Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Clustering Based on Similarity Measures Using Concept Data Analysis

Bhavana Jamalpur
Asst. Prof, CSE, S.R Engineering College,
Hasanparthy, Warangal, Telengana, India

S. S. V. N Sarma
Professor, CSE, Vaagdevi Engineering College,
Bollikunta, Wararangal, Telengana, India

Abstract — Clustering aims to identify groups of similar data objects together such grouping is pervasive in the way humans process information which helps to discover of patterns and find interesting correlations in large data sets. It has been useful in many application domains in the field of engineering, business and social sciences. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points with same similarity are placed in one cluster are more similar to each other than points in different clusters. This paper evaluates the clustering validity measures based similarity on the dataset.

Keywords: clustering algorithms, unsupervised learning, cluster validity

I. INTRODUCTION

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting patterns in the underlying data. The goal of clustering is categorizing or grouping similar data items together. For example, consider a transactional database containing items purchased by customers. A clustering procedure is used to group the customers in such a way that customers with similar buying behaviour are in the same cluster. Clustering helps to partition a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters.

II. PROBLEM SPECIFICATION

The objective of the clustering methods is to discover significant groups present in a dataset. In general, identifying the data objects which are close to each other based on the similarity measure and are separated. A problem in clustering is to decide the optimal number of clusters in the data set and check for cluster validity.

III. STEPS IN CLUSTERING

The basic steps to develop clustering process are as follows

Feature selection. The goal is to select properly the features on which clustering is to be performed so as to encode as much information as possible concerning the task of our interest. Thus, pre-processing of data may be necessary prior to their utilization in clustering task.

- **Clustering algorithms.** This step refers to the choice of an algorithm that results in the definition of a good clustering scheme for a data set. A proximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set.

The different types of clustering algorithms are

- *Partition clustering* - attempts to directly decompose the data set into a set of disjoint clusters. More specifically, they attempt to determine an integer number of partitions that optimise a certain criterion function. The criterion function may emphasize the local or global structure of the data and its optimization is an iterative procedure.
- *Hierarchical clustering* - proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of the algorithm is a tree of clusters called dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained.
- *Density-based clustering* - The key idea of this type of clustering is to group neighbouring objects of a data set into clusters based on density conditions.
- *Grid-based clustering* - This type of algorithms is mainly proposed for spatial data mining.

The main characteristic is that they quantise the space into a finite number of cells and operates on the quantized space.

i) Proximity measure - is a measure that quantifies how "similar" two data points (i.e.feature vectors) are. In most of the cases we have to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others.

ii) Clustering criterion - In this step, we have to define the clustering criterion, which can be expressed via a cost function or some other type of rules. We should stress that we have to take into account the type of clusters that are expected to occur in the data set. Thus, it define a "good" clustering criterion, leading to a partitioning that fits well the data set.

iii) *Validation of the results* - The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering methods, the final partition of data requires some kind of evaluation in most applications .

iv) *Interpretation of the results* - In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.

Measures of Cluster Evaluation are based on

1. *Supervised* - Measures the extent to which the clustering structure discovered by clustering method matches some external criteria. Supervised measures are often called External indices.
   (Entropy , Purity a nd Jaccard Measure)
2. *Unsupervised* – Measures the goodness of a clustering structure without respect to external information such as SSE (Sum of the Squared Error). Unsupervised measures of cluster validity is divided into Cluster Cohesion (compactness ,tightness) and Cluster Separation (isolation) which determine how well /distinct clusters are from others. UnSupervised measures are often called internal indices.
3. *Relative* – Compares different clusterings or clusters. A relative cluster evaluation measure is a supervised or unsupervised for example K-means clustering can be compared using SSE or entropy.

