Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.85 – 89

Privacy Preserving Data Mining in Big Data by using K-means Clustering Algorithm

Anisha K[1], Sudheer Shetty[2]
[1]Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India
[2]Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India
[1] anishakrishn@gmail.com; [2] sudheer.cs@sahyadri.edu.in

Abstract— In current days there is a huge growth in information accumulation because of the advancement in the field of data innovation. It is extremely crucial that the information gets uncovered when the associations begin sharing the information for the mining procedure and protection might be ruptured. Privacy preserving methods gives another track to tackle this issue. The point of Privacy preserving data mining is the extraction of appropriate information from bulk quantity of advanced information while ensuring in the meantime delicate data. Four diverse sorts of clients required in data mining applications, to be specific, data provider, data collector, data miner, and decision maker. In this paper, we talk about the privacy preserving technique utilized for data collector while performing data mining procedure and we have broke down the utilization of normalization techniques in accomplishing privacy and depict an estimated calculation taking into account k- means. It is a novel technique for huge information examination which is quick, versatile and has high precision.

Keywords— Big data, Privacy preserving data mining, Z-Score normalization, K-means clustering algorithm, Centroid

I. INTRODUCTION

Today is the period of Google. The thing which is unfamiliar for us, we Google it, and in divisions of seconds we get the quantity of connections thus. This Big Data is the same thing than out general term information. Simply huge is a watchword utilized with the information to recognize the gathered datasets because of their huge size and complication. We can't organize them with our present systems or data mining programming instruments. The information gathering has become a great deal and is away from the capacity of generally utilized programming apparatuses to catch, organize, and prepare inside a fair time. The Big Data is

only information, accessible at heterogeneous, independent sources, in great vast sum, which get overhauled in portions of seconds.

Since we are in a time of data blast, it is essential to have the capacity to discover valuable data from huge amounts of information. As a result, different data mining strategies have been created. Data mining is regularly connected to fields, for example, marketing, sales, finance, and medical treatment. Furthermore, the fast progress in Internet and communications technology has prompted the development of information streams. In an information mining process, there four sorts of client are included, data provider, data collector, data miner, decision maker [1]. Each of them needs to protect the information while among the procedure of information mining. In this paper, we examine about how data collector can protect the information in data mining process. Data collector who is the client who gathers information from data providers and after that distribute the information to the data miner. The real worry of data collector is to ensure that the changed information contain no delicate data yet at the same time safeguard high utility. In the event that the data collector doesn't take adequate safety measures before discharging the information to open or data miner, those delicate data might be uncovered. Therefore, how to protect private information without exposure while getting a precise mining result during the time exact mining will turn out to be progressively troublesome, which thusly has prompted the advancement of Privacy-Preserving Data Mining methods. Besides, the potential vast number of information streams in addition to restricted memory space has obliged the customary techniques from acquiring the mining result with precision. In perspective of the above mentioned issues, thinks about on Privacy-Preserving Data Stream Mining as of late have ended up one of the vital issues in the field of data mining. A number of privacy preserving algorithms have been proposed and are utilized these days.

In this paper, we propose another strategy utilizing Z-Score normalization [2]. All in all, Z-Score normalization is utilized as a pre-preparing venture in data mining for change of information to a sought reach. Our motivation is to utilize it for preserving privacy through data mining in big data. We utilize K- means clustering algorithm [2], [3], [6] to accept the proposed approach and approve for precision.

II. LITERATURE SURVEY

In this paper [1], proposes a strategy k-anonymity. In k-anonymity modifies the estimations of semi identifiers in unique information table. Each tuple in the anonymized table is undefined from in any event k-1 different tuples along the semi identifiers. On the off chance that a table fulfils k- anonymity and the opponent just knows the semi identifier estimations of the objective individual, then the likelihood that the objective's record being distinguished by the opponent won't go beyond 1/k.

In this paper [2], analysed three normalization strategies specifically Min-Max, Z-Score and Decimal Scaling normalization. Min-max normalization performs a direct change on the actual data. The values are normalized inside the given extent. Z score normalization, likewise called as Zero mean normalization. Here the information is normalized in light of the mean and standard deviation. Decimal Scale Normalization in view of the development of decimal purpose of estimation of characteristic. The decimal point numbers are moved relies on upon the greatest total estimations of trait.

