www.ijiarec.com

ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Data Mining Techniques for Analyzing Crime

[1]Ms. N. ZahiraJahan, M.C.A., M.Phil.,Associate Professor/MCA
[2]Ms. K. Lavanya, III MCA
Department of MCA, Nandha Engineering College (Autonomous) Erode – 52.
Email ID: zahirajahan1977@gmail.com, lavanyakannan635@gmail.com

***Abstract*: The data mining is data analyzing techniques that used to analyze crime data previously stored from various sources to find patterns and trends in crimes. To solve the problems previously mentioned, data mining techniques employ many learning algorithms to extract hidden knowledge from huge volume of data. Data mining is data analyzing techniques to find patterns and trends in crimes. It can help solve the crimes more speedily and also can help alert the criminal detection automaticallyThe previous system focus in mining the crime data from the crime database for that the KNN clustering is used for clustering the data. The values are classified by using the crime value. Crime values are taken from the crime database. To identify crime falls under category, each crime data is provided with its parametric value. Only clustering of crime data is made and the crime is not been split as per the crime ratio. The proposed system provides security for the crime data during outsourcing. Clustering and Classification is made on information. While classifying the data, the watermark content is used. The watermark content is used for verifying the classification data. Based on clustering and classification, the data can be classified and kept secure. Both Clustering and Classification is made on crime data. Data is secure by applying water mark content on data. The crime is been split as per the crime ratio.**

***Index Terms - Crime data, data mining, Clustering, knn, watermark*.**

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Objective

- Crime prevention and detection become an important trend in crime and a very challenging to solve crimes.
- The data used for analysis require the accuracy and sufficiency.
- This proposed system focuses on Traffic Violation and Border Control, Violent Crime, the Narcotics, Cyber Crime.
- Issues and challenges on crime are Data Collection and Integration, Crime Pattern, Performance, Visualization.
- As per the occurrence of the crime that is split into the ratio for detecting the most occurrence is happened is done in this crime pattern analysis.
- Crime occurrence in that particular area is taken as the input. The ratio of person occurrence in particular area is been calculated as the output.

II. LITERATURE REVIEW

AritThammanoetalstated that Data classification is one of the fundamental problems in data mining. Classification, as described, is a process of finding a model that describes and distinguishes data classes, for the purpose of being able to use the model to predict the class of objects which class label is unknown. There are many classification techniques that have been used thus far such as Decision tree, Neural networks, Support vector machines, and Bayesian networks. This paper focuses on a type of classification model that is based on K-means clustering algorithm. K-means is the most popular clustering algorithm. It is very efficient and very easy to implement. Besides being used as a clustering technique, K-means has also been adapted for data classification.

Two main problems of K-means algorithm are that (i) the number of clusters is needed to be specified before running the algorithm, and (ii) the quality of the resulting clusters depends heavily on the selection of initialcentroids. Many researchers, such as, have tried to overcome the above problems by introducing efficient methods for selecting the initial cluster centers.

Ying Zhaoetal describes fast and high-quality document clustering algorithms play an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. In particular, clustering algorithms that build meaningful hierarchies out of large document collections are ideal tools for their interactive visualization and exploration as they provide data-views that are consistent, predictable, and at different levels of granularity.

This paper focuses on document clustering algorithms that build such hierarchical solutions and (i) Presents a comprehensive study of partition and agglomerative algorithms that use different criterion functions and merging schemes. (ii) Presents a new class of clustering algorithms called constrained agglomerative algorithms, which combine features from both partition and agglomerative approaches that allows them to reduce the early-stage errors made by agglomerative methods and hence improve the quality of clustering solutions.

Chun-Nan Hsu et al investigated when component wise extrapolation should be preferred. To conclude that, when the Jacobian of the EM mapping is diagonal or block diagonal, CTJEM should be preferred. Show how to determine whether a Jacobian is diagonal or block diagonal and experimentally confirm our claim. In particular, Show that CTJEM is especially effective for the semi-supervised Bayesian classifier model given a highly sparse data set.

