STABILITY-BASED CLUSTER ANALYSIS APPLIED TO MICROARRAY DATA

Ciprian Doru Giurcăneanu, Ioan Tăbuş

Institute of Signal Processing,
Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland

Ilya Shmulevich, Wei Zhang

Cancer Genomics Lab., Dep. of Pathology,
The University of Texas,
M.D. Anderson Cancer Center
Houston, TX 77030, USA

ABSTRACT

This paper studies the estimation of the number of clusters using the so-called stability-based approach, where clusters obtained for two subsets of the dataset are compared via a similarity index and the decision regarding the number of clusters is taken based on the statistics of the index over randomly selected subsets. We introduce a new similarity index $s(\cdot,\cdot)$, and analyze the consistency of the estimator of the number of classes when k-means algorithm is used in conjunction with $s(\cdot,\cdot)$. Various similarity indices are experimentally evaluated when comparing the "true" data partition with the partition obtained at each level of a hierarchical clustering tree. Finally, experimental results with real data are reported for a glioma microarray dataset.

1. INTRODUCTION

Two different approaches have been considered when applying the stability-based methods for finding structure in microarray data:
(1) After randomly splitting the dataset into two subsets, select one subset for learning, and the other one for testing. Firstly a clustering algorithm C_A is applied to the learning set, and the resulting classes are used to classify the samples which belong to the test set. Then the test set is clustered with the same algorithm C_A, and a similarity measure (index) is computed between the labels produced by classification and clustering, respectively [1, 2, 3];
(2) Apply the same clustering algorithm C_A to both subsets and calculate the similarity index on the samples belonging to the intersection of subsets [4]. A modified variant is introduced in [5]: C_A is applied to the whole dataset (reference clustering), and to a randomly chosen subset. The similarity index is computed for the samples contained in the selected subset.

In both cases, it is assumed that the number of clusters k belongs to $\{2, 3, \ldots, k_{max}\}$, and for each value allowed for k, after running many times the algorithm, the empirical distribution of the similarity index is collected. While various methods were proposed for estimating the number of clusters from the shape of the collected empirical distribution, less attention was paid to the selection of the clustering algorithm and of similarity index. We investigate the impact of various partitioning and hierarchical clustering algorithms when used in conjunction with well-known similarity indices like Fowlkes-Mallows, Jaccard, Rand [6], and Rand$_{HA}$ [7].

The definition for the partition-distance $D(\cdot,\cdot)$ is introduced in [8]: for any two partitions P and P' of an N object set T, $D(P,P')$ is the minimum number of elements that must be deleted from T, so that P and P' restricted to the remaining elements are identical. An *assignment* is defined as a selection of entries of the contingency matrix M such that no row or column contains more than one selected entry, and is called *optimal* when the sum of the selected cell values is the largest over all possible assignments [9]. Let $A(P,P')$ denote the value of the optimal assignment for the contingency matrix of partitions P and P'. An important result was proven in [9] stating that the partition distances and the assignments satisfy $D(P,P') = N - A(P,P')$, and the elements to be removed from T to induce identical partitions on P and P' are all those objects associated with the cells not selected in the optimal assignment. We define a new index of similarity between any two partitions, P and P', as follows: $s(P,P') \triangleq 1 - \frac{D(P,P')}{N-1} = \frac{A(P,P')-1}{N-1}$. It is a measure of similarity, ranging from $s(P,P') = 0$ when the two partitions have no similarities (i.e., when one consists of a single cluster and the other only of clusters containing single objects), to $s(P,P') = 1$ when the partitions are identical.

The paper is organized as follows. In Section 2 we investigate analytically the stability-based approach for estimating the number of clusters when k-means algorithm is used in conjunction with the similarity index $s(\cdot,\cdot)$. In Section 3 various similarity indices are experimentally evaluated when assume that the "true" structure of the data (the number of clusters and the membership) is known, and compare this partition with the partition obtained at each level of a hierarchical clustering tree. Experimental results are reported in Section 4 for a glioma dataset.