In this paper [3], min-max normalization is utilized as a pre-handling venture in data mining for change of information to a wanted extent. Our motivation is to utilize it for preserving privacy through data mining. We utilize K- means clustering to accept the proposed approach and approve for exactness.

In this paper [4], two approaches are used. One methodology is to change the information before conveying it to the data miner. The second approach accept the information is appropriated between two or more locales, and these destinations coordinate to take in the worldwide information mining results without uncovering the information at their individual locales.

In this paper [5], present a group of geometric data transformation methods (GDTMs) that contort confidential numerical values keeping in mind the end goal to meet privacy protection in clustering analysis.

In this paper [6], present a technique for k- means clustering when different locales contain different values for a common set of elements. Every site takes in the cluster of every substance, except adapts nothing about the properties at different destinations.

III. MODEL AND ARCHITECTURE

A. Proposed System

The proposed framework for the most part planned for preserving the privacy of data in the process of data mining. This framework can use in different applications such like in a hospital to keep up the patient records, in an organization to keep up employee details. These are a few case, there are such a variety of fields required this framework due to the reason of information blast is there in each field and privacy preserving data mining in big data is have more consideration.

B. System Architecture

The system architecture is given underneath in figure 1. The proposed framework modifies the delicate information by utilizing Z-Score normalization technique and the changed information will circulated over bunch by utilizing k-means clustering algorithm.

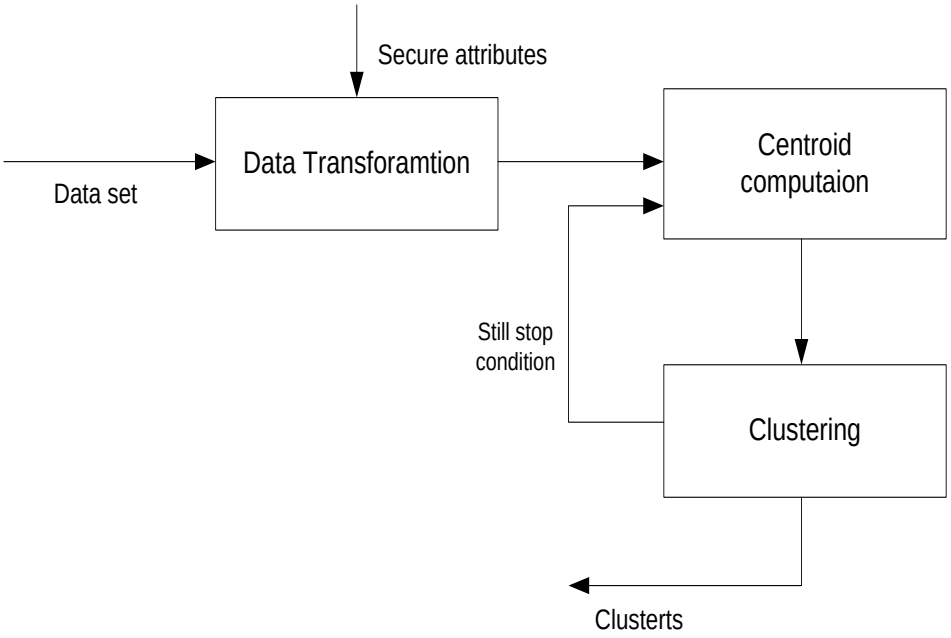

Fig .1 System Architecture

C. System Implementation

The framework can use in various field. From the huge measure of information we need to distinguish the delicate attribute .The determination of delicate attribute fluctuates from individual to individual. For instance, somebody considers pay to be delicate data while somebody doesn't; somebody thinks much about security while somebody cares less. After the choice of delicate attribute we need to perform normalization method to change the real information. In order to do perform this we utilize Z-score normalization technique.

1) Z-score normalization: *Z*-score normalization, additionally called as Zero mean standardization. Here the information is standardized in view of the mean and standard deviation. At that point the formula is,

$$d' = \frac{d - mean(p)}{std(p)}$$

Where mean (p) = total of the all property estimations of P

std (P) =Standard deviation of all estimations of P

2) K-means Clustering Algorithm: After change, we have to put the information in relating cluster. To do this, we utilize K-means clustering algorithm. In this calculation K demonstrate the quantity of cluster. The focal

thought of this Clustering is to minimize the entirety of squares of the separation amongst information and relating cluster centroid in that information set. The clustering procedure must be done until it gets balanced out. At that point, the items are assembled in light of the between relative separation among every entity and the centroid. The figure 2 explains the flow chart of K-means clustering algorithm.

