Analysis and Comparisons of Clustering Consensus Functions

XU Sen[1, a], ZHOU Tian[2, b] and YU Hua Long[3, c]

[1]Intelligent Information Processing Laboratory, Yancheng Institute of Technology, Yancheng 224051, China

[2]Science and technology on Underwater Acoustic Laboratory, Harbin Engineering University, Harbin 150001, China

[3]School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China

[a]xusen@ycit.cn, [b]zhoutian@hrbeu.edu.cn, [c]yuhualong@just.edu.cn

Keywords: consensus function; clustering; algorithm.

Abstract. This paper analyzes and compares different clustering consensus functions. We analyze different consensus functions and compare them over several different document datasets. Our experimental results indicate that some consensus functions are consistently superior to the others, and can produce the best results.

Introduction

As an unsupervised machine learning method, clustering analysis has been studied for many years in data mining areas. Clustering is a common tool widely used in many applications such as image segmentation and document clustering [1]. There have been thousands of clustering algorithms, but in practice no one is able to identify clusters with different sizes, shapes, and densities [1]. Recently, many studies show that the combination of multiple clustering algorithms could result in novel and robust clustering results [2-5]. The key problem in clustering combination is how to design a consensus function so to as obtaining a final superior result.

This paper mainly researches the characteristics of the various clustering consensus functions. In particular, we evaluate nine consensus functions which are commonly seen in literature. These consensus functions explore the data collection from different views, such as modeling the data collection as a graph or modeling the data collection as hierarchical tree. We experimentally compared the performance of these consensus functions using six different datasets obtained from various sources. Experimental results indicate that some consensus functions are consistently superior to the others, and can produce the best results.

Clustering Integration

In this section, we introduce clustering integration (clustering combination or cluster ensemble) and its problems. Clustering combination usually takes two stages. At the first stage, it stores the results of some independent runs of base clustering algorithms. Then, it uses the specific consensus function to find a final result from stored results.

The problem of designing clustering consensus functions is more difficult than classifier ensembles because cluster labels are symbolic and thus we must also solve a label correspondence problem. In addition, the cluster number provided by the individual clustering results may vary based on the clustering method as well as on the particular view of the data available to that method. Furthermore, the desired number of clusters is often unknown in advance.

Many methods have been proposed for solving clusterings combination problem. Strehl and Ghosh propose CSPA (Cluster-based Similarity Partitioning Algorithm), HGPA (Hyper-graph Partitioning Algorithm) and MCLA (Meta-Clustering Algorithm) based on graph partitioning algorithms [2]. Fred and Jain [3] use agglomerative clustering algorithms such as single linkage,

Published by Atlantis Press, Paris, France.

complete linkage, average linkage and ward linkage algorithms to produce a final clustering according to the similarity matrix. Recently, Xu et al [4, 5] introduce the key ideas of spectral graph theory and proposed two spectral clustering algorithms to solve document clustering combination problem.

The above methods all have some disadvantages. For example, graph partitioning algorithms usually impose strong constraint on the size of clusters so as to avoid trivial solution. Agglomerative clustering algorithms explicitly or implicitly impose a structure to the final clustering. For instance, single linkage algorithm will cause “chain effect”, and the shape of clusters found by complete linkage algorithm is spherically like.

Experimental Results

Experimental Dataset. In our experiments, we used six different datasets, and Table 1 summarizes the general characteristics. We obtained these datasets from different sources to ensure diversity in the datasets. For all datasets, we used a stop-list to remove common words. Moreover, any term that occurs in fewer than two documents is eliminated.

Table 1 Summary of datasets used to evaluate the various consensus functions.

Data	Source	# of documents	# of terms	# of classes
fbis	FBIS(TREC)	2463	12674	17
hitech	San Jose Mercury(TREC)	2301	13170	6
reviews	San Jose Mercury(TREC)	4069	23220	5
la12	LA Times(TREC)	6279	21604	6
tr31	TREC	927	10128	7
tr41	TREC	878	7454	10

The *fbis* dataset is from the Foreign Broadcast Information Service data of TREC-5 [6]. The *hitech* and *reviews* datasets were derived from the San Jose Mercury newspaper articles that are distributed as part of the TREC collection. Each one of these datasets were constructed by selecting documents that are part of certain topics in which the various articles were categorized (based on the DESCRIPT tag). The *hitech* dataset contained documents about computers, electronics, health, medical, research, and technology; and the *reviews* dataset contained documents about food, movies, music, radio, and restaurants. The *la12* dataset was obtained from articles of the Los Angeles Times that was used in TREC-5 [6]. The categories correspond to the desk of the paper that each article appeared and include documents from the entertainment, financial, foreign, metro, national, and sports desks. Datasets *tr31* and *tr41* are derived from TREC-6 [6], and TREC-7 [6] collections. The classes of these datasets correspond to the documents that were judged relevant to particular queries.

