EPiC Series in Computing

Volume 60, 2019, Pages 120–128

Proceedings of 11th International Conference
on Bioinformatics and Computational Biology

Robust Fuzzy Cluster Ensemble on Cancer Gene Expression Data

Yan Yan, Tin Nguyen, Bobby Bryant, and Frederick C. Harris, Jr.

Department of Computer Science and Engineering,
University of Nevada, Reno
Reno, NV, 89557, U.S.A
yan@cs.unr.edu, tinn@unr.edu, {bdbryant,fred.harris}@cse.unr.edu

Abstract

Noise remains a particularly challenging and ubiquitous problem in cancer gene expression data clustering research, which may cause inaccurate results and mislead the underlying biological meanings. A clustering method that is robust to noise is highly desirable. No one clustering method performs best across all data sets despite a vast number of methods available. Cluster ensemble provides an approach to automatically combine results from multiple clustering methods for improving robustness and accuracy. We have proposed a novel noise robust fuzzy cluster ensemble algorithm. It employs an improved fuzzy clustering approach with different initializations as its base clusterings to avoid or alleviate the effects of noise in data sets. Its results show effective improvements over most examined noisy real cancer gene expression data sets when compared with most evaluated benchmark clustering methods: it is the top performer on three of the eight data sets, more than any other methods evaluated, and it performs well on most of the other data sets. Also, our fuzzy cluster ensemble is robust on highly noisy synthetic data sets. Moreover, it is computationally efficient.

keywords: clustering, cluster ensemble, gene expression data analysis.

1 Introduction

Many clustering methods have been designed and applied to cancer gene expression data for the purpose of cancer classification. They aim to improve therapeutic results by diagnosing cancer types or sub-types with improved accuracy in comparison with traditional methods such as histopathology or immunohistochemistry.

A common and exploratory analysis is to perform clustering on the cancer or patient samples (tissues). Such kind of analysis was first carried out in late 1990s with promising results. In addition, bioinformaticians have proposed novel clustering methods that take intrinsic characteristics of gene expression data into account, such as noise and high-dimensionality, to improve the clustering results. However, different algorithms (or even the same algorithm with different parameters) often provide distinct clusterings. As a result, it is extremely difficult for users to decide which algorithm and parameters will be optimal for a given set of data set for a particular task. There is no single clustering algorithm that can perform the best for all data sets [9],

and discovering all types of cluster shapes and structures presented in data is impossible for any known clustering algorithm [4].

Cluster ensembles have recently emerged as simple and effective methods for improving the robustness and accuracy of clustering results. Cluster ensembles can perform many algorithms on a data set, and integrate the results to find the best clustering.

The paper is organized as follows. In Section 2, we provide related work and a brief literature review. In Section 3, we present our proposed method. Experimental results are detailed in Section 4. Finally, we conclude our study in Section 5.

2 Related Work

Clinical researchers usually use simple traditional clustering methods such as hierarchical [17], K-Means [16], and SOM [6] for cancer gene expression data cluster analysis. Such traditional methods have much better availability in standard software packages and are easy to implement.

Novel clustering methods have been proposed by bioinformaticians to improve the clustering results on gene expression data to address its intrinsic characteristics including noisy and high dimensional, such as Non-negative Matrix Factorization (NMF) method [2]. Such new methods are not getting enough attention from clinical researchers as they may require particular programming environments or more user-specified parameters, which is difficult for non-expert users.

Cluster ensembles combine multiple clustering decisions from base clusterings or ensemble members [13, 15]. There are two main steps in a clustering ensemble: generation step and consensus step. In the generation step, cluster ensemble methods use a variety of approaches to obtain diversity in base clusterings. Four ensemble generation methods have been commonly used: a) using a single clustering algorithm with different initializations [8], b) using multiple clustering algorithms [10], c) using different subsets of genes [1], and d) using data sampling techniques [5].

