IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Special Issue-3, May 2016 E-ISSN: 2347-2693

Grouping of Similar Handwritten Devanagari Scripts Using Different Distance Measures for Grid Based Approach

Prathima Guruprasad[1] and Vijayalakshmi B[2]

[1]*Department of Computer Science, Vivekananda Institute of Technology, Karnataka-India*

Available online at: www.ijcseonline.org

***Abstract*—** Due to increase in the amount of data, it is important to find useful information from data which is the main objective of data mining. Clustering is one of the techniques of data mining. Data clustering is the process of grouping similar data into same clusters. A Clustering Algorithm partitions a data set into several groups such that similarity within a group is larger than other groups. This paper gives the insight of grouping similar handwritten Devanagari words using STING algorithm. We take a wide view of the possible grouping using different distance measures on STING algorithm, compare their results and try to increase efficiency and decrease fault rate. The idea is to capture statistical information associated with spatial cells in such a manner that whole classes of queries and clustering problems can be answered. The most efficient implementation is one with least fault rate and that best distance measure to be considered to cluster the similar handwritten Devanagari scripts using STING algorithm**.**

***Keywords*—**Data mining, Distance measures, Clustering, Grouping, Devanagari, STING algorithm

I. INTRODUCTION

Data mining is a task of analyzing data and summarizing it into useful information. Data mining are used to collect, transform and store transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, provide data access to industries analysts and information technology professionals, Analysis of the data is done various application software, present the data in a useful format such as a graph or table [4]. Cluster Analysis is an automatic process to find same objects from a database. It is a fundamental operation in data mining.

Cluster is a grouping of data objects that are same to one another within the same cluster and are dissimilar to the objects in other clusters. In data mining the data is mined using two learning approaches.

Supervised Learning

In supervised learning data includes both the input and the desired output. These methods are fast and correct. The accurate results are known and are given in data inputs to the model during the learning process. Supervised models are neural network, Decision trees.

Unsupervised Learning

This model doesn't provide the correct results during the learning. It is used to cluster the input data in classes based on the basis statistical properties only. Unsupervised models are different types of clustering, distances and normalization, k-means, self organizing maps.

Data Clustering Techniques

- Partitioning Clustering

Partitioning algorithms arranges all the objects into various partitions. The reason of dividing the objects into several subsets is that checking all possible subset systems is computationally not feasible [5] [6].

The number of partitions k is less than the total number of objects n. Examples are K-means, K-medoids.

- Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent.

In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K number of clusters. Examples are AGNES (Agglomerative Nesting), CURE (Clustering Using Representatives), BIRCH (Balanced iterative reducing and clustering using hierarchies).

- Density based Clustering

Density-Based Clusters are referred as areas of higher density than the remainder of the data set. Objects in these dense areas that are required to separate clusters are usually considered to be noise and border points. It requires just two parameters and is mostly in sensitive to the ordering of the database [7].The quality of density-based clustering

Prathima Guruprasad,
Dept. of Computer Science, Vivekananda Institute of Technology, Karnataka
Vijayalakshmi B, vijayalakshmibj4@gmail.com
Dept.of Computer Science, Vivekananda Institute of Technology, Karnataka

depends on the distance measure used in the function. Examples are DBSCAN (Density-based spatial clustering of applications with noise), OPTICS (Ordering points to identify the clustering structure).

- Grid based Clustering

The Grid-Based type of clustering approach uses a multi resolution grid data structure. It minimizes the object space into a finite number of cells that form a grid structure on which operations for clustering are performed. The grid-based clustering approach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points. Examples are STING (Statistical Information Grid), CLIQUE (Clustering in Quest).

Section 1 deals with introduction of clustering techniques. Section 2 gives detail information about the existing work. Section 3 describes the proposed work which is carried out with STING algorithm and different distance measures and it is followed by conclusion in Section 4. Finally acknowledgment and references.

II. LITERATURE SURVEY

The grid-based clustering approach uses a multi resolution grid data structure [1]. It minimizes the space into a finite number of cells which form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time which is independent of the number of data objects and dependent on only the number of cells in each dimension.

