International Journal of Engineering Research and Development
e-ISSN : 2278-067X, p-ISSN : 2278-800X, www.ijerd.com
Volume 2, Issue 8 (August 2012), PP. 29-32

A Novel Distance Similarity Measure on Learning Techniques & Comparison with Image Processing

***Abstract*—Clustering techniques make it possible to search large amounts of data for characteristic rules and patterns. To monitoring each data recorded on a cluster or any data mining classification, they can be used to calculate the distance similarity. In this paper, we present "Supervised & Unsupervised learning" a method similarity measures which are used for analysis. The clustering method first partitions the training instances into two clusters using Euclidean distance similarity, on each cluster, representing a density region. To analyze any technique in the mining, our work studies the best measure by using classification association with supervised & unsupervised algorithms that have not been used before. We compare the Euclidean distance similarity image processing that have the best efficiency or the best learning.**

***Keywords*—Distance Similarity, Measure, Metric, Mining Classifications, Euclidean, Huffman Code.**

I. INTRODUCTION

Cluster analysis classifies data into useful groups, useful clusters are the goal then the resulting cluster should capture the natural structure of the data. For example cluster has been used to group related documents for browsing to find genes and proteins that have similar functionality and to provide a grouping of spatial locations prone to earthquakes. In other words cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. The scope of this paper is modest: to provide an introduction to cluster similarity measures in the field of data mining, where we define data mining to be the discovery of useful, but non-obvious, information or patterns in large collections of data. Much of this paper is necessarily consumed with providing a general background for cluster similarity measure.

Clustering in general is an important and useful technique that automatically [3] organizes a collection with a substantial number of data objects into a much smaller number of coherent groups [1]. In the particular scenario of text documents, clustering has proven to be an effective approach for quite some time and an interesting research problem as well. It is becoming even more interesting and demanding with the development of the World Wide Web and the evolution of Web 2.0. For example, results returned by search engines are clustered to help users quickly identify and focus on the relevant set of results. Customer comments are clustered in many online stores, such as Amazon.com, to provide collaborative [2] recommendations. In collaborative bookmarking or tagging, clusters of users that share certain traits are identified by their annotations. Object document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting. For example, when we are grouping the objects two documents are regarded as similar if they share similar thematic topics. When clustering is employed on web sites, we are usually more interested in clustering the component pages according to the type of information that is presented in the page. For instance, when dealing with universities' web sites, we may want to separate professors' home pages from students' home pages, and pages for courses from pages for research projects. This kind of clustering can benefit further analysis and utilize of the dataset such as information retrieval and information extraction, by grouping similar types of information sources together. Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair-wised similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity and the Jaccard correlation coefficient. Meanwhile, similarity is often conceived in terms of dissimilarity or distance as well [2]. Measures such as Euclidean distance and relative entropy have been applied in clustering to calculate the pair-wise distances.

II. SECTION

Related Work: The specific problem of clustering categorical data on cluster analysis that discuss the problem of determining similarity between categorical attributes. The problem has also been studied recently. However, most of these approaches do not offer solutions to the problem discussed and the usual recommendation is to "binarize" the data and then use similarity measures designed for binary attributes. Most work has been carried out on development of clustering algorithms and not similarity functions. Hence these works are only marginally or peripherally related to our work. Wilson and Martinez [4] performed a detailed study of heterogeneous distance functions (for categorical and continuous attributes) for instance based learning. The measures in their study are based upon a supervised approach where each data instance has class information in addition to a set of categorical/continuous attributes. There have been a number of new data mining techniques for categorical data that have been proposed recently. Some of them use notions of similarity which are neighborhood-based [5] or incorporate the similarity computation into the learning algorithm [7, 8]. These measures are useful to compute the neighborhood of a point and neighborhood-based measures but not for calculating similarity between a pair of data instances. In the area of information retrieval, Jones et al. [9] and Noreault et. al [6] have studied several similarity measures. Another comparative empirical evaluation for determining similarity between fuzzy sets was performed by Zwick et al. [10], followed by several others [11].

