RESEARCH ARTICLE OPEN ACCESS

Evaluation of Images Using Various Distance Metrics

Venkataramana Battula[1], Saritha Ambati[2]
Assistant Professor, CSE Department, MVSREC, Hyderabad
** Corresponding Author: Venkataramana Battula*

ABSTRACT
Due to the digitization of data and advances in technology, it has become extremely easy to Obtain and store large quantities of data, particularly Multimedia data. Image data plays vital role in every aspect of the systems like business for marketing, hospital for surgery, engineering for construction, Web for publication and so on. Fields ranging from Commercial to Military need to analyze these data in an efficient and fast manner. The need for image mining is high in view of such fast growing amounts of image data. Similarity and dissimilarity measures referred to as measures of proximity. Computing similarity measures are required in many data mining tasks. Categorical data, unlike numeric data, conceptually is deficient of default ordering relations on the attribute values. Devise distance metrics in data mining tasks for categorical data more challenging. Efficient extraction of low level features like color, texture and shapes for indexing and fast query image matching with indexed images for the retrieval of similar images by content . Features are extracted from images in pixel and compressed domains. The feature extraction and similarity measures are the two key parameters for retrieval performance. A similarity measure plays an important role in image retrieval. This paper compares different distance metrics such as Euclidean, Manhattan,Chebyshev, Canberra distances to find the best similarity measure for clustering the images.

Keywords: *Types of attributes, proximity measures, Clustering, K-Means.*

I. INTRODUCTION

Data mining is the process of finding interesting patterns in large quantities of data. This process of knowledge discovery involves various steps, the most of these being the application of algorithms to the data set to discover patterns as in, for example, clustering. The goal of clustering is to find natural groups in the data so that objects in a group are similar to other objects in the same group and dissimilar to objects in other groups. When implementing clustering (and also some classification algorithms such as k nearest neighbours), it is important to be able to quantify the proximity of objects to one another [1].

New distance/similarity measures for a variety of applications are introduced. Use of similarity measures in many different fields such as anthropology, biology, chemistry, computer science, ecology, information theory, geology, mathematics, physics, psychology, statistics, etc. There have been considerable efforts in finding the appropriate measures among such a plethora of choices because it is of fundamental importance to pattern classification, clustering, and information retrieval problems. Such endeavors have been conducted throughout different fields. Despite such comparative studies on diverse distance/similarity measures, further comprehensive study is necessary because even names for certain distance/similarity measures are fluid and promulgated differently. Synonyms for similarity include proximity and similarity measures are often called similarity coefficients. The choice of distance/similarity measures depends on the measurement type or representation of objects [6].

One of the biggest challenges of this decade is with databases having a variety of data types. Variety is among the key notion in the emerging concept of big data, which is known by the 4 Vs: Volume, Velocity, Variety and Variability. Currently, there are a variety of data types available in databases, including: interval-scaled variables (salary, height), binary variables (gender), categorical variables (religion: Jewish, Muslim, Christian, etc.) and mixed type variables (multiple attributes with various types). Despite data type, the distance measure is a main component of distance-based clustering algorithms. Partitioning algorithms are mainly dependent on distance measures to recognize clusters in a dataset.

Clustering used in medical image analysis, clustering gene expression data , investigating and analyzing air pollution data , power consumption analysis, and many more fields of study. Improving clustering performance has always been a target for researchers. Performance of clustering algorithms depends on the dissimilarity (distance) measures.

These algorithms use similarity or distance measures to cluster similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters. Examples of distance-based clustering algorithms include partitioning clustering algorithms, such as k-means as well as k-medoids and hierarchical clustering [2].

II. LITERATURE SURVEY

The performance of many learning and data mining algorithms depend on their being given a good measure over the input space. For instance, K-means, nearest-neighbors classifiers and kernel algorithms such as SVMs all need to be given good metrics that reflect reasonably well the important relationships between the data.

