BIOINFORMATICS REVIEW Vol. 21 no. 15 2005, pages 3201–3212
doi:10.1093/bioinformatics/bti517

Data and text mining

Computational cluster validation in post-genomic data analysis

Julia Handl*, Joshua Knowles and Douglas B. Kell

School of Chemistry, University of Manchester, Faraday Building, Sackville Street, PO Box 88, Manchester M60 1QD, UK

Received on March 24, 2005; revised and accepted on May 24, 2005
Advance Access publication May 24, 2005

ABSTRACT

Motivation: The discovery of novel biological knowledge from the *ab initio* analysis of post-genomic data relies upon the use of unsupervised processing methods, in particular clustering techniques. Much recent research in bioinformatics has therefore been focused on the transfer of clustering methods introduced in other scientific fields and on the development of novel algorithms specifically designed to tackle the challenges posed by post-genomic data. The partitions returned by a clustering algorithm are commonly validated using visual inspection and concordance with prior biological knowledge—whether the clusters actually correspond to the real structure in the data is somewhat less frequently considered. Suitable computational cluster validation techniques are available in the general data-mining literature, but have been given only a fraction of the same attention in bioinformatics.

Results: This review paper aims to familiarize the reader with the battery of techniques available for the validation of clustering results, with a particular focus on their application to post-genomic data analysis. Synthetic and real biological datasets are used to demonstrate the benefits, and also some of the perils, of analytical cluster validation.

Availability: The software used in the experiments is available at http://dbkgroup.org/handl/clustervalidation/

Contact: J.Handl@postgrad.manchester.ac.uk

Supplementary information: Enlarged colour plots are provided in the Supplementary Material, which is available at http://dbkgroup.org/handl/clustervalidation/

1 INTRODUCTION

The exploration of complex datasets, for which no or very little information about the underlying distribution is available, fundamentally relies on the identification of 'natural' group structures in the data, a task which may be tackled using clustering techniques (Duda *et al.*, 2001; Everitt, 1993; Hastie *et al.*, 2001; Jain *et al.*, 1999). A cluster analysis can be seen as a three step process as outlined in Figure 1. Cluster validation techniques (Dubes and Jain, 1979) are clearly essential tools within this process, and their frequent neglect in the post-genomic literature hampers progress in the field. In particular, this is of concern in two areas:

- *Algorithm development.* Many novel clustering algorithms are insufficiently evaluated, such that users remain unaware of their relative strengths and weaknesses. A more thorough use of quantitative, reproducible and objective cluster-validation techniques would permit one to alleviate this uncertainty, thus assisting the distinction between more and less useful methods, and encouraging the acceptance of novel advanced clustering techniques.
- *Verification of results.* Most current clustering algorithms do not provide estimates of the significance of the results returned.[1] The verification of clustering results is therefore often based on a manual, lengthy and subjective exploration process. Cluster-validation techniques have the potential to provide an

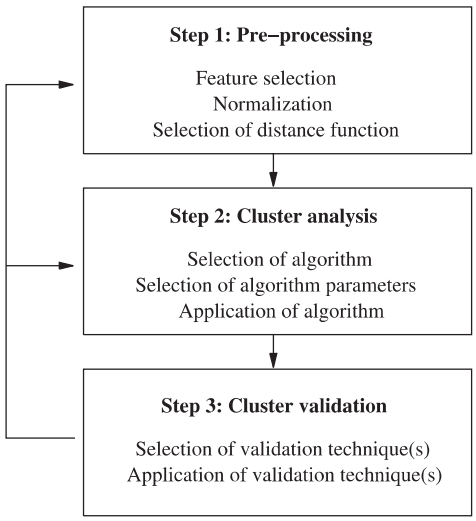

Fig. 1. The three main steps involved in a cluster analysis. The first of these involves a number of data transformations including feature selection, normalization and the choice of a distance function, to ensure that related data items cluster together in the data space. The second step consists of the selection, parameterization and application of one or several clustering methods. The resulting partitionings are evaluated in the third step, and it is at this stage that cluster-validation techniques are needed. The results of a cluster analysis may be crucially affected by the choices made in the first two steps, and information on the quality of the partitioning can (and should) therefore be used to revise these choices.

*To whom correspondence should be addressed.

[1]Notable exceptions include classic expectation–maximization (EM) algorithms as well as some newly developed methods such as adaptive quality-based clustering (DeSmet *et al.*, 2002), the self-organizing tree algorithm (Herrero *et al.*, 2001), and multiobjective clustering (Handl and Knowles, 2005).

analytical assessment of the amount and type of structure captured by a partitioning, and should therefore be a key tool in the interpretation of clustering results.

The aim of this review paper is to explain and to encourage the use of cluster-validation techniques in the analysis of post-genomic and other data. In particular, the paper attempts to familiarize researchers with some of the fundamental concepts behind cluster-validation techniques, and to assist them in making more informed choices of the measures to be used. The remainder of this paper is structured as follows. Section 2 provides a summary of essential background information. The different types of validation techniques are reviewed in Section 3, followed by a discussion of some of their fundamental biases and problems (Section 4). Section 5 attempts to give some guidelines regarding the effective use of validation techniques, and Section 6 demonstrates their use on a gene expression dataset. Finally, the conclusion is given in Section 7.

2 BACKGROUND

2.1 Clustering

Traditional classifications of clustering algorithms (Duda *et al*., 2001; Everitt, 1993; Hastie *et al*., 2001; Jain *et al*., 1999) primarily distinguish between hierarchical, partitioning and density-based methods. Here, a somewhat different categorization is used, based on the clustering criterion (implicitly or explicitly) optimized by each algorithm. This permits a better appreciation of the connections between clustering algorithms and cluster-validation techniques. Capturing the intuitive notion of a cluster by means of any explicit, formal definition is one of the fundamental difficulties of clustering (Estivill-Castro, 2002). There are several valid properties that may be ascribed to a good partitioning, but these are partly in conflict and are generally difficult to express in terms of objective functions. Despite this, existing clustering criteria/algorithms do fit broadly into three fundamental categories:

- *Compactness.* This concept is generally implemented by keeping the intra-cluster variation small. This category includes algorithms like *k*-means (MacQueen, 1967), average-link agglomerative clustering (Vorhees, 1985), self-organizing maps (SOMs) (Kohonen, 2001) or model-based clustering approaches (McLachlan and Krishman, 1997). The resulting methods tend to be very effective for spherical or well-separated clusters, but they may fail to detect more complicated cluster structures (Duda *et al*., 2001; Everitt, 1993; Hastie *et al*., 2001; Jain *et al*., 1999).
- *Connectedness.* This is a more local concept of clustering based on the idea that neighbouring data items should share the same cluster. Algorithms implementing this principle are density-based methods (Ankerst *et al*., 1999; Ester *et al*., 1996) and methods such as single-link agglomerative clustering (Vorhees, 1985). They are well-suited for the detection of arbitrarily shaped clusters, but can lack robustness when there is little spatial separation between the clusters.
- *Spatial separation.* Spatial separation on its own is a criterion that gives little guidance during the clustering process and can easily lead to trivial solutions. It is therefore usually combined with other objectives, most notably measures of compactness or balance of cluster sizes. The resulting clustering objectives can be tackled by general-purpose meta-heuristics such as simulated annealing, tabu search and evolutionary algorithms (Bandyopadhyay and Manlik, 2001; Rayward-Smith *et al*., 1996).

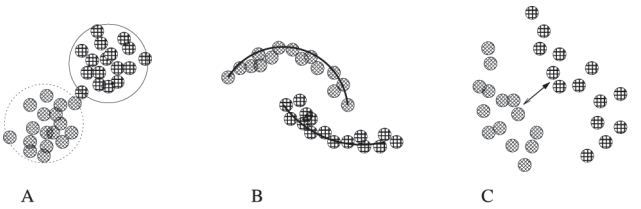

Fig. 2. Examples of datasets exhibiting compactness (**A**), connectedness (**B**) and spatial separation (**C**), respectively. Clearly, connectedness and spatial separation are related (albeit opposite) concepts. In principle, the cluster structure in the datasets B and C can be identified by a clustering algorithm based on either connectedness or on spatial separation, but not by one based on compactness.

Each of these categories is illustrated in Figure 2.

