International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 4, April 2014 ISSN 2319 - 4847

Finding similar case subset and hotspot detection in felonious data set using data mining algorithms: Weighted Clustering and classification

Ms. Apexa Joshi[1], Dr. Suresh M.B.[2]

[1]Research Scholar, School of Science, RK University, Rajkot, India.
[1]Assistant Professor, JVIMS-MCA-GTU, Jamnagar, India.

[2]Professor & Head, Dept. of ISE, East West Institute of Technology,
Viswaneedam Post, Bangalore

Abstract

Felonious activities have been increased terrifically. So safety has become a key issue to the societies. The control activities should be taken by the Security agencies (like FBI, CBI, CID, police Department and other law enforcement agencies). The security agencies have to take corrective activities for reducing the Felonious activities. To analyze the Felonious data, clustering and classification techniques are used. These data will be stored in the Felonious database. Spatial clustering algorithm and structured crime classification are used to categorize the Felonious activities. These algorithms assistance to identify the hotspot of felonious activities. Felons will be recognized based on the eyewitness or hint at the crime spot. The credentials of hotspot of Felons activities will help the Security agencies to provide more safety to the particular territory, this assistances to prevent crimes in future. When we apply this concept to all areas at least we can reduce Felonious activities. Entirely the criminalities cannot be controlled. In this paper we find the hotspot of the felonious activities and finding the felons by using clustering and classification algorithms. It can not only provide multiple hints to solve crimes but also improve efficiency to catch the felons.

Keywords: crime analysis and data mining, classification, K-Modes clustering, crime mapping, weighted attributes, GIS, information gain ratio, hotspot, spatial pattern.

1. INTRODUCTION

Security and crime predicting events are most vital worries for both citizens and government. In Gujarat State of India, zillions of cases have been stored in felonious database of department of public security and this number increases in millions every year. Massive felon data have been gathered in law enforcement organizations in felonious databases. To classify the felonious activities and felons there are methods available to decrease it. Where number of felonious activities has occurred that place is recognized as hotspot. By recognizing the hotspot of the felonious activities, this will support the police department to avoid such kind of movement in future in the all locations. The data mining concept is very much useful to analyze the crimes and felons. Classification and clustering algorithms are realistic. First we have done the classification after that applying clustering algorithm. Based on the type of crime, the felonious activity will be classified, after the classification is done based on the classification outcome, the related type of felonious activities will be clustered together. By using GIS the hotspot of the felonious activities will be viewed. Searching subsets of related cases from huge felonious data is a key task for intelligence analysts in law-enforcement societies. If such subsets are originate and provided to crime investigators, multiple inklings can be obtained from diverse cases. For example, the offender in case A stole bicycles by cracking the lock and in case B the offender stole bicycles around shopping malls. If case A and case B were confirmed to be similar cases, investigators can conclude that the offender often stole bicycles around shopping malls by cracking the lock. This additional information may help to solve the two cases together. Once these two cases were solved, so did all cases in the same subset. This example demonstrates that finding similar cases subsets not only assists in the process of crime investigation but also greatly improves efficiency. However, the conventional approach of finding similar cases subsets is relatively inefficient. Not until a “seed case”, either a new happened case or an important case, needed to be investigated did the intelligence analysts begin to use the query system to find the similar case subsets. Analysts usually input one or more keywords of the “seed case” to query the database. Then they had to review the query results one by one to confirm whether they are similar cases. The reviewing process often takes one or two hours. If a lot of cases have to be queried, the process takes even longer. How to use computer aided approach to automatically find the similar case from a large number of cases without a “seed case” is a great demand. Another drawback of manually querying the database is that all attributes are treated equally and the importance of some specific attributes in different case categories is not reflected. We find that when intelligence analysts query databases, they often query some specific attributes, such as location, victim, tools used etc. These attributes play a critical role in finding the similar case subsets because offenders tend to act similarly as they did before, like choosing the same

location or using the same tools. In addition, different behavioral attributes play the main part in different cases categories. For instance, according to trained analysts' experience, burglary offenders tend to focus on ways to break in a house while fraud offenders tend to focus on choosing their targets, either a person or a company. Therefore these different focuses must been well utilized in the process of finding similar cases subsets to reflect the nature of each case category. The importance of weighing attributes was first addressed in [7]. In [7], the author adopted expert-based methodology. The weight of attributes is given by experienced experts. The fatal defect of this method lies in the fact that different experts will give distinct weights, making the result subjective and less convincing. An automatic way of weighing the attributes is required. In this paper, we propose a two-phase clustering algorithm called AK-Modes to find the similar case subsets: first we compute weights of behavioral attributes related to a given case category; then cases are put into different subsets using AK Modes algorithm, which takes attributes' weights into consideration. The main contributions of our works are:

