APTIKOM Journal on Computer Science and Information Technologies
Vol. 2, No. 1, 2017, pp. 8~11
ISSN: 2528-2417
DOI: 10.11591/APTIKOM.J.CSIT.42

Detection of Crimes Using Unsupervised Learning Techniques

R. Buli Babu*[1], G. Snehal[2], P. Aditya Satya Kiran[3]
Department of Electronics and Computer Engineering, KLEF University, India
*Corresponding author, e-mail: babuklu123@kluniversity.in

Abstract

Data mining can be used to detect model crime problems. This paper is about the importance of data mining about its techniques and how we can easily solve the crime. Crime data will be stored in criminal's database. To analyze the data easily we have data mining technique that is clustering. Clustering is a method to group identical characteristics in which the similarity is maximized or minimized. In clustering techniques also we have different type of algorithm, but in this paper we are using the k-means algorithm and expectation-maximization algorithm. We are using these techniques because these two techniques come under the partition algorithm. Partition algorithm is one of the best methods to solve crimes and to find the similar data and group it. K-means algorithm is used to partition the grouped object based on their means. Expectation-maximization algorithm is the extension of k-means algorithm here we partition the data based on their parameters.

Keywords*: clustering, data mining, k-means, expectation-maximization*

1. Introduction

Data collection and storing that data during the past decades is difficult and referring it for new crime is also difficult as we have to refer all the crimes from the starting which crime can be similar to it. Document analysis can be difficult to solve a crime. If there are more documents to solve the crimes we have to study more documents and refer to them and understand them is difficult. To solve this problem we use computer forensics we can solve the crimes fast and it is fast growing field where it can easily be examine the evidence. In case if there is any damage to the computer we can recover the stored data there will be no loss of information. By using this digital content it is easy to solve the crime. Even this data can be hidden where others can't see this data. This can be done by using the user name and password. By this only the officers related to that investigation can see the data. We can also know identify that who logged in to the criminal data. Because of this it reduces the manual effort, time, redundancy and to solve the crimes easily. By using the digital device we can store large amount of data. There are some methods already presented by different researchers to analyze the multiple documents. Existing methods is DFI propose multi-level search approach, it gives the accurate results and also produce the evidence related to the current investigation.

The drawback of these methods has no provision for end user. Here the end user has to search the data relevant to that task or group the data on given subject. The DIF system takes the input as text file in unstructured format. This data will be converted to the structured form by using different data mining techniques. There are many clustering algorithms to group the relevant data can be used to analysis the crime. Such methods are used for the data analysis which has less or no prior information about the input data. All digital data produce applications of end results. Data sets are made up of unlabeled sets or classes of data which is identified as unknown initially. Even if we consider the availability of labeled dataset there is no certainty that classes that are available of labeled dataset in input dataset or next raw dataset which is being collected through different computers or related to different investigations. The unstructured data sample can be of different sources. To provide an efficient solution for such heterogeneous data, we use clustering techniques. These clustering techniques are used to find the related document by using patterns. This algorithm improves the performance by end users. The method of this cluster is to group the data related to each other with some similarities which we define it as cluster. Similarly we have different type of clusters grouping with some similarities.

Received December 3, 2016; Revised February 3, 2017; Accepted February 14, 2017

The investigators can easily find the related document from which cluster they required. By this we can scrutinize other documents with each cluster. By this we can easily solve the difficult tasks by analyzing the documents easily and it also saves the time compared to earlier. In the recent investigation we have studied the work done on different clustering algorithms such as k-means, single link, complete link, average link with different digital forensic datasets in [1]. In [1], author presented the methodology for the clustering algorithms which were used for the forensic analysis of data/evidence in the criminal cases are being investigated by detectives. In this paper we are analyzing the partition algorithm for the criminal data. From the studies there are different types of clustering methods. They are partitioning methods, grid-based method, hierarchical method, density-based method and model-based methods. This research paper is based on partition clustering.

