Clustering by heterogeneous data fusion: framework and applications

Shi Yu **Bart De Moor** **Yves Moreau**
Department of Electrical Engineering
Katholieke Universiteit Leuven
Leuven B3001, Belgium
{*shi.yu, bart.demoor, yves.moreau*}*@esat.kuleuven.be*

Extended Abstract

1 Introduction

Clustering is an important problem with many applications, and a number of different algorithms and methods have emerged over the years. The goal of clustering is to group data points into homogeneous groups, where the homogeneity is usually measured by distances or similarities among data points. Recently, may applications face the requirement of clustering by data fusion. This is because that information contained in single data source is limited by its specific observation, therefore, combining multiple observations might facilitate the comprehensive understanding of the problem. For instance, in order to investigate memory persistence (long term or short term memory) of bacteria, a bacterium is observed at different experimental conditions and evolutional times [11]. Then the multiple observations are categorized by clustering algorithms. In scientometrics, a strategy has been proposed to combine text mining data and bibliometrics data (hybrid clustering) to explore the structure mapping of journal sets [8]. In bioinformatics, high throughput techniques produce numerous genomic data. The challenge to endow clustering algorithm with the ability to retrieve correlated or complementary information about the underlying functional partitions of genes and proteins has attracted many interests [2,14]. Unfortunately, though the machine learning community has already focused on data fusion for classification [7] and novelty detection [5], the extension to unsupervised learning such as clustering, is still an unresolved and ongoing problem.

In this paper, we present a unified framework to obtain partitions from heterogeneous sources. When clustering by data fusion, the determination about the "relevance" or "usefulness" of the source is a vital issue to statistically guarantee a lower bound of the performance. In other words, if the clustering algorithm is able to detect the most "relevant" data source, we can expect that the fusion approach works at least as good as the best individual data. In order to achieve the above objective, two different strategies are applied in the framework. The effectiveness of the clustering performance is evaluated on several experiments and applications.

2 A framework of clustering by data fusion

Currently, clustering algorithms for data fusion can be concluded into two main categories. The first approach is clustering ensemble, also known as clustering aggregation or consensus clustering, that combines multiple partitions into a consolidate partition by consensus function. Different ensemble algorithms have the same conceptual framework as shown in Figure 1, they mainly vary on the choice of consensus functions. Clustering ensemble is originally applied on single source where various partitions are generated by different

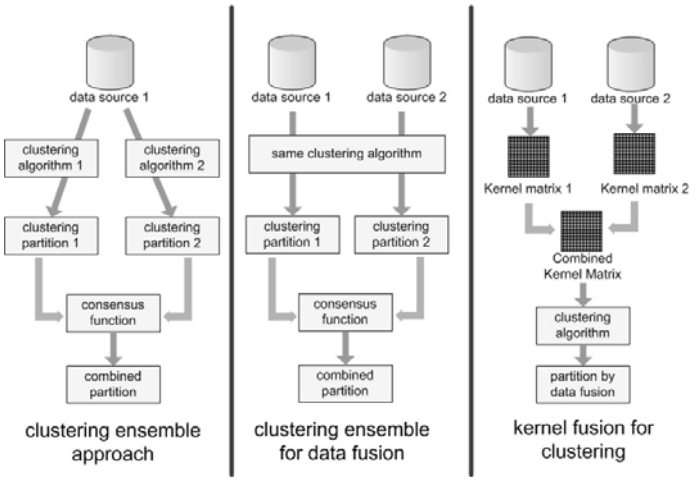


Figure 1. The conceptual framework of clustering from multiple sources

representations, dimensionality reduction techniques and clustering algorithms. The strategy of clustering ensemble can be extended to data fusion, where the main difference is that partitions now are varied by sources. An underlying assumption is: If the information contained in multiple sources is highly correlated, their partitions should also still contain "common agreement" thus a consolidated partition can also be obtained.

