ISSN: 2321-7782 (Online)
Volume 2, Issue 8, August 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Survey on Crime Data Analysis of Data Mining Using Clustering Techniques

Dr. A.Bharathi[1]
Professor
Dept. of IT
Bannari Amman Institute of Technology
Sathyamangalam – India

R. Shilpa[2]
Research Scholar
Dept. of IT
Bannari Amman Institute of Technology
Sathyamangalam – India

Abstract: ***Data mining is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data and it is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining plays an important role in terms of prediction and analysis. Crimes are a social nuisance and cost our society dearly in several ways. Crime investigation has very significant role of police system in any country. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. This paper presents detailed study on clustering techniques and its role on crime applications. This study also helps crime branch for better prediction and classification of crimes.***

Keywords: Data mining, Clustering, Crime data analysis, Prediction, Classification.

I. INTRODUCTION

Crime is one of the dangerous factors for any country. Crime analysis is the activity in which analysis is done on crime activities. Today criminals have maximum use of all modern technologies and hi-tech methods in committing crimes. It is impossible to find a country which has a crime- free society. As long as human beings have feelings they incline on attempting crimes. So the present society has also filled with various kinds of crimes.

Hence, creation of data base for crimes and criminals is needed. Developing a good crime analysis tool to identify crime patterns quickly and efficiently for future crime pattern detection is challenging field for researchers.

Data mining techniques have higher influence in the fields such as, Law and Enforcement for crime problems, crime data analysis, criminal career analysis, bank frauds and other critical problems. In recent years, data clustering techniques have faced several new challenges including simultaneous feature subset selection, large scale data clustering and semi-supervised clustering. Mostly, cluster analysis is an important human activity which indulge from childhood when learn to distinguish between animals and plants, etc by continuously improving subconscious clustering schemes. It is widely used in numerous applications including pattern recognition, data analysis, image processing, and market research etc [16].

Recent researches on these techniques link the gap between clustering theory and practice of using clustering methods on crime applications [17].

Cluster accuracy can be improved to capture the local correlation structure by associating each cluster with the combination of the dimensions as independent weighting vector and subspace span which is embedded on it [14,15].

A web page involving a crime can be thought of as a chain of actions with series of background attributes. We can analyze web information from the perspective of events and apply some research results related to the events to solve the problem of web crime mining.

Duplicates (or matches) refer to applications which share common values. There are two types of duplicates: exact (or identical) duplicates have the all same values; near (or approximate) duplicates have some same values(or characters), some similar values with slightly altered spellings, or both. This paper argues that each successful credit application fraud pattern is represented by a sudden and sharp spike in duplicates within a short time, relative to the established baseline level. Duplicates are hard to avoid from fraudsters' point-of view because duplicates increase their' success rate. The synthetic identity fraudster has low success rate, and is likely to reuse fictitious identities which have been successful before. The identity thief has limited time because innocent people can discover the fraud early and take action, and will quickly use the same real identities at different places.

It will be shown later in this paper that many fraudsters operate this way with these applications and that their characteristic pattern of behavior can be detected by the methods reported. In short, the new methods are based on white-listing and detecting spikes of similar applications. White-listing uses real social relationships on a fixed set of attributes. This reduces false positives by lowering some suspicion scores. Detecting spikes in duplicates, on a variable set of attributes, increases true positives by adjusting suspicion scores appropriately.

The organization of the paper is as follows. Section 2 describes data mining on crime domain. Crime Reporting Systems discussed in Section 3. Overview of crime data mining is presented in section 4. Clustering methods have discussed in section 5 and section 6 concludes the paper.

II. Data Mining on Crime Domain

Recent developments in crime control applications aim at adopting data mining techniques to aid the process of crime investigation. COPLINK is one of the earlier projects which is collaborated with Arizona University and the police department to extract entities from police narrative records [9]. Bruin, Cocx and Koster et al. presented a tool for changing in offender behavior. Extracted factors including frequency, seriousness, duration and nature have been used to compare the similarity between pairs of criminals by a new distance measure and cluster the data accordingly [2].

Brown proposed a framework for regional crime analysis program (ReCAP) [1]. The data mining was adopted as an algorithm for crime data analysis. J.S.de Bruin et.al compared all individuals based on their profiles to analyze and identify criminals and criminal behaviors [2]. Nath et.al used K-means clustering to detect crime pattern to speed up the process of solving crimes [5]. Adderley and Musgrove applied Self Organizing Map (SOM) to link the offenders of serious sexual attacks [11]. Recently, Ozgul et.al proposed a novel prediction model CPM (Crime Prediction Model) to predict perpetuators of unsolved terrorist events on attributes of crime information that are location, date and modus operandi attributes [7]. LianhangMa, Yefang Chen, and Hao Huang et.al presented a two phase clustering algorithm called AK-modes to automatically find similar case subsets from large datasets [13]. In the attribute-weighing phase, the weight of each attribute related to an offender's behavior trait using the concept Information Gain Ratio (IGR) in classification domain is computed. The result of attribute weighing phase is utilized in the clustering process to find similar case subsets.

