Analysis and Design of an Algorithm Using Data Mining Techniques for Matching and Predicting Crime

[1]Anshu Sharma, [2]Raman Kumar
[1,2]Dept. of CSE, DAV Institute of Engineering and Technology, Jalandhar, Punjab, India

Abstract
Crime analysis uses past crime data to predict future crime locations and times. Criminology is an area that focuses the scientific study of crime and criminal behavior. It is a process that aims to identify crime characteristics. It is one of the most important fields where the applications of data mining techniques can produce important results. The exponentially increasing amounts of data being generated each year make getting useful information from that data more and more critical.. Analysis of the data includes simple query and reporting, statistical analysis, more complex multidimensional analysis, and data mining. The wide range of data mining applications has made it an important field of research. Criminology is one of the most important fields for applying data mining. Criminology is a process that aims to identify crime patterns. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. An approach based on data mining techniques is discussed in this paper to extract important patterns from reports gathered from the city police department. The reports are written in simple plain text. The plain texts are converted into the format understandable by the tool. Then, exiting data mining techniques are applied to get patterns of crime data and a new algorithm is proposed to improve the accuracy of the crime pattern detection system. The various data mining techniques such as clustering and classification are used to get the patterns of crime data. This paper presents a new algorithm for K-Means using weighted approach. The results of proposed algorithm are compared with existing K-means clustering algorithm. The weighted approach proves to be better approach than existing K-means.

General Terms
Crime pattern detection.

Keywords
Data mining, criminology, clustering, classification.

I. Introduction

Crime analysis involves exploiting data about crimes to enable law enforcement to better apprehend criminals and prevent crimes. The manual extraction of patterns from data has occurred for centuries. The proliferation, ubiquity and increasing power of computer technology has increased data collection, storage and manipulations. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. So it can be defined as "Data Mining is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence but also database management" [1].

Crime is a behavior disorder that is integrated result of social, economical and environmental factors. Crimes are the social nuisance and cost our society dearly in several ways. Crime variables and crime matching are the two main components which are usually involved in crime analysis process. Crime variables are the parameters that can describe the crime characteristics uniquely. These are the main subject of crime analysis process. Crime matching is the process of assigning crimes or criminals to the previously solved or unsolved crime incidents.

Crime analysis basically includes leveraging a systematic approach for identifying, discovering and predicting crime patterns. The input of a crime analysis system consisted of the data and information gathered from city police department. The large volumes of crime related data existed in police departments and also the complexity of relationships between these kinds of data has forced the traditional crime analysis methods to become obsolete. These methods require large amount of resources and human effort to get pattern of data. Data mining overcomes the above problems by transforming the gathered data into useful knowledge to get the desired result. Once the data is transformed into the useful knowledge, various data mining techniques are applied such as clustering and classification technique so as to get patterns of data.

II. Literature Review

In the recent decade, a great deal of scientific researches and studies have been performed on crime data mining.

Yu et al. [2] discussed the approach to architecting datasets from original crime records. The dataset contains aggregated counts of crime and crime related events categorized by the police department. An ensemble of data mining classification techniques is employed to perform crime forecasting.

Phua et al. [3] proposes a multilayered detection system complemented with two additional layers: Communal detection and Spike detection. Communal detection finds real social relationships to reduce the suspicion score and is temper resistant to synthetic social relationships. Spike detection finds spikes in duplicates to increase the suspicion score and is probe resistant for attributes. It is an attribute oriented approach on a variable size set of attributes.

Xue et al. [4] analyzes criminal incidents as spatial choices processes. Spatial analysis processes can be used to discover the distribution of people behavior in space and time. Two adjusted spatial choice model that includes model of decision making processes are presented.

Hussain et al. [5] presents a micro simulation model that can be drawn out by interlinking the universal principles with the attributes of the individual for profiling of the criminal behavior. This paper elaborates the criminal behavior analysis by using data mining techniques.

