Experimental Verification of the Quality of Clusterings Produced by Hard Clustering Algorithms After the Removal of Unstable Data Elements

Wim De Mulder*
Department of Electrical Energy,
Systems and Automation
University of Ghent
Ghent, Belgium
Email: wim.demulder@ugent.be

Zahra Zavareh*,
Konika Chawla
and Martin Kuiper
Department of Biology
Norwegian University of Science and Technology
Trondheim, Norway
Email: martin.kuiper@ntnu.no
*These authors contributed equally

***Abstract*—Many different clustering algorithms have been developed to detect structure in data sets in an unsupervised way. As user intervention for these methods should be kept to a minimum, robustness with respect to user-defined initial conditions is of crucial importance. In a previous study, we have shown how the robustness of a hard clustering algorithm can be increased by the removal of what we called unstable data elements. Although robustness is a main characteristic of any clustering tool, the most important feature is still the quality of the produced clusterings. This paper experimentally investigates how the removal of unstable data elements from a data set affects the quality of produced clusterings, as measured by the mutual information index, using three biological gene expression data sets.**

Keywords-hard clustering; cluster quality; unstable elements; mutual information context; microarray data.

I. INTRODUCTION

A. Introduction to cluster analysis

Clustering is an important approach for the analysis of large-sized data sets. Clustering partitions data sets into groups or clusters, such that data elements within the same group share a higher similarity than data elements that are member of different groups. Similarity is typically expressed in terms of a user-defined distance measure.

For large data sets, cluster analysis is often a necessary preprocessing step, since it organizes the data in manageable subsets. The field of bioinformatics widely uses clustering approaches for the analysis of often huge biological data sets with millions of data elements which today can be easily, cheaply and accurately measured with advanced functional genomics technologies (e.g., inferring the expression levels of all genes of an organism on a microarray [1]). However, these so-called high throughput technologies confront the biologist with the daunting challenge of analyzing massive data sets. Yet, for researchers in cluster analysis these data sets offer an interesting alternative to the low dimensional toy data sets that all too often are used to perform cluster analysis experiments on and to validate a certain hypothesized behavior of a clustering algorithm. Here, we focus on gene expression data sets, for which cluster analysis is more often than not a necessity before other specific biological analysis tools can be applied [2].

Since cluster analysis is an unsupervised method, the values for the parameters are typically chosen by simple rules of thumb. Examples of clustering algorithm parameters include the initial cluster centers, in case of k-means [3]; the maximal number of maintained edges, in case of the Memory Constrained-Unweighted Pair Group Method with Arithmetic Mean (MC-UPGMA) clustering algorithm [4]; the fuzzifier, in case of fuzzy c-means [5], etc. Since it is hard to accept that the intrinsic structure of a data set depends on some hit or miss values for these parameters chosen by a user, robustness with respect to these values is of crucial importance. In previous work [6], we introduced the concept of 'unstable data element' and showed that removing such data elements from data increases the robustness in terms of the measure called instability, introduced in the same paper. This previous work is shortly discussed in Section I-B, for convenience. A question we did not consider is how the quality of the result of a clustering algorithm is influenced by the removal of unstable data elements. The theorems we have proven only show that robustness is increased when the most unstable data elements are removed, but they do not exclude the possibility that as a side-effect the quality of the produced clusterings is adversely affected. In informal words, removing unstable data elements implies that a data set can be better clustered by a clustering algorithm, but it is possible that the better separable clusters are worse in terms of cluster validation measures. In this paper, we compare the quality of clusterings, produced by k-means, before and after the removal of unstable data elements, using three relatively large biological data sets that describe the activity of genes from an organism. The quality is measured using the mutual information index, a theoretically well-founded measure that is often used as cluster validation measure if external labels are available [7], [8], [9], [10].

The paper is organized as follows. Section I-B outlines our previous work. In Section II, we describe the three biological data sets that are used to investigate the research question mentioned above. Section II-C explains how the mutual information index can be used as cluster validation measure if gene annotations are available. In Section II-D, we recall from our

 ISBN: 978-1-61208-335-3

previous work what we mean by the most robust clustering from a given sample of clusterings. Section III contains the experiments where we compare the quality of the most robust clustering of the data set after removal of unstable genes to clusterings of the data set before the removal of such genes.

