Data Mining in the Context of Legality, Privacy, and Ethics

Amos Okomayin
Department of Computer Science
Middlesex University
London. United Kingdom

Tosin Ige
Department of Computer Science
University of Texas at El Paso
Texas, USA

Abosede Kolade
Department of Marketing and Bus.
Texas A&M University, Commerce
Texas, USA

Abstract

Data mining possess a significant threat to ethics, privacy, and legality, especially when we consider the fact that data mining makes it difficult for an individual or consumer (in the case of a company) to control accessibility and usage of his data. Individuals should be able to control how his/ her data in the data warehouse is being access and utilize while at the same time providing enabling environment which enforces legality, privacy and ethicality on data scientists, or data engineer during data mining process. This paper review issues of legality, privacy, and ethicality in data mining, review processes of Data mining, and also proposes solution to current ethical and privacy issue in data mining. It introduces a new method which enforces data mining without infringing on the privacy of individual or consumer whose data are being used. The sole aim of this paper is to propose a new method of mining data which restricts scientists within the constraints of legality, privacy, and ethicality.

Introduction

In an ethical sense, database security is closely related to privacy as it inhibits the unauthorized dissemination of personal data thus further enhancing, albeit indirectly, an individual's capacity to regulate access to their data. When data can be viewed from many angles and at abstraction levels, it threatens the goal of protecting data security and guarding against the invasion of privacy (Ige & Adewale 2022a). It is important to study when knowledge discovery may lead to an invasion of privacy, and what security measures can be developed for preventing the disclosure of sensitive information (Chen et al. 1996). The development of data warehouses has increased the importance of database security. Prior to this, data were typically held in separate databases to which access was controlled and limited to people with a specific functional role. Data warehouses bring together data from multiple sources and therefore more complex factors need to be considered when establishing security measures. In terms of database security, two forms of mining operation need to be considered:

1. Those operating as authorized applications by an individual or organization that owns and has full access to the data.
2. Those operating as unauthorized applications by an individual or organization that has access to the data only insomuch as has been permitted for other allowable purposes. Note that an individual need not be external to the organization that owns the data for the second point to occur. Conventional database security protects data via user authorization techniques (O'Leary 1991) making no distinction between the degrees of sensitivity present in the database (Mills 1997). A more sophisticated model, Multi Level Security (MLS), extends conventional security measures by classifying data according to its confidentiality (Elmasri & Navathe 2004, Ige & Adewale 2022b).

This chapter is analysis current model to privacy and ethical issues in data mining, implementation of the model, as well as limitations of the existing model which eventually calls for a new model that eventually guarantee and address problem of legality, ethically, and privacy in a well secured environment for knowledge discovery in database.

1.1 Limitation of existing Model

Firstly, existing data mining techniques involves exportation of database data to a file which can be in SQL, XML, JSON file format, and are then mined by querying the file using programming such as python, R, or SQL. This method has a disadvantage because data in such file format are not dynamic or in real time.

Let assume a database was exported by the database administrator on Sunday November 1, 2020 at 12:00 GMT for data mining, by the time we start the process of data mining like classification, association, clustering, regression, analysis, prediction and so on which can take several hours to days to complete. The record will not had been an updated record as our prediction would had been based on the exported file as of Sunday November 1, 2020 at 12:00 GMT which was when it was exported.

So, since the exported data in the file is in offline mode, not connected to the database data source for update, is not consistently updated real time as records are entering from the database from different source our analysis and prediction will not be accurate.

So existing data mining technique are only effective for static data but fails to address dynamic data which constantly changes.

Another limitation to existing data mining method is that it does not consider the unpredictability of mankind. For instance, a survey that shows the daily number of customer complaints for different stations like BBC, CNN, Aljazeera. If as at the time the database was exported for scientist to work on it, there might be more preferable for BBC, more complaints for CNN and Aljazeera cable network. Meanwhile, CNN might improve on its service after just few days, while the data scientist is still cleansing, sorting, classifying, and analyzing the exported database file. While the process of data mining is ongoing, more people changings their mind and moves away from BBC towards CNN due to it improve services.