Similar work which compares four similarity measures on a collection of Yahoo! News pages. The present study differs in two aspects. First, we extended the experiments by including the averaged KL divergence. Broadly agree with Strehl et al's [12]. We both found that the performance of the cosine similarity, Jaccard correlation and Pearson's coefficient are very close, and are significantly better than the Euclidean distance measure. In addition, we found that the KL divergence measure is comparable and in some cases better than the others. Second, we also experimented with other types of data sets in addition to the web page documents. This measure was more frequently used to assess the similarity between words, especially for such applications as word sense disambiguation. It was not until recently that this measure has been utilized for document clustering. Information theoretic clustering algorithms such as the Information Bottleneck method [13] rely on this measure and have shown considerable improvement in overall performance. Meanwhile, enhanced representation of documents has been a promising direction recently, especially the incorporation of semantic information and taking account of the semantic relatedness between documents. A number of researchers have reported results on these aspects. For example, Hotho et al. propose to extend the conventional bag of word representation with relevant terms from WordNet. Experiments on document clustering task show the effectiveness of the extended representation. Moreover, the effectiveness of different representation strategies also depends on the type of task at hand. For example, when clustering journalistic text, proper names have been found to be a more appropriate representation for the text content. This investigation differs from these strategies in that we use only the basic bag of words representation. However, combining this extended representation is likely to improve performance and this is planned for future work.

III. SECTION

Problem Definition: The number of objects in a dataset, each point is denoted by P so on grouping the number of mining technique to find the similarity between each point of object and to represent the subsets of dimensions of the points which used as indices. Our proposed work introduces the distance similarity function which efficient to implement this in novel analysis.

3.1. Data Mining Distance Measures: Number of points in the data set is denoted by N. each object is denoted by P_i P_j , K denotes the number of clusters and d denotes the number of dimensions of a point. D denotes the set of dimensions of the points P_i and P_j respectively l, m and x are used as indices.

3.1.1. Euclidean Distance for cluster similarity: An N*N matrix M_e is calculated for points with d dimensions the Euclidean distance Me(P_i,P_j) between two points Pi and Pj is defined as follows

$$\text{Me}(P_i\ P_j) = \sqrt{\sum_{x=1}^{d} (P_{i_x} - P_{j_x})^2}$$

Where P_{ix} and P_{jx} represent the xth dimension values of P_i and P_j respectively M_e is a symmetric matrix.

3.1.2. Huffman Code Similarity: A method to encode data in the form of bits based on the frequency of various values is ordered that is less the number of bits in the code greater is the frequency of that particular value and more the number of bits less is the frequency. Determining the maximum value of each dimension we divide the 0 to maximum range into a suitable number of bins. Each of the bins is assigned a unique number of string values. Based on the values of the dimensions of the points the string is assigned a frequency value. The frequencies of the strings are used to get the Huffman codes for each bin for every dimension. Hence for each dimension of every point, a Huffman code is assigned based on the frequency of the string representing the bin to which it belongs to then the Huffman codes are converted to integers. When two bins of a dimension have the same frequency the integer representation of the Huffman codes reflects the relative values of the dimension for two different points so for every point the original data values are mapped to a new set of integers, for distance similarity use the Euclidean with Huffman code.

3.2 Similarity measure for Association Rule: The basic concepts for association mining i_1,i_2,…i_m be a set of m distinct attributes also called items. A set of items is called an item set where for each non-negative integer k an item set with exactly k items is called k-item set. A transaction is set of items that has a unique identifier TID the support of an item set A in database D denoted supD(A) is the percentage of the transactions in D containing A as the subset. The item sets that meet a user specified minimum support are referred as frequent item set or as associations. An association rule is an expression of the form A => B where A and B are disjoint item sets.

