Enhancement of Privacy-Preserving k-means*

Adrián Csiszárik[a], András Lukács[a]

[a]Inter-University Centre for Telecommunications and Informatics,
Eötvös Loránd University
{csadrian,lukacs}@cs.elte.hu

Abstract

A common method to transform a data mining algorithm into its privacy-preserving variant is to make it work in the distributed setting and replace particular components of the algorithm with secure multi-party computation (SMC) primitives. The efficiency of the resulting algorithm depends on the computational and communication costs come from the distributed nature of the data and the SMC primitives. In this work, we present improvements of the distributed privacy-preserving k-means algorithm based on the modification of calculations used to find the nearest centroid. In order to reduce the communication costs and to improve the speed of the calculations, our first approach is omitting unnecessary branches of the calculations in the distance calculation steps. Another method is to replace the current centroid to a not necessarily nearest but at least nearer centroid. In the latter case more steps of iterations are needed. To analyse the efficiency of the above methods we present a number of measurements on synthetic and real data sets.

Keywords: data mining, clustering, k-means, privacy

MSC: 62H30, 68M14

References

[1] Vaidya, J. and Clifton, C. Privacy-Preserving K-Means Clustering over Vertically Partitioned Data, *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003), 206–215.

[2] Elkan, C. Using the Triangle Inequality to Accelerate k-Means, *ICML* (2003), 147–153.

*This work was supported by the TAMOP 4.2.2.C-11/1/KONV-2012-0001 project.