GIS based Serial Crime Analysis using Data Mining Techniques

S. Sivaranjani
Asst.Prof – Dept of CSE,
Faculty of Engineering
Avinashilingam University

S. Sivakumari, PhD
Prof & Head - Dept of CSE
Faculty of Engineering
Avinashilingam University

S. Maragatham, PhD
Prof & Dean –
Dept of Civil Faculty of
Engineering
Avinashilingam University

ABSTRACT

Serial crimes are the problems which are mostly committed by minority of offenders. The law enforcement people are forced to find out the serial crime which is considered as complex task. In order to find out the serial crimes, one needs to investigate the large number of crimes which are unlinked with each other. These problem need to be analyzed in different ways to find out the link that exist among each crimes which happened in various locations. In the previous work, cut clustering algorithm were used to cluster the similar type of crime happened in various locations. However the existing work lacks from the number of labeled classes used for clustering which will limit the number of data points to be used. Also the cut clustering algorithm used in the existing research leads to be more complex that does not specify the characteristics of the sub graphs that are created hence it leads to the complexity of deciding a node to be added or deleted. To overcome these problems, in our work majority weighted minority over sampling technique is used to handle class imbalance problem and the dynamic cut clustering is introduced which overcomes the limitation of graph cut clustering algorithm. In this work Quantum Geographic Information System (QGIS) tool is used to visualize, navigate, manipulate, and analyze geographic crime datasets.

Keywords

Geographic Information System (GIS), Serial Crime, Dynamic Cut Clustering.

1. INTRODUCTION

Researchers have found that resolving serial crimes has been the privilege of the criminal justice and law enforcement officials. The decision support system started helping the law enforcement officers and detectives to track crimes in order to reduce the crime rates. The detection of linked crimes is helpful to law enforcement for several reasons. First, the collection of information from crime scenes increases the amount of available evidence. Second, the joint investigation of multiple crimes enables more efficient use of law enforcement resources. Law enforcement agencies need to handle a large amount of report in the detection of series of crimes manually. An integration of decision support system and GIS (Geographic Information System) enable the law enforcement agencies to review the series of crimes in an efficient way. The system which is used to capture, store, retrieve, analyze and display spatial information is called as GIS.

In the proposed work, we develop a framework where data mining concepts are used to link crime locations and non crime areas which is useful to investigate the relationship between linked crimes and unlinked crimes. Graph based Clustering algorithm is used to cluster the input data based upon the similarity or distance measures. A QGIS is used to analyze and map the series of crime locations.

2. BACKGROUND STUDIES

Anton Borg et al., proposed a decision support system for comparing and analyzing residential burglaries. A large amount of crime records must be reviewed by the law enforcement officers in order to find similar crimes. The prospective usage of the cut clustering algorithm for grouping crimes towards reducing the amount of residential burglary crimes for analysis based on characteristics was investigated. The characteristics used were residential characteristics, modus operandi, stolen goods, and spatial temporal similarity [1].

Krunal Patel et al., proposed a GIS based Decision Support System (DSS) which integrate and access massive amount of location based information.GIS based DSS allows police personnel to plan effectively for emergency response, determine mitigation priorities, analyze historical events, and predict future events.GIS based DSS can furthermore used to get significant information to emergency responders upon dispatch of police personnel [2].

Nick et al., discussed about the application of a Decision support system (DSS) that simulates both human and environmental factors. Although other agent-based models of crime do exist, this research represents the initial working example of integrating a behavioral framework into an agent based model for the simulation of crime [3].

Dawei Wang et al., introduced a framework for spatial data mining to study crime hotspots using Geospatial Discriminative Patterns (GDPatterns) to find the significant difference between hotspots and typical areas in a geo-spatial dataset. By utilizing the GD patterns the author developed a unique model called Hotspot Optimization Tool (HOT) for the improved efficiency in identification process of crime hotspots. The real world crime dataset collected from a northeast city in the United States were used for the study. GDPattern clusters were grouped and visualized based on the similarity measure [4].

M. Ahmed and R. S. Salihu proposed the use of Geographic Information Systems (GIS) in determining crime hotspots in Dala LGA of Kano State and also analyzed the challenges for the police departments in implementing computerized crime mapping systems [5].