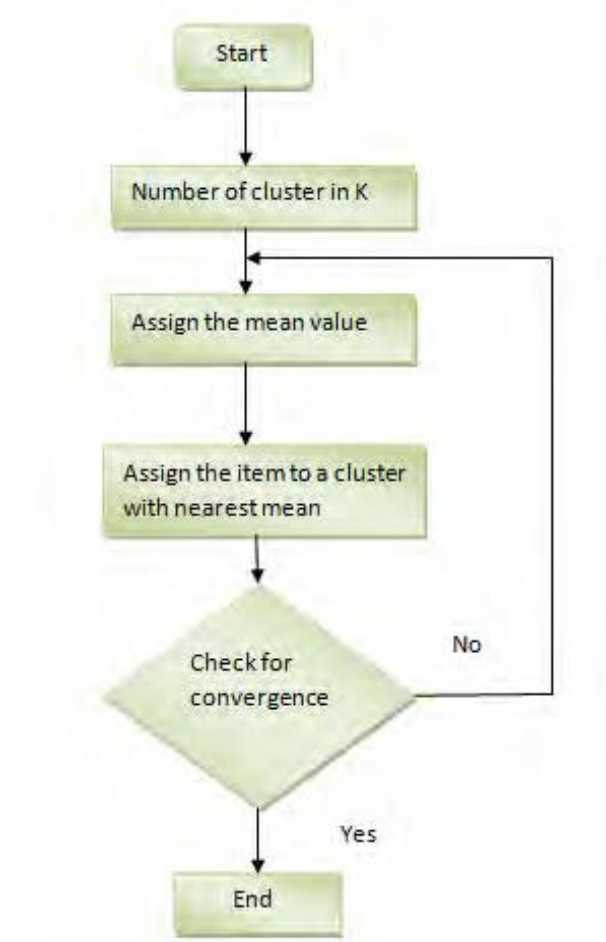

Fig. 2 Flow chart of K-means clustering algorithm

IV. CONCLUSIONS

There are such a large number of privacy preserving techniques are actualized for protect delicate entity in data mining process. In this paper we have discussed about the privacy preserving data mining in big data. The significance of this framework more essential now a days, in light of the reason of data explosion. The Z-Score normalization changes the actual information to privacy preserved structure. The K-means clustering algorithm performs on information and guarantees the correctness.

ACKNOWLEDGEMENT

I would like to thank God Almighty for blessing me to complete this work.

I am profoundly indebted to my guide, Mr. Sudheer Shetty, Associate Professor, Department of Computer Science and Engineering, Sahyadri College of Engineering and Management, for innumerable acts of timely advice and encouragement.

I would like to express my sincere thanks to my beloved family members and friends for their wishes and encouragement throughout the work.

REFERENCES

[1] Lei Xu, Chunxiao Jiang, Jiang Wang, Jian Yuan and Yong Ren, "Information Security in Big Data: Privacy and Data Mining", IEEE, 2014, vol. 2.
[2] C.Saranya and G.manikandan, "A Study on Normalization Techniques for Privacy Preserving Data Mining", International Journal of Engineering and Technology, vol. 5, No.3, 2013
[3] Syed Md. Tarique Ahmad, Shameemul Haque and Prince Shoeb Khan, "Privacy Preserving in Data Mining by Normalization," International journal of Computer Application, vol. 96, No.6, June 2014.
[4] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M. Y., "Tools for Privacy Preserving Distributed Data Mining", ACM SIGKDD Explorations Newsletter, Vol. 4, pp.28-34, 2002
[5] Stanley R. M. Oliveir, Osmar R. Zaiane "Privacy Preserving Clustering By Data Transformation", 18th Brazilian Symposium on Databases, Manaus, Brazil, pp. 304 -318 (2003).
[6] Jaideep Vaidya, and Chris Clifton, "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data," SIGKDD'03, August 24-27, 2003

