International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 10 ǁ October 2022 ǁ PP. 409-413

High Dimensional Data analysis base on Machine Learning methods

Dao Thi Phuong Anh
Faculty of Information Technology, Hanoi University of Natural Resources & Environment, Hanoi, Vietnam, 100000.

Abstract

In this article, research in the characteristics of Big Data. High dimensional data has always been a processing challenge for algorithms in data mining and knowledge discovery about data. With the explosion of the Internet and data generation systems such as social networks, newspapers, the new data text generated every day is huge. Moreover, these types of data are often unstructured, high dimensionality requires efficient algorithms to handle. Research and analyze high dimensional data by some machine learning methods such as Kmeans, Dbscan, SVM for the purpose of data mining effectively. When we master big data, we will have a greater chance of success in today's competitive landscape, people will benefit more from extracting information more accurately, more usefully at a cost with lower fees.

Keywords: *Big Data, Kmean, Dbscan, SVM, Machine learning algorithms*

Date of Submission: 02-10-2022 Date of acceptance: 15-10-2022

I. INTRODUCTION

Big data has increased the demand of information management specialists so much so that. Developed economies increasingly use data-intensive technologie. Big Data is an incredibly high data system, to the point where it cannot be stored in traditional database systems. The complexity and impossibility of forming a unified whole of Big Data data is also a factor that makes it difficult to synchronize in a database system.

Data is collected from many different sources including: unlimited data from internet, web 2.0, from research devices (astronomical data, medical services...), data from communication devices smart (also known as smart device). Hence it has a non-stationary structure. To effectively exploit and use that huge data source, machine learning algorithms make a significant contribution. On the basis of the design as well as the topic's objectives, the important new issues raised in this topic that need to be solved are: Overview of big data, research on some machine learning algorithms. to analyze high dimensional data such as Kmean, Dbscan, SVM. Then, I proceed to build a program to apply machine learning algorithms to analyze high dimensional data and based on the results of the application, I conduct empirical analysis and evaluation.

II. CONTENTSOFTHESTUDY

1. Big Data Overview

Big Data is characterized by 5V which is Volume- Data volume, Velocity - Data processing speed, Variety - Data diversity, Veracity- Accuracy, Value - Information value. High dimensional data has always been a processing challenge for algorithms in data mining and knowledge discovery about data. With the explosion of the Internet and data generation systems such as social networks, newspapers, the new data text generated every day is huge. Moreover, these types of data are often unstructured, high dimensionality requires efficient algorithms to handle [2].

The importance of big data doesn't hinge on how much data we have, but what we do with it. We can take data from any source and analyze it to find answers enabling cost reduction, time reduction, new product development and optimized services, and intelligent decision making. When we combine big data with powerful analytics, we can perform business-relevant tasks such as: Identify the root cause of problems, issues and defects in near real time. Generate coupons at the point of sale based on customer buying habits. Recalculate the entire risk portfolio in minutes. Detect fraud before it affects our organization. Problems such as target customer analysis, business process optimization, public health industry, banking and finance business, security, politics, and law. Optimizing machinery and equipment and building smarter cities [2].

With the current trend of 4.0 technology, Big Data is widely applied and useful in many fields. Companies in the world and Vietnam soon applied Big Data such as Amazon, IBM, Microsoft, HP, Dell, Facebook, FPT…[4].

2. Big Data management technology

2.1.Platform technology applied to Big Data

Big Data has a very high amount of data to store and often stores data streams of different types at high speed. Virtualization is a foundational technology applied to the implementation of cloud computing and big data. It provides the basis for many of the foundational properties needed to access, store, analyze, and manage distributed computing components in big data environments. In addition, the power of the cloud makes it possible for users to access necessary computing and storage resources with little or no IT support or purchase additional hardware or software. One of the key features of the cloud is elastic scalability: Users can add or subtract resources in near real-time based on changing requirements. Clouds play an important role in the world of big data. Major changes occur when infrastructure components are combined with advances in data management. Horizontal scaling and infrastructure optimization support the actual implementation of big data [3].