2. ON THE CONSISTENCY OF THE STABILITY-BASED ESTIMATOR

We consider the simple case of well-separated clusters used in [3] to give a theoretical justification for the prediction strength of the method: a distribution that is spread uniformly over K unit balls in p-dimensional space ($p > 1$): $B(\underline{a}_1)$, $B(\underline{a}_2)$, ..., $B(\underline{a}_K)$ where $\underline{a}_i$ denotes the center of the i-th ball, or equivalently the mean of the i-th population. Assume that the distance between any two different centers $\underline{a}_i$ and $\underline{a}_j$ is not smaller than four. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N$ be a sample from this distribution. Following the same procedure as in [1, 2, 3] we randomly split the sample into two sets: one used for learning, and the other one for testing. For simplicity we analyze the case when the cardinality of each set is $N/2$. Hypothesizing that k different clusters exist in the observed data, the centroids $\{\underline{\hat{a}_i}^{tr} : 1 \leq i \leq k\}$ and $\{\underline{\hat{a}_i}^{te} : 1 \leq i \leq k\}$ are found after running the k-means algorithm over the learning

C.D. Giurcăneanu and I. Tăbuş were supported by Academy of Finland.

0-7803-7946-2/03/$17.00 ©2003 IEEE

set, respectively the test set. The centroids determined during the learning step induce a partition on the test set. Let us denote $s(k)$ the similarity index computed between this partition and the partition obtained during the test step. Repeat the sample splitting N_t times, and let $\bar{s}(k)$ be the average of $s(k)$ over all splittings. We make the hypothesis that the cardinality of the sample set is large ($N \to \infty$), and also $N_t \to \infty$.

Proposition 2.1 *Under the hypotheses listed above, $\bar{s}(k)$ verifies*

$$\begin{aligned} k = K &\quad : \quad \bar{s}(k) = 1 + o_p(1), \\ k > K &\quad : \quad \bar{s}(k) \leq 1 - \frac{1}{8}\frac{1}{K} + o_p(1) \end{aligned}$$

where $o_p(\cdot)$ is employed for the stochastic order symbol.

The proof is deferred to the Appendix. According to the Proposition, we expect that $\bar{s}(k)$ is close to 1 when $k \leq K$, and has a sharp drop at $k = K + 1$. This allows us to estimate K by checking at which k the average $\bar{s}(k)$ significantly drops below 1. Note that the sharpness of the transition at K decreases with K, which makes the method most suitable for the cases when there exist a priori knowledge that K is relatively small.

3. MEASURING SIMILARITY IN HIERARCHICAL CLUSTER ANALYSIS

In this section we want to understand the behavior of similarity indices and hence we assume that the "true" structure of the data is known, and compare this partition with the partition obtained at each level of the hierarchical solution. This approach was originally used in [10] to compare some similarity indices. It is a well-known fact that the hierarchical clustering does not yield a discrete number of clusters, but rather a hierarchical arrangement between objects.

We perform similar experiments to those described in [10] in order to evaluate the newly introduced index $s(\cdot,\cdot)$, and for comparison we also compute Rand[6], Rand$_{HA}$[7] and Jaccard[6]. For the first set of experiments, each generated data set consists of 50 points uniformly distributed in a hypercube in 4,6, or 8-dimensional Euclidean space. There is no significant cluster structure in the data, but a "criterion" solution is assumed: a hypothetical number of clusters (set at either 2,3,4, or 5), and a particular distribution pattern of the points to the clusters, the so-called 60% density condition, namely one cluster contains 60% of the total number of objects, while 40% of objects are uniformly assigned across the other clusters. For each selected number of clusters, 15 data sets are generated. The hierarchical clustering is performed by using the following methods: single link, complete link, group average, and Ward's method [6]. The computed similarity index is averaged over the data sets and over the hierarchical clustering methods, and the mean statistics (and the borders at two standard deviation) are plotted in Figure 1a) versus the hierarchy level. The only index for which the mean plot is flat, and close to zero, is Rand$_{HA}$. For $s(\cdot,\cdot)$ and Jaccard the computed mean is decreasing when the number of clusters in hierarchical clustering is increasing. Rand takes values larger than the other indices, and the mean is increasing slowly when the number of clusters in hierarchical clustering is increasing.

