TRANSACTIONS ON DATA PRIVACY 4 (2011) 127–128

Foreword for the special issue of selected papers from the 1st ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning

Aris Gkoulalas-Divanis*, **Yücel Saygin****, **Vassilios S. Verykios*****

*Information Analytics, IBM Research-Zurich, Switzerland.

**Faculty of Engineering and Natural Sciences, Sabanci University, Turkey.

***Hellenic Open University, Greece.

E-mail: agd@zurich.ibm.com, ysaygin@sabanciuniv.edu, verykios@eap.gr

The first Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML 2010) was organized on September 24, 2010 at Barcelona, Spain, in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). Privacy and security-related aspects of data mining and machine learning have been the topic of active research during the last decade due to the existence of numerous applications with privacy and/or security requirements. Privacy issues have become a serious concern due to the increasing collection and sharing of personal data for purposes like data publishing or data mining. This has led to the development of privacy-preserving data mining and machine learning methods. More general security considerations arise in applications such as biometric authentication, intrusion detection and malware classification. This has led to the development of adversarial learning algorithms, while parallel work in multi-agent settings and in low regret learning algorithms has revealed interesting interplays between learning and game theory. Although significant research has so far been conducted, we are still far from addressing the numerous theoretical and practical challenges that arise in this domain. Firstly, several emerging research areas in data analysis, decision making and machine learning, require new theoretical and applied techniques for the offering of privacy or security. Secondly, there is an urgent need for learning and mining methods with privacy and security guarantees. Thirdly, there is an emerging demand for security applications such as biometric authentication and malware detection. In all cases, the strong interconnections between data mining and machine learning, cryptography and game theory, create the need for the development of multidisciplinary approaches on adversarial learning and mining problems.

The aim of the PSDML workshop was to bring together scientists and practitioners who conduct research on privacy and security issues in data mining and machine learning to discuss the most recent advances in these research areas, identify open problem domains and propose possible solutions. The workshop was organized along four core subjects: (a) data privacy and security issues, (b) theoretical aspects of machine learning for secu-

rity applications, (c) privacy-preserving data mining, machine learning and applications, and (d) security applications of machine learning. The present special issue contains three extended papers that have been selected among the papers presented at PSDML 2010 focusing mainly on the privacy aspects of data mining and machine learning.

The first paper by Manas A. Pathak and Bhiksha Raj is titled "Efficient Protocols for Principal Eigenvector Computation over Private Data". It presents an efficient secure multiparty computation protocol for computing the principal eigenvector of a collection of data matrices that belong to semi-honest parties, coordinated by a semi-honest arbitrator. The protocol is augmented with randomization, data padding and oblivious transfer to conceal the information which the parties can learn from the intermediate results. The authors provide an analysis of correctness, security and efficiency of the protocol together with experimental results from a prototype implementation.

The second paper by Henrik Grosskreutz, Benedikt Lemmen and Stefan Rüping is titled "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data". The paper studies the problem of subgroup discovery and presents new secure protocols which support distributed subgroup discovery on horizontally partitioned data. The authors analyze the properties of their top-1 subgroup discovery protocol and prove that it leaks only little information, namely the size of the database and the share of positive records. They also report experimental results which demonstrate the feasibility of the approach by using a prototype implementation of the protocol.

The third paper by Gérald Gavin, Julien Velcin and Philippe Aubertin is titled "Privacy Preserving Aggregation of Secret Classifiers". This work considers a number of parties, each having its own dataset, who build a private classifier for predicting a binary class variable. The authors develop protocols which allow combining these private classifiers in a privacy-preserving way in order to improve individual predictions. The aggregation is achieved through a secure computation of linear combinations (weighed votes) over the private classifiers. The proposed protocols are shown to be correct and private against any active polynomial adversary.

We believe that the selected papers are a good representative of the research in privacy aspects of data mining and machine learning. We hope that you will enjoy reading them.

Acknowledgements

We would like to thank Prof. Vicenç Torra and Prof. Josep Domingo-Ferrer, Editors in Chief of Transactions on Data Privacy for their strong support on the organization of this special issue. We also want to thank the reviewers of the papers for their helpful comments.

Aris Gkoulalas-Divanis
IBM Research-Zurich
agd@zurich.ibm.com

Yücel Saygin
Sabanci University
ysaygin@sabanciuniv.edu

Vassilios S. Verykios
Hellenic Open University
verykios@eap.gr