2. Related Work

In this section we went through the different types of document clustering in crime data. Crime data describes the use of different clustering algorithms. Most describes the study of the use of classical algorithms for clustering data. For example., Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-Means (FCM) and Self-Organizing Maps (SOM). These algorithms are widely used and have well known properties. In [1] document clustering algorithms are implemented by using various datasets. Manish Gupta et. al. [2]. highlights the existing systems which are used by Indian police as e-governance initiatives and also here the author proposes an interactive query based interface as crime analysis tool to assist police to complete their activities. Data storage, networking and security are the important point which plays a major role in investigation [3]. A. Malathi et al.[5] proposed system by use of missing value and clustering algorithm to predict the crime patterns and by this fast up the process of solving crime for data mining approach.]. Kadhim B. Swadi Al-Janabi [8] presents a proposed system for the crime analysis to detect the data using simple K-Means algorithm for data clustering and decision tree algorithms for data classification. K. Zakir Hussain et al. [11] captured the human experience for years into data mining and to design a simulation model. J. Han and M. Kamber [12] widespread concepts of data mining and tools which are required how to manipulate data are described detailed. Fault prediction using quad tree and Expectation Maximization clustering algorithm, limits the research in this book to the section of ―Cluster Analysis. This book describes different type of clustering methods in the cluster analysis section. In [13], mixture models chapter in explained in detail and Expectation-Maximization Algorithm are related to the EM introduces these concepts. M. Steinbach, G. Karypis, and V. Kumar [14] clustering techniques document are compared and discussed.

3. Data Mining and Crime Patterns

We have seen that in crime investigation we use clustering to group the data. Clusters or a subset of clusters have correspondence of one to one to crime patterns. These algorithms in data mining are used to identify the similar records which are different from other data. The detection of crime, allows the detectives to concentrate on crime sprees and solve one of the crimes results to solve the whole spree. In some cases if the group's of incidents are suspected to be one spree, the complete data can be built from different bits of information from each of the crime investigations. For example, one crime investigation reveals that the suspect has brown hair, the witness reveals that suspect is middle aged and next reveals that there is a mole on his/her face all these together can give much complete picture because all these evidences can be matched to the any of the cluster where we can easily solve the crime. But by using these evidences individually we can have many documents where it takes more time.

By grouping all these there will be some documents compared to the individual going through the evidence. By this we can easily solve the crimes by using less time. Instead of clustering we can also classification algorithm. But this is not that much predictable quality for future crimes when compared to the clustering. Classification algorithm is applicable to the existing and solved crimes. But we may have unsolved crimes these also can be solved by using the clustering. Clustering is grouped based on the first objects. But in classification it is dependent on the particular object. Now, a day we have different type of crimes like theft, murder, Traffic Violations, Fraud, Drug Offences, Cyber Crime. Previously they used paper to file a crime but now, a day they are using the computers to file a crime and recording the data grouping that data by using the different techniques. By the study clustering is the best technique to solve the crimes in a simple way.

4. Approach used by Clustering Algorithms

4.1. Types of Clustering

Clustering is a unsupervised task without having a priori knowledge by discovering groups of similar documents. There are two types of categories in clustering algorithms; they are the partitional and the hierarchical. K-Means algorithm and the link clustering they come under these two categories. K-Means and hierarchical clustering have many comparisons. In hierarchical clustering the size of data increases as the computational expansive, since to merge small clusters and D_D similarity matrix by using the certain link functions. By comparing with them K-Means is faster. It updates the centroid clusters with each iteration and reallocates each document by its nearest centroid by this we can say that it is an iterative algorithm. Comparison of K-Means and hierarchical algorithm can be found in [15].

4.2. K-Means Algorithm

K-Means clustering analysis aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Algorithm, for partitioning the K-Means algorithm, where the mean value the objects in the cluster is represented by each cluster.

Input: Number of groups.

1. Place K focuses into the space represented by the data that are being grouped. These focuses
2. Represent the group centroids initially.
3. Assign every data to the cluster that has the nearest centroid.
4. Recalculate the positions of the K centroids, when the objects have been assigned.
5. Rehash Steps 2 and 3 until the centroids have no more move. This produces a partition of the data into gatherings from which the metric to be minimized can be calculated.

Output: An arrangement of k groups. K-Means calculation is a base for all other grouping calculations to locate the mean qualities. The K-Means algorithm does not find the corresponding to the global objective function minimum, to find the most optimal configuration. We can reduce the effect by running the K-Means for multiple times

4.3. Expectation-Maximization Algorithm

Expectation-Maximization is a type of model based clustering method. Expectation-Maximization calculation is an expansion of K-means calculation which can be utilized to discover the parameter gauges for every cluster.

