Data Clustering with Principle Component for the Complete Must-Link Constraints

Chao-Lung Yang*
Department of Industrial Management
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C
Tel: (+886) 2-2733-3141 ext. 3621, Email: clyang@mail.ntust.edu.tw

Nguyen Thi Phuong Quyen
Department of Industrial Management
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C
Tel: (+886) 2-2733-3141 ext. 7111, Email: quyen.ntp@gmail.com

Maisyatus Suadaa Irfana
Department of Industrial Management
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C
Tel: (+886) 2-2733-3141 ext. 7111, Email: Fanamoet@gmail.com

Abstract. This research aims to develop an integrated method to solve a special constrained clustering problem constructed by Complete Must-Link (CML) constraints. Constrained clustering analysis is a semi-supervised learning to accommodate the information while it is available, to improve efficiency and purity of clustering. The CML clustering problem can be considered as aggregating pre-defined data groups. Through the transitive closure process of data aggregation, the data in each group are replaced by their centroid for clustering analysis. This causes information missing issue which means the data distribution or shape of the original groups are omitted, especially when the groups are intersected each other. In this research, a new method named CML-PCA is proposed for CML constrained clustering problem. The principal component analysis (PCA) which provides the supplemental information describing original partition blocks is suggested to be included in the distance matrix of the constrained clustering algorithm if they are intersected each other. The intersected ratio is invented to determine whether CML data partitions are intersected or not. The proposed algorithm is tested by using the simulated dataset and real-world data sets. From the experimental result, the proposed CML-PCA outperforms the traditional agglomerative clustering method when multiple validation indices were compared

Keywords: maximum five keywords should be included

1. INTRODUCTION

In recent years, the semi-supervised clustering method emerges and attracts a lot of attention from the data mining community. In contrast to traditional (unsupervised) clustering, the semi-supervised clustering conducts the clustering process under the guidance of some supervisory information to improve efficiency and purity of clustering (Zhao et al., 2012; Jiang et al., 2013). The supervisory information can be represented by two kinds of instance constraints: Must-Link (ML) constraint and Cannot-Link (CL) constraint (Wagstaff and Cardie, 2000; Wagstaff et al., 2001). Essentially, a ML constraint specifies that two instances must be placed in the same cluster while a CL constraint enforces that two instances should not be placed in the same cluster. ML and CL constraints can be formed by the pre-determined information relieved in the collected dataset or by the expert in the domain.

In addition, the constrained ML data instances might possess the prior information such as the known member groupings or associations between data in order to construct group-level constraints. The pre-existing knowledge can capture larger building blocks of instances in the dataset *D* to form the group-level constraints. The group-level constraint can be treated as the union of some ML constraints based on the transitive and combinable characteristics of the ML constraint. For instance, two ML constraint $\{x_1, x_2\}$ and $\{x_2, x_3\}$ implies that an ML constraint $\{x_1, x_3\}$ exists and they can be combined into a group-level constraint

$\{x_1, x_2, x_3\}$ (Johnson and Wichern, 2002).

Prior information such as known member groupings or associations between the data instances might be available to construct group-level constraints. For example, the nationalities or departments of participants are usually considered as the group-level constraints, if participants from the same country or department should be grouped together. Or, the products manufactured at the same batch need to be clustered together. This kind of data feature can provide a complete set of must-link constraints in which each data instance has at least one must-link constraint with another data instance. In other words, it is a special case of a constrained clustering problem in which the r pre-existing partitions ($M_1,...,M_r$) among n items is provided. Each data instance x_i in the dataset D must belong to one of $M_1,..., M_r$ partitions exclusively where $i= 1,..., n$, n is totaling number of data instances and r is the number of pre-existing partitions or groups. If the clustering is performed upon those pre-determined partitions with r≫k (k is the number of clusters) and no CL constraint is involved, in this research, we named this particular restriction as complete must-link (CML) constraints for clustering.