The Expectation-Maximization (EM) algorithm is one of the most popular algorithms for data mining from incomplete data. However, when applied to large data sets with a large proportion of missing data, the EM algorithm may converge slowly. The triple jump extrapolation method can effectively accelerate the EM algorithm by substantially reducing the number of iterations required for EM to converge. There are two options for the triple jump method, global extrapolation (TJEM) and component wise extrapolation (CTJEM).Tried these two methods for a variety of probabilistic models and found that in general, global extrapolation yields a better performance, but there are cases where component wise extrapolation yields very high speed up.

ShyamVaranNath et al applied the techniques to real crime data from a sheriff's office and validated our results. It also uses semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy. To developed a weighting scheme for attributes here to deal with limitations of various out of the box clustering tools and techniques. This easy to implement data mining framework works with the geospatial plot of crime and helps to improve the productivity of the detectives and other law enforcement officers. It can also be applied for counter terrorism for homeland security.

Michael Chauetal discussed an Efficient and effective access of criminal-justice data is critical for law enforcement personnel to perform investigations and fight crimes. Currently, most criminal-justice data are stored in structured relational databases, in which data are represented as tables consisting of various fields, such as the attributes of a suspect, the address of a crime scene, and so on. Detectives and crime investigators can search for useful information in such databases by providing specific search queries.

MeghaMudholkar et al described most aspect of life move to digital networks, crimes comes with them. Our lives increasingly depend on the internet and digital networks, but these create new vulnerabilities and new ways for criminals to exploits the digital environment. Not only can many existing crimes be replicated in online environments, but novel crimes that exploit specific features of digital networks have emerged as well. With new crimes come new forms of policing and new forms of surveillance and with these come new dangers for civil liberties.

These kinds of cybercrime information present on web pages are in the form of text. Because a lot of crime information in documents is described through events, event-based semantic technology can be used to study the patterns and trends of web-oriented crimes. So for cyber crime mining, event ontology is constructed to extract the attributes and relations in web pages and reconstruct the scenario for crime mining.

III. SYSTEM METHODOLOGY

A. Problem Definition

The data mining is data analyzing techniques that used to analyze crime data previously stored from various sources to find patterns and trends in crimes. In additional, it can be applied to increase efficiency in solving the crimes faster and also can be applied to automatically notify the crimes.

Crime prevention and detection become an important trend in crime and a very challenging to solve crimes.
Several studies have discovered various techniques to solve the crimes that used too many applications. Such studies can help speed up the process of solving crime and help the computerized systems detect the criminals automatically.

The crime data previously stored from various sources have a tendency to increase steadily. As a consequence, the management and analysis with huge data are very difficult and complex. To solve the problems previously mentioned, data mining techniques employ many learning algorithms to extract hidden knowledge from huge volume of data.

Data mining is data analyzing techniques to find patterns and trends in crimes. It can help solve the crimes more speedily and also can help alert the criminal detection automatically. The proposed system gives the brief reviews of researches on various implementations of data mining and the guidelines to solve the crimes by using data mining techniques. It also discusses research gaps and challenges in the area of crime data mining.

B. System Model

The system performs clustering on the crime database. The system uses k-means clustering method to form cluster on crime data. In this system, the values are categorized and denoted by the label to identify the cluster. As the crime type varies with the crime value, each crime is separated from another one. To identify crime falls under category, each crime data is provided with its parametric value.

The current way of communication can be affected by attackers in the middle. The data being sent by the sender to the receiver can be tracked by anyone using an application or trapping in between the nodes in the network. To avoid such drawbacks, cryptography is required to encrypt the content.

C. Crime Dataset

(i) Add Crime Profile

In this module contains crime details such as person id, name, address, mobile number, gender, DOB and occupation. It also includes passport number, a adhere number, case details for unique identification of a person who is in crime. These details are stored in 'crime' table.

