Discrimination Aware Data Mining in Internet of Things (IoT)

Asmita Gorave
Assistant Professor
Department of computer Engineering
MIT College of Engineering,
Pune, India

Vrushali Kulkarni
Head of Department
Department of computer Engineering
Maharashtra Institute of Technology,
Pune, India

ABSTRACT
IoT is a technology where objects around us will be able to connect to each other and communicate via Internet. Recently, IoT has become an important technology in the world of Internet. However, it comes across problem of big data and extracting knowledge from such data using data mining techniques. Recently, it is observed that data mining increases the risk of violation of fundamental human right, called non-discrimination. It is obvious that data mining tasks in IoT will also face the risk of discrimination. Discrimination Aware Data Mining is an area which deals with finding methods to discover and/or prevent discrimination. This paper is the first step towards finding discrimination related issues in IoT. This paper describes discrimination discovery and prevention issues faced by IoT. This paper also specifies the huge future research avenue related to discrimination aware data mining in IoT.

Keywords
Discrimination aware data mining; data mining; Internet of Things (IoT)

1. INTRODUCTION
Internet of Things (IoT) is getting attention of lots of researchers recently. As Internet is becoming very important thing in day-to-day life of human being, IoT has come into picture. IoT means connecting the things (it may include smart devices, smart phones, cameras, sensing devices, etc.) to the Internet [1]. IoT faces many problems while connecting such things [2]. Such problems include massive scaling, architecture and dependencies, robustness, openness, security, privacy and human-in-the-loop, creating knowledge and big data. Lots of research is going on in the field of IoT in order to resolve these problems. Broadly these problems can be categorized into three areas: networking, data and/or information processing and human-computer interaction. Survey of researches to resolve these problems is presented in [3] [4] [5] [6] [7]. As IoT involves connecting millions of things together, it is required to deal with large amount of data, which is called "Big data". New architectures must be created in order to handle connection of millions of things. Security and privacy are also the important issues that need to be considered while transferring data and information among large number of interconnected things. Lots of methods and/or techniques are already taken into consideration by many researches to solve the above mentioned problems. Many new architectures and/or techniques are proposed to connect large number of devices and handle "big data" in IoT.

As mentioned above, creating knowledge and big data is one of the problems faced by IoT. This paper focuses around the problem of creating/extracting knowledge from big data in IoT. Data mining is a technique to extract useful knowledge from raw data. Either existing data mining techniques can be used or new data mining techniques can be created to extract knowledge from big data. However, recently it is observed that data mining increases the risk of violation of fundamental human right: non-discrimination [8]. As data mining is done on the "big data" collected from things in IoT, discrimination can occur as a side effect of data mining tasks in IoT also. In IoT, data will be collected from many devices such as sensors. Such sensors work on different platforms and in different physical conditions. If data are collected in wrong way from such sensors or analyzed in wrong way, discriminatory decisions can be generated from such data. If the collected data is discriminatory towards particular community and the same discriminatory data is used to extract decisions using data mining tasks, then the extracted decisions will also become discriminatory. In this way discriminatory decisions can be generated as a side effect of data mining tasks in IoT.

Research is going on in order to avoid the risk of discrimination due to data mining tasks from the year 2008 [8]. The area where the related research is going on is called Discrimination Aware Data Mining (DADM). It deals with finding methods to discover, prevent and/or measure discrimination.

As discussed above, it can be concluded that, DADM must be incorporated with IoT in order to make IoT free from the risk of discrimination. To prevent or discover discrimination in IoT will be an important research avenue. This paper is the first step towards finding research directions related to DADM and IoT.

The rest of the paper is organized as follows. Section 2 provides overview of the work related to DADM and IoT. Section 3 provides the discussion about DADM and IoT. Section 4 provides our views related to DADM and IoT under the heading "Discussions" and section 5 specifies conclusion and future research directions in the field of DADM and IoT.

2. RELATED WORK
Initially the problem of discrimination is presented by the researcher D. Pedreschi in 2008[8]. In [9] different methods are proposed to discover direct and indirect discrimination. Discrimination can be prevented in one of the three ways [10]: pre-processing, in-processing and post-processing. In pre-processing the original discriminatory data is transformed such that no discriminatory decision is made. In case of in-processing approach, standard data mining algorithms are changed in such a way that no discriminatory decisions are made. Post-processing approach deals with changing the final results of data mining tasks in order to remove discrimination.

Research in DADM deals with developing different discrimination prevention methods using any of the above three approaches. In [10] [11], different methods for discrimination prevention using preprocessing approach are shown. In [12], modifications are done to standard decision tree techniques to prevent discrimination. Here both in-processing and post-processing discrimination prevention approaches are used. Standard Naïve Bayes algorithms are modified to prevent discrimination in [13]. It also uses both in-processing and post-processing approaches. Different metrics to measure amount of discrimination is given in [14].