Cluster Validity Assessment

Cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. This is the main objective of cluster validity. In general terms, there are three approaches to investigate cluster validity

1. external criteria
2. internal criteria

There are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme as
Compactness, the members of each cluster should be as close to each other as possible.
A common measure of compactness is the variance, which should be minimized.
Separation, the clusters themselves should be widely spaced. There are three common approaches measuring the distance between two different clusters:

- *Single linkage:* It measures the distance between the closest members of the clusters.
- *Complete linkage:* It measures the distance between the most distant members.
- *Comparison of centroids :* It measures the distance between the centers of the clusters.

The two first approaches are based on statistical tests and their major drawback is their high computational cost. Moreover, the indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme. On the other hand, the third approach aims at finding the best clustering scheme that a clustering algorithm can be defined under certain assumptions and parameters.

IV. METHODOLOGY USED

Fig.1 shows the basic steps to find the similarity among the data objects and their attributes. The objects with similar properties are grouped/clustered in the following way as:

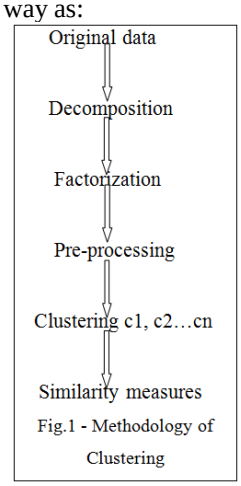


Fig.1 - Methodology of Clustering

Step1:- The original data set contains 101 objects and attributes which is available in UCI data set.

Step2:-The data is decomposed into 0/1 representation considering on useful attributes of our interest and ignoring the remaining without any loss of information by using the concept of attribute subset selection.

Step3:-The attributes are factored by using the concept of Boolean matrix factorization where the attributes are factored .

Step4:-Then preprocess steps of data mining such as data cleaning, data transformation, integration , reduction and discretization

Step5:-Data objects are partitioned into groups called clusters that represent proximate collections of data elements based on a similarity calculation.

Illustration

Given two objects, *A* and *B*, each with *n* binary attributes, the Jaccard coefficient is a useful measure of the overlap that *A* and *B* share with their attributes. Each attribute of *A* and *B* can either be 0 or 1. The total number of each combination of attributes for both *A* and *B* are :

M_{11} indicates the total number of attributes where *A* and *B* both have a value of 1.

M_{01} indicates the total number of attributes where the attribute of *A* is 0 and the attribute of *B* is 1.

M_{10} indicates the total number of attributes where the attribute of *A* is 1 and the attribute of *B* is 0.

M_{00} indicates the total number of attributes where *A* and *B* both have a value of 0.

Each attribute must fall into one of these four categories,

$M_{11} + M_{01} + M_{10} + M_{00} = N$

The Jaccard similarity coefficient, *J*, is given as

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

V. EXPERIMENTAL STUDY

Consider the sample dataset which is factorized and by using attribute subset selection the following lattice is constructed by taking the relevant attributes and and data bjects and then clustering the objects by their behaviour of the concepts and calculating the similarity among them