Sung-Hyuk Cha summarizes — Distance or similarity measures are essential to solve many pattern recognition problems such as classification, clustering, and retrieval problems. Various distance/similarity measures that are applicable to compare two probability density functions, pdf in short, are reviewed and categorized in both syntactic and semantic relationships. A correlation coefficient and a hierarchical clustering technique are adopted to reveal similarities among numerous distance/similarity measures[6].

Vaishali R. Patel & Rupa G. Mehta compare the results of modified k-Means with different distance measures like Euclidean Distance, Manhattan Distance, Minkowski Distance, Cosine Measure Distance and the Decimal Scaling normalization approach. Result Analysis is taken on various datasets from UCI machine dataset repository and shows that Mk-Means is advantageous and improve the effectiveness with normalized approach and Minkowski distance measure[9].

Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem study various distance measures and their effect on different clustering techniques. In addition to the standard Euclidean distance, they use Bit-Vector based, Comparative Clustering based, Huffman code based and Dominance based distance measures. They cluster both synthetic datasets and one real life dataset using the above distance measures by employing k-means, matrix partitioning and dominance based clustering algorithms. Finally analyze the results of their study using a real life dataset of cricket and compare the accuracy of various techniques using synthetic datasets[5].

Sudipta Acharya and Sriparna Saha have developed an automatic clustering technique using the search capability of multiobjective optimization which can automatically determine the relevant distance measure and the corresponding partitioning from a given data set. Proposed automated framework is generic in nature *i.e.*, any number of different distance measures can be incorporated into it[11].

III. TYPES OF DATA AND PROXIMITY MEASURES

Data sets are made up of data objects. A data object represents an entity. Data objects are typically described by attributes. An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature. The type of an attribute is determined by the set of possible value such as nominal, binary, ordinal, or numeric. Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. These also referred to as categorical. The values do not have any meaningful order. In computer science, the values are also known as enumerations. A binary attribute is a nominal attribute with only two categories or states: 0 or 1. A binary attribute is symmetric if both of its states are equally valuable and carry the same weight. binary attribute is asymmetric if the outcomes of the states are not equally important. An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled. Interval-scaled attributes are measured on a scale of equal-size units. A ratio-scaled attribute is a numeric attribute with an inherent zero-point[13].

Classification algorithms developed from the field of machine learning often talk of attributes as being either discrete or continuous. Each type may be processed differently. A discrete attribute has a finite or countable infinite set of values, which may or may not be represented as integers. If an attribute is not discrete, it is continuous. The terms numeric attribute and continuous attribute are often used interchangeably in the literature[12].

Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers. Central tendency which measure the location of the middle or center of a data distribution. Intuitively speaking, given an attribute, where do most of its values fall? In particular, we discuss the mean, median, mode, and midrange. Dispersion of the data will explain how are the data spread out? The most common data dispersion measures are the range, quartiles, and interquartile range; the five-number summary and boxplots; and the variance and standard deviation of the data. [13]

Table 1: Comparison of different types of attributes

Types		**Description**	**properties it possesses**	**Transformation**	**Operations**	**Examples**
qualitative	Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	distinctness	Any permutation of values	mode, entropy, contingency correlation, χ^2 test	zip codes, employee ID numbers, eye color, sex: {male, female}
	ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	distinctness & order	An order preserving change of values, i.e., new_value = f(old_value) where f is a monotonic function.	median, percentiles, rank correlation, run tests, sign tests	hardness of minerals, {good, better, best}, grades, street numbers
quantitative;	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	distinctness, order & addition	new_value =a * old_value + b where a and b are constants	mean, standard deviation, Pearson's correlation, t and F tests	calendar dates, temperature in Celsius or Fahrenheit
	Ratio-Scale	For ratio variables, both differences and ratios are meaningful. (*, /)	distinctness, order & addition, Multiplication	new_value = a * old_value	geometric mean, harmonic mean, percent variation	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current

Similarity and dissimilarity measures referred to as measures of proximity. Similarity and dissimilarity are related. A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unalike. The higher the similarity value, the greater the similarity between objects. A dissimilarity measure works the opposite way. It returns a value of 0 if the objects are the same. The higher the dissimilarity value, the more dissimilar the two objects are.

