase in which the noise and immaterial features are pruned from contention .attribute selection and dimensionality reduction are directly related .In feature selection ,original subsets of the features are selected[10].

b) Probabilistic and Generative Models: In probabilistic models, the core initiative is to replicate the data from a production process. First a exact form of the generative model (e.g., mixture of Gaussians) is unspecified, use of the hidden Markov model [11].

c) Distance based Algorithms: Many special forms of generative algorithms can be shown to reduce to distance-based algorithms. This is because the mixture components in generative models often use a distance function within the probability distribution. For example, the Gaussian distribution represents data generation probabilities in terms of the Euclidian distance from the mean of the mixture. As a result, a generative model with the Gaussian distribution can be shown to have a very closely related, throughout the k-means clustering methods. In actuality, several distances-based algorithms can be shown to be reductions from or simplifications of different kinds of generative models. Distance-based methods are often desirable because of their simplicity and ease of implementation in a wide variety of scenarios.

d) Density and Grib based Methods

This density- and grid-based methods are closely related classes, which try to explore the data space at high levels of granularity. The density at any particular point in the data space is defined either in terms of the number of data points in a pre specified volume of its locality or in terms of a smoother kernel density estimate. A major advantage of these methods is that since they explore the data space at a high level of granularity. Two classical methods of density-based methods and grid-based methods are DBSCAN [12] and STING [4], respectively.

TABLE I. COMPARISON TABLE OF COMPLEXITY

Clustering Algorithm	Complexity	Working Process And Functionality
Hierarchical Clustering	O(N)^2(time) O(N)^2(space)	It forms a tree structure use agglomerative and divisive no The input parameter is required
Partitioning Clustering	O(N*K*d)(time) O(N + d)(space)	K-means algorithm is built by classifying data to groups of objects based on their attributes based clustering/feature into K no. of groups.
CLARANS	Quadraticintotal performance	CLARANS use randomized search to facilitate the clustering of a large number of objects. It is more efficient of PAM and CLARA.
DENSITY BASED	O(N)(time)	Density based scanning is based on density reach ability and density connectivity.
DBSCAN	O(N log N)	It use medoids and means for cluster
FAST FARTHEST	O(N)time O(N)space	Farthest first algorithm builds by classifying data to group of objects.

VI. LITERATURE REVIEW

Fey We Georges [13] an important problem in clustering is how to efficiently answer ranges queries given an object of and search distance do we want to retrieve all objects whose distance to O are less than to D.

Martin Easter, Han's Peter et al[14] The problem of incremental updating mined patterns after making changes to the database has just recently started to receive more attention. The concept of association rule mining has been introduced by [As94],an association rules are a disjoint subset of the set of an item set.

Abbas Bahrololoum Hossein Nezamabadi [3] Recently more and more attention has been focused on nature inspired algorithm to solve clustering problems. Furthermore, there are many clustering algorithms, which have been made hybridizing different types of the evolutionary algorithm into k-means algorithm for overcomes the disadvantages of k-means algorithm. Evolutionary algorithm such as Ant colony optimization, Genetic algorithm, Bees algorithm, Gravitational search algorithm, is usually nature inspired algorithms.

B. Suresh Kumar, H. Venkateswara et al [15] a clustering algorithm based discussion on various algorithms on clustering of data labeling techniques are taken up in this section. In numerical data, cluster representation is used to characterize and summarize the results were as in categorical data; this is not in a similar manner.

R. G. Ultra [16] – In Data Mining crime mining is defined as the discovery of interesting structure in data, wherever the structure designates patterns, statistical or predictive models of the data, and relationships represented in the middle part of the data. In the crime mining for some results using data mining methods . This technique is applied to study criminal cases, which mainly concerned entity extraction, pattern clustering, classification and social network analysis. This method used to get the data of criminals by using frequency rate of the incident.

Mohamed lafar and R.Sivakumar [17] have discussed about Ant-based clustering as a biologically inspired data clustering technique. Clustering methods, aims is the unsupervised classification of patterns in different groups .Clustering problem has been approached from different disciplines. In recent years, several algorithms have been developed for solving numerical and combinatorial optimization problems, mainly promising among them are swarming intelligence.