2.2 Clustering in post-genomic data analysis

Unsupervised classification has many applications in post-genomics. In particular, clustering plays a crucial role in the analysis of gene-expression data (Eisen, 1998; Golub *et al*., 1999; Quackenbush, 2001; Slonim, 2002). Clustering can also be applied directly to sequence the data, for example to group genes based on shared *cis*-regulatory regions (Bilu and Linial, 2002). It serves as a data-mining tool to analyse both proteomics and metabolomics data (Goodacre *et al*., 1998), and can be applied in the context of protein comparison and structure prediction (Kaplan *et al*., 2004; Krasnogor and Pelta, 2004). Recently, there have been numerous advances in the development of improved clustering techniques for post-genomic data analysis. Prominent examples include biclustering techniques (Madeira and Oliveira, 2004) and gene shaving (Hastie *et al*., 2000), which have both been specifically designed to deal with the particular challenges posed by gene expression data. Despite such advances, traditional clustering techniques such as hierarchical clustering algorithms (Eisen, 1998), *k*-means (Tavazoie *et al*., 1999), fuzzy *c*-means (Gasch and Eisen, 2002), finite mixture models (Yeung *et al*., 2001b) and SOMs (Tamayo *et al*., 1999) remain the predominant methods in post-genomics—a fact that is arguably more owing to their conceptual simplicity and their wide availability in standard software packages than to their intrinsic merits.

2.3 The need for cluster-validation measures in post-genomic data analysis

While cluster analyses are, potentially, a tool to speed up and semi-automate data processing, the majority of cluster analyses carried out on post-genomic data to date are quite far from this end. This is partly owing to the difficulties of the data tackled. Post-genomic data are typically high-dimensional, contain many more variables than samples, have high levels of noise and may have multiple missing values; these properties pose problems to many traditional clustering methods and makes the cluster analysis very challenging. However, there is hardly any consensus on the best distance function, clustering method or method of feature selection to be used for the different types of post-genomic data. As a consequence, it is common practice

among researchers to employ a variety of different clustering techniques to analyse a dataset, and to use visual inspection and prior biological knowledge to select what is considered the most 'appropriate' result. Clearly, this process of data analysis is highly subjective, and may be a dangerous endeavour. In particular, researchers may unwittingly overrate clusters that reinforce their own assumptions, and ignore surprising or contradictory results. This is, of course, counterproductive to the unspoken aim of unsupervised classification, which is to identify surprising or unexpected patterns in the data that may then serve for hypothesis generation (Kell and Oliver, 2004). Most importantly, while the use of prior biological knowledge and assumptions may be necessary and important in the final interpretation of a cluster analysis, it is not an acceptable means of replacing an unsupervised validation step, in which the significance of individual clusters in terms of the underlying data distribution is verified.

The fact that a validation step is needed follows from the following two issues that arise when using clustering algorithms:

- *Bias of clustering algorithms towards particular cluster properties.* Clustering algorithms are biased towards partitions that are in accordance with their own clustering criterion. This is at the bottom of the fundamental discrepancies observable between the solutions produced by different algorithms.
- *Non-significance of results in the absence of natural clusters.* Unsupervised classification relies on the existence of a distinct structure within the data. However, most clustering algorithms return a clustering even in the absence of actual structure, leaving it to the user to detect the lack of significance of the results returned.

Either of the above may lead to a lack of compliance between a partitioning and the underlying data distribution, a situation which can be detected using computational cluster-validation techniques.

3 SURVEY OF CLUSTERING-VALIDATION TECHNIQUES

The data-mining literature provides a range of different validation techniques, with the main line of distinction between external and internal validation measures (Halkidi *et al.*, 2001). These two groups of techniques differ fundamentally in their focuses, and find application in distinct experimental settings. External validation measures comprise all those methods that evaluate a clustering result based on the knowledge of the correct class labels. Evidently, this is useful to permit an entirely objective evaluation and comparison of clustering algorithms on benchmark data, for which the class labels are known to correspond to true cluster structure. In cases where no class labelling is available, or the available labels are dubious, an evaluation based on internal validation measures becomes appropriate. Internal validation techniques do not use additional knowledge in the form of class labels, but base their quality estimate on the information intrinsic to the data alone. Specifically, they attempt to measure how well a given partitioning corresponds to the natural cluster structure of the data.

The following survey of validation techniques is limited to those for crisp partitionings, that is, partitions in which each data item is assigned exactly one label (cf., for example, Pal and Bezdek, 1995 for more information regarding the evaluation of fuzzy partitionings). Mathematical definitions for selected validation techniques are provided in the Supplementary Material.

3.1 External measures

3.1.1 Type 1: Unary measures Standard external evaluation measures take a single clustering result as the input, and compare it with a known set of class labels (the 'ground truth' or 'gold standard') to assess the degree of consensus between the two. Traditionally, the gold standard would be complete and unique, in the sense that exactly one class label is provided for every data item, and that the label is unequivocally defined. A partitioning can then be evaluated both with regard to the purity of individual clusters and the completeness of clusters. Here, purity denotes the fraction of the cluster taken up by its predominant class label, whereas completeness denotes the fraction of items in this predominant class that is grouped in the cluster at hand. Clearly, both these aspects provide a limited amount of information only, and trivial solutions for both of them exist such as a partitioning consisting of singleton clusters (scoring maximally under purity), and a one-cluster solution (scoring maximally under completeness). In order to obtain an objective assessment of a partition's accordance with the gold standard, it is therefore important to take both purity and completeness into account. Comprehensive measures like the F-measure (see Supplementary Material) (van Rijsbergen, 1979) provide a principled way to evaluate both of these and are therefore preferable over simpler techniques.

Note that techniques like the F-measure provide a means to assess the quality of a clustering result at the level of the entire partitioning, and not for individual clusters only. In principle, such measures can also be adapted for use with a 'partial labelling' (i.e. for use in a setting where only incomplete labelling information is available—such as functional annotation for a fraction of genes in a microarray experiment) by applying the measure to the labelled data and their respective cluster assignments only. This can provide a more comprehensive way of assessing clustering quality than does the computation of corrected significance levels of the 'enrichment' (Gat-Viks *et al.*, 2003; Tavazoie *et al.*, 1999; Toronen, 2004) of individual selected clusters.

3.1.2 Type 2: Binary measures In addition to measures based on purity and completeness, the data-mining literature also provides a number of indices, which assess the consensus between a partitioning and the gold standard based on the contingency table of the pairwise assignment of data items. Most of these indices are symmetric, and are therefore equally well-suited for the use as binary measures, that is, for assessing the similarity of two different clustering results.

Probably the best known such index is the Rand Index (see Supplementary Material) (Rand, 1971), which determines the similarity between two partitions as a function of positive and negative agreements in pairwise cluster assignments. A number of variations of the Rand Index exist, in particular the adjusted Rand Index (Hubert, 1985), which introduces a statistically induced normalization in order to yield values close to zero for random partitions. Another related index is the Jaccard coefficient (Jaccard, 1908), which applies a somewhat stricter definition of correspondence in which only positive agreements are rewarded. Note that not all indices based on contingency tables are symmetric. The Minkowski Score (see Supplementary Material) (Jardine and Sibson, 1971), for example, is

asymmetric [i.e. $M(U,V) \neq M(V,U)$ for two partitionings U and V] and therefore less suited for assessing the similarity between clustering results.

3.2 Internal measures

Internal measures take a clustering and the underlying dataset as the input, and use information intrinsic to the data to assess the quality of the clustering. Using the same categorization as for clustering methods (see Section 2.1), the first three types of internal measures can be grouped according to the particular notion of clustering quality that they employ.

3.2.1 Type 1: Compactness A first group comprises validation measures assessing cluster compactness or homogeneity, with intra-cluster variance (see Supplementary Material) (also sum-of-squared-errors minimum variance criterion, the measure locally optimized by the k-means algorithm) as their most popular representative. Numerous variants of measuring intra-cluster homogeneity are possible such as the assessment of average or maximum pairwise intra-cluster distances, average or maximum centroid-based similarities or the use of graph-based approaches (Bezdek and Pal, 1998).

