Improving Accuracy using different Data Mining Algorithms

Pooja Pandey
Research Scholar
CSE & IT Department
Baba Banda Singh Bahadur
Engineering College, Fatehgarh Sahib

Ishpreet Singh
Assistant Professor
CSE & IT Department
Baba Banda Singh Bahadur
Engineering College, Fatehgarh Sahib

ABSTRACT
Mining large data set is an important issue to deal with as data is growing as the field grows. Today, crime rate is a menace that each country faces. With the increase in crime rate the data is increasing and it is such a critical field that accuracy is important at the same time. This paper shows the comparison in the results between clustering and the classification. K means is used in clustering and in classification decision tree is used. The process of applying decision tree and clustering one after the other is used CDDT(clustered data of decision tree) in this paper.

General Terms
Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords
CDDT, clustering, classification,decision tree

1. INTRODUCTION
Data mining *is* the extraction of hidden predictive information from large databases. Its tools predict future trends and behaviours and different organizations to focus on the most important information in their data warehouses. This technique can be implemented rapidly on existing software and can be integrated with the other products. Most commonly used techniques may include decision tree, artificial neural network, genetic algorithm, nearest neighbour method, rule induction etc. In this paper focus will be on clustering and classification as their outputs will be compared.

Clustering in data mining is very important to discover distribution patterns. Clustering is a method that organizes data into different classes of similar characteristics. It is the way of searching hidden patterns. Clustering is a little similar to classification. classification categorize the similar data into same group. In this paper comparison of two different algorithms of data mining will be given. The comparison will be between clustering and classification. In clustering k means algorithm is used while decision tree classifier is used in classification. Firstly the data is mined using k-means algorithm and clusters are made using that. After that decision tree classifier is applied on the same data again and then k-means algorithm will be applied again. The accuracy of the clusters of the simple k-means algorithm and the clustered data of decision tree (CDDT) classifier will be compared. This paper will include five parts. In the second part standard k means algorithm will be explained , in the next parts decision tree classifier will b explained , the fourth part will give the experimental results and the last portion will give the conclusion

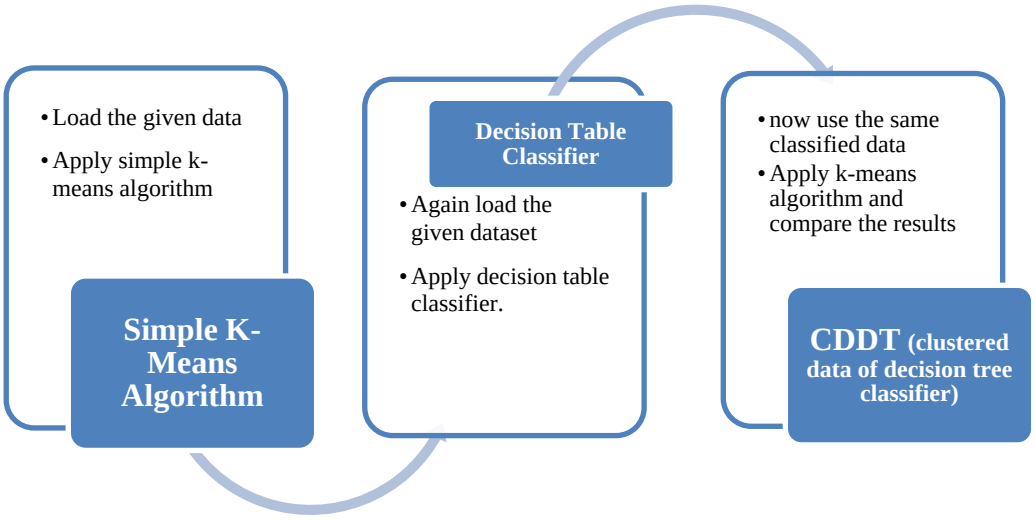


Fig.1.1 flow of process

2. STANDARD K MEANS CLUSTERING ALGORITHM
A. The process of k means algorithm
K means is simple and widely used technique of clustering. It is completely based partitioning methodology. It partitions n-data items into k groups where k indicates number of clusters specified by the user. Clusters are formed such that each item in the cluster has minimum distance from the centroid. For calculating distance between item and the centroid, k means algorithm uses the Euclidean distance measurement. It aims to

minimize the sum of squared distances between all points and the cluster center. This procedure consists of following steps:

Input: K: the number of desired clusters.