G.Li et al., proposed a Bayesian spatio-temporal model for the space–time variation in burglary risk and also presented a novel two-stage method in order to classify the areas into crime cold spots, hot spots or neither and the applications of temporal dynamics within each risk category. A further contribution was the enclosure of covariates with the model in order to elucidate the space–time classification of areas [6].

G., Lin, J., &Zheng, W. designed and implemented a Web based Geographic Information System (GIS) for crime mapping and decision support. Four hotspot mapping techniques, i.e., chloropleth mapping, grid mapping, spatial ellipse mapping and kernel density mapping, were implemented in the system [7].

Guanli Huang and Guanhua Huang proposed three new methodologies for geographical profiling, namely, Bayesian—Factor analysis model, Time series analysis model and GIS(Geographic Information System) Decay model, to study geographical profiling problems. The author compared their accuracy, efficiency, sensitivity and robust according to 11 historical serial crime samples and Monte Carlo simulations [8].

Jiaji Zhou and Long Chen discussed about Geographic Profiling which has successfully assisted investigations for serial crimes. This profiling considers the multi-cluster feature of serial criminal spots and also proposed a Multi-point Centrography model as a natural extension of Single-point Centrography for geographic profiling. K-means clustering is first performed on the data samples and then Single-point Centrography is adopted to derive a probability distribution on each cluster. Finally, a weighted combination of each distribution is formed to make next-crime spot prediction [9].

Nelson Devia and Richard Weber proposed an agent-based simulation model to test diverse policing strategies in a virtual environment that generates artificial street crime data. The model can thus evaluate the strategies effectiveness and collateral effects in supporting police personnel for decision-making process [10].

Shyam Varan Nath formulated crime pattern detection as a machine learning task and thereby use data mining to support police detectives in solving crimes. The author identified the significant attributes using expert based semi-supervised learning method and developed the scheme and discussed about k-means clustering with some enhancements to aid in the process of identification of crime patterns. Thus these data mining methods have promising future [11].

Yang Zhou Hong et.al., proposed a novel graph clustering algorithm, SACluster, through a unified distance measure based on both structural and attribute similarities. The large graph associated with attributes is partitioned into k clusters so that each cluster may contain densely connected sub-graphs with identical attribute values. The proposed methodology automatically learns the degree of contributions of attribute similarity and structural similarity. Theoretical analysis was done to show that SA-Cluster is converging [12].

AhamedShafeeq B. M and Binu V.S analyzed the correlation between various crimes. The whole work was divided into two parts: 1) to check spatial autocorrelation between various crimes 2) to compare various attribute clusters and its relation. The spatial distribution of various crimes in the states of India and also the correlation between the attributes and crimes in 2012 were analyzed using exploratory spatial analysis methods. The clustering is used to identify the patterns with different crime densities, Employment and Police force distribution. Finally, the thematic maps of clusters were used to compare its correlation. The crime cluster scan was used for planning various security measures in the states [13].

Ohino,Alan and T.Murray provided a better understanding of spatial distribution of crime based upon GIS and quantitative technique. Police departments, city officials and policy makers all recognize the importance of a better understanding of the dynamics of crime. The establishment of an analytical and theoretical framework for evaluating the relationship between aspects of place and the clustering of crime will undoubtedly lead to enhanced crime prevention strategies [14].

Giles C Oatley and Brian W Ewart developed a software system that enables the trending of historical data for crime reduction based upon victim, offender, location and details of victimizations. The software utilizes visualization tools and is capable of mapping a range of sophisticated predictions. [15].

3. METHODOLOGY

In order to cluster serial crimes various methods and techniques have been used which are discussed in the proceeding section. Fig 1 shows the statistical results of crime data.

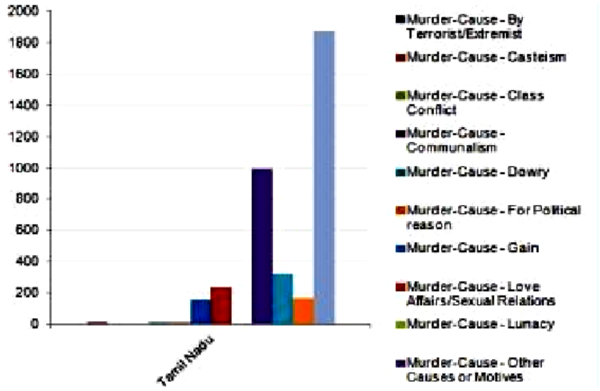

Fig. 1: Statistical results of crime data

Fig 1 depicts the statistical report of crime which were committed for various causes and motives.

