IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 9, September 2017

Survey of Crime Patterns Analysis Using Classification Data Mining Models

V. Vishnupriya, MC.A. [1], M. Valarmathi, M.C.A., M. Phil.[2],

Research Scholar, Computer Science, Vivekanandha College for Women, Tiruchengode, India[1]

Head of the Department, Computer Science, Vivekanandha College for Women, Tiruchengode, India[2]

Abstract: Crime prevention and detection become an important trend in crime and a very challenging to solve crimes. The crime data previously stored from various sources have a tendency to increase steadily. As a consequence, the supervision and analysis with huge data are very tricky and complex. To solve the problems, data mining techniques employ many learning algorithms to extort 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. Clustering is a data analyzing technique in unsupervised type. This technique is used to divide the same data into the same group and the different data into the other group. For the simple and effective clustering techniques, there are several algorithms such as K-means clustering. This approach is supervised learning scheme that used to dispense objects to one of many pre-determined categories. The algorithms of categorization have been widely applied to the numerous problems that include many various applications. Crime are characterized which change over time and increase continuously. The changing and increasing of crime direct to the issues of understanding the crime behavior, crime predicting, precise detection and managing large volumes of data obtained from various sources.

Keywords: Data Mining, Clustering, K-means clustering, Association Rule Mining.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from special perspectives and summarizing it into useful information - information that can be used to enlarge revenue, cuts costs or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to study data from many special dimensions or angles, classify it, and review the interaction identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in huge data sets relating methods at the connection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

The previous studies 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 assessment. 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 gives the brief reviews of investigates on various implementations of data mining and the strategies to resolve the crimes by using data mining techniques. It also discusses research gaps and confronts in the area of crime data mining. Crime prevention and detection become an imperative trend in crime and a very challenging to solve crimes. 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. The knowledge results obtained from data mining processes are used to assist in decision making and to solve the problems.

II. LITERATURE SURVEY

Arit Thammano [1] describes the most popular clustering algorithm because of its efficiency and superior performance. However, the performance of K-means algorithm depends heavily on the selection of initial centroids.

DOI10.17148/IJARCCE.2017.6912

This paper proposes an extension to the original K-means algorithm enabling it to solve classification problems. First, the entropy concept is employed to adapt the traditional K-means algorithm to be used as a classification technique. Then, to improve the performance of K-means algorithm, a new scheme to select the initial cluster centers is proposed. The proposed models are tested on seven benchmark data sets from the UCI machine learning repository. 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.

Ying zhao, George karypis [2] describe a 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, Han-Shen Huang , Bo-Hou Yang [3] describe 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 extraplolation yields a better performance, but there are cases where component wise extrapolation yields very high speed up. However, when applied to large data sets with a large number of parameters to estimate, the EM algorithm may converge slowly. If the data sets also contain a large proportion of missing data or there are a large number of hidden variables in the model, the convergence of EM can be even slower

Shyam Varan Nath [4] describe model crime detection problems. Crimes are a social nuisance and cost our society dearly in several ways. Any research that can help in solving crimes faster will pay for itself. About 10% of the criminals commit about 50% of the crimes. Here we look at use of clustering algorithm for a data mining approach to help detect the crimes patterns and speed up the process of solving crime. Take a k-means clustering with some enhancements to aid in the process of identification of crime patterns. To applied these 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.

Michael Chau, Jennifer J. Xu, Hsinchun Chen [5] describe a Valuable criminal-justice data in free texts such as police narrative reports are currently difficult to be accessed and used by intelligence investigators in crime analyses. It would be desirable to automatically identify from text reports meaningful entities, such as person names, addresses, narcotic drugs, or vehicle names to facilitate crime investigation. This paper, report our work on a neural network-based entity extractor, which applies named-entity extraction techniques to identify useful entities from police narrative reports. Preliminary evaluation results demonstrated that our approach is feasible and has some potential values for real-life applications. Our system achieved encouraging precision and recall rates for person names and narcotic drugs, but did not perform well for addresses and personal properties. Our future work includes conducting larger-scale evaluation studies and enhancing the system to capture human knowledge interactively. Efficient and effective access of criminal-justice data is critical for law enforcement personnel to perform investigations and fight crimes.

III. SYSTEM METHODOLOGY

a) Association Rule Mining

This technique is unsupervised learning method that used to find the hidden knowledges in unlabeled data. It is used to solve the issues if the learners get the unlabeled example data. In additional, association rule can discover the

DOI10.17148/IJARCCE.2017.6912

interesting co-occurrences of objects in large data sets. In the basic of association rule, the rule consists of two parts. 1) The predecessor, which is on the left side or called the left hand side (LHS). 2) The subsequent, which is on the right side or called the right hand side (RHS). A form of general association rule is LHS ! RHS, where LHS and RHS are disjoint item-sets.

If the LHS item-set arises then the RHS item-set will be likely to occur. For the efficient innovation of association rules, the imperative statistical measurements, the support and confidence measures, should be used together. A value of such measures is in he range of 0-1. If a association rule has very low support, this rule is likely to be uninteresting. As a consequence, the support measure is often used to dispose the uninteresting association rules. The confidence measure is used to gauge the reliability of association rules. Apriori algorithm is used to help prune the candidates explored during frequent item-set generation to reduce the processing time.

b) K-means Algorithm

Clustering is a data analyzing technique in unsupervised type. This technique is used to divide the same data into the same group and the different data into the other group. First, the user specifies the kcentroids number. The K is the number of the wanted clusters. Each cluster must have a centroid that is a mean of a cluster. Then each data record is assigned to the nearest centroid.