In the second set of experiments, the test data are generated according to the algorithm described in [11]; the clusters contained in the data are separated in the variable space, and internally cohesive. It was observed that the means of similarity indices are close to 1.0 when the number of clusters in hierarchical solution is equal to the true number of clusters, for all considered structures. We plot in Figure 1b) the mean statistics for the similarity indices in the case of the "60% density condition" for 4 clusters.

All plots in Figure 1 for Rand, Rand$_{HA}$, and Jaccard are very close to similar plots in [10]. The new index $s(\cdot,\cdot)$ has almost the same performance pattern as Jaccard; however, generally the variance of $s(\cdot,\cdot)$ is smaller than the variance of Jaccard index, while the mean is larger. After performing similar experiments for "equal density condition" and "10% density condition" (data not shown), we can extend the conclusions from [10]: a value of at least 0.9 for the Rand, 0.7 for the Jaccard, and 0.8 for $s(\cdot,\cdot)$ is likely to reflect the recovery of some part of the true structure, and not only an agreement due to chance.

4. CLUSTER ANALYSIS FOR A GLIOMA DATASET

We consider here a microarray dataset containing gene expressions from patients with various types of glioma (brain cancer)[12]. The glioma dataset contains measurements of 588 genes for 25 patients: 4 cases of anaplastic astrocytoma (AA), 3 cases of anaplastic oligodendroma (AO), 6 cases of oligodendroma (OL), and 10 cases of glioblastoma multiforme (GM) [12]. The measurements include also 2 patients with pathological attributes close both to anaplastic oligodendroma and to glioblastoma multiforme (AO/GM).

Note that the dataset is summarized as a $N \times p$ matrix where $N = 588$, while $p = 25$. Before applying the clustering algorithms, the data are processed as described in [13] by quantization to four levels, corresponding to four values of the gene expressions: "very low", "low", "high" and "very high". The well-known Lloyd quantizer is used, and we emphasize here that the quantization does not rely on a priori knowledge about how the samples are assigned to different cases of gliomas.

As it was already shown in many studies, the selection of genes (feature selection), plays an important role in sample (patient) clustering. Since feature selection is not the aim of this paper, we resort to use the four genes (IGFBP2, GNB2, UBE2A, CTGF) found to be discriminative for the glioma types [13]. Now the dataset reduces to a 4 × 25 matrix containing values quantized at four different levels. Applying a hierarchical algorithm C_A, we cluster all 25 samples to obtain a reference dendrogram. Then 23 randomly chosen samples are clustered with C_A, and a new dendrogram is built; a similarity index is computed between the partitions obtained by cutting the reference dendrogram, respectively the new dendrogram at level k corresponding to the hypothesized number of clusters. For every algorithm C_A, the experiment is repeated $N_t = 100$ times, and the median for every considered similarity index is computed. Observe in Table 1 that for $k = 5$ non-singleton clusters the agreement between the reference dendrogram and 23 samples-based dendrogram is perfect for all similarity indices and all clustering algorithms. Observe also that median value varies strongly with k, depending on the used clustering algorithm. Based on cluster stability criterion, one can easily decide from Table 1 that the number of distinct clusters present in the glioma dataset is $\hat{K} = 5$, which is in good agreement with the known pathological classification of that data set. Table 2 shows how the samples are assigned to the clusters when cutting the reference dendrogram at level $\hat{K} = 5$. We remark that the best version of the hierarchical clustering method is Ward algorithm which produces the optimal assignment closest to the known pathological discrimination.

Conclusion The hierarchical agglomerative algorithms can be successfully applied for the estimation of the number of (sample) clusters in microarray data, in a very efficient computational scheme, since the same tree can be used for all values of $k \in \{2, 3, \ldots, k_{max}\}$. Once $\hat{K}$ is estimated, partition methods can be further employed for assigning the objects to the clusters.

5. APPENDIX

Proof of Proposition 2.1: The method of proof follows closely [3]. When the hypothesized number of clusters equals the true number of unit balls for the underlying distribution ($k = K$), relying on the main theorem in [14] we conclude that, after an appropriate relabelling, $\sup_{1 \le i \le K} ||\hat{\underline{a}}_i^{tr} - \underline{a}_i|| = o_p(1)$, respectively $\sup_{1 \le i \le K} ||\hat{\underline{a}}_i^{te} - \underline{a}_i|| = o_p(1)$. Reasoning as in [3], it results that $\bar{s}(k) = 1 + o_p(1)$, or equivalently $\bar{s}(k)$ converges in probability to 1 when $k = K$.

