PRIVACY-PRESERVING DATA MINING

INITIAL PLAN

Jasmin Beckford

Supervisor: Dr Jianhua Shao

Moderator: Padraig Corcoran

CM3203

ONE SEMESTER INDIVIDUAL PROJECT

40 Credits

Project Description

Nowadays, the amount of personal data is ever increasing as companies and organisations store more and more data regarding their consumers and individuals. With the rise of data that can be transferred and shared between organisations, the question is posed as to what extent is the privacy protected of those individuals contained within a dataset?

The healthcare industry, in particular, collects and processes vast amounts of patient data. The law in the UK states that data within the healthcare system used for 'secondary purposes' must not identify individual patients.[1] It could be considered that removing attributes such as the patient's name or NHS number is sufficient to provide anonymity and therefore protect the privacy of the patient's data. However, this is not necessarily the case as the volume and availability of personal data nowadays can be exploited in order to carry out attacks on data.

Multiple datasets containing data about a particular individual can be collated in order to cross-reference attributes and positively identify an individual in a dataset which may have had attributes removed such as the individual's name. Privacy-preserving data mining aims to tackle this and puts forward a number of different techniques for transforming data in order to protect the privacy of the individuals concerned.

However, privacy-preserving data mining also produces the challenge of information loss. Through the transformation of the data in order to preserve its privacy, its granularity can sometimes be greatly reduced. As a result of this, any subsequent data processing or data mining algorithms may be a lot less useful and therefore may not give as much meaningful information as the original data would. This is the natural trade-off between preserving the privacy of data and experiencing information loss and throughout this project it is this relationship that will be explored. [2]

This project will focus on existing privacy-preserving techniques and will investigate their effectiveness with regards to the performance of subsequent data mining processes. Through comparing data mining results of an original dataset and the same dataset in a privacy-preserved form, the extent to which privacy preservation affects data mining will be scrutinised and will address the question of whether the amount of information lost outweighs the ethical principles of privacy.

Project Aims and Objectives

The primary aim and objective of this project is as follows:

AIM	INVESTIGATE THE EFFECTIVENESS OF A DATA MINING ALGORITHM ON PRIVACY-PRESERVED DATA.
OBJECTIVE	Using a decision tree algorithm, compare the accuracy of classifying unseen test cases in an original dataset and the same dataset in a privacy-preserved form.

This can be broken down into a number of different aims and objectives which are shown below.

AIM: EXPLORE EXISTING PRIVACY-PRESERVING TECHNIQUES

- Become familiar with and gain knowledge of techniques and methods that are currently used to preserve the privacy of data. This will include an understanding of how they operate and their strengths and limitations.

AIM: IMPLEMENT A PRIVACY-PRESERVING ALGORITHM

- Through the use of a language such as Python or Java, implement a chosen privacy-preserving technique which will be used to specify a dataset as input and produce the same dataset that adheres to the principles of the chosen privacy-preserving technique as output.

AIM: RUN A DATA MINING ALGORITHM ON THE PRIVACY-PRESERVED DATA

- After the privacy-preserving techniques have been applied to the data, it will not be able to be inputted into a decision tree algorithm in the usual way due to the grouping of data. Therefore, an algorithm needs to be produced that can handle this new data. This can be done in two ways:

    1. Implementing a complete algorithm that is designed to take privacy-preserved data as input.
    2. Modifying an existing open-source algorithm to ensure it is able to handle privacy-preserved data as input.

Work Plan

Below is the work plan that I will adhere to in order to carry out this project. In addition to this, I will be carrying out the following tasks on an ongoing basis:

- Weekly meetings with my supervisor to discuss progress and any struggles
- Continued development of knowledge of privacy-preserving data mining throughout the course of the project

I have dedicated the last four weeks of the project to writing up any remaining sections of the report and for finalisation and submission. Moreover, I have a designated contingency week to allow for any previous tasks earlier in the timeline which may have needed an extra week to complete. This ensures that even if I do have to allow additional time for a task, I will not fall behind in my overall project timeline.

WEEK	COMMENCING	TASKS	MILESTONES
1	23RD JANUARY 2017	▪ Conduct background reading on privacy-preserving data mining as an introduction to the topic ▪ Develop the initial plan	**DELIVERABLE:** SUBMIT INITIAL PLAN – MONDAY 30TH JANUARY 2017
2	30TH JANUARY 2017	▪ Continue background reading on the topic ▪ Research existing privacy-preserving techniques such as k-anonymity and Mondrian.	
3	6TH FEBRUARY 2017	▪ Become familiar with the chosen dataset and its attributes. ▪ Explore how privacy-preserving algorithms can be coded.	
4	13TH FEBRUARY 2017	▪ Decide on a privacy-preserving technique. ▪ Design a privacy-preserving algorithm for this technique.	
5	20TH FEBRUARY 2017	▪ Implement the privacy-preserving algorithm in a chosen language.	**REVIEW MEETING 1:** CHECK THE ALGORITHM DESIGN.

6	27TH FEBRUARY 2017	▪ Test the privacy-preserving algorithm on datasets.	MILESTONE: PRIVACY-PRESERVING ALGORITHM IS IMPLEMENTED.
7	6TH MARCH 2017	▪ Explore how data mining algorithms can be manipulated to handle privacy-preserved data. ▪ Consider the implementation of a decision tree algorithm from scratch.	
8	13TH MARCH 2017	▪ Implement either a method of manipulating existing data mining algorithms or the creation of a decision tree algorithm from scratch. ▪ Test that the algorithm can input the privacy-preserved data as expected.	MILESTONE: DATA MINING ALGORITHM IS IMPLEMENTED AND WORKS AS EXPECTED.
9	20TH MARCH 2017	▪ Run the data mining algorithm on both the original and privacy-preserved datasets and record the results. ▪ Begin evaluation.	REVIEW MEETING 2: CHECK IF THE RESULTS ARE AS EXPECTED.
10	27TH MARCH 2017	▪ Continue evaluation of the obtained results.	
11	3RD APRIL 2017	▪ Write up final stages of report	
12	10TH APRIL 2017	▪ Write up final stages of report	
13	17TH APRIL 2017	▪ Write up final stages of report	
14	24TH APRIL 2017	▪ Proofread and finalise report	MILESTONE: FINAL REPORT COMPLETE.
15	1ST MAY 2017	CONTINGENCY WEEK IF NEEDED.	DELIVERABLE: SUBMIT FINAL REPORT – FRIDAY 5TH MAY 2017

DELIVERABLES

The deliverables for this project will be:

1. An initial project plan
2. A final report including the approach, implementation, results and evaluation of the project.
3. A privacy-preserving algorithm (to be included in the final report)
4. A decision tree algorithm (to be included in the final report)

References

[1] https://www.england.nhs.uk/ourwork/tsd/ig/

[2] Aggarwal, Charu C and Philip S Yu. *Privacy-Preserving Data Mining*. 1st ed. New York: Springer, 2008. Print.