2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

PVis - Partitions' Visualizer: extracting knowledge by visualizing a collection of partitions

Katti Faceli and Tiemi C. Sakata
Departamento de Computação de Sorocaba
Universidade Federal de São Carlos
Sorocaba, Brazil
Email: {katti, tiemi}@ufscar.br

André C. P. L. F. de Carvalho
ICMC
Universidade de São Paulo
São Carlos, Brazil
Email: andre@icmc.usp.br

Marcilio C. P. de Souto
Univ. Orléans
INSA Centre Val de Loire, LIFO EA 4022
Orléans, France
Email: marcilio.desouto@univ-orleans.fr

***Abstract*—Recent advances in cluster analysis highlight the importance of finding multiple meaningful partitions and point out to the need for approaches to evaluate them. They also suggest that the evaluation should consider knowledge of a domain expert. In this paper, we present a visualization method, called PVis[1] (Partition's Visualizer), that allows the integrated visualization of a collection of partitions. PVis allows to compare the content of a set of partitions. The comparison can be done with respect to priori knowledge provided by an expert. PVis can be useful in the discovery of relevant information to the domain experts performing cluster analysis. In order to illustrate our approach, we give an example of how to perform an exploratory analysis of collections of partitions. In order to do so, we use a well-known dataset from the Bioinformatics domain, regarding molecular classification of cancer.**

I. Introduction

Clustering techniques are suited to explore and verify structures that are present in the data, by grouping the objects according to some sort of similarity [1], [2], [3]. The idea is to reveal the hidden intrinsic structures with great potential of practical utility for the domain experts.

A great number of applications of cluster analysis can be found today in both academic and commercial areas. The solutions can range from application of traditional clustering algorithms to the use of the more recent approaches in cluster analysis, which encompasses the ensembles, multi-objective clustering, subspace clustering, multi-view clustering, among others. Important areas such as biology, medicine, engineering, marketing, remote sensing and bioinformatics can be benefited from cluster analysis. In bioinformatics, for example, cluster analysis has been successfully applied to gene expression data with the aim of gathering insights for the understanding of biological processes and diseases mechanisms [4].

In general, cluster analysis comprises several steps, ranging from data preparation to the validation and interpretation of the results [5]. These last tasks are of ultimate importance to the experts, as they will guarantee the usefulness of the knowledge extracted. The interpretation of the results involves the observation of the clusters' contents and could benefit from their comparison with domain knowledge that is available. The inspection of the contents of one or two partitions is a manageable task. However, if a higher number of partitions is to be inspected, a graphical representation that allows the simultaneous visualization of the partitions would be necessary.

In this paper, we propose a visualization method, called PVis (Partitions' Visualizer). It is suited for the simultaneous visualization of a collection of partitions, allowing the comparison of their clusters' contents. PVis can be used in several ways. We will illustrate one of them. More specifically for using domain knowledge, in the form of a known partition of the data, to guide the discovery of new knowledge. To illustrate the use of PVis, we will present a case study in bioinformatics domain.

In Section II, we provide an overview of existing techniques for visualization in the context of gene expression's data analysis. In Section III we describe the algorithm to generate the visualization for PVis. Next, in Section IV, we present a case study that illustrates the use of PVis for exploring gene expression data aiming the identification of subtypes of cancer.

II. Related work

We decided to restrict the related work section to the techniques closely related to our case study's domain. Thus, in this section, we will provide an overview of the techniques for visualization of clustering results in the context of Bioinformatics.

One of the most frequently used graphical representation in this context is the heatmap [6], [7]. A heatmap is a two-dimensional colored grid used to display data that are arranged in a matrix. The value of each entry in the matrix determines the color of the corresponding position in the grid. The rows and columns of the matrix can be independently reordered to display the similar rows and/or columns next to each other. In [6], the authors use the result of an hierarchical clustering algorithm to order the genes in the original data matrix. Next, they graphically represent the primary data: each attribute of each data point is represented with a color that quantitatively and qualitatively reflects the original observations. The dendrogram showing the hierarchy obtained by the clustering algorithm is also represented in the display.

Several other visualization tools have been proposed to aid the exploration of either one partition or a hierarchy of partitions. In [8], the authors compare several dimensionality reduction methods for visualization of microarray data, which helps in identifying clusters of samples. In [9], the authors

This work was partially supported by FAPESP.