In the consensus step, cluster ensemble methods use a variety of consensus functions to combine base clusterings. Four ensemble consensus methods have been commonly used: a) using a pairwise similarity-based consensus function [8], b) using a graph-based consensus function [18], c) using a mutual information-based consensus function [20], and d) using voting based consensus function [5].

The noise problem remains particularly challenging in clustering applications, even if there are many pre-processing techniques such as logarithmic transformation or standardization. For bioinformatics applications, noise can make it difficult to detect the true clusters and obscure or mislead the underlying biological meanings.

We present a novel clustering algorithm Improved Fuzzy Cluster Ensemble (IFCE) to improve robustness against noise.

3 Methodology

Our proposed IFCE methodology is illustrated in Figure 1 (adapted from [19]). It includes two major steps: (1) the ensemble generation step generating base clusterings to form a cluster ensemble; and (2) the ensemble consensus step producing the final clustering result using a consensus function.

In the ensemble generation step, diversity is often artificially introduced in order to improve the output results of an ensemble. In homogeneous ensembles, based clusterings are created

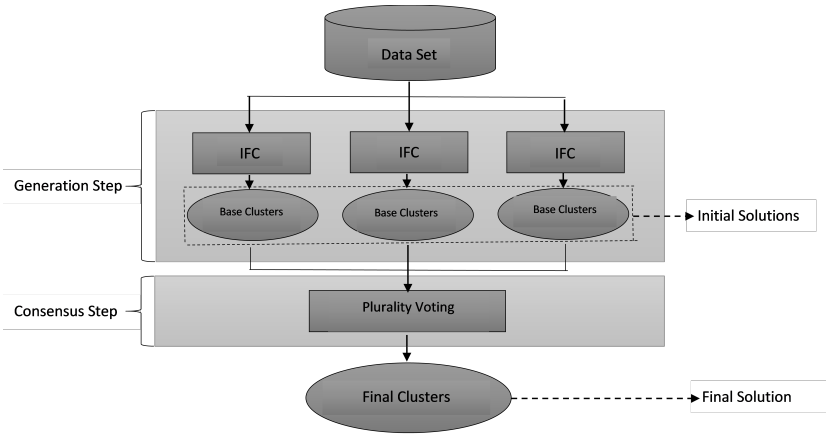

Figure 1: IFCE (adapted from [19])

using a single clustering algorithm. In heterogeneous ensembles, base clusterings are created using different clustering algorithms. IFCE is a homogeneous ensemble, and its three base clusterings are Improved Fuzzy Clustering (IFC) using Modified Weighted Fuzzy Expected Value (MWFEV) [12] with different initializations.

With IFC, first a large number (twice the total existing feature vectors) seed cluster centers are generated in the feature space. Then those that are too close to obtain a reduced but uniformly distributed set of initial seeds are eliminated. For elimination, IFC uses the average distance between centers as a threshold, so that half of the seeds are eliminated. After IFC obtains the initial cluster centers, it applies K-Means clustering (together, it is called Improved K-Means [11]).

After applying the Improved K-Means clustering, we use the MWFEV method for computing the cluster centers. The MWFEV is given by the equation [12]:

$$\vec{\mu}^{(r+1)} = \sum_{p=1}^{P} \alpha_p^{(r)} x_p \tag{1}$$

Where,

$$\alpha_p^{(r)} = \frac{exp[-\frac{(x_(p)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]}{\sum_{m=1}^{P} exp[-\frac{(x_(m)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]} \tag{2}$$

$$(\delta^2)^{(r+1)} = \sum_{p=1}^{P} \alpha_p^{(r)} (x_p - \vec{\mu}^{(r)})^2 \tag{3}$$

After a number of iterations, IFC converges to many relatively small clusters ready for merging. In order to produce more natural shaped clusters as the results instead of forcing them into normed balls due to use of the distance function, IFC merges the closest clusters until the Xie-Beni validity measure does not decrease anymore or until the number of clusters is reduced to two. It finds the two clusters with the minimum distance between their centers, calculates a new center with the average of the two centers. It then reduces the number of clusters by one accordingly [12].

