KYBERNETIKA — VOLUME *45* (2009), NUMBER 3, PAGES 548-560

On the Comparison of Some Fuzzy Clustering Methods for Privacy Preserving Data Mining: Towards the Development of Specific Information Loss Measures

Vicenç Torra; Yasunori Endo; Sadaaki Miyamoto

Abstract: Policy makers and researchers require raw data collected from agencies and companies for their analysis. Nevertheless, any transmission of data to third parties should satisfy some privacy requirements in order to avoid the disclosure of sensitive information.

The areas of privacy preserving data mining and statistical disclosure control develop mechanisms for ensuring data privacy. Masking methods are one of such mechanisms. With them, third parties can do computations with a limited risk of disclosure.

Disclosure risk and information loss measures have been developed in order to evaluate in which extent data is protected and in which extent data is perturbated. Most of the information loss measures currently existing in the literature are general purpose ones (i. e., not oriented to a particular application). In this work we develop cluster specific information loss measures (for fuzzy clustering). For this purpose we study how to compare the results of fuzzy clustering. I. e., how to compare fuzzy clusters.

Keywords: privacy preserving data mining; statistical disclosure control; fuzzy clustering; fuzzy c-means; fuzzy c-means with tolerance;

AMS Subject Classification: 68T05; 68T37; 68T99;

References

[1] J. C. Bezdek: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York 1981.

[2] CASC: Computational Aspects of Statistical Confidentiality, EU Project, http://neon.vb.cbs.nl/casc/ (Test Sets)

[3] J. Domingo-Ferrer and V. Torra: Disclosure control methods and information loss for microdata. In: Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies (P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. M. Zayatz, eds.), Elsevier 2001, pp. 91–110,

[4] J. Domingo-Ferrer and V. Torra: A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure, and

Data Access: Theory and Practical Applications for Statistical Agencies (P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. M. Zayatz, eds.), Elsevier 2001, pp. 111–133.

[5] G. Duncan, S. Fienberg, R. Krishnam, R. Padman, and S. Roehrig: Disclosure limitation methods and information loss for tabular data. In: Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies (P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. M. Zayatz, eds.), Elsevier 2001, pp. 135–166.

[6] G. Duncan, S. Keller-McNulty, and S. Stokes: Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report No. 121 of National Institute of Statistical Sciences 2001, www.niss.org.

[7] G. Duncan, S. Keller-McNulty, and S. Stokes: Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility Through the R-U Confidentiality Map. Technical Report No. 142 of National Institute of Statistical Sciences 2004, www.niss.org.

[8] Y. Hasegawa, Y. Endo, Y. Hamasuna, and S. Miyamoto: Fuzzy *c*-means for data with tolerance defined as hyper-rectangle. In: Proc. MDAI 2007 (Lecture Notes in Artificial Intelligence 4617), pp. 237-248.

[9] J. Lane, P. Heus, and T. Mulcahy: Data access in a cyber world: Making use of cyberinfrastructure. Trans. Data Privacy 1 (2008), 2–16.

[10] P. Medrano-Gracia, J. Pont-Tuset, J. Nin, and V. Muntés-Mulero: Ordered data set vectorization for linear regression on data privacy. In: Proc. MDAI 2007 (Lecture Notes in Artificial Intelligence 4617), Springer, Berlin 2007, pp. 361–372.

[11] S. Miyamoto and K. Umayahara: Methods in gard and fuzzy clustering. In: Soft Computing and Human-Centered Machines (Z.-Q. Liu and S. Miyamoto, eds.), Springer, Tokyo 2000, 85–129.

[12] S. Mukherjee, Z. Chen, and A. Gangopadhyay: A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms. The VLDB Journal 15 (2006), 293–315.

[13] R. Murata, Y. Endo, H. Haruyama, and S. Miyamoto: On fuzzy *c*-means for data with tolerance. J. Advanced Computational Intelligence and Intelligent Informatics 10 (2006), 5, 673–681.

[14] J. Nin, J. Herranz, and V. Torra: Rethinking rank swapping to decrease disclosure risk. Data and Knowledge Engrg. 64 (2008), 1, 346–364.

[15] A. Oganian and J. Domingo-Ferrer: On the complexity of optimal microaggregation for statistical disclosure control. Statistical J. United Nations Economic Commission for Europe 18 (2000), 4, 345–354.

[16] V. Torra and J. Domingo-Ferrer: Record linkage methods for multidatabase data mining. In: Information Fusion in Data Mining (V. Torra, ed.), Springer 2003, pp. 101–132.

[17] V. Torra and J. Nin: (2008) Record linkage for database integration using fuzzy integrals. Internat. J. Intel. Systems 23 (2008), 715–734.

[18] M. Trottini: Decision Models for Data Disclosure Limitation. Ph.D. Dissertation, Carnegie Mellon University 2003, http://www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

[19] W. E. Yancey, W. E. Winkler, and R. H. Creecy: Disclosure risk assessment in perturbative microdata protection. In: Inference Control in Statistical Databases 2002 (Lecture Notes in Computer Science 2316), Springer, Berlin 2003, pp. 135–152.

[20] A. C. Yao: Protocols for secure computations. In: Proc. 23rd IEEE Symposium on Foundations of Computer Science, Chicago 1982, pp. 160–164.