IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VIII (Nov – Dec. 2014), PP 13-17
www.iosrjournals.org

Comparative Study of Clustering Algorithms Used in Counter Terrorism

[1]Sanjay Dwivedi, MCA, MPhil, [2]Prabhat Pandey, PhD, DSc. OSD,
[3]Manmohan Singh Tiwari, PhD, Mohd. Athar Kalam[4]
[1]Research Scholar, APSU Rewa (MP)- India
[2]Additional Directorate, Rewa Division, Rewa (MP)- India
[3]HOD-Computer Science, SGS Govt. PG College, Sidhi (MP)- India
[4]Research Scholar, APS University, Rewa, MP- India

Abstract: *Data mining can be used to model crime detection problems, detect unusual patterns, terrorist activities and fraudulent behaviour. We will look at k-means clustering with some enhancements to aid in the process of identification of crime patterns. The k-means algorithm is one of the frequently used clustering method in data mining, due to its performance in clustering massive data sets. The final clustering result of the k-means clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In this paper a new method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity.*

Keywords: *Clustering, data mining, k-means, counter-terrorism, predictive analytics*

I. Introduction

Data mining is concerned with the automatic discovery of patterns and relationships in large databases.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. Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step . Numerous methods have been proposed to solve clustering problem. One of the most popular clustering method is k-means clustering algorithm developed by Mac Queen in 1967. The easiness of k-means clustering algorithm made this algorithm used in several fields. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups [1], [7], [9], [10]. The k-means clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently.Data mining and predictive analytics are gaining acceptance in criminal investigations and public safety. The predictability of violent crime is the foundation for the behavioural analysis of violent crime. In many ways, terrorism is violence with a larger agenda. Terrorism and efforts to support it also encompass other crimes including fraud, smuggling, money laundering, identity theft, and murder, which have been investigated successfully with the use of data mining and predictive analytics. Like many others, we have been exploring the use of data mining and predictive analytics in crime and intelligence analysis with some very promising preliminary successes. Described as "knowledge discovery" or "sense making" tools, data mining and predictive analytics give us an opportunity to manage and make sense of the information coming in, with the output being actionable intelligence products [6].

II. Background and related work

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. Crime control tends to use knowledge from the analysed data to control and prevent the occurrence of crime, while the criminal suppression tries to catch a criminal by using his/her history recorded in data mining.

K. A. Abdul Nazeer et al. [1] proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used, one method for finding the better

initial centroids. And another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time.

Zhang Chen et al. [3] proposed the initial centroids algorithm based on k-means that have avoided alternative randomness of initial centre. Fang Yuan proposed the initial centroids algorithm. The standard k-means algorithm selects k-objects randomly from the given data set as the initial centroids. If different initial values are given for the centroids, the accuracy output by the standard k-means algorithm can be affected. In Yuan's method the initial centroids are calculated systematically.

Koheri Arai et al. [5] proposed an algorithm for centroids initialisation for k-means. In this algorithm both k-means and hierarchical algorithms are used. This method utilizes all the clustering results of k-means in certain times. Then, the result transformed by combining withHierarchical algorithm in order to find the better initial cluster centres for k-means clustering algorithm.

De Bruin et. al. [6] introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. This method also provided a visual clustering of criminal careers and identification of classes of criminals.

Some results on crime mining have been made through using data mining techniques. Chen et al. applied data mining techniques to study crime cases, which mainly concerned entity extraction, pattern clustering, classification and social network analysis. Abraham et al. [4] proposed a method to employ log files as history data to search relationship by using the frequency occurrence of incidents.

III. k-Means Clustering Algorithm

k-Means Clustering algorithm attempt to find groups in the data. The following pseudo code shows the procedure:

Initialize $\mathbf{m}_i$, i = 1 to k random $\mathbf{x}^t$

Repeat

For all $\mathbf{x}^t$in X

$b_i^t \leftarrow 1$ if $\| \mathbf{x}^t - \mathbf{m}_i \| = \min_j \| \mathbf{x}^t - \mathbf{m}_j \|$

$b_i^t \leftarrow 0$ otherwise

For all $\mathbf{m}_i$, i = 1,…,k

$\mathbf{m}_i \leftarrow$ sum over t ($b_i^t \mathbf{x}^t$) / sum over t (b_i^t)

Until $\mathbf{m}_i$ converge

The vector m contains a reference to the sample mean of each cluster. x refers to each of our examples, and b contains our "estimated [class] labels" [10].

More simply in words, the algorithm roughly follows this approach:

1) Choose the way in which to initialize the mi to be the mean of each groupand do it.

2)For each example in the set, assign it to the closest group (denoted by mi).

3)For each mi, recalculate it based on the examples that are currently assigned to it.

4) Repeat steps 2-3 until mi converge.

