Indian Journal of Science and Technology, *Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87910, February 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Attribute Weighted Fuzzy Clustering Algorithm for Mixed Crime Data

Revatthy Krishnamurthy[1] and J. Satheesh Kumar[2]

[1]Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; revatthykm11@yahoo.com
[2]School of Computer Science and Engineering, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; jsathee@rediffmail.com

Abstract:

Background/Objectives: The main objective is to design and develop a clustering algorithm for finding similar sub sets from crime data. This paper focuses a method for developing an algorithm and modify the existing technique in three ways, such as i) new attribute weightage scheme instead of IGR, ii) suitability to mixed data and iii) using FCM-based clustering instead of k-means. **Methods/Statistical analysis**: Generally, the effectiveness of clustering algorithm is completely based on distance matching that finds the similarity between data records and centroid. Giving equal importance for all the attributes is not much effective in clustering process. Instead, attribute weightage could be included in distance matching. A weight vector is generated based on mutual information. The method for attribute weightage is common for both numerical and categorical data. Finally, the grouping of similar sub sets is done based on FCM-based clustering procedure in which the distance matching is carried out based on the attribute weights. **Findings:** The experimental analysis has done using crime and hepatitis datasets where the performance of the proposed clustering algorithm has been analyzed. Results show that proposed FCM method has good accuracy than the AK-mode. **Application/Improvements:** Proposed method plays an important role in crime domain for better prediction. Type II fuzzy can also be used for better closeness analysis.

Keywords: Crime Data, Categorical Data, FCM Clustering, Numerical Data, Non-Overlapping Interval, Overlapping Interval

1. Introduction

One of the major applications of clustering is crime analysis that has become one of the most essential activities in the world, since the technology development and the high growth of community have results high magnitude of crimes, most of the time with bizarre patterns[1,2]. Crime data are categorized based on the crime type, analyze and identify important crime hot-spots and prediction and prevention over the protection safety of future crimes are very needful in crime analysis. There are number of analysis methods are available for identifying and assessing potential hot spots in crime analysis[3, 4]. Among the recent algorithms for this application, AK-Modes that have turn into and successfully progressed the efficiency compared with the established one and can help in the decision-making process for categorical data. The number of clusters, K, must be supplied as a parameter is the main disadvantage in the k-means algorithm[5].

Self-Organization Map (SOM) and K-means methods have been used to evaluate the patterns for clusters. Furthermore, the lack of pattern quality assessment in spatial clustering could lead to meaningless or unknown information. Validity of SOM and K-means methods[6] has been examined using compactness and separation criteria. In this crime application, data has been separated into two parts. First part contains simulated data which has 2D x, y coordinates and subsequent part of data was real data corresponding to crime investigation.

In this paper, a new approach for attribute weighted dissimilarity that is proposed to design and develop a clustering algorithm for finding similar sub sets for crime

*Author for correspondence

data is presented. This weightage includes numerical and categorical dataset. The distant values are measured individually for both numerical and categorical dataset weightage values and are used to find an accurate distance matching of the data set. In the attribute weightage, overlapping and non-overlapping interval methods are adopted to find the weightage function. Then, the fuzzy clustering method is used to cluster the data records. The experimentation is done using the uci crime and hepatitis datasets and the performance of the proposed clustering algorithm is analyzed.

The organization of this paper is as follows: Problem definition and contribution of the paper is presented in Section II. The proposed technique of new attribute weightage method for finding similar sub sets from crime data is calculated in Section III. An experimental result is discussed in Section IV and section V concludes the paper.

2. Problem Identification and Contributions

K-means algorithm is used as an initial process for many other algorithms. But the main disadvantage of k-means algorithm is the problem of choosing similarity measures. Traditionalalgorithms[7,8] are used for numerical data and must be modified to take into consideration of the specific characteristic of categorical data. Recently, Ak-mode was proposed based on attribute weighted scheme and k-means clustering procedure. In the existing work, information gain ratio was used to find out the attribute weightage and consider the mutual information among the various classes for attribute weightage computation the performance would be significantly increases proposed algorithm. A new attribute weightage formula with the idea of mutual information is proposed. Along this, the dissimilarity measure is designed for both numerical and categorical data sets. In addition, grouping of similar subsets was done using Fuzzy clustering procedure[9,10] in which it is adapted to handle both numerical and categorical data.