(ii) Crime Observation Details

In this module contains information about crime observation transaction details. It includes crime person id, name, entry date and three parameter value which describe crime details such as type of crime, severity and punishment. These details stored in 'Crime Observation' table and viewed by using data grid view control.

Authentication and key distribution are the two main problem associated with the communication between two systems over a TCP/IP communication. A protocol is an agreed-upon sequence of actions performed by two or more principles. Cryptographic protocols make use of cryptography to accomplish some task securely.

(iii)Embed Watermark Data in Crime Pattern

In this module, the watermark content details are converted into bytes and stored in crime observations third column along with a numeric value 301 is added. The first observation values are taken as X position and second observation values are taken as Y position and are pointed initially. Then the first watermark byte value is added with X and then the X Position is modified. This repeats for all watermark bytes (each one watermark byte is stored in one crime all observation data). The process is listed in embed steps of algorithm 1 of algorithms section.

(iv) Extract Watermark Data in Crime Pattern

In this module, the modified crime data set is taken and record values for third observation column values greater than 301 is filtered out. Then for each crime, the third observation column value is subtracted with 301 and the numeric value's character is found out (ascii value). This repeats for all the crime and watermark content is appended. The result watermark content is displayed. The KNN value is found out before and after watermarking and checks for same result display. The process is listed in extract steps of algorithm 1 of algorithms section.

D. K-Nearest Neighbors Algorithm

The k-Nearest Neighbors algorithm (k-NN) is a non-parametric technique utilized for classification and regression. In both cases, the input consists of the k closest training exemplar in the characteristic space. The output Depends on whether k-NN is employed for classification or regression:

- In k-NN classification, the output is a class label. An object is classified by a greater part of vote of its neighbors, with the object being dispensed to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In k-NN regression, the result is the property value for the object. This value is the average of the values of its k nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

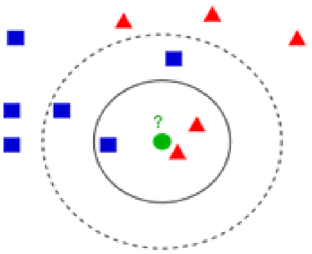

Fig.1 K-NN

For example, a common weighting system consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is requisite.

K-NN Algorithm

```
Input: Crime Data, Watermark Data
Output: Modified Crime Observation Data

Add the Crime Profiles (P).
Add the Crime Observation Data (O).
Enter watermark content (W).
Convert the watermark data to bytes and find the length of watermark data (L).
Sort the Crime Observation Data (O) Crime wise.
I=0
For Each Crime's Observation Set in (O)
Alter the Observation Data's third value such that OD(3) = 301 + W(I)
Change the OD(1) position = OD(1) position + W(I)
        I=I+1
If I>=L Then
    Break
  End If
Next
Output the New Crime Data Set.
```

IV. RESULTS AND DISCUSSION

The following Table describes experimental result for proposed system analysis in patient observation data set. The table contains patient id and patient data set surrounded data point details are shown. The data pint is called as a class label. The experimental work analysis the class label surrounded in pervious algorithm work.

S. No.	Crime Id.	Class Label
1	00001	A
2	00002	B
3	00003	C
4	00004	A
5	00005	B
6	00006	C
7	00007	A
8	00008	B
9	00009	C
10	000010	A
11	000011	B
12	000012	C
13	000013	A
14	000014	B
15	000016	C
16	000017	A
17	000018	B
18	000019	C
19	000020	A
20	000021	A

Table 1. Crime Observation Data set

The Fig describing the crime dataset analysis in KNN and KNN with water marking algorithm process. The figure contains data point X position value Y position values and its crime data set. The comparison chart is effective water marking process compare to proposed system.

