ON ETHICAL AND LEGAL ASPECTS OF DATA MINING

Mehmet Cudi OKUR*

ABSTRACT

Data mining technology allows large volumes of data to be exploited for discovering previously unknown,possibly useful knowledge. The speed and extent of developments in information technologies have increased the power and potential of data mining.However, the privacy of personally sensitive information is not respected generally in the process,which creates ethical and legal problems in some applications.In this study,the ethical and legal aspects of data mining are explored and a critical evaluation of the protective methods is presented.

Keywords: Data mining , knowledge discovery,computer ethics,privacy reservation,sensitive information.

1.BACKGROUND

The users of information technology tools and techniques are well aware that, data about almost every aspect of daily life are being gathered,stored,manipulated and used by public or private organizations and governments. The extent of these applications has been growing since the introduction of data processing methodologies such as data warehousing,on line analytical processing and data mining.Major software vendors including Oracle,IBM,Microsoft and most statistical software packages offer increasingly powerful tools and options for intelligent use of large volumes of data for decision support and fact finding.Several scientific and commercial variations of the related methodologies are used in diverse application areas.The interest is still high in this field and will continue to be so in the foreseeble future .Among the current intelligent technologies,data mining has been specifically popular in most application areas, where knowledge extraction from large data sets is the pimary goal.The driving force behind all these developments is the ubiquitous

* Department of Computer Engineering, Yasar University, Izmir,Turkey. mehmet.okur@yasar.edu.tr

nature of the information technologies and their ever increasing data storage ,processing and communication capabilities.

Most data mining applications aim to extracting useful new information from large volumes of structured and unstructured data, originating from various sources. The extracted information may be used for prediction and planning or improving the quality of business or public services.Governments were among the first to consider data mining for discovering and preventing criminal and terrorist activities.The potential in the intelligent use of database and data warehouse records have attracted most public agencies into the field of data mining to improve their services.Common application areas which involve public services include the following(Two Crows Corporation,2005):

- Preventing abuse,fraud and waste in public finance,agriculture and environmental protection.
- Fighting crime,discovering and preventing terrorist activities.
- Detecting false claims,assessing the performance of healthcare,social security and insurance programs.
- Improving the e-government applications.
- Early detection of large scale security threats and improving national defense system.

Commercial,industrial and scientific data mining cover a wide range of application areas which,among the others, include banking ,marketing, e-commerce,credit scoring, increasing customer loyalty, entertainment,insurance,manufacturing,medicine and so on.The immediate availability of large volumes of data and unhindered use of them for these kind of applications make data mining more and more popular in practice.

Most data mining applications inevitably proceses data or information of personal nature.It is where the techniques and some results of data mining become ethically or even legally questionable.The privacy of an individual can be violated when specific information is obtained,manipulated and disseminated by other entities wihout his or her knowlwedge or consent(Olson,2007). Although data mining tools are poweful enough from a technical point of view,they still require domain expertise for a successful and harmless interpretation of the results. In the next section a discussion of the related issues is presented.

2. PRIVACY ETHICS AND DATA MINING

Absolute pricvacy is not possible to achieve in the information age.Because, individuls disseminate data in their common daily activities such as web browsing , e-commerce and e-government dealings, e-mail and mobile phone communications , credit card and ATM transactions etc.(Wel and Royakkers,2004).Major data sources include server and cookie logs,customer information,intelligent Internet agents and centralized demographic and other oficial records.Powerful data processing,storage and communication technologies allow these data to be manipulated and used by the other people and agencies freely and in most cases indiscreetely.The reason is that,data mining does not supply information about the social and ethical consequences of the results and also it does not necessarily discover causal relationships.Therefore,how and where to use the results is largely a choice of the "miner".This mechanism is inevitably prone to have negative impacts on privacy and individual rights.

In any data processing system,it is fair to expect that sensitive personal data must be protected from misuses and abuses by the outside entities.The position of data mining is even more critical in this respect, because of its power and potential.Naturally,every individual should have control over his or her personally sensitive data.But what constitutes "personal" or "sensitive" ? What are the limits? Of course these are questions whose answers may differ from individual to individual.Another important problem arises when individual's rights and public interest contradict.In most cases,public or business oriented data mining applications are claimed to produce beneficial outcomes for individuals,organizations and the society. Unfortunately,it is very difficult to impose universally acceptable gudelines and rules for distinguishing the right from the wrong.Because,data mining applications are generally open ended and they can as well lead to unpredictable and possibly harmful personal results.For example,very detailed data about buying habits,times and locations are extracted from customer transactions and data mined regularly by most supermarket chains.It is impossible for an average customer to be knowledgeable on possible uses of sensitive data and their consequences(Peace et al.,2002).

Data mining for crime prevention is based on the records of activities of the individuals, such as travel,electronic and phone communications,shopping and encounters with the other people. Data mining in these applications can produce inaccurate and faulty results, which usually constitute breaches of privacy and can be harmful for the individuals.For example,a person may be classified into a “suspect” short list while he or she has nothing to do with the affair at hand.The consequences could be serious in such a case and the whole process is certainly unethical if not illegal .The negative impacts such as litigation,adverse publicity,loss of reputation,discrimination could be further aggravated if the data are unreliable,faulty or even fabricated (Fule and Roddick,2004).This is where an inherently unethical process could also turn into an “illegal” one.

