Comprehensive survey on clustering algorithms and similarity measures.

Dipali deshmukh, N.B.Pokale.

Abstract:- Text mining is the structure of data contained in natural language text . It is the series of action of extracting information from text. It consist of information retrieval, lexical analysis, pattern recognition, information extraction. The main targete in text mining is to find the similarity between documents in order to group the similar documents. The word recurrence distributions are identified to find the similarity between various documents. The overreckon goal is, essentially, to turn text into data for analysis. Vector space model was used to categorise the relevant documents. Long documents are poorly represented because they have poor similarity values and keywords must exactly match the document terms. Documents may have similar context but different term vocabulary won't be associated which tendsds to less accuracy. Illusory based approach was used and it was suitable only to short queries. In high dimensional data, the common distance. measures can be influenced by noise. Existing clustering algorithms are implemented based on partitioning, hierarchical, density based and grid based. All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper. So, in this paper we propose "multi view- point based clustering methods with similarity measure by using incremental algorithm" approach for clustering high dimensional data. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured.

INTRODUCTION:-

Clustering is an important task in data mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. A cluster is a group of data objects that are similar to one another within the same cluster and are not similar to the objects in other clusters. Clustering is been a widely studied problem in a various of application domain including neural networks. Clustering divides data into meaningful or useful groups or clusters. If meaningful clusters are the target, then the output clusters have to capture the natural structure of the data. Clustering is an most important area of research, which finds applications in all fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics etc. In this paper the analysis of cluster is doing with method K-means clustering algorithm. K-means a clustering algorithm that deals with numerical attribute values (NAs) firstly, although it could also be applied to datasets with binary values, by considering the binary values as numerical. The K-means clustering algorithm for numerical datasets requires the user to specify the number of clusters to be produced and the algorithm builds and refines the supposedly number of clusters. But due to number of repitation in the loop, the basic K-means is computationally extra time consuming and also it outcomes different results with different dataset . So the proposed K-means clustering algorithm will minimise the number of repitations and the time complexity.

2. CLUSTERING ALGORITHMS

2.1. SIMPLE KMEANS CLUSTERING

KMeans is an repitative clustering algorithm in which items are forword among set of clusters instill the desired set is got. This can be viewed as a type of squared error algorithm. The cluster mean of Ki={ti1, ti2, ….., tim}is defined as, mi= (1) Algorithm 1: K-Means Clustering Input: D={a1, t2, ….., tm} // set of elements. k // number of desired clusters. Output: K // set of clusters. Procedure: assign initial values for means a1, a2, ….ak; repeat assign each item ai to the cluster which has the closest mean; calculate new mean for each cluster; until convergence criteria is met; In almost all cases, the simple KMeans clustering algorithm takes more time to form clusters. So it is not suitable to be employed for large datasets.

2.3. EFFICIENT KMEANS CLUSTERING .

In every iteration, the k-means algorithm computes the distances between data point and all centers; this is computationally very expensive especially for huge datasets. For each data point, the distance can be kept to the nearest cluster. At the next iteration, compute the distance to the previous nearest cluster. By comparing the old distance with new distance, and if it is less than or equal, then the point will be in the same cluster. This saves the time required to compute distances to k−1 cluster centers. Two functions are written to implement efficient KMeans clustering algorithm . The first is the simple KMeans, which calculates the nearest point of center. This is done by computing the distances to all centers. Each data point keeps its distance to the nearest center.

C. FARTHEST FIRST CLUSTERING .

Farthest first is a different than K Means. it placed the cluster center at the point other than the present cluster. This point have to lies within the data area. The points that are further are clustered together first. This component of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed.

2.4. MAKEDENSITYBASED CLUSTERING .

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular. The make density based clustering algorithm can also be used in noise and when outliers are encountered. The points with same density and present within the same area will be connected to form clusters. Algorithm 3: Density based Clustering .

1. Compute the ε -neighborhood for all objects in the data space.
2. Select a core object CO.
3. For all objects co □ CO, add those objects y to CO which are density connected with co. Proceed until no further y are encountered.
4. Repeat steps 2 and 3 until all core objects have been processed.

