INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)
www.vmsindia.org

DETERMINATION OF NUMBER OF CLUSTERS IN CLUSTER ANALYSIS AND SIMILARITY MEASURE FOR THE CLUSTERS: A SURVEY

Pranjal S Bogawar
Priyadarshini College of Engineering, Nagpur
pbogawar@gmail.com

Abstract
Clustering plays vital role in data mining. Clustering is unsupervised classification technique. It is used to group the elements with same features. This is very useful to group the large amount of data. Newly created groups are used to analyze, summarize the huge data. Big companies can take large decisions from this summarized data. Clustering also useful in image processing.(E.g. Medical image, face detection, thumb impression etc.) This paper gives the idea about what is cluster, clustering. Clusters can easily created by finding the similarity between the data. This paper gives idea about some similarity measures and their comparative study. There are various clustering techniques. Each technique is suitable to solve particular problem as every technique has its pros and cons. Data should be classified into proper number of clusters. There are some methods which calculate proper number of clusters. This paper gives the idea about these methods.
Keywords Cluster, Cluster Analysis, EM, K-means.

Introduction

Process of Searching useful information from the large volume of raw data is called as Knowledge discovery in database. Data Mining is technique used for KDD to get hidden information. Data mining techniques are used for time-series analysis, predication, summarization, association, and Sequence Discovery. This paper gives the idea about clustering, its application, Similarity measures, Clustering algorithms and techniques to determine the number of clusters.

Material and Methods

2. What is Cluster?: Cluster is collection of objects which are â€œsimilarâ€ between them and are â€œdissimilarâ€ to the objects belonging to other clusters. In another way it is said as closely packed group (of people or things) pictorially it is shown as figure1

Result and Discussion

Methods to Determine the Number of Clusters: We discussed various methods for clustering the data. But how many clusters is always a problem. There are various techniques to determine the proper number of clusters. Easiest method for determining the number of clusters k is as follows k = â^šn where n is the number of objects to be clustered. But this technique is not the universal as sometimes all data get clustered into two clusters only. Milligan and Cooper (1985) discussed 30 different techniques to determine the number of clusters in Hierarchical Clustering. Calinski and Harabasz index, Je(2)/Je(1), C-Index, Gamma, Beale, Cubic clustering criterion, Point- Biserial, G(+) index, Mojena, Davies and Bouldin, Stepsize, Likelihood ratio, |logâ ¡(p)| , Sneath, Frey and Van Groenewoud, Log(SSB/SSW), Tau, (C) Ì…/ k^(.5) , n log(|T|/|W|), k^2 /|W|, Bock, Ball and Hall, Trace Cov W, Trace W, Lingoes and Cooper, Trace W-1 B, Generalized Distance, McClain And Rao, Mountford and |T| / |W| are the methods discussed by them. According to the clustering technique it varies. Reallocation methods like EM and K-means Clustering which moves observations iteratively from one cluster to another, number of clusters has to specify in advance. Bayesian Information Criterion is used for Reallocation methods and Hierarchical Clustering methods [17, 18, 19, 21, 22]. Sugar and James (2003) gives new idea based on distortion theory with limited parametric assumptions for finding the number of clusters using K-means Clustering algorithm. Xie-Beni Index and Kwon Index is used for partitioning while using fuzzy clustering algorithm [20].

Conclusion

Clustering is the very basic task of the data mining techniques. It is used for the automatic information retrieval from the given documents. Clustering is the unsupervised techniques. This paper gives the various clustering techniques used for text documents and images. This paper gives various distance finding methods to find the similarity between images or text documents. Comparative study of various similarity measures is done. Clustering techniques, their pros and cons are studied here. The methods which determines the number of clusters are also discussed in this paper.

Acknowledgement

Dr. K.K. Bhoyar, Professor, Information Technology, Y.C.C.E., Nagpur

Reference

1. Yun Yang,Ke Chen, â€œTemporal Data Clustering via weighted clustering ensemble with different representationsâ€ , IEEE transactions on Knowledge and Data Engineering, Vol- 23 , No 2, February 2011. pp 307-320.

2. Subhash Sharma, Ajith Kumar, â€œCluster Analysis and Factor Analysisâ€ ,

3. Liping Jing, Lixin Zhou Michael K. Ng, Joshua Zhexue Huang, â€œOntology based Distance Measure for text Clusteringâ€ , www.siam.org/meetings/sdm06/workproceed /Text%20Mining/jing1.pdfâ€Ž, pp1-8 4. Jana Vembunarayanan, â€œTF-IDF and Cosine Similarityâ€ ,October 27, 2013, http://janav.wordpress .com /2013 /10/27 / tf-idf-and-cosine-similarity

5. Michael D. Shapiro, Matthew B. Blaschko, â€œ On Hausdorff Distance Measuresâ€ , https://web.cs.umass.ed u/publication /docs/2004/UM-CS-2004-071.pdfâ€Ž, Computer Vision Laboratory University of Massachusetts Amherst, MA 01003, USA