2.2 Big data management

Big data is becoming a key element in how organizations leverage high-volume data at high speed to solve data-specific problems. However, big data does not exist in isolation. To be effective, companies often need to combine the results of big data analysis with existing data in the business. In other words, we cannot think of big data in isolation from operational data sources. There is a wide range of important operational data services.

One of the most important services provided by operational databases (stores also called data) is persistence. Persistence ensures that the data stored in the database will not be changed without permission and that it will be available as long as it is important to the business. What good is a database, if it cannot be trusted to protect the data that we put into it? With this important requirement, we have to think about what kind of data we want to keep, how can we access and update it, and how can we use it to bring professional decision making. At this fundamental level, the choice of database engine is critical to our successful big data implementation. There are three types of databases: relational databases, non-relational databases, key-value pairs databases, document databases, column databases, graph databases, and spatial databases.

Unlike traditional operational database systems and applications, data warehouses have been used by business lines and financial analysts to help make decisions about the direction of a business. business strategy. The data was collected from many different relational database sources, then made sure that the metadata was consistent, and the data was error-free and then integrated well. So, what we need to do is: Integrate big data with traditional data warehouses then Conduct big data analysis and data warehouses.

2.3. Analyze high dimensional numerical data dimensional numerical data using some machine learning algorithms

High dimensional data has always been a processing challenge for algorithms in data mining and knowledge discovery about data. In addition, these types of data are often unstructured, high dimensionality requires effective processing methods. That is using machine learning algorithms to analyze high dimensional datasets, here we learn two basic machine learning algorithms: DBSCAN and Support Vector Machine. The results after analysis will greatly affect the decisions of companies and businesses in business strategy.

K-MEANS algorithm

K-means algorithm belongs to the category of very simple unsupervised algorithms and is widely applied to clustering problem samples introduced by MacQueen in the document “J. Some Methods for Classification and Analysis of Multivariate Observations” in 1967.

Clustering is the process of grouping a group of data points into a small number of clusters. In general terms of mathematical representation, we have n data points xi, i=1...n that need to be classified into k clusters. The goal of the problem is one data point for a cluster. The K-means algorithm provides us with a method to find the positions of points μi,i=1...k of clusters such that the distance function from the points to the clusters is minimal [1].

$$\arg\min_{\mathbf{c}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in c_i} d(\mathbf{x}, \mu_i) = \arg\min_{\mathbf{c}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in c_i} \|\mathbf{x} - \mu_i\|_2^2$$

where ’ci’ is a set of points inside the cluster ’i’. The K-means algorithm uses Euclidean distances.

Advantage:

- With a high number of variables, the K-means algorithm can compute faster than other hierarchical clustering algorithms (if K is small).

- K-means can cluster clusters more tightly than hierarchical clustering, especially if the clusters are spherical.

Shortcoming:

- The initialization of the central element of the initial clusters affects the division of objects into the cluster in case the data is not high.

- The number of clusters k must always be determined in advance.

- The area of clusters is not clearly defined, with the same object, it may be included in this cluster or another cluster when the data capacity changes.

Initialization conditions have a great influence on the results. Different initialization conditions can give different clustering results.

- The influence of the attribute on the process of creating clusters is not determined.

SVM algorithm

Support Vector Machine (SVM), researched and introduced by Vapnik in 1995, is a statistical theory-based supervised learning method used for classification and object recognition problems. The basic idea of the algorithm is as follows: for a training set represented in d-dimensional space {x1, x2, …, xn}, without loss of generality, consider the problem of 2 classes, each data element xi will belong to one of two classes denoted by +1 or -1. The SVM method will find the best hyperplane to be able to divide the points on this space into two separate layers, respectively layer +1 and layer -1. With training data in d-dimensional space, the hyperplane function will be a d-variable polynomial [1,7].