3.EXISTING SIMILARITY MEASURES

3.1 CHI - SQUARED .

Karl Pearson in 1900 proposes the chi-squared Statistic . It examines whether there exist, any association between the categorical variable. Range exists between -1 to +1 for two variables and 0 to +1 for larger number of variable. The value more close to 1 indicates a strong relationship between variables. The chi square ($x2$) formula is defined as,

$$X^2 = \sum_{i=1}^{n} (O_i - E_i)^2 / E^2$$

where, O_i represent observed value and E_i represent Expected value.

Steps in Chi Square Test:

1. Given Observed frequency

2. Note the Expected frequency

3. Apply the chi square formula

4. Find the degree of freedom($df = N - 1$)

5. If the obtained value is equal or greater than the chi square table reject the null hypothesis.

Advantage of Chi square is it requires no assumptions about the shape of the population distribution from which a sample is drawn. It can be applied to nominal or ordinal measured variables. Limitation of Chi square similarity are, 1) need quantitative data, 2) sensitive to sample size, 3) does not give much information about the strength of the relationship and 4) Expected frequency should not be less than 1.

3.2COSINE SIMILARITY .

Cosine similarity is a popular method for text mining. It is used for comparing the document (word frequency) and finds the closeness among the data points in clustering. Its range lies between 0 and 1. The similarity between two terms X and Y are defined as follows.

CosineSim (X, Y) = (X^0 Y) / I I XI I I I YI I

3.3 OVERLAP .

The overlap is measure counts the number of attribute that matches the two data instance. It uses only the diagonal entries of the similarity matrix and sets off diagonal entries to 0 [5]. The range of per attribute value is 0 to 1. 0 indicate no match exist between the attribute and 1 indicates match exist between the attribute. The overlap similarity is defined as,

$$S_k(X_k, Y_k) = \begin{cases} \mathbf{1} & \text{if } X_k = Y_k \\ \mathbf{0} & \text{0therwise} \end{cases}$$

3.4DISC.

Data Intensive Similarity Measure for Categorical Data analysis (DISC) [6]. It makes use of a data structure called categorical information table (CI Table). CI table stores the co-occurrence statistics for the categorical data. The similarity between two attribute is measured using the cosine similarity measure.

3.5 DILCA

DIstance Learning in Categorical Attribute is the measure used by the author . Co-occurrence table is formed for all the features using symmetric uncertainty a matrix is generated and conditional probability is applied, the results are given to the Euclidean measure to find the similarity between the attributes.

CONCLUSION

The paper describes a review on different clustering methodologies and similarity measure associated with the categorical data clustering. The factor that affects various clustering algorithm, its advantage and limitation are discussed. Time complexities of various categorical clustering algorithms are discussed. Cluster accuracy and error rate for real world data set using different categorical clustering algorithms, parametric and non parametric version of DILCA and categorical similarity measure are illustrated.

REFERENCES

1. G. Salton, M. McGill, Eds. Introduction to Modern Information Retrieval.New York: McGraw-Hill, 1983.

2. X. B. Xue and Z. H. Zhou, "Distributional features for text categorization,"IEEE Trans. Knowl. Data Eng., vol. 21, no. 3, pp. 428–442, Mar. 2009.

3. L. A. F. Park, M. Palaniswami, and K. Ramamohanarao, "A novel document ranking method using the discrete cosine transform," IEEE Trans.Pattern Anal. Mach. Intell., vol. 27, no. 1, pp. 130–135, Jan. 2005.

4. Park L A F, Ramamohanarao K and Palaniswami M, "Fourier domain scoring: A novel document ranking method," IEEE Trans. Knowl. Data Eng., vol. 16, no. 5, pp. 529–539, May 2004.

5. K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

clustering," IEEE Trans. Knowl. Data Eng., vol. 16, no. 10, pp. 1279–1296, Oct. 2004.

