Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data

ABSTRACT

Intrusion Detection System (IDS) provides an important basis for the network defense. Due to the development of the cloud computing and social network, massive amounts of data are generated, which inevitably brings much pressure to IDS. And therefore, it becomes crucial to efficiently divide the data into different classes over big data according to data features. Moreover, we can further determine whether one is normal behavior or not based on the classes information. Although the clustering approach based on Kmeans for IDS has been well studied, unfortunately directly using it in big data environment may suffer from inappropriateness. On the one hand, the efficiency of data clustering needs to be improved. On the other hand, differ from the classification, there is no unified evaluation indicator for clustering issue, and thus, it is necessary to study which indicator is more suitable for evaluating the clustering results of IDS. In this study, we propose a clustering method for IDS based on Mini Batch Kmeans combined with Principal Component Analysis. Firstly, a preprocessing method is proposed to digitize the strings and then the dataset is normalized so as to improve the clustering efficiency. Secondly, the Principal Component Analysis method is used to reduce the dimension of the processed dataset aiming to further improve the clustering efficiency, and then Mini Batch Kmeans method is used for data clustering. More specifically, we use Kmeans++ to initialize the centers of cluster in order to avoid the algorithm getting into the local optimum, in addition, we choose the Calsski Harabasz indicator so that the clustering result is more easily determined. Compared with the other methods, the experimental results and the time complexity analysis show that our proposed method is effective and efficient. Above all, our proposed clustering method can be used for IDS over big data environment.
CONCLUSION

In this study, we proposed a clustering method based on Mini Batch Kmeans with PCA (PMBKM) for Intrusion Detection System. Taking IDS classic dataset KDDCUP99 for example, both 10% dataset and full dataset are tested. Firstly, we preprocess the given dataset and then the PCA method is used to reduce the dimension so as to improve the clustering efficiency. Additionally, the Mini Batch Kmeans method is used for the clustering of the processed dataset. Compared with Kmeans (KM), Kmeans with PCA (PKM), as well as Mini Batch Kmeans (MBKM), the experimental results show that our proposed PMBKM is effective and efficient. Above all, PMBKM can be used for intrusion detection system over big data environment. In our future work, we will engage in the research of clustering method over fog computing.

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb

SOFTWARE REQUIREMENTS:

- Operating system : Windows 7/UBUNTU.
- Coding Language : Java 1.7, Hadoop 0.8.1
- IDE : Eclipse
- Database : MYSQL
REFERENCES