Zhong et al. [6] presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relative and interesting information.

Malathi et al. [7] presents a clustering algorithm for crime data using data mining. They used MV Algorithm and Apriori Algorithm with some enhancement to aid in the process of filling missing values and identification of crime patterns. They applied these techniques to real crime data from a city police department.

Gupta et al. [8] highlights the existing system used by Indian police as e-governance initiative and also proposes an interactive query based interface as crime analysis tool to assist police in their activities. The authors used proposes interface to extract information from vast crime database mainatianed by National Crime Record Bureau (NCRB) and find crime hotspots using crime data mining techniques such as clustering etc.

Nath et al. [9] uses the clustering algorithm for data mining approach to help detect the crime patterns and speed up the process of solving crimes. Authors used K-Means clustering with some enhancements to aid in the process of identification of crime patterns. The used semi supervised learning technique for knowledge discovery from the crime records and to help increase the predictive accuracy.

Keyvanpour et al. [10] discussed an approach based on data mining technique to extract important entities from police narrative reports which are written in plain text. They have applied SOM clustering method in the scope of crime analysis and finally used the clustering results in order to perform crime matching process.

III. Proposed System Architecture

From the above literature review it has been concluded that in order to get crime patterns the two techniques are commonly used. K-Means clustering is used for the patterns and neural networks are used for classification. The improvement in K-Means clustering can produce better results than simple K-Means clustering algorithm. Improvement in clustering can improve the classifier performance. K-Means clustering algorithm can be improved by assigning weights to the seeds [11].

The proposed system comprises of two parts. In the first part we will use simple K-means clustering technique and RBF network to get the patterns of data. In the second part, we will propose a new algorithm using weighted approach and RBF network to get the patterns of data. The two techniques are then compared for accuracy using the performance parameters like precision, recall and F1 measure.

The steps for the first stage are as follows:

- Select the crime data.
- Apply K-Means clustering algorithm on the crime data and obtain the patterns of data.
- Now apply the attributes on the obtained clusters.
- Apply the RBF algorithm on the obtained clusters.
- Check the results of the approach used.

The steps for the second stage are as follows:

- Select the crime data.
- Apply Weighted K-Means Updated clustering algorithm on the crime data and obtain the patterns of data.
- Now apply the attributes on the obtained clusters.
- Apply the RBF algorithm on the obtained clusters.
- Check the results of the approach used.

Now compare the results of the above stages based on the following parameters:

- Precision
- Recall
- F1 Measure.

A. Steps for Data Transformation

Step 1: Collect the crime data

Experiments are conducted on real world crime data of Jalandhar city obtained from the office of Deputy Commissioner. The data obtained is in the Excel format as shown below:

The collected data is distributed into two categories, two third of crime data is used for training and remaining is used as testing.

1. Training Set

A training set is a set of data used in various areas of information science to discover potentially predictive relationships. Training sets are used in artificial intelligence, machine learning, genetic programming, intelligent systems, and statistics. In all these fields, a training set has much the same role and is often used in conjunction with a test set.

2. Test Set

A test set is a set of data used in various areas of information science to assess the strength and utility of a predictive relationship. Test sets are used in artificial intelligence, machine learning, genetic programming and statistics. In all these fields, a test set has much the same role.

Step 2: Select training sample.

The profile size can be chosen appropriately. In the case of existing system, large profile size increases the accuracy but the same time it increases the training time. But profile size doesn't affect the speed of operation of the RBF classifier.

Step 3: Data cleaning.

Select only the required attributes and discard others. Only the relevant information is used while the other information is discarded..

Step 4: Data Transformation

The next step is the data transformation. In this the data is transformed according to the needs and rules to obtain the hidden patterns and to discover the hidden relationship among the data. In this research to obtain the hidden pattern from the data, WEKA is used as a tool. In Weka (Waikato Environment for Knowledge Analysis) the algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

B. Simple K-Means Clustering Algorithm

The k-means [11] algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters [12].

