Asynchronous Distributed Clustering in Mesh Networks

Centre for Data Analytics

Insight

Student: Cheng Qiao, Supervisor: Professor Kenneth N. Brown

Motivation

Data is generated simultaneously by many different sensors or agents. How should the community learn a global picture, without the cost (privacy, energy, time) of transmitting all the raw data?

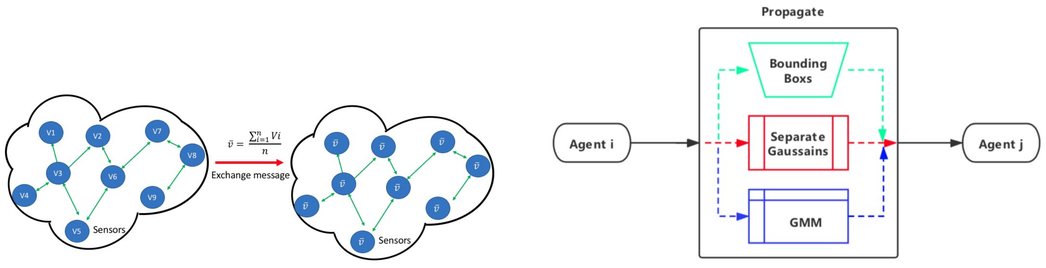

Figure: Problem definition and proposed information propagation

Background

The state-of-art technology in this category is proposed by Datta [1], Fatta [2], Benezit [3] and Bendechache [4].

- Datta et al. [1] proposed a synchronous distributed k-means algorithm, exchanging centroids and counts each round.
- Fatta et al. [2] and Benezit et al. [3] offered a similar approach, but using gossip to exchange information.
- Bendechache et al. [4] suggested that represent the cluster by boundary points in a tree-based network.

The previous algorithms synchronise the behaviour of the sensors [1, 2, 3] and ignore the communication cost [1, 2, 3, 4].

Proposed approach

Each agent clusters its own data, and then announces its centroids, counts and distribution, using: 1),kmeans, and nested bounding boxes, 2), kmeans, and gaussians, or 3), a single Gaussian mixture model (refer to figure 1). Two scenarios are considered:

- The number of agents is known. Agents simply relay new descriptions, until one agent receives info from all others.
- Agents only know their neighbours. Agents generate new clusters after each new information, and re-broadcasts.

To assimilate new information, an agent samples from the received descriptions to get new data, and then clusters again. Agents account for repeated information by sampling and model subtraction, issuing requests for reduced models if needed. All communication is asynchronous.

Name	Measures	Mean	S.D
MMF1-P	Time	8.6	3.09
	Accuracy	98.42 [72.5 100.0 99.7]	4.02
	R-msg	20.89 [11.5,25.68, 23.43]	9.12
MMF1-SG	Time	2.24	0.1
	Accuracy	98.46 [72.6 99.9 99.8]	5.33
	R-msg	33.85 [25.5, 40.53, 33.34]	8.69
MMF1-GMM	Time	1.68	0.08
	Accuracy	94.29 [70.7 99.7 97.15]	8.15
	R-msg	36.51 [27.80, 43.89, 36.0]	7.26
[1]	Time	12.23	6.71
	Accuracy	87.1 [66.1 99.6 90.45]	10.53
	R-msg	42.22 [27.5, 52.65, 42.29]	26.54
	R-poll	45.91 [31.73, 57.61, 44.73]	26.32
[2]	Time	61.61	26.52
	Accuracy	89.21 [20.7, 100.0, 98.7]	15.83
	R-msg	104.27 [96.84, 110.63, 104.73]	3.67
[3]	Time	20.56	19.68
	Accuracy	78.9 [33.1, 99.9, 81.05]	16.35
	R-msg	139.67 [127.72, 148.56, 139.81]	5.42

Name	Measures	Mean	S.D
MMF1-P	Time	8.72	1.85
	Accuracy	98.73 [72.9 99.9 99.7]	3.82
	R-msg	21.54[21.52, 17.78, 26.57]	3.47
MMF1-SG	Time	3.01	0.13
	Accuracy	96.92 [72.7 99.9 99.8]	8.08
	R-msg	29.74 [17.07, 46.15, 28.96]	2.52
MMF1-GMM	Time	2.57	0.16
	Accuracy	96.23 [73.0 99.7 98.65]	6.20
	R-msg	30.94[18.02, 48.82, 30.25]	2.56
[1]	Time	16.62	6.38
	Accuracy	88.56 [60.5 99.9 92.05]	10.58
	R-msg	29.59[16.03, 50.92, 28.37]	13.13
	R-poll	31.33[17.42,51.84,30.50]	12.59
[2]	Time	42.38	15.59
	Accuracy	81.21 [19.40 100.0 81.4]	15.48
	R-msg	87.56 [77.32,96.56,88.13]	6.41
[3]	Time	71.68	45.36
	Accuracy	78.43 [41.6 100.0 79.95]	14.01
	R-msg	204.16 [180.54, 224.47, 203.82]	12.47

Figure: Comparison of scenario 1 on dense and sparse network

Our proposed algorithm is:

- More accurate. Scores higher than [1, 2, 3].
- Faster. Way more better than [1, 2, 3].