B. Previous work

1) Introductory notions: The concepts and methods discussed in our previous work [6] apply to hard clustering algorithms. Such algorithms produce hard clusterings, meaning that every element is member of exactly one cluster, to full degree. K-means is the best known example of such an algorithm.

Given a data set $D = \{g_1, \ldots, g_n\}$, any hard clustering can be represented as a matrix C with elements $C(j,k), j = 1...n, k = 1...n$:

$$\begin{aligned} C(j,k) &= 1 \quad \text{if } g_j \text{ and } g_k \text{ are placed in different clusters} \\ &= 0 \quad \text{if } g_j \text{ and } g_k \text{ are placed in the same cluster} \end{aligned}$$

We defined the expected clustering $E[C]$ as the matrix that contains as elements $E[C](j,k) = E[C(j,k)]$, where the expected value is taken over all hard clusterings of the data produced by a given hard clustering algorithm and where randomness arises from the random selection of initial conditions. This matrix can be considered as independent of any specific choice of initial conditions, and thus maximally robust, since it is the uniquely defined probability-weighted sum over all possible clusterings generated by the given clustering algorithm. It is clear that the expected clustering is only a theoretical concept, i.e., it cannot be determined in practice. In practice, a sample of clusterings $\{C_1, \ldots, C_N\}$ is generated and the expected clustering is approximated by the average clustering $\bar{C}$ with elements $\bar{C}(j,k) = \frac{1}{N}\sum_{i=1}^{N} C_i(j,k)$.

2) Instability: We introduced the instability of a data element g_k:

$$\mu(g_k) = \frac{1}{n-1}\Big(\sum_{j=1}^{k-1}\sigma(\bar{C}(j,k)) + \sum_{j=k+1}^{n}\sigma(\bar{C}(k,j))\Big) \quad (1)$$

with

$$\begin{aligned} \sigma(a) &= 1-a \quad 0.5 \le a \le 1 \\ &= a \quad\quad 0 \le a < 0.5 \end{aligned}$$

for $a \in [0,1]$.
We define the instability of a given clustering algorithm for a data set, as

$$\mu = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j<k\le n}\sigma(E[C](j,k)) \quad (2)$$

In practice, the instability is approximated using the average clustering $\bar{C}$ corresponding to a sample of clusterings generated by the given clustering algorithm, for a data set, as follows:

$$\hat{\mu} = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j<k\le n}\sigma(\bar{C}(j,k)) \quad (3)$$

For convenience, we will write μ instead of $\hat{\mu}$. The intuition behind the instability of a given clustering algorithm, for a given data set, is that it represents a measure for the difference between clusterings generated with different initial conditions. As such, it is an inverse measure for robustness. The concept of instability is more extensively described in [6].
It was proven that the instability of a clustering algorithm equals the average instability of the data elements:

Theorem 1.

$$\frac{1}{n}\sum_{k=1}^{n}\mu(g_k) = \mu$$

3) Cluster stability variance: We extended the variance of a random variable taking values on $\mathbb{R}$ to the variance of a hard clustering algorithm C, for a given data set: $\sigma^2(C) = E[d(C, E[C])^2]$ where $d(C, E[C])$ denotes the 'distance' from C to $E[C]$ which we defined as:

$$d(C, E[C]) = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j<k\le n}|C(j,k) - E[C](j,k)| \quad (4)$$

Randomness arises from the random choice of initial conditions. In practice, the variance is approximated by what we called the cluster stability variance (CSV) using a sample of clusterings $\{C_1, \ldots, C_N\}$:

$$CSV = \frac{1}{N-1}\sum_{i=1}^{N} d(C_i, \bar{C})^2 \quad (5)$$

The CSV is at least zero, and it is only zero when the produced clusterings are independent of the choice of initial conditions. The larger the CSV, the more dependent on initial conditions the produced clusterings are, for a given data set. In other words, the CSV is also an inverse measure for robustness.