When $k > K$, there exist at least one population (out of K) for which the k-means algorithm finds two different centroids during the training, respectively the test stage. To fix the ideas we consider $k = K+1$, and without loss of generality we can assume that the training data laying in $B(\underline{a}_1)$ are split into two clusters by the boundary of a halfspace H_{tr}. As in [3] we analyze the important case when the split of the test set occurs in the same population, and denote H_{te} the respective halfspace. For computing $s(k)$, we focus on the particular structure of the $k \times k$ contingency matrix M corresponding to the partitions induced on the test set by the centroids $\{\hat{\underline{a}}_i{}^{tr} : 1 \le i \le k\}$, respectively $\{\hat{\underline{a}}_i{}^{te} : 1 \le i \le k\}$. Relying on similar arguments as in the case $k = K$, remark that all entries of M except m_{12}, m_{21} and $\{m_{ii}\}_{1 \le i \le k}$ converge to zero, thus

$$s(k) = \frac{\max(m_{11} + m_{22}, m_{12} + m_{21})}{N-1} + \frac{N - n_1 - 1}{N-1 \cdot} + o_p(1)$$

where n_1 is the cardinality of the intersection between $B(\underline{a}_1)$ and the test set. Let $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{N/2}\}$ be the test set, while the training set is $\{\underline{x}_{N/2+1}, \underline{x}_{N/2+2}, \ldots, \underline{x}_N\}$.

For any set $S \subset \Re^p$ we define the indicator function $\phi_S : \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{N/2}\} \to \{0, 1\}$ for which $\phi_S(\underline{x}_i)$ takes value 1 if and only if $\underline{x}_i \in S$. Now we can write the expression for m_{11}: $m_{11} = \sum_{1 \le i \le N/2} \phi_{B(\underline{a}_1) \cap H_{tr} \cap H_{te}}(\underline{x}_i)$. We investigate the behavior of the ratio $\frac{m_{11}}{N/2}$ when $N \to \infty$. First consider the following arguments from [3]: the random halfspaces H_{tr} and H_{te} are independent, their normal directions are distributed uniformly on the unit sphere $\mathcal{S}^{p-1}$, and the distance of the bounding hyperplane to $\underline{a}_1$ converges to zero. Using the identity $\frac{m_{11}}{N/2} = \frac{\sum_{1 \le i \le N/2} \phi_{B(\underline{a}_1) \cap H_{tr} \cap H_{te}}(\underline{x}_i)}{N/2}$, observe that for given halfspaces H_{tr} and H_{te}, $\frac{m_{11}}{N/2}$ is the relative frequency of the event $\{\mathcal{E} : \underline{x}_{i, 1 \le i \le N/2}$ falls into $B(\underline{a}_1) \cap H_{tr} \cap H_{te}\}$, or equivalently the sample average of the indicator function, and the strong law of large numbers implies that $\frac{m_{11}}{N/2}$ converges almost sure (a.s.) to $Pr(\mathcal{E})$. Reasoning as in [3] leads to $Pr(\underline{x}_{i, 1 \le i \le N/2} \in B(\underline{a}_1) \cap H_{tr} \cap H_{te} | H_{tr}, H_{te}) = \frac{1 - \theta/\pi}{2K}$ where $\theta \in (0, \pi)$ is the angle between the normals of H_{tr} and H_{te}. We conclude that, with probability 1, $\frac{m_{11}}{N/2} \to \frac{1-\theta/\pi}{2K}$, $\frac{m_{12}}{N/2} \to \frac{\theta/\pi}{2K}$, $\frac{m_{21}}{N/2} \to \frac{\theta/\pi}{2K}$ and $\frac{m_{22}}{N/2} \to \frac{1-\theta/\pi}{2K}$. Similarly $\frac{n_1}{N/2} \to \frac{1}{K}$ a.s. Moreover, observe that $\max(m_{11} + m_{22}, m_{12} + m_{21})$ takes the value $\frac{1-\theta/\pi}{K}$ when $\theta \in (0, \pi/2]$, respectively $\frac{\theta/\pi}{K}$ when $\theta \in (\pi/2, \pi)$. From the asymptotic expression of $s(k)$, it is elementary to obtain:

$$s(k) \le 1 + \frac{1}{2} \frac{\max(m_{11} + m_{22}, m_{12} + m_{21})}{N/2} - \frac{1}{2} \frac{n_1}{N/2} + o_p(1)$$