Output: A set of k clusters

Algorithm:

1) Randomly select k objects as initial centroids naming(m_1,m_2,m_3)

2) Calculate the distance between each object O_i and

each centroid, then assign each object to its nearest cluster center,

formula for calculating distance as:

$d(O_i,M_i) = \sqrt{\sum_{j=1}^{d}\left(O_{i1-M_{j1}}\right)}$, i=l.....N; j=l.....k;

d (O_i, M_j) is the distance between data i and cluster j;

3) Calculate the mean in order to create the new cluster centers

$M_i = \frac{1}{Z_i}\sum_{j=1}^{Zi} x_{ij}$, i=l.....k; Z_i is the number of samples of current cluster i;

4) Repeat step 2 and 3 until the criterion function E converged,

return (m1, m2.....mk). Algorithm terminates.

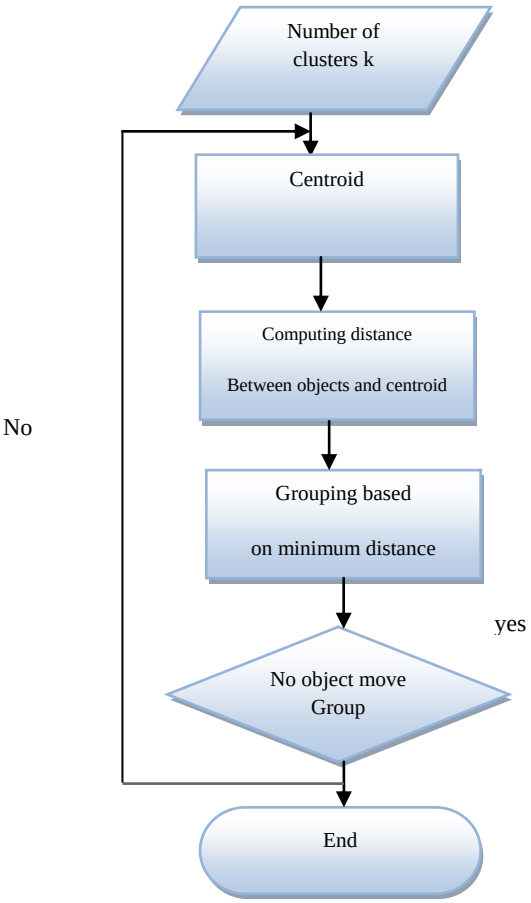


Fig. 2.1 The K- means algorithm process

3. DECISION TABLE CLASSIFIER

Classification techniques are most suited for predicting data sets with binary or nominal categories. It is a systematic approach for building classification models from an input data set. Decision table classifier, rule based classifiers, neural networks, support vector machines and naive byes classifier are its examples.

Decision tree is simple yet widely used classifier. Explaining classification with decision tree become simpler using an example. As in this paper crime analysis is done, so suppose an unknown dead body is found and it is damaged completely .It is needed to find whether the body is of male or of female. One approach is to pose a series of questions about the characteristics of gender difference. On question may be whether the body is has long hair or short. But it may not be sufficient. So next question may be about their body parts. Each time we get an answer, a follow up question is asked until we get the desired conclusion. The tree has three type of nodes.

- **A root node** that has no incoming edges and zero or more outgoing edges.
- **Internal node** each of which has exactly one incoming edge and two or more outgoing edges.
- **Leaf or terminal node** each of which has exactly one incoming edge and zero outgoing edges.

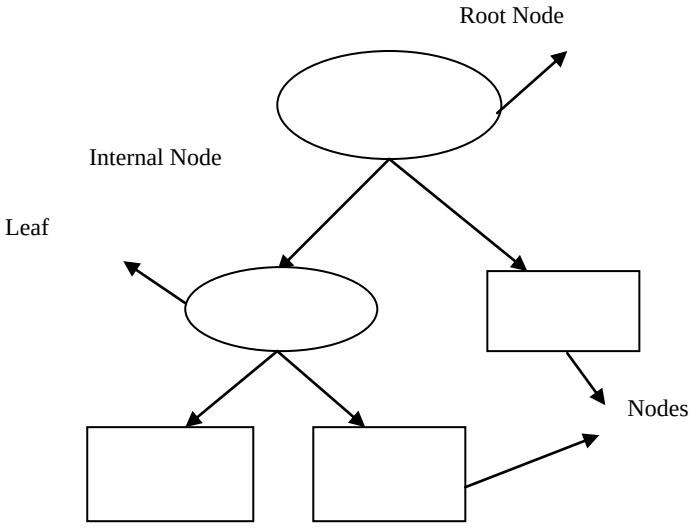


Fig.3.1 Decision Table Classifier

The main algorithm that generates the decision table:

Algorithm GAC (generating decision table by grouping and counting)

(minsup, minconf: real)

1 begin

2 best Split Attr = Best Split Attr;

3 cand D Table = Candidate D Table (bestSplitAttr);

4 decision Table = Prun D Table (cand D Table, minsup, minconf);

5 end.

minsup and minconf as input parameters.