6. Jeff Henrikson, â€œCompleteness and Total Boundedness of the Hausdorff Distanceâ€ , www-math.mit.edu/phase2 /UJM/vol1 /HAUSF.PDF pp-69-79

7. Marie Pierre Dubuisson,Anil Jain, â€œA modified Hausdorff Distance for Object matchingâ€ , Proc. International Conference on Pattern Recognition, Jerusalem, Israel, 1994, pp566-568

8. P.C. Mahalanobis, â€œ On the generalized Distance in Satasticsâ€ , In Proceedings National Institute of Science, India, Vol. 2, No. 1. (16 April 1936), pp. 49-55

9. Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik, â€œ Similarity between Euclidean and Cosine angle distance for Nearest Neighbor queriesâ€ , ACM1-58113-812-1/03/04, 2004, pp 1-6

10. N.A. Thacker, F.J.Aherne , P.I. Rockett,â€œThe Bhattacharya Metric as an Absolute Similarity Measure for Frequency Coded Dataâ€ , TIPRâ€™97, Prague1997, PP- 1-11

11. Konstantinos G.Derpanis,â€œThe Bhattacharya Measureâ€ , March 2008,PP 1-3

12. Hongbo Du, â€œ Data mining Techniques and Applicationsâ€ , Cengage Learning

13. Shunzhi Zhu,â€œInformation Retrieval using Hellinger distance and sqrt-cos similarityâ€ , Computer Science & Education(ICCSE), 2012, IEEE, PP-925-929

14. Cuixia Li, Yingjun Tan, Jinsheng Kong,â€œAn Mahalanobis distances based text clustering algorithmâ€ , Automatic Control and Artificial Intelligence (ACAI 2012)IEEE, PP-465-468

15. Sung-Hyuk Cha, â€œComprehensive Survey on Distance Similarity measures between Probability Densityâ€ ,International Jouranl of Mathematical Models and Methods in Applied Science, Issue 4,Volume -1, 2007, PP- 300-307

16. Glenn W. Milligan and Martha Cooper, â€œAn examination of procedures for determining the number of clusters in the data setâ€ , Phycometrica, Vol-50 ,No: 2, June 1985, PP 159-179

17. Chris Fraley and Adrian E. Raftery, â€œHow many cluster? Which Clustering Method? Answer vai Model based Cluster Analysisâ€ ,The Computer Journal, Vol-41, No-8, 1998, PP-578-588

18. Kyu J Han,Shrikant S. Narayan, â€œA Robust Stopping Criterion for Agglomerative Hierarchical Clusteringâ€ , Speech Analysis and Interpretation Laboratory (SAIL) Ming Hsieh Department of Electrical Engineering ,University of Southern California, Los Angeles, CA, USA, PP -1-4.

19. Qinpie Zhao, Ville Hautamaki, and Pasi FrÃ¤nti ,â€œKnee point detection in BIC for detecting the number of Clustersâ€ , ACIVS 2008, LNCS 5259, Springer 2008 pp. 664â€“673

20. Yuangang Tang, Fuchun Sun, Zengqi Sun, â€œImproved validation Index for Fuzzy Clusteringâ€ , American Control Conference June 8-10, 2005. Portland, OR, USA, PP 1120-1125

21. Scott Shaobing Chen, P.S. Gopalkrishnana,â€œClustering via the bayesian information criterion with application in speech recognitionâ€ , IEEE international Conference on Acoustics, Speech and Signal Processing, ISSN 1520-6149 Vol-2, 1998 PP-645-648

22. Stan Salvador,Philip Chan,â€œDetermining the number of Clusters/Segments in Hierarchical Clustering/ Segmentation Algorithmâ€ , ICTAIâ€™04 Proceeding of 16th IEEE international Conference on Tools with Artificial Intelligence, ISBN:0-7695-2236-X,2004, PP-576-584

23. Catherine A. Sugar And Gareth M. James, â€œFinding number of clusters in Dataset An Information Theoretic Approachâ€ , Journal of the American Statistical Association,2003, PP-750-763

24. Data Mining Techniques and Applications by Hongbo Du

25. Data Mining by Margarest H. Dunham, S. Sridhar

26. Florian Beil, Martin Ester, Xiaowei Xu, â€œFrequent term based text clusteringâ€ , SIGKDD 02 Edmonton, Alberta, Canada, ACM 1-58113-567-X/02/0007, 2002, PP-1-7

27. Anna Huang ,â€œSimilarity measure for text Document Clusteringâ€ , NZCSRSC , April 2008, PP-49-56

28. Aminul Islam And Diana Inkpen ,â€œSemantic Text Similarity using corpus based word Similarity and String Similarityâ€ , ACM Transactions on Knowledge Discovery from Data, Vol. 2, No. 2, 2008, PP-1-25