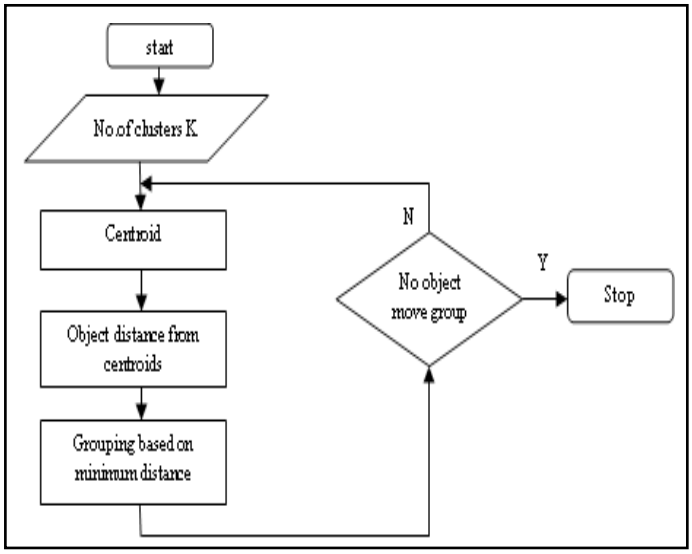


Fig. 1: K-means Flow Chart

C. RBF Networks

Radial functions are simply a class of functions. In principle, they could be employed in any sort of model (linear or nonlinear) and any sort of network (single-layer or multi-layer). In the field of mathematical modeling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control.

D. Proposed Algorithm

(Weighted K-means) WEIGHTEDKMEANSUPDATED (α, α_a, W(c), W(c)sal, cs,cl, W(c_{all}))

Here α represents the minimum acceptable variation between the clusters. α_a represents the changed variation after the process is applied and is initialized to 0. W(c) represents the cluster being examined for the variation and W(c)sal represent the variation between the smallest and the largest cluster. cs represents the cluster with the smallest weight and cl represents the cluster with the largest weight. W(call) represents the factor that all the clusters are in the acceptable format.

Step 1: Initialize the centers and α and set α_a==0.

Step 2: Calculate Average distance of centroid from datasets.

Step 3: Assign all data to the nearest centre and calculate the centre positions using Euclidian Distance.

Step 4: Calculate Weight of each Centroid based on no. of datasets assigned and average distance of the cetroid from datasets assigned.

Step 5: Check the fitness of each centre.

Step 6: Find cs and cl, the centre that has the smallest and the largest value of weight.

Step 7: If w(c)w(c)s a l < α,

- Assign the members of cl to W(c), and leave the rest of the members to cl.
- Recalculate the positions of cs and cl.

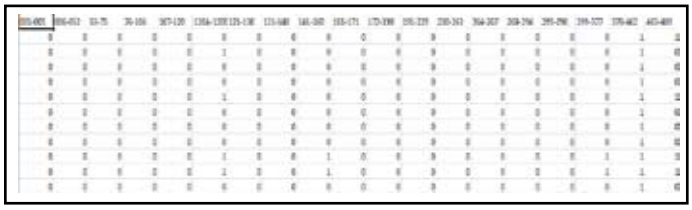
Step 8: Calculate weight of cs and cl.

Step 9: Update α_a according to updated centroids and repeat step (4) and (5) until W (C_{all}) is acceptable.