An alternative approach of clustering by data fusion is achieved by fusing similarity matrices. If the similarity matrices are positive semi-definite, the data integration problem can be formulated as a kernel fusion problem. The main difference of kernel fusion approach is that the integration is carried in kernel space before clustering algorithm is applied (early integration) while clustering ensemble aggregates partitions after clustering (late integration). Kernel method elegantly resolves the heterogeneities of data sources by representing them as same-size kernel matrices. Moreover, if we assume that the importance of each data source is equivalent, we can combine the kernels in an average manner, thus the issue of data integration is transparent to the clustering algorithm. The averagely combined kernel is a new data source thus the partition can be obtained by standard clustering algorithms in kernel space. Furthermore, we can also apply a more machine-intelligent approach by coupling the optimization problem of kernel learning with the objective function of pattern analysis. In that case, the weights assigned on each data source can be adjusted adaptively during the clustering procedure [4]. In this paper, we propose a novel adaptive kernel K-means clustering (AKKC) algorithm to obtain partitions and optimal weights simultaneously.

In conclusion, we survey 13 different algorithms from two main approaches and crossly compare them in the unified framework. We implement 6 clustering ensemble algorithms: HGPA [9], CSPA [9], MCLA [9], QMI (Quadratic Mutual Information) [10], EACAL (Evidence accumulation clustering - average linkage) [6] and AdacVote (Adaptive cumulative voting) [3]. We also implement 7 kernel fusion algorithms: AKKC [8,13], K-means clustering on averagely combined kernels, hierarchical clustering on averagely combined kernels (4 linkage methods) and spectral clustering on averagely combined kernels.

3 Experiments and Applications

In this section, we present a number of experiments to show the effectiveness of clustering framework.

3.1 Synthetic data

The first experiment is carried on 3 synthetic data sources with intuitive patterns. We assume

that 400 samples (equally divided into 2 labels as blue & red points in Figure 2) are represented in 3 different data sources. In the first source (denoted as normal data), the 2 clusters form two normal distributions with some degree of overlapping. In the second source (denoted as ring data), the 2 clusters show ring distributions. In the third one (denoted as random data), the labels are randomly assigned which represents a very irrelevant data. We combine them for clustering, first in linear space (combine 3 data sources in their original dimensions) and then extend them in nonlinear space (create 1 linear kernel and 2 RBF kernels for each data source then combine 9 kernels in total). The performance is evaluated by comparing the clustering partitions with the labels of the samples.

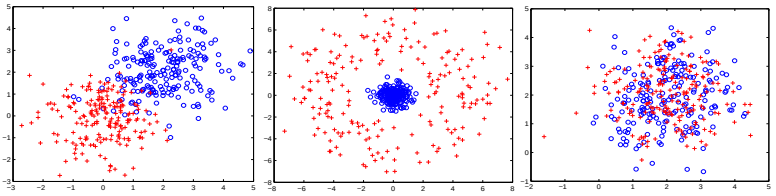


Figure 2. Three data sources in synthetic data

3.2 Digit recognition data

In the second experiment, we adopt two UCI digit recognition data sources for clustering analysis. The first source is pen based recognition data, which is created by trainers input on a tablet with 500 by 500 pixels resolution. The second source is optical recognition data, which is scanned by NIST OCR system. The labels of 10 digit classes from the original data are used for evaluation. Both data sources have been used separately as benchmark data for classification and clustering analysis. In this paper we randomly select 1500 samples (150 for each digit class) from each source and combine them for clustering. The experiment is randomly permuted and repeated 50 times.

3.3 Journal sets clustering using text mining and bibliometrics data

The third experiment arises from a real application in scientometrics analysis. We adopt a data set containing (articles, letters, notes and reviews) from year 2002 till 2006. The data set is obtained from the Web of Science (WoS) by Thomson Scientific. From this data set, 1869 journals are selected and the titles, abstracts and keywords of their papers are indexed by a text mining program. Thus we obtain a journal-by-term data source (text mining source). In the text mining source, the journals are expressed in the vector space model containing 669,860 terms and the weights of terms are calculated by TF-IDF weighting scheme. We also apply citation analysis as the supplementary of text mining analysis. For the same 1869 journals, we aggregate all paper-level citations into journal-by-journal citations. The direction of citations is ignored and a symmetric citation data is obtained. Therefore, the present study combines cross-citation analysis with text mining for clustering. To evaluate the performance, we reference the Essential Science Indicators (ESI) classification created by Thomson Scientific as the ground-truth labels of journal assignment (7 labels).