3.2.2 Type 2: Connectedness The second type of internal validation technique attempts to assess how well a given partitioning agrees with the concept of connectedness, i.e. to what degree a partitioning observes local densities and groups data items together with their nearest neighbours in the data space. Representatives include k-nearest neighbour consistency (Ding and He, 2004) and connectivity (see Supplementary Material) (Handl and Knowles, 2005), which both count violations of nearest neighbour relationships.

3.2.3 Type 3: Separation The third group includes all those measures that quantify the degree of separation between individual clusters. For example, an overall rating for a partitioning can be defined as the average weighted inter-cluster distance, where the distance between individual clusters can be computed as the distance between cluster centroids, or as the minimum distance between data items belonging to different clusters. Alternatively, cluster separation in a partitioning may, for example, be assessed as the minimum separation observed between individual clusters in the partitioning.

3.2.4 Type 4: Combinations The literature provides a number of enhanced approaches that combine measures of the above different types. In this respect, combinations of type one and type three are particularly popular, as the two classes of measures exhibit opposing trends: while intra-cluster homogeneity improves with an increasing number of clusters, the distance between clusters tends to deteriorate. Several techniques therefore assess both intra-cluster homogeneity and inter-cluster separation, and compute a final score as the linear or non-linear combination of the two measures. An example of a linear combination is the SD-validity Index (see Supplementary Material) (Halkidi *et al.*, 2001);[2] well-known examples of non-linear combinations are the Dunn Index (see Supplementary Material) (Dunn, 1974), Dunn-like Indices (Bezdek and Pal, 1998), the Davies–Bouldin Index (Davies and Bouldin, 1979) or the Silhouette Width (see Supplementary Material) (Rousseeuw, 1987).

[2]SD refers to the fact that this index measures the scattering and the distance of clusters.

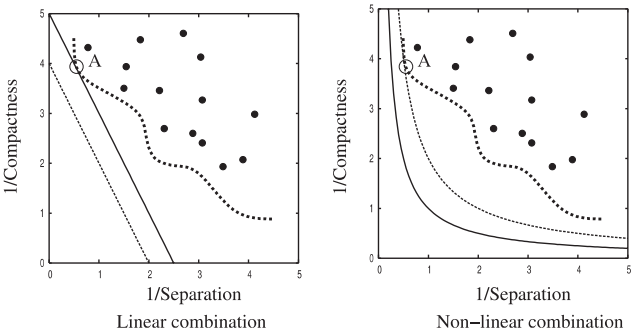

Fig. 3. Illustration of the linear and non-linear combination of two objectives. Here, the two objectives are to be minimized. On the left-hand side the solid lines indicate lines of equal measure under $c = \alpha \cdot x + \beta \cdot y$ (linear combination), where α and β are the relative weights assigned to objectives x and y. On the right-hand side the solid lines indicate lines of equal measure under $c = x \cdot y$ (non-linear combination). In both cases, the solutions on lines/curves closer to the origin are rated higher (c is smaller). A possible Pareto front, that is, the set of solutions that are Pareto optimal with respect to the two objectives, is shown as a dashed line. It can be seen that in both cases Solution A scores best under c, while all other Pareto optimal solutions are overlooked.

While the above methods are relatively popular, the linear or non-linear combination of the measures inevitably results in a certain information loss (Fig. 3), and can therefore lead to incorrect conclusions. An alternative and more principled way of evaluating N measures simultaneously is the evaluation of the resulting N-tuples with respect to Pareto optimality (Pareto, 1971): a clustering result is judged to dominate (be superior to) another partitioning, if it is equal or better under all measures, and is strictly better under at least one measure. A recent study using Pareto optimality for clustering and validation with regard to a type one and a type two measure can be found in (Handl and Knowles, 2005).

3.2.5 Type 5: predictive power/stability Validation techniques assessing the predictive power or stability of a partitioning form a special class of internal validation measures. They are clearly not external since as they do not make use of label information. However, they are quite different from traditional internal measures in that their use requires additional access to the clustering algorithm. Measures of this type repeatedly re-sample or perturb the original dataset, and re-cluster the resulting data. The consistency of the corresponding results provides an estimate of the significance of the clusters obtained from the original dataset.

The methods described in (Ben-Hur *et al.*, 2002, http://psb.stanford.edu/psb-online/; Bittner *et al.*, 2000; Breckenridge, 1989; Fridlyand and Dudoit, 2001, http://www. stat.berkeley.edu/sandrine/tecrep/600.pdf; Kerr and Churchill, 2001; Lange *et al.*, 2004; Levine and Domany, 2001; Li and Wong, 2001; McShane *et al.*, 2002; Tibshirani *et al.*, 2001a, http://www-stat.stanford.edu/tibs/ftp/predstr.ps) employ the concept of self-consistency, that is, the idea that a clustering algorithm should produce consistent results when applied to data sampled from the same source. In order to assess the degree of stability of a partitioning, several papers (Ben-Hur *et al.*, 2002; Levine and Domany, 2001) repeatedly draw overlapping subsamples of the same dataset (the individual subsamples are drawn without replacement). Each subsample is clustered individually, and the resulting partitions are compared by applying an external validation index to the partial

partitions obtained for the overlapping shared set of points. A slightly different approach has been taken in (Breckenridge, 1989; Fridlyand and Dudoit, 2001; Lange *et al*., 2004; Tibshirani *et al*., 2001a). Here the data are split repeatedly into a training and a test set (typically of equal size and with no overlap), and both sets are clustered. The partitioning on the training set is then employed to derive a classifier to predict all class labels for the test set. The disagreement between the prediction and the partitioning on the test set can then be computed using an external binary validation index. Obviously, the classifier used for prediction has a significant impact on the performance of this method and should comply with the modelling assumptions made by the clustering algorithm. Lange *et al*. (2004) recommend the use of a nearest-neighbour classifier for single link, and of centroid-based classifiers for algorithms such as *k*-means that assume spherically shaped clusters. Finally, the stability of a clustering result can also be assessed by comparing the partitions obtained for perturbed data (Bittner *et al*., 2000; Kerr and Churchill, 2001; Li and Wong, 2001). For this purpose, a number of bootstrap datasets are generated from the original data: using a simple error model (Bittner *et al*., 2000) or more advanced methods such as ANOVA (Kerr and Churchill, 2001), a noise component is added to each data item. The resulting datasets (in which data items are slightly perturbed with respect to their original position) are subjected to a cluster analysis. The partitions obtained can then be directly compared using external binary indices (i.e. by comparing the cluster assignments for data vectors derived from the same original data point).

3.2.6 Type 6: Compliance between a partitioning and distance information An alternative way of assessing clustering quality is to estimate directly the degree to which distance information in the original data is preserved in a partitioning. For this purpose, a partitioning is represented by means of its cophenetic matrix C (Romesburg, 1984), where C is a symmetric matrix of size $N \times N$ and N is the size of the dataset. In a crisp partitioning, the cophenetic matrix contains only zeros and ones, with each entry $C(i, j)$ indicating whether the two elements i and j have been assigned to the same cluster or not. For the evaluation of a hierarchical clustering, the cophenetic matrix can also be constructed to reflect the structure of the dendrogram. Here, an entry $C(i, j)$ represents the level within the dendrogram at which the two data items i and j are first assigned to the same cluster.

The cophenetic matrix can then be compared to the original dissimilarity matrix using Hubert's Γ Statistic (essentially the dot-product between the two matrices), the Normalized Γ Statistic, or a measure of correlation such as the Pearson correlation (Edwards, 1967) (in cases where the prime emphasis is on the preservation of absolute distance values) or the Spearman rank correlation (Lehmann and D'Abrera, 1998) (in cases where the prime emphasis is on the preservation of distance orderings). The correlation between the two matrices is commonly referred to as cophenetic correlation, matrix correlation or standardized Mantel Statistic (Halkidi *et al*., 2001). As an aside, cophenetic correlation can also be used as a binary index to assess the preservation of distances under different distance functions and within different feature spaces, or to compare the dendrograms obtained for different algorithms.