Imagine we have a data set consisting of blue and red points placed on the same plane.
We can find a line to separate the sets of blue and red points separately as shown below.

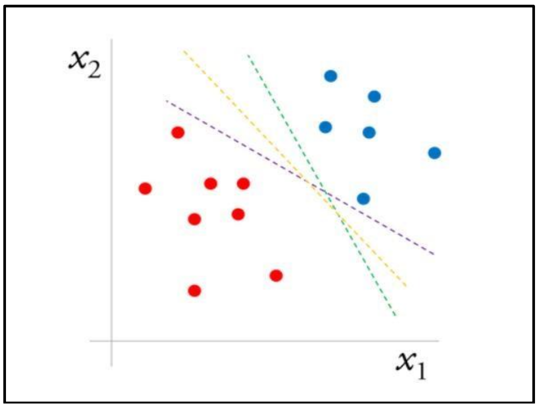

Figure 1: Example of data division in the plane.

The algorithm has the following pros and cons:

Advantage:

-With a high number of variables, the K-
means algorithm can calculate faster than other hierarchical grouping algorithms (if K is small).

- K-means can cluster clusters more tightly than hierarchical clustering, especially if the clusters are spherical.

Shortcoming:

The problem of height number: In case the number of attributes (p) of the data set is much highr than the number of data (n), then SVM gives quite bad results. Probability not yet obvious: SVM's classification is just an attempt to separate objects into two classes separated by the SVM hyperplane. This does not explain the probability of a member appearing in a group [5].

DBSCAN algorithm

DBSCAN algorithm was proposed by Ester, Kriegel and Sander in 1996 when researching spatial data clustering algorithms based on the definition of a cluster as the maximum set of connected points in terms of density. DBSCAN algorithm detects clusters of arbitrary shape, good noise detection ability. DBSCAN performs well on multi-dimensional space, suitable for databases with densely distributed density even with noisy elements.

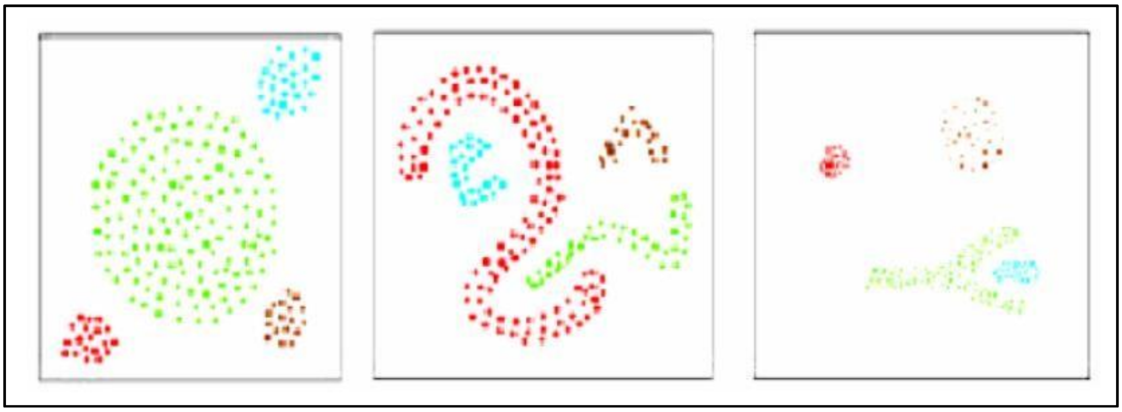

Figure 2: Shape of clusters explored by DBSCAN algorithm.

