PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

ABSTRACT

As one important technique of fuzzy clustering in data mining and pattern recognition, the possibilistic c-means algorithm (PCM) has been widely used in image analysis and knowledge discovery. However, it is difficult for PCM to produce a good result for clustering big data, especially for heterogeneous data, since it is initially designed for only small structured dataset. In PPHOPCM, the functions for updating the membership matrix and clustering centers are approximated as polynomial functions to support the secure computing of the BGV scheme. Experimental results indicate that PPHOPCM can effectively cluster a large number of heterogeneous data using cloud computing without disclosure of private data.

EXISTING SYSTEM

AS personal computing technology and social websites, such as Facebook and Twitter, become increasingly popular, big data is in the explosive growth. Big data are typically heterogeneous, i.e., each object in big data set is multi-modal. Specially, big data sets include various interrelated kinds of objects, such as texts, images and audios, resulting in high heterogeneity in terms of structure form, involving structured data and unstructured data. Moreover, different types of objects carry different information while they are interrelated with each other. Images pass on different information from the surrounding texts, they describe the same objects from different perspectives. Furthermore, big data are usually of huge amounts. For example, Facebook, the famous social websites, collects about 500 terabytes (TB) data every day .These features of big data bring a challenging issue to clustering technologies. Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group

DRAWBACKS

- It is difficult to cluster big data effectively, especially heterogeneous data.
- High time complexity.
- Only applicable to small data sets.

PROPOSED SYSTEM

This paper proposes a privacy preserving high-order PCM scheme (PPHOPCM) for big data clustering. PCM is one important scheme of fuzzy clustering . PCM can reflect the typicality of each object to different clusters effectively and it is able to avoid the corruption of noise in the clustering process .However, PCM cannot be applied to big data clustering directly since it is initially designed for the small structured dataset. Specially, it cannot capture the complex correlation over multiple modalities of the heterogeneous data object. The paper proposes a high-order PCM algorithm by extending the conventional PCM algorithm in the tensor space.

Tensor is called a multidimensional array in mathematics and it is widely used to represent heterogeneous data in big data analysis and mining. In this paper, the proposed HOPCM algorithm represents each object by using a tensor to reveal the correlation over multiple modalities of the heterogeneous data object. To increase the efficiency for clustering big data, we design a distributed HOPCM algorithm based on MapReduce to employ cloud servers to perform the HOPCM algorithm. However, the private data tends to be in disclosure when performing HOPCM on cloud. Take the medical data which is a typical type of big data for example. A large amount of private information such as personal email address and diagnostic data is included in the medical records. The disclosure of the private information will threaten people's lives and property greatly. Therefore, to protect the private data on cloud, we propose a privacy preserving HOPCM scheme by using the BGV technique that is of high efficiency .Unfortunately, BGV does not support the division operations and square root operations that are the necessary computation in the functions for updating the membership matrix and clustering centers in the HOPCM algorithm although it is a fully homomorphic encryption scheme

ADVANTAGES

- It avoids the corruption of noise in the clustering process.
- It protects the sensitive data.

SYSTEM REQUIREMENTS

H/W System Configuration:-

- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

S/W System Configuration:-

- Operating System : Linux
- Application Server : Tomcat5.0/6.X
- Backend coding : Java
- Tool : Virtual Box
- Environment : Ubuntu
- Technology : Hadoop