3.2.7 Type 7: Specialized measures for highly correlated data This last category of internal validation measures includes a number of techniques that explicitly exploit redundancies and correlations such as those inherent to post-genomic data. The first of these, the figure of merit (Yeung *et al*., 2001a), is motivated by the jacknife (Efron and Tibshirani, 1993) approach. For a dataset with D features, the figure of merit of Yeung *et al*. (2001) requires the computation of D partitions, each of them based on only $D - 1$ out of the D features. For each partitioning, its figure of merit is then computed as the average intra-cluster variance within the unused feature, and the aggregation of these values provides an estimate of the overall performance of the algorithm. Datta and Datta (2003) extend this approach to the computation of a figure of merit by means of different internal validity indices—specifically, one measure of pairwise co-assignment, one of cluster separation and one of cluster compactness.

The second approach, overabundance analysis (Ben-Dor *et al*., 2002, http://www.cs.huji.ac.il/nirf/Abstracts/BFY2Full.html; Bittner *et al*., 2000) assesses the frequency of discriminatory variables for a given partitioning, that is, it identifies those variables that show significant differences between the identified clusters. The observed frequencies are compared with a null-model to assess the significance of the partitioning.

Clearly, both of the above approaches are only applicable to datasets with correlated (dependent) variables, but this is likely to be true for most types of post-genomic data (such as gene expression data and protein and metabolome profiles).

3.3 Number of clusters

Most of the internal measures discussed above can be used to estimate the number of clusters in a dataset, which usually involves the computation of clustering results for a range of different numbers of clusters, and the subsequent plot of the performance under the internal measure as a function of the number of clusters. If both the clustering algorithm employed and the internal measure are adequate for the dataset under consideration, the best number of clusters can often be identified as a 'knee' in the resulting performance curve. See, e.g., the Gap Statistic (Tibshirani *et al*., 2001b) for a formalized approach. This type of use in model selection has been the most common application of internal validation measures in bioinformatics (Bolshakova and Azuaje, 2003; Bolshakova *et al*., 2005; Fridlyand and Dudoit, 2001; Lange *et al*., 2004; Tibshirani *et al*., 2001b), yet, their more broad applicability in the validation of cluster quality has been frequently neglected.

3.4 Statistical tests of clustering tendency

The previous sections have been concerned with the validation of a clustering result obtained on a given dataset. However, when in doubt about the quality of a raw dataset, it may be useful to apply statistical tests to examine the clustering tendency of the data prior to conducting a cluster analysis (McShane *et al*., 2002). For this purpose, the distribution of nearest neighbour distances can be examined, and compared with that under a suitable null model. However, the sparseness of data in high dimensions may lead to instabilities, and it is therefore recommended to apply this type of test to low-dimensional data only. An example is provided by McShane *et al*. (2002), where the data are subjected to a principal components analysis and projected to the first three principal components. The principal components can further be used to generate an appropriate null model, in the above case a three-dimensional Gaussian distribution with means and standard deviations in each dimension estimated from the original data.

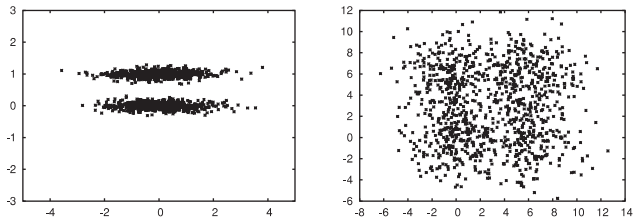

Fig. 4. Two-dimensional datasets 'Long' and 'Square', both generated from Normal Distributions. Square contains four clusters with strong overlap that are difficult to detect for algorithms and measures based on connectivity or spatial separation. Long contains two elongated clusters that are difficult to detect for algorithms and measures based on compactness. Note that these two datasets are simplistic and have been selected for demonstration purposes only. While, for these particular data, the problems on Long could be overcome by normalization, it is realistic to assume that this may not always be possible for real datasets (which usually contain several differently shaped clusters).

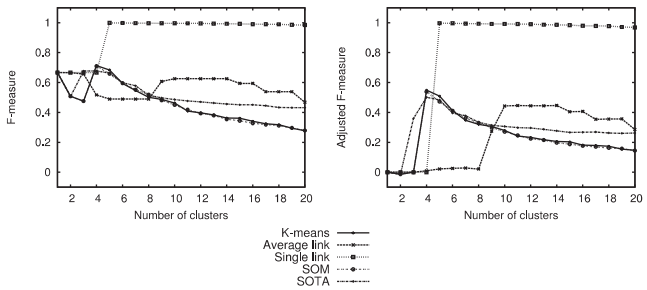

Fig. 5. Illustration of the biases of external validation measures. Shown are the results for k-means, SOM, SOTA, average link and single link on the Long dataset under (left) the F-measure, and (right) the adjusted F-measure (averages over 21 runs). The original F-measure values indicate that both single and average link clearly outperform k-means, SOM and SOTA for $k = 2$, by a margin of up to 0.2. However, this conceals the fact that for this cluster number all five algorithms have equally failed to identify the correct cluster structure on Long. While k-means, SOM and SOTA have split both clusters in the middle (minimizing variance), both agglomerative clustering algorithms have isolated outliers in one cluster and merged the bulk of the data in the second cluster (see Fig. 6). Only for $k \geq 5$ does single link succeed in separating the two core clusters. However, the poor performance of all five algorithms for $k = 2$ becomes evident in the plot of the adjusted F-measure. In general, the normalization may not only correct the estimated absolute degree of quality, but may also correct the ordering between the solutions obtained for different numbers of clusters (e.g. average link for $k = 2$ and $k = 10$) or for different algorithms (e.g. average link and k-means for $k = 9$).

4 MEASURE-SPECIFIC BIASES

In the following section, some of the major factors affecting certain validation techniques are discussed, and are demonstrated on two exemplary two-dimensional datasets (shown in Fig. 4). For the purpose of this study, several advanced techniques are selected from the categories described above. These are the F-measure (takes values in [0,1], to be maximized), the adjusted Rand Index (takes values in [0,1], to be maximized), variance (to be minimized), connectivity (to me minimized), Silhouette Width (takes values in $[-1,1]$, to be maximized), the Dunn Index (to be maximized) and a stability-based method (takes values in [0,1], to be maximized). Experiments are conducted using five different clustering algorithms namely the partitioning method k-means, SOM, the self-organizing tree algorithm (SOTA) and two agglomerative hierarchical algorithms based on the linkage criteria of single link and average link, respectively. Enlarged colour plots and implementation details for the individual measures and algorithms are provided in the Supplementary Material.

4.1 Biases of external measures

External validation techniques suffer from biases with respect to the number of clusters, the distribution of cluster sizes and the distribution of class sizes in a partitioning (Halkidi *et al.*, 2001; Hubert, 1985). For example, the completeness of a cluster trivially obtains the maximum possible value of 1 for a one-cluster partitioning and tends to decrease with an increasing number of clusters. On a different line, the F-measure and the Rand Index tend to be overly optimistic in situations where relatively small clusters have been overlooked.

These unwanted effects can be alleviated through normalization by the results expected for random data. The adjusted measure E_a is obtained as $E_a = (E - E_e)/(E_m - E_e)$, where E_e is the expected value for random data and E_m is the maximum attainable value of the index. Ideally, this adjusted index E_a will then be limited to the interval [0,1]. Most commonly, this procedure has been applied to the Rand Index for which the expected value (and thus the adjusted Rand Index) can be computed by exact statistical methods (Hubert, 1985). However, the use of normalization is generally useful for any external measure. In cases where the expected value cannot be statistically derived, it can be approximated using Monte Carlo simulation. For a given partition, a number of 'random partitions' are generated, which agree with the original partition in the number of clusters, the distribution of cluster sizes and the distribution of class sizes. Each of these partitions is evaluated under the validation measure considered, and the average value is taken as an approximation of the expected value E_e. The required random partitionings may simply be obtained by permuting the cluster labels in the original partitioning.

This normalization of external indices is crucial to obtain an objective picture of the real, absolute and relative performance of an algorithm. Figure 5 illustrates the degree to which the values returned by the F-measure can be misleading, if un-normalized. In order to facilitate the interpretation of these performance curves, a selection of the actual clustering results is shown in Figure 6.

