Abstract

Classification algorithms under the category of data mining have widespread applications in the modern world finding their use in almost every field and area that aims at predicting an outcome class for some data instance. As a result of which many supervised classification algorithms have been studied in the field of machine learning. Many classification algorithms can be used to serve the purpose, K-Nearest Neighbor, Gaussian Naive Bayes, Decision Tree to name a few. However even today it is a time consuming and complex task to decide the most suitable algorithm for the data under consideration. This article discusses an approach that predicts an algorithm that would produce best accuracy for the given data, depending upon internal data parameters: size of data, ratio of numerical attributes, count of outliers, average correlation, number of classes in target and average number of classes in attributes. This paper analyses the relation between the performance of K-Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes and Decision Tree classification algorithms and internal data parameters thereby evaluating a generic approach to determine the most accurate algorithm and also studies some limitations, like the inability of incorporating external factors namely memory requirement and
others.

References

9. Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

Index Terms

Computer Science

Algorithms

Keywords

Supervised Learning, Classification Algorithm, Decision Tree, Logistic Regression, K Nearest Neighbors, Naive Bayes