International Journal of Research in Computer Science
eISSN 2249-8265 Volume 2 Issue 1 (2011) pp. 29-31

www.ijorcs.org

A COMPARATIVE STUDY ON DISTANCE MEASURING APPROACHES FOR CLUSTERING

Shraddha Pandit[1], Suchita Gupta[2]
*Assistant Professor,Gyan Ganga Institute of Information Technology and Management,Bhopal

Abstract: *Clustering plays a vital role in the various areas of research like Data Mining, Image Retrieval, Bio-computing and many a lot. Distance measure plays an important role in clustering data points. Choosing the right distance measure for a given dataset is a biggest challenge. In this paper, we study various distance measures and their effect on different clustering. This paper surveys existing distance measures for clustering and present a comparison between them based on application domain, efficiency, benefits and drawbacks. This comparison helps the researchers to take quick decision about which distance measure to use for clustering. We conclude this work by identifying trends and challenges of research and development towards clustering.*

Keywords: *Clustering, Distance Measure, Clustering Algorithms*

I. INTRODUCTION

Clustering is an important data mining technique that has a wide range of applications in many areas like biology, medicine, market research and image analysis etc.It is the process of partitioning a set of objects into different subsets such that the data in each subset are similar to each other. In Cluster analysis Distance measure and clustering algorithm plays an important role [1].

An important step in any clustering is to select a distance measure, which will determine how similarity [1] of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

It is expected that distance between objects within a cluster should be minimum and distance between objects within different clusters should be maximum. In this paper we compare different distance measures. Comparison of these distance measures show that different distance measures behave differently depending on application domain. The rest of the paper is organized as follows:

In section II, we discuss distance measures and its significance in nutshell; in section III, we present the comparison between these distances measures in TABLE I; In section IV, we describe how the accuracy can be measured and interpretation of the comparison; And we conclude the report.

II. DISTANCE MEASURES AND ITS SIGNIFICANCE

A cluster is a collection of data objects that are similar to objects within the same cluster and dissimilar to those in other clusters. Similarity between two objects is calculated using a distance measure [6].Since clustering forms groups; it can be used as a pre-processing step for methods like classifications.

Many distance measures have been proposed in literature for data clustering. Most often, these measures are metric functions; Manhattan distance, Minkowski distance and Hamming distance. Jaccard index, Cosine Similarity and Dice Coefficient are also popular distance measures. For non-numeric datasets, special distance functions are proposed. For example, edit distance is a well-known distance measure for text attributes.

In this section we briefly elaborate seven commonly used distance measures.

A. Euclidean Distance

The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. It is the straight line distance between two points.

In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$.

In N dimensions, the Euclidean distance between two points p and q is $\sqrt{(\Sigma i=1N (pi-qi)^2)}$ where pi (or qi) is the coordinate of p (or q) in dimension i.

B. Manhattan Distance

The distance between two points measured along axes at right angles. In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is |x1 - x2| + |y1 - y2|.

This is easily generalized to higher dimensions. Manhattan distance is often used in integrated circuits where wires only run parallel to the X or Y axis. It is also known as rectilinear distance, Minkowski's [7] [3] L1 distance, taxi cab metric, or city block distance.

C. Bit-Vector Distance

An N x N matrix Mb is calculated. Each point has d dimensions and Mb (Pi, Pj) is determined as d-bit vector. This vector is obtained as follows:

If the numerical value of the xth dimension of point is greater than the numerical value of the xth dimension of point pj,y then the bit x of Mb(Pi ,Pj) is set to 1 and bit x of Mb(Pj, Pi) is set to 0.All the bit vectors in Mb are then converted to integers.

D. Hamming Distance

The Hamming distance between two strings of equal length is the number of positions for which the corresponding symbols are different.

Let x, y A^n. We define the Hamming distance between x and y, denoted dH(x, y), to be the number of places where x and y are different.

The Hamming distance [1] [6] can be interpreted as the number of bits which need to be changed (corrupted) to turn one string into other. Sometimes the number of characters is used instead of the number of bits. Hamming distance can be seen as Manhattan distance between bit vectors.

E. Jaccard Index

The Jaccard index, also known as the Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets.