3.4 Disease genes clustering by genomic data fusion

The fourth experiment comes from bioinformatics research. We investigate a real biological problem of clustering human disease-causing genes. The ground truth labels of genes come from domain knowledge, which is also adopted as a benchmark data previously in gene prioritization system [1,12]. It consists of 537 human genes (620 occurrences) categorized in 29 diseases. Most of these diseases are complex, resulting from the interplay of both environmental and numerous distinct genetic factors. It is thus often difficult to identify disease clusters with a single data set. We adopt ten heterogeneous genomic sources (expressed in 26 kernels - several kernels can be derived from one data source) as representatives of various available genomic data sets. Due to the length restriction of the abstract, introduction about the data sources and the kernel functions are omitted.

To investigate the clustering performance, we enumerate all the pairwise combinations of the 29 diseases (resulting in 406 binary clustering tasks). All 406 tasks are repeated 20 times (resulting in 8,120 runs of clustering per kernel). In each repetition, the average performance of all 406 tasks is used as the performance indicator of this run. Then the mean value of 20 repetitions is reported as final result.

4. Conclusion and discussions

In Figure 3 we present the performances evaluated by two external validations (rand index and normalized mutual information). Our motivation in this paper is not to find the best algorithm in all experiments. We are interested in evaluating different approaches in a unified framework and to get insight about the challenges for the new emerged problem of clustering from multiple sources.

The main discovery is that data fusion indeed improves the performance. The improvement is quite significant on synthetic data, digit recognition data and disease gene data. Moreover, if we rank the performances of different algorithms and crossly compare them in four experiments, we come up with some nice candidate algorithms, for instance, our proposed AKKC algorithm, Ward linkage and spectral clustering based on averagely combined kernel. It seems that kernel fusion method generally works better than ensemble method for heterogeneous data fusion. This is probably because clustering ensemble method is quite sensitive to the number and quality of input partitions. If the number of data sources is small and the disagreement of the partitions is large, clustering ensemble method usually works worse than kernel fusion method. To avoid this disadvantage, we can first generate more partitions on single source, then combine all partitions in the fusion framework. In this way, clustering ensemble algorithms are able to find “agreement” among sufficient partitions and obtain stable consensus partitions. We can also go beyond the simple "partition generation" strategy and apply different data mining models to retrieve information from the same data source. For instance, in disease gene clustering we vary the biological text mining model and obtain 15 different textual gene profiles by changing the domain vocabularies and weighting schemes. In our early study, we found that with the same corpus collection, the choice of text mining configuration is a significant factor determining the quality of textual gene profiles in biological validations [12]. In clustering problem we find the same phenomenon (the orange bars in Figure 3 are performance obtained by different textual profiles). The merit here is, by combining these textual profiles together with biological data, the clustering performance by data fusion is strongly improved (much better than the best individual one, lddb-idf).

Clustering by data fusion is a new topic and there are still many remaining challenges. In our paper we mainly combine data sources in linear space. Nonlinear space integration is a very interesting problem (as the result of nonlinear fusion on synthetic data) but it involves a new issue of how to identify the optimal kernel mapping. Another problem is internal validation (for example, the mean silhouette value) often behaves differently on data sources. Since internal validation is often used as an indicator to find the optimal cluster number, how to extend clustering model prediction and comparison techniques to multiple sources is hence an ongoing issue.

Acknowledgement

The results presented in this paper are based on joint works with Leo Charles Tranchevent, Xinhai Liu, Frizo Janssens and Wolfgang Glänzel.

References

[1] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. & Moreau, Y. (2006) Gene prioritization through genomic data fusion, Nature Biotechnol., 24(5), 537-544.

[2] Asur, S., Ucar, D. & Parthasarathy, S. (2007) An ensemble framework for clustering protein-protein interaction networks, *Bioinformatics*, 23(13), i29-i40.

[3] Ayad, H.G. & Kamel, M.S. (2008) Cumulative Voting Consensus Method for Partitions with a

Variable Number of Clusters, *IEEE Trans. PAMI*, 30,160-173.

[4] Chen, J.H., Zhao, Z., Ye, J.P. & Liu, H. (2007) Nonlinear Adaptive Distance Metric Learning for Clustering, *ACM KDD07*.

