iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** V **Month of publication:** May 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering

Deepak Sinwar[#1], Rahul Kaushik[*2]

[#]*Assistant Professor*, [*]*M.Tech Scholar*

Department of Computer Science & Engineering

BRCM College of Engineering & Technology, Bahal

Abstract*— Clustering is the task of assigning a set of objects into groups called clusters in which objects in the same cluster are more similar to each other than to those in other clusters. Generally clustering is used to find out the similar, dissimilar and outlier items from the databases. The main idea behind the clustering is the distance between the data items. The work carried out in this paper is based on the study of two popular distance metrics viz. Euclidean and Manhattan. A series of experiments has been performed to validate the study. We use two real and one synthetic datasets on simple K-Means clustering. The theoretical analysis and experimental results show that the Euclidean method outperforms Manhattan method in terms of number of iterations performed during centroid calculation.

***Keywords* - Clustering, Euclidean, Manhattan, Distance, K-Means, Outliers**

I. INTRODUCTION

Data mining is one of the essential step of the "Knowledge Discovery from Databases (KDD)" process, a pretty young and interdisciplinary field of computer science, is the process that attempts to discover interesting yet hidden patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management ways, data processing, model and inference considerations, complexity considerations, post-processing of discovered structures, visualization, and online updating. Commonly used data mining tasks [1] are classified as:

Classification– is the task of generalizing well-known structure to apply to new data for which no classification is present. For example, classification of records on the bases of the 'class' attribute. Prediction and Regression are also considered as a part of classification methods.

Clustering– 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. Anomaly detection (Outlier/change/deviation detection) is also considered as a part of clustering techniques. This step generally used for identification of unusual/ abnormal data records or errors, which can be interesting sometimes. In both the cases outliers may require further investigation and processing.

Association rule mining (Dependency modelling) – It is the task of finding interesting associations between various attributes of the dataset. The associations are generally based on the newly, interesting yet hidden patterns. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

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. But

how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). Distance metric is the main criteria in clustering like 'Simple K-means'. Generally we may prefer to choose either Euclidean or Manhattan for this purpose.

In the next section of the paper we describe the background work related to the clustering and distance metrics, whereas section III will elaborate the experimental work and conclusion with future work will be discussed in section IV.

II. Background Work

Several distance metrics, such as the L1 metric (Manhattan Distance), the L2 metric (Euclidean Distance) and the Vector Cosine Angle Distance (VCAD) have been proposed in the literature for measuring similarity between feature vectors [6]. Measures of distance have always been a part of human history. Euclidean Distance (ED) is one such measure of distance and its permanence is a testament to its simplicity and longevity. Although other methods to determine distance have emerged, Euclidean Distance remains the conventional method to measure distance between two objects in space, when that space meets the prerequisites [9]. The Euclidean distance [7] of two n-dimensional vectors, x and y, is defined as:

$$d(i,j)=\sqrt{\sum_{i=1}^{n}(x_{i1}-y_{i1})^2+(x_{i2}-y_{i2})^2+.....+(x_{in}-y_{in})^2}$$

and Manhattan (or city block) distance [7] is defined as :

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+.....+\left|x_{in}-x_{jn}\right|$$

Manhattan and Euclidean are popularly used distance metrics to determine the similarities between the data items of a cluster. Vadivel et al. [2], has performed their comparative study of both these methods in the application of image retrieval system. From their experimental results they conclude that the Manhattan distance gives the best performance in terms of precision of retrieved images. There may be cases where one measure performs better than other; which is totally depending upon the criterion adopted, the parameters used for validation the study etc. There are two categories of clustering algorithms (Kaugfman and Rousseeuw, 1990): Partitioning Clustering and Hierarchical Clustering. Partitioning Clustering (PC) (Duda and Hart, 1973, Kaugfman and Rousseeuw, 1990) starts with an initial partition, then tries all possible moving or swapping of data points from one group to another iteratively to optimize the objective measurement function. Each cluster is represented either by the centroid of the cluster (KMEANS), or by one object centrally located in the cluster (KMEDOIDS). It guarantees convergence to a local minimum, but the quality of the local minimum is very sensitive to the initial partition, and the worst case time complexity is exponential. Hierarchical Clustering (HC) (Duda and Hart, 1973, Murtagh, 1983) does not try to find the 'best' clusters, instead it keeps merging (agglomerative HC) the closest pair, or splitting (divisive HC) the farthest pair, of objects to form clusters. With a reasonable distance measurement, the best time complexity of a practical HC algorithm is O(N2). On the other hand an efficient and scalable data clustering method was proposed [13] named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), based on a new in-memory data structure called CF-tree, which serves as an in-memory summary of the data distribution. They have implemented it in a system and studied its performance extensively in terms of memory requirements, running time, clustering quality, stability and scalability; they also compare it with other available methods. Finally, BIRCH is applied to solve two real-life problems: one is building an iterative and interactive pixel classification tool, and the other is generating the initial codebook for image compression. Most data clustering algorithms in Statistics are distance-based approaches. That is, they assume that there is a distance measurement between any two instances (or data points), and that this measurement can be used for making similarity decisions; and (2) they represent clusters by some kind of 'center' measure.

