Australian Journal of Basic and Applied Sciences, 5(8): 594-599, 2011
ISSN 1991-8178

A New Method for Clustering Using Learning Automata

[1]Morteza Saberi Kamarposhti, [2]Ehsan Mousavian Jandaghi, [3]Kamal Jadidy Aval

[1]Department of Computer Science, Islamic Azad University, Firoozkooh Branch, Firoozkooh, Iran.
[2]Department of Computer Science, Islamic Azad University, Mashhad Branch, Mashhad, Iran.
[3]Department of Computer Science, Islamic Azad University, Firoozkooh Branch, Firoozkooh, Iran.

Abstract: Development in storage and retrieval of information and the fast growth of their application in some areas like internet search engines, digital photography and video surveillance has caused a huge amount of data with a great volume and in big size. Different methods are proposed to cluster this data. CURE is one of the common methods for this approach. In this paper, an improved hierarchical method is proposed based on CURE clustering using learning automata. The results of this method are showing big improvements in comparison to basic CURE method. Comparing to similar methods, this method has also better performance.

Key words: *Clustering; CURE Clustering; Learning Automata; Pattern Recognition.*

INTRODUCTION

Development in storage and retrieval of information and the fast growth of their application in some areas like internet search engines, digital photography and video surveillance has caused a huge amount of data with a great volume and in big size. It is estimated that about 281 exabytes are used in the world in 2007 and this number will become ten times more on 2011 (one exabyte is approximately 10^{18} bytes) (Gantz, John F., 2008). Most of this data will be stored digitally and this leads to a great potential in developing automated data analysis, clustering and retrieval techniques.

In pattern recognition, data analysis is related to predictive modeling: several training data are fed into the system and the system is going to predict the behavior of unseen data. This is called Learning. There is often an obvious difference between learning issues that are divided into two categories:
Supervised (Classification): Training data is labeled.
(Clustering): Training data is not labeled. (Duda, 2001)

Clustering is a more difficult than classification. Now, there is a lot of interest on an intermediate concept named semi-supervised learning (Chapelle, 2006). In semi-supervised classification, a little portion of data is labeled. The other unlabeled data is not discarded and is used in learning process.

The goal of data clustering (also known as data analysis) is detection of natural groups in a set of patterns, points or objects. Webmaster (Merriam-Webster, 2008) is defining the cluster analysis as "A statistical classification technique for detecting that each member of population will be in which group of several groups due to the multiple properties". An example of clustering is shown in figure 1 (Anil K., 2010).

The final goal is finding an automated algorithm to detect natural groups (figure. 2. b) in the data space without labeling (figure. 2.a).
Generally clustering is being used in these three aspects (Anil K., 2010):

1. Underlying structure: to increase overview of data, hypothesis generation and anomaly detection.
2. Natural classification: to detect similarity degree about organisms or forms.
3. Compression: as a method of organizing data and summarization it to sample data.

Previous Work:

Different people from different fields have been working on development of clustering methods. Psychologists, sociologists, biologists, statists, mathematicians, engineers, computer scientists, medical researchers and others that are gathering real data, are collaborating to develop clustering methods. As stated in (JSTOR, 2009) "data clustering appeared for the first time in a paper about anthropological data". Data clustering was also known as clumping, Q-Analysis and Taxonomy according to the field it was applied (Jain, 1988). There are several books about clustering (Jain, 1988; Sokal, 1963; Anderberg, 1973; Hartigan, J.A., 1975) . Clustering algorithms are also used widely in data mining (Han, 2000; Tan, 2005; Bishop, 2006; Seyed Amin Hosseini Seno, 2009; Reza Entezari-Maleki, 2009).

There are two main classifications for clustering algorithms:

1. Hierarchical

Corresponding Author: Morteza Saberi Kamarposhti, Department of Computer Science, Islamic Azad University, Firoozkooh Branch, Firoozkooh, Iran.
E-mail: Morteza.saberi@iaufb.ac.ir

2. Partitional

In the agglomerative hierarchical clustering each data is considered as a cluster and then similar clusters are merged in a recursive manner to achieve larger clusters. This operation will continue until the required cluster is achieved. On the other hand, in top-down hierarchical clustering, all data is considered as a cluster and then larger clusters are divided into smaller ones recursively to achieve required cluster count. One of the significant works that is done on hierarchical clustering is CURE that has the following benefits in comparison to others:

- Clustering algorithm can detect arbitrary shaped clusters.
- This algorithm has no problem with points away.