[5] De Bie, T., Tranchevent, L.C., Van Oeffelen, L.M. & Moreau, Y. (2007) Kernel based data fusion for gene prioritization, *Bioinformatics*, 23 (13), i125-i132.

[6] Fred, A.L.N. & Jain, A.K. (2005) Combining Multiple Clusterings Using Evidence Accumulation, *IEEE Trans. PAMI*, vol.27, 835-850.

[7] Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L.E. & Jordan, M.I. (2004) Learning the kernel Matrix with Semidefinite Programming, *JMLR*, 5, 27-72.

[8] Liu, X.H., Yu, S., Moreau, Y., De Moor, B., Jassens, F. & Glänzel, W. (2008) Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets, *Internal Report, SCD-SISTA, ESAT, K.U.Leuven*, in submission.

[9] Strehl, A. & Ghosh, J. (2002) Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions, *JMLR*, 3, 583–617.

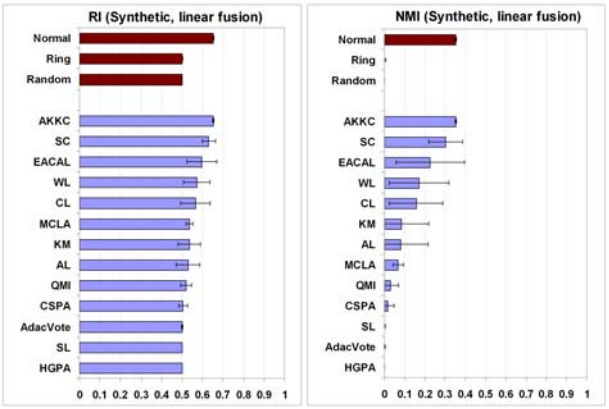
[10] Topchy, A., Jain, A.K. & Punch, W. (2005) Clustering Ensembles: Models of Consensus and Weak Partitions, *IEEE Trans. PAMI*, 27, 1866-1881.

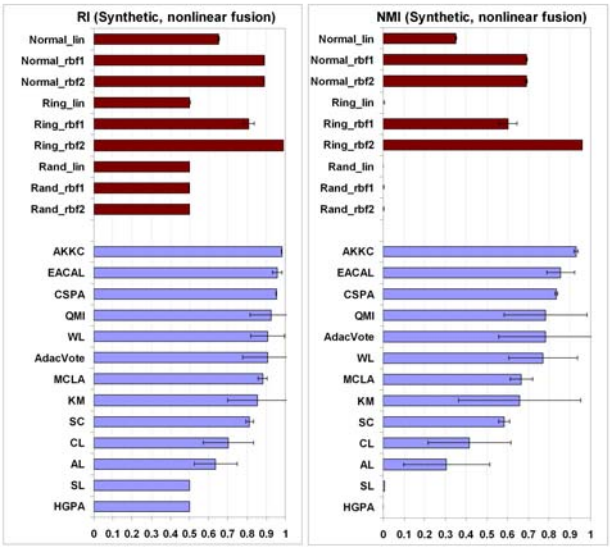
[11] Wolf, D.M., Bodin, L.F., Bischofs, I., Price, G., Keasling, J. & Arkin, A.P. (2008) Memory in Microbes: Quantifying History-Dependent Behavior in a Bacterium, *PLOS one*, 3(2), e1700.

[12] Yu, S., Van Vooren, S., Tranchevent, L.C., De Moor, B. & Moreau, Y. (2008) Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining, *Bioinformatics*, 24 (16), i119-i125.

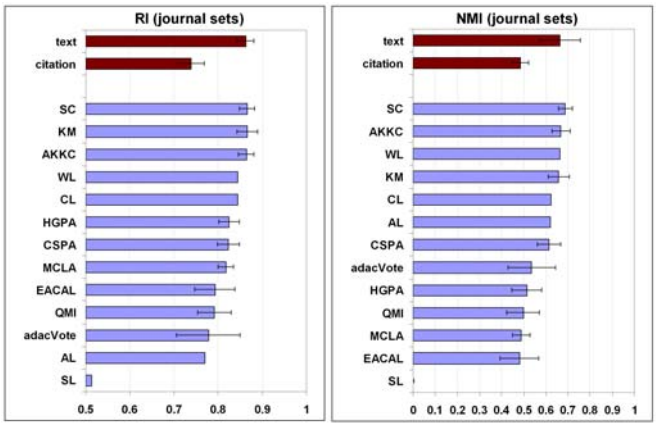
[13] Yu, S., Tranchevent, L.C., Liu, X.H., De Moor, B. & Moreau, Y. (2008) Clustering analysis by heterogeneous data fusion, *Internal Report, SCD-SISTA, ESAT, K.U.Leuven*, in submission.