III. Experimental Work

A series of experiments has been performed to validate the study. We have used two real and one synthetic data sets viz. Iris, Diabetes and BIRCH respectively for our purpose. The details of the datasets are given below:

ISSN: 2321-9653

Relation Name	Number of Instances	Type of Data
Iris	150	Numeric
Diabetes	768	Numeric
BIRCH (Synthetic)	136	Numeric

All the experiments were performed on Intel® Core™i3-370M, with 2GB DDR3 Memory. We have used WEKA 3.7.10 as our development tool for clustering of data items. The Table-I depicts the summary of the experiments.

Table-I: Summary of iterations performed by Simple K-Means Clustering using Euclidean and Manhattan distances

Num_ Cluster	Number of Iterations					
	BIRCH		IRIS		DIABITIES	
	Eucli-dean	Man-hattan	Eucli-dean	Man-hattan	Eucli-dean	Man-hattan
2	2	3	7	6	4	6
3	3	5	3	3	13	21
4	6	6	4	4	12	16
5	6	7	4	4	16	9
6	4	7	6	7	9	11
7	4	11	6	7	18	13
	25	39	30	31	72	76
Total No. of Iterations performed using Euclidean Distance:						127
Total No. of Iterations performed using Manhattan Distance:						146

As shown in Table-I, these three datasets were tested for studying the two basic distance metrics viz. Euclidean and Manhattan on Simple K-Means clustering method provided within the WEKA data mining tool. Other configurations are unaltered except the '*numClusters'* parameter, whose value is taken between 2 to 7 for all experiments. Clustering has been performed on these datasets one by one, by setting once the *distanceFunction* as *EuclideanDistance* function and *ManhattanDistance* on the other hand.

