Repulsive Affinity-Based Clustering

Laurens van der Maaten
TiCC, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
lvdmaaten@gmail.com

Clustering is a fundamental problem in machine learning that has applications in a wide variety of domains. It is concerned with finding groups of objects that are closely related under some dissimilarity measure. Most current clustering techniques suffer from a fundamental problem: they do not consider the data density between a point and its corresponding cluster center in the evaluation of the quality of a clustering. Instead, they only consider the dissimilarity between each point and its corresponding cluster center. As a result, they may fail to identify clusters that correspond to a wide mode in the data distribution, because in these clusters, points in the periphery of the mode are relatively far away from the center of the mode. This *density problem* hampers the performance of, e.g., k-means clustering and affinity propagation [1].

The density problem may be resolved by using (dis)similarities based on diffusion distances [3], as diffusion distances integrate over all paths in a neighborhood graph on the data. However, diffusion distances are computationally expensive to compute, and require the appropriate selection of parameters such as the number of steps in the random walks. Mixture models resolve the density problem by assuming a specific form of the clusters, which allows mixture models to deal with, e.g., elongated clusters. However, as a result of the assumption on the shape of the clusters, mixture models cannot successfully deal with the large variety in clusters shapes that characterizes real-world data.

We present work in progress on a new clustering technique called Repulsive Affinity-Based Clustering (RAC) that attempts to address the density problem. The objective of RAC is to learn a collection of mixing proportions $\pi_i^{(c)}$ (with $\sum_c \pi_i^{(c)} = 1$) in such a way that $\pi_i^{(c)}$ measures the extent to which datapoint i is a member of cluster c. RAC first computes joint probabilities p_{ij} that measure the similarity between datapoints i and j. A standard approach to compute these probabilities is to use the stochastic neighborhood measure [2]. Under this measure, the joint probability p_{ij} is proportional to the density under a Gaussian that is centered on datapoint i

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2/2\sigma^2)}. \tag{1}$$

As the similarities of the input data are represented in terms of the stochastic neighbor measure, it seems sensible to define a similar stochastic neighbor measure under the clustering model as well. Assume we have a cluster c in which datapoints i and j have mixing proportions $\pi_i^{(c)}$ and $\pi_j^{(c)}$. Assuming the mixing proportions are independent (i.e., assuming the data is iid), the probability of randomly picking i and j from cluster c is proportional to $\pi_i^{(c)}\pi_j^{(c)}$. Marginalizing out the clusters c, we obtain the stochastic neighbor measure q_{ij} under the clustering model

$$q_{ij} = \frac{\sum_c \pi_i^{(c)}\pi_j^{(c)}}{\sum_{k \neq l} \sum_{c'} \pi_k^{(c')}\pi_l^{(c')}}. \tag{2}$$

In order to make sure that $\pi_i^{(c)} \in [0, 1]$ and that $\sum_c \pi_i^{(c)} = 1$, we represent the mixing proportions $\pi_i^{(c)}$ in terms of mixing weights $w_i^{(c)}$ as follows

$$\pi_i^{(c)} = \frac{\exp(-w_i^{(c)})}{\sum_{c'} \exp(-w_i^{(c')})}. \tag{3}$$

Maximum likelihood learning of the parameters $w_i^{(c)}$ amounts to minimizing the Kullback-Leibler divergence $KL(P||Q)$. Although ML learning of the parameters may seem like a sensible idea, it suffers from a weakness that is due to the asymmetry of the KL divergence. Specifically, the KL divergence is hardly increased if two dissimilar datapoints i and j are given high mixing proportions in the same cluster, because their pairwise similarity p_{ij} is infinitesimal. In this case, the only contribution to the KL divergence comes from wasting some probability mass in the Q-distribution.

Hence, it is much better to learn the mixing proportions $\pi_i^{(c)}$ by minimizing the *inverse* KL divergence $KL(Q||P)$.

