2F1-3in

Semi-supervised Evolutionary Distance Metric Learning for Clustering

Wasin Kalintha[*1] 福井 健一[*2] Ken-ichi Fukui 小野 智司[*3] Satoshi Ono 女鹿野 大志[*3] Taishi Megano 森山 甲一[*2] Koichi Moriyama 沼尾 正行[*2] Masayuki Numao

Graduate School of Information Science and Technology, Osaka University[*1]
The Institute of Scientific and Industrial Research, Osaka University[*2]
Graduate School of Science and Engineering, Kagoshima University[*3]

Exising method for supervised clustering called Evolutionary Distance Metric Learning (EDML) has never been compared to other clustering method. This work conducted experiments to compare EDML with other semi-supervised clusterings, such as COP-Kmeans and other DML methods. The result empirically confirms that EDML gives better clustering structure than the candidate clustering methods-i.e. K-means, COP-Kmeans, and MPC-Kmeans. Also, we justify the effect of the number of constraints, effect of smoothing, and the feasibility to evaluate EDML in various criteria. Therefore, EDML is assured that it has potential to improve clustering quality and is capable of using various clustering indices.

1. Introduction

Many methods have been proposed for clustering problems. Despite the inspiration vary, their aspiration is to group similar instances together in the same cluster and vice versa[Yin 12]. However, there are no "right"answers for clustering[Xing 02]. The performance of clustering is critically determined by definition of similarity between the data points. The similarity can be not only the Euclidean distance between instances, but it can also be the Mahalanobis distance which satisfies the axiom of distance. Generally, clustering algorithms are unsupervised learning. On the other hand, in real application, some background knowledge are coincidentally provided. The distance metric learning (DML)[Yang 06] attempts to optimize a metric to improve clustering or classification by taking advantage of these given knowledge to transform the data space and stretch the partitions.

In contrast to the conventional semi-supervised clustering methods[Wagstaff 01, Bilenko 04], Fukui et al. have proposed Evolutionary Distance Metric Learning (EDML)[Fukui 13] that optimizes any cluster validity index such as Purity, F-measure, or Entropy, depending on the clustering purpose. Furthermore, by using smoothed cluster validity indices[Fukui 12] to consider neighbor relations in the data space, EDML successfully avoids the problem of over-fitting. Since an objective function based on cluster validity is massively multimodal in changing of the distance metric, EDML uses an evolutionary algorithm of Self-Adaptive Differential Evolution (jDE)[Brest 06], which can deal with multimodality without manual adjustment of its control parameters. Also, EDML is compatible to any clustering algorithms.

In this paper, we apply EDML for clustering in four experiments: (1) compare EDML with unsupervised and semi-supervised conventional clustering algorithms, (2) study the effect of number of labeled data, (3) study the effect of neighborhood smoothing in clustering index and (4) evaluate EDML with assorted criteria.

Contact: Kalintha Wasin, 〒567-0047 大阪府茨木市美穂ヶ丘 8-1, 06-6879-8426, and wasin@ai.sanken.osaka-u.ac.jp

2. Related work

In this section, we introduce two of typical semi-supervised clustering algorithms related to EDML.

2.1 Constrained K-means Clustering with Background Knowledge: COP-Kmeans

Wagstaff et al. proposed COP-Kmeans[Wagstaff 01] that is a modification of K-means by using background knowledge that can be expressed as a set of pairwise constraints— i.e. Must-link constraints indicate that two instances have to place in the same cluster, and Cannot-link constraints indicate that two instances must not be in the same cluster. It proceeds as K-means which ensures that none of specified constraints are violated.

2.2 Metric pairwise constrained K-means: MPC-Kmeans

MPC-Kmeans[Bilenko 04] was proposed by Bilenko et al. This clustering algorithm is a combination of constraint-based clustering of COP-Kmeans which allows constraint violation if it leads to a more cohesive clustering, and a distance metric learning methods.

3. Evolutionary Distance Metric Learning

3.1 Global distance metric learning

A Mahalanobis-based distance is used as many global DML methods. Given a dataset $\mathcal{D} = \{\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,v})^t \in \mathbb{R}^v\}_{i=1}^N$, the Mahalanobis-based distance

can be defined as:

$$d^2_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)^t \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j), \quad (1)$$

where $\mathbf{M} = (m_{k,l})$ is a $v \times v$ matrix. The elements of $\mathbf{M}$ $(m_{k,l})$ are variables to be learned that represent a transformation of the input data, in this case, $\mathbf{M}$ must be a symmetric positive semi-definite matrix to satisfy the distance propositions. Specifically, when only diagonal element in $\mathbf{M}$ (where $k = l$) is used, we denote as "EDML-D".