IFCE comprises three IFCs with different initializations as its base clusterings. After diverse clustering results have been produced by the multiple base clustering algorithms, they need to be integrated into a single result. A voting method is commonly used as the consensus function in integrating clustering results for ensemble. IFCE uses plurality voting as its integration function to obtain the final clustering using clustering results from its three base clusterings. With plurality voting, each data object is assigned to one cluster in each base clustering, and the cluster who has more votes (plurality) than any other cluster is the winner. It is different than majority voting, with which the winner has more than half of the votes.

4 Experimental Results

In this section, we first describe our experimental design and settings including data sets, clustering validity measure, and other experimental conditions. We then present the experimental results on clustering criteria comparison analysis, parameters analysis, complexity analysis, and noise robustness analysis.

4.1 Experiment Design and Settings

The goal of our experiments is to compare the performance of IFCE to a number of clustering algorithms. We measure the performance in terms of clustering criteria, parameter sensitivities, complexity, and noise robustness. We choose eight real cancer gene expression data sets and ten synthetic noisy data sets for our comparison experiments.

The eight real cancer gene expression data sets used for experiments are summarized in Table 1. They are filtered data sets from the empirical study of de Souto et al. [3] with uninformative genes removed. They were originally obtained from published microarray studies.

Data Set	Cancer Type	Samples	Genes	Clusters	Chip
Golub1999v1	Leukemia (bone marrow)	72	1,877	2	Affy.
Golub1999v2	Leukemia (bone marrow)	72	1,877	3	Affy.
Amstrong2002	Leukemia (bone marrow)	72	2,194	3	Affy.
Chowdary2006	Breast-Colon Tumors (breast and colon)	104	182	2	Affy.
Nutt2003	Brain Tumor (brain)	50	1,377	4	Affy.
Pomeroy2002	Central Nervous System (brain)	42	1,379	5	Affy.
Chen2002	Hepatocellular Carcinoma (liver)	180	85	2	cDNA
Khan2001	Small, Round Blue-cell Tumors (multi-tissue)	83	1,069	4	cDNA

Table 1: Cancer gene expression data sets

The algorithms used for comparison are as follows: a) four traditional or simple clustering algorithms: KM (K-Means), SL (Single-Linkage), CL (Complete-Linkage), AL (Average-Linkage), and b) six state-of-the-art cluster ensemble methods: MULTI-K, CCHC (Consensus Clustering with Hierarchical Clustering), GCC (Graph-Based Consensus Clustering), CSPA (Cluster-Based Similarity Partitioning Algorithm), HGPA (Hyper-Graph Partitioning Algorithm), MCLA (Meta-Clustering Algorithm).

To evaluate the clustering results of IFCE against the other clustering algorithms, external clustering validity measure of Classification Accuracy (CA) [14] is chosen because of the importance of domain meaningfulness. CA calculates the percentage of accurately clustered data

objects among all data objects clustered. Let Q be the number of total clustered objects, and a be the number of accurately clustered objects.CA is defined by the equation:

$$CA = \frac{a}{Q} * 100\% \tag{4}$$

Higher value of CA means higher clustering accuracy.

CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM from the study of Iam-on et al. (supplementary data) [7] are adopted for evaluating against CA results of IFCE. Cluster labels are available in the data sets, but they are not used in any clustering process in the experiments. Cluster labels are only used to calculate CA after clustering is finished. We run IFCE over the data sets. The outputs include cluster assignment for each data object and the CA value for each run.

For each of the six cluster ensembles chosen to evaluate against IFCE, fixed number of clusters (K) with full space data is used. For IFCE, automatic calculated K with full space data is used. IFCE uses Improved K-Means method that can find K automatically by reducing a large initial K via merging small clusters. IFCE no longer needs a user specified fixed K to perform clustering as many other clustering methods require.