The main idea to detect clusters of DBSCAN algorithm is that inside each cluster there is always a higher density than outside the cluster (the neighborhood of each point in a cluster has a number of points greater than the minimum threshold). Furthermore, the density in noisy regions is lower than the inner density of any cluster. In each cluster must determine the radius of the neighborhood and the minimum number of points in the vicinity of a point in the cluster. The shape of the neighborhood depends on the distance function between the points (if using Manhattan distance in 2D space the neighborhood is rectangular, if using Eclidean distance in 2D space then the neighborhood circular). Points in each cluster are classified into two types: points inside the cluster (core point: core feature) and point lying on the border of the cluster (border point: border feature). The neighborhood shape of a point is determined based on choosing the distance function between two points p and q, denoted dist(p,q) [1,6].

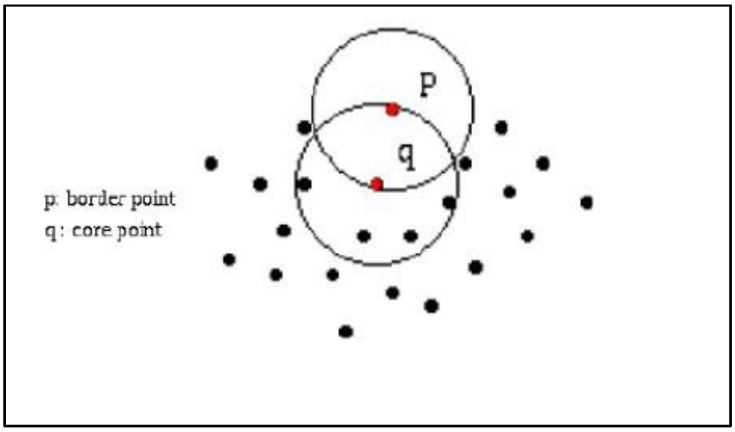

Figure 3: Boundary objects and core objects.

III. EVALUATION OF EMPIRICAL DATA AND RESULTS

The author conducted experimental data analysis on Weka software for two algorithms KMEANS and SVM. And test on data set: Student list data with high dimension (25). First, create two datasets: one for training and one for testing.

Table 1: Experimental results with K - Means

Ordinal number	Training data	KMEANS
1	Student Score Data	97.74 %

Table 2: SVM empirical results for student score data

Ordinal number	Training data	Test data	Accuracy
1	279	31	96.77 %

In this table, with high dimensional numerical data, it will take a long time to store and process.

IV. CONCLUSION

The birth of Big Data marked a breakthrough in technological development. Big Data is a system. The data system is so high that it cannot be stored in database systemstraditional material. Complexity and impossibility to form a unified whole of dataIs Big Data also a factor that makes it difficult to synchronize in a systemtraditional database system. Data is collected from a variety of sourcesincluding: unlimited data from internet, web 2.0, from research devices (natural data)documents, medical services, etc.), data from smart devices (also known as smart devices),that it has a non-stationary structure. To effectively exploit and use data sources.

That huge, machine learning algorithms make a big contribution.This report analyzed the importance of Big Data and its applications in practice, especially in production and business. Research some machine learning algorithms in data analysis. Analyze high dimensional data through several machine learning algorithms. However, some contents of the topic need to be improved and supplemented, such as researching and installing more machine learning algorithms for more diverse data sets, reducing data dimensionality, etc.

REFERENCES

[1]. Big Data economic analysis group - Big data collection, Knowledge Publishing House, 2017.

[2]. Bernard Marr, Big Data – Big Data, Industry and Trade Publishing House, 2017.

[3]. Vu Viet Vu, Selection of training data set for Support Vector Machine method, Proceedings of the National Conference on Electronics, Communication and Information Technology (REV), ISBN: 978-604-931- 253-3, December, pp: 3.28-3.32, 2016.

[4]. Rui Xu, Donald C. Wunsch II: Survey of clustering algorithms. IEEETrans. Neural Networks 16(3): 645-678, 2005.

[5]. Anil K. Jain: Data clustering: 50 years beyond K-means. Pattern Recognition Letters (PRL) 31(8):651-666, 2010.

[6]. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In the proceeding of SIGKDD Conference on Knowledge Discovery and Data Mining, pp: 226-231, 1996.

[7]. [V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.