Work related to data mining in IoT is specified in [4]. It presents survey of different data mining techniques which can be used to extract knowledge from "big data" in IoT. Different research directions in IoT is specified in [2]. These research directions include creating knowledge and big data. Information processing for IoT is given in [15].

3. DADM AND IOT

As mentioned in the section 1, data mining in IoT can come across the risk of discrimination. So it is important to find ways in order to handle discrimination in IoT. This section puts light on some of the important issues in order to handle discrimination in IoT. Issues related to discrimination in IoT are mentioned below:

3.1 Discrimination Discovery in IoT

No work has been done till the date to discover discrimination in IoT data mining tasks.

3.1.1 Discrimination discovery process [9]

The method to discover direct discrimination is as follows:

3.1.1.1 Identify the potentially discriminatory (PD) attributes

Anti-discrimination laws of different countries prohibit discrimination based on certain attributes such as age, gender, nationality, religion, color, marital status etc. Such attributes are called potentially discriminatory attributes as they have high chances of discrimination.

3.1.1.2 Identify potentially discriminatory classification rules

Knowledge is extracted from the data using data mining tasks in terms of classification rules. If the extracted classification rule contains PD attribute, then the rule is called potentially discriminatory rule.

3.1.1.3 Apply discrimination measure and discrimination threshold to identify discrimination

Discrimination measure such as elift and discrimination threshold (α) are used in order to see whether the extracted classification rule is really discriminatory.

3.1.2 Discrimination discovery issues faced by IoT

Above section specifies common discrimination discovery process. Some issues must be considered while doing discrimination discovery in IoT data mining tasks:

3.1.2.3 Deciding discrimination threshold (α)

Discrimination discovery is highly dependent on discrimination threshold (α). Such α is calculated from laws of different countries. Deciding α is the most difficult problem faced by discrimination discovery in IoT, as data may be collected from different countries. IoT will have international data and α will be different in different countries, so it will be necessary to agree upon some common rules to find a common α.

3.1.2.4 Deciding PD attributes

Same problem as that of deciding α can occur in case of deciding PD attributes. Laws of different countries will have different PD items. E.g. if color is PD in one country, in other country it may be non-PD. So research must be done in order to deciding α and PD items for data in IoT.

3.2 Discrimination Prevention in IoT:

Discrimination can be prevented in one of the three ways: pre-processing, in-processing, post-processing.

3.2.1 Discrimination Prevention (pre-processing)

Here discrimination is discovered in terms of discriminatory classification rules and then the original discriminatory data is transformed. Now the discriminatory rules will not be mined from the transformed data [10].

3.2.2 Discrimination prevention issues faced by IoT (pre-processing)

While transforming the original dataset, IoT may face many issues:

3.2.2.1 Heterogeneous data

In IoT, data will be collected from many things and from many heterogeneous platform, the data must be converted into a common format in order to prevent discrimination. It will be altogether a different search topic.

3.2.2.2 Removing noise from original data

It will be necessary to remove noise from the original data before pre-processing it, otherwise wrong results will be mined.

3.2.3 Discrimination Prevention (in-processing)

Here original data is not transformed, but the standard data mining algorithms are changed in such that no discriminatory decisions are made. Following are the techniques in order to prevent discrimination using in-processing techniques:

3.2.3.1 Discrimination prevention using decision tree creation [12]

Here while choosing splitting criteria to create decision tree, not only accuracy of the split is considered, but also the discrimination caused by the split is considered. The split which causes minimum discrimination is selected.

3.2.3.2 Discrimination prevention using Naïve Bayes [13]

A hidden variable is used to train Baysian model.

3.2.4 Discrimination prevention issues faced by IoT (in-processing)

While changing standard data mining algorithms in order to remove discrimination, many issues can be faced by IoT:

3.2.4.1 Choosing the optimum data mining algorithm

Choosing the standard data mining algorithm to be modified is an important issue. Standard data mining algorithms need to be modified in order to remove discrimination in IoT. Most of the work is related to modifying classification algorithms.

3.2.4.2 Performance of the algorithm

Performance of the algorithm will be dependent on the aim of the IoT application. Depending upon the nature or objective of

the IoT application, the standard data mining technique must be chosen.

3.2.5 Discrimination Prevention (post-processing)

Here neither original data are transformed nor is standard data mining algorithm changed. Results of data mining tasks are changed such that discrimination is removed. Following are the techniques in order to prevent discrimination using post-processing techniques:

3.2.5.1 Leaf relabeling

Class labels at the leaves of constructed decision tree are changed in order to remove discrimination.

3.2.5.2 Naïve Bayes

Naïve Bayes classifier is used and probability of positive decision is modified.

3.2.6 Discrimination prevention issues faced by IoT (post-processing)

Some issues are needed to be considered while using post-processing approach for discrimination prevention in IoT:

3.2.6.1 Publishing of modified results

As modified results are published, data mining can be performed only by data holder. It is necessary to decide format of the published data while publishing the IoT application results.

Table 1 summarizes the issues related to discrimination aware data mining in IoT.