The Jaccard coefficient [11] measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$\sqrt{}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

F. Cosine Index

It is a measure of similarity between two vectors of n dimensions by finding the angle between them, often used to compare documents in text mining. Given two vectors of attributes, A and B, the cosine similarity [11], θ, is represented using a dot product and magnitude as

$$\theta = arccos \frac{A.B}{\|A\|\|B\|}$$

For text matching, the attribute vectors A and B are usually the tf-idf vectors of the documents. Since the angle, θ, is in the range of [0, π], the resulting similarity will yield the value of π as meaning exactly opposite, π / 2 meaning independent, 0 meaning exactly the same, with in-between values indicating intermediate similarities or dissimilarities

G. Dice Index

Dice's coefficient [11] (also known as the Dice Coefficient) is a similarity measure related to the Jaccard index.

For sets X and Y of keywords used in information retrieval, the coefficient may be defined as:

$$S = \frac{2|X \cap Y|}{|X| + |Y|}$$

When taken as string similarity measure, the coefficient may be calculated for two strings, x and y using bigrams as follows:

$$S = \frac{2n_t}{n_x + n_y}$$

Where nt is the number of character bigrams found in both strings, nx is the number of bigrams in string x and ny is the number of bigrams in string y.

III. ACCURACY AND RESULT INTERPRETATION

In general, the larger the number of sub-clusters produced by the clustering the more accurate the final result is. However, too many sub-clusters will slow down the clustering. The above comparison table compares 5 proximity measures. This comparison is based on 4 different criteria which are generally required to decide upon distance measure and clustering algorithms.

All above comparisons are tested using standard synthetic dataset generated by Syndeca [3] Software and few of it is tested using open source clustering tool CLUTO.

IV.CONCLUSION

This paper surveys existing proximity measures for clustering and presents a comparison between them based on application domain, efficiency, benefits and drawbacks. This comparison helps the researchers to take quick decision about which distance measure to use for clustering. We ran our experiments on synthetic datasets for its validation. Future work involves running the experiments on larger and different kinds of datasets and extending our study to other proximity measures and clustering algorithms.

V. REFERENCES

[1] Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem , “An Experiment with Distance Measures for Clustering” , Technical Report: IIIT/TR/2008/132

[2] John W. Ratcliff and David E. Metzener, Pattern Matching: The Gestalt Approach, DR. DOBB’S JOURNAL, 1998, p. 46.

[3] Martin Ester Hans-Peter Kriegel Jrg Sander and Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, AAAI Press, 1996, pp. 226–231.

TABLE I: COMPARISON OF DISTANCE MEASURE

Distance Measure	**Formula**	**Algorithms In which it is Used**	**Benefits**	**Drawbacks**	**Application Area**
Euclidean	$\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$	-Partitional Algorithms -K Modes - AutoClass -ROCK	Easy to implement and Test	Results are greatly influenced by variables that have the largest value. Does not work well for image data, Document Classification	-Appl. Involving Interval Data - In health psychology analysis - DNA Analysis
Manhattan	$\|x_1 - x_2\| + \|y_1 - y_2\|$.	Partitional Algorithms	Easily generalized to higher dimensions	Does not work well for image data and Document Classification	In Integrated Circuits
Cosine		Ontology and	Handles both Continuous and categorical variables	Does not work well for nominal data	Text Mining
Similarity	$\Theta = \arccos \frac{A \bullet B}{\|\|A\|\| \|\|B\|\|}$	Graph based			
Jaccard Index	$\sqrt{(A,B)} = \frac{\|A \cap B\|}{\|AUB\|}$	Neural Network	Handles both Continuous and categorical variables	Does not work well for nominal data	Document classification

[4] Bar-Hilel, A., Hertz, T., Shental, N., & Weinshall, D. (2003). Learning distance functions using equivalence

[5] Fukunaga, K. (1990). Statistical pattern recognition. San Diego: Academic Press. 2nd edition.

[6] Rui Xu, Donald Wunsch "Survey of Clustering Algorithms" IEEE Transactions on Neural Networks , VOL. 16, NO. 3, MAY 2005. doi:10.1109/TNN.2005.845141

[7] http://en.wikipedia.org/wiki/Data clustering

[8] http://en.wikipedia.org/wiki/K-means

[9] http://en.wikipedia.org/wiki/ DBSCAN

[10] http://en.wikipedia.org/wiki/Jaccard index

[11] http://en.wikipedia.org/wiki/Dice coefficient

How to cite

Shraddha Pandit, Suchita Gupta, "A Comparative Study on Distance Measuring Approaches for Clustering". *International Journal of Research in Computer Science*, 2 (1): pp. 29-31, December 2011. doi:10.7815/ijorcs.21.2011.011