[1]PVis is available at http://lasid.sor.ufscar.br/coloring/index.html

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

present a tool for genomic data exploration. Its tool, VISual Statistical Data Analyzer (VISDA), integrates hierarchical clustering and visualization supported by hierarchical mixture modeling, supervised/unsupervised informative gene selection, supervised/unsupervised data projection and prior knowledge. In [10], the authors present other useful visualization tool that represents a partition that is a consensus of several other partitions, facilitating the determination of the number of clusters, clusters' membership and boundaries.

In [11], the authors mention a variety of approaches and tools for data visualization and briefly discuss their application scenarios. They argue that most of the existing approaches are not intended for cluster analysis, are only suitable for analysis of a hierarchical clustering or are based on a specific clustering algorithm. Given these arguments, the authors in [11] propose several visualization techniques that can be used with the results of any clustering algorithm applied to protein and gene expression data. Their several methods complement each other, allowing different types of analysis of the data.

All methods previously mentioned are suited for the visualization of the data or for the visualization of a single partition at a time. At the best of our knowledge, in the literature, there is no method for the integrated representation and visualization of a broad collection of independent partitions.

The tool of Eisen et al. [6] is the most directly related to PVis. As already mentioned, it represents an hierarchy of partitions and is based in the original data to produce the visualization. In contrast, PVis is purely based on objects assignments to the clusters. It does not consider the data itself (like the expression profiles). In this way, our approach and the methods from [6] and [11], for example, complement each other.

III. PVIS: PARTITIONS' VISUALIZER

A simplified version of the visualization method implemented in PVis was first presented in [12]. Before describing the method, we will introduce some useful notation.

A clustering algorithm looks for structures hidden in the data [1], [3]. One common kind of structure is a hard partition of the data. In this kind of partition, each object should be assigned to only one cluster, and all objects must be assigned to a cluster. More formally, given a set of objects $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, a partition of X in K^i clusters can be defined as: $\pi^i = \{c_1^i, c_2^i, ..., c_{K^i}^i\}$ with $2 \leq K^i < n$, such that [3]:

1) $c_j^i \neq \emptyset, j = 1, ..., K^i$,
2) $\bigcup_{j=1}^{K^i} c_j^i = X$ and
3) $c_j^i \cap c_l^i = \emptyset, j, l = 1, ..., K^i$ and $j \neq l$.

Let $\Pi = \{\pi^1, \pi^2, ..., \pi^r\}$ be a set of partitions to be visualized, where r is the number of partitions. Let $\pi^{\mathtt{ref}} = \{c_1^{\mathtt{ref}}, c_2^{\mathtt{ref}}, ..., c_{k^{\mathtt{ref}}}^{\mathtt{ref}}\}$ be a *reference partition*[2].

The visualization method is based on a coloring scheme that assigns colors to each $c_j^i \in \pi^i$, according to $\pi^{\mathtt{ref}}$. The coloring scheme is presented in the algorithm of Fig. 1.

[2]The meaning of a *reference partition* will be detailed later in this section.

Input: Π, $\pi^{\mathtt{ref}}$
Output: color and intensity for each $c_j^k \in \pi^k$ where $\pi^k \in \Pi$ and for each $c_i^{\mathtt{ref}} \in \pi^{\mathtt{ref}}$

1: **for all** $c_i^{\mathtt{ref}} \in \pi^{\mathtt{ref}}$ **do**
2: color($c_i^{\mathtt{ref}}$) $\leftarrow$ newColor
3: intensity($c_i^{\mathtt{ref}}$) $\leftarrow$ maxIntensity
4: **end for**
5: **for all** $\pi^k \in \Pi$ **do**
6: **for all** $c_j^k \in \pi^k$ **do**
7: $c_{maj} = \mathtt{argmax}_{c_i^{\mathtt{ref}} \in \pi^{\mathtt{ref}}} |c_i^{\mathtt{ref}} \bigcap c_j^k|$
8: color(c_j^k) $\leftarrow$ color(c_{maj})
9: **end for**
10: **for all** $c_i^{\mathtt{ref}} \in \pi^{\mathtt{ref}}$ **do**
11: C $\leftarrow$ $\{$ c_j^k | color(c_j^k) = color($c_i^{\mathtt{ref}}$), $\forall$ $c_j^k \in \pi^k$ $\}$
12: sort C in descending order according to $|c_i^{\mathtt{ref}} \bigcap c_j^k|$, $c_j^k \in$ C
13: auxIntensity $\leftarrow$ maxIntensity
14: **for all** $c_j^k \in$ C **do**
15: intensity(c_j^k) $\leftarrow$ auxIntensity
16: auxIntensity $\leftarrow$ auxIntensity - 1
17: **end for**
18: **end for**
19: **end for**