IV. Implementation Details

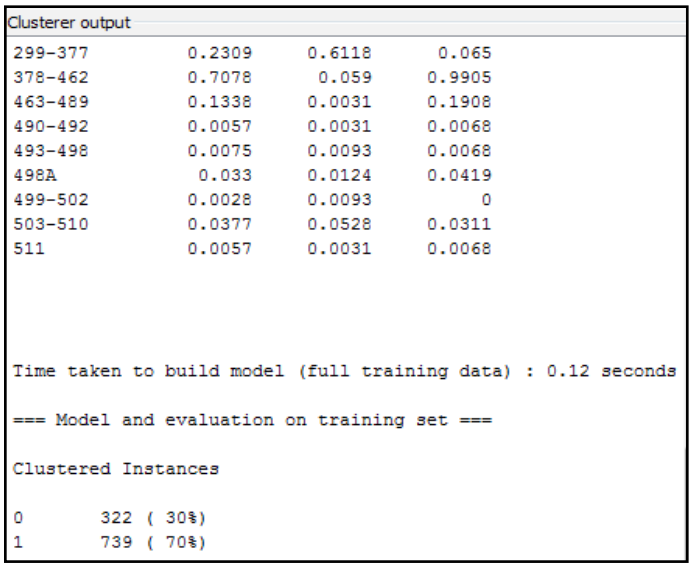
Implementation of the proposed system has three phases collection of data, training and testing and then validation. Experiments are conducted on real worlds crime data obtained from the office of deputy commissioner, Jalandhar. The collected data is converted into the format understandable by the tool.

Table 1: Sample Data



A. Implementation Based on K-Means Clustering and RBF Network

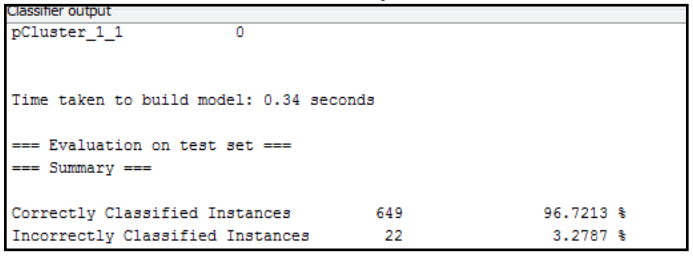
Once the data is transformed into the form understandable by the tool, the training and testing samples are selected. After than for the first case simple K-Means algorithm for clustering and RBF algorithm for classification is applied on the crime data obtained.



```
Clusterer output
299-377          0.2309      0.6118       0.065
378-462          0.7078       0.059      0.9905
463-489          0.1338      0.0031      0.1908
490-492          0.0057      0.0031      0.0068
493-498          0.0075      0.0093      0.0068
498A              0.033      0.0124      0.0419
499-502          0.0028      0.0093           0
503-510          0.0377      0.0528      0.0311
511              0.0057      0.0031      0.0068



Time taken to build model (full training data) : 0.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       322 ( 30%)
1       739 ( 70%)
```

Fig. 2: Implementation of Simple K-Means Clustering

The fig. 2 is the screen shot of the K-means clustering algorithm's output. Now this output will be used as input for the classification. Training of a system is done using RBF classification. Now when the complete data is clustered into two groups, for the detection of crime pattern, we pay attention to datasets belonging to cluster1. We train the system using RBF networks. The fig. 3 shows the result of classification. The accuracy obtained is 96.7213%



```
Classifier output
pCluster_1_1            0


Time taken to build model: 0.34 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         649               96.7213 %
Incorrectly Classified Instances        22                3.2787 %
```

Fig. 3: Classifier Output

B. Implementation based on Proposed Algorithm and RBF Network

Now the above same steps are applied on proposed algorithm i.e. weighted K-means updated. The fig. 3 is the screen shot of the weighted K-means updated clustering algorithm's output.

```
Clusterer output
493-498          0.0075      0.0085          0          0      0
498A              0.033      0.0128          0     0.0157      0
499-502          0.0028      0.0043          0          0
503-510          0.0377      0.0427      0.012          0      0
511              0.0057           0          0     0.0079      0



Time taken to build model (full training data) : 0.16 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       234 ( 22%)
1        83 (  8%)
2       127 ( 12%)
3       485 ( 46%)
4        65 (  6%)
5        67 (  6%)
```