- Emphasize the importance of attribute weighing in crime investigation and propose the use of Information Gain Ratio (IGR) in classification domain for the calculation of attributes' weights;
- Propose a two-phase clustering algorithm AK-Modes which combines the attribute-weighing phase and the clustering phase together;
- Experiments show AK-Modes is effective and can find significant results.

The rest of the paper is organized as follows. Section 2 discusses crime analysis Section 3 contains crime mapping section 4 defines crime classification section 5 contains crime clustering (problem of finding the similar case subsets as a clustering problem, two-phase algorithm AK-Modes for finding the similar case subset, experiment result) and finally section 6 offers the conclusion of our work.

2. CRIME MAPPING

Crime mapping helps the police department to protect the people from the crime more effectively. An understanding of where and why crimes occur will help to fight against with the crime. Simple map shows where the crimes have been occurred.

2.1 Visualizing the Crime Location

Digital maps visualize the crime scenario in quicker manner. At which place the crimes have occurred that will be visualized. Instead of searching from the list of events, mapping is easy to visualize the crime hot spot.

2.2 Integrate the Community Characteristics

Community characteristics mean the most possible places for occurring the crime activities. For example slums, schools, parks, colleges, alcohol permit location and etc.

2.3 Producing the Maps

At any geographical level the maps can be produced. Where the crimes have occurred that particular place will be shaded darker. The number of crime incidents percentage change will be displayed by shading the area's location.

3. CRIME ANALYSIS

Crime analysis is a set of systematic and analytical process for providing the information regarding crime patterns at the particular time. Crime investigation is an important activity for identifying the crime hotspot. It supports the number of department functions that includes patrol deployment, special operations, tactical units, investigations, planning and research, crime prevention and administrative services. Crime analysis can be divided into three categories, these are following as,

3.1 Tactical

Tactical is an analytical process for providing the information to assist operations personnel (patrol and investigative officers) for identifying the crime trends, patterns, series and hotspot. It includes at which time crime is occurred and associating the criminal activities by crime method.

3.2 Strategic

It includes the preparation of crime statistical summaries, resource acquisition and allocation studies.

3.3 Administrative

It focuses on provisioning on economic, geographic or social information to administration.

3.4 Identification of clusters

GIS identifies the areas that contain the more number of clusters (hotspot). The similar type of crime activities will be grouped together. Based on the clusters' result, which cluster contains the more number of criminal activities that will be called as crime hotspot for the particular crime.

3.5 Comparison of location of crime hotspot

The crime hotspot that have been identified over several months.

3.6 Comparison of hotspot with different crime types

The identified crime hotspot will be compared with the other type of crime hotspot. For example burglary type of crime hotspot will be compared with the murder type of crime hotspot.

4. CRIME CLASSIFICATION

To classify the crime incidents based on the similarity between the crime objects stored in the class, structure crime classification is used. Classification is the hierarchy of these attributes. These attributes are represented by classification in three ways,

1) Classification of crime place
2) Classification crime types
3) Classification of crime time

The structured crime classification algorithm is used to identify the more similar objects in the data sets. Algorithm, to find the hotspot and coldspot from the dataset.

Input: Database DB

Output: hotspot or coldspot

1. Assign S=DB
2. Apply purification attribute Ai by Cn]
3. Repeat
   a. Find the similarity of crime attribute objects (C.Ai, C.Ai+1)
   b. Find the probability of particular crime classification = Probability (C.Ai,Classification)
   c. Threshold T=(Cluster Area-Sparse Area)
   d. Find F(C) =classification U Pi (Ci)
   e. If F(C)>positive description
      Produce a hot spot
   Else
      Produce a cold spot
4. Go to step 3

Let S denotes a set of crime incidents. Ai is an attribute of crime incidents and Ci is a classification of each crime attribute Ai. For two elements x1, x2 in the tree of Ci, if there is a path from x1 to x2 is called the parent of x2. Furthermore, x1 is a generalization of x2. In the structure crime classification algorithm, the national dissipation between the events is similar and the events are more similar. Choose the crime attribute Ai in the crime class C. Find the similarity of each crime attribute of crime objects if both objects have the same similarity, join these two objects have the same crime attribute incident and put into the same class C. And finally find the F(C) based on the probability of crime incident occurring in the particular class to which it is merged. If F(C) is greater than the positive description, it produces a crime hot spot, and otherwise it produces a crime cold spot.