Each clustering method repeats for 50 runs and the average of the CA values is adopted. This approach helps to reduce the effect of stochastic variation with clustering methods and achieve consistency in clustering results.

4.2 Clustering Criteria (CA) Comparison

The CA results of IFCE and other investigated clustering algorithms on real cancer gene expression data sets are presented in Figure 2. As we mentioned earlier, the CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM from the study of Iam-on et al. (supplementary data) [7] are adopted for evaluating against CA results of IFCE. Figure 2 illustrates that IFCE is the top performer on three of the eight data sets, more than any other methods examined. Also, it performs well on most of the other data sets.

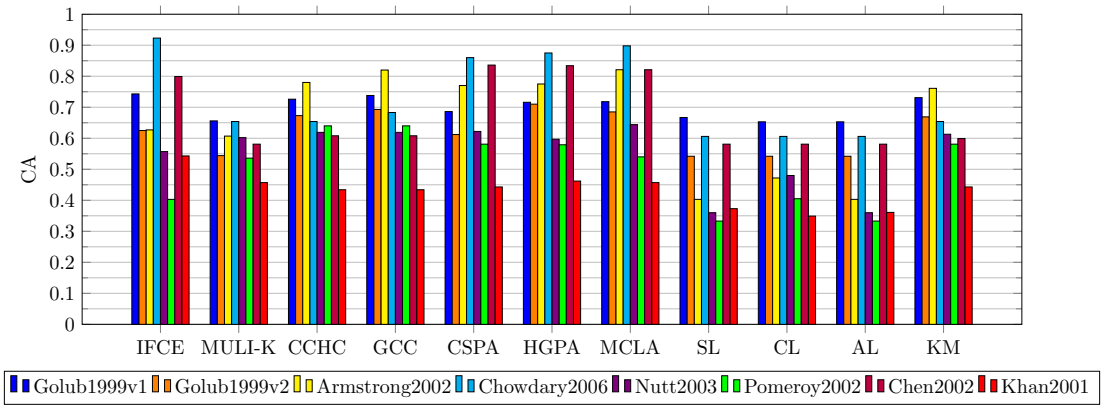

Figure 2: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM across eight real cancer gene expression data sets over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et. al (supplementary data) [7].

4.3 Parameter Analysis

IFCE provides the option of defining the values of multiple parameters. The initial values we define in our experiments are based on empirical experience or general estimation, and they work well on our data sets. However, in an explorative study, it is helpful to experiment with various values. This way, we can assess if IFCE has high degree of dependency on any particular values of some parameters. We can also find out the relation between IFCE and its parameters.

To evaluate IFCE's performance on various parameter values, parameter analysis are examined next. We choose two of the eight data sets for parameter analysis due to page limit constraints. Based on Figure 2, we selected data set Chowdary2006 because IFCE produces the highest CA value on it. In addition, we select data set Chen2002 because IFCE produces one of the relatively average CA values on it.

The first parameter examined is the number of clustering runs N. Smaller number of runs saves computing time, however there may not be enough runs to achieve the desired accuracy due to stochastic variation. A larger number of runs have the potential to increase clustering accuracy, but it increases the expense in run time. Therefore, various values of N are chosen. Results with N = 1, 5, 50, 100, 200 for data sets Chowdary2006 and Chen2002 are presented in Figure 3. Figure 3 also shows that CA values of IFCE are relatively stable with varying N values. It also shows that higher N values usually produce small increase in clustering accuracy.

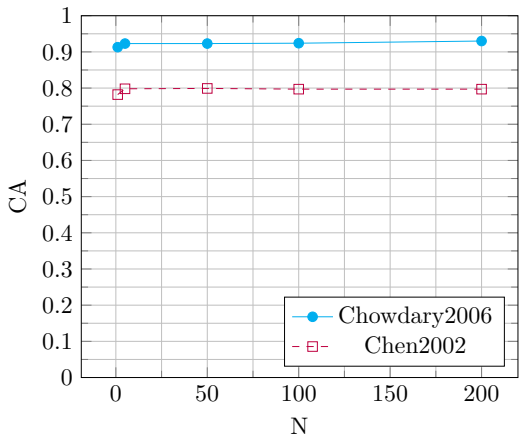

Figure 3: CA of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.