Fig. 1. Algorithm of the coloring scheme

In order to illustrate the description of the method, we will consider a dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{10}\}$, and $\pi^{\mathtt{ref}}$ and $\Pi = \{\pi^1, \pi^2\}$ as shown in Table I. To avoid confusion in the explanation, we will make use of the word *class* to refer to the clusters of the reference partition, reserving the term cluster to the clusters of the partitions in Π.

TABLE I. PARTITIONS FOR THE EXAMPLE

Partition	Clusters
$\pi^{\mathtt{ref}}$	$c_1^{\mathtt{ref}} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, $c_2^{\mathtt{ref}} = \{\mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$
π^1	$c_1^1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $c_2^1 = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7\}$, $c_3^1 = \{\mathbf{x}_6\}$, $c_4^1 = \{\mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$
π^2	$c_1^2 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9\}$, $c_2^2 = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$, $c_3^2 = \{\mathbf{x}_6\}$, $c_4^2 = \{\mathbf{x}_{10}\}$

The first step in the procedure of generating the visualization is the assignment of a different color and an intensity to each *class* (Fig. 1, Lines 1 to 4). In our example, we assign the color green to $c_1^{\mathtt{ref}}$ and the color orange to $c_2^{\mathtt{ref}}$, and a maximum value for intensity of 3 for both cases. These assignments can be seen in Table II. The maximum intensity should be the highest number of clusters sharing the same color in the partitions in Π. In the example, the value is 3 as there is at most three clusters with the same color in each of the partitions in Π (the green clusters in π^1).

Next, the majority *class* of each cluster (c_{maj}) is found and each cluster $c_k^i \in \pi^i$ is labeled with the same color of its majority *class* (Fig. 1, Lines 6 to 9). The majority *class* of a cluster is the *class* with the largest number of objects in the cluster. Table II shows the results of these steps applied to π^1 and π^2. Column $|c_i^{\mathtt{ref}} \cap c_j^k|$ shows the number of objects shared between each cluster c_j^k and each *class* $c_i^{\mathtt{ref}}$. The largest number for each cluster is highlighted in boldface, which

TABLE II. STEPS OF THE COLORING SCHEME APPLIED TO THE EXAMPLE

Partition	Cluster	$\lvert c_i^{\mathtt{ref}} \cap c_j^k \rvert$		c_{maj}	Color	Intensity
		$c_1^{\mathtt{ref}}$	$c_2^{\mathtt{ref}}$			
$\pi^{\mathtt{ref}}$	$c_1^{\mathtt{ref}}$	-	-	-	green	3
	$c_2^{\mathtt{ref}}$	-	-	-	orange	3
π^1	c_1^1	**3**	0	$c_1^{\mathtt{ref}}$	green	3
	c_2^1	**2**	1	$c_1^{\mathtt{ref}}$	green	2
	c_3^1	**1**	0	$c_1^{\mathtt{ref}}$	green	1
	c_4^1	0	**3**	$c_2^{\mathtt{ref}}$	orange	3
π^2	c_1^2	2	**3**	$c_2^{\mathtt{ref}}$	orange	3
	c_2^2	**3**	0	$c_1^{\mathtt{ref}}$	green	3
	c_3^2	**1**	0	$c_1^{\mathtt{ref}}$	green	2
	c_4^2	0	**1**	$c_2^{\mathtt{ref}}$	orange	2

indicates the choice of the majority *class* (column c_{maj}). The column "Color" shows the color assigned to each cluster.

