Journal of Ambient Intelligence and Humanized Computing (2023) 14:14977–14978
https://doi.org/10.1007/s12652-023-04699-2

EDITORIAL

Special issue on data mining and data privacy

Vicenç Torra[1,2] · **Yasuo Narukawa**[3]

Published online: 19 September 2023

Abstract
This special issue encompasses fourteen papers that focus on different aspects related to data science. This ranges from supervised and unsupervised machine learning algorithms to extract knowledge from data, to data protection procedures to ensure that the knowledge extracted is privacy-preserving.

1 Preface

Data science is the science of data and its goal is to explain processes and objects trough available data. According to (Said and Torra 2019) its goal is "to make informed decisions based on the knowledge extracted from the underlying data". Machine learning and data mining provide tools to extract this knowledge. However, data can be sensitive. Privacy protection mechanisms are required to avoid the disclosure of confidential information. Neither data releases, nor data-driven ML models, nor decisions made from these models should release sensitive information.

In this special issue we present several papers that focus on different aspects of data mining and data privacy.

The first papers are related to tools and methodologies to extract knowledge from data.

The first paper by Endo et al. is on clustering. The authors propose a variation of k-means algorithm and establish its relation to a weighted alpha complex. The variation is also objective-based and its goal is that the clusters have a topological structure with filtration.

✉ Vicenç Torra
vtorra@ieee.org

Yasuo Narukawa
nrkwy@eng.tamagawa.ac.jp

[1] Hamilton Institute, Maynooth University, Maynooth, Ireland

[2] Umeå University, Umeå, Sweden

[3] Tamagawa University, Tokyo, Japan

The second paper by Nguyen et al. is also on k-means. In the paper Huynh et al. study the problem of applying k-means to categorical data. The authors consider an information-theoretic definition of similarity for objects and a kernel-based representation for clusters.

The third paper by Armengol is related to building classifiers. The paper presents an approach to find patterns based on the minimization of the distance between two indistinguishability relations (Recasens 2011).

The fourth paper by Nin and Tomás propose a model to study the details of default propagation in customer-supplier networks. Authors' model is a a variation of the Microscopic Markov-Chain Approach (MMCA) model.

The fifth paper by Steinhauer et al. propose the use of topic modelling tools in a new setting. The authors develop the use of topic modelling for anomaly detection in telecommunication networks.

The remaining papers in the issue have a strong focus on topics related to data privacy.

The sixth paper by Kaaniche et al. focuses on assessment opinion polls scenarios. The authors propose an anonymous certification scheme to ensure that each user cannot hand in more than one poll.

The seventh paper by Chen et al. is on predictable opportunistic networks where end-to-end connectivity is not guaranteed. These networks can be modelled as dynamic graphs (Jain et al. 2004). Solutions to provide anonymity for messages in this type of networks are proposed.

The eight paper by Stokes introduces a new model graph database model. One of the goals of the model is to be a privacy-preserving mechanism. Parallels are drawn

to concepts in combinatorics (e.g., clique complexes, flag geometries, and graph covers).

The ninth paper by Inuiguchi et al. proposes a data anonymization method. The authors consider decision rule induction based on rough sets, and how privacy-preserving solutions can be found in this setting.

The tenth paper by Ito et al. is also about data anonymization. In this case, the authors focus on attacker models to de-identified data, and different models are considered based on different types/amounts of background knowledge.

The eleventh paper by Salas introduces an anonymization approach for large sparse data sets, the type of data that is used to build recommender systems. The anonymization approach is to obtain a protected data set that is compliant with k-anonymity.

The twelfth paper by Nuñez-del-prado and Nin studies online anonymization. The authors want to provide a stream processing algorithm to anonymize a data set in real time. In this case the scenario relates to location privacy.

The thirteenth paper by Casas focuses on graphs. The author considers the problem of producing a privacy-preserving graph and analyse the vertex and edge modification techniques found in the literature.

The fourteenth paper by Senavirathne et al. is also about data anonymization. In this case the authors focus on rounding as a way to coarsen continuous data, and, thus, a way to achieve privacy-preserving databases for numerical data.

References

Jain S, Fall K, Patra R (2004) Routing in a delay tolerant network. In: Proc of SIGCOMM '04, p 145–158

Recasens J (2011) Indistinguishability operators. Modelling Fuzzy Equalities and Fuzzy Equivalence Relations, Springer

Said A, Torra V (2019) Data science in practice. Springer