An Information Theoretic Divergence for Microarray Data Clustering

Nguyen Xuan Vinh, *student member, IEEE*
and Nguyen Minh Phuong

***Abstract*— The Kullback-Leibler (KL) divergence used in conjunction with the unsupervised Self-Organizing Map (SOM) algorithm has been previously shown to be effective for the gene clustering problem: the patterns of the gene clusters obtained were found to be superior to those obtained by the hierarchical clustering algorithm using the uncentered Pearson correlation measure. Motivated by this initial finding, in this research we study the effectiveness of the KL-divergence in a more general setting where the data points are not necessarily projected to the unit simplex but to a parallel simplex in the positive orthant. Two novel hard and soft clustering algorithms based on the so-called generalized KL-divergence are proposed. We tested the algorithms on both gene and sample clustering problems. Experimental results on real microarray datasets with known class labels (for genes or samples) show that the generalized KL-divergence based algorithms produce comparable or better results to those obtained by similar algorithms based on popular distance measures for microarray data clustering, such as the squared Euclidean distance and the Pearson correlation. Two validation indices, namely the Adjusted Rand Index and the newly developed Variation of Information metric, have been used to validate the results.**

I. INTRODUCTION

Clustering is one of the most important data mining techniques used to extract useful information from microarray data. The aim of clustering is to group together similar objects (genes or tumor samples) thus allowing biologists to identify potential relationship between objects. For a clustering method to work well the choice of a suitable similarity (or dissimilarity) measure plays a very important role. The definition of similarity is domain specific and for the gene clustering problem, the similarity is often taken to be the similarity in *shape* between the two gene profiles. Genes with similar profile patterns often have similar functions, participate in a particular pathway or respond to a common environmental stimulus and thus should be grouped together [4]. For microarray data analysis, popular choices of distance metric are the squared Euclidean distance, the cosine similarity or correlation coefficients such as the Pearson correlation. The latter two measures have been found to work quite well on microarray data ([9], [21]). The squared Euclidean distance on the other hand often fails to spot genes with very similar patterns but with large differences in magnitude. The situation is demonstrated in fig. 1(a) where genes 2 and 3 are closer than are genes 1 and 2, although genes 1 and 2 have more similar pattern profiles. One of the normalization procedures often applied to microarray data is to make each gene profile have unit norm and zero mean. With this data normalization scheme, several distance and similarity measures, such as the squared Euclidean distance, the cosine similarity and the Pearson correlation coefficient essentially coincide and depend only on the dot product of the normalized gene profiles. We shall refer to the unified measure as the normalized squared Euclidean distance. The effect of data normalization is illustrated in fig. 1(b) where similar gene patterns are placed closer together.

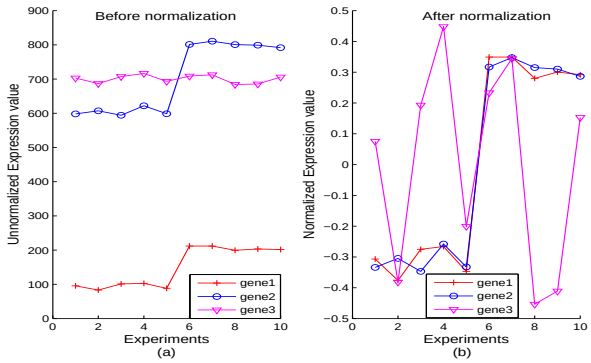

Fig. 1. Unnormalized and normalized gene expression profiles: after normalization, genes with more similar patterns are placed closer together

