Internal Evaluation Measures as Proxies for External Indices in Clustering Gene Expression Data

Milan Vukićević, Boris Delibašić, Miloš Jovanović, Milija Suknović
Center for Business Decision Making
Faculty of organizational sciences University of Belgrade
Belgrade, Serbia
vukicevicm@fon.bg.ac.rs

Zoran Obradović
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, PA, USA
zoran@temple.edu

Abstract*—Several external indices that use information not present in the dataset were shown to be useful for evaluation of representative based clustering algorithms. However, such supervised measures are not directly useful for construction of better clustering algorithms when class labels are not provided. We propose a method for identifying internal cluster evaluation measures that use only information present in the dataset and are related to given external indices. We utilize these internal measures for the construction of representative based clustering algorithms. Both identification and utilization steps of the proposed method are enabled by use of a component-based clustering algorithm design. Experiments on 432 algorithms using gene expression data sets provide evidence that some internal measures could be used as surrogates for external indices proposed in the literature. Moreover, the obtained results suggest that internal measures correlated to selected external indices can guide the algorithms toward significantly better cluster models.

***Keywords*-component-based clustering, evaluation measures, microarray data**

I. INTRODUCTION

Cluster analysis on DNA microarray data is essential for identifying biologically relevant groups of genes. Many clustering algorithms are often successfully used in clustering of microarray data as single algorithms (e.g. [1]), as improved algorithms (e.g. [2]) or as a part of consensus schemas (e.g. [3]). However, evaluation of cluster models is a difficult task because of the unsupervised nature of clustering [4].

Internal evaluation measures are usually used for measuring cluster quality after clustering is done. On the other hand, there are algorithms that optimize internal measures during algorithm execution (e.g. K-means optimizes intra-cluster distance), however their quality is measured with another internal measure (e.g. using Silhouette index). As an alternative, it was suggested to use an internal measure both for quality evaluation and as the objective function for clustering [5]. It was observed that this may be possible when the objective function captures what is desirable in a particular application and there is a feasible algorithm for finding optimal clustering. Such algorithms can be designed, but the problem is that users usually do not know precisely what the desired clustering property (internal evaluation) for their application area is.

On the other hand, external indices for cluster evaluation can provide objective information on how good the cluster model is according to real classes in data (so called "ground truth" or "golden standard"). The problem is that external indices cannot be optimized during the execution of a clustering algorithm because of the unsupervised nature of the clustering process (true classes are not known in advance). As a step towards the solution for this problem, we propose a method to detect internal measures that are highly correlated to selected external indices, and we suggest a way to utilize these internal measures in representative based clustering algorithms. Both detection and utilization steps are achieved by using a component-based clustering algorithm design [6] together with recently recommended external measure for evaluation of clustering algorithms [7].

We propose solving this problem and generalizing conclusions by using component based design we produced 432 clustering algorithms that utilize various components for algorithm flexibilities. Analyzing such a large variety of algorithms allowed us to study correlations of internal and external quality measures for a family of representative based algorithms rather than just a few specific algorithms (e.g. K-means).

II. RELATED WORK

One of the most important issues in cluster analysis is the evaluation of clustering results [8]. In recent times there have been many efforts to identify adequate methodologies for using internal and external evaluation measures (e.g. [9]). Still this is almost always done separately for internal and external evaluation measures. Many proposed algorithms use internal cluster validity measures as explicit objective functions that the clustering algorithms attempt to optimize (e.g. [10]) and showed very promising results. However, selecting the right objective functions (internal evaluation measures) is a difficult task because of the unsupervised nature of clustering. As a potential solution to this problem the K-star algorithm is used to investigate the correlation between several internal evaluation measures and F-measure and concluded that some of them could be used to improve the performance of text clustering algorithms [11].

Recently, it is suggested by [4] that internal evaluation measures should be evaluated with external indices. They used K-means to generate different partitions and compared the best

partitions by internal measures to the best partitions by external measures (real partitions).

Our proposal follows to some extent the work [4][11]. We explore correlation of several internal measures with recently recommended external measures for evaluation of clustering algorithms [7]. Additionally, our aim is to generalize the results over representative based clustering algorithms using component-based algorithm design.

III. COMPONENT-BASED ALGORITHM DESIGN

In this paper we use the framework for component based design of representative based clustering algorithms [6] to construct a large class of representative-based clustering algorithms. Clustering algorithms used in our study are designed by varying RCs from the following four sub-problems: initializing representatives, measuring distance, updating representatives and evaluating clusters.

