ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

A Review of Different Types of Clustering Algorithm & Analysis

A.Udhaya Kunam [1], Dr.N.Sujatha [2]

Research Scholar, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivagangai, TamilNadu, India [1]

Assistant Professor, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivagangai, TamilNadu, India [2]

ABSTRACT: Data mining (also known as Knowledge Discovery from data, or KDD for short).Data mining is one of the number of analytical tools for analyzing data. Data mining is an essential process where intellectual methods are applied to extract data patterns. Cluster technique is used to group a set of data into multiple group. But a very dissimilar to objects in other clusters. Clustering is the critical part of data mining. In this paper we are study the various clustering algorithms**.** Performance of these clustering algorithms are discussed and analyzed utilizing a clustering algorithm using Weka tool.

KEYWORDS: Data mining algorithms, Weak tools.

I. INTRODUCTION

Data mining is an important process where intelligent methods are applied to extract data patterns. Data mining is used to find the data warehouse data's. There are three types of data mining techniques, they are clustering, regression and classification,. Theoretically, data mining is the process of detecting correlations or pattern among dozens of fields in huge relational database. Cluster analysis or clustering is the main task of assigning a set of items into group (called clusters). It allow user to evaluate data from many different dimension or angles, categorize it, and review the relationship. Store and manage the information in a multidimensional database system. It give data access to business forecaster and information technology expert. Evaluate the data by application software. The data can be represented in a graphical format such as graph or tree.The algorithms are applied directly to a dataset. Pollution dataset is taken in this paper in order to analyze the performance of clusters .such as dbscan, optics, expectation maximization, simple k-means clustering algorithms. It is taken from the UCI machine learning repository website. This pollution dataset comprise the details of a pollution in 16 attributes .The above mentioned four algorithms are applied on the dataset to evaluate and compare the performances.

II. WHAT IS CLUSTER ANALYSIS?

Cluster analysis [1] or basically clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet similar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a Clustering. Cluster analysis has been widely used in a lot application such as

- ✓ Business intelligence
- ✓ Image pattern recognition,
- ✓ Web search
- ✓ Biology
- ✓ Security

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Cluster analysis[2] groups objects (observations, measures) based on the information found in the data describing the objects or their relationships. The purpose is that the objects in a group will be related to one other and unrelated to the objects in other groups.
The definition of what constitute a cluster is not well defined, and, in many applications clusters are not well separated from one another. However, most cluster analysis seeks as a result, a hard classification of the data into non-overlapping groups.

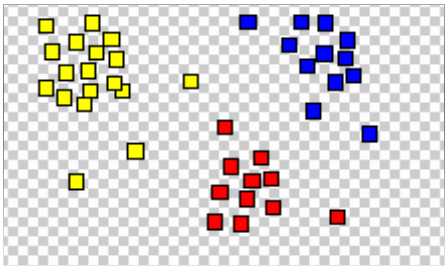

Figure 1 shows three clustering groups mentioned by squares in different colours.

III. DATASET

A dataset is a collection of data. For performing the comparisons and analysis we must project datasets. In this survey I am taking data from UCI machine learning repositories. This should have been taken from different nature.

IV. METHODOLOGY

My methodology is very easy. In the Weka, I am applying different clustering algorithms and predict the results that will be very useful for the new researchers.

V. CLUSTERING ALGORITHMS

1. DBSCAN clustering algorithm:

DBSCAN:[4] DBSCAN (Density –Based Spatial Clustering of application with Noise) find core items ,that is objects that have dense neighbourhoods. It connects core objects and their neighbourhoods to form dense regions as clusters.
A user –specified parameter $\epsilon > 0$ is used to specify the radius of a neighbourhood we consider for every object. The ϵ-neighbourhood of an object o is the space within a radius ϵ centred at o.

