iMass: An Approximate Adaptive Clustering Algorithm for Dynamic Data Using Probability Based Dissimilarity

Panthadeep Bhattacharjee, Pinaki Mitra
Dept. of CSE, IIT Guwahati

Supplementary Material

1 Evaluation of the cluster quality

We provide three masures: Normalized Mutual Information(NMI), Rand-Index and F1-score for judging the cluster quality. For the five class labeled data, we adopted the aforementioned metrics while for remaining three unlabeled data, a cluster accuracy percentage is depicted.

Table 1: Clusters quality evaluation metrics

Dataset	***#Classes***	***iMass***			**MBSCAN**		
		NMI score	**Rand- index**	**F1-score**	**NMI score**	**Rand- index**	**F1-score**
Libras	15	**0.272**	**0.80438**	**0.10896**	0.272	0.80438	0.10896
Segment	7	**1.0**	**0.17848**	**0.30290**	1.0	0.17848	0.30290
Wine	3	**1.0**	0.38488	0.55583	0	**0.52105**	**0.68512**
Seeds	3	**1.0**	0.38684	0.55787	1.0	**0.38743**	**0.55848**
Iris	3	**1.0**	0.38743	**0.55848**	0.81497	**0.54598**	0.42210

For the other three datasets: Aggregation, S1 and S2, a mean cluster accuracy of 66.67% were observed.

1.1 Normalized Mutual Information(NMI):

Normalied Mutual Information[1] provides the reduction in entropy of class labels given that the custer labels are already known. It is an external measure because we need the class labels of the instances to determine the NMI. Since NMI is normalized, itenables us to measure and compare the NMI values between different clusterings. The following formula gives the NMI measure for a

given algorithm.

$$NMI = \frac{2 * I(Y;C)}{[H(Y) + H(C)]} \tag{1}$$

where Y is the number of class labels. C represents the cluater labels, $H(.)$ is the entropy. $I(Y;C)$ is given by the follwing relation:

$$I(Y;C) = H(Y) - H(Y|C) \tag{2}$$

$H(Y|C)$ represents the entropy of class labels within each cluster.

For most of the class labeled datasets in Table 1, we observe that $iMass$ either retains or has a better NMI value than that of MBSCAN algorithm. However for datasets: Wine and Seeds dataset, MBSCAN has higher Rand-index and F1-score as compared to $iMass$.

1.2 Rand-Index:

Rand-Index measures the percentage of correct decisions. Its calucaltion is based on the evaluation of TP(True Positive), FP(False Positive), True Negative(TN), False Negative(FN). A TP decision allocates two similar items within same cluster. A TN puts two diffeternt items in different clusters. FP allocates two dissimilar objects to same cluster while FN assigns two similar items to dissimilar clusters. The Rand-Index is given by:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

1.3 F-measure:

The F-measure pemalizes FN more than FP contary to the Rand-Index measure. For our cluster evaluation purpose, we measure the F1-score given by:

$$Precision(p) = \frac{TP}{TP + FP} \qquad Recall(r) = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2pr}{p + r} \tag{5}$$

2 Some related works:

Apart from the related works of DBSCAN and MBSCAN algorithm as cited in the main paper, we also present a brief idea about relevant works on mass-based dissimilarity: mp dissimilarity[2] approach. This work investigates about the ways of exploiting the data distribution for finding dissimilarity between two data items. Instead of depending on geometric distance, it computes the proximity between two objects in each dimension as a probability mass in a region enclosing the objects. The dissimilarity of a pair of data objects can be computed in $\mathcal{O}(d \log n)$ time where d represents the numberof dimensions. Also

the algorithm[2] involves $\mathcal{O}(dn)$ time to construct the binary tree. In comparison $iMass$ constructs $iForest$ on $\mathcal{O}(t \log \Psi)$ time with mass-matrix being built in $\mathcal{O}(n^2)$ time. Here n, t, Ψ represents the dataset size, number of $iTrees$ and the subsample size per $iTree$.

In another study related to image retreival [3], a novel dissimilarity measurement technique was proposed which can calculate both the distance and perceptual similarity of two images in multi-dimensional space. It combines the properties of both m_p[2] and l_p[2] dissimilarity having a $\mathcal{O}(rd)$ time with d being the number of dimension and r is the number of points in a given dimension used for finding the mass of any pair(x, y).

References

[1] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[2] Sunil Aryal, Kai Ming Ting, Gholamreza Haffari, and Takashi Washio. mp-dissimilarity: A data dependent dissimilarity measure. In *2014 IEEE International Conference on Data Mining*, pages 707–712. IEEE, 2014.

[3] Hamid Shojanazeri, Dengsheng Zhang, Shyh Wei Teng, Sunil Aryal, and Guojun Lu. A novel perceptual dissimilarity measure for image retrieval. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2018.