Algorithm

1. Find the initial input by finding the initial centroid.
2. By using the cosine distance formula or any other distance formula calculate the distance between each data point and each centroid.
3. Based on the probability of membership of a data pint to a particular cluster assign the weights for each combination of data point and cluster
4. Repeat:

a) Which has highest weight reassign each data point to the cluster i.e., highest probability.
b) if a data point belongs to more than one cluster with the same probability, then (re)assign the data point to the cluster based on minimum distance.
c) Update the cluster means for every iteration until clustering converges.

5. Result

In the original data there are groups which are not exactly clustered, some data is not grouped here. In K-Means clustering the data is grouped based on the algorithm, here it finds the nearest centroid and groups the data according to the nearest centroid as shown in Figure 1. In Expectation-Maximization is the extension of K-Means here the data is expected and it is grouped to the exactly nearest cluster

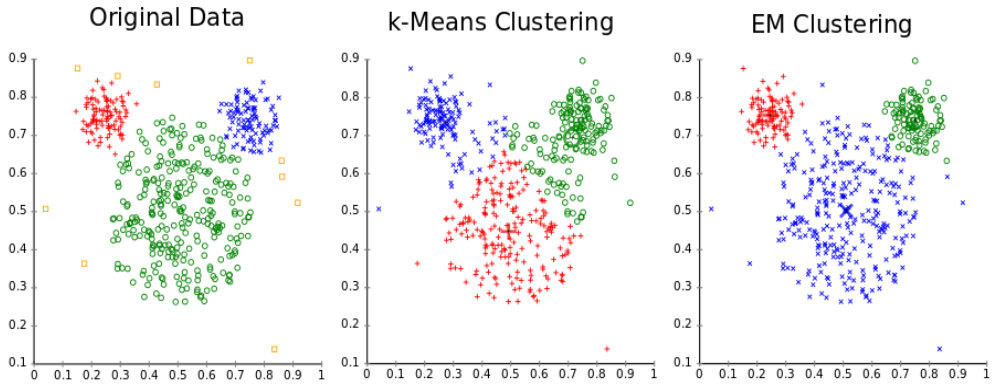

Figure 1. Comparison Of The Original Data With K-Means Clustering And EM Clustering Results

6. Conclusion

In crime data clustering techniques plays a vital role to investigate the crime and it helps for solving the unsolved crimes easily. By grouping the data with similar objects we can easily solve the unsolved crimes. For finding similarity objects partitioning clustering algorithm is one of the best method. It is observed that finding similar words and collect them in a single cluster which helps in crime analysis. This paper deals with the study of clustering techniques and similarity measures in crime data.

References

[1] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection‖, *IEEE Transactions On Information Forensics And Security*, Vol. 8, No. 1, January 2013.
[2] Manish Gupta1, B.Chandra1 and M. P. Gupta1, 2007 Crime Data Mining for Indian Police Information System.
[3] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, The expanding digital universce: A forecast of worldwide information growth through 2010,‖ *Inf. Data*, vol. 1, pp. 1–21, 2007.
[4] Faith Ozgul, Claus Atzenbeck, AhmetCelik, Zeki, Erdem, *Incorporating data Sources and Methodologies for Crime Data Mining*, IEEE proceedings, 2011.
[5] A.Malathi, Dr.S.Santhosh Baboo. D.G. Vaishnav College, Chennai, 2011 Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters.
[6] J. Mena, "Investigative Data Mining for Security and Criminal Detection", Butterworth Heineman Press, pp. 15-16, 2003
[7] Hao Cheng, Kien A. Hua and Khanh Vu, *Constrained Locally Weighted Clustering*, Journal proceedings of the VLDB Endowment, vol. 1, no.2, 2008
[8] Kadhim B.Swadi al-Janabi. Department of Computer Science. Faculty of Mathematics and Computer Science. University of Kufa/Iraq, 2011 A Proposed Framework for Analyzing Crime Data Set using Decision Tree and Simple K-means Mining Algorithms.
[9] C M Bishop, ―Pattern Recognition and Machine Learning‖ NewYork Springer-Verlag 2006
[10] Kilian Stoffel, Paul Cotofrei and Dong Han, Fuzzy Methods for Forensic Data Analysis, *European Journal of Scientific Research*, Vol.52 No.4, 2011.
[11] K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana, 2012 Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model
[12] Han, Kamber, Pei, ‖Data Mining: Concepts and Techniques‖, MK Third Edition
[13] C M Bishop, Pattern Recognition and Machine Learning‖ NewYork Springer-Verlag 2006
[14] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques‖. Technical Report 00-034, University of Minnesota, 2000