Detecting sub-patterns

The methods above assume that all agents are receiving data from the same distribution, and so there is only one pattern. But in many applications, there might be sub-groups of agents that are receiving different patterns of data, and this must be identified. Existing work on anomaly detection assumes the anomalies are rare. How should we detect general sub-patterns?

Our approach

The agent that does the final clustering is responsible for determining what sub-patterns exist. For problems where there are two or more agent-patterns, we attempt to cluster the agents from the individual agent descriptions.

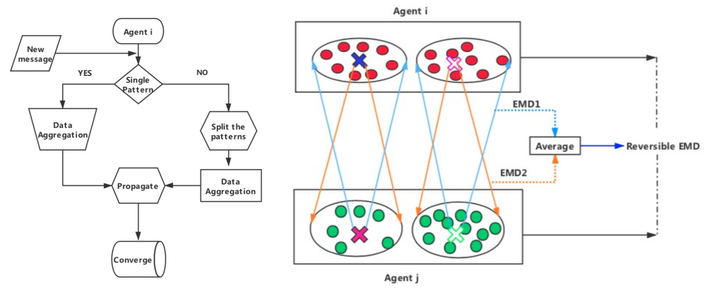

Figure: Flow chart for agent and reversible EMD we proposed

We considered Earth move distance (EMD) between points, Weighted EMD between centroids, reversible EMD between centroids and points (refer to the figure above), wavelet EMD and robust EMD. And accuracy in putting points in the right clusters, and which agent are in which pattern are shown.

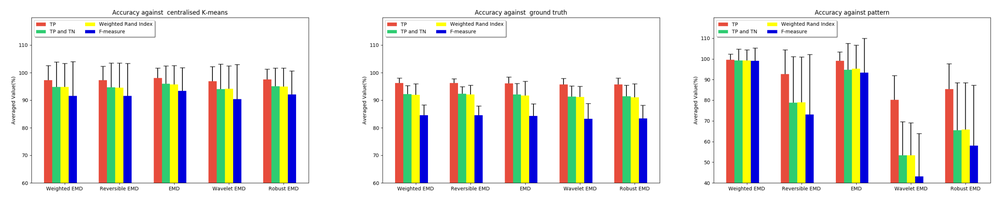

Figure: Comparison of various methods on regular multi-dimensional Gaussian dataset

The above methods all assume that there is more than one pattern. Before we can apply them, we need to decide whether or not multiple patterns exist. We considered G-means, Kernel density estimation (KDE), KDE with optimised bandwidth, DBSCAN, AIC and shift EMD.

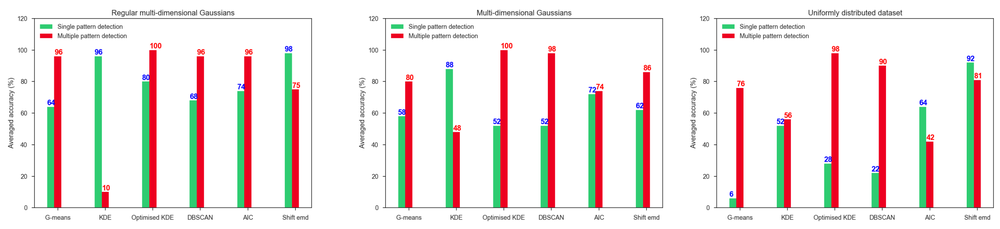

Figure: Comparison of various methods to detect patterns

Conclusion

- Kernel Density Estimation with optimised bandwidth outperforms other methods to detect the patterns except the underlying datasets are uniformly distributed. The reason lies in the fact that the clustering algorithm do not fit well with the underlying datasets.
- Weighted EMD between centroids outperforms other methods in putting the right agent into the right pattern. In accuracy against centralised k-means and ground truth, there are little difference among all methods.

Reference

S.Datta, C.Giannella, and H.Kargupta. newblock Approximate distributed k-means clustering over a peer-to-peer network. IEEE Transactions on Knowledge and Data Engineering, 21(10), 2009.

G.D. Fatta, F. Blasa, S.Cafiero, and G.Fortino. Epidemic k-means clustering. pages 151–158, 2011.

Benezit F, V. Blondel, P. Thiran, J. Tsitsiklis and M. Vetterli. Weighted gossip: Distributed averaging using non-doubly stochastic matrices, 2010 IEEE Int Symposium Information Theory, Austin, TX, 2010, 1753-1757.

Bendechache M, Le-Khac N A, Kechadi M T. Hierarchical aggregation approach for distributed clustering of spatial datasets. IEEE 16th ICMD Workshops. 2016: 1098-1103.

Publication

Cheng Qiao and Kenneth N Brown, Asynchronous Distributed Clustering Algorithm for Wireless Sensor Networks, International Conference on Machine Learning Technologies (ICMLT), 21st-23rd June, 2019 (Accepted).

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289

DCU · NUI Galway OÉ Gaillimh · UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh · UCD Dublin · sfi Science Foundation Ireland