The relative entropy or Kullback-Leibler (KL) divergence has been previously shown to be effective for gene clustering. Kasturi *et al.* [13] have shown that when gene profiles are projected to the "probability space", i.e., the unit simplex, the KL-divergence is able to effectively detect the pattern dissimilarity between gene profiles. KL-divergence when used in conjunction with the Self-Organizing Map (SOM) algorithm produce better result compared to the hierarchical clustering algorithm with the uncentered Pearson correlation measure (the cosine similarity). This initial finding reported in [13] is encouraging, however in our opinion there are still several concerns. First, the paper compared the SOM, a partitional clustering algorithm, with the hierarchical clustering algorithm, which is not optimized to generate partitional clusters. Meanwhile, all the clustering validation indices used therein require partitional clustering as input. In the case of using the same SOM algorithm with the KL-divergence and the uncentered Pearson correlation, the relative performance of the two measures was not clear. Second, the internal validation method used, the Davies-Bouldin index, made use of the Euclidean distance. As shown in [18] internal validation methods also need to employ a certain distance measure and thus may favor that distance measure over others. Hence a different comparison method need to be performed to ensure fairness. Finally the KL-divergence was

The authors are with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. Corresponding author: n.x.vinh@unsw.edu.au

applied in a rather ad hoc manner without much theoretical insight given. Since the work in [13], to our knowledge, no more work has been done to further evaluate the KL-divergence for gene clustering.

In this research we study the effectiveness of a more general type of divergence that we call the generalized Kullback-Leibler (gKL) divergence when compared to the more popular normalized Squared Euclidean (nSE) distance. Since both the gKL-divergence and the nSE-distance are instances of a large class of divergences namely the Bregman divergences, we used the hard and soft Bregman clustering algorithms [3] as the framework for comparison. In addition to the gene clustering problem, the potential usefulness of the gKL-divergence for the sample clustering problem is also studied for the first time.

II. Bregman divergence and Bregman clustering

A. Bregman divergences

Definition: Let ϕ: $\mathcal{S} \mapsto \mathbb{R}$ be a strictly convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ such that ϕ is differentiable on the relative interior of $\mathcal{S}$ (denoted by ri($\mathcal{S}$)), assumed to be nonempty. The Bregman divergence d_ϕ: $\mathcal{S} \times$ ri($\mathcal{S}$) $\mapsto [0, \infty)$ is defined as: $d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$, where $\nabla\phi(y)$ represents the gradient vector of ϕ evaluated at y.

A wide variety of distortion functions such as the Euclidean distance, Mahalanobis distance, Itakura-Saito distance and KL-divergence (relative entropy) fall into the class of Bregman divergences. These distortion functions have been used widely for clustering in various fields [3].

B. Normalized Squared Euclidean distance

Consider the following strictly convex function $\phi_E = \|x\|^2$ on $\mathbb{R}^d$. The corresponding Bregman divergence of ϕ_E is the Squared Euclidean distance: $d_{\phi_E}(x, y) = \|x - y\|^2$. If the data is mapped to the manifold $S_R = \{x : \|x\| = R, R > 0\}$ by the scaling operation $x \leftarrow Rx/\|x\|$ then $d_{\phi_E}(x, y) = 2R^2 - 2\langle x, y\rangle$, which we shall refer to as the normalized Squared Euclidean (nSE) distance. A special case is when $R = 1$ all the data points will have unit norm (and so equal variance) which is a common normalization step for microarray data analysis. It is noted that in practice, gene profiles are also often shifted to have zero mean prior to norm-normalization ([6], [7], [14], [16]). In our experiments the mean-normalization step is also performed for the nSE-distance based algorithms.

C. Generalized KL-divergence

Consider the following strictly convex function $\phi_{KL} = \sum_{j=1}^{d} x_j \log x_j$ on $\mathbb{R}_+^d$. The corresponding Bregman divergence of ϕ_{KL} is the Generalized I-divergence:

$$d_{\phi_{KL}}(x, y) = \sum_{j=1}^{d} x_j \log(\frac{x_j}{y_j}) - \sum_{j=1}^{d}(x_j - y_j) \qquad (1)$$

If the data is mapped to the manifold $S_\alpha = \{x : \sum_{i=1}^{d} x_i = \alpha, \alpha > 0\}$ then $d_{\phi_{KL}}(x, y) = \sum_{j=1}^{d} x_j \log(x_j/y_j)$. We shall call it the generalized KL (gKL) divergence. A special case is when $\alpha = 1$ we get the well known KL-divergence.

Since microarray data naturally contain only positive numeric values, either in the form of absolute gene expression level (as with one color microarray) or in the form of ratio (as with two color microarray), the gKL-divergence will be naturally applicable. The data normalization process is just a scaling operation $x \leftarrow \alpha x/\sum_{j=1}^{d} x_j$ so that each gene (or sample) profile will not change direction but just magnitude and will lie on the same manifold S_α.

D. Bregman hard clustering

Let $\{x_i, i = 1 \ldots n\}$ be the set of gene profiles. We need to find the k cluster centers $\mu = (\mu_1; \mu_2; \ldots; \mu_k)$ so that the following total Bregman divergence objective function is minimized:

$$f_\phi(\mu) = f_\phi(\mu_1; \mu_2; \ldots; \mu_k) = \sum_{i=1}^{n} \min_{l=1..k} d_\phi(x_i, \mu_l) \qquad (2)$$

The following heuristic K-means type Bregman hard clustering algorithm [3] can be used to solve the problem. The algorithm loops through the following two steps until no further improvement can be made, i.e., no data point can change membership:

1) *The membership assignment step:* each data point is assigned to the nearest cluster. Let X_h be the set of indices of points that belong to the cluster with the corresponding center μ_h then $X_h = \{i : d_\phi(x_i, \mu_h) = \min_{l=1\ldots k} d_\phi(x_i, \mu_l)\}$
2) *The center adjustment step:* the cluster centers are relocated to the gravity center of its new members:

$$\mu_l = \frac{\sum_{i \in X_l} x_i}{|X_l|}, l = 1 \ldots k \qquad (3)$$

Since in this research all the data points are normalized to the same manifold, either S_R or S_α, it is natural to require that all the cluster centers also lie on the same manifold as the data. It can be seen that for the hard gKL-clustering algorithm, this requirement is automatically satisfied, i.e. all the centers defined by (3) will lie on S_α. For the hard nSE-clustering, the additional constraints $\{\|\mu_h\| = R, h = 1 \ldots k\}$ need to be added to the optimization problem (2) and the center adjustment step in (3) will be consequently changed to $\mu_l = R(\sum_{i \in X_l} x_i)/\|\sum_{i \in X_l} x_i\|, l = 1 \ldots k$. This variant of the algorithm has already been known in the literature under the name of "Spherical K-means" (since the data points lie on a sphere) and has been demonstrated to be efficient for clustering directional data such as text or microarray data [8]. In this research we take the Spherical K-means as the counterpart of the hard gKL-clustering algorithms.

E. Bregman soft clustering

It has been shown that there is a bijection between regular Bregman divergences and regular Exponential families [3]. Thus for each Bregman divergence d_ϕ there exists an exponential probability distribution of the form $p(x) =$

$\exp(-d_\phi(x,\mu))b_\phi(x)$ where μ is the location parameter and $b_\phi(x)$ is a function independent of μ. The Bregman soft clustering problem is defined as that of learning the maximum likelihood parameter $\Theta = \{\mu_h, \pi_h\}_{h=1}^k$ of a mixture model of the form:

$$p(x|\Theta) = \sum_{h=1}^{k} \pi_h \exp(-d_\phi(x,\mu_h))b_\phi(x) \quad (4)$$

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{n} \log \sum_{h=1}^{k} \pi_h \exp(-d_\phi(x_i,\mu_h))b_\phi(x_i) \quad (5)$$

Finding the global solution for the above optimization problem is hard, thus the iterative EM algorithm is often applied to find a local optimizer. The iterations for EM are:

$$E-step: \quad p(h|x_i,\Theta^{(l)}) = \frac{\pi_h^{(l)} \exp(-d_\phi(x_i,\mu_h^{(l)}))}{\sum_{h'=1}^{k} \pi_{h'}^{(l)} \exp(-d_\phi(x_i,\mu_{h'}^{(l)}))} \quad (6)$$

$$M-step: \quad \pi_h^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} p(h|x_i,\Theta^{(l)}) \quad (7)$$

$$\mu_h^{(l+1)} = (\sum_{i=1}^{n} p(h|x_i,\Theta^{(l)})x_i) / \sum_{i=1}^{n} p(h|x_i,\Theta^{(l)}) \quad (8)$$

Again for the soft gKL-clustering and soft nSE-clustering, it is preferable to require that all the cluster centers lie on the same manifold as the data points. For the soft gKL-clustering, the requirement is automatically satisfied, i.e. all the centers defined by (8) will lie on S_α. For the soft nSE-clustering problem, the introduction of the additional constraints $\{\|\mu_h\| = R, h = 1 \dots k\}$ to the optimization problem (5) will result in an additional norm-normalization step after (8): $\mu_h^{(l+1)} \leftarrow R\mu_h^{(l+1)}/\|\mu_h^{(l+1)}\|$ as shown in [17]. This slight variant of the soft nSE-clustering algorithms has been shown to be effective for the gene clustering problem [17] and will be taken as the counterpart for the soft gKL-clustering algorithms.

F. Choosing the normalization constant α and R

Suppose that we have normalized our data to the manifold S_α and S_R. Now consider the effect of mapping the data to some other manifolds $S_{\beta\alpha}$ and $S_{\beta R}$ with $\beta > 0$. We shall see that the hard gKL-clustering and nSE-clustering algorithms are not affected by the scaling coefficient β. This is because the K-means type hard Bregman clustering algorithm used a piecewise linear decision function based on the distance matrix $D = [d_\phi(x_i,\mu_h)]_{i=1\dots n}^{h=1\dots k}$. With the introduction of β the new distance matrices are just linear scaled versions of the old distance matrices, i.e. βD and thus the decision taken at each step will be unchanged. The situation is however totally different for the EM-type soft clustering algorithm where the soft decision at each step is taken based on the soft membership matrix $P = [p(h|x_i,\Theta^{(l)})]_{i=1\dots n}^{h=1\dots k}$. The introduction of β will change the soft membership matrix nonlinearly and thus will affect the decision at each step.

Two extreme values of β can be noted: when $\beta \to \infty$ the posterior probability in the E-step takes values in $\{0, 1\}$ and hence the soft EM algorithm becomes the hard clustering algorithm. Another extreme is when $\beta \to 0$. In this case all the posterior probabilities $p(h|x_i)$ will tend to $1/k$, all the mixture proportions π_h will tend to $1/k$ and the algorithm will converge to the degenerate solution where all the cluster centers coincide and equal to $\mu_o = \sum_{i=1}^{n} x_i/n$. Neither of the two extremes is of interest when using the soft clustering algorithms. A suitable value in between which retains a suitable level of fuzziness is of interest. However, parameter tuning is more an art than a science. Similar to the problem of choosing the learning rate for the SOM algorithm, the fuzzy parameter for the fuzzy C-means algorithm, the temperature parameter for the Simulated Annealing algorithm and the cross over and mutation probability for the Genetic algorithms, the best knowledge can only be in the form of general recommendation. In some rare cases, recommendation can be in the form of explicit formulae as in [6] but no theoretical foundation of such choice could be given. Optimal values for the normalization constant R and α for the soft nSE-clustering and gKL-clustering algorithms are also data dependent. We have empirically found that choosing α in the range $[50, 300]$ and R in the range $[4, 20]$ would give reasonably good clusterings for typical gene clustering problems with a few hundred genes.

III. Comparison method

Microarray data clustering has attracted much research effort. This is reflected by the large number of new algorithms that have been developed specifically for microarray data. Also a number of existing clustering algorithms have been applied into this new area. With a variety of currently available algorithms, there is a great need for clustering algorithm evaluation methods. In this section we first summarize some of the clustering validation indices. These indices measure the goodness of a clustering and serve as the basis for comparing clustering algorithms. We then choose a suitable comparison method to assess the performance of the gKL-divergence and nSE-distance based algorithms.

