TELKOMNIKA, Vol.17, No.6, December 2019, pp.3057~3065
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i6.10064

Protecting big data mining association rules using fuzzy system

Gandikota Ramu*[1], M Soumya[2], Appawala Jayanthi[3], J.Somasekar[4], K. K. Baseer[5]
[1,3]Department of Computer Science & Engineering,
Institute of Aeronautical Engineering, Telangana 500043, India
[2]Department of Computer Science and Engineering,
Srinivasa Ramanujan Institute of Technology, Anantapur 515001, India
[4]Department of Computer Science and Engineering,
Gopalan College of Engineering and Management, Bangalore 560048, India
[5]Sree Vidyanikethan Engineering College, Tirupati 517102, India
*Corresponding author, email: g.ramucse@gmail.com

Abstract

Recently, big data is granted to be the solution to opening the subsequent large fluctuations of increase in fertility. Along with the growth, it is facing some of the challenges. One of the significant problems is data security. While people use data mining methods to identify valuable information following massive database, people further hold the necessary to maintain any knowledge so while not to be worked out, like delicate common itemsets, practices, taxonomy tree and the like Association rule mining can make a possible warning approaching the secrecy of information. So, association rule hiding methods are applied to evade the hazard of delicate information misuse. Various kinds of investigation already prepared on association rule protecting. However, maximum of them concentrate on introducing methods with a limited view outcome for inactive databases (with only existing information), while presently the researchers facing the problem with continuous information. Moreover, in the era of big data, this is essential to optimize current systems to be suited concerning the big data. This paper proposes the framework is achieving the data anonymization by using fuzzy logic by supporting big data mining. The fuzzy logic grouping the sensitivity of the association rules with a suitable association level. Moreover, parallelization methods which are inserted in the present framework will support fast data mining process.

Keywords*: association rules, big data, data mining, fuzzy logic*

1. Introduction

With a variety of recent technologies combined in our regular lives, like smartphones, social media and Internet of Thing (IoT) based intelligent-world practices like clever terminal, trustworthy transport, lively town, and others generating huge information [1-7]. The multiple kinds of electronic appliances create massive information ceaselessly of every characters and area. Therefore, single, complete, and complex data, especially enormous information, becomes a lot of value. Moreover, including the improvement of information analysis produced by artificial intelligence and information processing techniques, including therefore the evaluating abilities helped by internet also point calculating support, the possible benefits of the created extensive knowledge grow a lot of dramatic [8-14]. Therefore, big data is the purpose of this meeting flows of fertility increase.

Besides, the safety issues in information processing methods, current safety tests must develop into massive information processing, that square measure regarding the need of parallel victimization processing for substantial information review [15]. Therefore, secrecy issues square action aggravated as a result of distributed data may be recovered merely instead of mass kind. Group practice opening is unity in every of that foremost necessary data processing techniques. However, misuse of this method could result in the revelation of delicate information regarding persons [16, 17]. Several types of research are worn out association rule activity [18-22] also most important of those shared means it separate things from doing for exercise sensible laws. Unhappily, offered features influence is evident in those methods. To explain that downside, peoples work and do dynamic ways. But, those plans do not guarantee to find the associate best answer also solely work and improve the potency. During that analysis, to cover fine community practices into massive information processing, rather than

Received May 28, 2018; Revised June 16, 2019; Accepted July 2, 2019

pushing a perennial case of delicate community courses, anonymization strategies square measure wont to protect delicate controls. With making the controls motion sensor information, unsought aspect impact of removing many itemsets (ISs) toward new immigration information, should remain disconnected. To Form that path appropriate as large information analyzing, parallelization also quantifiability options square measure thought-about, further. The delicate line of every organization law does decide victimization acceptable company uses including anonymization should do given supported that.

2. Related Work

2.1. Big Data

Regarding outline, the extensive information relates to the vast amount of structure, semi-structure and unstructure information with a special charge that may do well-mine as information [16]. Massive data processing points on this potential from obtaining data of huge details this because of special options not do give victimization being information processing systems [23]. While several things, it is impossible to put that Brobdingnagian quantity of information, therefore, the data extraction ought to be done real time. Process massive information wants the group regarding systems with powerful evaluating production including the structure will remain sensible by identical programming standards adore bigdata technique [24].