3.1 Data Representation

The instances are inserted into an n x n adjacency matrix, and for each pair in the adjacency matrix a similarity index is computed as an edge representation. The minimum cut tree algorithm, when given complete graphs or near-complete graphs, can produce trees that are star-shaped, i.e. each node is connected directly to the root node, or unary.

3.2 Cut Clustering

The cut clustering algorithm is a graph-based clustering algorithm based on minimum cut tree algorithms to cluster the input data based upon similar characteristics. Any changes made between the edges leads to the re-processing of the entire structure of the graph. The procedure involves repeated coarsening of the original graph to smaller size graph, partitioning it and then remapping the partitions back.

3.3 Majority Weighted Minority Oversampling Technique (MWMOTE)

Class over sampling is done in order to support the different classes of crimes which are having different number of items. In the real time scenario, the crimes that are predicted in the different location may consist of different number. It needs to be analyzed for the efficient prediction of the crime behaviour. But it will lead to performance degradation where the numbers of data’s are different in number. In order to overcome this problem, the methodology called the synthetic over sampling methodology is implemented. The algorithm of MWMOTE is described below:

The nearest neighbourhood samples are clustered by using K-Nearest Neighbourhood algorithm. The distances between the two nodes are calculated by using Euclidian distance. Hence the distance between two nodes acts as a weight between the majority and minority samples.

Algorithm 1:
MWMOTE(Gmaj,Gmin,N,n1, n2;,n3)

Input:

1. Gmaj: Set of majority class samples
2. Gmin: Set of minority class samples
3. N: Number of synthetic samples to be generated
4. n1: Number of neighbours used for predicting noisy minority class samples
5. n2: Number of majority neighbours used for constructing informative minority set
6. n3: Number of minority neighbours used for constructing informative minority set

End

3.4 Dynamic Cut Clustering Algorithm

Let us construct a graph G=(V,E)were V represents the vertices and E represents the edges. The number of vertices is denoted as n. The order of graph is given by n=|V|.

After sampling classes, grouping of crimes are done in order to predict same types of crimes that are happening in different location. It is done to identify the place of crime locations where the particular crime is happening frequently. To do so dynamic cut clustering algorithm is introduced which supports the deletion of data into the clusters. Once the edge gets added or deleted from the clusters we can retrieve the clusters of the original graph from the coarsened graph. Thus the computation time of our algorithm is very less. We can obtain the set of clusters seen so far, without any further processing. Let L = {L1, L2... Ls}, be the clusters of the graph G (V, E). Let A be the adjacency matrix of G.

3.4.1. Edge Deletion

Intra-Cluster Edge Deletion.

The connectivity within the clusters gets broken, when an edge gets deleted, which arises the need of re-clustering. The edges whose end vertices belongs to the same cluster is deleted from the existing graph. Here. So except L_u, all the other clusters are contracted to form the coarsened graph.

Hence the computation time took for clustering of serial crime is efficient when compared to cut clustering algorithm.

Algorithm 2:

Input: W (G), Θ (G), G_α^θ=($V_\alpha, E_\alpha\{b,d\}, c_\alpha^\theta$), edge {b,d} with weight Δ

Output: W (G_θ), Θ(G_θ) regarding G_θ

```
1.  $\theta_{b,d}$ ← first  min-t-$v_{b,d}$-cut given by FLOWALGO (t, $v_b$,d)
2.  If $c_\alpha^\theta$ ($\theta_{b,d}$) = $c_\alpha$ ($\theta_\alpha^\theta$) then
3.  return W (G), Θ (G)
4.  else
5.  set (R, $V_\alpha$\R):= $\theta_{b,d}$ with tєR
6.  For all $C_i$ != $C_{b,d}$ do $\theta_{b,d}$ = (R U $C_i$, ($V_\alpha$\R)\$C_i$)
7.  L (t) ← Ø, l(t) ← Ø
8.  Θ ($G_\Theta$) ← {$\theta_{b,d}$}, W ($G_\Theta$) ← {$v_{b,d}$}
9.  For I =1,…,z do
10. Add $v_i$ to L(t)
11. D ($v_i$) ←Ø
12. W ($G_\Theta$), Θ ($G_\Theta$) ← Check Cut-vertices (W(G), Θ(G), W ($G_\Theta$), Θ($G_\Theta$), $G_\alpha^\theta$ , {b,d}, D, L(t))
13. W ($G_\Theta$) ← W ($G_\Theta$) U $v_{b,d}$, Θ ← Θ U {$\theta_{b,d}$}
14. Resolve all crossings in Θ ($G_\Theta$)
15. isolate the sink t from all remaining un-clustered vertices
16. Return W ($G_\Theta$), Θ ($G_\Theta$).
```

