Modified K-Means Algorithm for Effective Clustering of Categorical Data Sets

Abstract

Traditional k-means algorithm is well known for its clustering ability and efficiency on large amount of data sets. But this method is well suited for numeric values only and cannot be effectively used for categorical data sets. In this paper, we present modified k-means algorithms that can perform clustering very effectively on mixed data sets. The main intuition behind our proposed method is that all prototypes are the potential candidates at the root level. For the children of the root node, we can prune the candidate set by using simple geometrical constraints. The experimental results show that this method is well suited for categorical data sets and overall time of computation is very minimal.

References

- Chien, L. J., Chang, C. C. and Lee, Y. J., ”Variant methods of reduced set selection for reduced support vector machines”, Journal of Information Science and
Modified K-Means Algorithm for Effective Clustering of Categorical Data Sets


- Huang, Z., "Clustering large data sets with mixed numeric and categorical values", Proceedings of The First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, 1997.
- Hsu, C. W., C. C. Chang and C. J. Lin, "Practical guide to support vector classification", Department of Computer Science and Information Engineering National Taiwan University, 2003.

Index Terms

Computer Science

Artificial Intelligence
Keywords
Clustering Large Data Sets K-Means algorithm CLARANS DBSCAN Data
Mining Pattern Mining
Rule Mining