4) Relationship between instability and cluster stability variance: We proved the following relationship between instability and CSV, given a sample of N clusterings:

Theorem 2.

$$CSV \le \frac{N}{N-1}\,\mu$$

5) Reducing the instability: We showed that the instability is reduced by removing the most unstable data element from the data set:

Theorem 3.

$$\mu(g_l) = \max\{\mu(g_k) | 1 \le k \le n\} \Rightarrow \Delta\mu_l \le 0$$

where $\Delta\mu_l$ represents the change in instability after removing g_l.
In other words, the instability of a clustering algorithm on any data can be increased by removing the most unstable data element. Due to Theorem 2 the CSV is also possibly reduced as a side effect. This process of removing the most unstable

 ISBN: 978-1-61208-335-3

data element can be repeated until the CSV attains a minimum or has stabilized.

6) Novelty of the described method: Our work has introduced two concepts, namely the instability of a data element and the cluster stability variance. The instability of a data element refers, loosely speaking, to the uncertainty about the cluster to which this element should be assigned. Stated another way, a data element has a high instability if the considered clustering algorithm is not able to reliably assign it to a cluster. From the instability of a data element, we have defined the instability of a given hard clustering algorithm. The cluster stability variance can be interpreted as an inverse measure for the robustness of a clustering algorithm, for a given data set.

We have introduced theorems that show how the instability of a given hard clustering algorithm can be reduced, i.e., by removing appropriate unstable data elements.

II. METHODS

A. Data sets

Three biological data sets with measurements of the level of expression of genes are used for our experimental study. Briefly, genes are segments of the genome of an organism, encoding the functional components (most often proteins) of that organism. The first step in the decoding of this information is the production of messenger ribonucleic acid (mRNA) molecules from these genes, the quantity of which provides information about the activity of said genes. Measuring the quantity of all mRNAs of all genes is a common approach in modern biology, and the ensuing data is called 'gene expression data set'. In our assessment of clustering performance we chose gene expression data sets that contain a series of measurements (essentially constituting a data vector for genes) covering a certain time range (time series experiment), usually spanning several hours after a particular stimulus, with measuring points some minutes or hours apart. Experiments of this type usually include a control group not undergoing the stimulus, allowing to express the observed expression as a ratio relative to a control. We preprocessed these data to get unique genes with significant expression values.

1) Rat data set: The first data set represents the gene expression response to a stimulus by the stomach hormone gastrin, a data set produced by Selvik et. al. [11], and available at the Gene Expression Omnibus (GEO) database [12] (accession ID GSE32869). This data set concerns a time series experiment on a cell line obtained from rats. Following treatment with gastrin, cells were sampled at 11 intervals during a 14 hour period to record how expression responses evolve over time. The experiment was done twice to obtain more robust data (average of replicates). The time series data was analyzed using an extended dimension reduction framework for significance analysis of gene expression data as described [13]. The extended framework uses partial least squares regression (PLS) in combination with a priori defined time curves based on hypothesized network motifs (unpublished data). This resulted in 2292 genes that were differential expressed in response to gastrin, while also displaying a relatively smooth profile. This data set is further referred to as the Rat data set.

2) Human data set: The second dataset is a time series gene expression analysis based on human breast cancer cells stimulated by the growth hormone epidermal growth factor [14]. These data were obtained from the GEO database (accession ID GSE13009). Although the data covers a 72 hour period, we considered the first 14 time points covering 24 hours. The data was preprocessed to select genes that were significantly affected by the hormone stimulus. This involved normalization by the Robust Multi-array Average (RMA) method [15] and filtering for high variation using the Genefilter package [16]. This resulted in 2194 genes representing the most significant variation in the data set. We refer to this data set as the Human data set.

