{tag}

{/tag}

International Journal of Computer Applications

Number 1 - Article 7

Year of Publication: 2011

Authors:

Alaa H. Ahmed

Wesam Ashour

10.5120/2999-4030

{bibtex}pxc3874030.bib{/bibtex}

Abstract

Since K-means is widely used for general clustering, its performance is a critical point. This performance depends highly on initial cluster centers since it may converge to numerous local minima. In this paper a proposed initialization method to select initial cluster centers for K-means clustering is proposed. This algorithm is based on reverse nearest neighbor (RNN)

search and coupling degree. Reverse nearest neighbor search retrieves all points in a given data set whose nearest neighbor is a given query point, where coupling degree between neighborhoods of nodes is defined based on the neighborhood-based rough set model as the amount of similarity between objects. The initial cluster centers computed using this methodology are found to be very close to the desired cluster centers for iterative clustering algorithms. The application of the proposed algorithm to K-means clustering algorithm is demonstrated. An experiment is carried out on several popular datasets and the results show the advantages of the proposed method.

Reference

- S. Theodoridis and k. Koutroumbaspattern, 2003. Pattern Recognition, 2nd edition, Elsevier.
- G. Gan et. Al. 2007 Data Clustering Theory, Algorithms, and Applications, Siam.
- J. Han and M. Kamber. 2006 Data Mining: Concepts and Techniques, 2nd edition, Elsevier.
- J. MacQueen 1967 "Some methods for classification and analysis of multivariate observation". In: Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.
- L. Kaufman and P. Rousseeuw. 1990 Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
- T. Zhang et. al. 1996 "BIRCH: An efficient data clustering method for very large databases". In Proceedings of the 1996 ACM SIGMOD international conference on management of data, pp. 103–114. New York: ACM Press.
- M. Ester et. al. 1996 "A density-based algorithm for discovering clusters in large spatial databases with noise," In Second international conference on knowledge discovery and data mining", pp. 226–231. Portland, OR: AAAI Press.
- M. Ankerst et. Al. 1999 "OPTICS: Ordering points to identify the clustering structure," In Proc. 1999 ACM-SIGMOD Int. Conf.Management of Data(SIGMOD'99), pp. 49–60, Philadelphia.
- W. Wang et. al, 1997 "STING: A statistical information grid approach to spatial data mining," In Twenty-third international conference on very large data bases, pp. 186–195.
- R. Agrawal et. Al. 1998 "Automatic subspace clustering of high dimensional data for data mining applications," In SIGMOD Record ACM Special Interest Group on Management of Data, pp. 94–105. New York: ACM Press.
- D. Fisher. 1987 "Improving inference through conceptual clustering," In Proc. 1987 Nat. Conf. Artificial Intelligence (AAAI'87), pp. 461–465, Seattle,WA.
- T. Kohonen 1990 "The self-organizing map," Proceedings of the IEEE, 78(9):1464–1480,.

- Q.J. Mac 1967 Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium, vol. 1. pp. 281297.
- J.M. Pen, J.A. Lozano, P. Larraaga, 1999 An empirical comparison of four initialization methods for the K-means algorithm, Pattern Recognition Letter 20, 10271040.
- Fukunaga K. 1990 Introduction to Statistical Pattern Recognition
- M. San Diego: Academic Press.
- S.Z. Selim, M.A. Ismail 1984 K-means-type algorithms: a generalized convergence

theorem and characterization of local optimality, IIEEE Transactions on Pattern Analysis and Machine Intelligence 6 8187.

- Jain A K, Dubes R C. 1988 Algorithms for Clustering Data
- M. Englewood Cliffs: Prentice Hall.
- F. Caoa et. al. 2009 "An initialization method for the K-means algorithm using neighborhood model", Computers and Mathematics with Applications, vol. 58, pp. 474 – 483.
- Z. Pawlak, 1991 "Rough Sets-Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, Dordrecht, Boston, London.
- S. Khan and A. Ahmad, 2004 "Cluster center initialization algorithm for K-means clustering," Pattern Recognition Letters, vol. 25, pp. 1293–1302,.
- X. Junling et. al, 2009 "Stable Initialization Scheme for K-means Clustering," Wuhan University Journal Of Natural Sciences, vol.14, no.1.
- J. Lu. et. al, 2008 "Hierarchical initialization approach for K-means clustering," Pattern Recognition Letters, vol. 29, pp. 787–795.
- Ken C.K. Lee, Baihua Zheng and Wang-Chien Lee 2008 "Ranked Reverse Nearest Neighbor Search" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 7, JULY.
- Blake C L, Merz C J. UCI Repository of Machine Learning Database
- EB/OL.
- 20011-03-15. http://www.ics.uci.edu/ MLRepository.html.
- Y.M. Yang, 1999 An evaluation of statistical approaches to text categorization, Journal of Information Retrieval 1 (1-2) 6788.
- A. Asuncion and D.J. Newman, University of California, Dept. of Information and Computer Science. The UCI Machine Learning Repository http://mlearn.ics.uci.edu/ MLRepository.html . Last visit May. 3, 2011

Index Terms

Computer Science Artificial Intelligence

Key words

Clustering reverse nearest neighbor search coupling degree

K-means

initialization