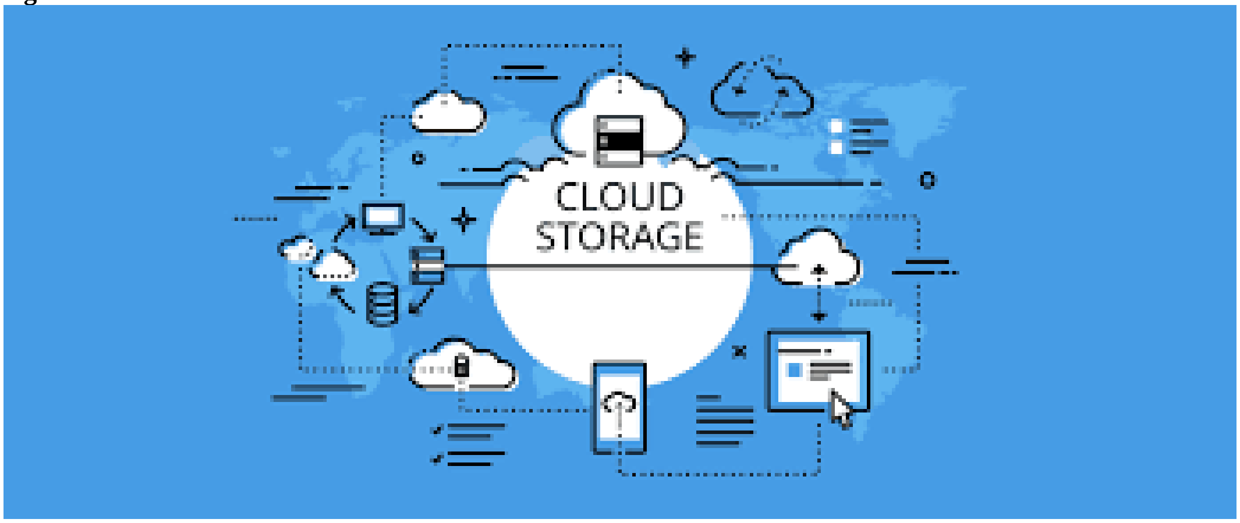
The fact that data mining process is still using the extracted data will render the result output inaccurate including the pattern and prediction.

Secondly, all existing measures to safeguard privacy in data mining are based on principle, meaning that data engineer or data scientist can decide to overlook them and infringe on individual privacy.

“ The laws regarding privacy are generally focused on protecting privacy rights from government actions. Currently, in the United States, policies related to privacy and business activities are largely voluntary. In contrast to the European Union, the only information privacy legislation to date is the Children's Online Privacy Act, which protects children from marketing research, and the Health Insurance Portability and Accountability Act, which protects medical information (Laczniak & Murphy 2006). This is to the benefit of businesses. However, it leaves businesses open to grave mistakes that can lead to significant financial losses and loss of consumer trust. Compliance with the letter of the law alone is not sufficient to prevent consumer dissatisfaction at the least and legal action at the most. Cary et al. (2003) suggest that the spirit of the privacy policy must be followed, not just the letter of the policy or the letter of the law. ‘Although legality generally stems from what society believes is morally right or wrong, an issue's legality does not always reflect the totality of its perceived morality. This differentiation reflects the classic distinction between the spirit of the law (morality) and the letter of the law (legality) (Raiborn & Payne 1990)’.

This article focuses on the issue of data mining as it relates to the consumer and to the issue of whether the consumer's private information has any proprietary status. A brief review of data mining is provided as a background for a better understanding of the purposes and uses of data mining. Also examined are several issues of the ethics of data mining, including a review of stakeholders, who they are, and which may be most seriously affected by unethical data mining practices. Several suggestions for the improvement of data mining as it relates to the consumer are further presented: suggestions that would allow for data mining that would be beneficial to both the business community and the consumer.”

Fig 2



Cloud Storage illustration (getfilecloud.com)

Thirdly, the existing solution to ethical issue and privacy concern in data mining fails to address third party remote access to data warehouse. It only address internal or in-house implementation without any external source in focus. This is another big flaws that had not yet been address in the world of big data.