Grid-based clustering algorithms are best in mining large multidimensional datasets. These algorithms divide the data space into a finite number of cells to form a grid structure and then form clusters from the cells in the grid structure. Clusters correspond to regions that are denser in data points than their surroundings [3]. The great advantage of grid-based clustering is reduction in time complexity, especially for very large data sets. Rather than clustering the data points directly, grid-based approaches cluster the neighborhood surrounding the data points represented by cells.

Some grid-based clustering algorithms also combine hierarchical clustering or subspace clustering in order to organize cells based on their density, for example Sting works with numerical attributes [2]. It is a multi-resolution clustering technique. Information such as mean, maximum and minimum is pre computed and stored in rectangular cells. Parameters at the higher level cells are drawn from the parameters of the bottom level cells.

For each cell, there are attribute independent parameters and attribute dependent parameters. First, a layer is determined from which query processing is to begin. This layer may consist of small number of cells. For each cell in this layer we check its pertinence by computing confidence internal. Irrelevant cells are removed and this process is repeated until the bottom layer is reached.

STING does not consider the spatial relationship between the children and their neighboring cells for construction of the parent cell. As a result, the shapes of the resulting clusters are isothetic, that is, all of the cluster boundaries are either horizontal or vertical and no diagonal boundary is detected. This may lower the quality and accuracy of the clusters despite the fast processing time of the technique.

III. PROPOSED WORK

In the present work we focus on the design of identifying the Devanagari words using grid-based clustering with STING algorithm. Different distance measures used as comparative analysis to find the most efficient distance measure that works well with STING algorithm [5].

The STING algorithm is implemented in different ways using different distance measures at a time. The inputs are fed to algorithm implemented in all different ways and the results are compared. The distance measure that yields good results with the algorithm is considered to be the best.

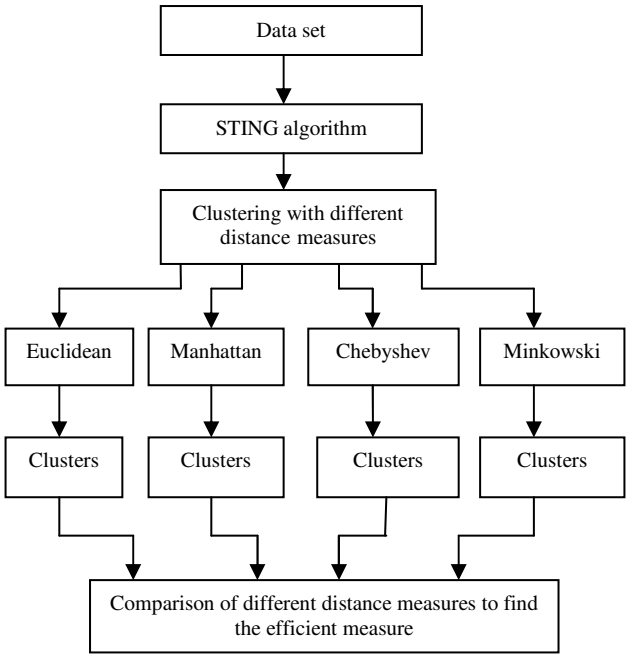

Fig.1 Flow of the proposed model from input data set to result comparison

3.1 Algorithm: STING Algorithm

1: Determine a level to begin with.
2: For each cell of this level, we calculate the confidence interval (or estimated range) of probability that this cell is relevant to the query.

3: From the interval calculated above, we label the cell as relevant or not relevant.
4: If this level is the leaf level, go to Step 6; otherwise, go to Step 5.
5: We go down the hierarchy structure by one level. Go to Step 2 for those cells that form the relevant cells of the higher level.
6: If the specification of the query is met, go to Step 8; otherwise, go to Step 7.
7: Retrieve those data fall into the relevant cells and do further processing. Return the result that meets the requirement of the query. Go to Step 9.
8: Find the regions of relevant cells. Return those regions that meet the requirement of the query. Go to Step 9.
9: Stop.

