Privacy preserving big data mining: association rule hiding using fuzzy logic approach

Abstract

- Recently, privacy preserving data mining has been studied widely. Association rule mining can cause potential threat toward privacy of data.
- So, association rule hiding techniques are employed to avoid the risk of sensitive knowledge leakage.
- Many researches have been done on association rule hiding, but most of them focus on proposing algorithms with least side effect for static databases (with no new data entrance), while now the authors confront with streaming data which are continuous data.
- Furthermore, in the age of big data, it is necessary to optimise existing methods to be executable for large volume of data.
- In this study, data anonymisation is used to fit the proposed model for big data mining. Besides, special features of big data such as velocity make it necessary to consider each rule as a sensitive association rule with an appropriate membership degree.

Existing

- association rule hiding techniques are employed to avoid the risk of sensitive knowledge leakage.
- Many researches have been done on association rule hiding, but most of them focus on proposing algorithms with least side effect for static databases (with no new data entrance), while now the authors confront with streaming data which are continuous data.
- Furthermore, in the age of big data, it is necessary to optimise existing methods to be executable for large volume of data

Disadvantage

- it is necessary to optimise existing methods to be executable for large volume of data
- it is infeasible to store this huge amount of data, so the knowledge extraction should be done real time.
- Boolean logic (versus fuzzy logic) approach to determine
- whether the association rule is sensitive or not.
- • Undesired side effect of hiding sensitive association rules on
- non-sensitive rules.
- • The impossibility of using in big data analysis.

Proposed

- data anonymisation is used to fit the proposed model for big data mining.
- Besides, special features of big data such as velocity make it necessary to consider each rule as a sensitive association rule with an appropriate membership degree.
- Furthermore, parallelisation techniques which are embedded in the proposed model, can help to speed up data mining process.

Advantage

- parallelisation techniques which are embedded in the proposed model, can help to speed up data mining process
- None of existing approaches can be executed in a parallel and scalable manner, to be appropriate for big data mining

HARDWARE REQUIREMENTS

- Processor – Pentium –III
- Speed – 1.1 Ghz
- RAM – 256 MB(min)
- Hard Disk – 20 GB
- Floppy Drive – 1.44 MB
- Key Board – Standard Windows Keyboard
- Mouse – Two or Three Button Mouse
- Monitor – SVGA

SOFTWARE REQUIREMENTS

- Operating System : Windows 8
- Front End : Java /DOTNET
- Database : Mysql/HEIDISQL

Conclusion

- Association rule mining, besides its benefits in discovering unclear relationships between data, will result privacy violation. Association rule hiding can help to protect sensitive association rules to be discovered.
- Many different techniques have been considered to hide certain association rules, but most of them try to select ISs in order to decrease the confidence value to be less than the defined threshold.
- None of existing approaches can be executed in a parallel and scalable manner, to be appropriate for big data mining.
- Besides, removing ISs from the database can cause serious information loss as new data stream arrive. In this research, new big data association rule hiding technique is presented, which uses fuzzy logic approach, tries to decrease undesired side effect of sensitive rule hiding on non-sensitive rules in data streams.
- Features such as parallelism and scalability are embedded in the proposed model to provide the facility of implementing this model for huge volume of data.
- Results show that the proposed model can be more effective in big data mining than existing rule hiding approaches. As future work, we will try to decrease undesired side effect of the proposed model to gain ... formation loss.

Reference

[1] Philip, C.L.C., Zh, C.-Y.: 'Data-intensive applications, challenges, techniques and technologies: a survey on big data', Inf. Sci., 2014, 275, pp. 314–347

[2] Ohbyung, K., Namyeon, L., Bongsik, S.: 'Data quality management, data usage experience and acquisition intention of big data analytics', Int. J. Inf. Manage., 2014, 34, (3), pp. 387–394

[3] Alfredo, C., Carson, K.S.L., Richard, K.M.: 'Mining constrained frequent item-sets from distributed uncertain data', Future Gener. Comput. Syst., 2014, 37, pp. 117–126

[4] Xuyun, Z., Chang, L., Surya, N.S., et al.: 'A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud', J. Comput. Syst. Sci., 2014, 80, (5), pp. 1008–1020

[5] Yaping, L., Minghua, C., Qiwei, L., et al.: 'Enabling multilevel trust in privacy preserving data mining', IEEE Trans. Knowl. Data Eng., 2012, 24, (9), pp. 1589–1612