EDML approach optimizes a clustering index *Eval* as follows:

$$\text{Maximize } Eval(Clustering(d^2_{i,j})), \quad (2)$$

where $Clustering(d^2_{i,j})$ denotes a clustering result by using a distance metric $d^2_{i,j}$ such as Entropy, F-measure, and Purity.

3.2 EDML framework

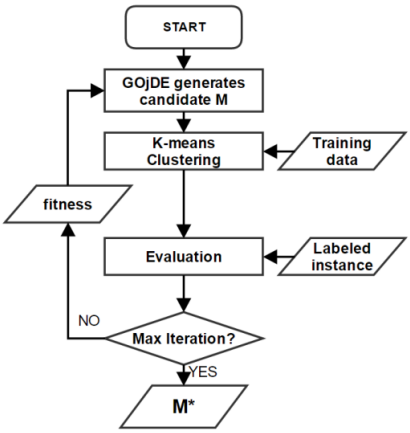

Figure 1: Diagram of the evolutionary distance metric learning (EDML) framework

The EDML framework is summarized in Fig. 1[Fukui 13]. First, the objective function in eq. (2) is optimized by using an evolutionary algorithm. Differential Evolution with self-adapting control parameters and generalized opposition-based learning(GOjDE)[Wang 13], one of evolutionary algorithm, is employed to generate candidates of the metric transform matrix $\mathbf{M}$ candidates. Next, obtained $\mathbf{M}$ is manipulated to transform the data space via eq. (1). Then, cluster structure is archived by K-means clustering. After that, the quality of cluster structure is evaluated with weighted Pairwise F-measure (which is introduced in the next section) as Eval() in eq. (2). Then, feed this evaluation value back into GOjDE as the fitness for new candidates of $\mathbf{M}$. GOjDE selects candidates on the basis of the fitness to evolve and generates the next candidates by mutation and crossover with certain probabilities. These steps are repeated until the limit iteration. Finally, we archieve the best metric transform matrix $\mathbf{M}^*$ in terms of the smoothed clustering index among the overall generations of candidates.

3.3 Evaluation criteria

The clustering result is evaluated using the clustering index. Because of pairwise constraints, we adopt the extension of one of pairwise-based cluster validity index call weighted Pairwise F-measure(wPFM)[Fukui 12]. wPFM is defined as followings.

	t(i) = t(j)	t(i) ≠ t(j)
c(i) = c(j)	a	b
c(i) ≠ c(j)	c	-

Table 1: Class and cluster confusion matrix of data pairs

Originally, Pairwise F-measure(PFM) is a harmonic average of the precision, which is a measure of the same class among each cluster, and the recall, which is a measure of the same cluster among each class. Whereas, wPFM is based on a degree that the data pairs belong to the same cluster.

Given $c(k)$ and $t(k)$ denoting the cluster/class assignment for the instance x_k. Fukui et al. proposed $likelihood(c(i) = c(j))$ indicating a degree that data pair x_i, x_j belongs to the same class instead of the actual number of data pairs. The $likelihood(c(i) = c(j)$ or $h_{c(i),c(j)}$ is given by the inter-cluster distance of data pair. Therefore, each value in PFM's class and cluster confusion matrix of data pairs show in Table 1 is replaced by summation of likelihoods as follows:

$$a' = \sum_{\{i,j|t(i)=t(j)\}} h_{c(i),c(j)}, \quad (3)$$

$$b' = \sum_{\{i,j|t(i)\neq t(j)\}} h_{c(i),c(j)}, \quad (4)$$

$$c' = \sum_{\{i,j|t(i)=t(j)\}} (1 - h_{c(i),c(j)}) = a + c - a'. \quad (5)$$

With these extended a', b', and c', extended Precission (P') and Recall(R') are defined as follows:

$$P' = \frac{a'}{a' + b'}, R' = \frac{a'}{a' + c'}. \quad (6)$$

Finally, wPFM which is a harmonic average of extended precision and recall, is derived as follows:

$$\text{wPFM} = \frac{2P'R'}{P' + R'}. \quad (7)$$

4. Experiment

In this section, we conduct experiments to show the comparison between the EDML and the conventional semi-supervised clustering algorithms.

4.1 Methodology

We compared EDML with the conventional K-means (KMN), COP-Kmeans, and MPC-Kmeans. K-means is an unsupervised clustering algorithm and regarded as the baseline for comparison here. Since COP-Kmeans and MPC-Kmeans are constraint-based clustering, we manipulate all labeled instances to produce pairwise constraints to carry out sufficient comparison.