4.2 Biases of internal measures

Just like external measures, most internal measures suffer from biases with regard to the number of clusters. More importantly, internal measures may additionally exhibit biases with regard to the shape of the underlying data manifold and the structure of a partitioning. Some of these biases can be detected through a comparison with the results expected for random data. Computation of the expected value I_e for an internal measure I can be done by Monte Carlo simulation: a number of random control datasets are generated under an appropriate null model. They are then clustered (using the same clustering algorithm as applied to the original data) and the resulting partitions are evaluated under I. The average value obtained is taken as an approximation of the expected value I_e. Importantly, different possibilities for the choice of the null model

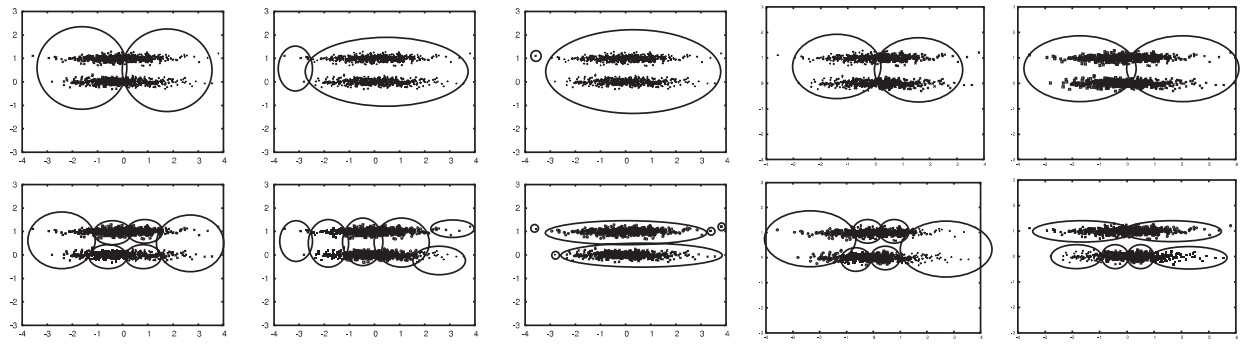

Fig. 6. Clustering results on Long for (from left to right) k-means, average link, single link, SOM and SOTA for (top) $k = 2$ and (bottom) $k = 6$.

exist, and the specific model used may have a crucial impact on the final outcome (Gordon, 1999). The main lines of distinction are between data-independent null models (e.g. a Poisson model or a Unimodal model) and data-influenced null models (e.g. an ellipsoidal model) (Gordon, 1999). The success of this technique strongly depends on the ability of the null model employed to capture the shape of the data manifold. Yet, it may help to detect those biases that result from a change in the number of clusters or the shape of the underlying data distribution.

The biases of internal measures are illustrated in Figure 7, where the application of several internal measures on the Long dataset is shown. The results obtained by different measures are only partially consistent, which is owing to several factors. First, for most internal measures that assess cluster compactness or ratios of inter-cluster and intra-cluster distances, the shape of the data manifold in this dataset introduces a bias towards a vertical split. This bias could be identified through the comparison with the results for uniformly random control data. Second, the classification of outliers in their own clusters can have a significant impact on the final result of a performance measure, in particular, if minimum or maximum pairwise distances are taken into account (this is the case for the Dunn Index). This problem can only to a certain degree be tackled by the elimination of outliers prior to clustering. Third, more fundamentally, k-means, SOM, SOTA and average link strive for spherically shaped compact clusters and are likely to perform reasonably well under the Silhouette Width even without the discovery of any cluster structure. Fourth, the clusters in the dataset are elongated and the correct partitioning therefore does not score highly under the Silhouette Width (or any other measure based on cluster compactness). For this reason the good performance of single link for $k \geq 5$ is not manifested in the plot of the Silhouette Width.

While underlining the benefits of the comparison with a null model, the above example also makes clear that such a step cannot entirely remove the biases of a measure with regard to particular cluster structures. Owing to the conflict between the assumptions of the Silhouette Width and the real cluster structure, it is impossible to identify the correct number of clusters for single link in the Silhouette plot. The results under variance and connectivity contain clues as to the best clustering solution, but the results are largely dominated by the measures' biases towards particular algorithms, making it hard to arrive at the correct conclusion. In the plot of variance for single link, the approximation of the correct solution (for $k = 5$) manifests itself in a small 'knee'—yet, the objective values obtained by k-means, SOM, SOTA and average link are far better. Simultaneously, single link

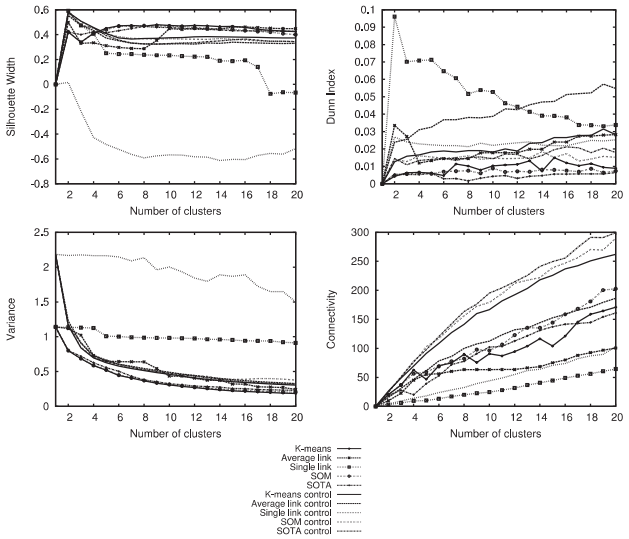

Fig. 7. Illustration of the biases of internal validation measures. A simple null model is used: the random control data have been generated from a uniform random distribution within the bounds of the original data. Shown are the results for k-means, SOM, SOTA, average link and single link on the Long dataset under (top left) the Silhouette Width, (top right) the Dunn Index, (bottom left) variance, and (bottom right) connectivity (averages over 21 runs) on the original data and uniformly random control data. The performance curves obtained for both original and uniformly random control data are plotted to permit visual comparison. Considering the Silhouette Widths only, the plot for the original data seems to indicate that $k = 2$ yields a good partitioning for all five algorithms, with the agglomerative algorithms being the best performers. Moreover, single link is assessed to perform worse than all other methods for all $k \geq 5$, a result that stands in obvious contrast to its true performance as verified by the adjusted F-measure (see Fig. 5). Only a comparison with the values obtained for random control data reveals that, for $k = 2$ and $k = 3$, k-means, SOM, SOTA and average link do in fact perform worse than for random data. In comparison with its performance on random control data, single link seems to perform well, but it is not possible to derive the correct number of clusters from the plot (there is no peak at $k = 5$). Interpretation of the other three measures is given in the text.

largely outperforms k-means, SOM, SOTA and average link under connectivity, and its best solution manifests itself in a weak plateau in the performance curve—yet, connectivity is clearly strongly biased towards single link, and may not be trustworthy. Without additional

knowledge, it will not be clear as to which algorithm and solution to choose.

The above data demonstrates the difficulty of selecting one internal validation measure that permits the objective quantification of a range of conceptually different algorithms. Both clustering algorithms and internal validation measures are based on certain assumptions about the cluster structure, which results in biases of measures with regard to specific algorithms (owing to shared underlying assumptions). An understanding of these biases is therefore crucial to ensure a valid assessment of clustering results by means of internal measures. In particular, it is essential to comprehend the working principles of the algorithms and the evaluation measures used and to select a combination of measures and algorithms that permits one to draw meaningful conclusions.

4.2.1 Complementary validation measures Evidently, type-4-validation techniques like the Silhouette Width constitute an attempt to combine measures and thereby reduce their individual biases. However, as outlined in Section 3.2.4, these existing methods are restricted to one fixed (linear or non-linear) combination of the two measures and may therefore still exhibit strong biases towards one or the other measure (as seen in the previous example). A more rigorous approach is the independent use of two or three complementary measures and the subsequent visualization of solutions in two- or three-objective space. In principle, such plots can be generated for any pair of measures, but they are particularly useful for the visualization of the results obtained using conflicting measures such as compactness versus separation, or compactness versus connectivity. Figure 8 demonstrates this approach using the measures of variance and connectivity.

This visualization has several advantages over traditional performance curves. First, it allows one to summarize information regarding the algorithms' performance under both internal validity measures. Second, the set of solutions returned by the different algorithms can be automatically reduced using the concept of Pareto optimality, and all Pareto optimal solutions can be identified. Third, it clearly demonstrates the behaviour of the different algorithms with respect to the two objectives. Figure 8 reveals both single link's tendency to isolate singleton clusters (reflected in the lack of improvement in variance) and k-means' tendency to partition the data into equally sized chunks without consideration of the underlying data distribution (reflected in the quick deterioration in connectivity). Moreover, the best solution—generated by single link for $k = 5$—is identifiable here.