5. CRIME CLUSTERING

Clustering is data mining technique for grouping the similar type of crimes will be grouped together. In this paper the burglary crime will be clustered, based on the clusters' result the crime hotspot will be identified.

5.1 Problem Statement

In this section, we will first introduce some terminologies used in AK-Modes and then define our problem as a clustering problem.

Categorical attribute: A categorical attribute is the one whose values do not have a natural ordering. Some typical categorical attributes are: gender, education level, marriage status etc. In our crime databases, behavioral attributes are categorical attribute because they usually describe an offender's trait, such attributes include victim, modus operandi, and location and so on.

Attribute weight: The weight of an attribute is a real value which indicates the importance of the attribute in different case categories. The larger the weight is, the more important the attribute is in that case category.

Case: A case is a record in a database that consists of some numerical and categorical attributes. In our context, we only consider the categorical attributes since they contain the most useful information of a case. For example, two burglary cases and a fraud case are shown in table 1:

TABLE1. Three Cases with Categorical Attributes

Case No.	000001	000002	000003
Case Category	Burglary	Fraud	Burglary
Time	2006-3-28	2006-4-3	2006-4-6
Location Category	Dwelling House	Hotel	Dwelling House
Modus Operandi	Invade from Window	Temptation of money	Invade from window
Victim/Target	Old woman	Middle-aged man	Middle-aged woman
Motivation	For money	For money	For money
Characteristic	Individual	Gang	Individual

Distance of two cases: The Distance of two cases A and B is calculated by summing all distances between respective attributes, that is:

$$Dist(A,B)=\sum_{i=1}^{m} d(a_i,b_i) \qquad (1)$$

m is the number of attributes and d() is a distance measure for two attribute values ai and bi.

Similar Cases: Two cases A and B are similar if and only if their distance is less than a predefined threshold α (0<α <1). That is, A and B are similar if and only if Dist(A, B) < α . In the example of table 1,case No.000001 and case No.000003 are similar cases.

The finding similar case subset problem: Given a dataset containing n cases (D1….,Dn) with m categorical attributes (A 1, ..., Am)find k similar case subset with centers C1, ..., Ck , such that

$$\min_{k} \sum_{j=1}^{k} \sum_{D_i \in C_j} dist(D_i,C_j) \qquad (2)$$

$$\text{s.t} \quad dist(D_i,C_j) < \alpha \ , D_i \in C_j \qquad (3)$$

We assume that we do not have to assign every case into a subset, only those cases satisfying the similar case condition are assigned.

5.2 AK-Modes Algorithm

The AK-Modes algorithm includes attribute-weighing phase and finding similar case subsets phase. Attributes are first weighted in the attribute-weighing phase. The weighted attributes are then integrated into clustering phase and finally the result is obtained. The process of the two-phase algorithm is shown in Fig.1.

5.2.1 Attribute Weighing Phase

Attribute Weighing is a critical phase before finding the similar case subsets since different attributes play the main role in different case categories. The task we should do in this phase is:

Given a list of attributes A1, ...,Am, use a weighing function F to every attribute Ai (1≤ i ≤ m) and compute their F-values, find top-k (1≤ k < m) attributes with larger F values.

Having studied various techniques regarding on attribute selection and dimension reduction, we decided to adopt the concept of “Information Gain (IG)” to support our algorithm. IG is an important concept in information theory and has been widely used in classification, especially in Decision Tree classification. IG can reflect the amount of information that an attribute contains in classification. The attribute with larger IG implies that the root node of the decision tree should be split on this attribute first.