The second parameter examined is the initial merging threshold (IMT) used to merge small clusters. Theoretically, when the IMT is small, clusters with shorter distance in between are merged while clusters with longer distance in between are not. Although the merging threshold increases during next clustering iteration to merge clusters with longer distance, the run time is longer than if we had chosen a larger initial merging threshold. However if we use a larger value as the initial merging threshold, we may risk missing small clusters by merging them at the beginning. Therefore, various values of IMT are chosen. Results with IMT = 1.0, 2.0, 3.0, 4.0 for data sets Chowdary2006 and Chen2002 are presented in Figure 4. Figure 4 also shows that CA values of IFCE change within about 0.008 with varying IMT values.

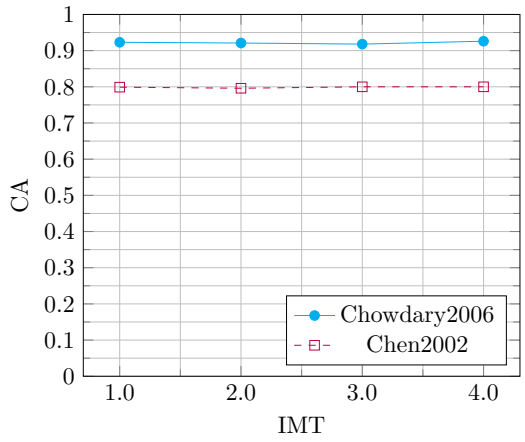

Figure 4: CA of IFCE on Chowdary2006 and Chen2002 with IMT (initial merging threshold) = 1.0, 2.0, 3.0, 4.0.

4.4 Complexity Analysis

To evaluate IFCE's performance on various complexity conditions, time and space complexity analysis are examined next.

Let Q be the number of examples in the data set, N be the number of dimensions in the data set, K be the number of clusters, and M be the number of base clusterings.

When we examine time complexity, IFCE involves three stages. **Stage 1:** initialization with I iterations of Improved K-Means. One vector distance costs $O(N)$, and complexity for KQ distances is $O(KQN)$. Complexity for I iterations is $O(IKQN)$. **Stage 2:** three base clusterings. One vector distance costs $O(N)$, and complexity for KQ distances is $O(KQN)$. Cost for computing the weights for K cluster centers is $O(KQ)$. The complexity for M base clusterings is $O(MKQN)$. **Stage 3:** relabeling and plurality voting ensemble. Relabeling and voting approach is proved to be $O(K^3)$ [5]. The time complexity of IFCE is $O(IKQN)$+$O(MKQN)$+$O(K^3)$. So, IFCE's time complexity converges to $O(N)$.

When we examine space complexity, for each base clustering, the cost of storing a matrix of QxN in memory is $O(QN)$. For M base clusterings, the total cost is $O(MQN)$. The cost of storing relabeling matrix is $O(K^2)$. The space complexity of IFCE is $O(MQN)$+$O(K^2)$, and converges to $O(N)$.

4.5 Noise Robustness Analysis

To examine the boundaries of IFCE's ability to maintain homogeneous clusters under conditions involving high noisy-to-signal ratio data sets, we have created ten synthetic noisy data sets. The clustering process is repeated 50 times on each data set and the resulting clusterings at different noise levels were examined.

The ten synthetic noisy data sets are based on real data sets Chowdary2006 and Chen2002 with increasing noise-to-signal ratios. Noise is incorporated by adding a constant (the maximum value in the gene) to the expression of cancer samples for that gene, such that the percentage of cancer samples with such added noise is 10%, 20%, 30%, 40%, 50%. Such cancer samples represent outliers in the data sets. The performance of IFCE on synthetic noisy data sets are

presented in Figure 5.