4.3 Biases of stability-based techniques

Stability-based techniques employ a less stringent definition of clustering quality than do traditional internal validation techniques and, therefore, do not suffer from the same biases towards particular algorithms. However, while being reliable indicators of clustering quality in many cases, stability-based techniques may also be misleading under certain circumstances. This predominantly concerns their application to datasets in which the shape of the data manifold causes a given clustering algorithm to converge reliably to certain suboptimal solutions. Under such conditions a clustering may appear stable under re-sampling/perturbation, while not corresponding to the real structure of the dataset. Figure 9 demonstrates these issues on the Long and Square dataset. Further issues with stability-based techniques have been pointed out in (Breckenridge, 2000; Krieger and Green, 1999).

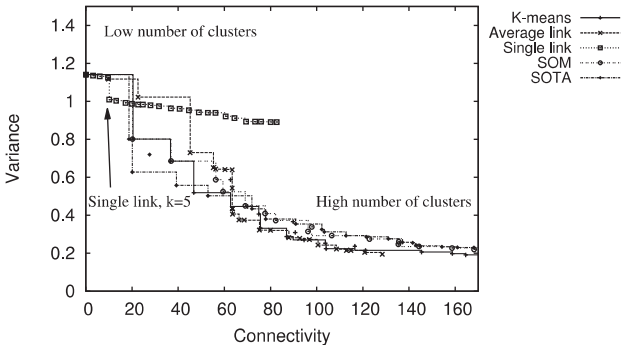

Fig. 8. Illustration of the solution visualization in two-objective space. Shown are the solutions (averages over 21 runs) for k-means, SOM, SOTA, average link and single link on the Long dataset in a plot of connectivity versus variance (both to be minimized). The set of solutions returned by each clustering algorithm is summarized by an attainment surface (Fonseca and Fleming, 1996), which is the boundary in objective space that separates the region dominated by the attained solutions from the region that is not dominated. Owing to the trends of the two objectives (variance decreases for a higher number of clusters, while connectivity increases), the number of clusters in the solutions generally increases from the top left to the bottom right. The correct number of clusters for a given algorithm is expected to show as a strong 'knee' in its attainment front. This is because for the correct number of clusters we expect a relatively large drop in variance at little additional cost in connectivity. In the above plot, a clear 'knee' can be observed for single link at $k = 5$.

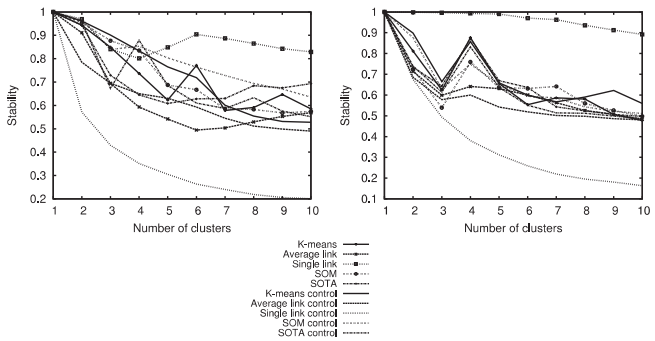

Fig. 9. Stability-based analysis (averages over 21 runs) for (left) the Long and (right) the Square dataset. On Long, the stability-based method is evidently able to estimate the presence of a good single link clustering solution for $k \geq 5$ (the stability plot peaks at $k = 6$), and the absence of good SOM and average link solutions (no significant peak in the stability plot). Yet, the results obtained for k-means, SOM and SOTA are disconcerting. For k-means, the stability-based analysis pinpoints a highly stable six-cluster solution. Further analysis shows that this solution is stable only due to k-means' tendency to converge to spherically shaped clusters, and is in fact highly sub-optimal (see Fig. 6). For SOM and SOTA, the stability-based analysis pinpoints a stable four-cluster solution for the random control data. For Square, where k-means, SOM, SOTA and average link perform comparably well (data not shown), the stability-based analysis correctly identifies the four-cluster solution for average link, k-means, SOM and SOTA (the stability plots for all four algorithms peak at $k = 4$). However, a comparison with the control curves reveals that, for k-means, SOM and SOTA, a four cluster solution is recommended with comparable confidence for uniformly random data. This shows that the conspicuous stability peak obtained for k-means, SOM and SOTA on the Square data is mainly an artefact of the square shape of the underlying data manifold.

5 GUIDELINES FOR EFFECTIVE CLUSTER VALIDATION

In the previous section, the strengths and weaknesses of different validation techniques have been discussed. Two sample datasets were used to demonstrate that the results returned by individual validation techniques can be biased and misleading under certain circumstances, but also that there exist means of detecting several of these biases. Ultimately, despite their imperfections, validation measures do provide significant amounts of information that cannot be obtained using visual inspection alone. Different and complementary validation tools exist, and the use of a set of such tools can minimize the risk of misinterpreting results, and thereby maximize confidence in the results obtained.

Cluster analysis is a complicated interactive process, which makes it impossible to provide an entirely clear-cut prescription on how to do clustering or to perform cluster validation. In general, the experimental set-up should fundamentally differ depending on the primary aim of a study. Cluster validation aimed at the evaluation of a novel algorithm or the comparison of several algorithms should be quite different to the type of cluster validation used during the analysis of a novel biological dataset. This section attempts to give some general guidelines on the conduct of an effective cluster validation in both scenarios.

5.1 Cluster validation for the evaluation/comparison of algorithms

When evaluating algorithms, the choice of datasets is a primary issue. Certainly, several datasets should be used, not just one—especially not only the dataset the algorithm was initially developed on. It is fundamental to appreciate that algorithms make different assumptions about the cluster structures, and are, consequently, more or less suited for particular datasets: no single algorithm can therefore be expected to perform well for all types of data (Gordon, 1999). Thus, the aim of any evaluation study should not be to show that a particular algorithm is the best overall, but to show what the particular strengths and weaknesses of a given algorithm are. For this purpose, it is important to test on benchmarks with interesting known data properties. In this scenario, two types of questions are then of interest, which are both essential to understand fully the outcome of an experiment.

- *How well does the algorithm perform on a given dataset?* On benchmarks, this type of question can be objectively answered using external cluster validation. The use of adjusted validity measures is preferable.
- *Why is the algorithm not performing well? What is going wrong?* Internal validation technique can be used to highlight these issues, particularly those of Type 1, Type 2 and Type 3, and their combination in Pareto plots, as these have straightforward interpretations in terms of data properties.

5.2 Cluster validation for a novel dataset

When clustering a novel biological dataset, cluster validation plays a very different role. A completely objective validation of cluster quality is usually impossible in such a case, but the use of cluster validation at different steps during the clustering process can help to improve the quality of results, and increase the confidence in the final result. Cluster analysis usually involves a first exploratory step, where the data are visualized (projected) to two- or three-dimensions (using methods such as principal components analysis or multi-dimensional scaling) in order to check for clustering tendencies. At this stage, a statistical test of clustering tendency (see Section 3.4) may help to quantify the visual impressions obtained.

No entirely reliable method exists to identify the number of clusters in a dataset, and the choice of the best number of clusters may well depend on the clustering method used. A cluster analysis should therefore always be performed for a (sensible) range of different numbers of clusters. Access to such a sequence of solutions is essential to understand the operation of a clustering algorithm and to identify trends in the data.

The core cluster analysis should preferably be conducted using several conceptually distinct clustering algorithms, i.e. algorithms that are not biased towards the same type of clusters. Binary external indices can then be used to quantify analytically the similarity between clustering results (including those with different numbers of clusters). If conceptually different algorithms generate highly similar partitions, this is a good indicator that actual structure has been discovered. On the other hand, coinciding clustering results returned by k-means, partitioning-around medoids or SOMs are less significant, as these algorithms share many concepts. If the partitions generated by different algorithms are highly dissimilar this is often an indication of poor structure in the data, and may point to defects in the pre-processing. In high-dimensional biological data, the structures in the data cannot often be perceived in the full feature space, and a drastic reduction of variables may be necessary in order to reduce the impact of noise (Shaw *et al*., 1997). This process of feature selection is often necessary but should preferably be based on unsupervised methods (e.g. by selecting the variables with the highest variation across the dataset). If the features are selected using the knowledge of the real class labels (e.g. by selecting the variables which are best correlated with the known class structure), a subsequent cluster analysis will trivially yield the desired result (even for random data).