Proposed Solution to ethical and privacy Issue in data Mining

Application of One way encryption

One way encryption are ways of encoding identity using complex algorithm calculation for maximum protection. Encrypting the name and address of an individual will give maximum protection and privacy of the individual. With one way encryption, the encrypted words can never be decrypted.

Modern way of Encrypting a word using combination of IVkey and Salt

```
   public Encrypter() {
      try {
         // Create the key
         KeySpec keySpec = new
PBEKeySpec(phrase.toCharArray(), salt, iterationCount);
         SecretKey key =
SecretKeyFactory.getInstance("PBEWithMD5AndDES").
generateSecret(keySpec);
         ecipher = Cipher.getInstance(key.getAlgorithm());
         dcipher = Cipher.getInstance(key.getAlgorithm());

         // Prepare the parameter to the ciphers
         AlgorithmParameterSpec paramSpec = new
PBEParameterSpec(salt, iterationCount);

         // Create the ciphers
         ecipher.init(Cipher.ENCRYPT_MODE, key,
paramSpec);
         dcipher.init(Cipher.DECRYPT_MODE, key,
paramSpec);
      } catch (NoSuchAlgorithmException |
InvalidKeyException | InvalidKeySpecException |
NoSuchPaddingException |
InvalidAlgorithmParameterException e) {

      }
   }

   private String encrypt(String str) {
      try {
         // Encode the string into bytes using utf-8
         byte[] utf8 = str.getBytes("UTF8");

         // Encrypt
         byte[] enc = ecipher.doFinal(utf8);

         // Encode bytes to base64 to get a string
         return new Base64().encodeAsString(enc);
      } catch (BadPaddingException |
IllegalBlockSizeException |
UnsupportedEncodingException e) {
      }
      return null;
   }

   private String decrypt(String str) {
      try {
         // Decode base64 to get bytes
         byte[] dec = new Base64().decode(str);
         byte[] utf8 = dcipher.doFinal(dec);

         return new String(utf8, "UTF8");
      } catch (BadPaddingException |
IllegalBlockSizeException |
UnsupportedEncodingException e) {

      }
      return null;
   }
}”
```

Calling the encryption method in the above class to encrypt "**Tosin Ige**" will give "**1wmPhXR0hYBhXVDoR/kBWHKtb/IKESS13aLv35d/8ys=".**

This will undoubtedly hide and protect the identity of an individual during data mining operation. Once, the name, and address of

CONCLUSION

We can see that it is possible to perform data mining on cloud database or data warehouse without infringing on the privacy of an individual by using a window service or web api to constantly doing one way encryption to identity part of an individual using dynamic data through hosted file without compromising the security and integrity of the database, as there is no direct connection to the database but the hosted file which is constantly being updated with records from the database by either a web API which acts as link bridge or background window service running at specific time interval to pull records from database and update it on the hosted file.
This is a better method of approach to prevent infringement on an individual privacy for three (3) reasons.

1. *The database detail remains secure and obscure from black hat hacker to a great length as there is no direct connection to the data warehouse but to the hostel file constantly updated by web API or background window service which constantly check every update records and then perform one-way encryption to the part that can lead to identity of an individual record in the database..*
2. *The records are not static but constantly changing inline with activities in the data warehouse to ensure scientist have the most updated data to work with.*
3. *The predictions and patterns are automatically generated in real time without any human intervention which in turn will help industries, companies, governments, consumers and so on to make informed decision and policy about products and services at appropriate time.*

This research is very much applicable in the field of data science and artificial intelligence especially machine learning being with high degree of data security as there is no direct connection to the data warehouse but the hosted file.

Reference

[1] Prakash, Hanumanthappa and Kavitha (2014), Big Data in Educational Data Mining and Learning Analytics, IJIRCCE.

[2] Nakamura, Nozaki, Nakayama, Morimoto and Miyadera (2015), Sequential Pattern Mining System for Analysis of Programming Learning History, IEEE.