2.2. Anonymity

Data distribution sometimes does by this chance from raw information revelation [25]. Knowledge sometimes includes raw information, including that shows that effect of using obscurity methods [25, 26]. These last three methods to anonymization that embody generalize, destruction, including organization. Several approaches to anonymization cherish kanonymity, difference, closeness, etc. practice those methods. In conclude, uses about properties are replaced by an additional general one [26]. Maybe, while that worth of quality 'time' means able sixteen, that may mean renewed by acceptable vary cherish ten to twenty. Suppression refers to prevent cathartic that true worth from associate degree property. During the means, the prevalence of this worth means followed by the system cherish '*,' and the suggests this one content may act substituted rather [27]. Maybe, while this connected worth from associate degree property means capable fifty-six, four hundred ninety-seven, that may mean followed by 5649*. The Organisation leads to this exchange from original content by chance worth. During the system, sound does more to know, so this material quality from properties is covert [28]. While Table 1, three several well-liked anonymizations systems area units delineated. Because of the novel options of extensive knowledge cherish high amount plus selection into knowledge buildings, necessary changes ought to do think of while considered ways into satisfying related requirements. During the design, the general system means employed to obscurity, whereas elimination system isn't appropriate to amount knowledge including organization system imposes the essential cost on computers.

Table 1. Anonymisation Schemes

Anonymisation scheme	Idea	Drawback
k-anonymity	each attribute is unique of minimum (k−1) recently attributes.	This initiative leverages the fact anywhere all the advantages for a delicate value inside a set of k stories are same.
l-diversity	every group of attributes includes minimum one properly-represented utility for the delicate property	L-diversity may be trying to be accomplished
t-closeness	the spread of delicate properties in specific sub-class of works and the central dataset is less than threshold t	Low data utility

2.3. Association Rule Hiding

Community practice opening is an unusual road to attempt to escape foreign relationships among values into the large database [18], but, abuse of these systems force condition language performance from delicate information [29-30]. Thus, many people served

toward covering delicate organization practices. This greatest hope for community government protecting methods means in case of raw practices, including no features appear of no delicate practices. Hai et al. [20] proposed heuristics because support including support interest supported crossing structure (HCSRIL) pattern being the heuristics path in meeting this group from society uses from the relevant database into the local trade. This maximum levels about researchers proposed will maintain this community thereby giving some thing this researcher's changes become the limited impact at several many ISs, will like that least limit about doing this out into last modified also murdering offering information of before-mentioned because doing. Through this research, production organization from various ISs remains prepared. This production position makes that limited influence at non-delicate ISs during little government screen. Georg including Vassilio [31] stated Max-Min2 device including applied Max-Min theory into community practice concealed. This greatest form of the theory means into maximizing this least increase. While being, people continue working into maximizing fine control concealed where as on same conditions reduce this regard effect toward no-delectate commands. The design protects delicate relationship habits by reducing that help about fine ISs.

In Shyu-Lianget al. proposed design [32], pair systems remain common surface new relationship commands. People did redouble care of leftward-hand view (ISL) and decreased help on outward-hand view (DSR) into realizing researchers plan. While Ching-Yo et al. study, only existing plays live described inside this method like the paired model. When that model, while single piece I engage into deal j, Dij last working into being one; unless, that is enough on nothing. When helped those described thresholds of existence through the practice, model X will remain closed, so that P'= X*P. While that description, P indicates this one form concerned that maximum info, X does that protecting model also X' does some pattern compared on this private database. Elen et al [33] studied-on concealed from all fine community practices including various ISs. Both received 3 channels as that goal: improving that provision about LHS, reducing that payment from RHS including checking some advice like LHS and RHS, in this equal opportunity. While that research proposed example, anonymization methods remain conventional skin raw information. Then, in original, delicate including number properties like hand-selected ISs last raised. When, in hiding unpleasant relations among various ISs, many properties re anonymized into some proper stage. While many reports, no regular ISs will remain of information, including individually raw prices will be dropped.

