International Journal of Engineering Research and Development
eISSN : 2278-067X, pISSN : 2278-800X, www.ijerd.com
Volume 4, Issue 8 (November 2012), PP. 69-72

Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data

Mahendra Tiwari[1] Randhir Singh[2]
[1]*Research Scholar (UPRTOU, Allahabad)*
[2]*Asstt. Professor (UIM,Allahabad)*

Abstract:—The data clustering is a big problem in a wide variety of different areas,like pattern recognition & bio-informatics. Clustering is a data description method in data mining which collects most similar data . The purpose is to organize a collection of data items in to clusters, such that items within a cluster are more similar to each other than they are in other clusters. In this paper, we use k-means & k-medoid clustering algorithm and compare the performance evaluation of both with IRIS data on the basis of time and space complexity.
Keywords:—clusters, pattern recognition, k-medoid.

I. INTRODUCTION

Cluster analysis is the organization of a collection of patterns in to clusters based on similarity. Intuitively ,patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster.
Data clustering is an unsupervised learning process, it does not need a labeled data set as training data, but the performance of the data clustering algorithm is often much poorer. Although the data classification has better performance , it needs a labeled data set as training data & labeled for the classification is often very difficult to obtain . In the case of clustering ,the problem is to group a given collection of unlabeled patterns in to meaningful clusters. In a sense, labels are data driven, that is they are obtained solely from the data.

II. CLUSTERING

Clustering methods are mainly suitable for investigation of interrelationships between samples to make a preliminary assessment of the sample structure. Clustering techniques are required because it is very difficult for humans to intuitively understand data in a high-dimensional space.

Partition clustering:

A partitioning method constructs k(k<n) clusters of n data sets (objects) where each cluster is also known as a partition. It classifies the data in to k groups while satisfying following conditions.

- Each partition should have at least one object .
- Each object should belong to exactly one group.

The number of partitions to be constructed (k) this type of clustering method creates an initial partition. Then it moves the object from one group to another using iterative relocation technique to find the global optimal partition. A good partitioning is one in which distance between objects in the same partition is small(related to each other) whereas the distance between objects of different partitions is large(they are very different from each other).
k-means algorithm:-Each cluster is represented by the mean value of the objects in the cluster.
K-medoid algorithm:- Each cluster is represented by one of the objects located near the center of the cluster.

K-means algorithm:

(I) Choose k cluster centers to coincide with k randomly chosen patterns or k randomly defined points inside the hyper volume containing the pattern set.
(II) Assign each pattern to the closest cluster center.
(III) Recomputed the cluster centers using the current cluster membership.
(IV) If a convergence criterion is not met step 2. Typical convergence criteria are: no reassignment of patternsto new cluster center, or minimal decrease in squared error.

K-medoid method:
It is representative object-based technique. In this method we pick actual objects to represent cluster instead of taking the mean value of the objects in a cluster as a reference points.
PAM(partitioning around method) was one of the first k-medoid algorithm. The basic strategy of this algorithm is as follows:-

- Intially find a representative object for each cluster.
- Then every remaining object is clustered with the representative object to which it is the most similar.
- Then iteratively replace one of the medoids by a non-medoid as long as the “quality” of the clustering is imposed.

III. EVALUATION/INVESTIGATION STRATEGY

3.1 H/W tools:

We conduct our evaluation on Pentium 4 Processor platform which consist of 512 MB memory, Linux enterprise server operating system, a 40GB memory, & 1024kbL1 cache.

3.2 S/W tool:

The implementation of K-means and k-medoid algorithm is done on Iris data in Mat lab. The data contains 3 classes of 150 instances each. Where each class refers to a type of IRIS plant. One class is linearly separable from other two, the letter are not linearly separable from each other.

1.3 Input data sets:-

Input data is an integral part of data mining applications. The data used in experiment is either real-world data obtained from UCI data repository and widely accepted during evaluation dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset This dataset was chosen because it have different characteristics and have addressed different areas.

