Review on importance of distance measure in data mining and its application

*E. Kavitha**
University College of Engineering Villupuram,
Villupuram, Tamil Nadu, India
Email: ekavithavrs@gmail.com

R. Tamilarasan
University College of Engineering Pattukottai,
Thanjavur, Tamil Nadu, India.
Email: rrtamilk@yahoo.co.in

***Abstract*— This review paper deals with the importance of Distance Measure in Data Mining. It conjointly expresses the essential of distance measure in mining of knowledge by varied researchers. Here we tend expose the fundamentals of Data Mining techniques, its completely different approaches, difficulties, execs and cons of them. Completely different measures of distance or similarity are convenient for various varieties of analysis. Data Mining is an analytic process to explore information in search of consistent patterns, then to validate the findings by applying the detected patterns to new subsets of data. Retrieval of information, similarities / dissimilarities, finding and implementing the correct measure are at the heart of data mining. Distance measures are mathematical approaches to measure distance between the objects. Practically distance measures helps to compare the objects from three different standpoints such as Similarity, Dissimilarity and Correlation. The work carried out in this paper is based on the study of popular distance measures.**

Keywords—Data mining, distance measure, Euclidean distance, Manhattan distance, Correlation distance.

I. INTRODUCTION

Simply saying for easy understanding Data mining refers to extracting or mining knowledge from large amounts of data. Many data mining techniques are based on similarity measures between objects [1]. At heart, the goal of data mining is to extract knowledge from data. Data mining is an interdisciplinary field, whose core is at the intersection of machine learning, statistics, and databases. There are several data mining task, including classification, regression, clustering etc., [2]. Each of these tasks can be regarded as a kind of problem to be solved by a data mining algorithm.

II. DISTANCE MEASURE

Importance of Measurement aims to mine structured data is to discover relationships that exist in the real world – business, physical, conceptual etc. Instead of looking at real world we look at data describing it .Data maps entities in the area of concern to symbolic representation by means of a measurement procedure. Numerical relationships between variables confine relationships between objects. Measurement process is crucial. Similarity metric is the basic measurement and used by a number of data mining algorithms. It measures the similarity or dissimilarity between two data objects which have one or multiple attributes. Informally, the similarity is a numerical measure of the degree to which the two objects are alike. It is usually non-negative and are often between 0 and 1, where 0 means no similarity, and 1 means complete similarity [8]. Considering different data type with a number of attributes, it is important to use the appropriate similarity metric to well measure the proximity between two objects. For example, Euclidean distance and correlation are useful for dense data such as time series or two-dimensional points. Jaccard and cosine similarity measures are useful for sparse data like documents, or binary data.

Proximity is a general term to indicate similarity and dissimilarity [11]. Similarity [4] is a numerical measure of how alike two data objects are. Value is higher when objects are more alike. It often falls in the range [0,1]. Dissimilarity is numerical measure of how different two data objects are. Lower when objects are more alike .Minimum dissimilarity is often 0 and the upper limit varies.

Distance or Metric should satisfy

1. d(i,j) ≥ 0 Positivity

2 d(i,j) = d(j,i) Commutativity

3 d(i,j) < d(i,k)+d(k,j) Triangle Inequality

The different distance measures are

- Euclidean distance
- Manhattan distance
- Correlation distance
- Min- kowski distance

The Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. The Minkowski distance is as follows

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^{q}+|x_{i2}-x_{j2}|^{q}+\ldots+|x_{ip}-x_{jp}|^{q})}$$

where i = (x_{i1}, x_{i2}, …, x_{ip}) and j = (x_{j1}, x_{j2}, …, x_{jp}) are two p-dimensional data objects, and q is a positive integer .

If q = 1, d is Manhattan distance

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{ip}-x_{jp}|$$

If q = 2, d is Euclidean distance

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^{2}+|x_{i2}-x_{j2}|^{2}+\ldots+|x_{ip}-x_{jp}|^{2})}$$

*Corresponding author.

This paper was presented by the first author in the National Conference on *Advances in Mathematics and its Applications to Science and Engineering* (AMASE-2016) conducted in Department of Mathematics, University College of Engineering-Pattukkottai, Thanjavur, Tamil Nadu, India, on 22nd January 2016.

Euclidean distance [5], [6], [7], [10] is one of the most commonly-used methods to measure the distance between two data objects.

Correlation is often used as a preliminary technique to discover relationships between variables. More precisely, the correlation is a measure of the linear relationship between two variables. The correlation is a measure of how well two sets of data fit on a straight line. Correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship. If the correlation is 0, then there is no linear relationship between the attributes of the two data objects. However, the two data objects might have non-linear relationships. There are different types of correlation and few examples are Pearson Correlation[10] and Jackknife Correlation [6].

Pearson correlation is defined by the following equation. x and y represents two data objects.

$$d_C(x,y) = 1 - \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}}.$$

Distance is to indicate dissimilarity. The active use of any of these measures (Table 1) is highly influenced by the number of data records / instances, their dimensionality, the level of precision required and the nature of analysis required.