3.5 GIS Representation

Geographic Information System (GIS) uses geographical features and computer-generated maps as an interface for accessing enormous amount of locality based information [16]. It allows police personnel in analyzing and mitigating the historical events to plan accordingly for emergency response, and to predict the future measures. It is used worldwide by police departments, both large and small, in providing visualized solutions for crime analysis and tracking criminals. Maps that displays the hot spot locations are very helpful in crime mapping for the police patrol to discover the places that they are most wanted.

The serial crimes are detected using the efficient methodologies and it is displayed on the screen by using the GIS tools [17][18]. It is done to achieve the various GIS terminologies in order to make more user friendly environment. Here we use QGIS tool in order to map the crime locations. Here we use a raster layer to insert the maps in the GIS tool. Then the output obtained by clustering algorithm is marked on the maps by using delimiter text layer. Fig.2 shows the output of serial crimes which is mapped using QGIS Tool.

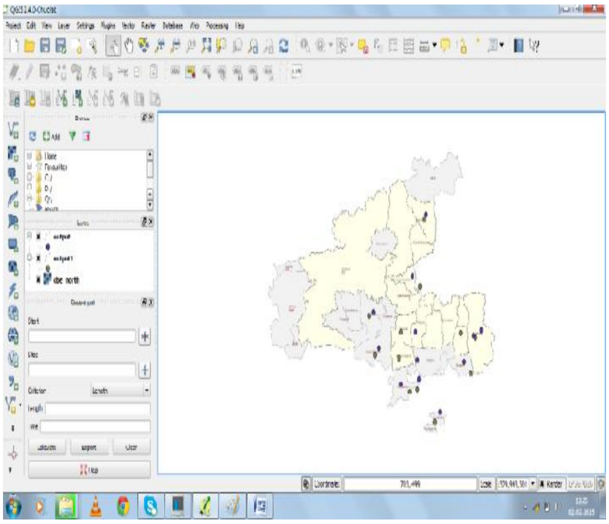

Fig.2. GIS Representation of Serial Crime

3.6 Cluster Validation Measurements

This section discuss about the metrics used to evaluate the performance of our algorithm. The metrics used to evaluate the accuracy and efficiency of clustering algorithm is rand index and modular index[19][20].

The percentage of correct decisions is defined by rand index (RI) i.e. how well the clustering algorithm has grouped the murders. RI for clustering can also be denoted Accuracy. The RI is computed as:

RI = TN + TP/TN + TP + FP + FN

Modularity is a cluster quality index that can be used to measure how well the clusters group and separate instances. It is based on the premise that the fraction of edges between nodes in a cluster should be higher than the expected fraction of edges between nodes in a cluster to indicate significant group structure, the modularity index maps onto [-1, 1].

The similarity between two pairs C and D is measured using jaccard index. A similarity value is computed between binary values, Where 0 represents that the two sets are identical. Fig 3. Shows the performance comparison of jaccard index.

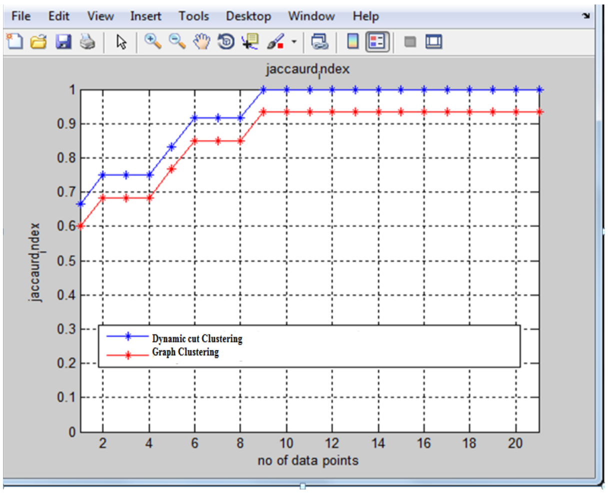

Fig.3. Jaccard Index

The correlation factor between time and space for the pairs of instances is computed by using mantel index. Fig.4. shows the performance comparison of Mantel index.