Essentially, CML constrained clustering can be considered as aggregating pre-defined groups to fewer number of clusters. Through the transitive closure construction which combines multiple pair-wise ML in a pre-defined group, each data instance will be associated with other data instances in the pre-defined group. For example, if A must link with B, and B must link with C, it will lead to A, B, and C must link with each other and form a pre-defined group contains A, B and C. This transitive closure construction in the constrained clustering methods such as constrained K-means or hierarchical clustering in fact use the centroid of a pre-defined group to represent data instances in that group. Therefore, the CML constrained clustering problem can be simplified as data clustering upon centroid of pre-defined groups

Existing methods for constrained clustering can be divided in two groups: constraint-based approach and distance-based approach. For both constraint-based approaches and distance-based approaches, the transitive closure construction in fact uses the centroid of a pre-defined group to represent data instances in that group. Therefore, the constrained clustering problem can be simplified as data clustering upon centroid of pre-defined groups. Although the computational performance of constrained clustering algorithms is promising, the existing drawback of using centroid is the information missing of the original dataset. Because all data points in a pre-defined group are represented by a single centroid, the information such as distribution or shape of original member will be neglected. This information loss, in fact, affects the clustering result when dealing with ML constraints. These affects are more significant especially when pre-defined groups contain the intersection instances with high density.

Figure 1 uses an example to illustrate this information missing issue. There are three partitions: Group 1, Group 2, and Group 3 intersect each other. These three partitions can be treated as pre-assigned groups constructed by CML constraints. The centers of Group 1, Group 2, and Group 3 are denoted by triangle, square, and circle symbols, respectively. When performing clustering on these three pre-assigned partitions (extremely simple case), centers of the groups will be represented the partitions based on transitive closure construction. Only considering the group centroid in performing clustering, group 1 and group 2 should be clustered together based on the location of centroid. However, from different perspective of clustering, such as considering the shape of data distribution, it seems group 1 and group 3 are more likely to be placed in the same cluster. Obviously, the existing clustering algorithm is not able to cluster three groups based on the distribution perspective under ML constraints, because the existing methods use the centroid of partition data to represent data points in the original group. This centroid representation will mislead the clustering result especially when the pre-defined groups are mixed together. Therefore, it is necessary to consider the information missing issue of the mixing dataset

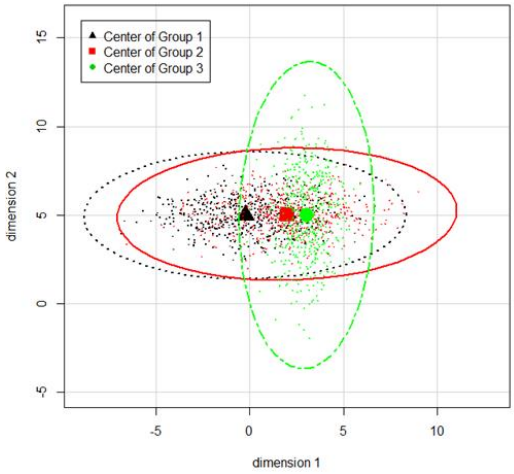

Figure 1: An illustration of intersection groups

This research proposes a method to deal with the constrained clustering problem. To avoid the information missing issue due to using only centroid in the existing methods for constrained clustering, the new method considers integrating the principal component analysis (PCA) in clustering procedure. The principal component analysis (PCA) which provides the supplemental information describing original groups is suggested to be added in the distance matrix of the constrained clustering algorithm. The research objective of this work is to develop an innovative clustering method which can combine PCA information to provide a better clustering result when pre-defined CML groups are mixed.

The rest of this paper is organized as follows. Section 2 provides the literature review about two most prevail constrained clustering methods: constrained K-means and agglomerative hierarchical clustering algorithms. Section 3 describes the proposed method– PCA-based CML constrained clustering. In section 4, the experimental results by applying proposed PCA-based CML clustering methods on a variety of datasets are shown. Finally, section 5 concludes this research and addresses the future research direction.