Memory usage for the algorithm is linear and time complexity for small data is $O(n^2)$.

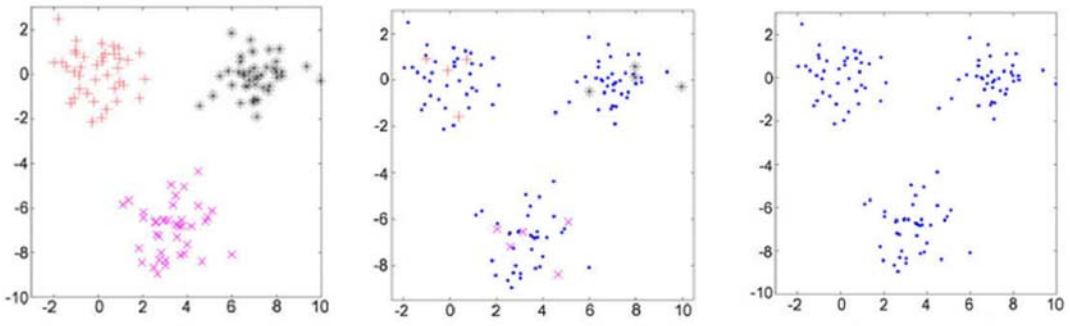

Fig. 1: (*a*) supervised clustering (*b*) semi-supervised clustering (*c*) unsupervised clustering.

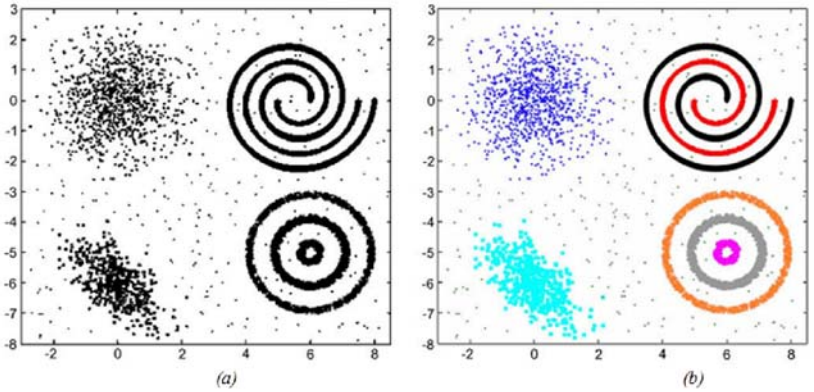

Fig. 2: (*a*) Input data (*b*) final goal of clustering.

CURE Algorithm Review:

The clustering algorithm classifies each input point in a separate cluster and every step in this algorithm is merging closest pairs of clusters. Representative point is stored for each cluster c to calculate the distance between a pair of clusters. To determine these points, c points are selected in the cluster and then reduced to the cluster average with α factor. The distance between two clusters is the distance between the nearest pair of points from them.

The goal is that the selected *c* points construct the physical and geometric shape of cluster. Furthermore, reduction of scattered points to the average is reducing the effect of outlying points. This is because the outlying points are often far from the cluster center and the reduction is causing the outlying points to be far from center while other points are not like this. α parameter can also be used to control the cluster shape. A small value for α will result in highly reduced number of scattered points. In other words, when α becomes bigger, the scattered points are closer to the average and clusters become more compact. Pseudo code that is used for CURE clustering is shown in figure 3. For more information about the algorithm refer to (S. Guha, 2001).

Proposed Algorithm:

In the next proposed algorithm, one phase is added to the clustering using CURE algorithm. In this phase, cluster center is calculated for each known cluster. Then distance from cluster center is calculated for all cluster members. In this phase, the further half of cluster members from cluster center are selected and for all selected members a *k* member probability vector is assumed having a value of 1/*K* for each member of vector.