The legal issues related to data mining applications are complex and difficult to evaluate and as such,can not be put into the framework of any particular law. Besides,The challenges of the information age have not yet been resolved properly by legal doctrines and by the legal systems of countries (Seifert,2007). Most countries choose to strenghten the privacy of personal information by a specific law.But such a law cannot be expected to foresee all possible vialotions and types of offences that might come about.In practice,legal court cases are usually resolved by applying some other available code or legistlation which aim to protect the individual’s rights in general . The major problem here is the growth rate of information and information technologies ,which render obsolate the rules,regulations and even laws in relatively short time spans.In many situations, it is very difficult to decide what constitutes “legal” or “private” let alone to establish consistent court rules.The main reason is that,most physical limitations of the hardware and software technologies loose their meaning rapidly and newer and more versatile ones replace them.This phenomenon influences the application methodologies and environments, including data mining.Various privacy preservation methods and measures aim to find solutions for this kind of legal and ethical problems by reducing their likelihoods of occurrence in data mining applications.

3. EFFECTIVENESS OF THE METHODS FOR PRIVACY PRESERVATION

A number of information technology oriented and statistical methods have been developed to find a balanced solution to the dilemma between public interest and individual’s rights.These methods aim to protect personally “sensitive” data from abuses and misuses in data mining applications.Since privacy and even ethics are mostly personal matters,commonly applicable protective solutions are difficult to formulate.Another important problem is that,large volumes of data originating from heterogenous sources are generally inaccurate and open to errors. Therefore,both data mining results as well as the effectiveness of privacy preservation methods suffer from this untrustwothy nature of the mined data.

The following measures and techniques are commonly used in data mining applications to protect sensitive personal information and thus to improve ethical quality of the results (Agraval and Sirkant,2000):

- Anonymisation.Any identifying attributes are removed from the source dataset or rules using these attributes are suppressed.
- Perturbation and augmentation.The values of some attributes are slightly altered or new noise data are added to prevent identification.
- Controlled access to data by query restictions and prohibitation of unothorised use .
- Adding information about the degree of sensitivity of certain attributes and using this information in the data mining process.

These methods and similar ones all have their own disadvantages and,in general, are easy to bypass .On the other hand, The algorithms for automated privacy protection lack generality and are applicable only for specific cases.However,research in this area is still vivid.Another common problem is the reduced confidence levels of data mining results due to the applications of privacy preservation techniques .

In most data miming cases, the ethical quality can only be improved at the expense of mining quality(Vaidya and Clifton,2004).Restricting access to only authorised parties also does not preclude the possibility of inapprpriate uses by the "authorithy".Especially,naive authorized users can create severe ethical and legal privacy violations as happens in data mining for law enforcement .

Data Mining practices can be more harmful for the individuals when the results are used for "screening" rather than "surveillance".Because,in screening ,the main motive is to perform identification, which is contrary to the anonymization goal of the privacy protection methods .Since the data mining techniques always have some margin for error, false positives and false negatives cannot be avoided altogether . For this reason, counter crime data mining projects TIA and MATRIX have been discontinued in the USA (Seifert,2007).Despite that ,it is safe to assume that, in most countries data mining is being actively used for public security and law enforcement purposes and this kind of applications will continue to increase in size and coverage.Unfortunately,the legal,ethical and technical protection mechanisms are either nonexistent or too weak to prevent misclassifications and other defametory concequences of data mining applications.The situation is even worse in developing countries where,individual’s rights can be ignored more easily for public security, or because of insufficient expertise .

A possible mitigation could be achieved by supporting and supplementing data mining by other artificial intelligence and neural nework technologies.Namely,validation through supplementary technologies should have a higher priority in dealing with sensitive personal information.This way, potentially harmful decisions could only be formed using mutually strenghtening outcomes from multiple decision support mechanisms.This would help to increase further the overall protection levels in data mining.

4. CONCLUSIONS

Data mining technology has proven to be useful in most commercial and public service applications.However,in practice, the ethical and legal consequences are generally overlooked.Average customer or citizen remains still at the mercy of very sophisticated and partially intelligent data mining software.Privacy protecting techniques are still in their infancy and do not seem to be efffective enough to prevent potential misuses and abuses of private information.Legal protection is also not sufficient despite the introduction of some laws that aim to strenghten the privacy of personally sensitive data.It seems that,in a society that is becoming more and more complex and information technology dependent, the individual will have to learn beter to live with the “Big Brother”. The pervasive nature of information technolgies and the intricacies of the methods used in data mining leave no more option for anybody.

REFERENCES

Agrawal, R, and R. Srikant.(2000). “Privacy-preserving DataMining,” Proceedings of the ACM SIGMOD Conference, Dallas,TX, May 2000.

Fule,P. And Roddick F.J.,(2004).Detecting Privacy and Ethical Sensitivity in Data Mining Results.ACSC2004,Dunedin,New Zealand.

Olson,D.J.(2007) .Ethical Aspects of web Log Data Mining.Int.J.of Information Tech. and Management. 10;1-11.

Peace,A.G.,Weber,J.,Hartzel,S.,Nightingale,J.(2002) Ethical Issues in ebusiness:aprorosal for creating the ebusiness principles.Business and Socity Review,vol 107,1,41-60.

Seifert,J,W. (2007) .Data Mining and Homeland Security:An Overview. CRS Report for Congress. Congressional Research Service.

Two Crows Corporation(2005) Introduction to Data Mining and Knowledge Discovery,Third Edition,available at : http://www.twocrows.com.

Vaidya,J.andClifton,C.,(2004).Privacy-preserving data mining: why, how, and when? Security & Privacy,IEEE vol 3,6.19-27.

Wel,L.van,and Royakkers,L. (2004) . Ethical Issues in Web Data Mining.Ethics and Information Technology,6; 129-140.