2. LITERATURE REVIEW

The instance-level ML and CL constraints have been integrated with traditional clustering algorithms, both on partitioned and hierarchical methods (Jain et al., 1999; Jain, 2010). Most of these existing semi-supervised clustering algorithms are designed for partitioned clustering methods and few research efforts have been reported on semi-supervised hierarchical clustering methods (Xing et al., 2002). Different from partitioned clustering where the clustering results can be easily represented using vectors, clustering indicators, or connectivity matrices for optimization, the results of hierarchical clustering are more complex and typically represented as dendrogram or trees. In Table 1, the input, output, complexity of algorithm, and advantage/disadvantage of traditional and constrained clustering algorithms are compared. Here, two prevailing algorithms: K-means and agglomerative hierarchical are investigated separately because of their fundamental differences. Note that all algorithms are compared based on handling CML constraints problem which is our interest in this research.

As can be seen, K-means algorithm has less time and space complexity than agglomerative hierarchical algorithm no matter in traditional or constrained versions. That is the main reason of K-means' popularity. However, K-means is sensitive to initial cluster selection and it only converges to local minimum. Therefore, it usually needs to restart several times for choosing the smallest value of the error. On the other hand, agglomerative hierarchical algorithm has more versatile because it allows determining the number of clusters later based on the generated dendrogram. It makes sense that this nature of hierarchical algorithm costs more complexity.

The efficiency of both constrained K-means (Wagstaff et al., 2001) and constrained hierarchical algorithms (Davidson and Ravi, 2005) can be significantly enhanced when CML constraints are provided. However, the CML datasets considered as pre-determined partitions of data are usually intersected in many real-world cases. How to deal with the intersection among CML groups is unknown. In addition, only using the centroids to represent original CML partitions and performing clustering on them may not be very effective due to the information missing issue mentioned in the first section.

Table 1: Comparison of clustering algorithms: K-means and agglomerative hierarchical. (Yang, 2009)

	K-means		**Agglomerative Hierarchical**	
Algorithm	**Traditional version**	**Constrained version (with CML constraints)**	**Traditional version**	**Constrained version (with CML constraints)**
Input	Number of clusters (k), Distance matrix ($n \times n$)	Number of clusters (k) Distance matrix of pre-determined groups ($r \times r$)	Distance matrix, size is ($n \times n$)	Distance matrix of pre-determined groups ($r \times r$)
Output	k clusters	k clusters	Dendrogram with all data points	Dendrogram with r centroid of connected components
Time Complexity	$O(knl)$	$O(krl)$	$O(n^2 \log n)$	$O(r^2 \log r)$
Space Complexity	$O(k+n)$	$O(k+r)$ If $k <= r$	$O(n^2)$	$O(r^2)$
Advantages	Simple and ubiquitous	ML constraints can improve efficiency	Consistency of clustering result Choose appropriate k later by dendrogram	ML constraints can improve efficiency
Disadvantages	Decide k in advance Sensitive to initial selection Converge to local minimum	Detailed information of original data is ignored	More time and space complexity	Detailed information of original data is ignored

r: # of connected components; k: # of clusters; n: # of instances; l: # of iterations

3. METHODOLOGY

This research considers the CML constrained clustering for intersected data problem. In order to deal with the information missing issue, the proposed algorithm called CML-PCA utilizes principal component loadings as the supplementary information to describe the pre-determined data groups constructed by CML constraints. Figure 2 illustrates the procedures of PCA-CML clustering algorithm. First, the CML constraints dataset is firstly applied the agglomerative clustering method to produce a dendrogram First, by the transitive closure construction, the partition blocks can be constructed in CML distance matrix. Then, the centroid of each partition block is calculated to form a centroid set based on the constrained clustering method. Dendrogram is generated based on the distance matrix constructed by centroid of each CML partition. Then, the intersected ratio which is invented in this research to measure the intersection of clusters constructed by CML groups is calculated. Starting from the top of dendrogram, if the CML groups intersect each other, the PCA loadings of the intersected CML groups are added in the distance matrix for extra clustering process. Based on this new distance matrix, the clustering is re-performed. If the CML groups do not intersect, the algorithm keeps searching down to the next level of the branch of dendrogram. The recursive process containing intersected ratio calculation and new distance matrix generation, and splitting process is repeated.

In this research, two PCs with the largest eigenvalues are selected as auxiliary components to supplement data coordination information in the intersected CML groups. The intersected ratio is then recursively calculated through the agglomerative clustering process to check if the clustering results under the certain branch of a dendrogram are intersected. The next sub-section will show the detailed information about how to calculate the intersected ratio of clustering result.

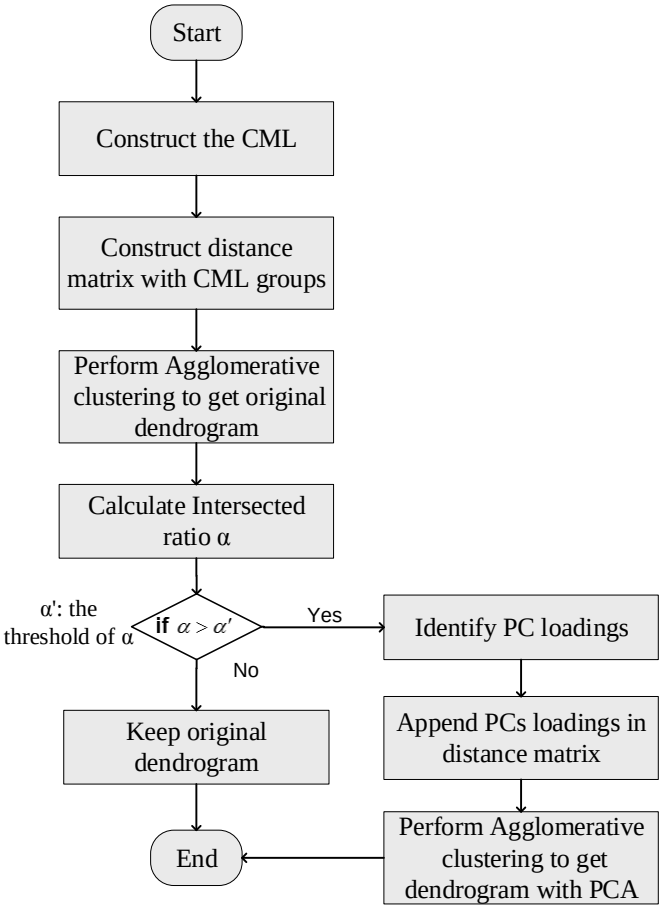

Figure 2: The proposed PCA-CML method

3.1 Intersected Ratio

In this research, intersected ratio (α) was purposed to determine if the CML groups based on the original data are intersected each other. Basically, the sum of square error (SSE) of a data cluster is used to specify the dispersion of a cluster. Larger SSE is, more dispersive the cluster is. The following subsection describes the definition of intersected ratio by using SSE and an example of calculation is addressed. Table 2 lists the notation used for intersected ratio calculation.

Table 2: Notation of intersected ratio

Symbol	Description
α	Intersected ratio
x	Data object
CML_i	The ith CML partition
$\boldsymbol{c_i}$	The centroid of CML_i partition
G	A set of CML partitions which are clustered together. Multiple CML partitions can be included in $\boldsymbol{G}$
CML_G	The SSE of CMLgroup with G components
c_G	The centroid of CML_G group

Through agglomerative clustering process upon CML constraints, multiple CML partitions can be clustered together. For a generated cluster with multiple CML partitions, the intersected ratio (α) which is the ratio of the summation of SSE of individual CML partitions in a cluster and the SSE of a cluster containing data points from all CML partitions, is proposed to measure the dispersion of the data cluster. The equation of intersected ratio is defined as Equation (2). The upper bond of the intersected ratio (α) is 1 which means the CML partitions are completely intersected when the partitions are exactly the same. If intersected ratio is larger, it means that CMLs partitions of the cluster are more intersected each other.

$$\alpha = \frac{\sum_{i \in G} SSE_i}{SSE_G}, \quad where\ SSE_G = \sum_{x \in CML_G} dist(c_G, x)^2 \qquad (1)$$

$$\alpha \in [0,1]$$

Figure 3 and Figure 4 illustrate two scenarios of intersected ratio calculation: intersected data and non-intersected data. As can been seen in Figure 3, there are 4 CML partitions specified by different colors that are intersected each other. The color circle indicates the range of a particular data distribution. Obviously, these four CML partitions have different dispersion. The intersected ratio is calculated at 0.9981 can be used to indicate the level of intersection is large in this case. Another case in Figure 4 shows 4 CML partitions which are far away from each other. The intersected ratio of this case is at 0.0233.

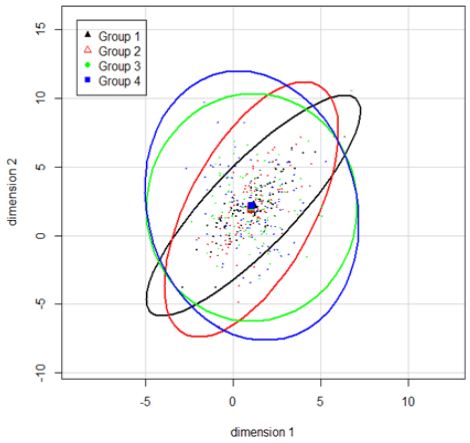

Figure 3: Illustration of intersected data

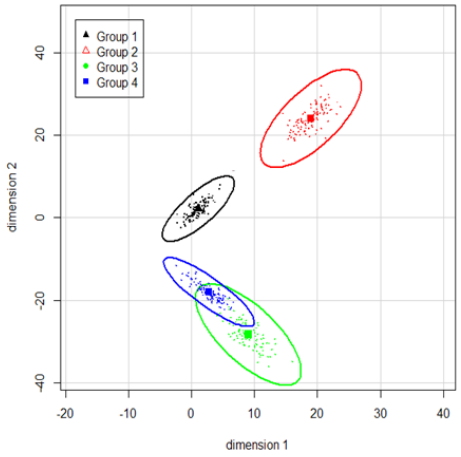

Figure 4: Illustration of non-intersected data

The two cases in Figure 3 and Figure 4 in fact demonstrate two extreme cases of the dispersion of CML partitions. By using the proposed intersected ratio, we can easily identify if the concerned CML partitions are intersected or not. To determine if CML partitions of a cluster are intersected or not, the threshold called α' need to be specified. If the intersected ration of the concerned CML partitions is larger than the threshold (α≥α'), the PCA loadings will be added in the clustering process to differentiate the intersected CML partitions. In this research, we set the threshold α'=0.5 to determine if CML partitions of a cluster are intersected because the 0.5 mean the middle level of intersected in CML data sets.

3.2 Principal Component Loadings

PCA is a multivariate mathematical technique to extract the most representative information among multiple data features. PCA method is commonly used in data dimension reduction and interpretation (Johnson and Wichern, 2002). According to Peres-Neto et al. (2005) the two non-trivial PCs are selected to integrate to the original

distance matrix in order to provide the supplementary information. The supplementary information about the partition blocks is added in the new distance matrix (or similarity matrix) for the intersected groups to alleviate the information missing issue caused by centroid replacement.

The PCA loadings are computed by using data points in the concerned CML partitions. The PCs are calculated based on the original data (Jolliffe, 2002). The new distance/similarity matrix is constructed based on the combined centroid feature set which includes the original centroid set and adding features.

$$\boldsymbol{m'} = \boldsymbol{[m, s]} \quad (2)$$

m: original similarity matrix

m': new similarity matrix

s: supplementary feature (as a matrix)

To describe the dispersion characteristics of each intersected data partition, the eigenvectors of each intersection data partition are computed by PCA method. The number of the adding features (loadings of the first two eigenvectors) is equal to the number of features due to the nature of PCA.

4. EXPERIMENT

The experiment is conducted to test the performance of the proposed CML-PCA on both the simulated datasets and real world data sets. The simulated dataset was generated by random number generator to evaluate the performance under the intersected CML partitions. The real world dataset consists of the yeast retrieved from UCI machine learning repository (http://archive.ics.uci.edu/ml/) and cartridge datasets which is the data about color printing quality measurement.

In this research, we used R language to implement the algorithm because the software has many predefined functions available. Another reason of using R is because it is an open source and modular software package supported by many communities. The intersected ratio 0.5 is chosen as the threshold in our experiment, because the 0.5 mean the middle level of intersection in CML groups.

To evaluate the efficiency of the proposed method, a set of validation indices which contain 8 widely used internal validation measures (Liu et al., 2012) is considered in this research. The Calinski_Harabasz index meausres the cluster validity based on the average between- and within-cluster sum of squares. The Davies-Bouldin index is calculated by averaging all the cluster similarities. The Dunn's index measures the minimum pairwise distance between objects in different clusters and the maximum diameter among all clusters. The SD index considers the concepts of the average scattering and the total separation of clusters. The index S_Dbw takes density into account to measure the inter-cluster separation. The Silhouette index counts the clustering performance based on the pairwise difference of between- and within-cluster distances. The index Xie-Beni considers the minimum square distance between cluster centers and the mean square distance between each data object and its cluster center. The proposed CML-PCA result is compared to the original clustering result which is constructed based on the agglomerative clustering process. The experimental results are shown in the next sections.

4.1 Simulated Data Experiment

The simulated data set consists of 10 CMLs which intersect with each other. Each CML partition has 100 data points. The data is generated randomly with the specified mean and standard deviation based on normal distribution. Figure 5 illustrates the simulated data. Obviously, the CML partitions in this data set are intersected. If performing the traditional hierarchical clustering method with group-level constrains, the algorithm, in fact, clusters the centroids of CML partitions. It means the dispersion of CML partitions will not be considered. If clustering this data set by the proposed CML-PCA algorithm, we expect the clustering result will not only consider the centroids of the CML partitions but also the dispersion of them.

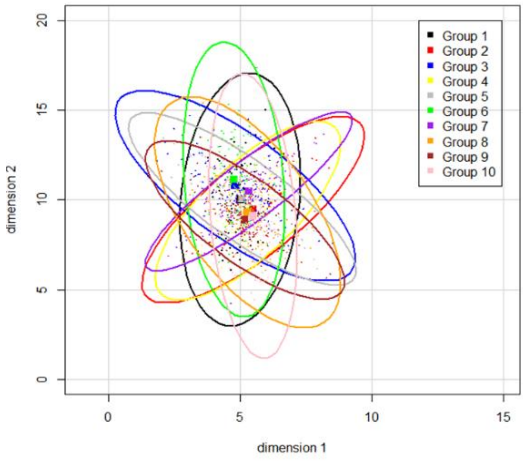

Figure 5: Illustration of intersected simulated data

The detailed process of the proposed CML-PCA is addressed. First, the original clustering result (dendrogram) is constructed based on the agglomerative clustering process. Then, the re-clustering process is performed on the whole dendrogram. If the intersected ratio of the dendrogram is smaller than $\alpha' = 0.5$ (non- intersected), we need to split the dendrogram into two parts based on the highest distance (height). In contrary, when the intersected ratio is higher than the threshold which means the data set on the dendrogram branch are intersected, we need to add PCA vectors to the distance matrix and re-cluster it based on CML-PCA method. The computational result for the simulated data showed that

the intersected ratio is 0.86 that exceeds the threshold. Therefore, PCA vector is added to the distance matrix of the root of clustering dendrogram. After including the PCA loading, the clustering result is able to consider their PCA loadings to perform clustering by not only their centroid distance but also the data distribution of each CML partition.

Figure 6 shows the comparison of the intersected dataset result with and without adding PCA. As can be seen on the left chart, the CML partition #2 is clustered in the different group against #4 and #7. After considering adding PCA loadings, the proposed CML-PCA clusters CML #2 together with CML #4 first and then with CML #7. This clustering result is consistent with the expectation because CML #2, #4, and #7 have very similar dispersion.

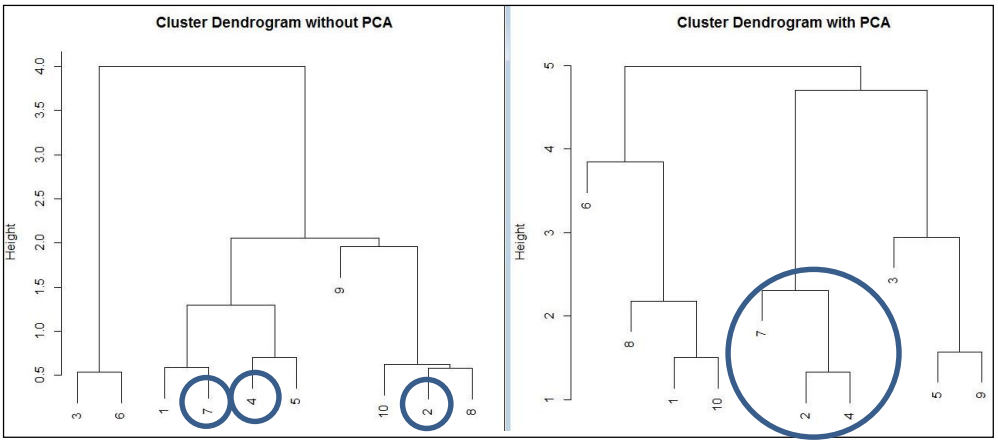

Figure 6: A comparison of clustering results on the simulated dataset with and without adding PCA loading

The comparison of the validation indices are shown in Table 3. Obviously, the proposed CML-PCA outperforms the existing agglomerative clustering method.

Table 3: Comparison of validation indices on simulated data

Index	CML without PCA	CML-PCA	Rule
Calinski-Harabasz	25.590	57.771	max
Davies-Bouldin	20.962	6.7127	min
Dunn	0.002	0.003	max
SD-scat	0.986	0.918	min
SD-dis	4.272	20.439	max
S-Dbw	4.272	4.054	min
Silhouette	-0.007	-0.001	max
Xie-Beni	484586	362159	min

4.2 UCI data sets

4.2.1 Yeast Dataset

Yeast dataset contains data related to the localization site of yeast protein. The data have 1484 instances with 9 attributes. The class distribution is the localization site which is used as a group-level constrains to form CML partitions. The data set consists of 14 CML partitions. The intersected ratio shows that all yeast data intersect each other ($\alpha = 0.97$). Therefore, we need to integrate the PCA loading in the distance matrix. Table 4 compares the clustering results with and without adding PCA.

4.2.2 Cartridge Dataset

Cartridge dataset contains the sensor information used for the calibration process of a color laser printer. The dataset contains three kinds of data: 1) sensor information during the calibration, 2) print-out measurements right after the calibration, and 3) cartridge information. The objective of collecting this dataset is to investigate factors which affect color printing and develop a new sensor mapping model to calibrate the printer. Similar to the experiments of the simulated and Yeast dataset, the PCA loading is added in the distance matrix when the intersected ratio exceeds the threshold.

The result of performing 8 validation indices on clustering result on the cartridge dataset is shown in Table 5. Once again, the proposed CML-PCA achieves significantly better performance than the original agglomerative clustering result.