4. EXPERIMENTAL RESULTS

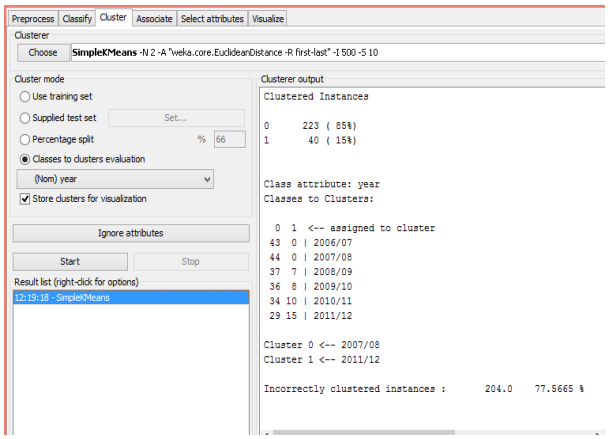
Crime rate is increasing at very fast speed, with that the data also becomes vast. So it is really difficult to handle such a big data manually. Therefore in this paper crime data set has been taken to determine the occurrence of crime. The sample dataset is shown below:

```
@relation data@attribute year
{2006/07,2007/08,2008/09,2009/10,2010/11,2011/12}@attribute
Homicide numeric@attribute 'Attempted murder' numeric@attribute
'Child destruction' numeric@attribute 'Causing death by
dangerous    or careless driving' {..,8.0,5.0,0.0,10.0,9.0,-
1.0,2.0,12.0,4.0,3.0,17.0,13.0,1.0,38.0,6.0,7.0,11.0,15.0,21.0,1
9.0,46.0,20.0,27.0,29.0,14.0}@data2006/07,8,7,0,..
2006/07,4,4,0,..
```

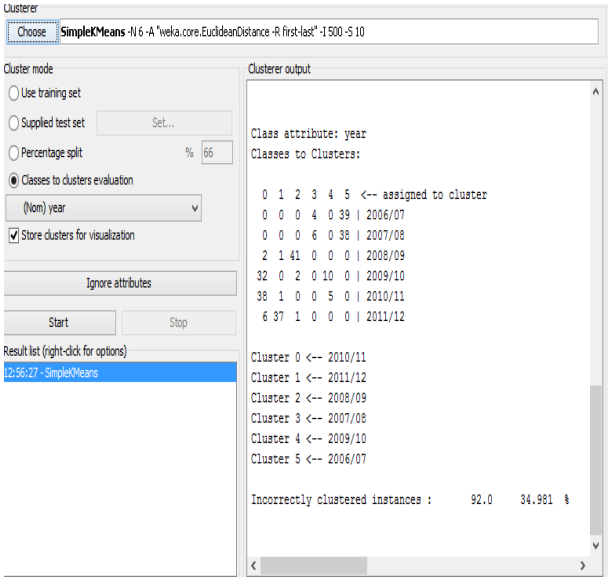
Fig.4.1 crime dataset

This data set include crimes of different years i.e. from 2006 to 2012. The crimes include in this data set are Homicide, Attempted murder, Child destruction, Causing death by dangerous or careless driving. The comparison will be done on the basis of accuracy. Accuracy here means the subtracting incorrectly clustered instances from 100. It will cluster the crimes that in which year which crime rate is high so that preventive measures can be taken accordingly.

In this paper firstly it is applied the K means clustering algorithm. The data was clustered on yearly bases. The incorrect cluster instances of simple k means algorithm were 204 and the accuracy was 22.4335% whereas when two algorithms were combined i.e. firstly the decision table classifier and then the k means clustering (CDDT) the results improved. The incorrectly clustered instances in CDDT were 90 and the accuracy was 65.019%.



Results of simple k means clustering



Results of decision table classifier and k means

5. ACKNOWLEDGMENTS

The authors are highly grateful and thankful to the Dr. Baljit Singh Khehra(Head Of CSE & IT Department) of the Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib and to the college also

6. CONCLUSION AND FUTURE SCOPE

As from the results it can be seen that (CDDT) clustering after applying the classifier can give the better results as compared to simple clustering technique or k-means algorithm. Accuracy is the serious issue in any data field. These results are showing the better accuracy and the improvised results. For the future scope results can also be compared using different distance calculating methods i.e. Manhattan distance and euclidean distance.

7. REFERENCES

[1] Rajeswari, K., Acharya, O., Sharma, M., Kopnar,M., & Karandikar, K.” Improvement in k-Means Clustering Algorithm Using Data Clustering”, In Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on,vol.3, no.15, pp. 367-369, IEEE.

[2] Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm Shi Na College of Information Engineering, Capital Normal

[3] Lima, M. F., Zarpelao, B. B., Sampaio, L. D., Rodrigues, J. J., Abrao, T., & Proença Jr, M. L.” Detection using baseline and K-means clustering”, In Software, Telecommunications and Computer Networks (softcom), 2010International Conference on, vol.3, no.5 pp. 305-309, IEEE.

[4] Ren, Q., & Zhuo, X. “ Application of an improved k-means algorithm in gene expression data analysis” In Systems Biology (ISB), 2011 International Conference on, pp. 87-91, IEEE.

[5] Z. Pawlak “Information systems - theoretical foundations”, Information Systems Journal 1981, Vol. 6, pp.205-218

[6] Y. Qiao, K. Zhong, H.-A.Wang and X. Li, “Developing event-condition-action rules in real-time active

database", Proceedings of the 2007 ACM symposium on Applied computing, ACM, New York, pp.511-516

[7] Z.W. Ra´s, A. Dardzi´nska, "Action rules discovery, a new simplified strategy, Foundations of Intelligent Systems", 2006 LNAI, No. 4203, Springer, pp.445-453

[8] Z.W. Ra´s, A. Tzacheva, L.-S. Tsay, O. G¨urdal, "Mining for interesting action rules", 2005 Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2005), Compiegne University of Technology, France, 2005, pp.187-193

[9] Wang, H., Qi, J., Zheng, W., & Wang, M. "Balance K-means algorithm. In Computational Intelligence and Software Engineering," Cise 2009 International Conference on, pp. 1-3, IEEE.

[10] Esteves, R. M., Hacker, T., & Rong, C. "Competitive k-means, a new accurate and distributed k-means algorithm for large datasets" In Cloud Computing Technology and Science (cloudcom), 2013 IEEE 5th International Conference on ,Vol. 1, pp. 17-24, IEEE.

[11] Tian, L., & Jianwen, W. "Research on network intrusion detection system based on improved k-means clustering algorithm", In computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on Vol. 1, pp. 76-79, IEEE.

[12] Singh, G., Antony, D. A., & Leavline, E. J" Data mining in network security-techniques & tools: a research perspective", Journal of theoretical & applied

[13] Yang, Q., & Wu, X. "10 challenging problems in data mining research" International Journal of InformationTechnology & Decision Making", vol.5, no.4,pp.597-604.

[14] Chen, C. H., Tseng, V. S., & Hong, T. P. ,"Cluster-based evaluation in fuzzy-genetic data mining. Fuzzy Systems", IEEE Transactions on, vol. 1, no.16,pp. 249-262.

[15] Liao, S. H., Chu, P. H., & Hsiao, P. Y," Data miningtechniques and applications–A decade review from 2000 to 2011", Expert Systems with Applications, vol.12,no.39, pp.11303-11311.

[16] Balabantaray, R. C., Sarma, C., & Jha, M. (2015).Document Clustering using K-Means and K Medoids.Arxiv preprint arxiv:1502.07938.

[17] Sujatha, M. S., & Sona, M. A. S.,"New fast k-means clustering algorithm using modified centroid selection method", Ininternational Journal of EngineeringResearch and Technology ,Vol. 2, No. 2 ,February-2013.

[18] Brar, R., & Sharma, N., "A NovelDensity Based KMeans Clustering Algorithm for Intrusion Detection", Journal of Network Communications and EmergingTechnologies (JNCET) www. Jncet. Org, vol.3, no.7

[19] W. Zhao, H. Ma, and Q. He, "Parallel K-Means Clustering Based on MapReduce," vol. 5931, Springer Berlin / Heidelberg, 2009, pp. 674– 679.

[20] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," Proc. VLDB Endow., vol. 5, no. 7, pp. 622–633, 2012.

[21] M.V.B.T.Santhi,V.R.N.S.S.V.SaiLeela,P. U.Anitha,D.Nagamalleswari" Enhancing K-Means Clustering Algorithm" International Journal on Computer Science And Technology(IJCST) Vol 2,Issue 4,Oct-Dec 2011

[22] Z.W. Ra´s, A. Wieczorkowska, "Action-Rules: How to increase profit of a company,in Principles of Data Mining and Knowledge Discovery",(2000) Proceedings of PKDD 2000, Lyon, France, LNAI, No. 1910, Springer, pp.587-592

[23] Z. Ra´s, E. Wyrzykowska, H. Wasyluk,"ARAS: Action rules discovery based on agglomerative strategy, in Mining Complex Data", Post-Proceedings of 2007 ECML/PKDD Third International Workshop (MCD 2007), LNAI, Vol. 4944, Springer, 2008, pp.196-208

[24] L.-S. Tsay, Z.W. Ra´s (2006), "Action rules discovery system DEAR3, in Foundations of Intelligent Systems", Proceedings of ISMIS 2006, Bari, Italy, LNAI, No. 4203, Springer, pp.483-492