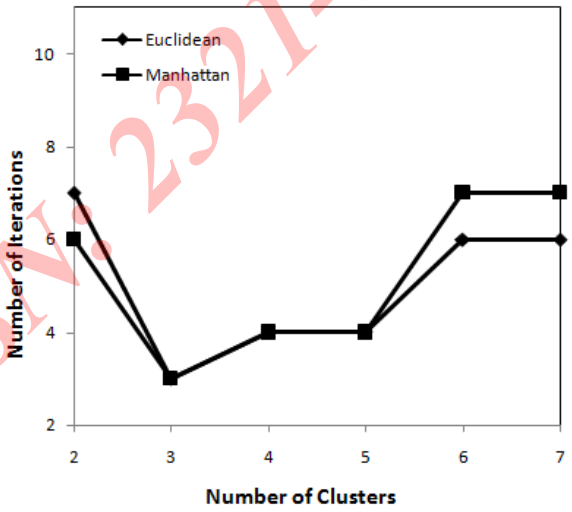


Fig. 1: Number of iterations performed on Iris Dataset

ISSN: 2321-9653

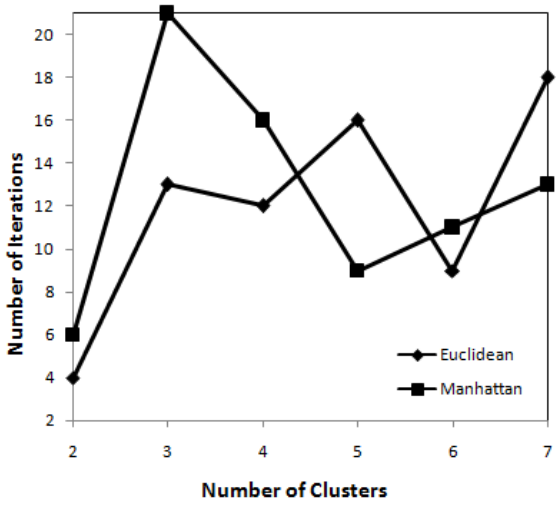


Fig. 2: Number of iterations performed on Diabetes Dataset

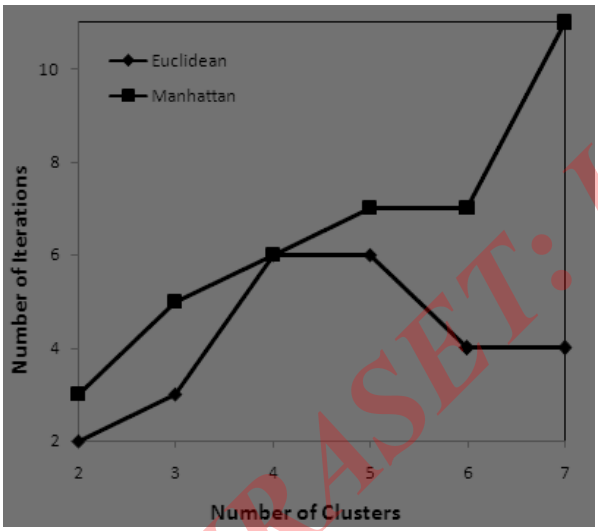


Fig. 3: Number of iterations performed on BIRCH Dataset

As shown in Figure 1, 2 and 3, that the number of iterations performed by Simple K-Means Clustering using Euclidean distance are less as compared to the Manhattan distance. These iterations are counted simply in the calculation of centroid points during the overall clustering process.

IV. Conclusions

This paper focus on the study of two popular distance metrics viz. Euclidean and Manhattan, which are generally uses during the clustering process. We have used Simple K-Means as the clustering mechanism to validate our study. The theoretical analysis and experimental results show that the Euclidean method outperforms Manhattan method in terms of number of iterations performed during centroid calculation. Two real and one synthetic datasets are used during the overall process of comparative study. This work may be extended by taking more clustering algorithms with high dimensional real datasets.

References

[1] A. Ghoting, S. Parthasarathy & M.E. Otey, " Fast mining of distance-based outliers in high-dimensional datasets", Data Mining & Knowledge Discovery, 2008, 16, pp. 349–364

[2] A. Vadivel, A.K.Majumdar & S. Sural, " Performance comparison of distance metrics in content based Image retrieval applications", Intl. Conference on Information Technology

[3] F. Angiulli, " Distance-based outlier queries in data streams: the novel task and algorithms", Data Min Knowl Disc, Springer, 2010, 20, pp. 290–324

[4] F. Angiulli, R. Ben-Eliyahu & L.Palopoli, "Outlier detection using default reasoning", Elsevier's Artificial Intelligence 172 (2008) 1837–1872

[5] H. Yu, J. Yang, J. Han & X. Li, " Making SVMs Scalable to Large Data Sets using Hierarchical Cluster Indexing", Data Mining and Knowledge Discovery, 11, 295–321, 2005

[6] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 7, pp. 729-736, 1995

[7] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, USA, 2001.

[8] K.A. Yoon, D.W. Bae, "A pattern-based outlier detection method identifying abnormal attributes in software project data", Elsevier's Information and Software Technology 52 (2010) 137–151

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

[10] M. Bouguessa, " Clustering categorical data in projected spaces", Data Mining & Knowledge Discovery, Springer, 2013, 10618-013-0336-8

[11] M. Zhou & J. Tao, "An Outlier Mining Algorithm Based on Attribute Entropy", Elsevier's Procedia Environmental Sciences 11 (2011) 132 – 138

