Abstract

High speed data streaming and data mining is the most contemporize challenges in machine learning. This demand methods displaying a high process effectiveness, with ability to continuously update their structure and handle ever-arriving big variety of instances. in this paper, we have a tendency to present a new incremental and distributed classifier based on the favored nearest neighbor algorithmic rule, adapted to such a exigent situation. This technique, enforced in Apache Spark, includes a distributed metric-space ordering to perform quicker searches. a vast live {of information of data of knowledge} containing useful data, referred to as big data, is created frequently. For handling such large volume of data, there's a necessity of big data structures, for example, Hadoop Map reduce, Apache Spark then on. Among these, Apache Spark performs up to one hundred circumstances speedier than ancient systems like Hadoop Map reduce. we have a tendency to concentrate on the plan of partition grouping calculation and its execution on Apache Spark.

Index Terms

Computer Science

Data Mining

Keywords

Nearest Neighbor, High-Speed Big Data, Data Streams