[3] Ige, T., & Adewale, S. (2022a). Implementation of data mining on a secure cloud computing over a web API using supervised machine learning algorithm. International Journal of Advanced Computer Science and Applications, 13(5), 1–4. https://doi.org/10.14569/IJACSA.2022.0130501

[4] Ige, T., & Adewale, S. (2022b). AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. International Journal of Advanced Computer Science and Applications, 13(5), 5–9. https://doi.org/10.14569/IJACSA.2022.0130502

[5] Ramakrishnan Srikant and Rakesh Agrawal, Mining Sequential Patterns: Generalizations and Performance Improvements, IBM Almaden Research Center, CA 95120.

[6] Nizar R. Mabroukeh and C. I. Ezeife (2010), A Taxonomy of Sequential Pattern Mining Algorithms, ACM.

[7] Perera D, Kay J, Koprinska I, Yacef K, Zaïane, Clustering and sequential pattern mining of online collaborative learning data (2009), Knowledge and data engineering, IEEE.

[8] Shanabrook DH, Cooper DG, Woolf BP, Arroyo I (2010), Identifying high-level student behavior using sequence-based motif discovery, In: Proceedings of the 3rd international conference on educational data mining.

[9] Kinnebrew J, Biswas G (2012), Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution, In: Proceedings of the 5th international conference on educational data mining.

[10] Baker and Inventado (2014), Educational Data Mining and Learning Analytics, SPRINGER.

[11] D. Gandhimathi and S. Gomathi (2015), “A Survey of Approaches and Tools Used in Educational Data Mining”, IJIRCCE.

[12] Shabandar, Hussain, Laws, Keight, Lunn and Radi (2017), Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses, IEEE.

[13] Bell, Beck, Miller and Herrera (2007), Video Mining - Learning Patterns of Behaviour via an Intelligent Image Analysis System, IEEE.

[14] Swati Singh Lodhi (2014), Development of Sequential ID3: “An advance Sequential mining Algorithm”, AJSE.

[15] Nakamura, Nozaki, Morimoto and Miyader (2014), Sequential Pattern Mining Method for Analysis of Programming Learning History Based on the Learning Process, IEEE.

[16] Ratnapala and Deegalla (2014), Students Behavioural Analysis in an Online Learning Environment Using Data Mining, IEEE.

[17] Yadav and Jain (2011), Analyses of Web Usage Mining Techniques To Enhance the Capabilities of E-Learning Environment, IEEE.

[18] Banu and Ravanan, Analysis of E-learning in Data Mining – A Dreamed Vision for Empowering Rural Students in India, IEEE, 2011.

[19] Conde and García, Learning analytics for educational decision making, ELSEVIER, 2015.

[20] Edona Doko and Lejla Abazi Bexheti,, Systematic mapping study of educational technologies based on Educational Data Mining and Learning analytics, MECO, 2018.

[21] Nejati, E., Jahangiri, A. & Salehi, M. R. (2018). The effect of using computer-assisted language learning (CALL) on Iranian EFL learners’ vocabulary learning: An experimental study. *Cypriot Journal of Educational Sciences, 13*(2), 351-362. https://doi.org/10.18844/cjes.v13i2.752

iJIM – Vol. 12, No. 4, 2018 121

Paper—Sequential Pattern Mining Model to Identify the Most Important or Difficult Learning Topics via…

[22] Soykan, E. & Ozdamli, F. (2016). The Impact of M-Learning Activities on the IT Success and M-Learning Capabilities of the Special Education Teacher Candidates. *World Journal on Educational Technology: Current Issues, 8*(3), 267-276. https://doi.org/10.18844/wjet.v8i3.1019

[23] Bahadir, C. & Karahoca, A. (2017). Airline revenue management via data mining. *Global Journal of Information Technology: Emerging Technologies, 7*(3), 128-148.

[24] Stosic, L. (2017). Does the use of ICT enable easier, faster and better acquiring of knowledge? *International Journal of Innovative Research in Education. 4*(4), 179–185.

[25] Abdugulova, Z. (2017). Allowing schools access to affordable computers: How schools can benefit from switching to inexpensive, cloud-based computing technologies. *International Journal of Learning and Teaching. 9*(3), 326-331. https://doi.org/10.18844/ijlt.v9i3.507

[26] Uzunboylu, H., & Tugun, V. (2016). Validity and Reliability of Tablet Supported Education

Attitude and Usability Scale. *Journal of Universal Computer Science, 22*(1),