[14] Yu, Z.W.., Wong, H.S. & Wang, H.Q. (2007) Graph-based consensus clustering for class discovery from gene expression data, *Bioinformatics*, 23(21), 2888-2896.

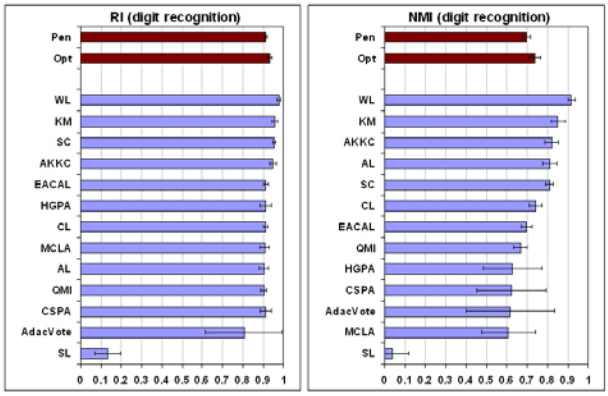




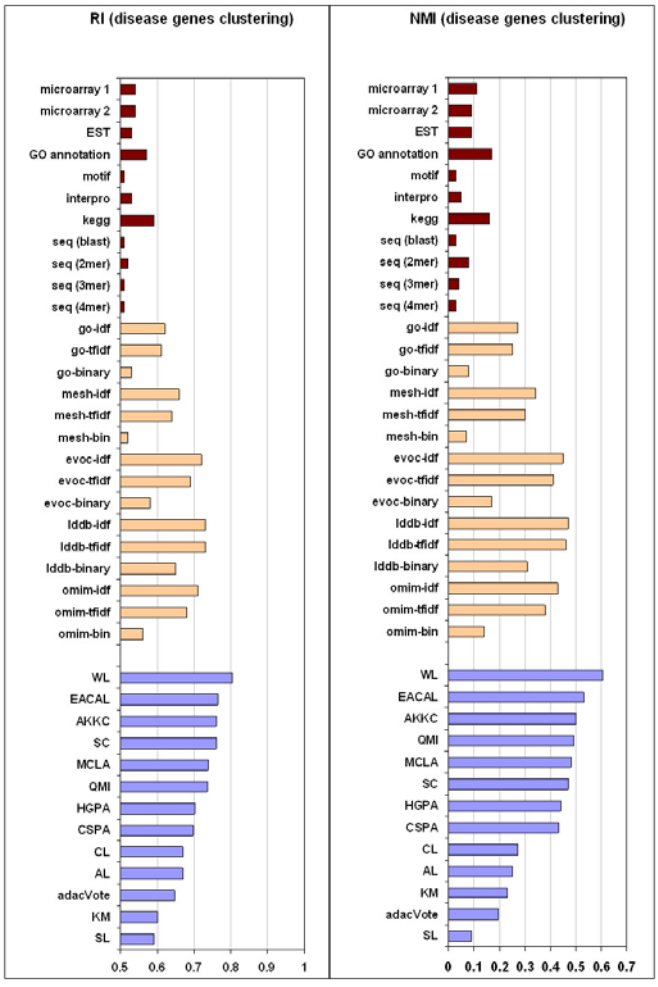
(a) synthetic data fusion



(c) journal sets clustering



(b) digit recognition data fusion



(d) disease genes clustering

Figure 3. Performance of all experiments evaluated by Rand Index & Normalized Mutual Information. *The brown bars and orange bars are performances obtained by clustering on single data (K-means). The blue bars are data fusion performances. In disease genes clustering, the data fusion results are obtained by combining 11 biological data sources (brown bars) and 15 text mining data sources (orange bars).*