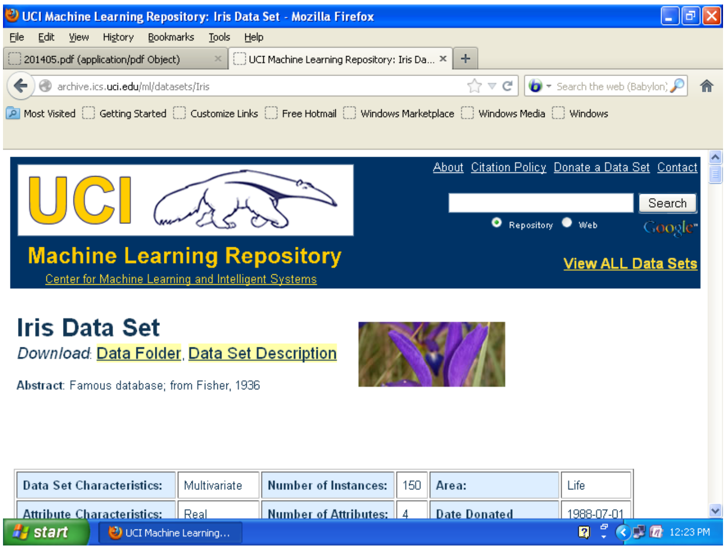

Fig.1 *Iris data set*

Relevant Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

- Predicted attribute: class of iris plant.
- This is an exceedingly simple domain.
- This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick@espeedaz.net)

The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature.
The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.
Number of Instances: 150 (50 in each of three classes)
Number of Attributes: 4 numeric, predictive attributes and the class
Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

Missing Attribute Values: None
Summary Statistics:

	Min	Max	Mean	SD	Class Correlation
sepal length:	4.3	7.9	5.84	0.83	0.7826
sepal width:	2.0	4.4	3.05	0.43	-0.4194
petal length:	1.0	6.9	3.76	1.76	0.9490 (high!)
petal width:	0.1	2.5	1.20	0.76	0.9565 (high!)

Class Distribution: 33.3% for each of 3 classes.

1.4 Experimental result and Discussion:-

To evaluate the selected tool using the given dataset, several experiments are conducted. For evaluation purpose, time and space complexities of k-means and k-medoid are measured. The time complexity of k-means is O(I*k*m*n) and time complexity of K-medoid is O(ik(n-k)2). Now assume that n=100,d=3,i=20 and number of clusters varying. We get the result and displayed in tables.

The space requirement for k-means are modest because only the data points and centroid are stored. Specifically, the storage required is O((m+k)n), where m is the number of points and n is the number of attributes. In particular the time required is O(i*k*m*n), where I is the number of iterations required for convergence, m is the number of points, k is number of clusters. K-medoid is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a k-means. K-medoid is relatively more costly ,complexity is O(ik(n-k)2), where I is the total number of iterations, k is total number of clusters, and n is the number of objects.

Table 1: Time complexity when number of cluster varying

No. of clusters	k-means time complexity	k-medoid time complexity
1	2000	2000
2	10000	6000
3	25000	9000
4	45000	11000

Table 2: Time complexity when number of iterations varying

No. iterations	k-means time complexity	k-medoid time complexity
5	5000	4000
10	10000	5000
15	15000	8000
20	25000	10000

Table 3: Space complexity when number of clusters varying

No. of cluster	k-means space complexity	k-medoid space complexity
10	500	7
15	700	8
20	900	9
25	1100	12

IV. CONCLUSION

From the above investigation, it can be said that partitioning based clustering methods are suitable for spherical shaped clusters in small to medium size data set.

k-means and k-medoids both methods find out clusters from the given data. The advantage of k-means is its low computation cost,and drawback is sensitive to nosy data while k-medoid has high computation cost and not sensitive to noisy data.

The time complexity of k means is O(i*k*m*n) and time complexity of k-medoid is O(ik(n-k)2.

REFERENCES

[1]. Peter M. chen and David A.(1993), storage performance-metrics and bench marks, Proceeding of the IEEE, 81:1-33

[2]. M.Chen, J. Han, and P. Yu. (1996) Data Mining Techniques for marketing, Sales, and Customer Support. IEEE Transactions on Knowledge and Data Eng., 8(6)

[3]. Agrawal R, Mehta M., Shafer J., Srikant R., Aming (1996) A the Quest on Knowledge discovery and Data Mining, pp. 244-249..

[4]. Chaudhuri, S.Dayal, U. (1997) An Overview of Data Warehousing and OLAP Technology. SIGMOD Record 26(1) 65-74

[5]. *John F. Elder et all, (1998)* A Comparison of Leading Data Mining Tools, Fourth International Conference on Knowledge Discovery & Data Mining

[6]. C. Ling and C. Li, (1998) "Data mining for direct marketing: Problem and solutions," in Proc, of the 4th international Conference on Knowledge Discovery & Data Mining, pp. 73-79

[7]. John, F., Elder iv and Dean W.(1998) A comparison of leading data mining tools, fourth International conference on Knowledge discovery and data mining pp.1-31

[8]. Michel A., et all (1998), Evaluation of fourteen desktop data mining tools , pp 1-6

[9]. Kleissner, C.(1998),, data mining for the enterprise, Proceeding of the 31st annual Hawaii International conference on system science

[10]. Brijs, T., Swinnen, G.,(1999), using association rules for product assortment decisions: A case study., Data Mining and knowledge discovery 254.