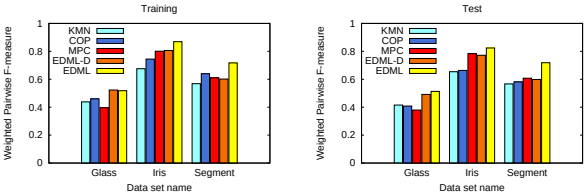

Figure 2: Cross validation results (Glass, Iris, and Segment dataset)

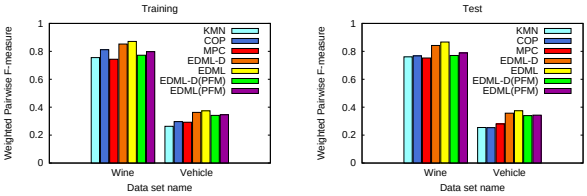

Figure 3: Cross validation results (Wine and Vehicle dataset)

4.2 Data sets

This experiment is conducted on five data sets from UCI repository*1: Glass, Iris, and Wine as small data sets, and Segment and Vehicle as relatively large data. The attribute values are normalized with mean is equal to zero and standard deviation is equal to 1 calculated from only training data. The properties of all data sets are sumarized in Table 2.

Data set	#Class	#Instance	#Attribute
Glass	6	214	9
Iris	3	150	4
Wine	3	178	13
Segment	7	2310	19
Vehicle	4	846	18

Table 2: Basic properties of the data sets

4.3 Experimental setup

We validated performance by a 10-fold cross validation with labeled sample rate $Ls_{rate} = 0.3$. Particularly, randomly selected 30% of all the instances were set as labeled instances. The numbers of clusters were set according to data set size-i.e., 20 and 50 to small and large data set respectively. For each data set, we performed a 10-fold cross validation on each algorithm for 25 times, which come from 5 times random initial centroids multiplied by 5 times random selection of labeled instances. Here, cluster centroids and distance metric learning matrix(M) are also obtained, these results were carried out in test data evaluation process. DML was used to transform the data space and continually cluster assignment using cluster centroids which obtained by cluster analysis over training data.

4.4 Comparision result

Fig. 2 and Fig. 3 show the cross validation results on all data set. Here, EDML-D and EDML can easily performed better than all other clustering method; however, MPC-Kmeans barely overcomes EDML-D.

4.5 Effect of the number of labeled data

Next, we checked varieties numbers of Ls_{rate} to observe a transition of the average fitness in terms of wPFM. The Ls_{rate} we used are 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1.0.

Fig. 4 shows transitions of the average of fitness over twentyfive trials with random initial values and constraints when changing labeled sample rate Ls_{rate}. Here, when $Ls_{rate} = 0.05$ or 5%, labeled samples are only 8 and 38 in Wine and Vehicle, respectively. Even in extremely small number of labeled instances conditions such as Wine data set, EDML could obtained roughly 0.78 in wPFM both training and test data. The evaluation values increase approximately over 10% in both data set-i.e. 13%, 16% in Wine, and 18%, 21% in Vehicle using EDML-D and EDML, respectively. Moreover, both EDML-D and EDML fitness are abruptly increase when Ls_{rate} is less than 0.3 and gradually increase when above 30%. Lastly, EDML-D and EDML tends to archive roughly the same wPFM in small number of labeled instance ($Ls_{rate} < 0.2$), this shows that when we have less than 20% of known instances, EDML-D can take part in EDML to reduce complexity.

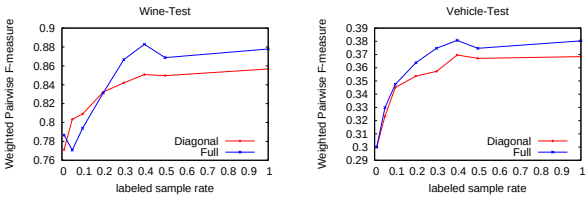

Figure 4: Transitions of the clustering index with different numbers of labeled instances

4.6 Effect of smoothing

The following presents the effect of smoothing in clustering index, which is specified by the likelihood function, is the comparison between the EDML that was trained with PFM and wPFM.

Fig. 3 shows the effect of smoothing in the clustering index, which is specified by the likelihood function in eq. (7). EDML without smoothing (EDML-PFM) provides lower result comparing to MPC-Kmeans, because MPC-Kmeans uses unlabeled instance to steer the clustering result when evaluating with clustering index without smoothing.