Internal validation measures should be used in addition to the above to provide feedback on the quality of the data and to check whether a given partitioning is justified in terms of the underlying data distribution. Here, it is important to use measures of the different basic types, Type 1, Type 2 and Type 3, and to check how well the solutions perform under each of them. A good clustering solution tends to perform reasonably well under multiple measures (Handl and Knowles, 2005). If a solution performs well only under one of them, this is likely to be an artefact of the biases of the employed algorithm. Type 4 measures and plots in two-objective space may be a valuable tool in identifying solutions that perform consistently well. Given the noisy nature of biological data, robust measures like the Silhouette Width are generally preferable to noise-sensitive measures such as the Dunn Index.

Owing to the many sources of noise and the high dimensionality of the data, the above internal validation techniques on their own may often be insufficient in biological data analysis. Frequently, the most conspicuous structure in the data may be artefacts due to experimental factors. On the one hand, cluster analysis can be a valuable tool in identifying such artefacts. On the other hand, the artefacts will ultimately have to be removed if a researcher is interested in biologically meaningful results. Towards this goal, external unary measures can be applied to assess the degree of preservation of replicate-relationships, or of prior biological knowledge. This information can then provide additional feedback on the quality of

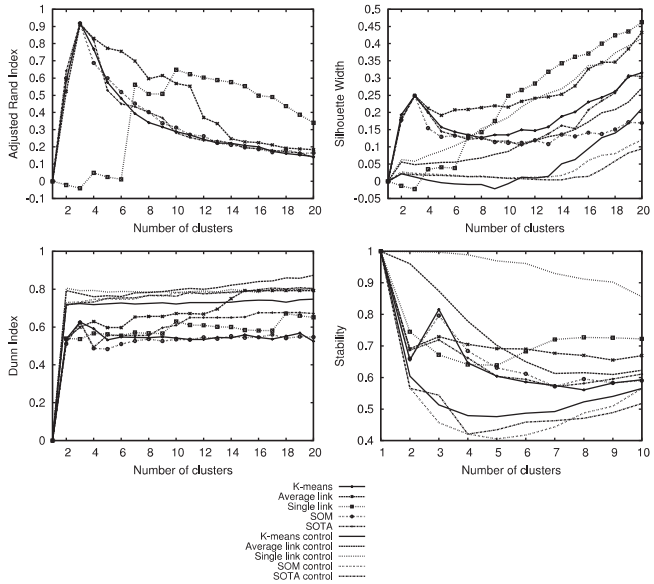

Fig. 10. Adjusted Rand Index, Silhouette Width, Dunn Index and stability (averages over 21 runs) for *k*-means, SOM, SOTA, average link and single link agglomerative clustering on the Leukemia test set. The evaluation under the adjusted Rand Index (comparing to the known class labels) shows that average link, *k*-means, SOTA and SOM perform robustly on these data. They identify the three main clusters (AML, B-lineage ALL and T-lineage ALL), and assign most of the samples correctly. Naturally, this is knowledge that would not be available in a real-life cluster analysis, and it is therefore interesting to see whether the results under the internal validation measures would have led to the same conclusion. The performance curves under the Silhouette Width clearly indicate the high quality of the three-cluster solution. This result is best seen when comparing with the results obtained under the null model and 'removing' the bias towards large numbers of clusters, which arises due to the small size of the dataset: for large numbers of clusters, the resulting partitionings contain many singleton clusters, which score highly under the Silhouette Width. The stability-based technique is less consistent: for *k*-means and SOM, the performance peak at $k = 3$ is well pronounced, but it is much weaker for SOTA and average link. Both the Silhouette Width and the stability-based method indicate the lack of structure in the single link solutions. The application of the Dunn Index is somewhat less successful: it fails to predict the insufficiency of single link, and it mis-estimates the number of clusters for average link.

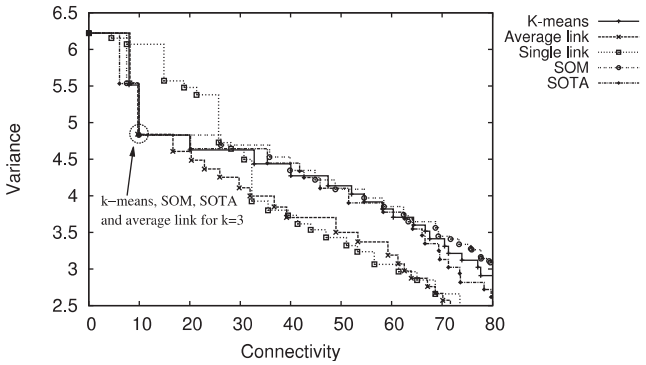

Fig. 11. Solution visualization in two-objective space. Shown are the solutions (averages over 21 runs) for *k*-means, SOM, SOTA, average link and single link on the Leukemia dataset in a plot of connectivity versus variance. The knee corresponding to the three-cluster solution is clearly pronounced. The visualization also shows the consistency among the *k*-means, SOM, SOTA and average link solutions for $k = 2$ and $k = 3$, which further increases the confidence in the correctness of these partitionings.

the data and of previous pre-processing steps. A good final clustering result will ideally combine validity under both internal and external measures, i.e., it will exhibit a distinct underlying cluster structure while being consistent with prior biological knowledge.

6 SAMPLE APPLICATION

In this last section, a brief example of a cluster analysis on gene expression data is given, in order to demonstrate the power of validation measures as a tool to provide insight into the structure of a dataset, and to assess the performance of individual clustering algorithms. The dataset employed is Golub *et al.*'s (1998) Leukemia dataset (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). The aim is to conduct an unsupervised analysis, and the genes used for the clustering are therefore selected in a completely unsupervised fashion.

The data are subjected to a series of standard pre-processing steps: lower and upper threshold values (raw expression values of 100 and 16 000, respectively) are applied, the 100 genes with the largest variation across samples are selected, and the remaining expression values are log-transformed. The resulting dataset of size 38×100 is subjected to a cluster analysis under Euclidean distance. The corresponding validation results are presented in Figures 10 and 11. Altogether, evidence accumulation over the set of employed validation techniques indicates a high quality of the three-cluster solution discovered by *k*-means, SOM, SOTA and average link. This three-cluster solution corresponds to an almost perfect separation of the samples of acute leukaemias into those arising from myeloid precursors (AML), and two sub-classes arising from lymphoid precursors (T-lineage ALL and B-lineage ALL).

7 CONCLUSION

The aim of this paper has been to familiarize researchers using post-genomic measurements with the multitude of validation techniques available for cluster analysis. For this purpose, the different types of validation measures have been reviewed, and specific weaknesses of individual measures have been addressed. It is hoped that the analysis provided has demonstrated not only the importance, but also the intricacy of cluster validation. It is fundamental to comprehend that the use of analytical validation techniques on their own is not sufficient, but that an understanding of the working principles of clustering algorithms, validation measures and their interactions is crucial to enable fair and objective cluster validation. Owing to the biases intrinsic to many internal validation techniques, a careful analysis of the results obtained is required, and results should always be double-checked using alternative complementary validation techniques.

Researchers should be aware that entirely objective cluster validation is possible only on the data with known well-defined cluster structures and the development and evaluation of new clustering algorithms should therefore always include such data. In this context, the development of synthetic datasets that realistically

mimic the properties of biological data [such as simulated gene-expression data (Mendes *et al*., 2003; Michaud *et al*., 2003)] are of particular importance as such an approach permits a controlled study of an algorithm's sensitivity with respect to specific data properties.

ACKNOWLEDGEMENTS

The authors would like to thank Oliver Sander and Roy Goodacre for proofreading and valuable feedback. J.H. acknowledges support of a scholarship by the Gottlieb Daimler- and Karl Benz-Foundation. J.K. is supported by a BBSRC David Phillips Fellowship. D.B.K. would like to thank the BBSRC, EPSRC, NERC and RSC for financial support.