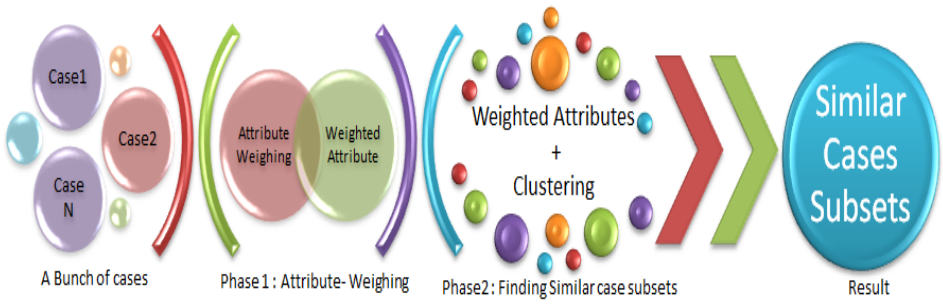

Figure 1: The processes of AK-Modes

Before computing IG we must know the class label that each record belongs to since IG is a concept in classification domain. In our work, we take the “case category” attribute as the class-label attribute because cases of different categories tend to favor different attributes. The results after taking “case category” as class-label attribute can reflect the nature of each case category.

Now we introduce some concepts and formulas for computation of IG.

The entropy for an attribute i A can be calculated as:

$$H(A_i)=-\sum_{a \in V_i} p(a)\log_2 p(a) \qquad (4)$$

Where a is the attribute value on Ai whose domain is Vi. p(a) is the probability that attribute Ai has the value a . Information Gain (IG): assume attribute Ai has m different values and the dataset X can be divided by i A into m different subsets X1,...,Xm, the IG of Ai is:

$$IG(A_i) = H(T) - \sum_{i=1}^{m} \frac{|X_i|}{|X|} H(X_i) \quad (5)$$

Where T is+ a class-label attribute. Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A notable problem is that it bias attributes that can take on a large number of distinct values, for example, customer ID in a business database. To overcome this drawback, information gain ration (IGR) is introduced by Quinlan [12].

Information Gain Ratio (IGR): The IGR of an attribute Ai is:

$$IGR(A_i) = IG(A_i) / IV(A_i) \quad (6)$$

Where IV(Ai) is the gain ration that can be computed as:

$$IV(A_i) = -\sum_{i=1}^{m} \frac{|X_i|}{X} \log_2 \frac{|Xi|}{|X|} \quad (7)$$

The value of IGR must be normalized so that it lies between 0 and 1.

5.2.2 Clustering Phase

The problem of finding the similar case subsets can be solved with the clustering technique. Nowadays most of current clustering algorithms focus on numerical data since it is easier to calculate their similarity in geometric space, for example, Euclidean distance measure. However, in our work, we are processing the crime cases whose attributes are categorical. Thus how to cluster these cases becomes a challenge.

Many algorithms have been proposed for clustering categorical data. Based on the classical K-Means clustering algorithm, a K-Modes algorithm is presented for categorical domain in [13], which introduces a simple similarity measure for categorical objects and sets modes instead of means for clusters. K-Modes are suitable for our problem since it has a good scalable capability for large datasets.

The main differences between AK-Modes algorithm and the K-Modes algorithm is that it integrates the results of attribute-weighing into the clustering process. The steps in AK Modes algorithm is as follows:

AK-Modes Algorithm

Input: Dataset D , Weighted Attributes A , threshold α

Output: clustering result

Step 1: Using the result of attribute-weighing phase A, find the Decisive Attribute and the number of clusters k. Then find initial k objects as the initial mode of every cluster.

Step 2: For every caseC in D , calculate its distance to every mode M and find the distance d to its closet mode. If d >α , case C is abandoned. Otherwise, put case C into the cluster with the closest mode.

Step 3: Update the mode of each cluster.

Step 4: Terminate the algorithm till all the modes do not change. If not, go back to step 2.

Here are some key points we must pay attention to in the AK-Modes algorithm.

5.2.2.1 The input setting of number of clusters k (in Step 1)

When dealing with a large bunch of cases, the intelligence analysts have little prior knowledge on how many subsets exist in them. Thus it is difficult for the intelligence analysts to determine the value of k before running the K-Modes algorithm. To solve this problem, an automatic mechanism for deciding k is required. Here we utilize the result of attribute weighing phase.

We define the attribute with the largest weight after the attribute-weighing phase as “Decisive Attribute”. Because of its largest weight, decisive attribute is the most important factor we must consider when finding similar case subsets. So we can let k be the number of subsets the data divided by decisive attribute.