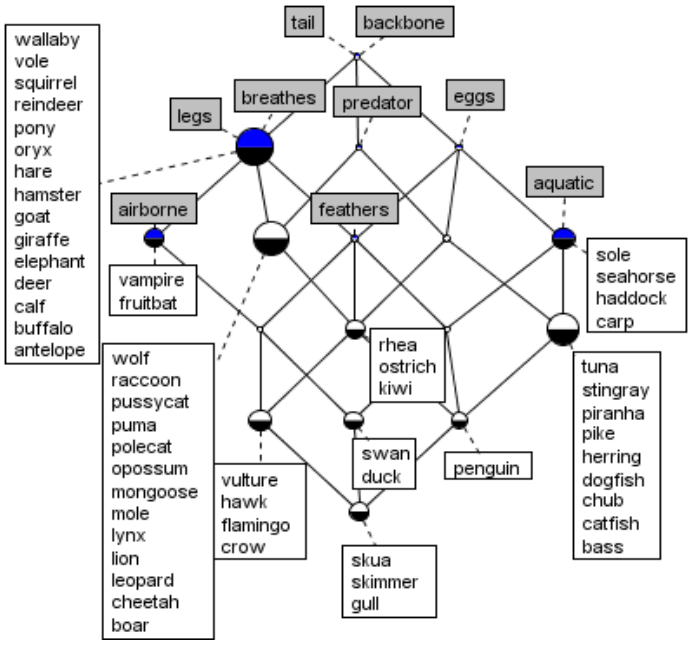


Fig-2 Lattice –Cluster

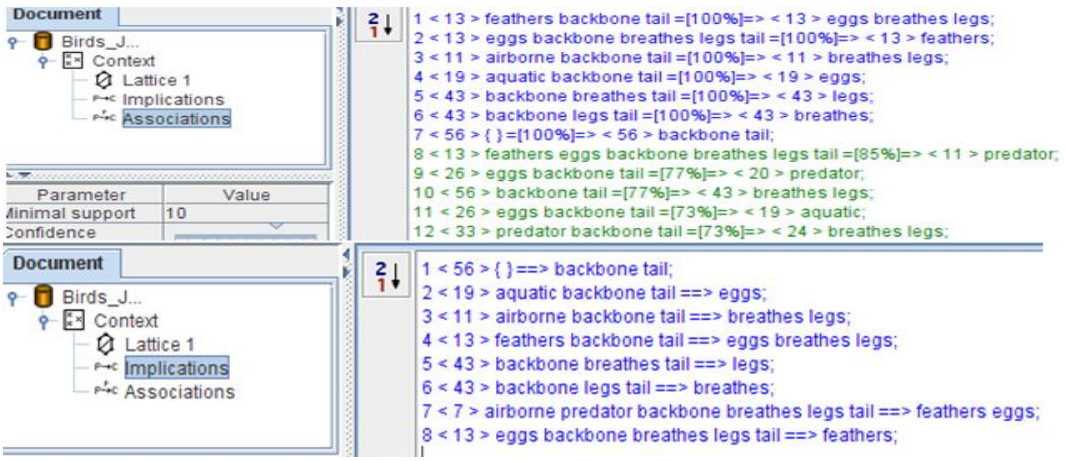


Fig-3 Association Rules and Implications

Table1-clusters and dataobjects

Sno	Clusters	Data Objects	Similarity (JC)	Concept Similarity
1	Cluster1(c1)	{1,2,3,4,5,6,7,8,9,10,11,12,13}	1.0	{feathers, eggs, airborne, aquatic }= 0 {predator ,backbone, breathes, legs ,tail}= 1
2	Cluster2(c2)	{14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30}	1.0	{feathers, eggs, airborne, aquatic ,predator} = 0 {backbone, breathes, legs, tail}= 1
3	Cluster3(c3)	{56,55,52,51,50,48,47,46,44}	1.0	{feathers, airborne, breathes, legs} =0 {eggs, aquatic, predator, backbone, tail}=1
4	Cluster4(c4)	{32,34,35}	1.0	{feathers, airborne, breathes, legs, eggs, aquatic, predator, backbone, tail}=1
5	Cluster5(c5)	{40,41,42}	1.0	{airborne, aquatic} =0 {feathers, breathes, legs, eggs, predator, backbone, tail}=1
6	Cluster6(c6)	{19,28}	1.0	{feathers, eggs, aquatic, predator}=0 {airborne, backbone, breathes , legs , tail}=1
7	Cluster7(c7)	{45,49,53,54}	1.0	{feathers, airborne, breathes, legs, predator}= 0 {eggs, aquatic , backbone , tail} =1
8	Cluster8(c8)	{31,36}	1.0	{ predator}= 0 { feathers, airborne, breathes, legs, eggs, aquatic, backbone, tail}=1
9	Cluster9(c9)	{33}	1.0	{airborne}=0 { feathers, breathes, legs, eggs, aquatic, predator, backbone, tail}=1

Entropy – The degree to which each cluster consists of objects of a single class. For each cluster i belongs to class j as Pij=mij/mi where mi is the no.of objects of class j in cluster i