The authors have declared no conflicts of interest.

REFERENCES

Ankerst,M., Breunig,M., Kriegel,H.-P. and Sander,J. (1999) OPTICS: ordering points to identify clustering structure. In Delis,A. *et al*. (eds), *Proceedings of the 1999 International Conference on Management of Data*. ACM Press, New York, pp. 49–60.

Bandyopadhyay,S. and Manlik,U. (2001) Nonparametric genetic clustering: comparison of validity indices. *IEEE Trans. Syst. Man Cybernet.*, **31**, 120–125.

Ben-Dor,A., Friedman,M. and Yakhini,Z. (2002) Overabundance analysis and class discovery in gene expression data. *Technical report*, Agilent Laboratories, Palo Aeto.

Ben-Hur,A., Elisseeff,A. and Guyon,I. (2002) A stability based method for discovering structure in clustered data. In Aetman,R.B. *et al*. (eds), *Pacific Symposium on Biocomputing*. World Scientific Publishing Co., New Jersey.

Bezdek,J. and Pal,N. (1998) Some new indexes of cluster validity. *IEEE Trans. Syst. Man Cybernet.*, **28**, 301–315.

Bilu,Y. and Linial,M. (2002) The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classification. *J. Comput. Biol.*, **9**, 193–210.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Bolshakova,N. and Azuaje,F. (2003) Cluster validation techniques for genome expression data. *Signal Processing*, **83**, 825–833.

Bolshakova,N. *et al*. (2005) An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, **21**, 451–455.

Breckenridge,J. (1989) Replicating cluster analysis: method, consistency and validity. *Multivar. Behav. Res.*, **24**, 147–161.

Breckenridge,J. (2000) Validating cluster analysis: consistent replication and symmetry. *Multivar. Behav. Res.*, **35**, 261–285.

Datta,S. and Datta,S. (2003) Comparison and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.

Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.*, **1**, 224–227.

Ding,C. and He,C. (2004) K-nearest neighbor consistency in data clustering: incorporating local information into global optimization. In Haddad,H.M. *et al*. (eds), *Proceedings of the 2004 ACM Symposium on Applied Computing*. ACM Press, New York, pp. 584–589.

Dubes,R. and Jain,A.K. (1979) Validity studies in clustering methodologies. *Pattern Recog. Lett.*, **11**, 235–254.

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*, 2nd edn. John Wiley and Sons Ltd.

Dunn,J.C. (1974) Well separated clusters and fuzzy partitions. *J. Cybernet.*, **4**, 95–104.

Edwards,A.L. (1967) *The Correlation Coefficient*. W.H. Freeman, pp. 33–46.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall.

Eisen,M.B. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Ester,M., Kriegel,H.P. and Sander,J. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis,E. *et al*. (eds), *Proceedings of the Second International Conference on Knowledge Discovery and Data-Mining*. AIII Press, Menlo Park.

Estivill-Castro,V. (2002) Why so many clustering algorithms: a position paper. *ACM SIGKDD Explor. Newslett.*, **4**, 65–75.

Everitt,B.S. (1993) *Cluster Analysis*. Edward Arnold.

Fonseca,C.M. and Fleming,P.J. (1996) On the performance assessment and comparison of stochastic multiobjective optimizers.In Voigt,H.M. *et al*. (eds), *Proceedings of the Fourth International Conference on Parallel Problem Solving from Nature*. Springer-Verlag, Berlin, pp. 584–593.

Fridlyand,J. and Dudoit,S. (2001) Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. *Technical report*, Department of Statistics, Berkeley.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering. *Genome Biol.*, **3**, 1–22.

Gat-Viks,I. *et al*. (2003) Scoring clustering solutions by their biological relevance. *Bioinformatics*, **19**, 2381–2389.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression. *Science*, **286**, 531–537.

Goodacre,R. *et al*. (1998) Rapid identification of urinary tract infection bacteria using hyperspectral whole organism fingerprinting and artificial neural networks. *Microbiology*, **144**, 1157–1170.

Gordon,A.D. (1999) *Classification*. 2nd edn. Chapman and Hall.

Halkidi,M. *et al*. (2001) On clustering validation techniques. *J. Intell. Inform. Syst.*, **17**, 107–145.

Handl,J. and Knowles,J. (2005) Exploiting the trade-off—the benefits of multiple objectives in data clustering. In Coello,L.A. *et al*. (eds), *Proceedings of the Third International Conference on Evolutionary Multicriterion Optimization*. Springer-Verlag, Berlin, pp. 547–560.

Hastie,T. *et al*. (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, 1–21.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag.

Herrero,J. *et al*. (2001) A hierarchical unsupervised growing neural network for clustering gene expression data. *Bioinformatics*, **17**, 126–136.

Hubert,A. (1985) Comparing partitions. *J. Classif.*, **2**, 193–198.

Jaccard,S. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, **44**, 223–270.

Jain,A.K. *et al*. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

Jardine,N. and Sibson,R. (1971) *Mathematical Taxonomy*. John Wiley and Sons.

Kaplan,N. *et al*. (2004) A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, **5**.

Kell,D.B. and Oliver,S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, **26**, 99–105.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kohonen,T. (2001) Self-organizing maps. *Springer Series in Information Sciences*. Vol. 30, Springer-Verlag.

Krasnogor,N. and Pelta,D.A. (2004) Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, **20**, 1015–1021.

Krieger,A.M. and Green,P. (1999) A cautionary note on using internal crossvalidation. *Psychometrika*, **64**, 341–353.

Lange,T. *et al*. (2004) Stability-based validation of clustering solutions. *Neural comput.*, **16**, 1299–1323.

Lehmann,E.L. and D'Abrera,H.J.M. (1998) *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall.

Levine,E. and Domany,E. (2001) Resampling method for unsupervised estimation of cluster validity. *Neural Comput.*, **13**, 2573–2593.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, **2**, 1–11.

MacQueen,L. (1967) Some methods for classification and analysis of multivariate observations. In de Cam,L.M. *et al*. (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 281–297.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.

McLachlan,G. and Krishman,T. (1997) *The EM Algorithm and Extensions*. John Wiley and Son Ltd.

McShane,L.M. *et al*. (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, **18**, 1462–1469.

Mendes,D.J. *et al*. (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19**, 122–129.

Michaud,D.J. *et al*. (2003) eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics*, **19**, 1140–1146.

Milligan,G.W. and Cooper,M.C. (1986) A study of the comparability of external criteria for hierachical cluster ananlysis. *Multivar. Behav. Res.*, **21**, 441–458.

Pal,N.R. and Bezdek,J.C. (1995) On cluster validity for the fuzzy *c*-means model. *IEEE Trans. Fuzzy Syst.*, **3**, 370–379.

Pareto,V. (1971) *Manual of Political Economy, 1971 Translation of 1927 Edition*. Augustus M. Kelley.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Rand,W. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

Rayward-Smith,V.J., Osman,I.H., Reeves,C.R. and Smith,G.D. (1996) *Modern Heuristic Search Methods*. John Wiley and Sons Ltd.

Romesburg,H.C. (1984) *Cluster Analysis for Researchers*. Belmont.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Shaw,A.D. *et al*. (1997) Discrimination of the variety and region of origin of extra virgin olive oils using C-13 NMR and multivariate calibration with variable reduction. *Anal. Chim. Acta*, **384**, 357–374.

Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, **32**, 502–508.

De Smet,F. *et al*. (2002) Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, **18**, 735–746.

Tamayo,P. *et al*. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Tibshirani,R., Walther,G., Botstein,D. and Brown,P. (2001a) Cluster validation by prediction strength. *Technical report*, Department of Statistics, Stanford University, CA.

Tibshirani,R. *et al*. (2001b) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B*, **63**, 411–423.

Toronen,P. (2004) Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics*, **5**, 34.

van Rijsbergen,C. (1979) *Information Retrieval*, 2nd edn. Butterworths.

Vorhees,E. (1985) *The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval*. PhD thesis, Department of Computer Science, Cornell University.

Yeung,K.Y. *et al*. (2001a) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Yeung,K.Y. *et al*. (2001b) Model-based clustering and data transformation for gene expression data. *Bioinformatics*, **17**, 977–987.