For $N \to \infty$ and $N_t \to \infty$ we compute $\bar{s}(k)$ applying a result from [15]: the density of the angle θ is
$g(\theta) = \frac{\Gamma(p/2)}{\Gamma((p-1)/2)\sqrt{\pi}} (\sin \theta)^{p-2}$. We focus on the contribution to $\bar{s}(k)$ of the term $\frac{\max(m_{11}+m_{22}, m_{12}+m_{21})}{N/2}$:

$$\mathcal{T} \triangleq \int_0^{\pi/2} \frac{1 - \theta/\pi}{K} g(\theta) d\theta + \int_{\pi/2}^{\pi} \frac{\theta/\pi}{K} g(\theta) d\theta$$

Elementary calculations lead to $\mathcal{T} \le \frac{3}{4} \frac{1}{K}$ for any $p \ge 2$, then is straightforward to show that $\bar{s}(k) \le 1 - \frac{1}{8} \frac{1}{K} + o_p(1)$.

6. REFERENCES

[1] J.N. Breckenridge, "Replicating cluster analysis: method, consistency, and validity," *Multivariate Behav. Res.*, vol. 24, no. 2, pp. 147–161, 1989.

[2] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biology*, vol. 3, no. 7, pp. research0036.1 – 0036.21, Jun. 2002, http://genomebiology.com/2002/3/7/research/0036.

[3] R. Tibshirani, G. Walther, D. Botstein, and P. Brown, "Cluster validation by prediction strength," Tech. Rep., Department of Biostatistics, Stanford University, Sep. 2001, http://www-stat.stanford.edu/~tibs/research.html.

[4] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Pac. Symp. Biocomputing*, 2002, vol. 7, pp. 6–17.

[5] A. Ben-Hur and I. Guyon, *Detecting stable clusters using Principal Component Analysis*, Methods in Molecular Biology. Humana Press, to appear, http://clopinet.com/isabelle/Papers/index.html.

[6] L. Kaufman and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, Wiley, 1990.

[7] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[8] A. Almudevar and C. Field, "Estimation of single generation sibling relationships based on DNA markers," *J. Agric. Biol. Environ. Stat.*, vol. 4, no. 2, pp. 136–165, 1999.

[9] D. Gusfield, "Partition-distance: a problem and class of perfect graphs arising in clustering," *Inform. Process. Lett.*, vol. 82, pp. 159–164, 2002.

[10] G.W. Milligan and M.C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behav. Res.*, vol. 21, pp. 441–458, 1986.

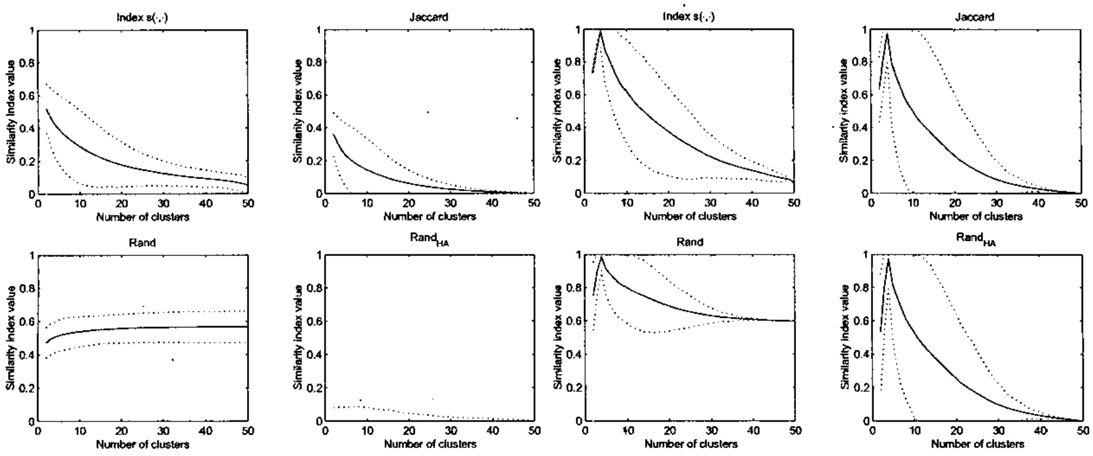
[11] G.W. Milligan, "An algorithm for generating artificial test clusters," *Psychometrika*, vol. 50, no. 1, pp. 123–127, Mar. 1985.

