When Privacy Meets Policy: Privacy Preserving Data Mining

Nicole K. Merritt

Niagara University

Table of Contents

Abstract

We live in a world where we are obtaining information way faster than we can figure out what to do with. This data is often tied to sensitive information such as credit card numbers, shipping addresses, medical records, social insurance numbers, and numerous other personal identifiers. Data mining has become a key topic of study in recent years. The question currently is how to gain valuable information for mined data while simultaneously protecting customer privacy. This paper discusses the key themes within current literature about data preserving data mining as well as looking at the GDPR and the role it plays in the future of data mining as well as proposes solutions to address the issue of protecting consumer privacy while usable knowledge.

Keywords: data mining, policy design, privacy standards, policy creation, GDPR

When Privacy Meets Policy: Privacy Preserving Data Mining

We are collecting more information than ever before, "customer interaction and learning relationships require capturing information everywhere" (Marakas, 2003). Data collection can be seen in a multitude of sectors. For example, in a retail setting customers purchase history can be used to predict trends as well as advertise additional products to customers who have similar buying patterns. Credit card companies can use mined data to watch for suspicious activities and track patterns that are out of the ordinary for the client. In the health sector, patient data can be complied to learn more about an illness or make breakthroughs with potential cures.

Problems arise when we have such a vast amount of user data tied to credit cards, shipping addresses, medical records, social insurance numbers, and numerous other personal identifiers. In a recently conducted survey it is noted that "97% of the population of the United States can be uniquely identified using the seemingly innocuous attributes of gender, date of birth, and 5-digit zip code" (Machanavajjhaka et al., 2007). Continually a growing concern, "technology appears to create new ways to acquire information faster than the legal system can handle the ethical property issues" (Marakas, 2003).

Unless properly addressed, tracking a customer's purchases could be a violation of consumer privacy agreements. As business continues to automate more and more of its processes, and in turn collecting increasing amounts of information on its customers, this will continue to become a larger issue. What we are left with is the question of how we can implement policies that both protect consumer's privacy while maintaining the information's value.

Literary Review

What is Data Mining?

Data mining, in short, is “the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” (Hand et al., 2001). Data mining allows us to gain valuable information from vast amounts of data faster than ever before. First emerging in the 1990’s, advances in data mining have occurred due to a couple of different factors. First, the automated of tasks such as credit card transactions, bank withdrawals, and web searches. Secondly, the growing ability for companies to store massive amounts of data.

The process of data mining follows a common path regardless of the type of data that is being mined. All data comes in as raw unprocessed information and is prepared, The preparation cleaning the data which includes looking for missing values, eliminating key identifies, and removing irrelevant or repeated values to ensure accurate results. From there, the data is then “passed to a data mining algorithm which produces and output in the form of rules or some other kind of ‘patterns’. These are interpreted to give – and this is the Holy Grail for knowledge discovery – new and potentially useful knowledge” (Bramer, 2016).

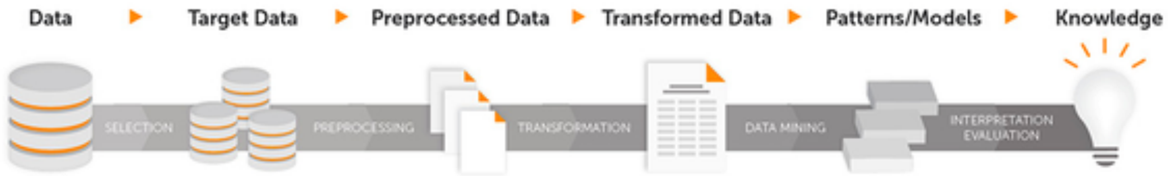

Figure 1: Process of data mining

Privacy & Productivity of Mined Data.

It is important to know where to draw the line when data extraction becomes an invasion of privacy and to understand “what security measures can be developed for preventing the disclosure of sensitive information” (Chen et al., 1996). However, privacy and productivity don’t exactly go hand in hand all the time, “privacy and accuracy are typically contradictory in nature, with the consequence that improving one usually incurs a cost in the other” (Rizvi & Haritsa, 2002).