Fig. 4: Pseudo code of Algorithm.

```
procedure cluster(S,K)
Begin
T:=build_kd_tree(S)

Q:=build_heap(S)

while size(Q)>k do

{
u:=extract_min(Q)
v:=u.closest
delete(Q,v)
w:=merge(u,v)
delete_rep(T,u)
delete_rep(T,v)
insert_rep(T,w)
w.closest:=x
for each X ε Q do
{
if dist(w,x)<dist(w,w.closest)
w.closest:=x
if x.closest is either u or v
{
if dist(x,x.closest)<dist(x,w)
x.closest:= closest_cluster(T,x,disst(x,w))
else
x.closest:=w;
relocate(Q,x)
}
else if dist(x,x.closest)>dist(x,w)
{
x.closest:=w;
relocate(Q,x)
}
}
insert(Q,w)
}
end
```

```
procedure merge(u,v)
begin
w:= u ∪v

                u u.mean + v v.mean
w.mean:=
                        u + v

tmpSet:=
for i:=1 to c do
{
maxDist:=0
 for each point p in cluster w do
{
if i=1

minDist:=dist(p,w.mean)
else
mindist:=min{dist(p,q): q} ∈tmpSet
if (minDist>=maxDist)
{
maxDist:=mindist
maxPoint:=p
}
}
tmpSet:=tmpSet ∪{maxPoint}
}
for each point p in tmpSet do
w.rep:=w.rep    {p+ α*( w.mean – p)    }
return w
end
```

In this phase, the distance from all members to all cluster centers is calculated and the minimum distance for each node is selected. After that, members of probability vector for each cluster member is updated as following:

Probability increase formula for the closest cluster center (l_m): $P_{n+1}(l_m) = P_n(l_m) + \frac{d_m}{\sum_{i=1}^{k} u_i}$

Probability decrease for other clusters (l_f): $P_{n+1}(l_f) = P_n(l_f) - \frac{d_f}{\sum_{i=1}^{k} u_i}$

Where P_{n+1} is the probability for the next phase, P_n is current probability, l_m is the probability amount for closest cluster, l_f is the probability amount for other clusters, d_m is the minimum distance to cluster center and d_f is the distance to other cluster centers. After doing this, all probabilities for each data is normalized separately to

be in the range of [0, 1]. Finally, 1% of data is selected randomly and is moved to the cluster having the maximum probability in the vector. This will continue till reaching two consecutive phases having the same cluster centers. At the end, all data having vector probability are moved to the cluster with the maximum value in the probability vector.

Experimental Results:

Adjusted Rand Index factor is used to evaluate the proposed algorithm. The formula is as following (Hubert, 1985):

$$J(C,K) = \frac{observed\ index - \exp ected\ index}{\max imum\ index - \exp ected\ index}$$

Data in the table I is used to test the algorithm (Xutao Li, 2010).

A comparison of proposed algorithm's results to other algorithms' is shown in table II and figure 4. Additionally, a comparison of correct cluster detection for the proposed algorithm and other algorithms is shown in table III.

Conclusion:

Many algorithms are proposed for data clustering. CURE is one of the famous algorithms. In this paper an improved hierarchical algorithm is issued based on CURE clustering using learning automata with improved results in comparison to the base CURE algorithm. Comparing to other similar algorithms, this algorithm has also a higher percentage in correct detection of clusters that shows better performance of proposed algorithm.

Table 1:

Data ID	Data Sets	No of Objects	No of genuine clusters
1	3D well-separated	4200	22
2	3D overlapped	3600	20
3	5D well-separated	2800	15
4	5D overlapped	3400	18
5	7D well-separated	7000	25
6	7D overlapped	7600	28
7	15D well-separated	2500	16
8	15D overlapped	2300	22
9	20D well-separated	2600	23
10	20D overlapped	2700	22
11	25D well-separated	3700	28
12	25D overlapped	3100	25

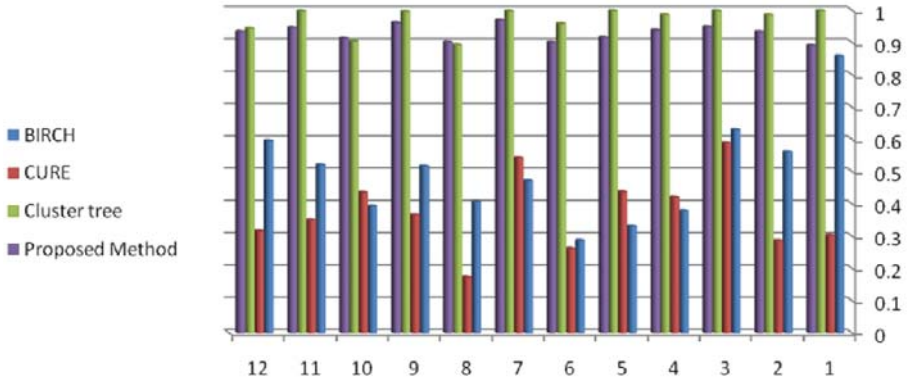

Fig. 3: A comparison of proposed algorithm to others.