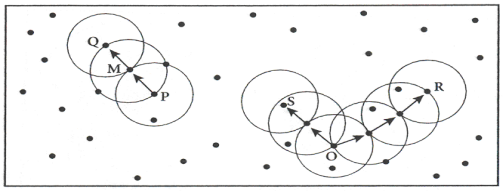

Figure 2: Function diagram

For a given ϵ represented by the radius of the circles. Let MinPts=3.Of the labelled points,m,p,o,r are core projects because each is in an ϵ-neighbourhood containing at least three points. To find the next cluster, DBSCAN arbitrarily selects an unvisited object from the remaining ones.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Algorithm: DBSCAN: a density-based clustering algorithm.
Input:

- D: a data set containing n objects,
- ϵ: the radius parameter ,and
- MinPts: the neighbourhood density threshold.

Output : A set of density-based clusters.

Advantage:

1. Discovery of arbitrary-shaped clusters with varying size.
2. Resistance to noise and outliers.
3. Time complexity.

Disadvantage:

1. High sensitivity to the setting of input parameters.
2. Poor cluster descriptors.
3. Unsuitable for high-dimensional datasets because of the curse of dimensionality phenomenon.

2. Optics clustering algorithm

Ordering points to identify the clustering structure (**OPTICS**) is an algorithm used to find density-based clusters in spatial data [5]. Its basic view is similar to DBSCAN,[3] but it represents one of DBSCAN's major weakness: The problem of identifying meaningful clusters in data of change density. In order to do so, the point of the database are (linearly) ordered such that point which are spatially closest become neighbours in the order. In addition, a special distance is stored for each point that represent the density that desires to be accepted for a cluster in order to have both points belong to the same cluster represented as a dendrogram. An attempt to overcome the necessity to supply different input parameters [6].

Algorithm OPTICS

```
OPTICS (SetOfObjects, e, MinPts, OrderedFile)
OrderedFile.open();
FOR i FROM 1 TO SetOfObjects.size DO
Object := SetOfObjects.get(i);
IF NOT Object.Processed THEN
ExpandClusterOrder(SetOfObjects, Object, e,
MinPts, OrderedFile)
OrderedFile.close();
END; // OPTICS
```

It illustrate the main loop of the algorithm OPTICS.[9]At the beginning, we open a file Ordered File for writing and close this file after ending the loop .Each object from a database Set Of Objects is simply handed over to a procedure Expand Cluster Order if the object is not yet processed.

Advantage:

1.When compared to the clustering algorithm do not limit to one global parameter .
2.It can take broad range of image.
3,Time complexity.
4. low cost compare to using a different outlier detection method

Disadvantage:

1. It is an improvement over that can find more complex hierarchies.

3. Expectation maximization

In common an expectation maximization (EM)[7] algorithm is a framework that deals with maximum likelihood or maximum a posterior estimates of parameters (MAP) in statistical model. Where the model depends on unobserved covert variables.[8]The k-means algorithm will continue the iteration until the clustering cannot be improved. Each iteration consists of two steps:

1. The expectation step (Exp-step): Given the current cluster centers, each object is assigned to the cluster with a center that is closest to the object. Here an object is expected to belong to the closest cluster.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

2. The maximization step(Max-step):Given the cluster assignment for each cluster the algorithm adjusts the center so that the sum of distances from the objects assigned to this cluster and the new center is minimized.That is, similarity of objects assigned to a cluster is maximized.

General EM Algorithm:
Alternate between steps until convergence:[9]
E step:

1.Maximize F wrt Q, keeping $!$ fixed.

2. Solution:

$Q^{k+1} = p(Z|D, \Theta^k)$

M step:

1. Maximize F wrt Θ, keeping Q fixed
2. Solution:

$Q^{k+1} = \arg_{\theta}^{\max} F(Q^{k+1}, \theta\Theta)$
$= \arg_{\theta}^{\max} \sum_Z p(Z|D, \Theta^k) \log p(X, Z| \Theta)$

Advantage:
1. Gives extremely useful result for the real world data set.
2. Use this algorithm when you want to perform a cluster analysis of a small scene or region-of -interest and are not satisfied with the results obtained from the k-means algorithm.