When all input data records have been assigned, the centroid changed of each cluster is updated by calculating the mean cluster. These processes will be repeated the assignment and improvement the centroids until the latest centroids do not change.

c) Classification

Classification technique is a supervised learning process that used to dispense objects to one of many pre-determined categories. The algorithms of classification have been extensively applied to the several problems that include many various applications. For example, it is used to solve the detecting of the suspect vehicles and intruders, the prediction of heart disease, the categorizing the document, etc. The basic concept of classification is described as the following: A collect data, also known as an input data, is used to process in a classification task. Each record consists of the attribute set and a class label. The class label is pre-determined category.

A collect data is divided into two sets.
1) Training set is paneled randomly that is used to generate a classification model, also known as a classifier, to envisage the class of the new unknown record.
2) Test set is a remaining set that is used to appraise the performance of the classification model.

d) Nearest Neighbor Approach

Nearest Neighbor approach is used to find the similarity between a new test record and a train record. When a train record closest to a new test record is discovered, the class label of a new test record is defined as the same class label of a train record. These processes can classify a new test record into the same group. However, nearest neighbor approach still has the limitations. If the number of records of train set is too less, train set does not cover all the possibilities of the attributes. To improve the performance of nearest-neighbor classification, the distance measurement may be useful to solve this problem such as euclidean distance. In addition to that the number of training records is more than one record contiguous to a new test record. K-Nearest Neighbor (KNN) method is used to solve the hitch. This method will use the greater part vote to hit upon the class label.

e) Crime Pattern

The issues of crime pattern are concerning with finding and predicting the hidden crime. Nowadays, the crime rate is increase continuously and the crime patterns are always changing. As a consequence, the behaviours in crime are difficult to be explained and predicted. The research interests on crime prevention and detection are concerning with finding and conducting the crime model to detect crimes. The challenge is modeling the crime attack behaviours that support crime detection although the crime patterns are changing. The predictive and statistic methods may be useful to find and conduct the crime model. The crime model should be able to predict and detect the criminal behaviors.

IV. 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:

DOI10.17148/IJARCCE.2017.6912

- 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 nearby and all calculation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. Both for sorting and regression, it can be helpful to load the contributions of the neighbors, so that the faster neighbors contribute extra to the average than the more distant ones.

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 in use from a set of items 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.

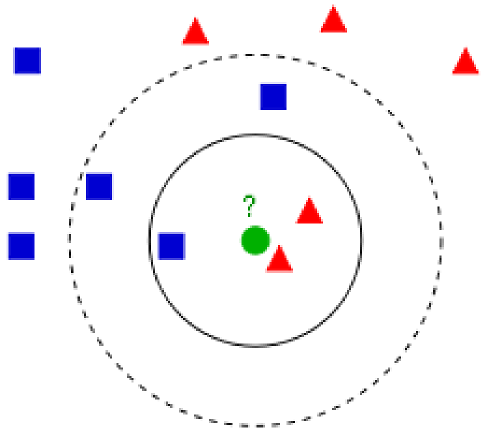

Fig 1.1 *k*-NN classification

In *k*-NN regression, the *k*-NN algorithm is meant for estimation permanent variables. One such algorithm uses a weighted average of the *k* nearest neighbors, weighted by the opposite of their distance. This algorithm works as follows:

1. Compute the Euclidean distance from the query example to the labeled examples.
2. Order the labeled examples by increasing distance.
3. Find a heuristically optimal number *k* of nearest neighbors. This is done using cross validation.
4. Calculate an inverse distance weighted average with the *k*-nearest multivariate neighbors.

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.

ACKNOWLEDGMENT

My heartfelt gratitude goes to my beloved guide **Mrs. M.Valarmathi, M.C.A., M.Phil.,** Head Of the Department, Department of Computer Science, Vivekanandha College for Women, Tiruchengode, India for dedication and patience in assigning me her valuable advice and efforts during the course of my studies.

DOI10.17148/IJARCCE.2017.6912

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’l Conf. Knowledge Discovery and Data Mining,pp. 551-556, 2004.

[10] T.N. Tran, R. Wehrens, and L.M.C. Buydens, “Knn Density-Based Clustering for High Dimensional Multispectral Images,”Proc. Second GRSS/ISPRS Joint Workshop Remote Sensing and Data Fusion over Urban Areas,pp. 147-151, 2003.

BIOGRAPHIES

Ms. V. Vishnupriya working as M. Phil Scholar in the Department of Computer Science at Vivekanandha College for Women, Tiruchengode, India. She has obtained her Under Graduate Degree in Computer Application from Vivekanandha college of Arts and Sciences for women, Tiruchengode, India and Master Degree in Computer Application from Vivekanandha Institute of Information and Management Studies, Tiruchengode, India.

Mrs. M. Valarmathi is currently working as Head of the Department in the Department of Computer Science at Vivekanandha College for Women, Tiruchengode, India. She did her Under graduate and Master degree at Vellalar College for Women, Erode. She did her M. Phil degree at Vinayaka Mission Deemed University, Salem, India.

DOI10.17148/IJARCCE.2017.6912