The data matrix (used to store the data objects) and the dissimilarity matrix (used to store dissimilarity values for pairs of objects) two data structures that are commonly used in the above types of applications. Object dissimilarity can be computed for objects(i,j) described by nominal attributes computed based on the ratio of mismatches. Dissimilarity and similarity measures for objects described by either symmetric or asymmetric binary attributes Simple Matching Coefficient ,Jacard Coefficient, Hamming distance. Dissimilarity of objects described by numeric attributes. Using measures such as Euclidean, Manhattan, and Minkowski, supremum distance (or) Chebyshev distance distances. Cosine similarity is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. For applications involving sparse numeric data vectors, such as term-frequency vectors, the cosine measure and the Tanimoto coefficient are often used in the assessment of similarity. Sung-Hyuk Cha listed various similarity measures formulae [6].

In many real databases, objects are described by a mixture of attribute types. In general, a database can contain nominal, symmetric binary, asymmetric binary, numeric, or ordinal attribute types. Compute the dissimilarity between these mixed objects follows a method is to group each type of attribute together, performing separate data mining (e.g., clustering) analysis for each type. This is feasible if these analyses derive compatible results. However, in real applications, it is unlikely that a separate analysis per attribute type will generate compatible results.[12]

IV. CLUSTERING

Cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Cluster analysis has extensive applications, including business intelligence, image pattern recognition, Web search, biology, and security. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters.

Clustering is a dynamic field of research in data mining. It is related to unsupervised learning in machine learning. Many clustering algorithms have been developed. These can be categorized from several orthogonal aspects such as those regarding partitioning criteria, separation of clusters, similarity measures used, and clustering space. Fundamental clustering methods of the following categories are partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Some algorithms may belong to more than one category.

Table-3: Comparison of different clustering methods

Method	**General Characteristics**
Partitioning	Find mutually exclusive clusters of spherical shape Distance-based May use mean or medoid (etc.) to represent cluster center Effective for small- to medium-size data sets
Hierarchical	Clustering is a hierarchical decomposition (i.e., multiple levels) Cannot correct erroneous merges or splits May incorporate other techniques like micro clustering or consider object “linkages”
Density-based	Can find arbitrarily shaped clusters Clusters are dense regions of objects in space that are separated by low-density regions Cluster density: Each point must have a minimum number of points within its “neighborhood” May filter out outliers
Grid-based	Use a multi resolution grid data structure Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

A partitioning method first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include k-means, k-medoids, and CLARANS.A hierarchical method creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (e.g., in Chameleon), or by first performing microclustering (that is, grouping objects into “microclusters”) and then operating on the microclusters with other clustering techniques such as iterative relocation (as in BIRCH). A density-based method clusters objects based on the notion of density. It grows clusters either according to the density of neighborhood objects (e.g., in DBSCAN) or according to a density function (e.g., in DENCLUE). OPTICS is a density-based method that generates an augmented ordering of the data’s clustering structure.A grid-based method first quantizes the object space into a finite number of cells that form a grid structure, and then performs clustering on the grid structure. STING is a typical example of a grid-based method based on statistical information stored in grid cells. CLIQUE is a grid-based and subspace clustering algorithm. Clustering evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method. The tasks include assessing clustering tendency, determining the number of clusters, and measuring clustering quality.

V. RESULTS & DISCUSSION

SPMF is an open-source data mining library written in Java, specialized in pattern mining.It is distributed under the GPL v3 license.It offers implementations of 143 data mining algorithms for association rule mining,itemset mining, sequential pattern,sequential rule mining, sequence prediction, periodic pattern mining, episode mining,high-utility pattern mining, time-series mining. clustering and classification, The source code of each algorithm can be easily integrated in other Java software. Moreover, SPMF can be used as a standalone program with a simple user interface or from the command line.SPMF is fast and lightweight [14].