6. Jiawei Han and Micheline Kamber Data Mining: Concepts and Techniques,Second Edition.

7. C. S. Li, "Cluster Center Initialization Method for K-means Algorithm Over Data Sets with Two Clusters", "2011 International Conference on Advances in Engineering, Elsevier", pp. 324-328, vol.24, 2011.

8. Mount, N. Netanyahu, C. Piatko and A. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation", "IEEE Transactions on Pattern analysis and Machine intelligence", vol. 24, no.7, 2002. [3] Y.M.

9. Cheung, "A New Generalized K-Means Clustering Algorithm","Pattern Recognition Letters, Elsevier",vol.24,issue15, 2883–2893, Nov.2003. [4] Z. Li, J.

10. Yuan, H. Yang and Ke Zhang, "K-Mean Algorithm with a Distance Based on the Characteristic of Differences", "IEEE International conference on Wireless communications, Networking and mobile computing", pp. 1-4, Oct.2008.

11. M. Erisoglu, N. Calis and S. Sakallioglu, "A new algorithm for initial cluster centers in K-Means algorithm", "Published in Pattern Recognition Letters", vol. 32, issue 14, Oct.2011.

12. D. Napoleon and P. G. Laxmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", "IEEE Trendz in Information science and computing", pp.42-45, Feb.2011.

13. Merz, P., 2003. An Iterated Local Search Approach for Minimum Sum of Squares Clustering. IDA 2003, p.286-296.

14. Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A Density- based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press, Portland, OR, p.226-231. [9] Sheikholeslami, G., Chatterjee,

15. S., Zhang, A., 1998. Wave- Cluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. Proc. 24th Int. Conf. on Very Large Data Bases. New York, p.428-439.

16. Hinneburg, A., Keim, D., 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining. New York City, NY.

17. Jain, A.K., Dubes, R.C., 1988. "Algorithms for Clustering Data". Prentice-Hall Inc. [12] Tapas Kanungo, David M. Mount, Nathan S.

18. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, VOL. 24, NO. 7, PP. 881-892, 2002. [13] Eduardo Raul

19. Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho ," A Survey of Evolutionary Algorithms for Clustering", *IEEE Trans. Syst., Man, Cybern.—Part C: Appl. And Review,* Vol. 39, No. 2,PP.133-155,2009.

20. Jiawei Han, Micheline Kamber,"Data Mining:Concepts and Techniques",Second Edition,*Elesvier Publications*,2006.

21. Zhang, T., Ramakrishnan, R., Linvy, M., 1996. BIRCH: "An Efficient Data Clustering Method for Very Large Databases". Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press, New York, p.103-114.

22. Guha, S., Rastogi, R., Shim, K., 1998. CURE: An Efficient Clustering Algorithms for Large Databases. Proc. ACM SIGMOD Int. Conf. on Management of Data. Seattle, WA, p.73-84.

23. Shi Na, L. Xumin, G. Yong, "Research on K-Means clustering algorithm-An Improved K-Means Clustering Algorithm", "IEEE Third International Symposium on Intelligent Information Technology and Security Informatics", pp.63-67, Apr.2010.

24. R. Xu and D. Wunsch, "Survey of Clustering Algorithms", "IEEE Transactions on Neural networks", vol. 16, no. 3, May 2005.

25. Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li," Automated Variable Weighting in k-Means Type Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* VOL. 27, NO. 5, PP. 657-668, 2005.

26. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proc. ACM SIGMOD Int. Conf. on Management of Data. Seattle, WA, p.94-105.

27. http://archive.ics.uci.edu/ml/datasets/Letter+Recognition
http://archive.ics.uci.edu/ml/datasets/Abalone

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

28. Yedla M, Pathakota SR and Srinivasa TM (2010) Enhancing K-means clustering algorithm with improved initial center. Intl Journal of Computer Sci and Info Tech 1 : 121–125.

29. Fahim AM, Salem AM, Torkey FA, Ramadan MA (2006) An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University SCIENCE A7:1626–1633. Available online at www.zju.edu.cn/jzus.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET