Two Phase Semi-supervised Clustering Using Background Knowledge

Kwangcheol Shin and Ajith Abraham

School of Computer Science and Engineering, Chung-Ang University
221, Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea
kcshin@archi.cse.cau.ac.kr, ajith.abraham@ieee.org

Abstract. Using background knowledge in clustering, called semi-clustering, is one of the actively researched areas in data mining. In this paper, we illustrate how to use background knowledge related to a domain more efficiently. For a given data, the number of classes is investigated by using the must-link constraints before clustering and these must-link data are assigned to the corresponding classes. When the clustering algorithm is applied, we make use of the cannot-link constraints for assignment. The proposed clustering approach improves the result of COP k-means by about 10%.

1 Introduction

In data mining, clustering is an unsupervised method to classify unlabeled data. Unsupervised means that it does not have any prior knowledge of the problem domain to formulate the number of clusters (classes). However, in some real situations, background knowledge that might help the clustering problem is available. One of representative back ground knowledge is the *must-link* and *cannot-link* constraints. Wagstaff and et al. [1] illustrated that their modified *k*-means algorithm (COP k-means), gives better results than the original k-means by using the *must-link* and *cannot-link* constraints.

Adami et al. [2] proposed a baseline approach that classifies documents according to the class terms, and two clustering approaches, whose training is constrained by the a priori knowledge encoded in the taxonomy structure, which consists of both terminological and relational aspects.

Shen et al. [3] proposed a novel approach, the so-called "supervised fuzzy clustering approach" that is featured by utilizing the class label information during the training process. Based on such an approach, a set of "*if-then*" fuzzy rules for predicting the protein structural classes are extracted from a training dataset. It has been demonstrated through two different working datasets that the overall success prediction rates obtained by the supervised fuzzy clustering approach are all higher than those by the unsupervised fuzzy c-means.

Zio and Baraldi [4] studied the Mahalanobis metric for each cluster for analyzing the complexity and variety of cluster shapes and dimensions. The a priori known information regarding the true classes to which the patterns belong is exploited to select, by means of a supervised evolutionary algorithm, the different optimal Mahalanobis metrics. Further, the authors illustrated that the diagonal elements of the

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 707 – 712, 2006.

matrices defining the metrics can be taken as measures of the relevance of the features employed for the classification of the different patterns.

Eick et al. [6] introduced a novel approach to learn distance functions that maximizes the clustering of objects belonging to the same class. Objects belonging to a dataset are clustered with respect to a given distance function and the local class density information of each cluster is then used by a weight adjustment heuristic to modify the distance function so that the class density is increased in the attribute space. This process of interleaving clustering with distance function modification is repeated until a "good" distance function has been found. We implemented our approach using the k-means clustering algorithm.

Some recent research [6] sought to address a variant of the conventional clustering problem called semi-supervised clustering, which performs clustering in the presence of some background knowledge or supervisory information expressed as pairwise similarity or dissimilarity constraints. However, existing metric learning methods for semi-supervised clustering mostly perform global metric learning through a linear transformation. Chang and Yeung [6] proposed a new metric learning method that performs nonlinear transformation globally but linear transformation locally.

In this paper, we present a novel algorithm which uses background knowledge more efficiently. At first, before the clustering process, graphs are constructed by using must-link constraints to find out how many classes do the dataset has and we make use of cannot-link constraints when the k-means clustering algorithm is applied. The proposed method could adaptively determine the k value empirical results illustrate about 10% of improvement over the COP k-means algorithm on some popular datasets.

2 Proposed Algorithm

We make use of basic background knowledge, must-link and cannot-link constraints. The algorithm is divided into two parts. The first part is for finding number of classes by using must-link constraints and assigns must-link data to the corresponding class and the other part is for applying the k-means algorithm effectively using cannot-link constraints.