Disadvantage:
1. Slow convergence.
2. Inability to provide estimation to the asymptotic variance-covariance matrix of the maximum likelihood estimator (MLE).

4. Simple K-Means algorithm

- Partitional clustering approach
- Every cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple.[10]

K-MEANS ALGORITHM

Select K Points as the initial cancroids

1. *Repeat*
2. *From K Cluster by assigning all points to the Closest cancroids*
3. *Recomputed the cancroids of each Cluster*
4. *Until The cancroids don't change*

K-means Clustering – Details[11]

- Initial centroids are often chosen at random.
  - -Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the Cluster.
- 'Closeness' is measured by Euclidean distance, cosine
- Similarity, correlation, etc.
- K-means will converge for common similarity measures
- Mentioned above.
- Most of the convergence happens in the first few
- iterations.
  - Often the stopping condition is changed to 'Until relatively few
- points change clusters'

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

- Complexity is O(n * K * I * d)
  n = number of points, K = number of clusters,
- I = number of iterations, d = number of attributes.

Advantage:

1.It is simple and robust

2.If large number of variables exists, then K-Means is computationally faster than hierarchical clustering, if we keep k smalls.

3If the clusters are globular-Means produce tighter clusters than hierarchical clustering

4.More efficient than k-mediod.

Disadvantages:

1. Difficult to predict k-value.
2. With global clusters it did not work well.

VI. EXPERIMENTAL RESULT

Clustering algorithm	Instances	Attributes	Clustered instances (100%)	Un clustered instances	Time(percentage split)
DBSCAN	60	16	0 54(100%)	0	0.02sec
OPTICS	60	16	0	60	0.25 sec
EM	60	16	0 13(22%) 1 9(15%) 2 38(63%)	0	0.94 sec Log likelihood: -53.00298
SIMPLE KMEANS	60	16	0 54(90%)	1 6(10%)	0.02 sec

Table 1: Comparison result of algorithms using Pollution dataset

VII. CONCLUSION

This paper aims to provide an overview of the algorithms used in different clustering techniques along with their respective advantages and disadvantages. The challenge with clustering analysis is mainly that different clustering techniques give substantially different results on the same data. From the table it is clearly shows that the number of cluster instances is higher using Expectation Maximization(EM) technique while compared to other techniques which provides optimal result for pollution data set. Our future work will concentrate on change of Clustering Method in this way enhancing the proficiency of order in a diminished time. Additionally a mix of characterization systems will be used to enhance the performance.

REFERENCES

[1]Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000

[2] International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 5, May 2012) 73 Comparison the various clustering algorithms of weka tools.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). Evangelos Simoudis, Jiawei Han, Usama M. Fayyad, eds. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp.226–231. ISBN 1-57735-004-9.

[4]Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000

[5] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". *ACM SIGMOD international conference on Management of data*. ACM Press. pp. 49–60.

[6] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). *OPTICS: Ordering Points To Identify the Clustering Structure*. ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[7] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000

[8] Clustering and the EM Algorithm
Susanna RiccoCPS 271 25 October 2007Material borrowed from: Lise Getoor, Andrew Moore, Tom Dietterich, Sebastian Thrun, Rich Maclin, ...(and Ron Parr)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

[9]Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: *"The R*-tree: An Efficient and Robust Access Method for Points and Rectangles"*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, ACM Press, New
York, 1990, pp. 322-331.
[10] Ester M., Kriegel H.-P., Sander J., Xu X.: *"A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press,1996, pp. 226-231.
[11] A. P. Dempster; N. M. Laird; D. B. Rubin ―Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.
[12]Data MiningCluster Analysis: Basic Concepts
and Algorithms Lecture Notes for Chapter 8 Introduction to Data Mining by Tan, Steinbach, Kumar.