III. Crime Reporting Systems

The data for crime often presents an interesting dilemma. While some data is kept confidential, some becomes public information. Data about the prisoners can often be viewed in the county or sheriff's sites. However data about crimes related to narcotics or juvenile cases is usually more restricted. Similarly, the information about the sex offenders is made public to warn others in the area, but the identity of the victim is often prevented.Thus as a data miner, the analyst has to deal with all these public versus private data issues so that data mining modeling process does not infringe on these legal boundaries.

Most sheriffs' office and police departments use electronic systems for crime reporting that have replaced the traditional paper-based crime reports. These crime reports have the following kinds of information categories namely - type of crime,

date/time, location etc. Then there is information about the suspect (identified or unidentified), victim and the witness. Additionally, there is the narrative or description of the crime and Modus Operandi (MO) that is usually in the text form.

The police officers or detectives use free text to record most of their observations that cannot be included in checkbox kind of pre-determined questions. While the first two categories of information are usually stored in the computer databases as numeric, character or date fields of table, the last one is often stored as free text. The challenge in data mining crime data often comes from the free text field. While free text fields can give the newspaper columnist, a great story line, converting them into data mining attributes is not always an easy job. We will look at how to arrive at the significant attributes for the data mining models.

IV. Overview of Crime Data Mining

Data Mining: Data mining deals with the discovery of unexpected patterns and new rules that are "hidden" in large databases. The use of data mining in this paper is to give the structured data from unstructured data of judge. In this paper the Data Mining techniques of crime in two directions they are

1. Classification of Crime
2. Clustering Technique of Crime

1. Classification of Crime

Crime: Crime is defined as "an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law". Crime is referred to as a comprehensive concept that is defined in both legal and non-legal sense.

Classification of Crime

- Traffic Violations
- Sex Crime
- Fraud
- Arson
- Drug Offences
- Violent Crime
- Cyber Crime

Traffic Violations: Driving under the influence of alcohol, fatal / personal injury / property damage traffic accident, road rage

Sex Crime: Sexual offences

Fraud: Forgery and counterfeiting, frauds, embezzlement, identity deception

Arson: Arson on buildings

Drug Offences: Narcotic drug offences (sales or possession)

Violent Crime: Criminal Homicide, armed robbery, aggravated assault, other assaults

Cyber Crime: Internet frauds, illegal trading, network intrusion / hacking, virus spreading, hate crimes, cyber piracy, cyber pornography, cyber-terrorism, theft of confidential information.

2. Clustering Technique of Crime

Clustering: Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Clustering can also be considered the most important unsupervised learning technique; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. There are so many techniques used in clustering, in this paper discuss about K-means algorithm, Ak-Mode Algorithm, Expectation-Maximization Algorithm is used.

V. CLUSTERING METHODS FOR CRIME DOMAIN

The partition clustering methods primarily classified into K-means, AK-mode and Expectation-Maximization algorithms. The partitioning method constructs 'k' partitions of the data from a given dataset of 'n' objects. After preprocessing, the operational data are undergoing the clustering techniques for grouping objects as different clusters.

5.1 K-Means Clustering Algorithm

K-means algorithm mainly used to partition the clusters based on their means. Initially number of objects are grouped and specified as k clusters. The mean value is calculated as the mean distance between the objects. The relocation iterative technique which is used to improve the partitions by moving objects from one group to other. Then number of iterations is done until the convergence occurs.

K-means algorithm steps are given as

Input: Number of clusters.

Step1: Arbitrarily choose k objects from a dataset D of N objects as the initial cluster centers.

Step 2: reassign each object which distributed to a cluster based on a cluster center which it is the most similar or the nearer.

Step 3: Update the cluster means, i.e. calculate the mean value of the object for each cluster.

Output: A set of k clusters.

K-means algorithm is a base for all other clustering algorithms to find the mean values.

5.2 Ak-Mode Algorithm

Ak- mode clustering algorithm is a two step process such as attribute weighing phase and clustering phase. In the attribute weighing phase, weights of the attributes are computed using Information Gain Ratio (IGR) value for each attribute. The greatest value of weight is taken as decisive attribute. The distance between two categorical attributes is computed as the difference between two data records gives the similarity measures. The analyst has set the threshold value α with the help of the computation result of similarity measures. This algorithm is mainly used for categorical attributes.