Level of Confidence for Association Rule A => B

$\text{sup}_D(\text{AUB})$ $\text{sup}_D(\text{A})$

Mining Task for discovering association rules consists of finding all frequent item set and finding all rules whose confidence levels are at least a certain value the minimum confidence.

3.2.1. Measure for Similarity Association: Let X and Y be two entities whose similarity want to measures denotes Sim(X,Y) to mean the similarity measure between X and Y with the below mentioned properties.

Identity of Item set : Sim (X,Y) =1 corresponds to the fact that two entities are identical.
Distinction of Item set: Sim(X,Y) =0 corresponds to the fact that two entities are different.

3.3. Similarity Measure for Classification: A mapping between a feature space and a label where the features represent characteristics of the elements to classify and the labels represent the classes. Supervised classifier is a set of labels or categories is known in advance and we have set of label.

3.3.1 Classification Nearest Neighbors: This classifier memorizes the entire training set and classifies only if the attributes of the new record match one of the training. For example given a point to be classified the KNN classifier finds the k closest points from the training records. It assigns the class label according to the class labels of its nearest neighbors.

3.3.2. Decision Tree classification: This classifiers on a target attribute or class in the form of a tree structure, in these nodes are a single attributes value is tested to determine to which branch of the subtree applies leaf nodes which indicate the value of the target attribute.

Once the partition induced by the test condition has been done, is recursively repeated until a partition is empty or all have the same target value.

The measure for the splits can be decided by maximizing the information gain

$$\Delta_i = I(parent) - \sum_{j=1}^{k_i} \frac{N(v_j)I(v_j)}{N}$$

IV. SECTION

Distance Similarity for Other Applications: Mining is extraction of relevant data, for these distance similarity measures not only used for data mining applications it can also apply for other application such Image processing, Recognition, MatLab, etc

4.1. Euclidean Similarity for Image: The M by N images are easily discussed in an MN dimensional Euclidean space called image space, natural to adopt the base e_1,e_2,…e_{MN} to form a coordinate system where e_{kN+1} corresponds to an ideal point source with unit intensity at location (k,l) thus an image x = (x^1,x^2,…x^{MN}), where x^{kN+1} is the gray level at the (k.l) the pixel is represented as a point in the image space and x^{kN+1} is the coordinate with respect to e_{kN+1}. The origin of the image space is an image whose gray levels are zero. The algebra of the image space can be easily formulated as the Euclidean distance of images such as distance between their corresponding points in the image space could not be determined until the metric coefficients of the basis are given.

Measure g_{ii} i,j =1,2,…MN as

$$g_{ij} = < e_i, e_j > = \sqrt{< e_i, e_i >}\sqrt{< e_j, e_j >} \cdot \cos\theta_{ij}.$$

On the Euclidean Distance of Images Liwei Wang, Yan Zhang, Jufu Feng Center for Information Sciences School of Electronics Engineering and Computer Sciences, Peking University
Beijing, 100871, China {wanglw, zhangyan, fjf}@cis.pku.edu.cn

4.2. Comparative Study: Data mining clustering is a data set as a group of clusters means collections of data points that belong together, segmentation as clustering represent an image in terms of clusters of pixels that belong together. Segmentation techniques are region based techniques connected regions grouping neighboring pixels of similar intensity levels such as fragmentation blurred boundaries overlooked and connectivity preserving relaxation based technique active contour model start with some initial boundary shape and iteratively modify to some energy function.

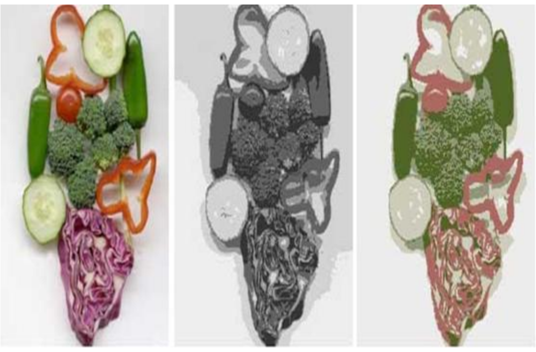

Figure 1 *shows the clustering segmentation.*

An image of mixed vegetables which is segmented using k-means to produce the images at center and on the right, each pixel is replaced with the mean value of its cluster in the center a segmentation obtained using only the intensity information and at the right a segmentation obtained using color information assumes five clusters.