5.2.2.2 The distance measure for computing distance between two cases (in Step 2)

As stated in Section 3, the distance between two cases A and B is calculated by summing all distances between respective attributes. For categorical attributes, we adopt the original similarity measure in K-Modes algorithm to compute the distance of two categorical attribute values in respective attributes, that is:

$$d(a_i, b_i) = \begin{cases} 0 (a_i = b_i) \\ 1 (a_i \neq b_i) \end{cases}$$

Where ai and bi are two categorical attribute values of A and B.

5.2.2.3 The threshold α for judging similar cases (in Step 2)

α is a parameter to determine whether two cases are similar. If α is too large, many cases will be similar leading to a large number of cases in the result, which is either impractical or little knowledge can be found within the result. But a small value of α may cause some similar cases to be abandoned by AK-Modes thus result in an information-loss situation. Therefore, a careful setting of α must be studied. We will see it in the experiment section.

5.3 Experiments

In this session, we evaluate the effectiveness of our algorithm. First we test the necessity of attribute weighing using UCI datasets. Then we conduct experiments on the real crime data to see how significant result we can find.

5.3.1 Experiment on UCI data

To show the necessity of attribute weighing, we conduct an experiment on datasets from the UCI Machine Learning Repository [14] to compare the accuracy of the K-Modes algorithm with and without weighted attributes.

5.3.1.1 Dataset

The dataset used are the Mushroom dataset and the Wisconsin Breast Cancer (original) dataset. The descriptions of the two datasets are as follows:

- The Mushroom dataset: It has 22 attributes and 8124records. Each record represents physical characteristics of a single mushroom. A classification label of poisonous or edible is provided with each record. The numbers of edible and poisonous mushrooms in the dataset are 4208 and 3916, respectively.
- Wisconsin Breast Cancer (original) dataset: It has 699 instances with 9 attributes. Each record is labeled as benign (458% or 65.5%) or malignant (241% or 34.5%). In our literature, all attributes are considered categorical with values 1,2,. . .,10.

5.3.1.2 Validation measure

Since each record in dataset we used is labeled, the accuracy of the result can be computed as follows. Given the final number of clusters k, accuracy r was defined as:,

$$r = \frac{\sum_{i=1}^{k} a_i}{n}$$

Where n is the number of records in the dataset and ai is the number of instances occurring in both cluster i and its corresponding class, which had the maximal value.

5.3.1.3 Experiment result

In the K-Modes algorithm with weighted attributes, we pick half of both datasets as training data for the calculation of attributes' weight since true labels exist. The remaining data are tested for accuracy.
After running the K-modes algorithm with and without weighted attributes, we get the results shown in Figure 2 and 3.

Figure 2: Accuracy vs. different number of clusters in Mushroom dataset

Figure 3: Accuracy vs. different number of clusters in Winsconsin Breast Cancer (original) dataset

From Fig. 2 and Fig. 3, we can see that in the Mushroom data set, K-modes algorithm using weighted attributes gets higher accuracy than the version without weighted attributes.

In the Wisconsin Breast Cancer (original) dataset, K-modes algorithm using weighted attributes has comparative accuracy to the version without weighted attributes. In a word, using weighted attributes in K-Modes algorithm can improve the accuracy.

5.3.2 Experiment on crime data

5.3.2.1 Dataset

The dataset used in the experiment was collected from the central database in Department of Public Security of Gujarat province of India. The database consists of all the cases occurring within 11 cities of the province since 2004. The crime data of one typical moderate-population city at year 2008 was provided as our data source.

We especially put our effort to do experiments on burglary offences because of their larger volume and higher occurring frequency. The experiment data includes 18428 records of burglary cases with 16 behavioral attributes. Moreover, the original data was cleaned and integrated as a preprocessing step since a lot of missing values or typing errors existed.

5.3.2.2 Experiment design and results

We partitioned the crime data into training data and test data. Training data is used as the source for attribute weighing and test data is used for the task of finding similar case subsets.

Training data: we also took about half (9172 cases) of burglary cases of 2008 as the training data to evaluate the weight of attributes. However, not all 16 attributes are evaluated for their weights. According to intelligence analysts' advice, we picked only 8 specific attributes: occurring area, location category, and victim/target, invade ways, modus operandi, crime motivation, crime characteristic and opportunities to commit a crime.