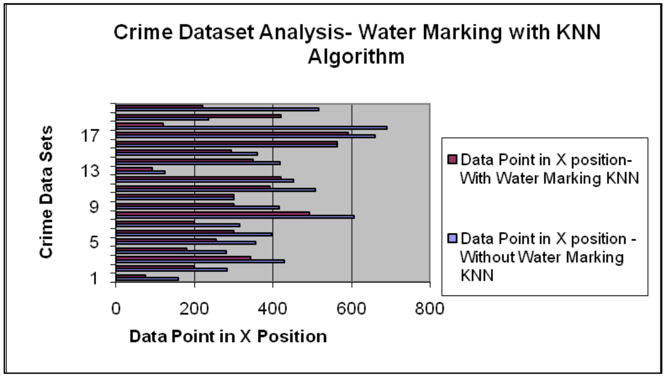

Fig 2. Crime Data Analysis of KNN without Watermarking

V. CONCLUSION

Crime are characterized which change over time and increase continuously. The changing and increasing of crime lead to the issues of understanding the crime behavior, crime predicting, precise detection, and managing large volumes of data obtained from various sources. Research interests have tried to solve these issues. In the crime investigation procedures, input data is very essential to use in training process and testing process. The training process is used to accomplish the crime model and the testing process is used to validate the algorithm. The issues of crime pattern are concerning with finding and predicting the hidden crime.

The proposed methodology provides security for the crime data during outsourcing. Clustering and classification is made on the crime information. While classifying the crime data, watermark content is added for the purpose of defense. The watermark content is used for verifying the classification data. Based on clustering and classification, the data can be classified and kept secured manner. Also the crime data is been split as per the crime ratio.

Scope for Future Enhancements

The proposed algorithm represents the feasible approach to using hubness for improving high-dimensional crime data clustering. And also have it in mind to explore other closely related research directions, including kernel mappings and shared-neighbor clustering for the crime dataset. This would allow us to overcome the foremost drawback of the proposed method detecting only hyper spherical clusters, just as K-Means. In addition, explore methods for using hubs to automatically resolve the number of clusters in the crime data.

Discussion

The following Table describes experimental result for proposed system analysis in patient observation data set. The table contains patient id and patient data set surrounded data point details are shown. The data pint is called as a class label. The experimental work analysis the class label surrounded in pervious algorithm work.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, second ed. Morgan Kaufmann, 2006.

[2] C.C. Aggarwal and P.S. Yu, “Finding Generalized Projected Clusters in High Dimensional Spaces,” Proc. 26th ACM SIGMOD Int’l Conf. Management of Data, pp. 70-81, 2000.

[3] K. Kailing, H.-P.Kriegel, P. Kro ¨ger, and S. Wanka, “Ranking Interesting Subspaces for Clustering High Dimensional Data,” Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003.

[4] K. Kailing, H.-P. Kriegel, and P. Kro ¨ger, “Density-Connected Subspace Clustering for High-Dimensional Data,” Proc. Fourth SIAM Int’l Conf. Data Mining (SDM), pp. 246-257, 2004.

[5] E. Mu ¨ ller, S. Gu ¨nnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data,” Proc. VLDB Endowment, vol. 2, pp. 1270-1281, 2009.

[6] E. Agirre, D. Martı´nez, O.L. de Lacalle, and A. Soroa, “Two Graph-Based Algorithms for State-of-the-Art WSD,”Proc. Conf.Empirical Methods in Natural Language Processing (EMNLP),pp. 585-593, 2006.

[7] K. Ning, H. Ng, S. Srihari, H. Leong, and A. Nesvizhskii,“Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology, ”BMC Bioinformatics,vol. 11, pp. 1-14, 2010.

[8] D. Arthur and S. Vassilvitskii, “K-Means++: The Advantages of Careful Seeding,”Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA),pp. 1027-1035, 2007.

[9] I.S. Dhillon, Y. Guan, and B. Kulis, “Kernel k-Means: Spectral Clustering and Normalized Cuts,”Proc. 10th ACM SIGKDD Int’lConf. Knowledge Discovery and Data Mining,pp. 551-556, 2004.