Given all $c_k^i \in \pi^i$ with the same majority *class* (and thus with the same color), the highest intensity value is assigned to the cluster with the greatest number of objects from the majority *class*. Decreasing values of intensities are assigned to the other clusters considering the number of objects from their majority *class* in decreasing order (Fig. 1, Lines 10 to 18). The intensities assignments can be seen in Table II.

Once all partitions have been colored, they are organized in a table. As a convention, the first column represents the identifiers of the objects in the dataset. The second column represents $\pi^{\mathtt{ref}}$. Each of the remaining columns represents one of $\pi^i \in \Pi$. Each cell in the table is filled with the color/intensity representing the cluster to which the object belongs to in the corresponding partition. Fig. 2 shows the result of PVis applied to the example.

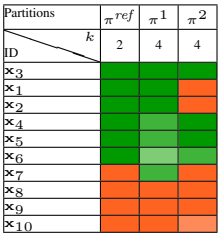

Fig. 2. Visualization of the example

The table with the visual information is as large as the number of objects of the dataset. Any number of objects can be represented. However, if this number is very large, the observation of the complete visual information by a human is hard. In order to minimize this difficulty we included in PVis a step of collapsing the objects with identical coloring pattern. That is, the rows with identical pattern can be represented as one single row. The user can set the minimal number of identical rows he/she wants to group. In the example, we can observe that $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same color and the same intensity in all columns and they can be collapsed. For the same reason, it is possible to collapse $\mathbf{x}_4$ with $\mathbf{x}_5$ and $\mathbf{x}_8$ with $\mathbf{x}_9$. Fig. 3 shows the result of collapsing identical lines of the example.

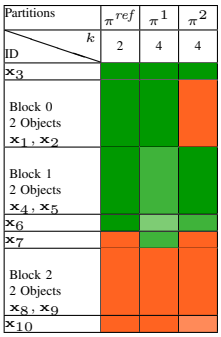

Fig. 3. Visualization of the example collapsing identical lines

In practical terms, the reference partition can be chosen in different ways, depending, for example, on the kind of priori information the user has or his/her aim in the analysis. For instance, the user could choose:

- A partition $\pi_i \in \Pi$. For example, (1) the partition with the smallest number of clusters, or (2) the partition that shows more similarity to each of the other partitions in Π, according to some partition similarity criterion.
- A known partition of the data. In this case, the expert can provide this priori knowledge in order, for example, to support the discovery of new knowledge.
- A partition that represents the consensus among the partitions in Π. Any ensemble method can be used to produce the consensus partition [13].

In the case study illustrated in this paper, the reference partition will be a known partition of the data.

IV. CASE STUDY

In order to show an application of the PVis method we use a classical dataset from the Bioinformatics domain. The aim of the analysis is to illustrate how new knowledge can be identified from sets of partitions if prior knowledge is given in the form of a reference partition. In order to do so, we first produced a set of partitions with traditional clustering algorithms. Then, we visualize these partitions using one known structure of the dataset as the reference.

A. Data

The dataset we used contains gene expression data from acute leukemia patients [14]. This dataset is interesting for our analysis as it has more than one underlying structure. The two structures of clinical interest (E_1 and E_2) refer to types and subtypes of acute leukemia:

- E_1: corresponds to the classification of the samples in Acute Lymphoblastic Leukemia (`ALL`) and Acute Myeloid Leukemia (`AML`);
- E_2: contains a refinement of the `ALL` class. In this case, the data are classified in `AML`, `Tcell` (T-lineage ALL) and `Bcell` (B-lineage ALL).

Besides E_1 and E_2, two other types of information are available: E_3 and E_4. They have no clinical relevance. We

will use them to check if some group could have appeared due to some sort of unexpected influence. For example, if a group emerged due to bias in the samples preparation or tissue type. The structures E_3 and E_4 refer to:

- E_3: a division of the samples according to the institution where they came from: DFCI (*Dana-Farber Cancer Institute*), CALGB (*Cancer and Leukemia Group B*), StJude (*St. Jude Children's Research Hospital*) and CCG (*Children's Cancer Group*);
- E_4: shows if the samples are from bone marrow (BM) or peripheral blood (PB).

All samples considered in [14] are used and the data was preprocessed in the same way as in the original work. After preprocessing we obtained a dataset containing 72 samples and 3571 attributes. To make things clearer, we labeled each sample according to its membership. For example, sample ALL-Bcell-BM-DFCI-1 belongs to ALL in E_1, to Bcell in E_2, to DFCI in E_3 and to BM in E_4. The final number was used just to enumerate the samples.