Fig. 4: Implementation of Weighted K-Means Updated Clustering

Now this output will be used as input for the classification. Training of a system is done using RBF classification. Now when the complete data is clustered into two groups, for the detection of crime pattern, we pay attention to datasets belonging to cluster1. We train the system using RBF networks. The fig. 5 shows the result of classification. The accuracy obtained is 98.6587%

```
Classifier output

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         662               98.6587 %
Incorrectly Classified Instances         9                1.3413 %
```

Fig. 5: Classifier Output

V. Results

From the above implementation details it is clear that the proposed approach provides better results than the simple K-Means clustering. Both the approaches can be compared on the basis of performance parameters. The performance measures i.e. precision, recall and F1 measures for the existing K-means comes out to be 0.962, 0.967 and 0.967 whereas in case of proposed K-means algorithm these comes out to be 0.989, 0.987, 0.987.

Table 2: Comparison based on performance parameters

Sr No.	Technique Used	Precision	Recall	F1 Measure.
1	Simple K-Means and RBF Networks	0.962	0.967	0.967
2	Proposed K-Means and RBF Networks.	0.989	0.987	0.987

The performance can be drawn from the cost benefit curve. The value is greater than the threshold value. The rising curve shows that system is performing well under the given threshold curve.

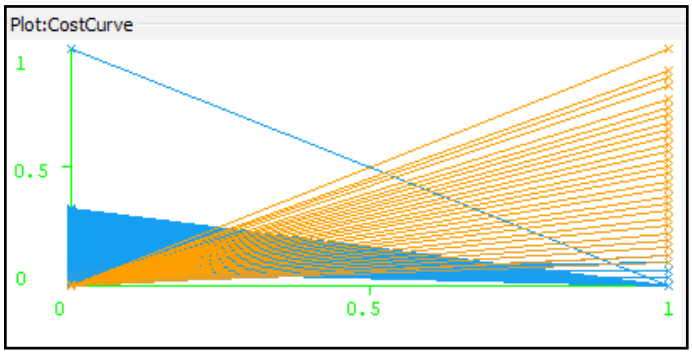


Fig. 6: Cost Curve for Proposed Approach

VI. Conclusion And Future Scope

Crime pattern detection is important in today's environment. The combination of facts such as extensive growth of terrorism, the vast financial possibilities of opening up in electronic trade, and the lack of truly secure system makes it an important field of research. The detection process should be adjustable to allow the system to deal with the constantly changing nature of crimes. This research proposes a modification in the K-Means algorithm using weighted approach. The result of the proposed approach comes better than the existing K-means approach. The accuracy of existing K-Means comes out to be about 96% whereas the accuracy using weighted K-means approach comes to be about 98%. This research is focused on the city level crime pattern detection. It can be advances to the state level or country level and the types of crimes.

References

[1] "IT Security Architecture", SecurityArchitecture.org, Jan, 2006.

[2] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding,"Crime forecasting using data mining techniques", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 03, pp. 779-786, 2012.

[3] Clifton Phua,"Resilient identity crime detection" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 03, pp. 533-546, 2012.

[4] Yefei Xue, Donald E. Brown,"A decision model for spatial site selection by criminals: A foundation for law enforcement decision support", IEEE Transactions on systems. Man. and Cybernetics-Part C: Applications and Reviews, Vol. 33, No. 01, pp. 78-85, 2003.

[5] K. Zakir Hussain, M. Durairaj, G. Rabia Jahani Farzana, "Criminal behavior analysis by using data mining techniques", IEEE Transactions on systems. Man. And Cybernetics-Part C: Applications and Reviews, Vol. 43, No. 05, pp 656-658, 2012.

[6] Ning Zhong, Yuefeng Li, Sheng- Tang W,"Effective pattern discovery for text mining" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 01, pp. 30-44, 2012.

[7] A. Malathi, Dr. S. Santhosh Baboo,"Algorithmic crime prediction model based on the analysis of crime clusters", Global Journal of Computer Science and technology Vol. 11, No. 11, pp 139-145, 2011.