Note that here step (d) ("evaluating clusters") is used during the execution of the algorithms (to be explained in Section 3). These algorithms represent the space of algorithms that are formed by reusing parts of various representative-based (k-means like) algorithms from the literature.

In particular, for initialization of cluster representatives we considered six RCs: RANDOM; (XMEANS); (GMEANS); (PCA); (KMEANS++) more detailed description of the RCs is available upon request.

The following four distances (similarities) were considered in this study as RCs: Euclidean distance (EUCLIDEAN); City block distance (CITY); Correlation similarity (CORREL); and Cosine similarity (COSINE).

We used the following three RCs for representatives update: MEAN, MEDIAN and ONLINE.

By combining RCs, original algorithms can be reproduced, but new algorithms can also be created. E.g. the K-means algorithms can be reconstructed as RANDOM-EUCLIDEAN-MEAN-COMPACT. An example of a new algorithm would be PCA-COSINE-ONLINE-SILHOU, that uses PCA to "initialize representatives", COSINE to "measure distance", ONLINE to "update representatives, and SILHOU to "evaluate clusters".

We integrated internal cluster evaluation measures' RCs into algorithms and in that way influenced the retrieved cluster models. These internal measures will be explained in more detail in the following section. By combining RCs from the four aforementioned sub-problems we designed 432 (6*4*3*6) algorithms, which form a space of representative-based cluster algorithms

IV. INTERNAL EVALUATION MEASURES AND EXTERNAL INDICES

Internal measures were developed to evaluate cluster models. However, representative based algorithms usually do not optimize these measures directly. In this research, we integrated internal evaluation measures into the cluster creation process using RC based design. In this way, algorithms were guided by a selected evaluation measure towards a desirable clustering solution. As representative based algorithms are usually optimized in several iterations, and rerun several times to avoid the local optima trap, in our approach after each iteration and at each restart of the clustering process the best model is kept according to a selected internal evaluation measure value. In addition, internal evaluation is also embedded in initialization RCs.

A good review of internal evaluation measures and their correlation can be found in [12]. For the experiments in this paper we used Compactness (COMPACT), XB-index (XB), Connectivity (CONN), Global silhouette index (SILHOU), AIC and BIC

External indices for evaluation of clustering results have recently received increasing interest [7]. It has been shown recently that besides metric properties, external indices should also be normalized and adjusted for chance. This is necessary because unadjusted measures give better results as the number of clusters increases (even when the number of constructed clusters is larger than true number of clusters). After exhaustive comparison between different information, the theoretic measure of adjusted mutual information (AMI) is recommended as a general purpose measure for clustering validation and algorithm comparison. This is why we are using this measure in following experiments.

V. EXPERIMENTAL RESULTS

We used four synthetic and six real world gene expression data sets that are drawn from [3] and are used in microarray data analysis, where biological samples are clustered based on gene expression data. Datasets used in this experiment are suggested as representative [3], and are often used as a benchmark in microarray data analyses (367 citations as of 13.06.2011). They are used because they represent real clusters (identified by experts or generated artificially), they are from the same area of research and are used at [7] where AMI index is proposed as a general purpose clustering evaluation measure. Table 1 shows a summary of datasets for synthetic and real world problems, respectively.

In our experiments, the number of clusters (K) is assumed to be already known and corresponding to the numbers shown at Table 1. On the *lung cancer* dataset, however, experts are not yet confident on the true number of clusters but they agree that there are at least four clusters in this dataset. Experiments were conducted using 432 representative algorithms designed with a component based approach (described in Section 2). All algorithms were run until cluster membership didn't change K (number of clusters) times but not more than for 20 iterations.

TABLE I. DATASETS SUMMARY, SOURCE: [6]

	Clusters	**Samples**	**Attributes**
Synthetic datasets			
Gaussian 3	3	60	600
Gaussian4	4	400	2
Simulated 5	5	500	2
Simulated 6	6	60	600
Real world datasets			
Leukemia	3	38	999
Novartis	4	103	1000
Lung cancer	4+	197	1000
CNS Tumors	5	48	1000
St. Jude	6	248	985
Normal	13	90	1277

A. *Correlation of internal measures to AMI*

In our experiments, the relation between external and internal measures is analyzed based on statistical correlation. Correlation was computed over 432 component-based clustering algorithms described in Section 2. Using such a large space of clustering algorithms instead of a single fixed clustering algorithm (e.g. K-means) greatly improves the generality of the conclusions, since the correlation is measured over a wider class of representative-based algorithms.

To select an appropriate internal evaluation measure (surrogate) for an external index, we investigated the statistical correlation between internal and external indices. The analysis of the correlation between the external index AMI and six internal measures is shown at Table 2.