The solution to a “perhaps infeasible goal of having both complete privacy and complete accuracy through approximate solutions that provide practically acceptable values for these metrics”. These approximate solutions can be utilised in the business sector since “cent-per-cent accuracy in the mining results is perhaps often not even a required feature” (Rizvi & Haritsa, 2002) when trying to track trends.

Policy Creation.

Grossman, Hornick, and Meyer (2002) suggests that there needs to be a common standard “for cleaning, transforming, and preparing data for data mining” as well as an agreement on a common set of web services for working with remote and distributed data. The authors note that there are currently a multitude of standards in the industry because data mining is used in many ways using a wide variety of different systems. Furthermore, “standards for cleaning and transformation data are only beginning to emerge” and because of this the implication of “using data mining in operation processes and systems are relatively immature”.

Privacy Meets Policy

Agawal et al., 2003 in the paper “Information sharing across private databases” proposes a couple of different solutions for protecting customer’s privacy, first being information shared only on a “need-to-know basis” and the second being the introduction of a trusted third party, “the main parties give the data to a “trusted” third party and have the third party [complete the data] compilation”. However, it is concluded that this solution is risky for both parties involved. This is because there is a very high level of trust required on both sides in order for this model to work.

Furthermore, organizations commonly “use some form of transformation on the data in order to perform the privacy preservation” however what usually happens is the process is a reduction in “granularity of representation in order to reduce privacy” which in turn causes the data to lose value. Methods usually involve some sort or randomization of the data, with “noise” added in order to ensure that “individual record values cannot be recovered” (Agawal et al., 2003). Questions raised from this method asks if the value of the data is lost in the process.

Another key method is that of k-anonymity model and l-diversity. Using the k-anonymity method, there is a reduction in the “granularity of data representation with the use of techniques such as generalisation and suppression”. L-diversity is employed to handle issues with k-anonymity, “the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme” (Agawal et al., 2003).

Proposed Privacy Solutions.

Iyenar (2002) notes that data is often “disseminated and share[d] within the organization collecting it and with other organizations. This dissemination could be to satisfy some legal requirements or as part of some business process”. One method of protecting personal information is to create a placeholder in the data, that is to take out a key identifying factor, such as a social insurance number, and replace it with a random token. It has however been noted that this “is not sufficient since the released data contains other information which when linked with other data sets can identify or narrow down the individuals or entities” (Iyenar, 2002).

Clifton et al. (2002) suggest that what is needed is a “toolkit of components that can be combined for specific privacy-preserving data mining applications”. What this basically means in practice is the creation of protocols that “can be used to make several standard data mining algorithms into privacy preserving distributed data mining algorithms”. These algorithms are used to process the data and insure that privacy is kept intact. This is done in a multitude of ways, such as EM clustering and vertically partitioning the data. While these approaches are a step in the right direction, it is concluded that “showing that this does not violate privacy may be difficult or impossible”. More research on this method is required to make this a more feasible option.

Another proposed solution is “the development of a class of data mining algorithms that try to extract the data patterns without directly accessing the original data” and thus, “guarantees that the mining process does not get sufficient information to reconstruct the original data (Kargupta et al., 2003). It is often concluded that “it is relatively easy to breach the privacy protection offered by the random perturbation based techniques”. Kargupta et al (2003) conclude by suggesting looking into “colored noise” in order to mask data and preserve privacy within data mining. Here, “randomization-based techniques are likely to play an important role” in privacy but need to be tweaked more to ensure the value of the data is not eliminated in the process.

If privacy wins over productivity data may be stripped to the point where nothing of value can be extracted. On the other side of the coin, if productivity outweighs privacy raw data may not be transformed enough to ensure individuals personal identifies are not linked together.

What has emerged in light of this conversation is privacy preserving data mining. Here, the basic idea being that the data is modified to ensure that individuals privacy is kept while gaining valuable information from the data mining algorithms, “the objective of privacy preserving data mining is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data” (Xu et al., 2014).