TABLE I. DIFFERENT DISTANCE MEASURE

Distance measure	Forms	Explanation
Min-kowski distance	$D_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert X_i - Y_i \rvert^p}$	Metric. Invariant to any translation and rotation only for n = 2 (Euclidean distance). Features with large values and variances tend to dominate over other features.
Euclidean distance	$D_{Euc} = \sqrt{\sum_{i=1}^{d} \lvert X_i - Y_i \rvert^2}$	The most commonly used metric. Special case of Minkowski metric at n = 2. Tends to form hyper-spherical clusters.
City-block distance	$D_C = \sum_{i=1}^{d} \lvert X_i - Y_i \rvert$	Special case of Minkowski metric at n = 1. Tend to form hyper- rectangular clusters.
Sup distance	$D_{Sp} = max\lvert X_i - Y_i \rvert$	Special case of Minkowski metric at n →∞.
Mahala-nobis distance	$D_M = (X_i - Y_i)^T S^{-1} (X_i - Y_i)$ where S is within group co-variance matrix.	Invariant to any nonsingular linear transformation. S is calculated based on all objects. Tends to form hyper-ellipsoidal clusters. When features are not correlated, squared Mahalanobis distance is equivalent to squared Euclidean distance. May tend to be computationally expensive.
Pearson correlation	$D_p(X,Y) = \sum_{i=1}^{d} \frac{(X_i - Y_i)^2}{Y_i}$	Not a metric. Derived from correlation coefficient. Unable to detect the magnitude of differences of two variables.

III. DATA MINING

Data mining is abstraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. Data mining tasks are classified into two Descriptive data mining and Predictive data mining. Prediction Tasks are used in some variables to predict unknown or future values of other variables whereas Description Tasks used to find human-interpretable patterns that describe the data. Some of the common data mining tasks are Classification [Predictive], Clustering [Descriptive], Association Rule Discovery [Descriptive], Sequential Pattern Discovery [Descriptive], Regression [Predictive], Deviation Detection [Predictive] etc. Commonly clustering is used to find out the similar, dissimilar and outlier items from the databases. The key idea behind the clustering is the distance between the data items. Clustering [12] is the task of discovering groups on the bases of the similarities of data items within the clusters and dissimilarities outside the clusters on the other hand from data set. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Clustering is a main task of

Explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.

IV. APPLICATIONS

Data mining is widely used in diverse areas [3]. Most of the application in data mining are Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

A. *Fraud Detection*

Predict fraudulent cases in credit card transactions. Use credit card transactions and the information on its account-holder as attributes. When a customer buys, what he buys, how often he pays on time, etc. Label past transactions as fraud or fair transactions. This forms the class attribute. Learn a model for the class of the transactions. Use this model to detect fraud by observing credit card transactions on an account.

B. *Market Segmentation*

Subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix. Collect different attributes of customers based on their geographical and lifestyle related information. Find clusters of similar customers. Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

C. *Document Clustering*

To find groups of documents that are similar to each other based on the important terms appearing in them. To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster. Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

V. CONCLUSION

In this paper we literature several papers to find the importance of distance measure in Data mining. Data mining plays an important role in many fields to retrieve useful information as the need of the customer. In retrieving or grouping of similar information under a common group the distance measure plays a major role. Grouping of similar items into a separate group is called as clustering. There are several papers dealing about the different distance measure and its support according to the nature of applications. The distance measure like Euclidean, Pearson, Jaccard etc are showing their benefits in grouping of similar data objects. They also have their own merits and demerits according to the place of using them.

REFERENCES

[1] David Hand, Heikki Mannila and Padhraic Smyth,. Principles of Data Mining, MIT press, 2002.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery: An overview,” in Advances in Knowledge Discovery & Data Mining

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Mateo, CA: Morgan Kaufmann, 2001.

[4] D. P. Berrar, W. Dubitzky, and M. Granzow, A Practical Approach to Microarray Data Analysis. Norwell, MA: Kluwer, 2003.

[5] Huai-Kuang Tsai, Jinn-Moon Yang, Yuan-Fang Tsai, and Cheng-Yan Kao,” An Evolutionary Approach for Gene Expression Patterns,”, IEEE Transactions on Information Technology in Biomedicine, VOL. 8, No. 2, June 2004 R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 1–34.

[6] Daxin Jiang, Chun Tang, and Aidong Zhang , “Cluster Analysis for Gene Expression Data: A Survey,” IEEE Transactions on Knowledge and Data Engineering, Vol. 16, NO. 11, November 2004.

[7] http://en.wikipedia.org/wiki/Euclidean_distance

[8] Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Published by Addison Wesley

[9] Georgina Stegmayer, Diego H. Milone, Laura Kamenetzky,Mariana G. Lo´ pez, and Fernando Carrari, “A Biologically Inspired Validity Measure for Comparison of Clustering Methods over Metabolic Data Sets,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, VOL. 9, NO. 3, MAY/JUNE 2012.

[10] Daxin Jiang, Jian Pei, Member, IEEE Computer Society, and Aidong Zhang, Member, IEEE,” An Interactive Approach to Mining Gene Expression Data,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, NO. 10, October 2005.

[11] Pablo A. Jaskowiak, Ricardo J.G.B. Campello, and Ivan G. Costa,” Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 10, No. 4, July/August 2013.

[12] Deepak Sinwar, Rahul Kaushik,Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering,” International Journal For Research in Applied Science and Engineering Technology (I JRAS ET), Vol. 2 Issue V, May 2014.

[13] L. D. Chase, " Euclidean Distance", College of Natural Resources, Colorado State University, Fort Collins, Colorado, USA, 824-146-294, NR 505, December 8, 2008.