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 disservering 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

  Pattern Recognition, vol. 33, pp. 1455-1465.
  Prentice-Hall, London.
  Englewood Cliffs, NJ.
  Reading.
  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
- Tang, L., Yang, Z., and Wang, M. 1997. Employing Genetic Algorithm to Improve the
  K-means Algorithm in Clustering Analysis. Mathematics Statistics and Application Probability,
  vol. 12.
  Representation Point Valid Clustering Algorithm. Pattern 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.
An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study


Index Terms

Computer Science

Data Mining
Key words
Genetic Algorithms Clustering
K-means algorithm
Squared-error criterion
Hepatitis-C Virus (HCV)