TABLE II. CORRELATION BETWEEN AMI AND INTERNAL MEASURES OVER REAL WORLD DATASETS

	Compactness	**XB**	**Silhouette**	**Connectivity**	**AIC/BIC**
Artificial datasets					
Gaussian3	0.258	0.254	0.771	0.535	**0.965**
Gaussian4	0.276	0.512	0.391	0.539	**0.988**
Gaussian5	-0.565	0.152	-0.388	-0.548	**0.926**
Simulated6	0.263	0.421	0.667	0.4	**0.9**
Real world datasets					
Leukemia	0.254	0.599	0.711	0.884	**0.916**
Novartis	0.099	0.73	0.722	0.732	**0.911**
Lung cancer	-0.320	-0.160	0.333	**0.806**	0.627
Normal	0.569	0.116	0.632	0.557	**0.736**
St. Jude	0.179	0.579	0.705	0.553	**0.949**
CNS Tumors	0.366	0.333	0.205	0.248	**0.414**

Table 2 shows that correlation between AMI and internal measures was consistent on all datasets. For all internal measures except *Silhouette,* lower values are better. For easier interpretation of the results in Table 2 we changed the sign in values that should be minimized. So, in Table 2 negative correlation means that AMI value decreases as the internal measure improves, which makes such an internal measure misleading. We see that considerable correlation exists between *AIC/BIC* and AMI throughout the datasets. It is also evident that other internal measures showed less correlation to AMI. One exception was the *lung cancer* dataset, where the *connectivity* measure looks more related to AMI.

However, after closer examination, we saw that for most algorithms on the lung cancer dataset, the AMI index was below 0.5 (also lowering the average, as seen on Fig. 1), which indicates poor compliance to the predefined "gold standard" desired clustering. This means that the algorithms we explored did not manage to reproduce the desired clustering. As mentioned before, this could be due to the uncertain true number of clusters. After selecting algorithms with AMI above 0.5 on *Lung cancer*, the correlation to the *AIC* internal measure (0.82) was greater than to the *connectivity* measure (0.64).

B. *Model selection based on internal evaluation measures*

Experiments conducted in Section 5.1 suggested that *AIC* (along with *BIC*) is the best suited internal measure for use as a replacement for AMI in absence of true clusters for gene expression data. As *AIC* and *BIC* behave similarly, we further analyzed the *AIC* more thoroughly and set the following hypothesis:

Hypothesis H1: "The AIC internal cluster evaluation measure could be used to guide cluster algorithms, and produce cluster models having significantly better AMI external evaluation index."

In this paper, 432 algorithms used internal cluster evaluation measures during algorithm execution. We divided these algorithms into six groups, each consisting of 72 algorithms, using a specific evaluation measure (model selection) RC, but varying other RC aspects as explained in Section 3. To see whether an internal evaluation measure used during algorithm run can produce significant differences in algorithm performance, we evaluated average AMI values on these groups. Since algorithm performance is strongly dependent on dataset properties, we evaluate algorithms on real world datasets.

It can be seen from Fig. 1 that on most real world datasets, the best or near best average AMI results are obtained in algorithm groups where AIC or BIC were used as cluster evaluation measure (model selection) RCs. On datasets where these two RCs did not give the best results, AMI values were very similar for all algorithms.

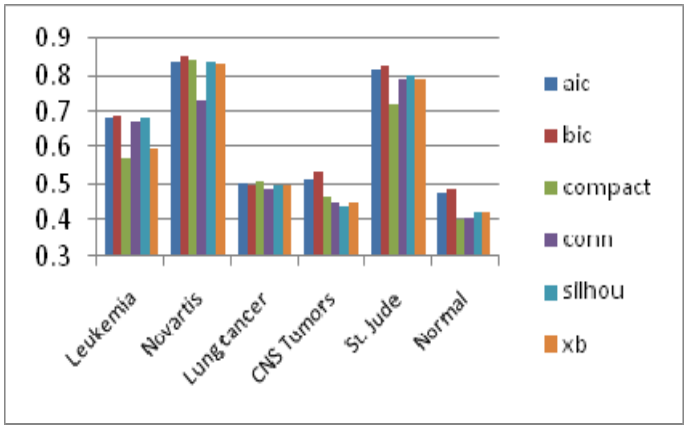

Figure 1. Average AMI values of algorithms with different model selection RCs over real world datasets

We used the Wilcoxon paired rank test to see if differences shown on Fig. 1 were significant and not produced by chance, or other sources of variation. Algorithms are compared in pairs where RCs for all sub-problems were the same, and only the "Evaluate clusters" RC was changed (e.g. RANDOM-MEAN-AIC was compared with RANDOM-MEAN-COMPACT). This way the influence of "Evaluate cluster" RCs on algorithm performance is isolated. Table 3 shows that in a number of cases algorithms that were using AIC produced significantly better cluster models as compared to other cluster evaluation RCs (here, * denote significance level 0.05, **0.01, and ***0.001).