3.2 Different distance measures used:

- Basic Euclidean Distance Measure

Euclidean distance is the most commonly used, it calculate the root of square differences between coordinates of two objects.

$$D_{XY} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})}^{2}$$

- Manhattan Distance

Manhattan distance or city block distance represents distance between two points in a city road grid. It computes the absolute differences between coordinates of two objects.

$$D_{XY} = \sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|$$

- Chebyshev Distance

Chebyshev distance is also known as Maximum value distance. It calculates absolute magnitude of the differences between coordinates of two objects.

$$D_{XY} = \max_{k}\left|x_{ik} - x_{jk}\right|$$

- Minkowski Distance

Minkowski distance is the generalized metric distance.

$$D_{XY} = (d\sum_{k=1}^{d}\left|x_{ik} - x_{jk}\right|^{1/p})^{p}$$

IV. CONCLUSION

Clustering is important in the process of data mining and data analysis. Here we present a Statistical Information Grid-based approach to spatial data mining. In the design we deployed many different measures and to determine the most efficient distance measure that works well with STING algorithm. This analysis also gives us the pros and cons of using different distance measures with STING algorithm.

ACKNOWLEDGMENT

This project is funded by VISION GROUP OF SCIENCE AND TECHNOLOGY (VGST), Department of IT, BT and S&T, Government of Karnataka. We are grateful to Dr. S. Ananth Raj, Consultant, VGST and Dr. G.K Narayana Reddy, President, VKIT for their whole hearted support.

REFERENCES

[1] Wei. Wang, Jiong Yang and Richard Muntz. "STING: A statistical information grid approach to spatial data mining". Proceeding VLDB '97 proceedings of the 23rd International conference on Very Large Data bases, 1987.
[2] T. Zhang, R. Ramakrishnan and M. Livny. "BIRCH: an efficient data clustering method for very large databases". Proc. 1996 ACMSIGMOD Int. Conf Management of Data, pp. 103-l14, Montreal, Canada, June 1996.
[3] Martin Ester, Hans-Peter Kriegal, Jorg S, Xiaowei Xu. "A Density-Based Algorithm for clustering in large spatial databases with noise", Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996.
[4] Jiawei Han and M Kamber, Data Mining: Concepts and Techniques, 2001 (Academic Press, San Diego, California, USA).
[5] T Soni Madhulatha. "An Overview of Clustering Methods" IOSR Journal of engineering Vol. 2(4) pp: 710-725.
[6] R Pushpalatha and Dr K. Meenakshi Sundaram. "Survey paper on clustering techniques in data mining" International journal of advanced research in data mining and cloud computing (ijarcsa) Vol. 3, Issue 2, 2015.
[7] Jaskaranjit Kaur and Gurpreet Kaur, "Clustering Algorithms in Data Mining: A Comprehensive Study", International Journal of Computer Sciences and Engineering, Volume-03, Issue-07, Page No (57-61), Jul -2015, E-ISSN: 2347-269.

Author's Profile

Mrs Prathima Guruprasad has over 17 years of rich teaching, practical and learning experience. Prathima Guruprasad received her B.E in Computer Science and Engineering from NMAM Institute of Technology, Nitte in 1997. She received her M.Tech in Computer science and Engineering from BMS college of Engineering, Bangalore in 2005. She is currently doing her Ph.D research work in the area of Document Image Processing at Mysore University under the Guidance of Dr. Jharna Majumdar, former Scientist–G, Dean (R & D), Prof. and Head, Department of CSE (P.G), NMIT, Bangalore. The topics dealt under Document image processing include recognition of textual content in Handwritten Ancient Manuscript scanned documents, understanding of Devanagari and Nandinagari scripts and their processing. She had delivered invited talk on Nandinagari Handwritten Character Recognition Systems for Researchers Conclave on machine Learning and computational Intelligence at AIT, Coimbatore and on Various feature extraction Techniques of Nandinagari Scripts at Workshop on Emerging Trends In Image &

Video Processing at NMIT, Bangalore. She has published more than 9 research papers on national and international conferences and Journals such as IEEE, Springer, McGraw-Hill publications. She also received a Grant of Rs 30.00 Lakhs from VGST under CISEE scheme in 2014.Her research area includes document Image Processing, Pattern Recognition, Machine Learning, Computer Vision and Data Mining.

Ms Vijayalakshmi B has received B.E in Computer Science and Engineering from East West Institute of Technology in 2008. She received her M.Tech in Computer Science and Engineering from AMC college of Engineering in 2012. She has over 5 years of teaching experience. She has published 3 research papers on national and international conferences. She is having good teaching and research interests. In addition she has been involved in guiding students projects. Her general research includes Data mining, Bio-informatics,Software Engineering.