A new attribute weighted dissimilarity measure is applied to the FCM Clustering algorithm for mixed crime data. The new dissimilarity measure is integrated with new attribute weighted dissimilarity measure to form a mixed weighted dissimilarity measure. Based on this, a mixed attribute weighting algorithm is proposed to cluster high-dimensional numerical data and categorical data. The performance of the mixed attribute weightage algorithm is investigated for both Numerical and Categorical data sets[11,12].

3. Proposed Attribute Weighted Fuzzy Clustering Algorithm for Mixed Crime Data

Here, a mutual information-dependent formula that is used to generate a weight vector for entire attributes of input mixed data is presented. Finally, the grouping of similar sub sets is done based on FCM-based clustering procedure in which the distance matching is carried out based on the attribute weights[13,14]. Finding similar sub sets from crime dataset can be done by the following three steps such as

Step 1: New attribute weightage scheme
Step 2: Proposed distance measure
Step3: Adapting FCM clustering for mixed data

3.1 New Attribute Weightage Scheme

In the new attribute weightage scheme, two weightage methods called the Weightage for Numerical data value (α_i) and Weight for categorical data value (β_i) with the help of overlapping and non-overlapping data values are used. Finally, the weightage values in the distance measure formula and the distance between the two attributes are calculated.

3.1.1 Weightage for Numerical Data Value

The weight of a numerical attribute (α_i) indicates the importance of attribute in different crime cases.

$$\text{Weightage}(\alpha_i) = \frac{\alpha_i^{NOL} + \alpha_i^{OL} + 1}{\alpha_i^{OL}} \quad (1)$$

In this crime data set, the weightage for numerical data value is computed with the help of overlapping and non-overlapping intervals. In numerical and categorical methods, overlapping and non-overlapping are defined in two different ways. In numerical method, each class contains minimum to maximum data value for using derived numerical data. In case of categorical method, non-overlapping is the maximum number of attributes and overlapping is the minimum number of attributes in the class.

$$a_i^{NOL} = \frac{p\left(1,2,3.....n\right)^{NOL}}{\sum_{i=1}^{n}\left(\frac{1}{1+p(i)}\right)^{NOL}} \quad (2)$$

In the Non overlapping for Numerical data(α^{NOL}), the numerator considers the probability of attribute values within the non-overlapping intervals together with all the classes and denominator considers the probability of attribute values within non overlapping interval individually in all the classes. The equation (2) only measures the Non overlapping value for numerical data.

The probability of Non-overlapping value in all the classes $\left(p\left(1,2,3.....n\right)^{NOL}\right)$ is defined as the ratio of the summation of found the non-overlapping value in each class $\left(\sum_{i=1}^{n} f_i^{NOL}\right)$ and the total number of attributes (F_T), the equation (3) is used to calculate the total number of non-overlapping values for crime datasets.

$$p\left(1,2,3.....n\right)^{NOL} = \frac{\sum_{i=1}^{n} f_i^{NOL}}{F_T} \quad (3)$$

The probability of individual classes for the non- overlapping values $p(i)^{NOL}$ which is defined as the ratio of non-overlapping attributes in individual classes $\left(f_{(i)}^{NOL}\right)$ and the total number of attributes in those classes $\left(f_{T(i)}\right)$. This equation(4) is used to find out the number of non-overlapping values of the individual classes.

$$p(i)^{NOL} = \frac{f_{(i)}^{NOL}}{f_{T(i)}} \quad (4)$$

In the Overlapping for Numerical data (a_i^{OL}), the numerator considers the probability of attribute values within the overlapping interval together with all the classes and denominator considers the probability of attribute values within overlapping interval individually in all the classes. The equation (5) measures the Overlapping value for numerical data.

$$a_i^{OL} = \frac{p\left(1,2,3.....n\right)^{OL}}{\sum_{i=1}^{n}\left(\frac{1}{1+p(i)}\right)^{OL}} \quad (5)$$

The probability of overlapping value is computed for all the classes $\left(p\left(1,2,3.....n\right)^{OL}\right)$. It is defined as the ratio of the summation of found overlapping value in each class $\left(\sum_{i=1}^{n} f_i^{OL}\right)$ and the total number of attributes (F_T). Here, the equation (6) is used to calculate the total number of overlapping value for the crime dataset.

$$p\left(1,2,3.....n\right)^{OL} = \frac{\sum_{i=1}^{n} f_i^{OL}}{F_T} \quad (6)$$

The probability of individual class is measured for overlapping value $p(i)^{NOL}$ which is defined as the ratio of non- attributes in individual classes $\left(f_{(i)}^{NOL}\right)$ and the total number of attributes in those classes $\left(f_{T(i)}\right)$. Here, the equation (7) is used to find out the number of overlapping individual classes.

$$p(i)^{OL} = \frac{f_{(i)}^{OL}}{f_{T(i)}} \quad (7)$$

3.1.2 Weightage for Categorical Data Dalue

The weight of Categorical attribute is a data value which indicates the importance of the attribute in different crime cases. The larger weight is the more important in case category.

$$\text{Weightage}\left(\beta_i\right) = \frac{\beta_i^{CNOL} + \beta_i^{COL} + 1}{\beta_i^{COL}} \quad (8)$$

β_i = weightage for Categorical data

Where, $\left(\beta_i^{CNOL}, \beta_i^{COL}\right)$ are non-overlapping and overlapping data for categorical attributes. A categorical attribute is one whose value does not have a natural ordering. Some typical categorical attributes are our behavioral attributes and they are usually described an offender's trait, location, often category, sub category and so on. The Non overlapping for categorical data (β_i^{CNOL}), the numerator considers the probability of maximum attribute values within the non-overlapping interval together in all the classes and denominator considers the probability of maximum attribute values within overlapping interval individually in all the classes. The equation (9) only measures the Non overlapping value for Categorical data.

$$\beta_i^{CNOL} = \frac{p\left(1,2,3.....n\right)^{MAX}}{\sum_{i=1}^{n}\left(\frac{1}{1+p\left(i\right)}\right)^{MAX}} \quad (9)$$

The probability of finding the maximum attributes in all the classes is $\left(p\left(1,2,3.....n\right)^{MAX}\right)$. It is defined as ratio of the summation of finding the maximum attributes in each class $\left(\sum_{i=1}^{n} f_i^{MAX}\right)$ to the total number of attributes (F_T). The equation (10) is used to calculate the total number of maximum attributes of crime dataset.

$$p\left(1,2,3....n\right)^{MAX} = \frac{f_{(i)}^{MAX}}{f_{T(i)}} \quad (10)$$

The probability of individual classes are measured as the maximum attributes $p\left(i\right)^{MAX}$ which is defined as the ratio of non-overlapping attributes in individual classes $\left(f_{(i)}^{MAX}\right)$ to the total number of attributes in those classes $\left(f_{T(i)}\right)$. The equation (11) is used to find out the number of overlapping values in the individual classes.

$$p\left(i\right)^{MAX} = \frac{f_{(i)}^{MAX}}{f_{T(i)}} \quad (11)$$

In the overlapping for categorical data (β^{COL}), the numerator considers the probability of minimum attribute values within the overlapping interval together with all the classes and denominator considers the probability of minimum attribute values within overlapping interval individually in all the classes. The equation (12) only measures the overlapping value for Categorical data.

$$\beta_i^{COL} = \frac{p\left(1,2,3.....n\right)^{MIN}}{\sum_{i=1}^{n}\left(\frac{1}{1+p\left(i\right)}\right)^{MIN}} \quad (12)$$

The probability of the minimum attributes in all the classes is $\left(p\left(1,2,3.....n\right)^{MIN}\right)$.It is defined as ratio of the summation of find out the maximum attributes in each classes $\left(\sum_{i=1}^{n} f_i^{MIN}\right)$ and the total number of attributes (F_T). The equation (13) is used to calculate the total number of minimum attributes of crime dataset.

$$p\left(1,2,3.....n\right)^{MIN} = \frac{\sum_{i=1}^{n} f_i^{MIN}}{F_T} \quad (13)$$

The probability of individual classes are measured with the maximum attributes $p\left(i\right)^{MIN}$ which is defined as the ratio of non-overlapping attributes in individual classes $\left(f_{(i)}^{MIN}\right)$ to the total number of attributes in those classes $\left(f_{T(i)}\right)$.

Here, the equation (14) is used to find out the number of overlapping values of individual classes.

$$p\left(i\right)^{MIN} = \frac{f_{(i)}^{MIN}}{f_{T(i)}} \quad (14)$$

3.2 Proposed Distance Measure

The distance matching is carried out based on attribute weights and two separate distance measures for numerical and categorical data for the crime dataset. The below equation has two separate sections containing the formula for Numerical and Categorical data values.

$$M_i = \sum_{i=1}^{m1} \alpha_i d_i^{nu}\left(x_i\ y_i\right) + \sum_{i=1}^{m2} \beta_i d_i^{ca}\left(x_i\ y_i\right) \quad (15)$$

Where, $d_i^{nu}\left(x_i\ y_i\right)$ = distance measure for numerical data

$d_i^{ca}\left(x_i\ y_i\right)$ = distance measure for categorical data

α_i = weightage for numerical data

β_i = weightage for categorical data

m1 = numerical object

m2 = categorical object

Definition A: (Distance Measure for numerical data)

$$d^{nu}{}_i(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_{m1}-y_{m1})^2} = \sqrt{\sum_{i=1}^{m1}(x_i-y_i)^2} \quad (16)$$

The distance measure is used for computing the distance between the two data points with respect to its numerical attributes is Euclidean distance. The Euclidean distance is computed as follows:

Definition B: (Dissimilarity Measure for categorical data)

The dissimilarity of two data points for the categorical attributes, $X=(y_1,y_2,\cdots,y_c)$ and $Y=(x_1,x_2,\cdots,x_c)$ is computed based on the following equation

$$d_i^{ca}(X,Y)=\sum_{j=1}^{m2}\delta(x_j,y_j) \tag{17}$$

$$\delta(x_j,y_j)=\begin{cases}0\ ; & if\ x_j=y_j\\ 1\ ; & otherwise\end{cases} \tag{18}$$

3.3 FCM Clustering Algorithm for both Numerical and Categorical Data

Let consider a mixed data of $D=\{X_{1,}\ldots,X_n\}$ set of categorical objects and $E=\{Y_{1,}\ldots,Y_n\}$ set of numerical objects. The Numerical objects are clustered where each object $X_i=(x_{i,1,}\ldots x_{1,m1})$, $1\le i\le m1$ and categorical objects are clustered, where each object $Y_i=(y_{i,1,}\ldots y_{1,m2})$, $1\le i\le m2$ is defined by m attributes that contains m1 categorical and m2 numerical. Then, this problem can be mathematically reformulated as follows:

$$R_m=\sum_{k=1}^{N}\sum_{l=1}^{C}u_{kl}{}^m\left\|w(x_k-c_l)\right\|^2 \tag{19}$$

Where m is any real number greater than 1, $u_{kl}{}^m$ is the degree of membership of x_k in the cluster l x_k, is the k of d-dimensional measured data, c_l - the d-dimension center of the cluster, and ||*|| -is the similarity between any measured data and the center.

An iterative optimization of fuzzy partitioning with the update of membership u_{kl} is

$$u_{kl}=\frac{1}{\sum_{t=1}^{C}\left(\frac{\|x_k-c_l\|}{\|x_k-c_t\|}\right)^{\frac{2}{m-1}}} \tag{20}$$

This iteration stops when, $\max_{ij}\left\{\left|u_{ij}^{(k+1)}-u_{ij}^{(k)}\right|\right\}<\varepsilon$ where ε is a termination criterion between 0 and 1, where as k is the iteration steps. Local minimum converges in this procedure.

3.4 Basic Steps of FCM-Mode Clustering Algorithm:

1. Initialize U = [u_{kl}] matrix, $U^{(0)}$
2. Update, (u^k), (u^{k+1})

$$u_{kl}=\frac{1}{\sum_{t=1}^{C}\left(\frac{\|x_k-c_l\|}{\|x_k-c_t\|}\right)^{\frac{2}{m-1}}}$$

3. Centroid computation for numerical data is based on definition C and categorical data is based on definition D.
4. If $\|\left(u^{k+1}\right),\left(u^k\right)\|<\infty$, then STOP.

Definition C: Centroid updating of numerical data

The center of the cluster for any numerical attribute is computed based on membership values and numerical data value.

$$C_j=\frac{\sum_{i=1}^{N}u_{ij}{}^m.y_i}{\sum_{i=1}^{N}u_{ij}{}^m} \tag{21}$$

Here, C_j- cluster dimension center.
y_i - Numerical data object

Definition D: (Updating of k-centroids for categorical data)

The centroids of categorical attributes within the cluster C_l are arranged in accordance with their relative frequency f_{xi}. The category x_i with high relative frequency of 'm_1' categorical attributes is chosen for the new representative. For example, gender is a categorical attribute having two categories (male and female) and location is also a categorical attribute having a number of categories (Australia, inner Sydney, Liverpool, Campbell town and more).