2.1 The Constraints

Must-link and cannot link constraints are defined as follows [1]:

- Must-link constraints specify that two instances have to be in the same cluster
- Cannot-link constraints specify that two instances must not be placed in the same cluster

Two instances within the constraints are randomly selected. The number of constraints is important to apply our algorithm.

2.2 Phase I : Find Number of Class by Using Must-link Constraints

In the first part of our algorithm, namely phase I, we make use of the must-link constraints to find the number of classes of the given dataset based on the following

assumption. If we have enough must-link constraints to get at least one of must-link constraints in each class, then we can find out the number of class.

Two instances (a, b) are in must-link if they should be assigned to a same class. This is illustrated as an undirected graph in Figure 1.

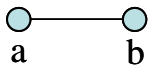

Fig. 1. Must-link graph

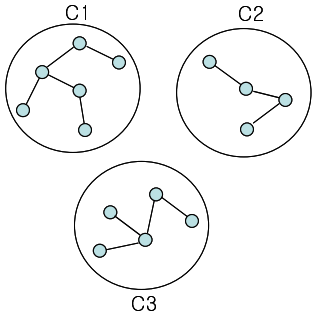

Fig. 2. Must-link graphs

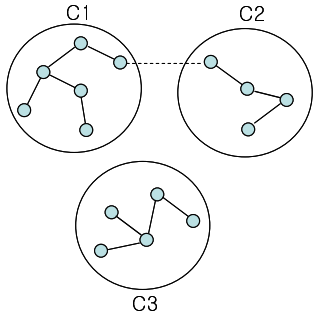

Fig. 3. Merging graphs according to class label

If there are lots of must-link constraint data, more graphs could be constructed as depicted in Figure 2.

Each graph has a class label because must-link data gives that information. We can merge graphs which has the same label as illustrated in Figure 3 (if C1 and C2 has the same label). By merging graphs, we can figure out the number of classes if we pick up sufficient number of must-link data. We also assign the must-link data to corresponding class before phase II.

2.3 Phase II : Applying Clustering Algorithm Effectively by Using Cannot-Link Constraints

In phase II, rest of the data is assigned to the preset-up classes by applying modified *k*-means clustering algorithm. 'Modified' means that we use cannot-link constraints to assign data. If two instances have a cannot-link relation, it cannot be in a same class. So, our algorithm does not assign data to the class which has the data that has cannot-link relationship with the assigning data. Suggested algorithm is depicted below.

1. Construct must-link graphs using must-link constraints
2. Merge graphs which has the same category ID (classes)
3. Construct clusters using graphs of step 2 (this determines proper number of clusters automatically).
4. For each point d_i in D, except must-link data already assigned at step 3, assign it to the closest cluster C_j such that C_j does not have cannot-link data with d_i. If no such

cluster exists, print "fail" and exit program.
5. For each cluster C_i, update its center by averaging all of the points that have been assigned to it.
6. Repeat step 4 and 5 until the whole data is covered.

3 Experimental Results

We used 4 datasets to verify our method as shown in Table 1.

Table 1. Test dataset used

Dataset	Instance	Attribute	Class
Soybean	47	35	4
Zoo	101	16	7
Glass	214	9	6
Image Segmentation	2100	19	7

We used the majority voting to evaluate the results.

$$MajorityVote(C_i) = \frac{Number\,of\,MajorityData(C_i)}{Number\,of\,Data(C_i)} \tag{1}$$

The majority voting formula gives high value when there are lots of same labeled data in the class.

We tested 5 times for COP k-means with user-providing the values for k and the proposed method could automatically determine the k value in phase I. We obtained an average value from 5 trials for each of the tested number of constraints. The proposed method gives better results about 13%, 4%, 12% and 8% over COP k-means as illustrated in Figures 4-7. An important observation here is COP k-means does not give better results than original k-means except for the Soybean dataset. The proposed method gives better results when compared to the original k-means and COP k-means.