Hesamoddin Salehian designed a method of image retrieval based on color and texture of the image which uses a relatively short feature vector. We have tested on SIMPIcity database consist of 1000 images of 10 different categories of Africa people & villages, Beach Buildings, Buses, Dinosaurs, Elephants, Flowers , Horses ,Mountains & glaciers, Food. They extracted features in 3 steps. The first step is the Color Extraction. It computes the feature vector consisting of 18 real numbers for each image. Please note that for each of the features, a normalization step is provided to fit the numbers between 0-10. Then the edge map of the given image is calculated. It is implemented in Canny Edge Detector class. It provides a binary image which consists of just zeros and ones. The last step is Texture Extraction. It uses co-occurrence matrix to extract texture features, and provides 48 real numbers as feature vector. At last, these two vectors are concatenated to each other and a 66 length feature vector for each image is computed. Extracted all features once and is stored in Features.txt[3].

SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0. Given two clusters, we can choose the one with the smallest error; one easy way to reduce SSE is to increase K, the number of clusters. A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

After applying K-Means Algorithm on features of images of SIMPIcity.

Table 3: Comparison analysis K-Means with different distance metrics

sno	Distance Metric	Total time(ms)	SSE (Sum of Squared Errors)	Max memory(in mb)	Iteration count	Final No. of Clusters
1	Euclidian	13379	528.4465887294608	3.9322738647460938	65	8
2	Manahan	257	24178.05100046572	3.074249267578125	29	7
3	Chebyshev	282	222.16276466267334	2.9964218139648438	22	10
4	Canberra	391	52993.550143874796	3.3418121337890625	38	6

VI. CONCLUSION

K-Means using Chebyshev distance measure takes less time, occupies less space with minimum number iterations and SSE.

REFERENCES

[1]. Madhavi Alamuri; Bapi Raju Surampudi; Atul Negi,," A survey of distance/similarity measures for categorical data",2014 International Joint Conference on Neural Networks (IJCNN),Year: 2014, Pages: 1907 - 1914

[2]. Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12):e0144059.doi:10.1371

[3]. Hesamoddin Salehian, Fatemeh Zamani, Mansour Jamzad, "Fast Content Based Color Image Retrieval System Based on Texture Analysis of Edge Map", September, 2011, www.scientific.net.

[4]. Vaishali R. Patel & Rupa G. Mehta ,Data Clustering: Integrating Different Distance Measures with Modified k-Means Algorithm, Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011 pp 691-700

[5]. Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem,"An Experiment with Distance Measures for Clustering", International Conference on Management of Data COMAD 2008, Mumbai, India,

December 17–19, 2008 , Computer Society of India,

[6]. Sung-Hyuk Cha “,Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions” ,International Journal Of Mathematical Models And Methods In Applied Sciences.

[7]. Dr. Meenakshi Sharma & Anjali Batra, "Analysis of Distance Measures in Content Based Image Retrieval",Global Journals Inc. (USA),Online ISSN: 0975-4172 & Print ISSN: 0975-4350,Volume 14 Issue 2 Version 1.0 Year 2014

[8]. Fazal Malik ,”Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain” ,Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia,2012.

[9]. Sanjay Patil ,Sanjay Talbar,"Content Based Image Retrieval Using Various Distance Metrics",Data Engineering and Management pp 154-161.

[10]. Dr. Meenakshi Sharma & Anjali Batra, "Analysis of Distance Measures in Content Based Image Retrieval",Global Journals Inc. (USA),Online ISSN: 0975-4172 & Print ISSN: 0975-4350,Volume 14 Issue 2 Version 1.0 Year 2014

[11]. Sudipta Acharya and Sriparna Saha,” Importance of proximity measures in clustering of cancer and miRNA datasets: proposal of an automated framework”, Molecular Bios stems Issue 11, 2016

[12]. Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.

[13]. Introduction to Data Mining Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison-Wesley,2005.ISBN : 0321321367.

[14]. http://www.philippe-fournier-viger.com/spmf/

*Venkataramana Battula. “Evaluation of Images Using Various Distance Metrics.” International Journal Of Engineering Research And Applications (IJERA), vol. 08, no. 01, 2018, pp. 29–34.