After setting the class label attribute to "case category", we got top-4 attributes shown in table 2 with larger IG values:

TABLE 2: Result of Attribute Weighing

Attribute Name	Weight
Victim/target	0.288
Modus Operandi	0.286
Location Category	0.240
Characteristic	0.186

Testing data: We used the remaining 9256 burglary cases as the testing data for finding similar case subsets.

5.3.2.2.1 The setting of threshold α

The threshold α is an important parameter in the AK Modes algorithm. It determines the number of cases in the result. We ran the AK-Modes algorithm with different value of α from range 1 to 0.1 and record the number of cases in the result to see how α affects the number of cases in the result.

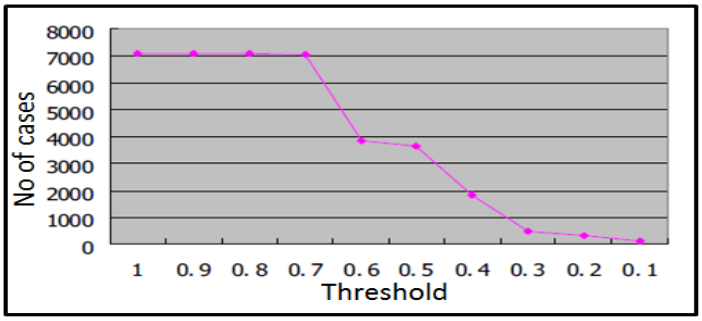

Figure 4: Similarity threshold vs. number of cases in the result

As shown in Fig. 4, the number of cases is decreased when the threshold changes from 1 to 0.1, which is accordance with the definition of similar cases. What's more, there are periods of dramatic decrease in the number of cases when the threshold changes from 0.7 to 0.6 and 0.5 to 0.3. This indicates that the threshold can effectively reduce the number of cases in the results if appropriately set.

5.3.2.2.2 The significant results of finding similar cases subsets

Inspired by the results discovered in a), we set the threshold α to 0.3 because it can get a suitable number of cases in the result and see how significant our result is after running the AK-Modes algorithm. There are totally 489 cases in the result and these cases are divided into 10 groups. We asked two experienced intelligence analysts to validate the result and they found some interesting discoveries in 2 groups.

In group 1 of 11 cases, the similarities of the behavior traits of the offenders are:

- Motorcycles were stolen in all cases.
- 100% offenders stole motorcycles by connecting the electric wires to start the engine.
- In 5 cases, offenders commit the crime at plazas or streets.

In group 2 of 45 cases, the similarities of the behavior traits of the offenders are:

- 100% victims were young women.
- 16 burglary offences occurred at commercial places, such as supermarkets, retail shops and hair salons.

By summarizing the similarities of cases in the same group, intelligence analysts can have a clearer clue of these crimes. Useful information obtained may help solving the cases or at least assist in the process of crime investigation. Also, the information can be provided to senior leaders for the prevention and prediction of crimes in the future. For instance, in the above results, the summing-up information tells people, especially young women, to be precautious to take care of their belongings/wallets at commercial places.

6. Conclusion

This paper presents the method to identify the hotspot of crime. Based on the type of crime the police department can easily identify the hotspot of the burglary crime. GIS is used to visualize the hotspot of burglary crime. Data mining concept is used to prevent and identify the crimes. Classification technique is used to classify the different crimes. Clustering technique is used to cluster the similar type of crimes together, based on the clusters' result the burglary type of crime hotspot will be identified. This result will help to reduce the burglary type crime. In future all type of crimes' hotspot will be identified; through this the crime activities will be reduced.

Finding similar cases subsets is an important task in crime investigation. Given a seed case, intelligence analysts often pay a lot of effort to review the results after querying the database based on their domain knowledge. In this paper, we proposed an AK-Modes algorithm to automatically find the similar case subsets without a given "seed case". AK-Modes combine the attribute-weighing phase with the process of clustering. Attribute-weighing phase is necessary because it highlights the importance and priority of different attributes in various case categories. The advantage of attribute weighing is shown using the UCI datasets and significant results have been discovered on the real crime dataset. We believe that the application of our model in practice can effectively improve the efficiency compared with the traditional manually reviewing approaches and can assist in the decision-making process.

In the future, there are still works to be done to improve our algorithm. First, the distance measure of two cases can be further studied. Semantic distance will be a good direction for this study. Second, by now the intelligence analysts have to set the threshold α in AK-Modes algorithm based on their experience. How to get a reasonable threshold automatically using some novel algorithm is still a challenging task for us.

