Privacy Preservation in Data Mining Through Noise Addition

Md Zahidul Islam

A thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

School of Electrical Engineering and Computer Science
University of Newcastle
Callaghan
New South Wales 2308
Australia

November 2007

Certificate of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree at any other University or Institution.

(Signed) ______________________________

Md Zahidul Islam

Acknowledgements

I would like to thank my supervisor A/Prof. Ljiljana Brankovic who is something more than just a supervisor to all her students. Whenever I was in a trouble she was there with her genuine suggestions and dependable directions. She introduced this research area to me. If I have learnt anything on how to do research then it is due to her wise supervision. She always led us to be independent researchers having high ethics and moral values.

I would also like to thank my co-supervisor Professor A.S.M. Sajeev for his support, encouragement and wisdom. I am also grateful to Dr Regina Berretta, Dr Michael Hannaford, Dr Alexandre Mendes, Professor Mahbub Hassan, Professor M. F. Rahman, Professor Mirka Miller and Professor Elizabeth Chang for their support and encouragement.

I would give my special thanks to my friends Helen, Mousa, Mouris and Tanya for their enormous moral support throughout my study. My thanks also to all Faculty and Staff members of the School of Electrical Engineering and Computer Science and all postgraduate students of the school during my study for being so kind and friendly to me.

Last but not least, I would like to thank my wife Moonmoon for her patience, care, support, trust, love and encouragement. My special thanks to my children Abdellah, Saifullah and Mahir for their love and support. I would like to thank my parents Harun and Nilu, my father in law, mother in law, sister and brother in law for their encouragement. They have been a very supportive family all the way.

This thesis is gratefully dedicated to

My Family:

Moonmoon, my wife

Abdellah, **Saifullah** and **Mahir**, my sons

My Parents, **My Sister**, **My Father in law**, **My Mother in law**

and **All Relatives**

for their patience, their unwavering support and their faith.

Say: "If the ocean were ink (wherewith to write out) the words of my Lord. Sooner would the ocean be exhausted than would the words of my Lord, even if we added another ocean like it, for its aid." (Qur'an 18:109)

List of publications arising from this thesis

1. M. Z. Islam, and L. Brankovic, Privacy Preserving Data Mining: A Framework for Noise Addition to all Numerical and Categorical Attributes, In *Data Mining and Knowledge Discovery.* (In Preparation)

2. M. Z. Islam, and L. Brankovic, Privacy Preserving Data Mining: Noise Addition to Categorical Values Using a Novel Clustering Technique, In *IEEE Transactions on Industrial Informatics*, 2007. (Submitted on the 3rd September, 2007)

3. L. Brankovic, M. Z. Islam and H. Giggins, Privacy-Preserving Data Mining, Security, Privacy and Trust in Modern Data Management, Springer, Editors Milan Petkovic and Willem Jonker, ISBN: 978-3-540-69860-9, Chapter 11, 151-166, 2007.

4. M. Z. Islam, and L. Brankovic, DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique in Privacy Preserving Data Mining, In *Proc. of the 3rd International IEEE Conference on Industrial Informatics*, Perth, Australia, (2005).

5. M. Z. Islam, and L. Brankovic, A Framework for Privacy Preserving Classification in Data Mining, In *Proc. of Australian Workshop on Data Mining and Web Intelligence (DMWI2004)*, Dunedin, New Zealand, CRPIT, **32**, J. Hogan, P. Montague, M. Purvis and C. Steketee, Eds., *Australian Computer Science Communications*, (2004) 163-168.

6. M. Z. Islam, P. M. Barnaghi and L. Brankovic, Measuring Data Quality: Predictive Accuracy vs. Similarity of Decision Trees, In *Proc. of the 6th International Conference on Computer & Information Technology (ICCIT 2003)*, Dhaka, Bangladesh, Vol.**2**, (2003) 457-462.

7. M. Z. Islam, and L. Brankovic, Noise Addition for Protecting Privacy in Data Mining, In *Proc. of the 6th Engineering Mathematics and Applications Conference (EMAC*

2003), Sydney, Australia, (2003) 457-462.

List of other publications during the candidature

1. M. Alfalayleh, L. Brankovic, H. Giggins, and M. Z. Islam, Towards the Graceful Tree Conjecture: A Survey, In *Proc. of the 15th Australasian Workshop on Combinatorial Algorithms (AWOCA 2004)*, Ballina, Australia, (2004).

Abstract

Due to advances in information processing technology and storage capacity, nowadays huge amount of data is being collected for various data analyses. Data mining techniques, such as classification, are often applied on these data to extract hidden information. During the whole process of data mining the data get exposed to several parties and such an exposure potentially leads to breaches of individual privacy.

This thesis presents a comprehensive noise addition technique for protecting individual privacy in a data set used for classification, while maintaining the data quality. We add noise to all attributes, both numerical and categorical, and both to class and non-class, in such a way so that the original patterns are preserved in a perturbed data set. Our technique is also capable of incorporating previously proposed noise addition techniques that maintain the statistical parameters of the data set, including correlations among attributes. Thus the perturbed data set may be used not only for classification but also for statistical analysis.

Our proposal has two main advantages. Firstly, as also suggested by our experimental results the perturbed data set maintains the same or very similar patterns as the original data set, as well as the correlations among attributes. While there are some noise addition techniques that maintain the statistical parameters of the data set, to the best of our knowledge this is the first comprehensive technique that preserves the patterns and thus removes the so called Data Mining Bias from the perturbed data set.

Secondly, re-identification of the original records directly depends on the amount of noise added, and in general can be made arbitrarily hard, while still preserving the original patterns in the data set. The only exception to this is the case when an intruder knows enough about the record to learn the confidential class value by applying the classifier. However, this is always possible, even when the original record has not been used in the training data set. In other words, providing that enough noise is added, our technique makes the records from the training set as safe as any other previously unseen records of the same kind.

In addition to the above contribution, this thesis also explores the suitability of prediction accuracy as a sole indicator of data quality, and proposes technique for clustering both categorical values and records containing such values.

Contents

List of Figures

List of Tables