3.5 AWFCM-Modes Algorithm

Input: Dataset D, Weighted Attributes for numerical and categorical data (α_i, β_i).
Output: clustering result.
Step 1: Initialize U= [u_{kl}] matrix, $U^{(0)}$
Step 2: Updating the membership (u^k), (u^{k+1}) for numerical and categorical data.

$$u_{kl}=\frac{1}{\sum_{t=1}^{C}\left(\frac{\|x_k-c_l\|}{\|x_k-c_t\|}\right)^{\frac{2}{m-1}}}$$

Step 3: Centroid computation for numerical data based on definition C.
Step 4: Centroid computation for categorical data based on definition D.

Step 5: Go to step2, until $\left(u^{k+1}\right)$, $\left(u^{k}\right)|<\infty$, then STOP.

4. Result And Discussion

4.1 Experimental Set Up

The proposed approach has been implemented in Matlab7.12. The experimentation was done with two widely applied datasets namely, crime and hepatitis. Here, Crime data is taken from publicly available source and Hepatitis is taken from the UCI machine learning repository[6].

4.2 Evaluation Metrics

The Cluster Accuracy is used to evaluate the performance of the proposed approach of the crime dataset. The evaluation metric is given below. Clustering accuracy[15, 16] (CA) is,

$$CA = \frac{1}{N}\sum_{i=1}^{T} X_i \qquad (22)$$

Where, N→ Number of data points in the dataset
T→ Number of resultant cluster
X_i→ Number of data points occurring in both cluster i and its corresponding class.

4.3 Performance Evaluation

The experimental results of the proposed approach are shown in Figure 1, Figure 2, Figure 3, Figure 4 and the clustering accuracy is calculated. In this work, the AWFCM mode for the clustering approach is adopted. The performance evaluation of the proposed AWFCM

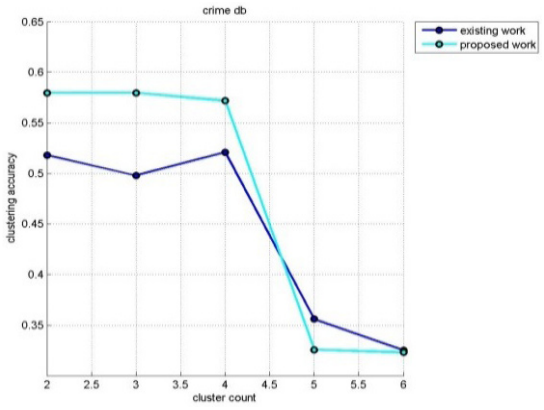

Figure 1. Clustering accuracy of dataset 1 (AK- mode) and data set 2 (AWFCM) for 50 iteration

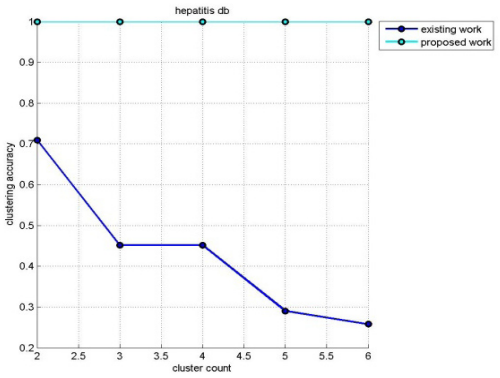

Figure 2. Clustering accuracy of dataset 1 (AK- mode) and data set 2 (AWFCM) for 50 iterations

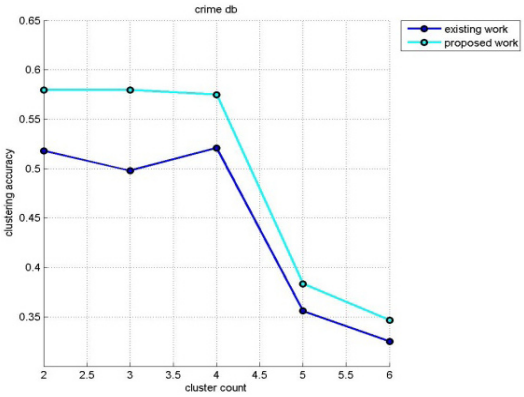

Figure 3. Clustering accuracy of dataset 1 (AK- mode) and data set 2 (AWFCM) for 50 iterations

