{tag}

{/tag}
International Journal of Computer Applications

Volume 34 - Number 6

Year of Publication: 2011

Authors:

M. H. Marghny

Rasha M. Abd El-Aziz

Ahmed I. Taloba

10.5120/4092-5420

{bibtex}pxc3875420.bib{/bibtex}

Abstract

Clustering analysis plays an important role in scientific research and commercial application. K-means algorithm is a widely used partition method in clustering. However, it is known that the K-means algorithm may get stuck at suboptimal solutions, depending on the choice of the initial cluster centers. In this article, we propose a technique to handle large scale data, which can

select initial clustering center purposefully using Genetic algorithms (GAs), reduce the sensitivity to isolated point, avoid dissevering big cluster, and overcome deflexion of data in some degree that caused by the disproportion in data partitioning owing to adoption of multi-sampling. We applied our method to some public datasets these show the advantages of the proposed approach for example Hepatitis C dataset that has been taken from the machine learning warehouse of University of California. Our aim is to evaluate hepatitis dataset. In order to evaluate this dataset we did some preprocessing operation, the reason to preprocessing is to summarize the data in the best and suitable way for our algorithm. Missing values of the instances are adjusted using local mean method.

Reference

- Yasin, H., JilaniT. A., and Danish, M. 2011. Hepatitis-C Classification using Data Mining Techniques. International Journal of Computer Applications.Vol 24– No.3.
- JilaniT. A., Yasin, H., and Yasin,M. M. 2011. PCA-ANN for Classification of Hepatitis-C Patients. International Journal of Computer Applications.Vol14– No.7.
- Wang,J. 2006. Encyclopedia of Data Warehousing and Mining. Idea Group Publishing.
- Filho,J.L.R.,Treleaven,P.C., andAlippiC. 1994. Genetic Algorithm ProgrammingEnvironments. IEEE Comput, vol.27, pp.28-43.
- Maulik, U., and Bandyopadhyay, S. 2000. Genetic Algorithm-BasedClusteringTechnique. Pattern Recognition, vol.33, pp.1455-1465.
- Anderberg,M.R. 1973. Cluster Analysis for Application. Academic Press,New York.
- Hartigan, J.A.1975. Clustering Algorithms. Wiley, New York.
- Devijver,P.A., and Kittler, J. 1982. Pattern Recognition: A Statistical Approach. Prentice-Hall, London.
- Jain, A.K., and Dubes, R.C. 1988. Algorithms for Clustering Data. PrenticeHall, Englewood Cliffs, NJ.
- Tou, J.T., and Gonzalez, R.C. 1974. Pattern Recognition Principles. Addison Wesley, Reading.
- Zhang,Y., Mao,J., and Xiong,Z. 2003: An Efficient Clustering Algorithm.International Conference on Machine Learning and Cybernetics, vol.1,pp.261-265.
- Han,J., and Kamber,M. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Forgy, E.1965. Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. Biometrics, pp.21-768.
- McQueen,J. 1967. Some methods for classification and analysis of multivariate observations. Computer and Chemistry, vol.4, pp.257-272.
- Tang, L., Yang, Z., and Wang, M. 1997. Employing Genetic Algorithm toImprove the K-means Algorithm in Clustering Analysis.MathematicsStatistics and Application Probability, vol.12.
- Cheng,E., Wang,S.,Ning, Y., and Wang, X. 2001. The Design and Realization of Using Representation Point Valid Clustering Algorithm. Patten Recognition and Artificial Intelligence, vol. 14.
- Duda, R. O., and Hart, P. E.1973. Patten Classification and Scene Analysis.New York: John Wiley and Sons.
- Selim,S. Z., and Alsultan, K. 1991. A Simulated Annealing Algorithm for theClustering

Problem. Patten Recognition, vol.24, pp.1003-1008.

- Krishna, K., and Narasimha, M. M. 1999. Genetic K-means Algorithm. Systems,IEEE Transactions on Man and Cybernetics, Part B, pp.433-439.
- Murthy, C. A., and Chowdhury, N. 1996. In Search of Optimal Clusters usingGenetic Algorithms.PatternRecog. Lett, pp.825-832.
- Tang, L., Yang, Z., and Wang, M. 1997. Improve K-means Algorithm of Cluster Method by GA. Mathematical Statistics and Applied Probability,pp.350-356.
- Fu, J.,Xu, G., and Wang, Y. 2004. Clustering Based on Genetic Algorithm.Computer engineering, pp.122-124.
- Lu, Q., and Yu, J. 2005. K-Means Optimal Clustering Algorithm Based onHybrid Genetic Technique. Journal of East China University of Scienceand Technology (Natural Science Edition), pp.219-222.
- Bandyopadhyay, S., and Maulik, U. 2002. An Evolutionary Technique Based onK-Means Algorithm for Optimal Clustering in RN. Information Sciences,pp.221-237.
- Yang, S., and Li, Y. 2006.K-means Optimization Study on k Value of K2meansAlgorithm. System Engineering Theory and Application, pp.97-101.
- Chittu, V.,Sumathi, N. 2011. A Modified Genetic Algorithm Initializing K-Means Clustering. Global Journal of Computer Science and Technology.
- Kumar, V., Chhabra, J.K., and Kumar, D. 2011. Initializing Cluster Center for K-Means Using Biogeography Based Optimization. Advances in Computing, Communication and Control. Vol.125. pp.448-456.
- Reddy, D., Mishra,D., and Jana, P. K. 2011. MST-Based Cluster Initialization forK-Means. Advances in Computing. Communication and Control.
- Min, W., andSiqing, Y. 2010. Improved K-means clustering based on genetic algorithm. IEEE, Computer Application and System Modeling.
- Li, X., Zhang, L., Li, Y., and Wang, Z. 2010.AImproved K-means Clustering Algorithm Combined with the Genetic Algorithm. IEEE, Digital Content,Multimedia Technology and its Applications.
- Fayyad, U., Reina,C., and Bradley, P.S. 1998. Initialization of Iterative Refinement Clustering Algorithms. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD98), pp.194-198.
- Pal, S. K., and Majumder, D. D. 1977. Fuzzy Sets and Decision Making Approaches in Vowel and Speaker Recognition. IEEE Trans. Systems, ManCybernet. SMC, vol.7, pp.625-629.
- Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems.Ann. Eugenics, vol.3, pp.179-188.
- Johnson,R. A., and Wichern, D. W. 1982. Applied Multivariate StatisticalAnalysis. Prentice-Hall, Englewood Cliffs, NJ, 1982.
- Virmajoki, O. 2004. Pairwise Nearest Neighbor Method Revisited. PhD thesis,University ofJoensuu, Joensuu, Finland.

Index Terms

Computer Science Data Mining

Key words

Genetic Algorithms Clustering

K-means algorithm

Squared-error criterion

Hepatitis-C Virus (HCV)