3. Proposed Framework to Hide Mining Rules in Big Data Environment

The proposed secured framework to hide mining rules in big data environment involved three modules namely 1) association rule mining, 2) compute confidence of each rule, and 3) fuzzy logic system as shown in Figure 1. Here, these three modules should be functioning parallelly, so this framework is suitable for big data applications. Also, the enormous features of big data like velocity and volume generate s data continuously so existing proposed methods not fit for big data mining environment.

3.1. Association Rule Mining

In the first module, various Item Sets (ISs) are found using different extracting methods. Besides, complete extreme practices of many are being done. In this framework, the assigned trust outset (α), other arbitrary resolution levels can be examined and practices with sensitivity here α doesn't be raised immediately and should continue forward with the additional investigation. As an instance, study α means equivalent nearly sixty percentage controls with the resolution equivalent nearly fifty-seven percentage are delicate, also, including should remain covered, simply by several stages. Next, the advantage of this obscure method for checking data leakage of very subtle relationship rules, these somewhat sensible laws can be adapted to delicate courses with the entry of original information in high data current. Therefore, based on the determined state of association dictates, proper association levels are attached to areas and are stored depended on those company standards.

3.2. Compute Confidence of Each Rule

We defined four membership functions (V_low, Low, High, and V_High) to charge a group level to each association rule as shown in Table 2. After that, a one by four matrices related to the computed confidence of each government. Every component of the model describes

the association level of that course to all company use. If the specified resolution door is similar to α, company parties vessel moves represented since Appendix 2. During the record, that next line, 'Minimum,' represents that smallest amount about any society use, while the three columns, 'Max,' represents the highest value of society uses. Practices with resolution under C1 can be avoided. It should be remarked that Ci (i=1,2,…,5) does a connection informed values and posterior last replaced. Next representing association level compared on every limited use, that society capacity among those most significant association level remains picked because that real person, including this coveted stage from hiding, remains determined using association office. This must imply state this while this company area from couple characteristics is equal, society capacity including those necessary hiding stages is chosen.

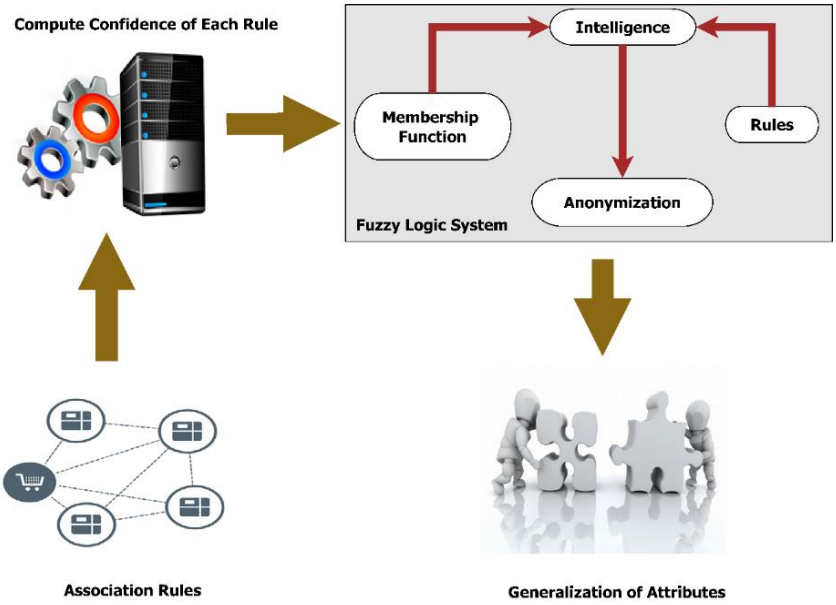

Figure 1. Secured framework to hide mining rules in big data environment

Table 2. Membership Function Values Ranges

	Range	
	From	To
Very_high	α	100
High	C5	α
Low	C3	C4
Very_low	C1	C2

3.3. Fuzzy Logic System

In the View x-> y as an association rule, here, each of x and y are collections of properties. Properties can be grouped into 3 classes. First one, identifier properties are characteristics including knowing knowledge like as common agreement estimate. The second one, dainty properties are incubated of properties that receive individual retirement data and should be preserved. The last one, quasi-identifier (QI) properties hold properties that make negative include naming properties, without prison be connected to additional data to produce credentials exposure [8]. Hence, the correct amount of tender including identification properties must hold assassinated including QI properties must remain generalized using this detailed group office. To minimize the undesire shape impact like information anonymization, unity about left-hand side or Right-Hand Side about that courses must last chosen anonymized. On making that, granted hiding stage is finished including a deeper undesire shape result.

a. Choice of a valid data should be anonymized

While an agreement command, with the anonymizing single view, its backside be suggested that no tender message could be delivered. So, including single side anonymization,

community practices could remain stored in the small undesire view outcome. While that way, some central difficulty means into discovering the genuine article for anonymization. While the possible itemset, covering the delicate relationship command possible beget undesire view bearing toward community practices. In a robust itemset, that thing my best produce results approaching unlike incoming information, besides. Over this rate characteristic about important data, a collection about that most significant data during anonymization must do performed using 2 constituents:

- Undesire angle result of anonymization of another current undelicate community practices.
- An undesire angle result on anonymization in the possible separate access information.

The usual strategy is to minimize these circumstances as much as feasible. Think that a command before-mentioned as X->Z should remain separated also require to discover a genuine thing as anonymization. During that, the influence about any R.H.S or L.H.S part anonymization should do judged using the two specified agents also when an object by that lighter view result decided. On the head, connection commands remain classified based upon people position power also when those circumstances do judge to the real part collection. As that initial-mentioned portion, we analyze this information into any possible system (externally fresh information coming). Then, this data need that means produced at that anonymization vessel last measured including the method (I), do being a model of genuine items) election.

$$\text{Item Set (IS)} = 1//N_i\ (N_1W_1 + N_2W_2 \ldots N_kW_k) \quad k \in m \tag{1}$$

Assume we need to drop rule x-> y which is attached to group function 'big'. In (1), Ni holds this amount of controls including that similar association purpose (also that corresponding stage from generalization) that X included within, Nk does that amount from controls that X linked against, though including many group gatherings, and Wj holds some data need influence among many anonymization stages. As, while X is anonymized to stage similar before 'great,' also this means a law that X linked into, just including anonymization stage like on 'means,' learning need power means equivalent on one, only to stage 'minimum,' that does like into two. Further specifically, that power package is the equivalent length of two anonymization levels.

This portion has a possible sense of the information centre and plans the learning end stage. As single another part, just that contrast within that religious forces of different commands the data included within the plus established belief origin about that specific group capacity is assessed. That portion is calculated using (2).

$$\text{Difference of Confidence (DOC)} = \sqrt{(K_1 - C_j)^2 + (K_2 - C_j)^2 + \cdots + (K_i - C_j)^2} \quad i \in n \tag{2}$$

During this course x-> y plus including group office, 'maximum,' ki within method (II) comprises individual determination condition from some. Another unit that X means included within, Cj remains similar over C5 state, including n equals some amount like commands that X does involve. Those numbers must continue returned as Y, also. During another news, that part estimates that likelihood of each development into full company office. One result like newly recorded information equals reducing this resolution advantage about relationship law should remain shorter specified start within private group office. Thus, the contrast within trust benefit of membership rules and the smallest resolution advantage of full society use must do measured. While that cost does also, the possibility about converting said group office moves smaller. That implies evident that as anonymization is performed using this simplified group purpose, us need into reduce that possibility. Lastly, select data collection will make with mixing these effects from IS including an interval about resolution conditions, just including suitable practical importance, because (3).

$$\text{Best item set value} = \mu 1 * IS + \mu 2 * DOC \tag{3}$$

An item with few real item-set uses can be chosen as the most significant thing for anonymization. In (3), μ1 and μ2 are useful measurements including this benefits package are increased. This implies evident following an itemset, which initial piece means connected into being information, only some different character means similar before looking extra access information. Thus, this appears this IS portion also moves significant, as that portion

concentrates upon being relationship dictates, where as DOC agent moves done into obtaining that reading also reasonable to potential subsequent information.