Mythili S1 and Madhiya E2 [18] Different approaches to clustering data can be described with the help of hierarchical (other axonometric representation of clustering methodology are possible are based is a discussion on Jain and Dubes) at the top level, there is a distinction between hierarchical and Partitioning.

VII. PROPOSED METHODOLOGY

s Clustering is the major field of the data mining has been used in a number of applications like engineering, biology, medicine and data multimedia. One of the most commonly used algorithms is K-Means algorithm of clustering. It partitions the objects into clusters by minimizing the sum of the squared distances between the centroid of the cluster objects. The K-Means clustering is simple, but it has high time complexity, so it is unsuitable for large data set. The algorithm allows early termination of the distance calculation by introducing an untimely exit condition in the search process.

As seen in the literature, the researchers contributed only to accelerate the algorithm, there is no involvement in cluster refinement. We have proposed Ant Colony Optimization (ACO) algorithm to improve the K-Means algorithms. The ant colony optimization algorithm (ACO) is used for solving computational problems to produce good results through graphs. In the natural world, ants (initially) wander randomly, and upon finding food to return their colony while laying down pheromone trails.

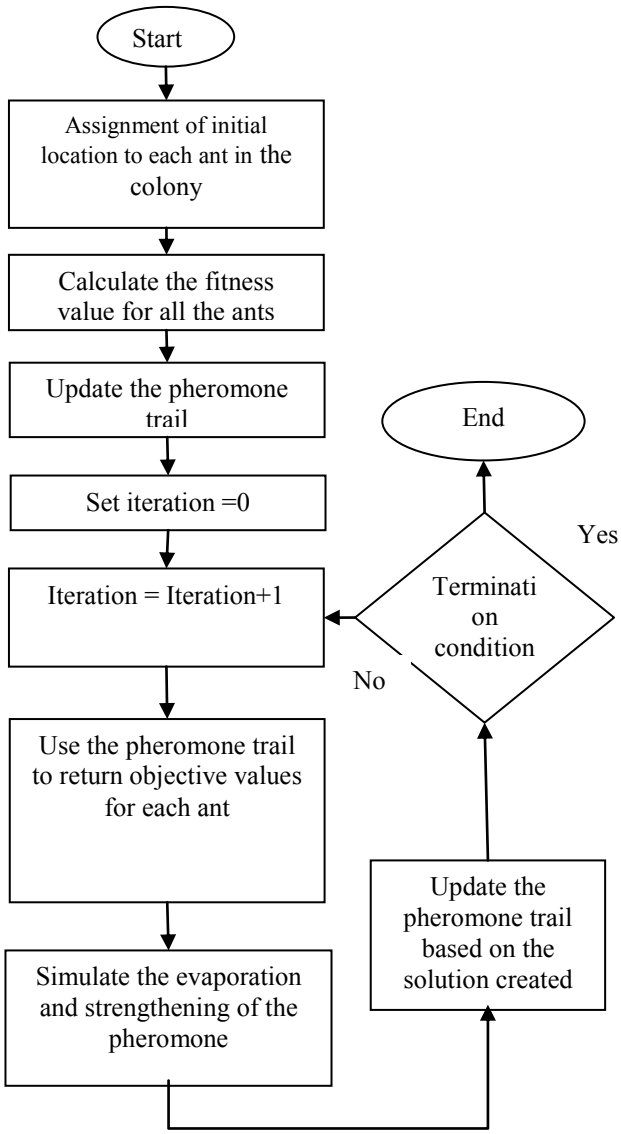

Fig. 3. Proposed Flow Chart of ACO

The proposed algorithm is applied to refine the cluster to

improve the quality. The basic reason for our refinement is, in any clustering algorithm the obtained clusters will never give us 100% quality. There will be some errors known as clustering. That is, a data item can be wrongly clustered. In K-Means algorithm, the initial cluster centers are normally chosen either sequentially or randomly as given in the standard algorithm. The quality of the final clusters based on those initial seeds may lead to local minimum; this is the disadvantage in K-Means clustering algorithm. The disadvantages of K- Means algorithm can be improvised by using ACO based cluster refinement algorithm Crime management is a significant and interesting application .It has a significant role in society. There was a serious increase in this area in the past few years, thus it has become significant for us to create a system for proper crime detection. In today's world, security is a manner which gives higher priority by all political parties and governments worldwide aiming to reduce crime incidence. As data mining is the appropriate field to apply for high- volume crime data set which knowledge gained from data mining approaches will be a useful and support police force for crime analysis. So In this paper crime analysis is done by performing k-means clustering algorithm on crime dataset. There is a wide variety of crimes are considered misdemeanors in most