GDPR

Leaving companies in charge of data privacy is like leaving a fox in charge of a hen house, because of this the European Union has taken steps to protect its citizens. Taking matters out of the hands of companies, the EU has established “The General Data Protection Regulation” (GDPR) and within it aims to “ensure a consistent level of protection for natural persons throughout the Union and to prevent divergences hampering the free movement of personal data within the internal market” and furthermore to “provide natural persons in all Member States with the same level of legally enforceable rights and obligations and responsibilities for controllers and processors, to ensure consistent monitoring of the processing of personal data” (Council of the European Union, 2016).

Findings

Companies run the risk of collecting too much data, putting into question the businesses intentions and strategy. Over mining your customer can make them suspicious and become more guarded with their information or in some cases refusing to give it at all. Data protection is extremely important. With database breaches happening regularly, name any big corporation and they have probably had a data breach in the last 10 years, companies need to ensure their customers that they are taking any and all necessary precautions to ensure that their information is being reasonably protected.

Currently, the GDPR aims its policy at organizations and requires them to be extremely transparent with individuals in regards to the use of their data as well as acknowledging when it has been compromised. There are however places where this protection becomes questionable. Assuming that alerting individuals of what they are consenting to goes far enough in protecting their rights is extremely naive. For example, when it comes to giving individuals the ability to consent to the use of their data, “most people just tick consent boxes without reading or understanding privacy statements, or that service providers sometimes assume that website visitors are somehow miraculously informed of the privacy statement and automatically give consent by merely visiting the website” (Koops, 2014). Making privacy statements that are more user friendly, straight forward, and quick to highlight how data will be used can help inform consumers and make the process more transparent.

Value of Data Mining.

Finding patterns and making connections in data is a critical element in our advancement. It is becoming common practice for organizations to “cooperate in certain areas and compete in others, which requires selective information sharing” similarly, “government agencies need to share information for devising effective security measures” as well as catching criminals as they cross state and country boarders. (Agrawal et al., 2003). Information cannot be valuable if it lives in a vacuum. It becomes valuable when it is connected with other sources of information to paint a broader picture of what is going on.

The value of data mining stretches beyond customer relationships and sales objectives, “anomaly direction is used to detect fraud or risks within critical systems … it can help to find extraordinary occurrences that could indicate fraudulent actions, flawed procedures or areas where a certain theory is invalid” (Rijmenam, 2014). Finding abnormal spending patterns and detecting flaws in systems help both the customer and credit card companies save money as well as help prevent leaks of critical information, “data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common every day network activity” (Big Data Made Simple, 2014).

Mined data has given organizations the opportunity to react quickly to customer needs. For example, a organization can “restructure a web site in order to better serve the needs of users of a site” (Cooley et al., 2014). Furthermore, the utilization of data mining has allowed “organizations [to] tailor their products and services and interact with the customers based on actual customer preferences, rather than some assumed general characteristics” (Shaw et al., 2001) which can make a large difference in building customer relationships.

However, customers may not be comfortable providing information because their privacy may be compromised and may in turn provide false information. This information could skew the data to the extent of making it unusable. To combat that, encouraging the customer to “provide correct information by ensuring that the mining process cannot, with any reasonable degree of certainty, violate their privacy” (Rizvi & Haritsa, 2002).

Recommendations

The introduction of a third party in a organization’s data mining process will become key in insuring privacy and with this, policy creation will shift on the integrity of the third party as it deals with customer’s information. Employing a third party allows experts in the field to prepare the data and create reports that can be of best use to the organization. Importantly, the goal is to outsource the task, not outsource control over the task. Putting clear policy and audits in place, the organization can ensure that data is being processed to their standards and is not being leaked to the wrong people.

To ensure privacy is met, organizations that employee a third party need to ensure that data transmission is seamlessly encrypted end to end, clear guidelines are established to outline the target of the data mine, and extensive transformation of the data is completed to ensure anonymity of data providers. Both the organization and the third party can mutually benefit from the outsourcing of data. For example, the organization itself can save time and money. The third party can add the raw data given to them to a larger data bank, which can serve them well in researching patterns moving forward.