IV. Enhanced K-means Algorithm

Input: data set x contains n data points; the number of cluster is k.

Output: k clusters of meet the criterion function convergence.

Program process:

Step 1. Initialize the cluster centre.

Step 1.1. Select a data point xi from data set X, set the identified as statistics and compute the distance between xi and other data point in the data set X. If it meet the distance threshold, then identify the data points as statistics, the density value of the data point xi add 1.

Step 1.2. Select the data point which is not identified as statistics, set the identified as statistics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

Step 1.3. Select data point from data set which the density value is greater than the threshold and add it to the corresponding high-density area set D.

Step 1.4. Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster centre set. Followed to find the k-1 data points, making the distance among k initial cluster centres are the largest.

Step 2. Assigned the n data points from data set X to the closet cluster.

Step 3. Adjust each cluster centre K by the formula (2).

Step 4. Calculate the distance of various data objects from each cluster centre by the given below formula (1), and redistribute the n data points to corresponding cluster [8].

$$d_{ji} = \left(1 - \frac{\sigma_j}{\sigma}\right) d_{ji} \quad (1)$$

Where j represents the j^{th} cluster c_j, i represents the i^{th} data point $x_{j,}d_j$ is the Euclidean distance between data point x_i and the cluster centre c_j, σ_i represents the square error of the cluster c_j, σ is the squares error sum of the K clusters c.

Step 5. Adjust each cluster centre K by the formula (2)-

$$k = \frac{d_{jh}}{D} xj_1 + \frac{d_{j(h-1)}}{D} xj_2 + \ldots\ldots\ldots + \frac{d_{j2}}{D} xj_{(h-2)} + \frac{d_{j1}}{D} xj_h (2)$$

Where j represents the j^{th} cluster, h is the number of data points in the cluster, d_{jh} represents the distance between the h^{th} data point which belongs to cluster c and cluster center with the condition that $dj_1 \leq dj_2 \leq \cdots \leq dj_h$, $\frac{d_{j1}}{D} + \frac{d_{j2}}{D} + \ldots + \frac{d_{jh}}{D} = 1$

Step 6. Calculate the criterion function E using formula (3), to determine whether the convergence, if convergence, then continue; otherwise, jump to step 4

Usually, the K-means algorithm criterion function adopts square error criterion, can be defined as:

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i - m_j||^2 \quad \text{where } x_i \in c_j (3)$$

In which, E is total square error of all the objects in the data cluster, xi bellows to data objectset, mi is mean value of cluster Ci (x and m are both multi-dimensional). The function of thiscriterion is to make the generated cluster be as compacted and independent as possible.

V. Terrorism Event Databases

Because of the challenges of collecting traditional crime data on terrorism, open source event databases have received increasing attention over the past four decades. While collection strategies have varied greatly, all of these databases rely on reports about terrorism from the print or electronic media. Event databases have serious limitations. In particular, because of the compelling interest that non-state terrorist groups have in media attention, open source information may be uniquely useful in the study of terrorism [5].

The evolving open source terrorist event databases have allowed for more rigorous analysis of terrorism and terrorist activity. However, a major limitation of most of these databases is that they have traditionally excluded domestic terrorist attacks. In general, international terrorist attacks are those involving a national or a group of nationals from one country crossing international borders and attacking targets in another country.

PGIS terrorism data was the only of the early open source databases on terrorism that attempted to track domestic as well as international attacks. The PGIS database was the original platform for the GTD. In the early days, PGIS relied especially on wire services and newspapers. By the 1990s, PGIS researchers were relying increasingly on the Internet.

VI. Data Collection Methodology

Data Integration is critical for the timing of a decision. The Global Terrorism Database (GTD) was developed to be a comprehensive, methodologically robust set of longitudinal data on incidents of domestic and international terrorism. Its primary purpose is to enable researchers and analysts to increase understanding of the phenomenon of terrorism. The GTD is specifically designed to be amenable to the latest quantitative analytic techniques used in the social and computational sciences[5].

The GTD was designed to gather a wide variety of etiological and situational variables pertaining to each terrorist incident. Depending on availability of information, the database records up to 120 separate attributes of each incident, including approximately 75 coded variables that can be used for statistical analysis. These are collected under eight broad categories, and include, whenever possible:

Table 1: Data Attributes

S.No.	Variables
	Incident date
	Region
	Country
	State
	City
	Latitude and longitude
	Perpetrator group name
	Tactic used in attack
	Nature of the target
	Identity of the target
	Type of weapons used

	Whether the incident was considered a success
	If and how a claim(s) of responsibility was made
	Amount of damage
	Total number of fatalities
	Total number of injured

Other variables provide information unique to specific types of cases, including kidnappings, hostage incidents, and hijackings.