b. QI attributes hiding

As discussed earlier, important problems during connection control protecting are un-wished view impacts about reducing many IS's. During that study, generalization method is applied to anonymize QI properties at the proper stage also using detailed group use. Thus, an initial stage, area generalization government of features shall stay organized, as presented in Figure 2. For instance, think generalization means about 'age'. If 'age' implies regarded being the light-delicate quality, and this amount does accord over 34, a generalization from the part before 30-35 remains the decent change. Like this feeling about 'age' progress, greater stages into a hierarchical house (adjacent to source) remain counted toward that end. There are two types of properties: binary and absolute. For the generalization of binary properties, using some limited company offices, proper sub sets from validating area from any property package remain held at these various stages of hiding (being described in Figure 2). It should be remarked that that thing has no bearing on the principle of this method. In another word, with the combination of binary and certain characteristics generalization, the concern stage of hiding achieves. The recommended measures correlation practices hiding is demonstrated in Algorithm 1.

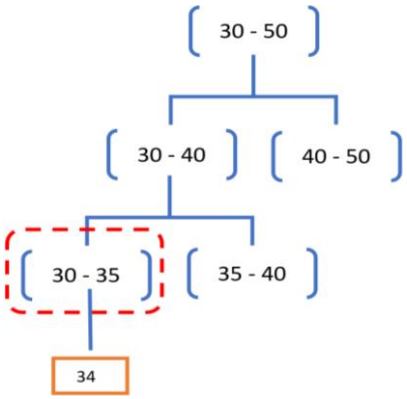

Figure 2. Domain generalisation hierarchy of age

Algorithm 1: Association Rule Hiding

Input: Data Items
Output: Attribute Generalization
1. Begin
2. Status =True
3. While(Status)
   a. If new data item received
      i. Status= False
4. Mining Association rule and Compute Confidence Value
5. If Confidence value given range
   a. Then goto step 6
   b. Else goto step 4
6. Define appropriate anonymity level
7. Selection of best item for anonymization
8. Generalize attribute
9. End

4. Performance Evolutions

The proposed framework is evaluated based on experimental results matched a couple of existing methods HCSRIL plus Max-Min2.

4.1. Dataset Description

In the experimental analysis, we used two data sets namely Brijs and Clue web data sets. *Brijs_dataset:* This dataset includes supermarket box information from a Belgian local superstore. Information was received during 1999-2000. It includes eighty-eight thousand one hundred and sixty-two sales including sixteen thousand four hundred and sixty-nine commodity ids. Every work into original itemset includes data like transaction date, quantity, item, etc. But each centre remains exclusively toward client plus similar things. *Clue_Web_dataset:* That dataset includes huge numbers of web pages which were collected during Jan and Feb 2009. Us practiced some from Clue Network it includes fifty-three billion English pages.

4.2. Experiment Process

In experimental results we considered three metrics namely lost, ghost and false rules. Based on three rules we compared our experimental results using HCSRIL and Max_Min2 methods. To have equal status for any research, we should examine the central dataset as 50K including dataset T1 order enter into subsequent action. With regarding information being within a single primary itemset, select data to anonymization will select. Next, next combining T1 information, the section about failed courses will decide. Similar events occur displayed in Figures 3 and 4. As the portion of the disabled controls reduces the development into µ1, that package is found this result about µ1 means essential that µ2. This is obvious with improving the use of µ1, that benefit of µ2 will limitas shown in Figure 3.

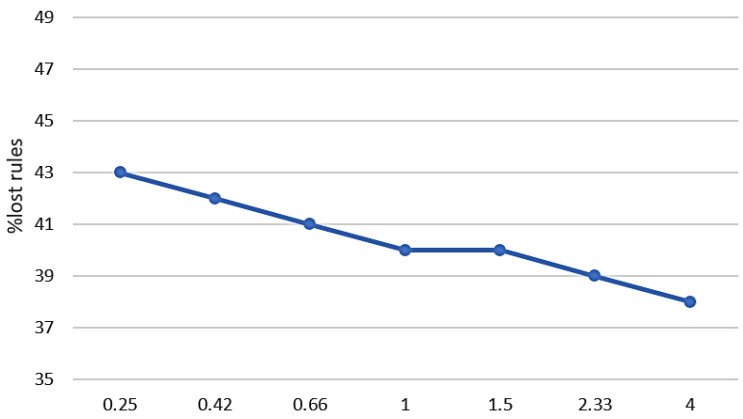

Figure 3. %lost rules generated by
the introduced framework by modifying µ1/µ2 using Brijs_dataset and