Ak-mode algorithm steps are as follows:

Input: Data set, weighted attributes and threshold value.

Output: cluster result

Step1: Find the number of clusters k and find initial mode of every cluster.

Step2: Calculate the distance for every mode and its closest mode.

Step3: update each cluster mode.

Step4: this process terminates when all the modes do not change. Else go to step 2.

Ak-mode algorithm has been used to find the similar subsets automatically from large datasets and mainly applied for categorical attributes.

5.3 Expectation-Maximization Algorithm

Expectation- Maximization algorithm is an extension of K-means algorithm which can be used to find the parameter estimates for each cluster. The entire data is a mixture of parametric probabilistic distribution. The weight of attributes is measured in the probability distribution and each object is to be clustered based on the weights instead of assign the objects to the dedicated clusters in K-means. To find parameter estimates, the two steps of iterative refinement algorithm are used.

Step1: Expectation step:

For each object of clusters, this step calculates the probability of cluster membership of object xi.

Step2: Maximization step:

Re-estimate or refine the model parameters using probability estimation from

step1.

This EM algorithm is easy to implement and it converges fast in practice.

VI. Conclusion and Future Works

Crime data is a sensitive domain where efficient clustering techniques play vital role for crime analysts and law-enforcers to precede the case in the investigation and help solving unsolved crimes faster. Similarity measures are an important factor which helps to find unsolved crimes in crime pattern. Partition clustering algorithm is one of the best method for finding similarity measures. This paper deals detailed study about importance of clustering and similarity measures in crime domain.

References

1. D.E. Brown, "The regional crime analysis program (RECAP): A Frame work for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol.3, pp.2848-2853, 1998.
2. J.S. de Bruin, T.K. Cocx, W.A. Kosters, J. Laros and J.N. Kok, “Data mining approaches to criminal career analysis,” in Proceedings of the Sixth International Conference on Data Mining (ICDM’06), pp.171-177, 2006.
3. J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann publications, pp. 1-39, 2006.
4. J. Mena, “Investigative Data Mining for Security and Criminal Detection”, Butterworth Heinemann Press, pp.15-16, 2003.
5. S.V. Nath, “Crime pattern detection using data mining,” in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 41-44,2006.
6. Brown, D.E and Hagen, S., “Data association methods with application to law enforcement.” Decision Support Systems, 34, 2000, 369-378.
7. Faith Ozgul, Claus Atzenbeck, AhmetCelik, Zeki, Erdem, “Incorporating data Sources and Methodologies for Crime Data Mining,” IEEE proceedings, 2011.
8. H.Chen, W.ChungXu, G.Wang, Y.Qin and M.Chen, “Crime Data Mining: A General Framework and Some Examples,” computer, vol.37, 2004.
9. H.Chen, W.Chung, Y.M.Chan, J.Xu, G.ang, R.Zheng and H. Atabakch,” Crime Data Mining: An Overview and Case Studies,” in proceedings of the annual national conference on digital government research, Boston, pp.1-5, 2003.
10. M.Chan, J.Xu,and H.Chen,”Extracting Meaningful Entities from Police Narrative Reports,” in proceedings of the National Conference on Digital Government Research,pp.271-275,2002.
11. R.Adderly and P.B. Musgrove, “Data Mining Case Study: Modeling the behavior of offenders who commit sexual assaults,” in proceedings of the 2006 IEEE/WIC/ACM Conference on Web Intelligent Agent Technology, pp.41-44, 2006.
12. Z.Huang,”Extensions to the K-means algorithm for clustering large datasets with categorical values, “Data Mining and Knowledge Discovery, vol.2, pp.283-304, 1998.
13. L.Ma, Y.Chen, H.Huang, “AK-Modes: A weighted Clustering Algorithm for Finding Similar Case Subsets,” 2010.
14. Hao Cheng, Kien A. Hua and Khanh Vu, “Constrained Locally Weighted Clustering,” journal proceedings of the VLDB Endowment, vol . 1, no .2, 2008
15. Guenael cabanes and Younes bennani, “A Simultaneous Two-Level Clustering Algorithm for Automatic Model Selection,” ICMLA '07 Proceedings of the Sixth International Conference on Machine Learning and Applications, p p. 316-321, 2007
16. Kilian Stoffel, Paul Cotofrei and Dong Han, “Fuzzy Methods for Forensic Data Analysis,” European Journal of Scientific Research, Vol.52 No.4, 2011,
17. Anuska Ferligoj, “Recent developments in cluster analysis,” Telecommunication Systems, vol .1,issue 4, 205–220, 2003