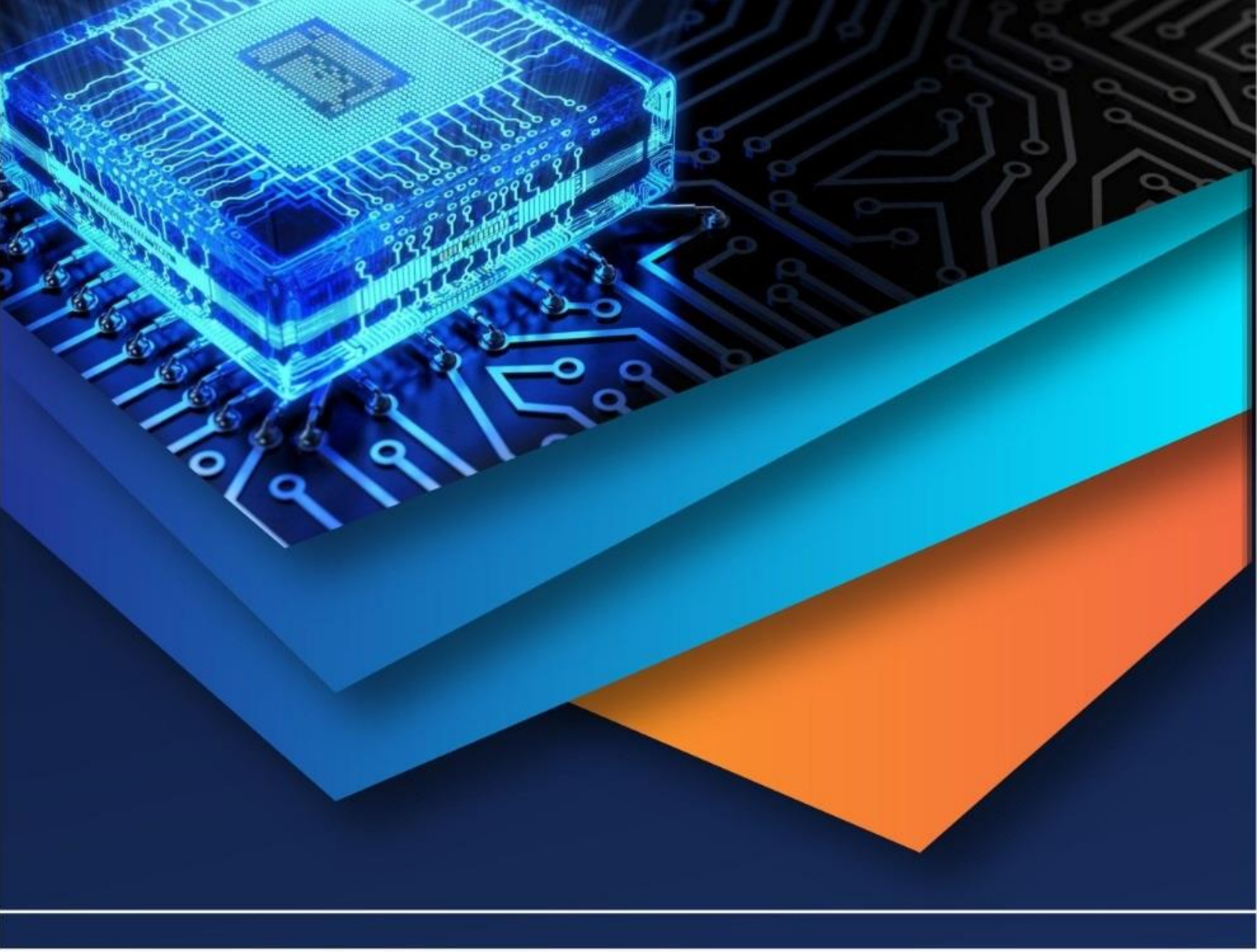
[12] S. Wu & S. Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013, pp. 589-603

[13] T.Zhang, R. Ramakrishnan & M. Livny, " BIRCH: A New Data Clustering Algorithm and Its Applications", Data Mining and Knowledge Discovery, 1, 1997, pp. 141–182

[14] W. Wong, W. Liu & M. Bennamoun, " Tree-Traversing Ant Algorithm for term clustering based on featureless similarities", Data Mining & Knowledge Discovery, 2007, 15, pp. 349–381

[15] Z. Xue, Y. Shang & A. Feng, " Semi-supervised outlier detection based on fuzzy rough C-means clustering", Elsevier's Mathematics and Computers in Simulation 80 (2010) 1911–1921

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY