RESEARCH ARTICLE OPEN ACCESS

A Survey on Crime Data Analysis Using Data Mining Techniques

*Gourav Govindaswamy[1], Vinod Kumar Kethineni[2,] Santhosh Kumar P[3]
[1,2]Department of Computer Science and Engineering, SCSVMV University, Kanchipuram, TamiInadu, India
Corresponding Author: Gourav Govindaswamy

ABSTRACT: In modern era, as the technology develops to a greater extent, the crime rates are increasing to a very vast level not only in India but also in major countries which possess a mighty threat to humans. So therefore these technologies can be used as an aid to reduce crime rates. An important concept called Data Mining can be used to cluster the data and analyze the different subsets of crime analysis. Clustering is the process of combining data objects into groups. This paper consists of important Data mining and clustering techniques and its role on crime applications.

***Keywords*:** Data Mining, Crime data analysis, Knowledge discovery process (KDD), Neural networks, Clustering.

Date of Submission: 13-07-2017 Date of acceptance: 31-08-2017

I. INTRODUCTION

The development of crime in recent years has increased to a greater extent and is becoming a serious problem in many countries. In today's world criminals and threat giving terrorist groups have maximum knowledge of all current hi-tech methods. Developing a good data mining tool for crime analysis would help the crime branch to form patterns and clusters efficiently and reduce crime rates.

The clustering technique is facing several challenges daily. Cluster analysis is an important activity for a human being. Some tough challenges such as clustering using simultaneous feature, large scale data clustering, etc. This clustering is also used in pattern recognition, image processing etc. Recent researches have been carried on clustering which links the gap between theory and clustering methods used crime applications [1].

The organization of the paper is as follows. Section II consists of some important Data Mining techniques. Section III consists of description about Knowledge discovery process. The clustering Methods are discussed in section IV. The methods Created and applied in crime domain are discussed in section V and the paper is concluded with section VI.

II. DATA MINING TECHNIQUES

A. Classification

Classification is a Data Mining technique which employs a set of pre-classified examples to develop a model to classify large records. Its main application is in fraud detection and credit card services. This approach frequently employs decision tree and neural network-based classification algorithms. The data classifications process involves learning and classification. In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to new data. Some of the classification models are listed below.

1. Neural Network

Neural network or an artificial neural network are a biological system that detects patterns and make predictions. The powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural networks are used in variety of applications are shown in fig.1. Artificial neural networks have become a powerful tool in tasks like pattern recognition, decision problem and predication application. So this can be used in our crime branch in predicting the next crime and even predict the plan of criminals and reduce the crime. This analytical network will also be useful in fraud detection and also in other management sectors where we can find the buying pattern of customers and sales and they can also find their interests to earn profit in the market. There are some other applications too.

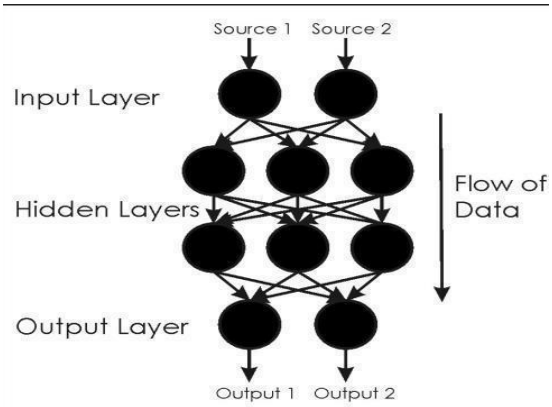

Fig 1. Neural Network with hidden layers

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network leans by adjusting weights so as to able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract pattern and detect trends that are too complex to be noticed by either humans or other computer techniques.

2. Decision Tress

A decision tress is a flow chart like structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distribution. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labeled with possible results of the test. The leaf node represents the cells and specifies the class to return if that lead node is reached. The classification of a specific input instance is thus performed by starting at the roof of a node, and depending on the results of the tests, following the appropriate branches until a leaf node is reached [2].

Decision tree is represented in figure 2.

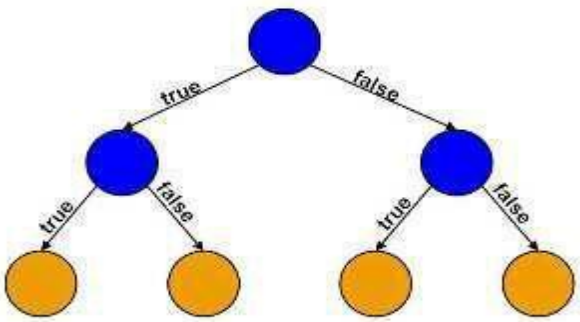

Fig. 2. Decision Tree

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

Partitioning methods

Hierarchical Agglomerative (divisive) methods Density based methods

Grid-based methods

Model-based methods.

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In Data Mining independent variables are attributed already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore more complex techniques may be necessary to forecast future values. The same model types often be used for both regression and classification, Neural networks too can create both classification and regression.

Types of association rule
Multilevel association rule
Multidimensional association rule
Quantitative association rule

D. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little value.

III. KNOWLEDGE DISCOVERYPROCESS (KDD)

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.

Knowledge discovery process plays a vital role in data mining as it says how to process the knowledge [3]. There are seven steps in KDD process. The first step is the data cleaning where missing values will be filled, noisy data will be smoothened, outlier's data will be removed and inconsistent data are resolved. Multiple data sourced, often heterogeneous are combined in a common source in the second step called data integration step. Third step is called as data selection step where the data relevant to the analysis is decided and retrieved from the data collection. Then the selected data is transformed into forms appropriate for the mining procedure in the fourth step called data transformation. Then the next step of data mining will undergo where clever techniques are applied to extract potentially useful pattern. Interesting patterns representing knowledge are identified based on given measures in the sixth step called pattern evaluation. Then the final step called as knowledge representation is the step where discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help us understand and interpret the data mining results.

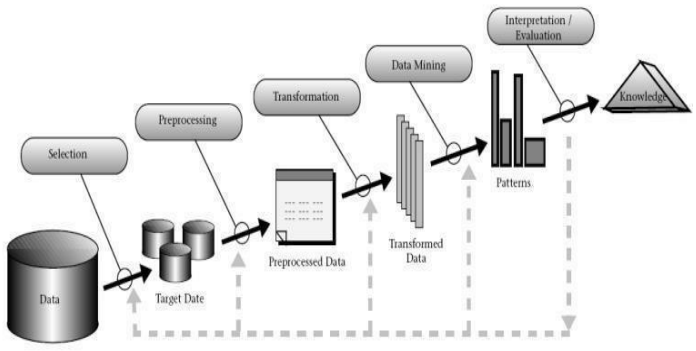

Fig 3. Knowledge discovery process

IV. CLUSTERING METHOD

Clustering methods are primarily classified into K-means, AK-mode and expectation maximization algorithms. The partitioning constructs 'k' partitions of data from a given dataset of 'n' objects.

1. K-means Clustering Algorithm

K-means algorithm partitions the clusters based on their means. The objects are grouped and specified as k clusters. The mean distance between the objects is calculated which is the mean value. To improve the partitions the object is moved from one group to another by relocation iterative technique.

Then until convergence occurs the iterations continue. The Algorithm steps are given as

Input: Number of clusters.
Step 1: 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 values 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 algorithm is a two step process such as attribute weighing phase and clustering phase, weights of the attributes are computed using Information Gain Ration (IGR) values for each attribute. The greatest value of weight is taken as decisive attribute. The distance between two categorical attributed is computed as the difference between two data records gives the similarity measures. The analyst has set the threshold values α with the help of the computation result of similarity measures. This algorithm is mainly used for categorical attributes. The steps are follows

Input: Data set, weighted attributes and threshold value.
Output: cluster result
Step 1: Find the number of clusters k and find the initial mode of every cluster
Step 2: Calculate the distance for every mode and its closest mode
Step 3: Update each cluster mode
Step 4: 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-Maximization algorithm

Expectation- Maximization algorithm is used as 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 step iterative refinement algorithms are used.

Step 1: Expectation step

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

Step 2: 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. METHODS CREATED AND APPLIED IN CRIME DOMAIN

Many recent developments are done in crime control applications to reduce crime rate. COPLINK is a project of Arizona university in collaboration with police department to extract the entities from police narrative records [4]. A tool was presented by Bruin, Cocx and Koster et al. for change in offender behavior. Extracted factors including frequency, seriousness, and duration have been used to compare the similarity between pairs of criminals by a new distance measure [5]. An analysis program called regional crime analysis program (ReCAP) was proposed by brown which adopted data mining algorithms [6].

J.S.de Bruin et al compared all individuals based on their profiles to analyze and identify criminals and criminal behavior [2]. Nath et.al used K-means clustering to detect crime pattern to speed up the process of solving crimes [7]. Adderly and Musgrove applied Self Organizing Map (SOM) to link the offenders of serious sexual attacks [8]. Ozgul et.al proposed a novel prediction model CPM (Crime Prediction Model) to predict perpetuators of unsolved terrorist events on attributed of crime information that are location, data etc [9]. LianhangMa, Yefang Chen, and also Hao Huang et.al presented a two phase clustering algorithm called AK-modes to automatically find similar case subsets from large datasets [10]. In 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.

VI. CONCLUSION

Crime data is a sensitive and large domain and therefore we need some efficient clustering techniques and algorithms which will help the crime analysts and law enforcers retrieve the data and information and draw patterns and conclude to a result which will help their investigation. The partition clustering algorithm can be developed in such a way that solves the unsolved crimes faster.

Partition clustering in very much helpful to draw patters and best method for finding similarity measures. This papers deals about the detailed study of crime data analysis using data mining and clustering and its important techniques. Presently in India the rape cases are going to a large extent. Many rapists are escaping giving false patterns and distractions to investigations. An artificial neural network called 'BLUE BRAIN' is carried out by many leading corporate and other companies which can think like and think more than humans. If the information from the network is extracted using proper tools, we can form appropriate patterns and we can capture all the rapists, criminals and other collar criminals using this technique. But this blue brain project is trying to make progress since 2005.

But it couldn't make even a prototype or model which they have promised by now. All they lag is algorithms. If perfect algorithms are used then it would lead to great success and the vision can be accomplished. So therefore my idea is to create a new algorithm for the blue brain project and provide them a steady progress.

REFERENCES

[1] Anuska Ferligo, "Recent developments in cluster analysis," Telecommunication Systems, Vol.1, issue 4, 205-220, 2003.

[2] Lior Rokach and Oded Maimon,"Data mining with Decision Tress: Theory and applications (Series in Machine perception and Artificial Intelligence)". ISBN: 981-2771- 791. World scientific publishing company, 2008

[3] Introduction to Data Mining and Knowledge Discovery, Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Faith Ozgul, Claus Atzenbeck, AhmetCelik, Zeki, Erdem, "Incorporating data Sources and Methodologies for Crime Data Mining," IEEE proceedings, 2011.

[10] L.Ma, Y.Chen, H.Huang, "AK-Modes: A weighted clustering Algorithm for Finding Similar Case Subsets," 2010

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

Gourav Govindaswamy. "A Survey on Crime Data Analysis Using Data Mining Techniques." International Journal of Engineering Research and Applications (IJERA), vol. 7, no. 8, 2017, pp. 30–34.