Distributed K-means over Compressed Binary Data

Elsa Dupraz[1]

[1] *Telecom Bretagne; UMR CNRS 6285 Lab-STICC Brest, France*

Keywords

Distributed source coding, K-means algorithm, Learning over compressed data, Sensor networks, LDPC codes

Summary

Networks of sensors have long been employed in various domains such as environmental monitoring, electrical energy management, and medicine [1]. In particular, inexpensive binary-valued sensors are successfully used in a wide range of applications, such as activity recognition on home environments [2]. In this paper, we consider a network of J binary-valued sensors that transmit their data to a fusion center. We assume that the fusion center has to perform K-means clustering on the binary data transmitted by the sensors.

In this context, the J sensors should send their measurements to the fusion center in a compressed form in order to greatly reduce the amount of data transmitted within the network. Low Density Parity Check (LDPC) codes have been shown to be very efficient for distributed compression in a network of sensors [3]. However, the standard distributed compression framework considers that the fusion center has to reconstruct all the measurements from all the sensors. Here, in order to avoid potentially complex decoding operations, we would like to perform K-means directly over the compressed data. For this problem, [4] considered real-valued measurement vectors compressed from Compressed Sensing (CS) techniques, and showed that applying the K-means algorithm in the compressed domain enables to recover the clusters of the original domain. However, K-means over compressed binary data was not considered in [4], nor in references therein.

In this work, we consider binary measurement vectors and we assume that the compression is realized from LDPC codes. We propose a formulation of the K-means algorithm that applies over binary data in the compressed domain. As for the original K-means, our clustering algorithm is divided into two steps, namely the cluster assignment step and the centroid estimation step.

We then carry a theoretical analysis of the performance of the proposed K-means algorithm in the compressed domain. We in particular derive analytical approximated error probabilities of each of the two steps of the K-means algorithm. In order to verify the accuracy of the analysis, we compare the obtained analytical expressions to the error probabilities measured from Monte Carlo simulations, see Figure 1. We observe that although the analytical expressions are only approximations, they predict accurately the error probabilities of the two steps of the algorithm. The theoretical analysis hence permits to verify that applying K-means in the compressed domain enables to recover the clusters of the original domain. It also serves to design the parameters of the LDPC codes that are used in the system.

At the end, we show from Monte Carlo simulations that the effective rate needed to perform K-means over compressed data is lower than the theoretical rate (the joint entropy of all the sensors measurements) that would be needed to reconstruct all the sensors measurements.

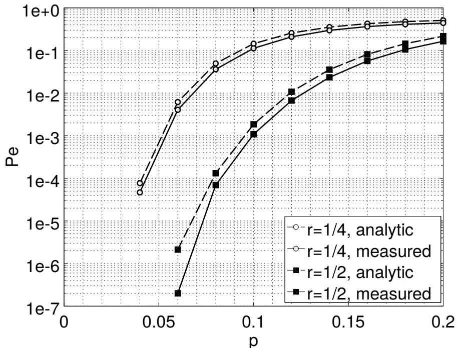

Figure 1: Approximated error probability of the centroid estimation step and error probability measured from Monte Carlo simulations, with respect to the binary symmetric source parameter p. The value r is the coding rate.

References

[1] J. Yick, B. Mukherjee, and D. Ghosal. Wireless sensor network survey. *Computer networks*, 52(12):2292–2330, 2008.

[2] F.J. Ordóñez, P. de Toledo, and A. Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, 13(5):5460–5477, 2013.

[3] Z. Xiong, A. Liveris, and S. Cheng. Distributed source coding for sensor networks. *IEEE Signal Processing. Magazine*, 21(5):80–94, 2004.

[4] C. Boutsidis, A. Zouzias, M.W. Mahoney, and P. Drineas. Randomized dimensionality reduction for K-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, 2015.