- Public intoxication
- Trespassing
- Speeding
- Prostitution
- Vandalism
- Use of a false ID

Thus, we are working in the field of crime detection using an optimized technique of ACO and also improved algorithm is applied to improve the crime detection capability. Using the above formed, defined flowchart of ant colony optimization, we are going to improve the optimization in the optimization in the field of crime detection.

CONCLUSION

Cluster analysis is the most significant data mining method used to discover pattern information and data segmentation. Through data clustering, people can find the data distribution, character observes of all clusters, and creates further study of the specific clusters. In addition, analysis of the cluster typically acts as preprocessing of other different operations of data mining. Therefore, analysis of the cluster has become most active topic of research in the data mining. As development of data mining, various clustering approaches have been founded, The study of the clustering method from the statistic's perspective, based on statistical theories, our paper creates an effort to combine statistical technique with computer algorithm method, and introduce the existing outstanding statistical approaches, including analysis of factor, correspondence analysis, and analysis of functional data, into the data mining. The paper discussed about the clustering and the various aspects of clustering that effects. A brief idea about the proposed work Is given. Which will be implemented in near future. We suppose that this field is less viewed and thus, we need to help the government in this field.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2nd ed., The Morgan Kaufmann Series in Data - Management Systems, Jim Gray, Series Editor 2006

[2] Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. "Knowledge Discovery and Data Mining: Towards a Unifying Framework."In KDD, vol. 96, pp. 82-88. 1996.

[3] Bahrololoum, Abbas, Hossein Nezamabadi-pour, and Saeid Saryazdi "A data clustering approach based on a universal gravity rule." Engineering Applications of Artificial Intelligence 45 (2015): 415-428.

[4] "Survey paper on clustering techniques in data Mining" International Archive of applied sciences and technology Volume 3 June 2012.

[5] Mohri, M., Rostamizadeh, A. and Talwalkar, 2012." Foundations of machine learning". MIT press.

[6] Bousquet, Olivier, Stephen Boucheron, and Gábor Lugosi. "Introduction to statistical learning theory" In Advanced lectures on machine learning, pp. 169-207. Springer Berlin Heidelberg, 2004.

[7] Han, J., Kamber, M. and Pei, J., 2011. "Data mining: concepts and techniques". Elsevier publication

[8] Patnaik, Sovan Kumar, Soumya Sahoo, and Dillip Kumar Swain, "Clustering of Categorical Data by Assigning Rank through Statistical Approach," International Journal of Computer Applications 43.2: 1-3, 2012.

[9] Jain Anil K., M.Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999)

[10] Gan, Guojun, Chaoqun Ma, and Jianhong Wu. "Data clustering: theory, algorithms, and applications". Vol 20 Siam, 2007. 264-323

[11] Verma, Manish, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, and Nidhi Gupta. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2, no. 3 (2012): 1379-1384

[12] Wu,Fei, and Georges Gardarin."Gradual clustering algorithms "In database system for Advanced Applications" 2001 Proceedings Seventh International Conference on, pp.48-55.IEEE 2001.

[13] B.Suresh Kumar, H.Venkateswara Reddy and T.Ankamna Raju, Preethi Vennam "clustering categorical data using rough membership function " International Conference on Computational Intelligence and communication network.

[14] Martin Ester and Hans-peter Kriegel, Jorge Sander, Michael Wimmer, Xiaowei xu" Incremental clustering for in a data warehousing environment"

[15] Uthra, R.G., "Data Mining Techniques to Analyze Crime Data" International Journal for technological research in engineering volume1, 2014.

[16] Mohamad Saree and Najmeh Ahmadian and Zahra Narimani "Data mining process using clustering techniques survey".

[17] MythiliS1, MadhiyaE2 "An analysis of clustering techniques in data mining" International Journal of Computer science and Mobile Computing in volume 3 in 2014.