3) Yeast data set: The third data set was produced for identification of genes showing activity changes during the process of cell division in the yeast Schizosaccharomyces pombe [17]. This data set is freely available [18]. We chose a subset of these data that had the lowest number of missing values, named 'elutriation 3'. We could link 374 of the 407 cell cycle related genes in this set through their systematic IDs but filtered out an additional 118 genes because they had a high incidence of missing values in their data vectors. The result is a data set containing expression profiles for 256 genes covering 20 time points. This data set is referred to as the Yeast data set.

B. Clustering algorithm

Throughout this paper k-means is used as clustering algorithm. The reason is that the application of our method is restricted to hard clusterings, as described in Section I-B1. In a hard clustering it holds that any two different elements either belong to the same cluster or belong to different clusters. K-means is the natural choice to produce such clusterings. Since our method does not refer to the hard clustering algorithm that produces the hard clusterings, it can equally well be applied to hard clusterings generated by other algorithms (e.g., for example, hard clustering algorithms based on multiple dissimilarity matrices [19]). We leave the application to hard clusterings produced by other algorithms than k-means to future research.

C. Mutual information index as cluster validation measure

1) Attribute matrix: The mutual information index is often used to validate clusterings, provided that external labels for qualifying the cluster elements are available. For gene expression data sets, we can rely on publicly available databases containing gene ontology (GO) annotations to be used as the external labels [20]. Gene ontology annotations are standardized terms that biological experts use to functionally describe various qualities of a gene such as their molecular function, biological process and cellular location that can be attributed to them. These qualifications represent attributes that can be collected for many genes through the BiomaRt package [21] of the Bioconductor analysis software [22]. Associations of the genes with these unique attributes are then represented by an attribute matrix T [23] such that if gene i is annotated with attribute j, we define $T(i, j) = 1$, otherwise $T(i, j) = 0$. We built attribute matrices for each data set.

Subsequently, the filtering method described in [24] was

 ISBN: 978-1-61208-335-3

applied. More concretely, the following type of genes and attributes were removed from the data:

- Unannotated genes, since they were not associated with any attribute (GO term).
- Attributes (GO terms) associated with fewer than 10 genes as these were considered to be not informative enough.

This procedure reduced the Rat data set to 1776 genes with a total of 257 different attributes, the Human data set to 1818 genes with 517 attributes and the Yeast data set to 253 genes with 30 attributes.

2) Calculation of mutual information index for clusterings of gene expression data sets: Given the attribute matrix, the mutual information index for a clustering C, containing clusters $C_1, \ldots, C_m$, is calculated in several steps.
First, we calculate the entropy $H(C)$:

$$H(C) = -\sum_{i=1}^{m} p(C_i) \log p(C_i) \tag{6}$$

where $p(C_i)$ denotes the probability of a gene belonging to C_i, which we approximate by the number of genes belonging to cluster C_i divided by the total number of clustered genes.
Secondly, the entropy of all attributes A_j is calculated. This is defined similarly as $H(A_j) = -p(A_j) \log p(A_j)$, where $p(A_j)$ is now estimated as the number of genes g_i for which $T(i,j) = 1$ divided by the total number of clustered genes.
Thirdly, we calculate the joint entropy $H(C, A_j)$ between the clustering C and each of the attributes A_j, defined as $H(C, A_j) = -\sum_{i=1}^{m} p(C_i, A_j) \log p(C_i, A_j)$, where $p(C_i, A_j)$ is the probability that a gene has attribute A_j and belongs to cluster C_i, estimated as the fraction of genes belonging to C_i and at the same time having attribute A_j, i.e., such that $\sum_i \sum_j p(C_i, A_j) = 1$.
Fourthly, the mutual information between C and each of the attributes A_j is defined as

$$I(C, A_j) = H(C) + H(A_j) - H(C, A_j) \tag{7}$$

Finally, the mutual information index is calculated as $\sum_j I(C, A_j)$.
The intuition behind the mutual information index is that it measures the degree to which the clustering of the genes is consistent with the information known about these genes as represented by the attribute matrix.

D. Most robust clustering from a sample of clusterings

In Section I-B1, and especially in our previous work [6], it was argued that the expected clustering can be considered as maximally robust. However, the expected clustering is only a theoretical concept and cannot be determined in practice. The average clustering is an approximation for the expected clustering, and thus can be considered as very robust, provided that the corresponding sample of clusterings is large enough. The problem is that the average clustering $\bar{C}$ is not a hard clustering, because it typically contains elements $\bar{C}(j,k)$ different from 0 and 1. This implies that the mutual information index cannot be calculated for $\bar{C}$, since this requires to calculate $H(\bar{C})$, which in turn requires to determine the number of genes belonging to each cluster. However, this number can only be determined for a hard clustering. A good candidate robust clustering is the clustering from the given sample of clusterings that is closest to the average clustering, in terms of the distance measure (4). That is, we define as most robust clustering, given a sample $\{C_1, \ldots, C_N\}$, the clustering C for which it holds that

$$C = \arg \min_{1 \le i \le N} \{d(C_i, \bar{C})\} \tag{8}$$

Evidently, this clustering is hard and it is legitimate to consider it as the most robust one among all clusterings belonging to the given sample.

III. Discussion

A. Experimental setup

1) Determination of the parameters: K-means was used with each data set to produce 500 clusterings with randomly chosen initial centers. From these samples of clusterings, the average clustering could be calculated, and this in turn allowed to calculate the instability of each gene, the instability of the clustering algorithm and the CSV (see Section I-B). The similarity measure was chosen as the correlation distance, since it has been argued that this measure is well suited to measure the coexpression between genes [25]. The number of clusters produced with k-means for the different Rat, Human and Yeast data sets was set to 9, 6 and 11, respectively. Determining the optimal number of clusters for a biological data set is not an exact science; many cluster analysis experts even argue that there does not exist something as 'the' optimal number of clusters. Rather, determining the optimal number of clusters is the subject of a continuing debate, and different optimization methods typically declare different numbers as being the optimal number of clusters. Therefore, we used a heuristic approach consisting of a visual inspection of the types of genes that are distributed over different clusters, while imposing a range of cluster numbers on k-means. This allowed us to set the above numbers based on the approximate distribution of genes over different clusters *vis a vis* the types forced together within single clusters, essentially checking the biological plausibility of separating genes or lumping them together. It is important to note, however, that this paper is not about defining the optimum numbers of clusters, but rather about the quality of clusterings before and after removal of unstable elements. The above mentioned numbers of clusters were kept fixed for all 500 clusterings of the different data sets.

2) Research topic: comparison of the quality of clusterings of the original data set with the quality of the most robust clustering of the reduced data set: As outlined above, our goal is to compare the quality of clusterings, as measured by the mutual information index, before and after the removal of unstable genes. This is done as follows. First, 500 clusterings are generated using k-means and their mutual information index is calculated as described in Section II-C2. Secondly, unstable genes are detected and removed from the data set until the CSV appears to reach a minimum or has stabilized (see Section I-B5). Thirdly, k-means is now applied on the reduced data set to produce again 500 clusterings and the mutual information

 ISBN: 978-1-61208-335-3

index is calculated for the most robust clustering (see Section II-D). The obtained 501 mutual information indices are plotted in a histogram, and the research question considered is how the mutual information index of the most robust clustering (of the reduced data set) compares to the mutual information indices of the clusterings of the original data set. This procedure is repeated for each of the three data sets.

B. Experimental results

For each data set we plot the mutual information indices as outlined above, and the CSV and instability after elimination of the most unstable gene (repeating this procedure until the CSV appears to have reached a minimum or has stabilized).