V. CONCLUSION

In this paper a general data mining classification techniques uses the distance measure to group or analyze for the objects in different datasets clustering method is first applied to partitioning the training instance into some number of disjoint clusters. The decision tree built on each training or test to learn the sub classifies decision space into classification regions, thereby improving the system classification performance. For Business analysis purpose association rule mining extracts frequent item sets.

Our future direction is to utilize dependency measure to end this we designed and implemented a novel explanation mechanism for the problem of having high false can also be resolved by one such approach that uses similarity to achieve a high rate of accuracy in the case of any business method with human- understandable can also improve the efficiency when analyzing the complex dataset.

REFERENCE

[1]. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Computing Surveys (CSUR), 31(3):264–323, 1999.

[2]. G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.

[3]. P. Willett. Recent trends in hierarchic document clustering: a critical review. Information Processing and Management: an International Journal, 24(5):577–597, 1988.

[4]. D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. J. Artif. Intell. Res. (JAIR), 6:1{34, 1997.

[5]. Y. Biberman. A context similarity measure. In ECML '94, pages 49{63. Springer,1994.

[6]. C. R. Palmer and C. Faloutsos. Electricity based external similarity of categorical attributes. In PAKDD '03 pages 486{500. Springer, 2003.

[7]. Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3):283{304, 1998.

[8]. V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS{clustering categorical data using summaries. In KDD '99, New York, NY, USA, 1999. ACM Press

[9]. W. P. Jones and G. W. Furnas. Pictures of relevance: a geometric analysis of similarity measures. J. Am. Soc. Inf. Sci., 38(6):420{442, 1987.

[10]. R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. International Journal of Approximate Reasoning, 1(2):221{242, 1987.

[11]. C. P. Pappis and N. I. Karacapilidis. A comparative assessment of measures ofsimilarity of fuzzy values. Fuzzy Sets and Systems, 56(2):171{174, 1993.

[12]. A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In AAAI-2000: Workshop on Artificial Intelligence for Web Search, July 2000.

[13]. N. Z. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In Proceedings of the 37th Allerton Conference on Communication, Control and Computing, 1999.

B. V. V. S. Prasad (Ph.D) is a Research Scholar under the Guidance of Dr.G.Manoj Someswar and completed M.Tech in Computer Science & Engineering. He is presently working as an Asst. Prof. in Malineni Lakshmaiah Women’s Engineering College, Pulladigunta, Guntur, India. He is having about 5years of teaching experience in different Engineering Colleges and an associate member of CSI and life member of ISTE, Member in IEEE. Undergone CIT program conducted by IIIT Hyderabad, published 4 international journals in the area of Data Mining.

Mr. B. RAJA SEKHAR working as Assistant Professor in CSE Department in Jawaharlal Nehru Institute of Technology, Hyderabad. He completed his B.Tech in Computer Science & Engineering and M.Tech in Software Engineering from JNTU Hyderabad. He is having 4years of Teaching Experience. He is a member of IACSIT and CSTA. His interested subjects are Data Mining, Algorithms, Data Structures using C, Software Engineering, Image Processing.

Mr. L.Maruthi Working as Asst. Professor in Nalla Malla Reddy Engineering College,Hyderabad. His qualification is M.tech. He is having 4.5 years of experience in different engineering colleges. His interested area is Data Mining. He is an associate member in Computer Society of India.

Mr. V. Uday Kumar working as Asst. Professor in Swamy Ramananda Thirtha Institute of Science and Technology. His educational qualification is M.Tech in Computer Science and Engineering. He is having 4.5 years of experience. His research area is Data Mining.