There is a perceived lack of control when it comes to third party involvement in computing and because of this the service is not currently being used to its full potential. With the third party potentially hosting a large amount of data you have the possibility of creating a honeypot, where attackers have the ability to gain a large amount of raw data quickly. This concern can be mitigated through the transformation and anonymization of data, “this means that some identifying data will be removed such as IP addresses and cookie information” (Chow et al., 2000).

As a proposed solution, the idea of information-centric security is emerging. With this, the data is protecting itself from within and “requires intelligence be put in the data itself”. The idea here is that the data can only be seen “if the environment is verified as trustworthy” (Chow et al., 2009). This takes the security out of the hands of the organization or the third party and places constant security within the data as well as eliminates this lack of control that is often perceived with the use of a third party service. This technique to be expanded to include data release on a need to know bases. Where the data itself will only release the information that is required at that time for the task to be completed.

The key is to ensure that the protection of consumer privacy does not undercut the data's value. Otherwise, there is little point in the process altogether. While there is a fine line between productivity and privacy, compromise can be made to ensure that values from both sides are intact. With information-centric security, raw data can be secured from point A to B but its value is still maintained in the process.

Truth be told, when data is given to others, whether it be an organization or a third party acting on behalf of the organization, there is no assurance that this information will not fall into the wrong hands, or be protected in such a way that is satisfactory to the data provider. However, if the proper precautions are taken, encrypting the data in transit along with the use of information-centric security for example, this risk can be dramatically minimised.

Conclusion

Privacy will continue to be a top concern in data mining. Customers are more tuned into what is going on, and because of this companies can no longer sweep breaches under the rug. With each breach customers become more uneasy about providing data. To counter act this fear it is important that companies establish a strong relationship with the customers.

Despite privacy concerns, data mining is a very valuable tool for organizations to utilize. Beyond sales, credit card companies can protect its customers from fraud more accurately, medical providers can prescribe medications with confidence, and advancements in disease treatments can be made. Policy still has a long way to come to create a comprehensive set of guidelines that covers all sectors, however significant advancements have been made with the introduction of the GDPR in ensuring the rights of the consumer are withheld.

With the creation of information-centric security, organizations can feel confident sending their data to a third party to process, reveal trends, and pass on usable knowledge to the organization more cost effectively than ever before. This approach allows the data to hold the power of authenticating its surroundings and only revealing its value if the system trying to open it passes the authentication process.

Moving forward, organizations have no choice but to adhere to stricter consumer privacy standards in order to operate in the international marketplace. With enforcement of the GDPR beginning in May of 2018, businesses are scrambling to be ready for its roll out to avoid hefty sanctions. Time will tell how much wait these sanctions hold and if they will be upheld to their full extent.

References

Aggarwal, C. C., & Yu, P. S. (2004). A condensation approach to privacy preserving data mining. *Internation Conference on Extending Database Technology*. Retrieved from https://www.researchgate.net/profile/Charu_Aggarwal/publication/221103454_A_Condensation_Approach_to_Privacy_Preserving_Data_Mining/links/02bfe514321af10a74000000.pdf

Agrawal, R., Evfimievski, A., & Srikant, R. (2003). Information sharing across private databases. *ACM SIGMOD International Conference on Management of Data*, *1*(1), 86-97. Retrieved from https://webpages.uncc.edu/~xwu/privacy/sigmod03.pdf

Alexander, D. (n.d.). Data mining. Retrieved December 12, 2017, from https://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

Bramer, M. (2007). *Principles of data mining*. https://doi.org/10.1007/978-1-84628-766-4

Chan, P. K., Fan, W., & Prodromidis, A. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems' Special Issue on Data Mining*, 1-17. Retrieved from https://pdfs.semanticscholar.org/9dea/762c875fe0e1db5b2fee76641b7cd7056381.pdf

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866-883. Retrieved from https://cs.nju.edu.cn/zhouzh/zhouzh.files/course/dm/reading/reading01/chen_tkde96.pdf

Chow, R., Golle, P., Jakobsson, M., Shi, E., Staddon, J., Masuoka, R., & Molina, J. (2009). Controlling data in the cloud: Outsourcing computation without outsourcing control. *Proceedings of the 2009 ACM Workshop on Cloud Computing Security*, pp. 85-90. Retrieved from http://masuoka.net/Ryusuke/papers/20091113-Controlling%20data%20in%20the%20cloud-