Table 4: Comparison of validation indices on Yeast dataset

Index	CML without PCA	CML-PCA	Rule
Calinski-Harabasz	57.792	121.112	max
Davies-Bouldin	1.835	1.943	min
Dunn	0.010	0.012	max
SD-scat	1.809	0.897	min
SD-dis	8.944	13.728	max
S-Dbw	4.327	3.638	min
Silhouette	0.038	0.072	max
Xie-Beni	372.35	328.89	min

Table 5: Comparison of validation indices on Catridge dataset

Index	CM-without PCA	CML-PCA	Rule
Calinski-Harabasz	243	825	max
Davies-Bouldin	0.8340	0.8826	min
Dunn	0.0011	0.0041	max
SD-scat	0.4200	0.2193	min
SD-dis	0.5956	0.7797	max
S-Dbw	3.4936	0.2541	min
Silhouette	0.1880	0.4683	max
Xie-Beni	390325	306824	min

In the cartridge data set, the CML-PCA method was performed to demonstrate the promising performance on the semi-intersected data distribution. When the CML groups intersect each other, CML-PCA method is able to consider the dispersion by adding PCA loading to cluster the CML partitions by their data distribution. The CML-PCA method is particularly useful when pre-determined data groups are intersected with each other

5. CONCLUSION

In this research, a new constrained clustering method called CML-PCA is proposed to deal with CML constraints which are group-level constraints by pre-determined factors or background knowledge. The proposed CML-PCA designs and implements the innovative algorithm to integrate PCA loadings of CML partitions in the distance matrix for clustering. This new distance matrix can consider the cluster data not only by centroid distance but also by the data dispersion which is important when data pre-partitions are intersected. Additionally, the intersected ratio is also defined in this research to measure the degree of the data intersection. The reason of utilizing intersected ratio is to evaluate the level of intersection among data partitions.

The experimental results are performed in simulated dataset and UCI datasets. The set including 8 widely used validation indices are used to compare the performance of the CML-PCA and traditional constrained clustering algorithm. Based on experimental result, the CML-PCA is significantly better than the traditional method. Particularly for cartridge dataset, the new clustering result performed by the CML-PCA shows the sensor partitions can be grouped not only by the centroids but also by the dispersion of data group characteristics. This result is useful for evaluating cartridge quality in the practical application. For future research, the proposed algorithm should be tested on dataset with higher dimensions or Big Data application. Especially, when the data partitions are highly intersected in the high dimension, exploring or clustering the data structures under certain knowledge about the data plays the important roles on data mining application.

ACKNOWLEDGMENTS

We appreciate the financial support from National Science Council of Taiwan, R.O.C. (Contract No. MOST-103-2221-E-011-116).

REFERENCES

Davidson, I. and Ravi, S.S. (2005) *Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho and J. Gama (ed), Knowledge Discovery in Databases*., Springer Berlin Heidelberg, chapter 11, 59-70.

Jain, A. K. (2010) "Data clustering: 50 years beyond K-means." *Pattern Recogn. Lett.*, **31**(8): 651-666.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) "Data clustering: a review." *ACM Comput. Surv.*, **31**(3): 264-323.

Jiang, H., Ren, Z., Xuan, J. and Wu, X. (2013) "Extracting elite pairwise constraints for clustering." *Neurocomputing*, **99**, 124-133

Johnson, R. A. and Wichern, D. W. (2002) *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, New Jersy 07458.

Jolliffe, I. T. (2002) *Principal Component Analysis*, Springer-Verlag New York, Inc.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., and Wu, S. (2012). Understanding and Enhancement of Internal Clustering Validation Measures. *IEEE Trans Syst Man Cybern B Cybern*. doi: 10.1109/TSMCB.2012.2220543

Peres-Neto, P. R., Jackson, D. A. and Somers, K. M. (2005) How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, **49**, 974 – 997.

Wagstaff, K. and Cardie, C. (2000) Clustering with instance-level constraints. *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 1103-1110.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001) Constrained K-means clustering with background knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 577-584.

Xing, E. P., Ng, A. Y., Jordan, M. I. and Russell, S. (2002) Distance metric learning, with application to clustering with side-information. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, MIT Press, 505–512.

Yang, C.-L. (2009) *Adaptive Clustering Model for Improving Color Consistency*, Doctor of Philosophy, Purdue University.

Zhao, W., He, Q., Ma, H. and Shi, Z. (2012) Effective semi-supervised document clustering via active learning with instance-level constraints, *Knowl. Inf. Syst*, 30(3), 569-587.