Table 1. Discrimination related issues faced by IoT

Discrimination Techniques	Issues faced by IoT
Discrimination discovery techniques [9]	Deciding discrimination threshold (α) Deciding PD attributes
Discrimination prevention techniques (pre-processing) [10][11]	Heterogeneous data Removing noise from original data
Discrimination prevention techniques (in-processing)[12][13]	Choosing the optimum data mining algorithm Performance of the algorithm
Discrimination prevention techniques (post-processing)[12][13]	Publishing of modified results

4. DISCUSSIONS

In section 3, different discrimination related issues faced by IoT are discussed. This discussion opens a large research avenue in the field of discrimination aware data mining and IoT. This is the first attempt to combine these two fields. From section 3, it is clear that, it is necessary to do combined study in these two fields. While deploying IoT applications and collecting data from IoT things, discrimination can happen. Either the existing discrimination discovery techniques can be used in IoT or new discrimination discovery techniques can be created for IoT.

IoT deals with large amount of social network data. Large amount of discrimination happens in such social network data. So it is very important to find new text mining techniques to discover and/or prevent discrimination in such data.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

As per the discussions in section 3 and 4, the two areas discrimination aware data mining and IoT are related to each other. It can be concluded that there is a large scope for combined research in these two fields. IoT is becoming very important technology in the world of Internet. It faces problem of extracting knowledge from big data. And data mining faces the risk of discrimination. So it is very obvious that IoT data mining tasks will also face the risk of discrimination. As research is going on to discover and/or prevent discrimination in data mining, the same research can be extended for IoT data mining applications. This paper is the first attempt to explore issues related to DADM in IoT. This paper can become a guide towards creating smart systems to prevent discrimination in IoT.

The future research directions related to DADM in IoT are mentioned below:

5.1 Discrimination discovery and/or prevention for social network data in IoT

IoT deals with large amount social network data collected from Internet [16]. In DADM discrimination prevention techniques are based on classification data mining techniques. However for IoT, it is necessary to discover and/or prevent discrimination using text mining techniques, as it deals with social network data.

5.2 Deciding discrimination parameter according to laws

As IoT deals with data from different sources from different countries, it will be a new research avenue to find ways to decide discrimination parameters such as discrimination threshold, discriminatory attributes etc.

5.3 Creating/modifying discrimination prevention techniques in IoT:

It is necessary to decide whether to use existing discrimination prevention techniques for IoT or to create new discrimination prevention techniques for IoT. This is also a new research area for IoT.

5.4 Deciding approach of discrimination prevention:

From discrimination prevention literature, it is clear that pre-processing techniques are better in discrimination prevention than in-processing or post-processing [10]. It will be interesting to check which techniques are better for IoT. This will be useful to create new discrimination prevention techniques for IoT.

5.5 Finding metrics to measure discrimination in IoT:

In the existing DADM literature, the metrics such as elift, slift, olift etc. are used to measure discrimination in the data. For IoT such metrics may not be useful. So new metrics are

required to be created according to laws to measure discrimination in big data in IoT.

6. REFERENCES

[1] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context-aware computing for the Internet of Things : a survey," IEEE Communications Surveys & Tutorials, submitted 2013.

[2] John A. Stankovic, "Research Directions for the Internet of Things," IEEE Internet of Things Journal, vol.1, no.1, pp. 3-9, February, 2014.

[3] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: vision, application and research challenges," Ad Hoc Networks, vol.10, no. 7, pp. 1497-1516, 2012.

[4] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang and Laurence T. Yang, "Data Mining for Internet of Things: A Survey," IEEE Communications Surveys & Tutorials, vol. 16, no.1, pp. 77-97, 2014.

[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, no.7, pp. 1645-1660, 2013.

[6] A. S. Elmaghraby and M. M. Losavio, "Cyber security challenges in Smart Cities: Safety, security and privacy," J. Adv. Res., vol. 5, no.4, pp. 491-497, 2014.

[7] S. Sicari, A. Rizzardi, L.A Grieco and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," Comput. Netw. 76, 146-164, 2015.

[8] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

[9] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.

[10] S. Hajian & J. Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination prevention in data mining," IEEE transaction on knowledge & data engg., pp. 1445-1459, 2013.

[11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Intl' Journal of Knowledge & Information Systems, Springer, Vol.33, no.1, pp. 1-33, 2012.

[12] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[13] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[14] D.Pedreschi, S.Ruggieri and F.Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09), pp. 581-592, 2009.

[15] Frieder Ganz, Daniel Puschmann, Payam Barnaghi and Francois Carrez, "A Practical Evaluation of Information Techniques for the Internet of Things," IEEE Internet of Things Journal, vol.2, no.4, pp. 340- 354, 2015.

[16] L. Wang, "Using the relationship of shared neighbors to find hierarchical overlapping communities for effective connectivity in IoT," in *Proc.* International Conference on Pervasive Computing and Applications, pp. 400–406, 2011.