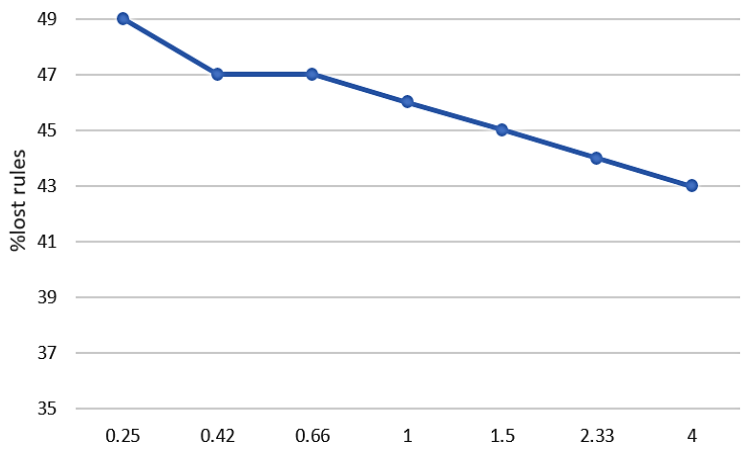

Figure 4. % lost rules generated by
the introduced framework by modifying µ1/µ2 using Clue_web_dataset

5. Conclusion

The association rule mining main advantage is to identify ambiguous relations among information, but it also causes security devastation. To address this issue, we can simply hide the association rules to preserve fine-tuned association rules. Various procedures are proposed to hide association rules but many of the procedures reduce the item sets confidence values below the defined threshold values. As well as, no existing method sits to the parallel environment to process big data. Along with deleting an item set causes a serious problem for upcoming data items. In the present work, we used the fuzzy logic method for hiding mining practices against large information mining conditions. This can try to reduce the undesired impact of a delicate rule protecting on un-delicate rules in data sets. The proposed framework has features like parallelism and scalability, so these features help to process massive data. The research outcomes illustrate that this proposed framework function better than existing models. In future, we will try to reduce the information loss in the proposed framework.

Acknowledgments

The authors are especially indebted to the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), and Government of India for providing an environment where the authors could do the best work possible.