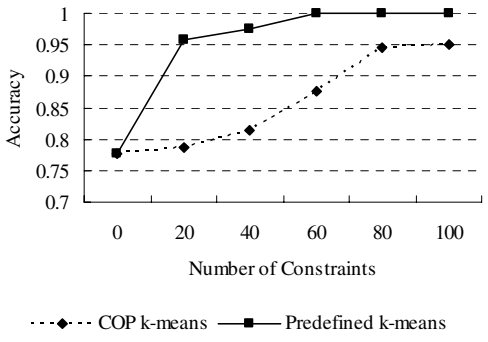

Fig. 4. Test result for soy bean dataset

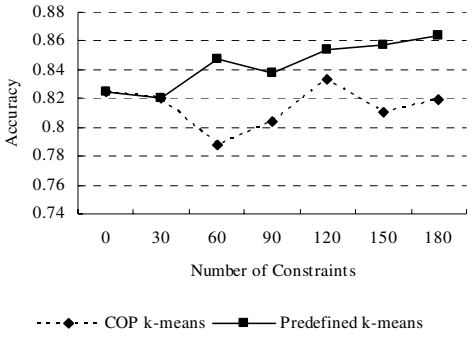

Fig. 5. Test result for zoo dataset

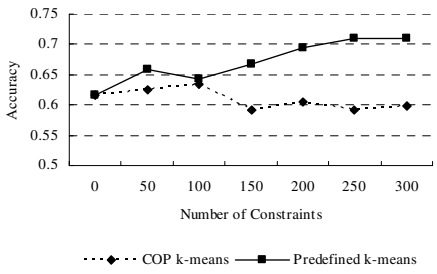

Fig. 6. Test result for glass

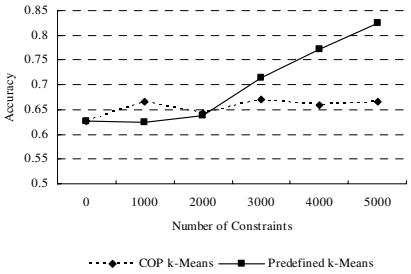

Fig. 7. Test result for image segmentation

4 Conclusions

Recent research has shown the importance of the conventional clustering problem called semi-supervised clustering, which performs clustering in the presence of some background knowledge or supervisory information. This paper proposed a new method to use background knowledge related to a domain more efficiently. For a given data, the number of classes is investigated by using the must-link constraints before clustering and these must-link data are assigned to the corresponding classes. The proposed clustering approach improves the result obtained by the direct COP k-means by about 10% on average.

Acknowledgments

Work supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

References

[1] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained k-means clustering with background knowledge", Proceedings of 18'th international conf. on machine learning, 2001, p. 577-584.

[2] Giordano Adami, Paolo Avesani and Diego Sona, Clustering documents into a web directory for bootstrapping a supervised classification, Data & Knowledge Engineering, Volume 54, Issue 3, pp. 301-325, 2005.

[3] Hong-Bin Shen, Jie Yang, Xiao-Jun Liu and Kuo-Chen Chou, Using supervised fuzzy clustering to predict protein structural classes, Biochemical and Biophysical Research Communications, Volume 334, Issue 2, 26 pp. 577-581, 2005.

[4] E. Zio and P. Baraldi, Identification of nuclear transients via optimized fuzzy clustering, Annals of Nuclear Energy, Volume 32, Issue 10, pp. 1068-1080, 2005.

[5] Christoph F. Eick, Alain Rouhana, A. Bagherjeiran and R. Vilalta, Using clustering to learn distance functions for supervised similarity assessment, Engineering Applications of Artificial Intelligence, Volume 19, Issue 4, pp. 395-401, 2006.

[6] Hong Chang and Dit-Yan Yeung, Locally linear metric adaptation with application to semi-supervised clustering and image retrieval, Pattern Recognition, Volume 39, Issue 7, pp. 1253-1264, 2006.