SURVEY OF DATA MINING TECHNIQUES ON CRIME DATA ANALYSIS

Revatthy Krishnamurthy,
Research Scholar in Computer Science,
Bharathiar University, Coimbatore
revatthykm11@yahoo.com

J. Satheesh Kumar,
Asst. Professor in Computer Applications,
Bharathiar University, Coimbatore
jsathee@rediffmail.com

Abstract

Data mining is a process of extracting knowledge from huge amount of data stored in databases, data warehouses and data repositories. Crime is an interesting application where data mining plays an important role in terms of prediction and analysis. Clustering is the process of combining data objects into groups. The data objects within the group are very similar and very dissimilar as well when compared to objects of 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.

Key words: Data mining, Crime data analysis, clustering.

Introduction

In recent years, volume of crime is becoming serious problems in many countries. In today's world, criminals have maximum use of all modern technologies and hi-tech methods in committing crimes [1]. The law enforcers have to effectively meet out challenges of crime control and maintenance of public order. 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].

The organization of the paper is as follows. Section II describes methods used in crime domain. Role of preprocessing in data mining is presented in section III. Clustering methods have discussed in section IV and section V 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. Role of data preprocessing

Data preprocessing techniques are mainly used for producing high-quality mining results. Raw data are being preprocessed before mining because data are in different format, collected from various sources and stored in the data bases and data warehouses. Major steps involved in data mining are data cleaning, data integration, data transformation and data reduction have shown in Figure 1.

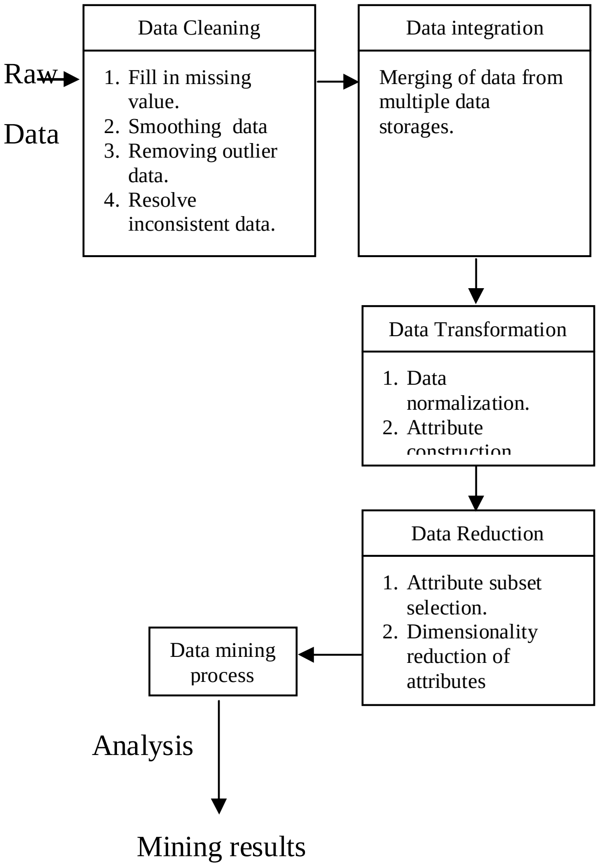
Raw
Data
Data Cleaning
1. Fill in missing value.
2. Smoothing data
3. Removing outlier data.
4. Resolve inconsistent data.
Data integration
Merging of data from multiple data storages.
Data Transformation
1. Data normalization.
2. Attribute construction
Data Reduction
1. Attribute subset selection.
2. Dimensionality reduction of attributes
Data mining process
Analysis
Mining results

Figure 1. Data preprocessing

In the data cleaning step, missing values will be filled, noisy data will be smoothened, outlier's data will be removed and inconsistent data are resolved. Data integration step undergoes merging of data. Data normalization and attribute construction are done in the data transformation for standardizing data. The data range falls under 0.0 to 1.0 during standardization. Attribute subsets are selected from large dataset and dimensionality has reduced. After preprocessing, finally standard data underwent the process of mining and hence better results have obtained.

IV. 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.

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.

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.

3. Expectation-Maximisation 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 $x_{i.}$

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.

V. Conclusion and future work

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 InternationaConference 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