Experimental Results Analysis. Since class labels are available, we adopt the NMI criterion to quantify the match between the categorization and the clustering result. NMI provides a measure that is impartial with respect to k as compared to purity and entropy. It reaches maximum value 1, only when the two sets of labels have a perfect one-to-one correspondence. Also, we use the F1 measure which is an oft-used measure in the information retrieval and natural language processing communities.

For each dataset, we use *k*-means with cosine similarity function for five times, each with random initialization and generating a different clustering solution. Each number reported below is obtained by averaging the NMI scores and the corresponding F1 measures.

Fig. 1 and Fig 2 show the NMI scores and F1 measures of the nine clustering integration algorithms over six datasets, respectively. In the figures, EASL, EACL, EAAL and EAWL represent the single linkage, complete linkage, average linkage and ward linkage algorithms proposed in [3], respectively, SGTA is the spectral graph theory-based algorithm proposed in [4], and SMSA is the similarity matrix- based spectral algorithm proposed in [5].

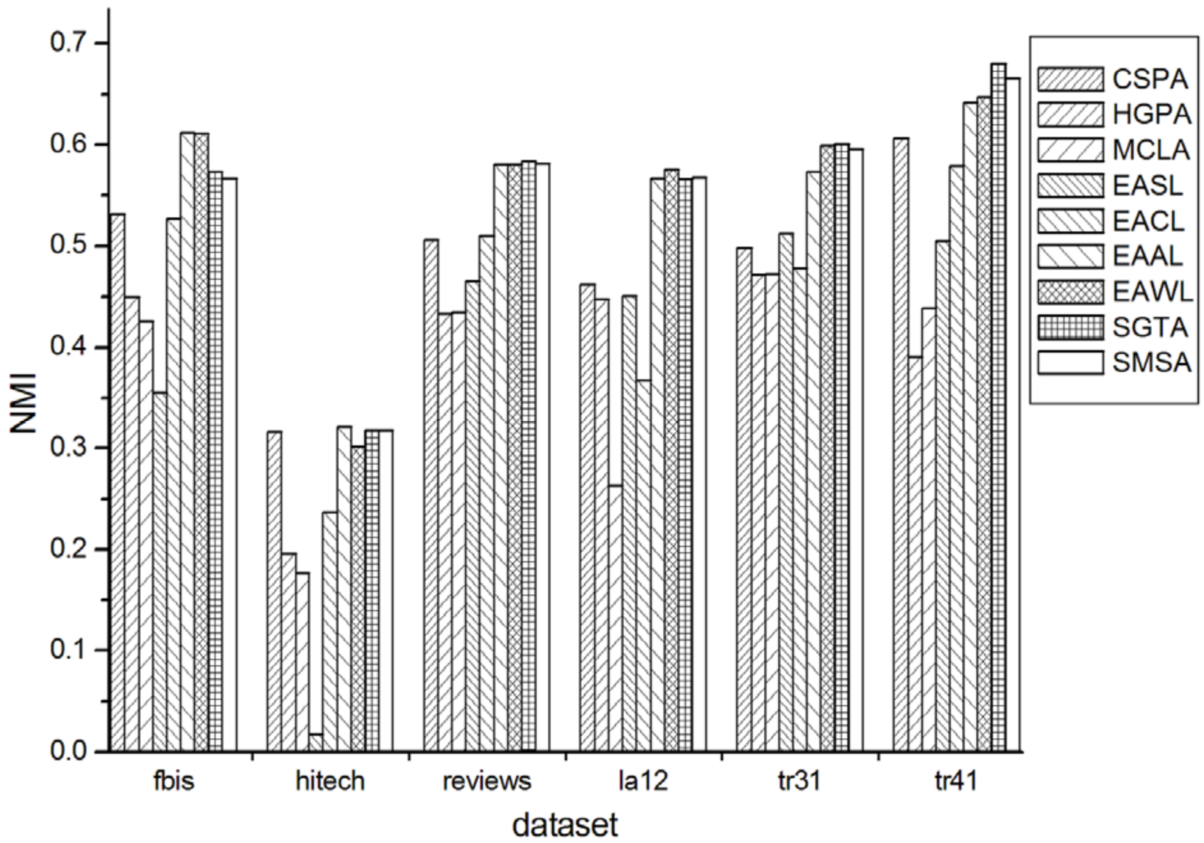

Fig. 1 NMI scores of different clustering integration algorithms

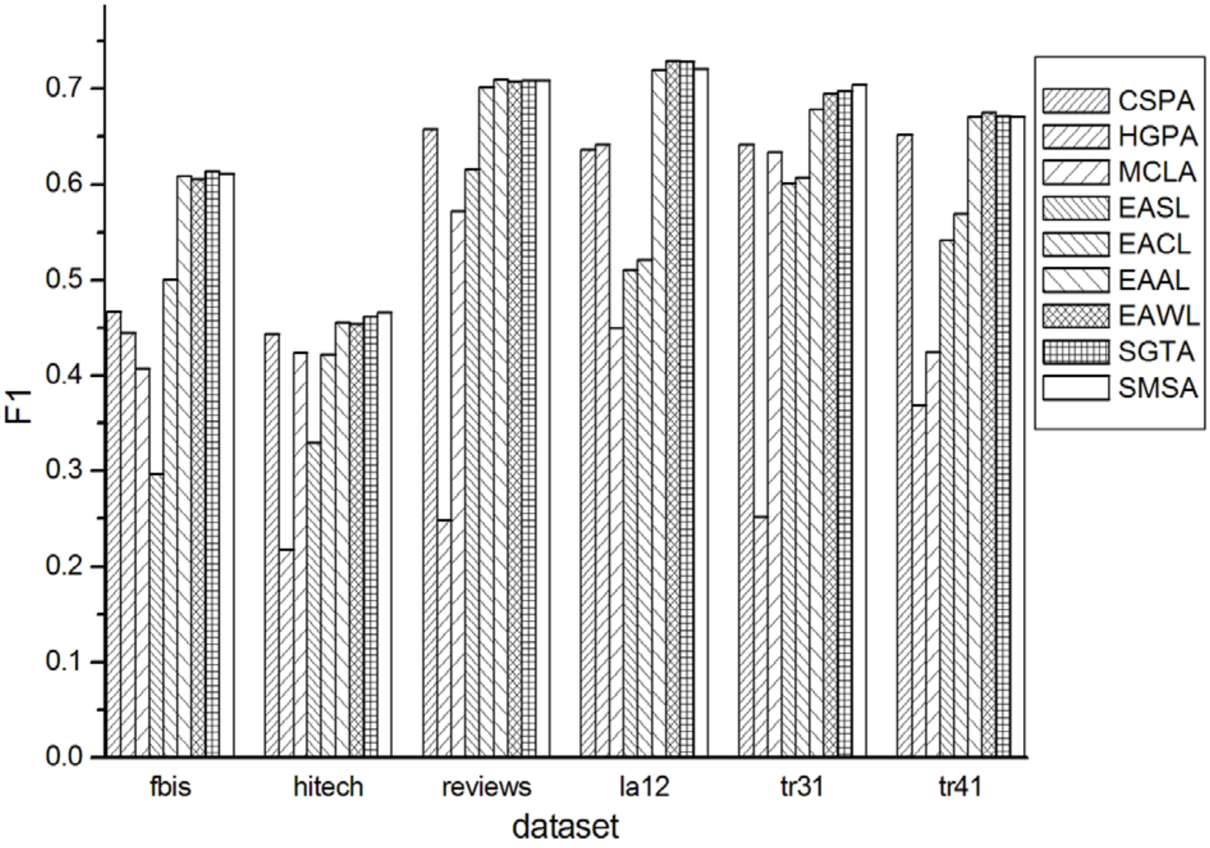

Fig. 2 F1 measures of different clustering integration algorithms

A number of observations can be made by analyzing these results. From Fig. 1, we can see EAAL get the highest score over dataset *fbis* and *hitech*; SGTA get the highest score over dataset *reviews*, *tr31* and *tr41*; SMSA get the highest score over dataset *la12*. In most cases, the above four algorithms outperforms the other five algorithms. From Fig. 2, we can see EAAL get the highest score over dataset *reviews*; EAWL get the highest score over dataset *la12* and *tr41*; SGTA get the highest score over dataset *classic* and *fbis*; SMSA get the highest score over dataset *tr31*. In most cases, the above four algorithms outperforms the other five algorithms.