B. Partitions' generation

To generate the collection of partitions Π, we ran three traditional and largely employed clustering algorithms [1]. The algorithms employed were: *k*-means (KM), complete-linkage (CoL) and average-linkage (AL) [1]. We run them with Euclidean distance and generate partitions with numbers of clusters $k \in [3, 8]$. In order to minimize the occurrence of suboptimal solutions, we run k-means 30 times for each k, with a random choice of initial centers. Among all 30 partitions produced for a given k, we selected the partition with the lowest squared error. For the algorithms AL and CoL, we generated the hierarchies and cut them in order to produce one partition for each value of k. With this procedure, we produced a set of partitions $\Pi = \{\pi^1, \pi^2, ..., \pi^{18}\}$.

C. Visualization and analysis

To visualize Π, we run PVis using E_1 (AML/ALL distinction) as the reference partition. We ordered the columns of the table by the values of the corrected Rand (CR) of the partitions with respect to $\pi^{\texttt{ref}}$ [1]. This means that π^1 is more similar to $\pi^{\texttt{ref}}$ than π^2 and so on. The visualization produced is shown in Fig. 4. In this table, π^1 to π^3 and π^{16} to π^{18} were generated with AL; π^6, π^9 and π^{12} to π^{15} were generated with KM; and π^4, π^5, π^7, π^8, π^{10} and π^{11} were generated with CoL.

Finding wrong assignments and outliers

The wrong assignments are objects placed with objects of a different *class* with respect to the reference partition. In the visualization, they are represented with a different color than that of the other objects of their *class*. In Fig. 4, the objects AML-BM-CALGB-29, AML-BM-CALGB-34, AML-BM-CALGB-35, AML-BM-CCG-66 were grouped with ALL-Bcell samples in several partitions (most of partitions from π_4 to π_{15}).

The outliers are isolated in small clusters or singletons, represented with a different color/intensity than that of all other objects. In Fig. 4, the most evident case of outlier is the object AML-PB-CCG-64 that was placed in a singleton in half of the partitions. Other cases are ALL-Bcell-BM-DFCI-17, ALL-Bcell-BM-DFCI-20 and ALL-Bcell-BM-DFCI-21, presenting a different color patter in partitions π_1 to π_3 and π_{16} to π_{18}.

The identification of wrong assignments and outliers is important to the expert. These objects could be noisy samples, incorrectly labeled objects, or samples that present an atypical behavior with respect to the other objects in the same *class*.

Finding new groups

The subdivision of a given original *class* of a dataset in more than one cluster can support the discovery of novel *classes* in the data. For example, this could lead to a molecular-based refinement of broadly defined biological *classes*, with implications in cancer diagnosis, prognosis and treatment [10]. In the visualization, the subdivision of a *class* in more than one cluster can be observed by the presence of different intensities of the color associated with the *class*.

In Fig. 4, we can observe several subdivisions of both *classes*, ALL and AML. These subdivisions could represent new knowledge that deserves investigation, as they appear several times as a clear data separation. For part of the subdivisions that we found, the information available in [14] (E_2, E_3 and E_4) gave us support to provide the cluster's interpretation, validating them as new knowledge discovered with PVis. For other cases, information was not available for this dataset, but it's known that several other relevant subdivisions of leukemia exists. For example, [15] mentions several subdivisions of the Bcell. In [16], the authors shows that besides the AML and AML types, a third type exists, revealed by a clear distinction from AML and ALL when applying clustering algorithms.

Observing Fig. 4, the first subdivision we can notice is the clear separation of the Bcell and Tcell samples. In π_3, the cluster in dark green (rows 1 to 35) corresponds to Bcell samples, and the cluster in light green (rows 36 to 44) contains all the Tcell samples. Moreover, the color pattern seen in rows 36 to 43 evidence a cluster containing the samples of the *class* Tcell (partitions π_3, π_{10} to π_{14}). This shows that the knowledge of E_2 could be discovered using PVis.

We can also clearly identify two other subdivisions of the ALL samples. In rows 1 to 28 and 29 to 35, we identified two groups of objects that contains the samples of the Bcell. Although we do not have enough information to provide the interpretation of these clusters, it is possible that they reflect meaningful subtypes of Bcell.