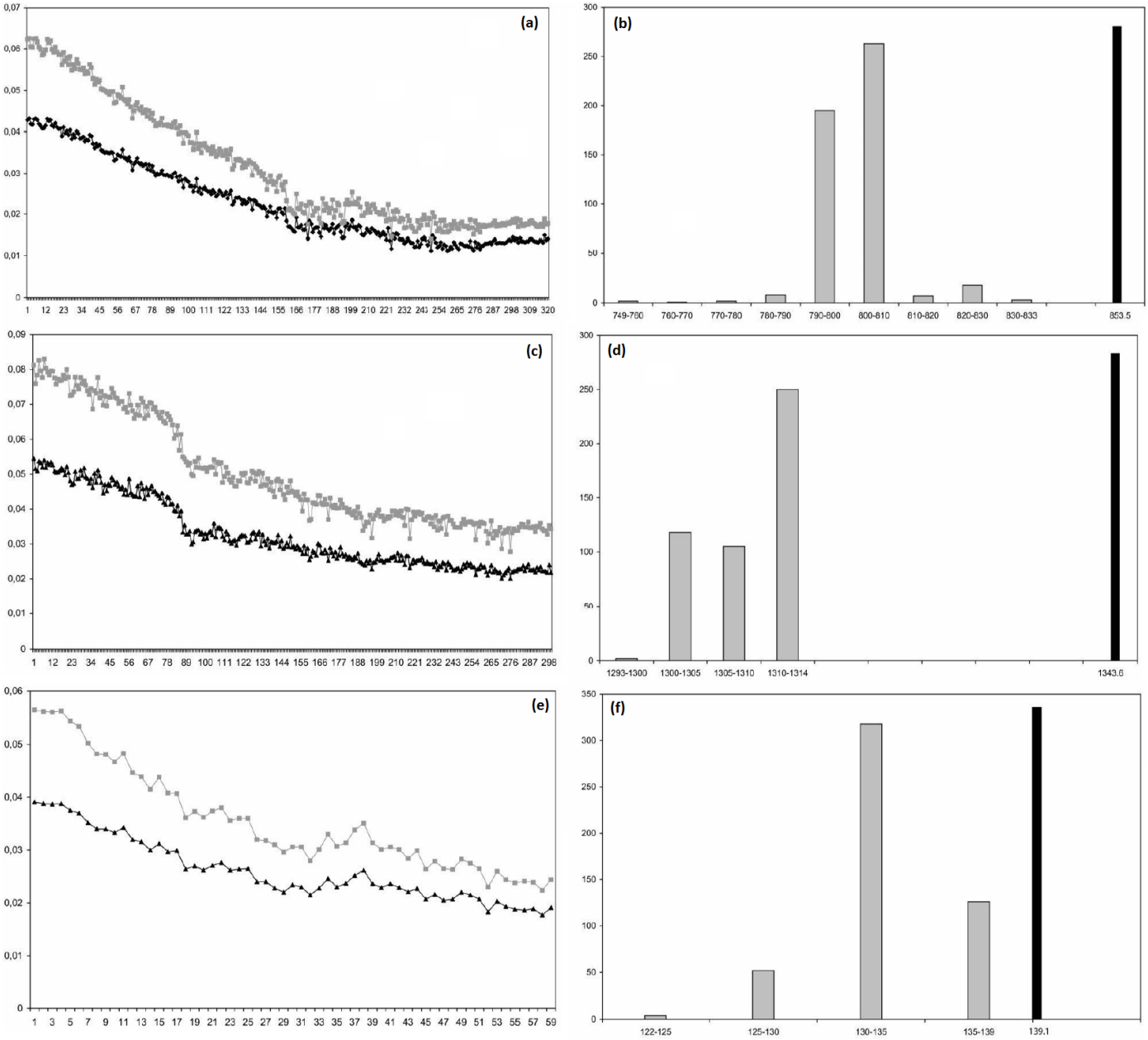

Figure 1: Left column (a,c,e): CSV (black) and instability (grey), the x-axis denotes the number of removed genes; Right Column (b,d,f): Mutual information indices for clusterings of original data set (grey) and of reduced data set (black), Y-axis denotes the number of clusterings, de x-axis denotes the MI; (a,b) Rat data set; (c,d) Human data set; (e,f) Yeast data set.

 ISBN: 978-1-61208-335-3

It will be noticed that the CSV is smaller than the instability, in accordance with Theorem 2. Theorem 3 states that the instability decreases after the removal of an unstable gene, although in the figures below one will see that although the general trend of the instability is decreasing, the instability sometimes *increases* after the current most unstable gene has been deleted. The explanation for this phenomenon is simply that practical restrictions in terms of time and memory force us to work with samples of clusterings rather than with the set of all possible clusterings as implicitly assumed by the theorems above.