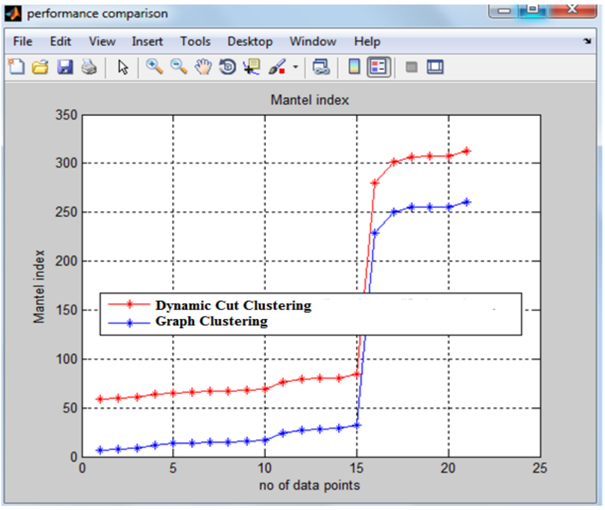

Fig.4. Mantel Index

4. EXPERIMENTAL RESULTS AND ANALYSIS

According to the dynamic clustering algorithm computation, time taken to calculate the entire graph is less when compared to cut clustering algorithm. The experimental results of our proposed system are based upon the accuracy of the clusters obtained. The graph below shows the comparison results of cut clustering algorithm and dynamic cut clustering algorithm based upon performance metrics such as rand index and modular index. The clustering result has obtained 90% of accuracy than the existing approach. The serial crime (i.e murder) can be viewed and displayed in maps by using QGIS tool. Fig (5&6) shows the comparison results of majority weighted oversampling with modified cut clustering algorithm and dynamic Graph cut clustering by rand index and modularity index.

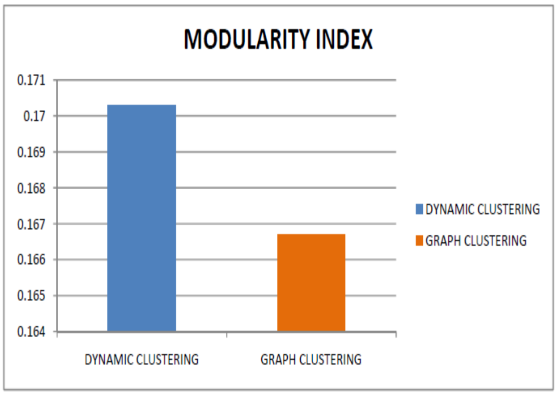

Fig.5. Modularity Index

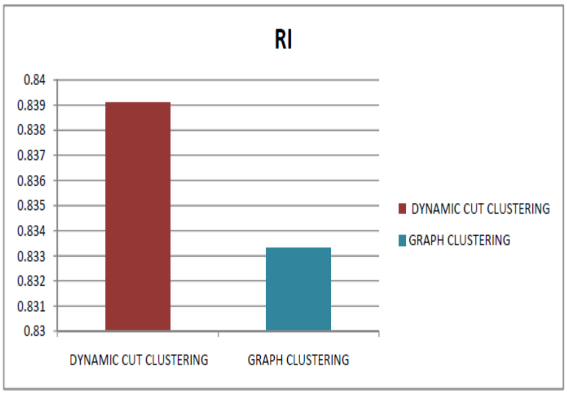

Fig.6. Rand Index

From the figures 5 and 6, it can be inferred that the Dynamic Cut Clustering outperforms Graph Clustering when compared with the performance metrics such as Modularity index and Rand index.

5. CONCLUSION

Hence in this work the integration of decision support system and GIS for managing and analyzing the crime data is presented. It is helpful for the law enforcement agencies to link the crime happened between different locations based upon the similarity coefficients such as spatial proximity, Modus Operandi(MO) etc., has been detected and analyzed. The serial crimes have been clustered by using dynamic cut clustering algorithm efficiently and also class imbalance problem have been used in order to avoid the unnecessary clusters which results in consumption of high computation time. We have identified two aspects for future work.