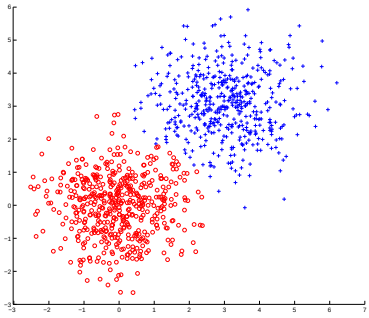
IRON	THINK	OBVIOUS	COMMON	SPORTS	MEMBER	DIRECTION	CRITICIZE
EXERCISE	BRAIN	UNCLEAR	NORM	BASKETBALL	GROUP	AGENDA	DEGRADE
FITNESS	PHILOSOPHY	UNKNOWN	ABNORMAL	BASEBALL	CLUB	NORTH	HATE
WEIGHTS	THOUGHT	VAGUE	UNUSUAL	ATHLETIC	GANG	LIST	INSULT
WORKOUT	MIND	CLEAR	STRANGE	PRO	GATHERING	EAST	DISOWN
WEIGHT	IDEA	UNSURE	REGULAR	FOOTBALL	PEOPLE	SOUTH	RIDICULE
FAT	OPINION	OBSCURE	USUAL	ARENA	PARTY	WEST	PUT DOWN
LIFT	LOGIC	UNDECIDED	DIFFERENT	ATHLETE	ORGANIZATION	SCHEDULE	HUMILIATE
THIN	THINKING	FOGGY	ROUTINE	REFEREE	COMMITTEE	ORDER	CRITICISM
GYM	REASON	KNOWN	NORMAL	JOCK	SOCIETY	ORGANIZE	EMBARRASS
HEAVY	DECISION	SURE	IRREGULAR	STADIUM	SOCIAL	DIRECTIONS	SHAME
SLIM	MEMORY	DOUBT	STANDARD	GAMES	TEAM	CRITERION	COMPLIMENT
SKINNY	CONCENTRATE	UNSEEN	ODD	GYMNAST	MEETING	COMPASS	CRUEL
DIET	ABSTRACT	VIVID	ORDINARY	ACTIVE	FRATERNITY	RULES	DISGRACE
LEAN	PONDER	INDIRECT	SAME	COMPETE	CREW	HEADING	PRAISE

Table 1: Most probable words of eight randomly selected 'topics'.

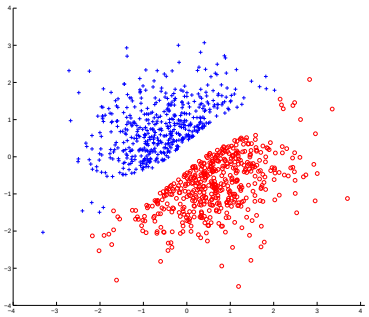
In contrast to ML learning, $KL(Q||P)$ strongly penalizes modeling a small p_{ij} by a large q_{ij}. The cost function thus has a strong repulsive nature: it is mainly concerned with making sure that dissimilar datapoints do not get high mixing proportions in the same cluster. In preliminary experiments, we found the repulsive cost function $C = KL(Q||P)$ to significantly outperform maximum likelihood learning.

Figure 1 shows the results of applying RAC on three artificial datasets (using $k = 2$). The results reveal RAC can successfully identify clusters with a variety of shapes, as it does not make assumption on the shape of a cluster. We also performed experiments on a dataset that contains word association data for $5,019$ words [4], which can readily be used as input into RAC. In Table 1, we present the results of an experiment in which we used RAC to cluster the word association data using $k = 100$ clusters. The table shows the 15 words with the highest probability in a 'topic' for 8 randomly selected topics.

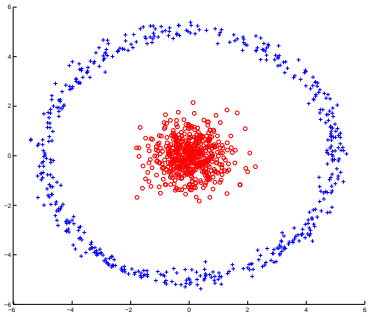
RAC attempts to resolve the density problem by considering all pairwise similarities between all points. Another interesting characteristic is its repulsive nature: in contrast to most other clustering techniques, it attempts to model dissimilar datapoints in different clusters.



(a) Gaussian data.



(b) Sin-sep data.



(c) Concentric data.

Figure 1: Result of applying RAC on three artificial datasets.

References

[1] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[2] G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, Cambridge, MA, 2002. The MIT Press.

[3] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[4] D.L. Nelson, C.L. McEvoy, and T.A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms, 1998.