1) Rat data set: The CSV, the instability and the mutual information indices for the Rat data set are shown in Fig 1a and Fig 1b. The number of genes that we removed was chosen rather heuristically, since we have to find a compromise between the desire to make the clustering algorithm as robust as possible (i.e., removing all unstable genes) and to limit the number of removed genes, since we want to end up with a clustering that contains a significant number of the genes from the original data set. We decided to remove 174 genes, since the sharpest decrease in the CSV appears up to this number of genes. This amounts to removing about 10% of all genes. Fig 1b shows the mutual information indices. The quality of the most robust clustering of the reduced data set (i.e., after removing the 174 genes) is significantly higher than the quality of any clustering of the original data set.

2) Human data set: The results for the Human data set are shown in Fig 1c and Fig 1d. We chose to remove 270 genes, since the CSV appears to stabilize after that number. This amounts to about 15% of all genes. The difference in quality between the most robust clustering of the reduced data set and the qualities of the clusterings of the original data set is striking.

3) Yeast data set: Fig 1e and Fig 1f display the results for the Yeast data set. Thirty-one unstable genes are removed, since the sharpest decrease in the CSV appears up to this number. In relative terms, about 12% of all genes are taken out of the given data set. The mutual information index of the most robust clustering of the reduced data set is higher than that of any of the clusterings of the original data set.

IV. Conclusion and future work

In previous work, we introduced the concepts 'instability' and 'cluster stability' variance as inverse measures of the robustness of a hard clustering algorithm, with respect to initial conditions. We showed how removing unstable data elements from a data set increases the robustness of the clustering algorithm. A question we did not consider is how the removal of such unstable data elements affects the quality of the produced clusterings. Although the reduced data set can be better clustered by a clustering algorithm in the sense that this clustering algorithm is able to recognize a more definite structure in the data set, irrespective of the initial conditions, it is possible that the produced clusterings are of a lower quality, meaning that the recognized structure is further away from the real structure. In this paper, the quality of clusterings of the original data set is compared to the quality of the most robust clustering of the reduced data set, i.e., after removing unstable genes, performed on three authentic biological gene expression data sets, where the quality is measured in terms of the mutual information index. Although our hope was that the quality of the most robust clustering of the reduced data set would be higher than that of most clusterings of the original data set, to our surprise it turned out that the robust clustering of the pruned data set significantly outperforms *all* clusterings of the original data set. The main conclusion that we therefore draw from this and our previous work is that it is beneficial to detect and remove unstable data elements from a data set, both in terms of robustness of the clustering algorithm with respect to initial conditions and in terms of the quality of the generated clusterings.
As future work, we plan to apply our method to hard clusterings produced by other hard clustering algorithms than k-means.

Acknowledgment

The authors would like to thank Endre Anderssen, Arnar Flatberg and Torunn Bruland from NTNU for their help in processing and filtering the Rat data set. These data were provided by the Genomics Core Facility (GCF), of the Norwegian University of Science and Technology (NTNU). GCF is funded by the Faculty of Medicine at NTNU and the Central Norway Regional Health Authority.

References

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," Science, vol. 270, pp. 467-470, Oct. 1995, doi: 10.1126/science.270.5235.467.

[2] D. Jiang, C. Tang and A. Zhang, "Cluster analysis for gene expression data: a survey," Transactions on Knowledge and Data Engineering, vol. 16, pp. 1370-1386, Nov. 2004.

[3] A. Alrabea, A.V. Senthilkumar, H. Al-Shalabi and A. Bader, "Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with PCA," Journal of Advances in Computer Networks, vol. 1, pp. 137-142, Jun. 2013, doi: 10.7763/JACN.2013.V1.28.

[4] Y. Loewenstein1, E. Portugaly, M. Fromer and M. Linial, "Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space," Bioinformatics, vol. 24, i41-i49., Jul. 2008, doi: 10.1093/bioinformatics/btn174.

