Applying Non-Parametric Clustering Algorithm on Clustering Analysis

范仕遠、陳木松

E-mail: 9304702@mail.dyu.edu.tw

ABSTRACT

Clustering analysis is not only major tools to uncover the underlying structure of a given data set, but also the promising ways to reveal the local input-output relations of a complex system. The clustering analysis is needed in a large variety of engineering and scientific problems, such as pattern recognition and classification, machine learning, computer vision, and system modeling and more. The goal of clustering analysis is to split data into a plausible number of subgroups such that the distance between objects within a subgroup is smaller than the distance between objects belonging to different subgroups. Unfortunately, without any prior knowledge the proper number of clusters is difficult to estimate. We use several useful information, such as distance matrix, covariance matrices derived by grid clustering algorithm, entropy, density and more are combined together in our research. In our research, we propose ellipse density entropy clustering algorithm because it only uses samples without any prior knowledge to reduce sensitivities about parameters and can automatically determine the number of clusters. It can avoid the situation which user set wrong parameters and wrong clustering results. In this thesis, we use gaussian distribution for (1) the separated clusters, (2) the overlapping clusters and (3) the containing noise clusters respectively and demonstrate it is feasible. We compare with Fuzzy C-means (FCM). The experiment results show that out scheme is reliable and stable.

Keywords : Data Mining, Knowledge Discovery, Clustering analysis,Entropy, Ellipse Density Entropy Clustering Algorithm

Table of Contents

REFERENCES

[1] 范仕遠、陳木松,“以熵值為主的模糊聚類分析”, 2003 人工智 慧、模糊系統及灰色系統聯合研討會.

[2] N.Zahid, M. Limouri and A. Essaid,“ A new cluster-validity for fuzzy clustering,” Pattern Recognition, vol.32 pp.1089-1097, 1999.

[3] C.E. Shannon,“ A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, pp. 379-423, 1948.

[4] I. Csiszk and J. Korner. Information Theory: Coding Theorems for Discrete Memoryless System. Academic Press, 1981.

[5] 范仕遠、陳木松,“以熵值為主的聚類個數預測法則”,彰雲嘉 2003 年研發成果聯合發表會 [6] Glenn Fung“ A Comprehensive Overview of Basic Clustering Algorithms,” May. 2001.

[7] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121-167, [8] McLachlan G. J., Basford K. E. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, 1988.

[9] Pan, W, Lin, J, & Le, C: A mixture model approach to detecting differentially expressed genes with microarray data. Technical Report 2001-011, Division of Biostatistics.

[10] S. M. Zabin and G. A. Wright, “ Non-parametric density estimation and detection in impulsive interference channels-Part II: Detectors, ” IEEE Trans. Commun., vol. 42, pp. 1698 – 1711, 1994.
[11] Ferguson, T.S., “ A Bayesian Analysis of Some Nonparametric Problems, ” The Annals of Statistics, 1, 209-230, 1973.
[12] C. Fraley and E. Raftery, “ How many clusters ? which clustering method ?, answers via model-based cluster analysis. ” Technical Report 329, Dept. of Statistics. University of Washington, Seattle, 1998.
[13] C. Bishop, Neural networks for pattern recognition. Oxford, U.K. Oxford Univ. Press, 1995.
[14] A. P. Dempster, N. M. Laird and D. B. Rubin, “ Maximum likelihood from incomplete data via the EM algorithm. ” Journal of the Royal Statistical Society Series B vol.39:1-38, 1977.
[15] Geoffrey J. McLachlan and Kaye E. Basford. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, 1988.
[16] S.J. Roberts, Dirk Husmeier, Iead Rezek, and William Penny, “ Bayesian Approaches to Gaussian Mixture Modeling, ” IEEE Trans. on PAMI, vol. 20, no. 11, Nov. 1998.
[17] J. C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. New York, 1981.
[18] J.S. Jang, C.T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing, Prentice Hall, 1997.
[19] D.E. Gustafson and W. Kessel, “ Fuzzy clustering with a fuzzy covariance matrix, ” Proc. IEEE-CDC 2, 761-766, 1979.
[20] Tran, D.; Wagner, M “ Fuzzy entropy clustering, ” Fuzzy Systems, The Ninth IEEE International Conference on, Vol.1, Pages:152 - 157 May 2000.
[21] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
[22] Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
[23] M. Ester, H. Kriegel, J. Sander, and X. Xu, ” A density-based algorithm for discovering clusters in large spatial databases with noise, ” in Proc. of the 2nd Int ’ l Conf. on Knowledge Discovery in Databases, Menlo Park, CA., pp. 226-231, 1996.
[24] W. Wang, J. Yang, and R. Muntz, “ STING : A Statistical Information Grid Approach to Spatial Data Mining, ” Technical Report CSD-97006, Computer Science Department, University of California, Los Angeles, Feb. 1997.
[25] G. Sheikholeslami, S. Chatterjee, and A. Zhang, “ WaveCluster: A multi-resolution clustering approach for very large spatial databases, ” In Proc. 1998 Int. Conf. Very Large Databases (VLDB ’ 98), pp. 428-439, New York, Aug. 1998.
[26] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “ Automatic subspace clustering of high dimensional data for data mining applications, ” SIGMOD 1998.
[27] 陳奕學,空間資料叢集演算法之設計, 義守大學資訊工程學系, 碩士論文, 2002.