References

[1] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: A General Framework and Some Examples," Computer, vol. 37, 2004.

[2] H. Chen, W. Chung, Y. Qin, M. Chau, J. Xu, G. Wang, R. Zheng and H. Atabakhsh, "Crime Data Mining: An Overview and Case Studies", in Proceedings of the annual national conference on Digital government research, Boston, pp.1-5, 2003.

[3] M. Chau, J. Xu, and H. Chen, "Extracting Meaningful Entities from Police Narrative Reports," in Procceedings of The National Conference on Digital Government Research, pp. 271-275, 2002.

[4] G. Wang, H. Chen, and H. Atabakhsh, "Automatically Detecting Deceptive Criminal Identities," Comm.ACM, Mar.2004, pp.70-76, 2004.

[5] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey, and Z. Huang, "COPLINK: Visualization for crime analysis," in Proceedings of The National Conference on Digital Government Research, pp.1-6, 2003.

[6] R. Adderley and P. B. Musgrove, "Data mining case study: Modeling the behavior of offenders who commit serious sexual assaults," in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, pp. 215-220, 2001.

[7] S. V. Nath, "Crime Pattern Detection Using Data Mining," in Proceedings of the 2006 IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology, pp. 41-44, 2006.

[8] Fatih Ozgul , Zeki Erdem and Chris Bowerman, "Prediction of past unsolved terrorist attacks," in Proceedings of the IEEE international conference on Intelligence and security informatics, Richardson, Texas,USA, pp.37-42, 2009.

[9] J. S. Bruin, T. K. Cocx, W. A. Kosters, J. F. J. Laros and J. N. Kok, "Data Mining Approches to Criminal Career Analysis," in Proceedings of th Sixth International Conference on Data Mining, pp. 171-177, 2006.

[10] G. C. Oatley, J. Zeleznikow, and B. W. Ewart, "Matching and predicting crimes," in Proceedings of the Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence, pp. 19-32, 2004.

[11] L. Ding, D. Steil, M. Hudnall, B. Dixon, R. Smith, D. Brown and A. Parrish, "PerpSearch: An Integrated Crime Detection System," in Proceedings of the 2009 IEEE international conference on Intelligence and security informatics, Richardson, Texas, USA, pp.161-163, 2009.

[12] Quinlan, J. R, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[13] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, vol. 2, pp.283–304, 1998.

[14] C. J. Merz, P. Merphy, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/~mlearn/MLRRepository.html, 1996.

[15] Sajendra Kumar, Ram Kumar Rana, Pawan Singh, "A Semantic Query Transformation Approach Based on Ontology for Search Engine" ISSN : 0975-3397 Vol. 4 No. 05 May 2012, International Journal on Computer Science and Engineering (IJCSE).

[16] Lizhen Li, Zhifeng Dong, Keming Xie, Ontology of general concept for Semantic Searching, Second International Conference on Computer Modeling and Simulation 2010.

[17] Yuangui Lei, Victoria Uren, and Enrico Motta, SemSearch: A Search Engine for the Semantic Web: IEEE Transactions on knowledge and Data engineering, VOL.19, NO. 2, FEBRUARY 2007.

[18] Thomas B. Passin, Explores Guide to the Semantic Web, Manning Publications Co. GreenWich 2004.

[19] Ronen Feldman, James Sanger, The text mining handbook: Advanced approach in analyzing unstructured data, Cambridge University Press USA2007.

[20] C.P.Johnson "Crime Mapping and Analysis using GIS", Geomatics 2000, Conference on Geomatics in Electronic Governence, January 2000, Pune.

AUTHOR

Ms. Apexa Joshi received the B.C.A and M.C.A degrees in Computer Science and application from Saurashtra University in 2005 and 2008, respectively. She is pursuing Ph.D in Computer science. Her area of specialization is data mining. She is having 10+ publications in international and national journals. She is Assistant professor at Jayshuklal Vadhar Institute of Management studies – GTU in MCA department from last 6 years.

Dr. Suresh M. B. received the B.E, M.Tech. and Ph.D. degrees in Information Science and Engineering. His area of specialization is Digital Image Processing. He is having the 16 years' wide experience in teaching industry. He is published 35 + number of papers in international and national journals. He is Professor & Head at East West Institute of Technology, Bangalore, Karnataka, India.