[5] R. Winkler, F. Klawonn and R. Kruse, "Fuzzy clustering with polynomial fuzzifier function in connnection with M-estimators," Applied and Computational Mathematics, vol. 10, pp. 146-163, 2011.

[6] W.D. Mulder, M. Kuiper and R. Boel, "Clustering of gene expression profiles: creating initialization-independent clusterings by eliminating unstable genes," Journal of Integrative Bioinformatics, vol. 7, Mar. 2010, doi: 10.2390/biecoll-jib-2010-134.

[7] J.M. Buhmann, "Information theoretic model validation for clustering," International Symposium on Information Theory, pp. 1398-1402, Jun. 2010, arXiv:1006.0375.

[8] S.A. Fattah, C.-C. Lin and S.-Y. Kung, "A mutual information based approach for evaluating the quality of clustering," IEEE International Conference on Accoustics, pp. 601-604, May 2011, ISSN: 1520-6149.

[9] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," Journal of Machine Learning Research, vol. 3, pp. 583-617, Apr. 2003.

[10] N.X. Vinh, J. Epps and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," Journal of Machine Learning Research, vol. 11, pp. 2837-2854, Oct. 2010.

[11] L.-K. Selvik, C. Fjeldbo, A. Flatberg et al., "The duration of gastrin treatment affects global gene expression and molecular responses involved in ER stress and anti-apoptosis," BMC Genomics, vol. 14, Jun. 2013, doi:10.1186/1471-2164-14-429.

 ISBN: 978-1-61208-335-3

[12] T. Barrett and R. Edgar, "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis," Methods Enzymol., vol. 411, pp. 325-369, Oct. 2006, doi: 10.1016/S0076-6879(06)11019-8.

[13] L. Gidskehaug, E. Anderssen, A. Flatberg, and B.K. Alsberg, "A framework for significance analysis of gene expression data using dimension reduction methods," BMC Bioinformatics, vol. 8, Sept. 2007, doi:10.1186/1471-2105-8-346.

[14] Y. Saeki, T. Endo, K. Ide et al., "Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells," BMC Genomics, vol. 10, Nov. 2009, doi:10.1186/1471-2164-10-545.

[15] R.A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," Biostatistics, vol. 4, pp. 249-264, Apr. 2003.

[16] http://www.bioconductor.org/packages/2.3/bioc/html/genefilter.html

[17] G. Rustici, J. Mata, K. Kivinen et al., "Periodic gene expression program of the fission yeast cell cycle," Nature Genetics, vol. 36, pp. 809-817, Aug. 2004.

[18] http://www.bahlerlab.info/projects/cellcycle/

[19] F. de A.T. de Carvalho, Y. Lechevallier and F.M. de Melo, "Partitioning hard clustering algorithms based on multiple dissimilarity matrices," Pattern Recognition, vol. 45, pp. 447-464, Jan. 2012.

[20] F.D. Gibbons and F.P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," Genome Research, vol. 12, pp. 1574-1581, Oct. 2002.

[21] S. Durinck, Y. Moreau, A. Kasprzyk et al., "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis," Bioinformatics, vol. 21, pp. 3439-3440, Aug. 2005.

[22] R.C. Gentleman, V.J. Carey, D.M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," Genome Biol., vol. 5, Sept. 2004.

[23] M. Ashburner, C.A. Ball, J.A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nature Genetics, vol. 25, pp. 25-29, May 2000.

[24] R. Steuer, P. Humburg and J. Selbig, "Validation and functional annotation of expression-based clusters based on gene ontology," BMC Bioinformatics, vol. 7, Aug. 2006, doi: 10.1186/1471-2105-7-380.

[25] L.J. Heyer, S. Kruglyak and S. Yooseph, "Exploring expression data: identification and analysis of coexpressed genes," Genome Research, vol. 9, pp. 1106-1115, 1999, doi: 10.1101/gr.9.11.1106.

 ISBN: 978-1-61208-335-3