4.7 Evaluation by various criteria

Lastly, for the fair evaluation and to show robustness of EDML in various criteria, we employed different clustering indices[Fukui 12] to evaluate the clustering result such as Entropy(ENT), Purity(Purity), F-meassure(FME), Pairwise F-meassure(PFM), and Pairwise Accuracy (PAC) both original index and with neighborhood-based smoothing(represented with prefix "w"-i.e. wENT, wPUR, wFME,

*1 http://www.ics.uci.edu/∼mlearn/MLRepository.html

wPFM, and wPAC).

Table 3 shows the cluster indices score in comparison of EDML and other clustering algorithm. When EDML is obtained better clustering index for less than 1%, 1%-5% or over than 5% , it is denoted by 0+, + and ++ and vice versa. In second row, D and F denote the EDML-D and EDML.

From these tables we can obtain insights as follows:

- Obviously, when we evaluate the clustering result with wPFM, EDML provides the highest result in every comparison, because EDML employs wPFM as the objective function. We also obtain the best result in every comparison when evaluating with wFME, since wPFM and wFME have similar calculation process.

- Since EDML attempts to transform the data space to make instances with the same class close together, it usually provides higher evaluation score when it is evaluated by clustering index with smoothing even in other criteia.

	KMN		COP		MPC	
	D	F	D	F	D	F
ENT	0+	+	0+	+	0-	0-
FME	++	++	++	++	-	-
PAC	0-	0-	0-	0-	0-	0+
PFM	++	++	++	++	0-	0-
PUR	–	–	–	–	-	0+
wENT	++	++	++	++	0-	-
wFME	+	++	++	++	+	++
wPAC	–	–	–	–	–	-
wPFM	++	++	++	++	++	++
wPUR	–	–	–	–	–	-

Table 3: Comparison of various cluster indices on Vehicle data set

5. Conclusion

In this paper, we evaluate EDML by comparing it with the conventional clustering algorithms: unsupervised clustering: K-means, semi-supervised clustering both constraint-based method: COP-Kmeans and distance-function method: MPC-Kmeans. The comparison results empirically showed that EDML is better than all other method by evaluate cluster structure with clustering index. Thus, EDML has a potential to improve clustering quality. Not only that, we illustrate the transition of clustering index. The higher number of labeled instance, the higher clustering index and EDML also successfully avoids the problem of overfitting. Furthermore, only 30% of labeled instance are enough to performs the EDML. Then we showed the effect of smoothing that EDML performed better when smoothing the clustering index. Last but not least, EDML is satisfy that it is capable of various clustering indices.

Acknowledgement

This work was partially supported by the Kayamori Foundation of Informational Science Advancement.

References

[Bilenko 04] Bilenko, M., Basu, S., and Mooney, R. J.: Integrating Constraints and Metric Learning in Semi-supervised Clustering, in *Proceedings of the 21st International Conference on Machine Learning*, pp. 81–88, ACM (2004)

[Brest 06] Brest, J., Greinero, S., Boskovic, B., Mernik, M., and Zumer, V.: Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems, *IEEE Transactions on Evolutionary Computation*, Vol. 10, No. 6, pp. 646–657 (2006)

[Fukui 12] Fukui, K. and Numao, M.: Neighborhood-based Smoothing of External Cluster Validity Measures, in *Proc. the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-12)*, pp. 354–365 (2012)

[Fukui 13] Fukui, K., Ono, S., Megano, T., and Numao, M.: Evolutionary Distance Metric Learning Approach to Semi-supervised Clustering with Neighbor Relations, in *2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 398–403 (2013)

[Wagstaff 01] Wagstaff, K., Cardie, C., Rogers, S., and Schrdl, S.: Constrained K-means Clustering with Background Knowledge., in *ICML*, pp. 577–584 (2001)

[Wang 13] Wang, H., Rahnamayan, S., and Wu, Z.: Parallel differential evolution with self-adapting control parameters and generalized opposition-based learning for solving high-dimensional optimization problems, *Journal of Parallel and Distributed Computing*, Vol. 73, pp. 62–73 (2013)

[Xing 02] Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J.: Distance Metric Learning with Application to Clustering with Side-Information, in *Advances in Neural Information Processing Systems (NIPS)*, pp. 505–512 (2002)

[Yang 06] Yang, L.: Distance Metric Learning : A Comprehensive Survey, Technical Report 16, Michigan State Universiy (2006)

[Yin 12] Yin, X., Shu, T., and Huang, Q.: Semi-supervised fuzzy clustering with metric learning and entropy regularization, *Knowledge-Based Systems*, Vol. 35, pp. 304 – 311 (2012)