First, in this work only individual edge representations have been investigated. Researchers have found that in some cases, combinations of edge representation scores better performance than the individual edge representation. A distance index using combined crime characteristics, such as spatial and MO characteristics needs to be developed. Further work on dynamic minimum cut tree clustering may include a systematic comparison to other dynamic clustering.

6. REFERENCES

[1] Anton Borg , Martin Boldt, Niklas Lavesson, "Detecting serial residential burglaries using clustering", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),2014,Volume 1., 5252–5266.

[2] Krunal Patel, Paru Thakkar, Leena Patel, Chandresh Parekh, " GIS based Decision Support System for Crime Mapping, Analysis and identify Hotspot in Ahmadabad City", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2014, 4(1), 32-35.

[3] Nick Malleson, Alison Heppenstall, and Linda See, "Crime Reduction Through Simulation Burglary", Computers, environment and urban systems, 2010, 34(3), 236-250.

[4] Dawei Wang, Wei Ding, Tomasz Stepinski, Josue Salazar, Henry Lo, Melissa Morabito, "Crime Hotspot Mapping Using the Crime Related Factors–A Spatial Data Mining Approach", International Journal of Advanced Research in Applied Artificial Intelligence, 2012, 553-562.

[5] M. Ahmed, R. S. Salihu, "Spatiotemporal Pattern of Crime Using Geographic Information System (GIS)", American Journal of Engineering Research (AJER), 2013, 2, 51-58.

[6] G. Li, R. Haining , S. Richardson , N. Best, "Space–time variability in burglary risk:A Bayesian spatio temporal modelling approach", Spatial Statistics, 2014, 9, 180-191.

[7] Zhou, G., Lin, J., Zheng, W. "A web-based geographical information system for crime mapping and decision support", International conference on computational problem-solving (ICCP), 2012, 147–150.

[8] Guanli Huang, Guanhua Huang, "Dynamic Analysis for Geographical Profiling of Serial Cases Based on Bayesian-Time Series", Journal of Software, 2012, 7, 1242-1249.

[9] Jiaji Zhou, Le Liang, Long Chen, "Geographic Profiling Based on Multi-point Centrography With K-Means Clustering" , World Academy of Science, Engineering and Technology, 2012, 3.

[10] Nelson Devia, Richard Weber, "Generating crime data using agent-based simulation", International conference on computational problem-solving (ICCP), 2011, 2, 121-160.

[11] ShyamVaranNath, "Crime Pattern Detection Using Data Mining", American Journal of Engineering Research (AJER), 2009, 2.

[12] Yang Zhou Hong, Cheng Jeffrey, Xu Yu, "Graph Clustering Based on Structural/Attribute Similarities", in Proc. VLDB Endowment, 2009, 2, 718-729.

[13] Ahamed Shafeeq B. M, Binu V.S, "Spatial Patterns of Crimes in India using Data Mining Techniques", International conference on computational problem-solving (ICCP), 2008, 2, 121-160.

[14] Ohino, AlanT.Murray, "Accessing spatial pattern of crime in Lima", International conference on computational problem-solving (ICCP), 2005, 1.

[15] Giles C Oatley, Brian W Ewart, "Crime analsis software: Pins in Map and Bayes net prediction", American Journal of Engineering Research (AJER), 2004, 2.

[16] Lucy Markson, Jessica Woodhams, John W. Bond, "Linking serial residential burglary: Comparing the utility of modus operandi behaviours, geographical proximity, and temporal proximity". Journal of Investigative Psychology and Offender Profiling, 2010, 7, 91-107.

[17] Sivaranjani S, Sivakumari S, "Mitigating serial hotspots on crime data using interpolation method and graph measures", International Journal of Computer Applications, 2015, Vol.126.

[18] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, Melissa Morabito, "Crime Hotspot Mapping Using the Crime Related Factors–A Spatial Data Mining Approach", Applied Intelligence, 2013, 39, 772-781.

[19] Santos, J. M., Embrechts.M, "On the use of the adjusted rand index as a metric for evaluating supervised classification", Artificial neural networks–ICANN, 2009, 175– 185.

[20] William Adderley. R, "The Use of Data Mining Techniques in Crime Trend Analysis and Offender Profiling", International conference on computational problem-solving (ICCP), 2007, 1.