For the *class* AML, we also observed a division of the samples in two groups. The cluster represented in dark orange encompasses the majority of the samples from CALGB. The other cluster encompasses most remaining AML samples originated from the other institutions (rows 61 to 65 and 68 to 71). The clusters found could also refer to the separation of adults and children, as all the sample from CALGB are from adults and all the samples of other institutions are from children.

From a clinical point of view, the separation of the samples by institution is not expected to occur. The clusters found could be due to differences in protocols used by the laboratories for sample preparation. Identifying such problems, the expert could, for example, propose approaches to avoid these artifacts such as the standardization of the protocols.

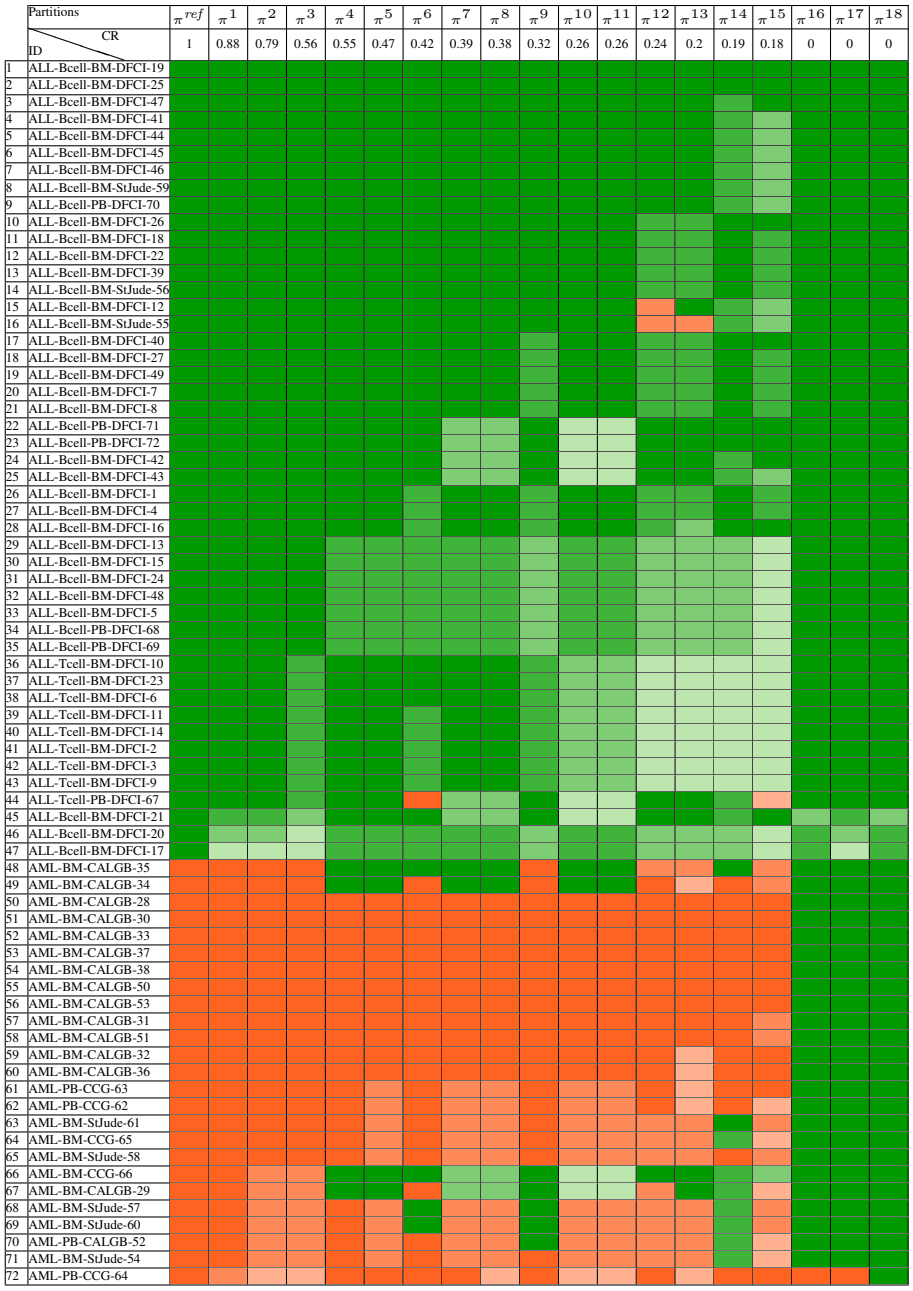

Fig. 4. Visualization of the partitions

V. Final remarks

In this paper, we presented a visualization method called PVis (Partition's Visualizer). PVis allows the visual inspection (comparison) of a collection of partitions of a given dataset. The aim is to visually identify useful information contained in the clusters belonging to the set of input partitions.

To illustrate the use of PVis, we provided a simple example of how to perform an exploratory data analysis. More specifically, we used a widely known dataset regarding molecular classification of cancer [14]. It is known that this dataset contains two structures of clinical interest that refer to types and subtypes of acute leukemia. Based on this, we considered one of the known structure as the *reference partition* for PVis — for example, as if it was prior knowledge provided by the user. To generate the partitions to be visualized/inspected, we ran the dataset with three traditional clustering algorithms: k-means, average-linkage and complete-linkage.

In this context, by using PVis, we could observe several subdivisions of the *classes* `ALL` and `AML` (*reference partition*). For some of the subdivisions that we could identify, the information available in [14] gave us support to provide the cluster's interpretation. For example, via PVis we could see clearly the other clinical subdivision of this data: `Bcell`, `Tcell`, and `AML`.

For other cases in which we could visualize subdivisions, the original paper ([14]) had no information for supporting an

interpretation. However, there are works in the literature that show that several other relevant subdivisions of leukemia exist. For example, [15] mentions several subdivisions of `Bcell`. Finally, by our visual inspection we found some partitions that had clusters representing a separation of adults and children — this kind of "structure" was not discussed in [14].