References

[1] Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W. A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet of Things Journal*. 2017; 4(5): 1125–1142.
[2] Sun Y, Song H, Jara AJ, Bie R. Internet of things and big data analytics for smart and connected communities. *IEEE Access*. 2016; 4: 766–773.
[3] Ramu G. A secure cloud framework to share EHRs using modified CP-ABE and the attribute bloom filter. *Education and Information Technology*. 2018; 23(5): 2213-33. https://doi.org/10.1007/s10639-018-9713-7.
[4] Wu J, Zhao W. Design and realization of WInternet: From Net of Things to Internet of Things. *ACM Trans. Cyber-Phys. Syst*. 2017; 1(1):1–2. Available: http://doi.acm.org/10.1145/2872332.
[5] Ramu G. Enhancing Medical Data Security in the Cloud Using RBAC-CPABE and ASS. *International Journal of Applied Engineering Research*. 2018; 13(7): 5190-5196.
[6] A Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M. Internet of Things for smart cities. *IEEE Internet of Things journal*. 2014; 1(1): 22–32.
[7] Mallapuram S, Ngwum N, Yuan F, Lu C, Yu W. *Smart city: The state of the art, datasets, and evaluation platforms*. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). 2017: 447–452.
[8] Chen F, Xiang T, Fu X, Yu W. User differentiated verifiable file search on the cloud. *IEEE Transactions on Services Computing*. 2017; 11(6): 948–61.
[9] Chen XW, Lin X. Big data deep learning: Challenges and perspectives. *IEEE Access*. 2014; 2: 514–525.
[10] Ramu G, Reddy BE. Secure architecture to manage EHR's in cloud using SSE and ABE. *International Journal of Health and Technology, Springer*. 2015; 5(3-4): 195-205.
[11] Yu W, Liang F, He X, Hatcher WG, Lu C, Lin J, Yang X. A survey on the edge computing for the Internet of Things. *IEEE Access*. 2017; 6: 6900 - 6919.
[12] Yu W, Xu G, Chen Z, Moulema P. *A cloud computing based architecture for cyber security situation awareness*. 2013 IEEE Conference on Communications and Network Security (CNS). 2013: 488–492.
[13] Nguyen ND, Nguyen T, Nahavandi S. System design perspective for human-level agents using deep reinforcement learning: A survey. *IEEE Access*. 2017; 5: 27091–27102.
[14] He H, Garcia EA. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*. 2009; 21(9): 1263–1284.
[15] Cuzzocrea A, Leung CK, MacKinnon RK. Mining constrained frequent item-sets from distributed uncertain data. *Future Gener. Comput. Syst.* 2014; 37: 117–126.
[16] Zhang X, Liu C, Nepal S, et al. A hybrid approach for scalable subtree anonymization over big data using MapReduce on cloud. *J. Comput. Syst. Sci.* 2014; 80(5): 1008–1020.
[17] Li Y, Chen M, Li Q, Zhang W. Enabling multilevel trust in privacy preserving data mining. *IEEE Trans. Knowl. Data Eng.* 2012; 24(9): 1589–1612.
[18] Wu YH, Chiang CM, Chen AL. Hiding sensitive association rules with limited side effects. *IEEE Trans. Knowl. Data Eng.* 2007; 19(1): 29–42.

[19] Gkoulalas-Divanis A, Verykios VS. Exact knowledge hiding through database extension. *IEEE Trans. Knowl. Data Eng.* 2009; 21(5): 699–713.
[20] Le HQ, Arch-Int S, Nguyen HX, Arch-Int N. Association rule hiding in risk management for retail supply chain collaboration. *Comput. Ind.* 2013; 64(7): 776–784.
[21] Li YC, Yeh JS, Chang CC. MCIF: an effective sanitization algorithm for hiding sensitive patterns on data mining. *Adv. Eng. Inf.* 2007; 21(3): 269–280.
[22] Keshavamurthy BN, Toshniwal D, Eshwar BK. Hiding co-occurring prioritized sensitive patterns over distributed progressive sequential data streams. *J. Netw. Comput. Appl.* 2012; 35(3): 1116–1129.
[23] Chen CLP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inf. Sci.* 2014; 275: 314–347.
[24] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* 2014; 26(1): 97–107.
[25] Nergiz ME, Gök MZ. Hybrid K-anonymity. *Comput. Secur.* 2014; 44: 51–63.
[26] Li B, Erdin E, Gunes MH, et al. An overview of anonymity technology usage. *Comput. Commun.* 2013; 36(12): 1269–1283.
[27] Monreale A, Andrienko GL, Andrienko NV, et al. Movement data anonymity through generalization. *Trans. Data Priv.* 2010; 3(2): 1–31.
[28] Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for K-anonymity. *IEEE Trans. Knowl. Data Eng.* 2010; 22(3): 334–347.
[29] Zhang G, Yang Y, Liu X, Chen J. *A time-series pattern-based noise generation strategy for privacy protection in cloud computing*. Int. Symp Cluster, Cloud and Grid Computing (CCGrid). Ottawa. 2012: 458–465.
[30] Wang H. Quality measurement for association rule hiding. *AASRI Procedia*. 2013; 5: 228–234.
[31] Moustakides GV, Verykios VS. A MaxMin approach for hiding frequent item sets. *Data Knowl. Eng.* 2008; 65(1): 75–89.
[32] Wang SL, Parikh B, Jafari A. Hiding informative association rule sets. *Expert Syst. Appl.* 2007; 33(2): 316–323.
[33] Dasseni E, Verykios VS, Elmagarmid AK, Bertino E. Hiding association rules by using confidence and support. *Inf. Hiding Lect. Notes Comput. Sci.* 2007; 2137: 369–383.