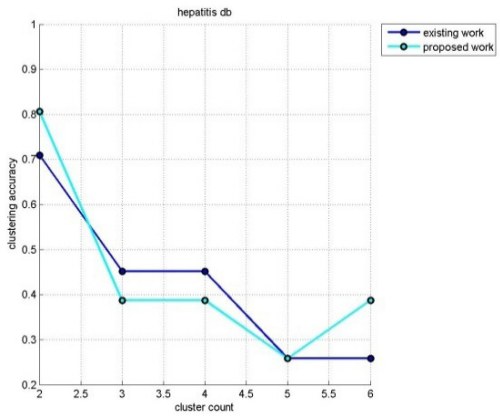

Figure 4. Clustering accuracy of dataset 1 (AK- mode) and data set 2 (AWFCM) for 50 iterations

mode is compared with AK mode and the experimental results are discussed. From figure (1), it is determined that crime dataset for AK mode has 53% accuracy, whereas AWFCM is with 58%.In Figure (2), the observation for hepatitis dataset gives 70% accuracy in AK mode and 100% accuracy in AWFCM. From Figure (3), the crime dataset for AK mode has 53% accuracy, whereas AWFCM with 58% were determined. In Figure (4), the observation of hepatitis dataset gives 70% accuracy in AK-mode and 80% accuracy for AWFCM.

5. Conclusion

In this paper, the method for attribute weightage of numerical and categorical data using overlapping and non-overlapping interval is suggested. Attribute weightage-based clustering algorithm was developed for finding similar sub sets from crime data. The motivational research and modified the existing technique in three ways, such as i) new attribute weightage scheme instead of IGR, ii) suitability to mixed data, and iii) Using FCM-based clustering instead of k-means. A FCM-based clustering approach is proposed in this paper to find similar sub sets from crime data. The performance of the proposed clustering algorithm is analyzed based on clustering accuracy with the AK-mode algorithm. From the experimental results it is observed that proposed FCM method has better accuracy 88% than the AK-mode.

6. References

1. Garcianocetti F, Solano Gonzalez J. Connectivity base dk-Hop Clustering in Wireless networks. Telecommunication Systems. 2003; 1(4):205–20.
2. Redmond SJ, Heneghan C. A method for initializing the K-means clustering algorithm using kd-trees. Pattern Recognition Letters. 2007; 28(8):965–73.
3. Marx Z, Dagan I, Buhmann JM, Shamir E. Coupled Clustering: A method for detecting structural correspondence. Journal of Machine Learning Research. 2003; 3:747–80.
4. Bicego M, Figueiredo MAT. Soft clustering using weighted one-class support vector machines. Pattern Recognition. 2009; 42(1):27–32.
5. Ma L, Chen Y, Huang H. AK-Modes: A weighted Clustering Algorithm for Finding Similar Case Subsets. Proceedings of IEEE International Conference on Intelligence Systems and Knowledge Engineering, (ISKE), Hangzhou. 2010. p. 218–23.
6. Mohammad khanloo M, Bashiri M. A Clustering based Location allocation Problem Considering Transportation Costs and Statistical Properties. 2013; 26:597–604.
7. Usman G, Ahmad U, Ahmad M. Improved K-Means Clustering Algorithm by Getting Initial Cenroids. World Applied Science Journal. 2013; 27(4):543–51.
8. Aliabadian A. A robust clustering approach based on KNN and modified C-means algorithm. World Applied Science Journal. 2013; 25(4):585–91.
9. Khatibi BV, Jawawi DNA, Hashim SZM, Khatibi E. A new fuzzy clustering based method to increase the accuracy of software development effort estimation. World Applied Science Journal. 2011; 14(9):1265–75.
10. Di Martino F, Sessa S. Implementation of the extended fuzzy C-means algorithm in geographic information systems. Journal of Uncertain Systems. 2009; 3(4):298–306.
11. Datasets, Center for Machine learning and Intelligent Systems, Donald Bren School of Information and Computer Sciences, University of California, Irvine. Available from: http://cml.ics.uci.edu/. 25/05/2015.
12. Krishnamurthy R, Kumar JS. Survey of data mining techniques on crime data analysis. International Journal of Data Mining Techniques and Applications. 2012; 1(2):117–20.
13. Brown DE. The regional crime analysis program (RECAP): A Frame work for mining data to catch criminals. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA. 1998; 3:2848–53.
14. Stoffel K, Cotofrei P, Han D. Fuzzy methods for forensic data analysis. IEEE Proceedings of Soft Computing and Pattern Recognition. 2010; 1–6.
15. Ferligoj A. Recent developments in cluster analysis. Telecommunication Systems. 2003; 1(4):1–9.
16. Iqbal R, Azmi Murad MA, Mustapha A, Shariat Panahy PH, Khanahmadliravi N. An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology. 2013; 6(3):4219–25.