Conclusion

In this paper we analyze and compare nine different clustering consensus functions. We perform these algorithms on several real-world datasets and the results indicate that different consensus functions will yield rather different results and that there exist some consensus functions which can produce the best overall clustering solutions.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant No. 60975042, 41006057, and 61105057, the 863 Key Program of China under Grant No. 2008AA092701, the Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20070217022 and the Talent Introduction Project of Yancheng Institute of Technology under Grant No. XKR2011019.

References

[1] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley Longman Publishing, Boston, USA, (2010).

[2] A. Strehl, and J. Ghosh, Cluster ensembles - a knowledge reuse framework for combining partitionings, Journal of Machine Learning Research (2002), pp. 583-617.

[3] A. Fred, A. Lourengo, Cluster ensemble methods: from single clusterings to combined solutions [M]. Supervised and Unsupervised Ensemble Methods and their Applications. Berlin: Springer (2008), pp. 3-30.

[4] S. Xu, Z.M. Lu, G.C. Gu, An efficient spectral method for document cluster ensemble, The 9th Intl. Conf. Young Computer Sci. (2008), pp. 808- 813.

[5] S. Xu, Z.M. Lu, G.C. Gu, A fast spectral method to solve document cluster ensemble problem, The 3rd Intl. Multi-Symposiums on Computer and Computational Sci. (2008), pp. 180-183.

[6] Information on http://trec.nist.gov

Published by Atlantis Press, Paris, France.