應用非參數型聚類法於聚類分析

范仕遠、陳木松

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

摘 要

聚 類分析不僅是擷取取樣資料結構與特性的重要工具，也可 應用於萃取複雜系統的輸入與輸出的關係。聚類分析已廣泛應用 於許多工程與科學的領域，如圖像辨識與分類、機器學習、電腦 視覺、與系統建模等。聚類分析試圖將取樣資料分割成若干個群 組或聚類，使得同聚類的資料距離總和愈小(或相似度愈大)，而且 不同聚類的距離愈大。然而在僅有取樣資料而無任何先前知識的 情況下，估測正確的聚類數目是非常困難的工作。 本 研究應用資料的距離資訊、密度、熵值、格子聚類法、及 共變數矩陣等相關資訊，提出橢圓密度熵值聚類演算法，以僅有 取樣資料而無任何先前知識(prior knowledge)的情況下，減少參數 的敏感度即可以自動決定各個聚類群組，以解決對於取樣資料結 構與各種類型的聚類演算法的認知不足，導致於設定無意義或錯 誤的參數，以致於對最後分析結果產生誤判的情況。本論文以高 斯分佈的資料結構分別對於(1).聚類間隔較大、(2).聚類資料部份 重疊，與(3).含有雜訊等情況，以驗證本研究方法的可行性。藉由與FCM(Fuzzy C-means)的比較，實驗結果展示本法的可靠性與穩 定性。

關鍵詞 : 資料探勘、知識發掘、聚類分析、熵值、橢圓密度熵值聚類演算法

目錄

參考文獻

[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.