[11]. Goebel M., L. Grvenwald(1999), A survey of data mining & knowledge discovery software tools, SIGKDD,vol 1, 1

[12]. Rabinovitch, L. (1999),America's first department store mines customer data. Direct marketing (62).

[13]. Grossman, R., S. Kasif(1999), Data mining research: opportunities and challenges. A report of three NSF workshops on mining large, massive and distributed data, pp 1-11.

[14]. Dhond A. et all (2000), data mining techniques for optimizing inventories for electronic commerce. Data Mining & Knowledge Discovery 480-486

[15]. Jain AK, Duin RPW(2000), statistical pattern recognition: a review, IEEE trans pattern anal mach Intell 22:4-36

[16]. Zhang, G.(2000), Neural network for classification: a survey, IEEE Transaction on system, man & cybernetics, part c 30(4).

[17]. X.Hu, (2002) "Comparison of classification methods for customer attrition analysis" in Proc, of the Third International Conference on Rough Sets and Current Trends in Computing, Springer, pp. 4897-492.

[18]. A. Kusiak, (2002) Data Mining and Decision making, in B.V. Dasarathy (Ed.). Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology TV, ol. 4730, SPIE, Orlando, FL, pp. 155-165.

[19]. Rygielski. D.,(2002) , data mining techniques for customer relationship management, Technology in society 24.

[20]. Anderson, J. (2002), Enhanced Decision Making using Data Mining: Applications for Retailers, Journal of Textile and Apparel, vol 2,issue 3

[21]. Madden, M.(2003), The performance of Bayesian Network classifiers constructed using different techniques, Proceeding of European conference on machine learning, workshop on probabilistic graphical models for classification, pp 59-70.

[22]. Giraud, C., Povel, O.,(2003), characterizing data mining software, Intell Data anal 7:181-192

[23]. Ahmed, S.(2004), applications of data mining in retail business, Proceeding of the International conference on Information Technology : coding & computing.

[24]. Bhasin M.L. (2006) Data Mining: A Competitive Tool in the Banking and Retail Industries, The Chartered Accountant

[25]. Sreejit, Dr. Jagathy Raj V. P. (2007), Organized Retail Market Boom and the Indian Society, *International Marketing Conference on Marketing & Society IIMK , 8-1*

[26]. T. Bradlow et all, (2007) Organized Retail Market Boom and the Indian Society, *International Marketing Conference on Marketing & Society IIMK, 8-10*

[27]. Michel. C. (2007), Bridging the Gap between Data Mining and Decision Support towards better DM-DS integration, International Joint Conference on Neural Networks, Meta-Learning Workshop

[28]. Wang j. et all (2008), a comparison and scenario analysis of leading data mining software, Int. J Knowl Manage

[29]. Chaoji V.(2008), An integrated generic approach to pattern mining: Data mining template library, Springer

[30]. Hen L., S. Lee(2008), performance analysis of data mining tools cumulating with a proposed data mining middleware, Journal of Computer Science

[31]. Bitterer, A., (2009), open –source business intelligence tool production deployment will grow five fold through2010, Gartner RAS research note G00171189.

[32]. Phyu T.(2009), Survey of classification techniques in data mining, Proceedings of the International Multiconference of Engineering and Computer Scientist(IMECS), vol 1

[33]. Pramod S., O. Vyas(2010), Performance evaluation of some online association rule mining algorithms for sorted & unsorted datasets, International Journal of Computer Applications, vol 2,no. 6

[34]. *Mutanen. T et all,* (2010), Data Mining for Business Applications , Customer churn prediction – a case study in retail banking , Frontiers in Artificial Intelligence and Applications, Vol 218

[35]. Prof. Das G. (2010), A Comparative study on the consumer behavior in the Indian organized Retail Apparel Market, ITARC

[36]. Velmurugan T., T. Santhanam(2010), performance evaluation of k-means & fuzzy c-means clustering algorithm for statistical distribution of input data points., European Journal of Scientific Research, vol 46 no. 3

[37]. Jayaprakash et all, performance characteristics of data mining applications using minebench, National Science Foundation (NSF).

[38]. Osama A. Abbas(2008),Comparison between data clustering algorithm, The International Arab journal of Information Technology, vol 5, N0. 3

[39]. www.eecs.northwestern.ed/~yingliu/papers/pdcs.pdf

[40]. *www.ics.**uci**.edu/~mlearn/*