VII. Results

We tested both k-means and enhanced k-means algorithms for the data sets with known clustering. The same data sets are used as an input for the original k-means algorithm. Both the algorithms need number of clusters as an input. In additional, for the original k-means algorithm the set of initial centroids also required. The enhanced method finds initial centroids systematically. The enhanced method requires only the data values and number of clusters as inputs and it does not take any additional inputs like threshold values.

Table 2: Comparison of the traditional and enhanced K-means clustering algorithm

Algorithm	**Data set**	**Accuracy of clustering**		
		Maximum	**Minimum**	**Average**
Traditional K-means	GTD	89.33	82.00	87.30
Enhanced K-means	GTD	90.45	83.26	88.51

It is clear from Table 2, as compared to the traditional K-means clustering algorithm, the enhanced K-means clustering algorithm on the data set in GTD clustering, accurate rate has improved significantly.

VIII. Conclusions and Future Trends

Given the geometrically increasing amounts of information, connecting the dots will be possible only with automated systems. Analysts are faced with a veritable tsunami of information that threatens to sweep them away. The ability to bring analytical and predictive models directly to operational personnel and into the operational environment holds the promise of allowing us to maneuver within the decision and execution cycle of our adversary, thereby gaining dominant battlespace awareness in the war on terrorism. Again, with the use of data mining and predictive analytics, information can serve as an interface between analytical and operational personnel. Perhaps the next step is to bring knowledge-discovery tools and the associated experts Data Mining and Predictive Analyticsto the frontlines of the war on terrorism. As a key clustering algorithm, K-means cluster algorithm has already become one of the hotspotsin the present. In this paper, through analysis the advantage and disadvantage of traditional K-meanscluster algorithm, elaborate two ways of improvement for K-means cluster algorithm, offerthe improved algorithm. However, researching on the improvementof K-means clustering algorithms are still not solved completely and the further attempt andexplore will be needed. We must exploit the technology available currently and begin anticipating the next move to achieve dominant battlespace awareness and victory in the war on terrorism.

References

[1]. K. A. Abdul Nazeer and M. P. Seba stian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, July 2009, London, UK.

[2]. Corcoran J.J., Wilson I.D. AND Ware J.A. (2003) Predicting the geo-temporal variations of crime and disorder, International Journal of Forecasting, Vol. 19, Pp.623–634.

[3]. Hsinchun Chen, Wingyan Chung, Yi Qin, et al. Crime Data Mining: An Overview and Case Studies. Proceeding of the 2003 annual national conference on Digital government research, Boston, M.A, 2003, pp 1-5.

[4]. T. Abraham and O. de Vel. Investigating profiling with computer forensic log data and association rules. Proc. Of the IEEE International Conference on Data Mining (ICDM'06), 2006, pp 11-18.

[5]. Gary LaFree and Laura Dugan.(2007). "Introducing the Global Terrorism Database," Political Violence and Terrorism 19:181-204G. P. Zarri. Semantic web and Knowledge Representation, Proc. Of the 13th International Workshop on Database and Expert System Applications (DEXA'02), 2002, pp. 1529-4188.

[6]. De Bruin, J.S. , Cocx, T.K. , Kosters, W.A. , Laros, J. and Kok, J.N. (2006) Data mining approaches to criminal career analysis," in Proceedings of the Sixth international Conference on Data Mining (ICDM'06), Pp. 171-177

[7]. Teknomo, Kardi, "K-means Clustering Tutorials".

[8]. Malathi. A, Dr. S. Santosh Baboo and Anbarasi. An intelligent analysis of city crime data using data mining. International Conference on Information and Electronics Engineering IPCSIT Vol 06. pp. 130-134.

[9]. Ke Sun, Jie Liu, Xueying Wang, K mean cluster algorithm with refined initial center point, in: Journal of Shenyang Normal University(Natural Science), 27(4), 2009, 448-451.

[10]. Alpaydin, Ethem. Introduction To Machine Learning. Cambridge, Massachusetts: MIT Press. 2004.

[11]. http://databases.about.com/od/datamining/a/kmeans.htm.

Dr. MMS Tiwari is the head of department of Computer Science at SGS Govt. PG College, Sidhi, MP, India.

He has received PhD in Physics from APS University Rewa, MP, India. He is an accomplished academician with over 30 years of rich experiences in the field of electronics & Computer Science.He has published a number of papers in both national & international journals and has presented these in various seminars.

Sanjay Dwivedi, Research Scholar of Computer Science at APS University, Rewa(MP), India. He has obtained

MPhil from the APSU Rewa and MCA from RGPV Bhopal (MP). He has worked with the **DRDO**, Ministry of defence, Govt. Of India's project - "An Intelligent Fire Detection System" , as a project trainee.His area of research is data mining and its applications. He has published several papers in both national & international journals. He has worked as reviewer of the International Journal IJCTE Singapore, and is a member of various International societies like IACSIT, IJCTE, ACM, IAENG.