Table 2:

Data ID	DB-SCAN	X-mean	BIRCH	CURE	NBC	OPTICS	Neural Gas	Tree-SOM	Cluster tree	Proposed Method
1	0.931	0.7767	0.8578	0.3028	0.9811	0.901	0.6134	0.8009	0.9995	0.8914
2	0.7012	0.4897	0.5611	0.2865	0.7848	0.6874	0.5844	0.671	0.9853	0.9343
3	0.7481	0.7759	0.6297	0.5894	0.9875	0.7473	0.6634	0.7577	0.999	0.9487
4	0.6491	0.8028	0.3789	0.4213	0.8166	0.5968	0.7063	0.694	0.9858	0.9391
5	0.8849	0.8251	0.3318	0.438	0.814	0.883	0.6452	0.7578	0.9989	0.9165
6	0.5021	0.5666	0.287	0.2619	0.4657	0.4655	0.6359	0.7185	0.9591	0.9012
7	0.7326	0.6897	0.4725	0.5433	0.7854	0.6661	0.4901	0.6654	0.998	0.9698
8	0.4331	0.58	0.407	0.1735	0.4385	0.4266	0.7298	0.7254	0.8931	0.9017
9	0.5269	0.8385	0.5173	0.3664	0.5513	0.5269	0.662	0.8615	0.9967	0.9623
10	0.5986	0.7012	0.3932	0.4358	0.637	0.5933	0.4666	0.6307	0.9051	0.9129
11	0.5854	0.6424	0.522	0.3507	0.6612	0.5854	0.6397	0.674	0.9981	0.9467
12	0.5562	0.6662	0.5962	0.3164	0.5998	0.5411	0.5474	0.688	0.9443	0.9349

Table 3:

Data ID	No of genuine clusters	DB-SCAN	X-mean	NBC	OPTICS	Cluster tree	Proposed Method
1	22	20	16	22	19	22	18
2	20	9	14	11	8	20	17
3	15	8	17	15	8	15	15
4	18	10	17	11	11	18	18
5	25	24	29	23	24	25	26
6	28	10	16	8	7	28	27
7	16	8	21	10	7	16	16
8	22	6	16	7	7	22	22
9	23	10	22	11	9	23	22
10	22	7	15	8	7	22	22
11	28	9	22	12	9	28	26
12	25	10	16	12	10	25	23

AKNOWLEDGEMENT

This Research was supported by Islamic Azad University, Firoozkooh Branch.

REFERENCES

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press.

Anil, K. Jain, 2010. Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31: 651-666.

Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer.

Chapelle, O., B. Schlkopf, A. Zien, (Eds.), 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.

Duda, R., P. Hart, D. Stork, 2001. Pattern Classi**W**cation, second ed. John Wiley and Sons, New York.

Gantz, John F., 2008. The diverse and exploding digital universe. Available online at: <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digitaluniverse.pdf>.

Guha, S., R. Rastogi, K. Shim, 2001. CURE: An Efficient Algorithm for Large Databases, 2001, Information systems, 26(1): 35-58.

Hubert, L., P. Arabie, 1985. Comparing partitions, Journal of Classi**W**cation, 2: 193-218.

Hartigan, J.A., 1975. Clustering Algorithms. John Wiley and Sons.

Han, Jiawei, Kamber, Micheline, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.

JSTOR, 2009. JSTOR. <http://www.jstor.org>.

Jain, Anil K., C. Dubes, Richard, 1988. Algorithms for Clustering Data. Prentice Hall. Merriam-Webster Online Dictionary, 2008. Cluster analysis. <http://www.merriam-webster-online.com>. Reza Entezari-Maleki, Seyyed Mehdi Iranmanesh and Behrouz Minaei-Bidgoli, 2009. An Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification Methods, World Applied Sciences Journal, Volume 7 (Special Issue for Computer & IT).

Sokal, Robert R., H.A. Sneath, Peter, 1963. Principles of Numerical Taxonomy. W.H. Freeman, San Francisco.

Seyed Amin Hosseini Seno, Rahmat Budiarto and Tat-Chee Wan, 2009. SHSDAP: Secure Hierarchical Service Discovery and Advertisement Protocol in Cluster Based Mobile Ad hoc Network, World Applied Sciences Journal, Volume 7 (Special Issue for Computer & IT).

Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin, 2005. Introduction to Data Mining, **W**rst ed. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA.

Xutao Li, Ye. Yunming, Li. Mark Junjie, K. Michael, Ng, 2010. On cluster tree for nested and multi-density data clustering, Pattern Recognition, 43: 3130-3143.