Figure 5 shows that IFCE demonstrates robustness to highly noisy data sets. It maintains cluster classification accuracy above 0.870 for Chowdary2006 and above 0.770 for Chen2002 even when signals are reduced by 50%.

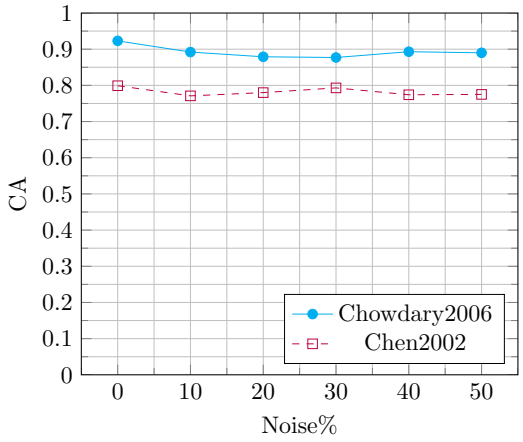

Figure 5: Noise Robustness with noise% = 0%, 10%, 20%, 30%, 40%, 50% added to Chowdary2006 and Chen2002.

5 Conclusion

We have presented a new fuzzy cluster ensemble method IFCE. We have also evaluated IFCE through comparisons with numerous existing benchmark ensemble clustering and simple clustering methods on eight real cancer gene expression data sets. IFCE is the top performer on three of the eight data sets, more than any other methods examined. Also, it performs well on most of the other data sets. IFCE is relatively stable with varying parameter values and is robust to highly noisy synthetic data sets. Moreover, IFCE is computationally efficient.

Acknowledgments

This research is in part supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Avogadri and G. Valentini. Fuzzy ensemble clustering based on random projections for dna microarray data analysis. *Artificial Intelligence in Medicine*, 45(2):173–183, 2009.

[2] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, March 2004.

[3] M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):1–14, 2008. 10.1186/1471-2105-9-497.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001.

[5] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[7] N. Iam-on, T. Boongoen, and S. Garrett. LCE: a link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, 26(12):1513–1519, 2010.

[8] E.-Y. Kim, S.-Y. Kim, D. Ashlock, and D. Nam. Multi-k: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC bioinformatics*, 10(1):260, 2009.

[9] L. I. Kuncheva and S. T. Hadjitodorov. Using diversity in cluster ensembles. *IEEE International Conference on Systems, Man & Cybernetics*, 2:1214–1219, 2004.

[10] L. I. Kuncheva and D. P. Vetrov. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1798 –1808, Nov. 2006.

[11] C. G. Looney. *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*. Oxford University Press, 1997.

[12] C. G. Looney. Interactive clustering and merging with a new fuzzy expected value. *Pattern Recognition*, 35(11):2413 – 2423, 2002.

[13] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen. Pinsplus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, page bty1049, 2018.

[14] N. Nguyen and R. Caruana. Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612, Oct 2007.

[15] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici. A novel approach for data integration and disease subtyping. *Genome research*, 27(12):2025–2039, 2017.

[16] R. Shai, T. Shi, T. J. Kremen, S. Horvath, L. M. Liau, T. F. Cloughesy, P. S. Mischel, and S. F. Nelson. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, 22(4918), 2003.

[17] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003.

[18] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[19] G. Teng, C. He, J. Xiao, Y. He, B. Zhu, and X. Jiang. Cluster ensemble framework based on the group method of data handling. *Appl. Soft Comput.*, 43(C):35–46, June 2016.

[20] A. P. Topchy, A. K. Jain, and W. F. Punch. Combining multiple weak clusterings. In *Proceedings of the IEEE International Conference on Data Mining*, pages 331–338. IEEE Computer Society, 2003.