In terms of extension and further work, the current version of PVis does not scale up for a large number of partitions. Other points to be investigated includes the application of PVis to a broader range of recent techniques for finding multiple clustering solutions and a deeper analysis of the possible types of reference partitions that can be used.

REFERENCES

[1] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[2] J. Handl, J. Knowles, and D. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[3] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[4] M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics*, vol. 9, pp. 497–520, 2008.

[5] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, September 1999.

[6] M. B. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proc. Natl. Acad. Sci. USA*, vol. 95, 1998, pp. 14 863–14 868.

[7] W. Huber, X. Li, and R. Gentleman, "Visualizing data," in *Bioinformatics And Computational Biology Solutions Using R And Bioconductor*, 2nd ed., R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, Eds. New York: Springer, 2005, vol. 2, pp. 161–179.

[8] C. Bartenhagen, H.-U. Klein, C. Ruckert, X. Jiang, and M. Dugas, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data," *BMC Bioinformatics*, vol. 11, p. 567, 2010.

[9] Y. Zhu, H. Li, D. Miller, Z. Wang, J. Xuan, R. Clarke, E. Hoffman, and Y. Wang, "cabigtm visda: Modeling, visualization, and discovery for cluster analysis of genomic data," *BMC Bioinformatics*, vol. 9, no. 1, p. 383, 2008. [Online]. Available: http://www.biomedcentral.com/1471-2105/9/383

[10] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[11] M. A. Hibbs, N. C. Dirksen, K. Li, and O. G. Troyanskaya, "Visualization methods for statistical analysis of microarray clusters," *BMC Bioinformatics*, vol. 6, p. 115, 2005.

[12] K. Faceli, A. C. F. L. F. Carvalho, and M. C. P. de Souto, "Evaluation of the contents of partitions obtained with clustering gene expression data," in *IV Brazilian Symposium on Bioinformatics. Lecture Notes in Computer Science*, vol. 3594, 2005, pp. 65–76.

[13] ——, "Cluster ensemble and multi-objective clustering methods," in *Pattern Recognition Technologies and Applications: Recent Advances*, B. Verma and M. Blumenstein, Eds. IGI Global, 2008, pp. 325–343.

[14] T. R. Golub, P. T. D. K. Slonim and, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. 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*, vol. 286, no. 5439, pp. 531–537, 1999.

[15] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[16] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.