[12] G. Fuller, K. Hess, C. Mircean, and I. Tabus et al., "Human glioma diagnosis from gene expression data," in *Computational And Statistical Approaches To Genomics*, W. Zhang and I. Shmulevich, Eds. Kluwer Academic Pub., 2002.

[13] I. Tabus, C. Mircean, W. Zhang, I. Shmulevich, and J. Astola, "Transcriptome-based glioma classification using informative gene set," in *Genomic and molecular neuro-oncology*, W. Zhang and G. Fuller, Eds. Jones and Bartlett Pub., 2003.

[14] D. Pollard, "A central limit theorem for k-means clustering," *Ann. Prob.*, vol. 10, no. 4, pp. 919–926, Sep. 1982.

[15] G. S. Watson, *Statistics on spheres*, Wiley, 1983.



a) b)

Figure 1: The case "60% density condition": mean of the similarity indices versus the number of clusters (solid line) with limits at two standard deviation (dotted line). a) No structure exists in the data. b) Data contains four distinct clusters.

Similarity index	Hierarchical clustering method	Number of clusters (k)					
		2	3	4	5*	6	7
$s(\cdot,\cdot)$	group-average	**1.0000**	0.9545	0.8095	**1.0000**	0.7727	0.7727
	complete-linkage	**1.0000**	0.9091	0.8500	**1.0000**	0.7727	0.8182
	Ward	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9318	0.9091
Jaccard	group-average	**1.0000**	0.8182	0.5763	**1.0000**	0.6092	0.5147
	complete-linkage	**1.0000**	0.6927	0.6190	**1.0000**	0.6232	0.5705
	Ward	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.7963	0.8298
Fowlkes-Mallows	group-average	**1.0000**	0.9001	0.7322	**1.0000**	0.7585	0.6822
	complete-linkage	**1.0000**	0.8188	0.7653	**1.0000**	0.7690	0.7354
	Ward	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.8866	0.9070
Rand$_{HA}$	group-average	**1.0000**	0.8528	0.6026	**1.0000**	0.6189	0.6077
	complete-linkage	**1.0000**	0.7295	0.6870	**1.0000**	0.7021	0.6784
	Ward	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.8624	0.8879

Table 1: Glioma dataset: the median of similarity indices, computed for $N_t = 100$ trials when the hypothesized number of clusters varies between 2 and 7. Bold characters are used to represent the maximum computed value for each similarity index, and for each clustering method. We estimate for all clustering algorithms and all similarity indices $\hat{K} = 5$, because five is the largest value of k for which the clustering is stable.

group-average					complete-linkage					Ward				
AA	AO	OL	GM	AO/GM	AA	AO	OL	GM	AO/GM	AA	AO	OL	GM	AO/GM
0	0	0	**5**	0	0	1	**6**	0	1	0	1	**6**	0	1
0	0	0	1	**0**	0	0	0	2	**0**	0	**0**	0	2	0
0	1	**6**	1	1	0	0	0	**5**	0	**4**	1	0	1	0
0	**0**	0	2	0	0	**0**	0	1	0	0	1	0	1	**1**
4	2	0	1	1	**4**	2	0	2	1	0	0	0	**6**	0

Table 2: The contingency tables for the glioma dataset: the true partition given by a priori knowledge on the type of disease for each patient is compared with partitions obtained by cutting hierarchical clustering trees at level $\hat{K} = 5$. For each contingency table, the entries associated to the optimal assignment are represented in bold. The optimal assignment takes value 15 for group-average and complete-linkage, respectively 17 for Ward algorithm. The 4 AA cases, respectively the 6 OL cases are correctly clustered by all algorithms. Group-average and complete-linkage cluster together only 5 GM cases, while Ward method group properly 6 GM cases.