2009%20ACM%20Workshop%20on%20Cloud%20Computing%20Security-CCSW-09-Paper.pdf

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. *Explorations Newsletter*, *4*(2), 28-34. Retrieved from http://helios.mm.di.uoa.gr/~rouvas/ssi/sigkdd/sigkdd.vol4.2/clifton.ps

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1-27. Retrieved from https://www.researchgate.net/profile/Bamshad_Mobasher/publication/2534306_Data_Preparation_for_Mining_World_Wide_Web_Browsing_Patterns/links/0f3175339a3a2c4271000000.pdf

Council of the European Union. (2016). *General Data Protection Regulation* (Report No. 2012/0011) (Council of the European Union, Author). Retrieved from http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

Dimkovska, K. (2017, June 28). The value of data mining. Retrieved December 12, 2017, from Vividus website: https://vividus.com.au/data-mining/value-data-mining/

Evfimievski, A., Gehrke, J., & Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. *ACM SIGMOD SIGACT SIGART symposium on principles of database systems*, *1*(1), 211-222. Retrieved from http://rsrikant.com/papers/pods03.pdf

Evimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2004). Privacy preserving mining of association rules. *Information Systems*, *29*, 343-364. Retrieved from http://www.cs.cornell.edu/aevf/research/InfoSystems_2004.pdf

Grossman, R. L., Hornick, M., & Meyer, G. (2011). Data mining standards initiatives. *HR Magazine*, 22-27. Retrieved from http://web.a.ebscohost.com.ezproxy.niagara.edu/ehost/pdfviewer/pdfviewer?vid=3&sid=e33f1e63-693b-4b86-9a21-5ad51287d8fa%40sessionmgr4008

Han, J. (2006). *Data mining: Concepts and techniques* [Lecture transcript]. Retrieved December 12, 2017, from http://liacs.leidenuniv.nl/~bakkerem2/dbdm2007/05_dbdm2007_Data%20Mining.pdf

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. *Principles of Data Mining*, *1*(1), 1-546. Retrieved from https://mitpress.mit.edu/books/principles-data-mining

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *1*(1), 279-288. Retrieved from https://pdfs.semanticscholar.org/2676/d77b4e4cc58250ed20b4f85576a9fb33ae5a.pdf

Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques. *IEEE International Conference on Data Mining*, 99. Retrieved from http://www.eecs.wsu.edu/~siva/icdm03.pdf

Koops, B.-J. (2014). The trouble with European data protection law. *International Data Privacy Law*, 250-261. Retrieved from http://www.isaca.org/Groups/Professional-English/privacy-data-protection/GroupDocuments/2014-08-24%20%20The%20Trouble%20with%20European%20Data%20Protection%20Law.pdf

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 1-52. Retrieved from https://pdfs.semanticscholar.org/004b/439ff1e6a15deedc7a7c4c6685f5ceafd237.pdf

Marakas, G. M. (2003). *Modern data warehousing, mining, and visualization: Core concepts* [Lecture transcript]. Retrieved December 12, 2017, from http://alainmaterials.webs.com/handouts/Data_Warehousing/Handout6.pdf

P, R. (2014, August 20). 14 useful applications of data mining. Retrieved December 12, 2017, from Big Data Made Simple website: http://bigdata-madesimple.com/14-useful-applications-of-data-mining/

Rizvi, S. J., & Haritsa, J. R. (2002). Maintaining data privacy in association rule mining. *International Conference on Very Large Data Bases*, 682-693. Retrieved from http://www.vldb.org/conf/2002/S19P03.pdf

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*, 127-137. Retrieved from http://sites.google.com/site/damonhu/km_dmForMarketing.pdf

Van Rijmenam, M. (2014). Five data mining techniques that help create business value. Retrieved December 12, 2017, from Datafloq website: https://datafloq.com/read/data-mining-techniques-create-business-value/121

Xu, L., Jiang, C., Wang, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access*, *2*, 1149-1176. Retrieved from http://ieeexplore.ieee.org/document/6919256/