[8] Manish Gupta, B. Chandra, M.P Gupta,"Crime data mining for Indian police information system", Computer Society of India, Vol. 40, No. 01, pp. 388-397, 2008.

[9] Shayam Varan Nath,"Crime pattern detection using data mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 09, pp. 41-44, 2010.

[10] Mohammad Reza Keyvanpour, Mostafa Javideh, Mahammad Reza Ebrahimi,"Detecting and investigating crime by means of data mining: A general crime matching framework" Procedia Computer Science, Vol. 03, pp. 872-880, 2011.
[11] Anshu Sharma, Raman Kumar,"The obligatory of an algorithm for matching and predicting crime- using data mining techniques", International journal of computer science and Technology (IJCST), Vol. 4, No. 2, pp. 289-292, 2013.
[12] Teknomo, Kardi,"K-Means Clustering Tutorials".
[13] [Online] Available: http://www.databases.about.com/od/datamining/a/kmeans.htm
[14] G. Holmes, A. Donkin, I.H. Witten,"Weka: A machine learning workbench", Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, 2009.
[15] [Online] Available: http://www.lifehacker.com/5237503/five-best-free-data-recovery-tools.
[16] [Online] Available: http://www.en.wikipedia.org/wiki/Weka_(machine_learning)
[17] S Benson Edwin Raj, A Annie Portia,"Analysis on credit card fraud detection methods", International Conference on Computer Communication and Electrical Technology ICCCET, IEEE, Vol. 02, No. 03, pp. 152-156, 2011.
[18] Sara Hajian, Josep Domingo-Ferrer, Antoni Martinez-Balleste,"Disrimination prevention in Dara Mining for intrusion and crime detection", IEEE, in proceeding of: Computational Intelligence in Cyber Security (CICS), 2011.
[19] Li Cunhua, Hu Yun, Zhong Zhaoman,"An event ontology construction approach to Web Crime Mining", Seventh International Conference on Fuzzy Systems and knowledge Discovery, IEEE, pp. 2441-2445, 2010.
[20] Shiguo Wang,"A comprehensive survey of data mining-based accounting-fraud detection research", International Conference on Intelligent Computation Technology and Automation, IEEE, pp. 50-53, 2010.
[21] Endy, Charles Lim, Kho I Eng, Anto Satriyo Nugroho, "Implementation of intelligent searching using self organizing maps for web mining used in document containing information in relation to cyber terrorism". International Conference on Advances in Computing, Control and Telecommunication Technologies, IEEE, pp 195-197, 2010.
[22] Bin Liu, Shu-Gui Cao, Xiao-Li Jia, Zhao-Hua Zhi, "Data Mining In Ditributed Data Environment", International Conference on Machine Learning and Cybernetics, IEEE, pp. 421-426, 2010.
[23] Hamidah Jantan, Abdul Razak Hamdan, Zulaiha Ali Othman,"Talent Knowledge Acquisition using Data Mining Classification Techniques", Conference on Data Mining and Optimization (DMO), IEEE, pp. 32-37, 2011.
[24] Soumadip Ghosh, Amitava Nag, Debasish Biswas, Jyoti Prakash Singh,"Weather Data Mining Using Artificial Neural Network", IEEE, pp. 192-195, 2011.
[25] Bo Wu, wandong Cai, Yongjun Li,"Association Analysis and case study framework based on the name distinction", International Conference on Computer Application and System Modelling (ICCASM 2010), IEEE, Vol. 04, pp. 285-289, 2010.
[26] Sherly K.K, R. Nedunchezhian,"BOAT adaptive credit card fraud detection system", International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, pp. 01-07, 2010.
[27] Shiguo Wang,"A comprehensive survey of data mining-based accounting-fraud detection research", International Conference on Intelligent Computation Technology and Automation, IEEE, Vol. 01, pp. 50-53, 2010.