$E_i=-\sum^{l} P_{ij} \log_2 P_{ij}$

Total entropy for a set of clusters is calculated as the sum of the entropies of each weighted clustered by the size of each cluster

$E=\sum^{k} (mi/m)$ ei where **k**- no of clusters

m- total no. of data

Considering in our dataset, calculate the entropy as

$E_i= -\sum^{9} P_{ij} \log_2 P_{ij}$

$=0.066 \log_2 0.066 +0.058 \log_2 0.058+0.111 \log_2 0.111+0.333 \log_2 0.333 +0.333 \log_2 0.333 +0.5 \log_2 0.5 +0.25 \log_2 0.25+0.5 \log_2 0.5+1.0 \log_2 1.0$

=1.512

e=(15/56+17/56+9/56+3/56+3/56+2/56+4/56+2/56+1/56)

e=(1/56)(15+17+9+3+3+2+4+2+1)(1.512)

e=1.512

Purity – contains the object of a single class .The purity of cluster is

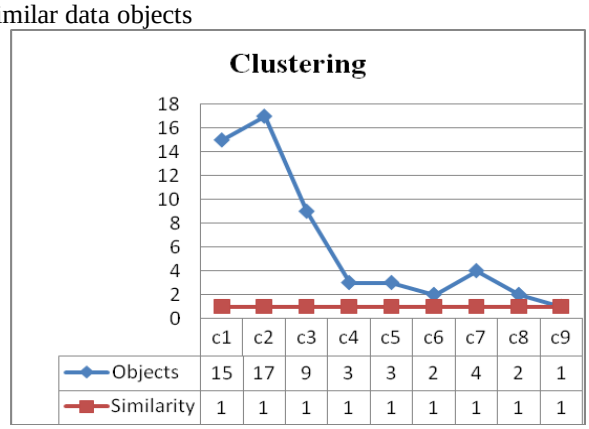
$$\mathbf{P_i=max_j P_{ij}}$$

Clustering Purity $=\sum^{k} (m_i/m)P_i$

Data Objects are grouped into clusters based on the similarity(s) and check for purity.

$$\mathbf{p_{ij}=m_{ij}/m_i}$$

All the above clusters (c1,c2,c3,c4,c5,c6,c7,c8,c9) have the cluster purity =1.0 where $p_i=\max P_{ij}$ consisting of similar data objects



VI. CONCLUSION

The major contributions of this paper is to provide a practical methodology to build the concept lattice and establish relationship between data mining embedded with FCA structure, by calculating the similarity measures (Jaccard measure) and also the concept similarity and then by clustering them based on the concepts .The experiment is conducted by lattice construction and finding the associations and implications based on the lattice. The experimental results show the cluster formed are pure with cluster purity=1. Hence clusters formed are pure and valid clusters with same behavior.

REFERENCES

[1] C. J. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, L. M.Hage, and W. E. Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse," American Medical Informatics Association Annual Fall Symposium (formerly SCAMC), 1997, pp. 101-5.

[2] K. Seki and J. Mostafa, "An Application of Text Categorization Methods to Gene Ontology Annotation," Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 138-145.

[3] T, Kardi. (2008). Similarity Measurement. http://people.revoledu.com/kardi\/tutorial/Similarity/.

[4] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository: Zoo Data Set. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[5] Chim, H., and Deng, X., "Efficient Phrase-Based Document Similarity for Clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1217-1229, 2008.

[6] Zamir, O. M. O., Etzioni, O., and Karp, R. M., "Fast and Intuitive Clustering of WebDocuments"

[7] Choi, S.-S, (2008), "Correlation Analysis of Binary Similarity Measures and Dissimilarity Measures", Doctorate dissertation, Pace University.

[8] Ganter, B., Wille, R., Formal Concept Analysis: Mathematical Foundations. Springer, Berlin.

[9] J.Han and M. Kamber. Data Mining : Concepts and Techniques, Morgan Kaufmann,2000