TABLE III. CORRELATION BETWEEN AMI AND INTERNAL MEASURES OVER REAL WORLD DATASETS

	SILHOU	CONN	XB	COMPACT
Leukemia	0.791	0.699	0.006**	0.001**
Novartis	0.645	0.01**	0.871	0.529
Lung cancer	0.732	0.219	0.585	0.459
Normal	0.000***	0.000***	0.003**	0.001**
St. Jude	0.168	0.050*	0.040*	0.000***
CNS Tumors	0.000***	0.000***	0.002**	0.000***

From previous results it can be seen that AIC and BIC RCs are consistently giving best or nearly best AMI values on all datasets. Still, using other internal measures as model selectors in some cases yields good results (Fig. 1), even if there is no strong correlation to AMI. However, in real world applications (without prior knowledge of real clusters) lack of correlation can produce misleading decisions. The user can not be sure whether a cluster model produced using an internal measure RC, badly correlated to AMI, will produce a real cluster structure (good external index). Strongly correlated internal measures, as AIC, give us confidence of achieving good AMI values, because they behaved consistently throughout the analyzed datasets.

VI. CONCLUSION AND FUTURE RESEARCH

The goal of this study was to identify internal evaluation measures that could be integrated and used during execution of a clustering algorithm to guide the cluster creation process in order to achieve good clustering performance, according to certain external evaluation measures.

The obtained results provide evidence that AIC (or BIC), used as a model selection component, can guide the algorithms toward significantly better cluster models (than other internal evaluation measures), measured by recently suggested AMI index.

This finding should motivate a more explicit optimization of internal measures that could be incorporated in algorithms to further improve produced cluster models. This could be accomplished using various meta-heuristic approaches that optimize any given optimization criteria.

With respect to the aforementioned problem, we plan to make deeper analysis and to make the conclusions based on data distributions rather than the application area (gene expression analysis).

ACKNOWLEDGMENT

This project is funded in part under a grant to Z. Obradovic with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

REFERENCES

[1] K. Dhiraj and S.K. Rath. Gene Expression Analysis Using Clustering. *Proc of 3rd International Conference on Bioinformatics and Biomedical Engineering*, 2009, pp. 154-163.

[2] F. Geraci, M. Leoncini, M. Montangero, M. Pellegrini and M. Renda. K-Boost: a scalable algorithm for high-quality clustering of microarray gene expression data. *Journal of computational biology a journal of computational molecular cell biology* 16, 2009, pp. 859-873.

[3] S. Monti S, 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* 52(1-2), 2003, pp.91-118.

[4] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, JM. Pérez and JI Martín. Towards a standard methodology to evaluate internal cluster validity indices.*Pattern Recognition Letters* 32(3), 2011, pp.505-515.

[5] B. Dom. An information theoretic external cluster validity measure. IBM Research Report RJ 10219. IBM's Almaden Research Center, San Jose, CA, 2001.

[6] B. Delibašic, K. Kirchner, J.Ruhland, M. Jovanović, M. Vukićević. Reusable components for partitioning clustering algorithms.*Artificial Intelligence Review* 32, 2009, pp.59-75.

[7] N. X. Vinh, J. Epps, J. Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res* 11, 2010, pp. 2837-2854.

[8] M. Halkidi, Y. Batistakis and M. Vazirgiannis. On clustering validation techniques, *Journal of Intelligent Information Systems* 17, 2001, pp.107-145.

[9] R. Campello. Generalized external indices for comparing data partitions with overlapping clusters, *Pattern Recognition Letters* 32, 2011, pp.966-975.

[10] Zhao Y, Karypis G. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning* 55(3), 2004, pp.311-331.

[11] D. Ingaramo, D. Pinto, P. Rosso and M. Errecalde. Evaluation of internal validity measures in short-text corpora. *Computational Linguistics and Intelligent Text Processing*. 2008, pp.555–567

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

[13] Q.H. Nguyen, and V.J. Rayward-Smith. Internal quality measures for clustering in metric spaces. *Int. J. Business Intelligence and Data Mining* 3 (1), 2008, pp. 4–29, 2008.