A. Comparing clusterings when external knowledge is available

The external knowledge here is the number of classes in the data and the class label for each data point (gene or sample). For the sample clustering problem external knowledge might be quite reliable (as the number of classes and the number of samples are often small, thus can be efficiently investigated by human experts). For the gene clustering problem the situation is quite different. Since the number of classes might be large and the number of data points might be huge, it is hard to obtain the correct class label for every gene. As a result external knowledge for the gene clustering problem (especially for large datasets) should not be considered as the "absolute gold standard" for comparing clustering algorithms but rather a guideline. Also when comparing clustering algorithms, a decent number of data sets is required for the comparison to be statistically meaningful. However in our observation, available (and reliable) data of this kind via publications in the area are not so abundant.

We summarize here two validity indices that can be employed to assess clusterings goodness when external knowledge is available, one is the popular Adjusted Rand Index and another is the recently developed Variation of Information. Let C be the true clustering with k clusters $C_1, C_2, \ldots, C_k$ and n_i be the number of points in cluster C_i, $\sum_{i=1}^{k} n_i = n$. Suppose C' is a clustering result given by some clustering algorithm with k' clusters $C'_1, C'_2, \ldots, C'_{k'}$ and n'_i is the number of points in cluster C'_i, $\sum_{i=1}^{k'} n'_i = n$.

1) Adjusted Rand Index (ARI): [2], [11] Let a, b, c and d respectively denote the number of gene pairs belonging to the same cluster in both C' and C, the number of pairs belonging to the same cluster in C' but to different clusters in C, the number of pairs belonging to different clusters in C' but to the same cluster in C, and the number of pairs belonging to different clusters in both C' and C. The Adjusted Rand Index assessing the concordance between the two clusterings is defined as follows:

$$ARI(C, C') = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \quad (9)$$

2) Variation of Information (VI): [15] This is an information theoretic criteria for comparing clusterings. The entropy associated with the clustering C (and similarly for C') is: $H(C) = -\sum_{i=1}^{k} P(i) \log P(i)$, where $P(i) = n_i/n$. The mutual information between two clusterings C and C' is:

$$I(C, C') = \sum_{i=1}^{k} \sum_{i'=1}^{k'} P(i, i') \log \frac{P(i, i')}{P(i)P'(i')} \quad (10)$$

where $P(i, i') = |C_i \cup C'_{i'}|/n$ represents the probability that a point belongs to C_i in the clustering C and to $C'_{i'}$ in C'. The Variation of Information between two clusterings C and C' is then defined by:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \quad (11)$$

The VI is a metric on the space of clusterings. It measures the "distance" between the two clusterings C and C'. Both the ARI and VI have certain advantages and disadvantages as discussed in [15]. In this research we use both indices with the VI chosen as the primary index.

B. Comparing clusterings when external knowledge is unavailable

In the case where external knowledge is unavailable, some internal validity indices such as the Silhouette index, the Davies-Boudlin index, the Dunn index or the FOM (Figure Of Merit) index can be employed to compare clusterings. Those indices however, require the definition of a distance measure (normally the Euclidean distance is employed) and thus may be inappropriate for comparing clustering algorithms which are based on different distance measures. For example, the FOM has been previously shown to be biased toward the Euclidean distance [18]. For this reason in this research we used only data sets with external knowledge to assess the performance of the gKL-divergence and nSE-distance based algorithms.

C. Summary of comparison method

In order to ensure a fair evaluation of the two measures, a comparison method has been carried out as follows:

- Distance measures do not stand alone but are always coupled with a certain clustering algorithm prototype to form a complete algorithm. In this research, a common algorithm prototype based on the hard and soft Bregman clustering algorithms has been chosen as the framework for comparison.
- To minimize the effect of parameter tuning (α and R), each soft algorithm is tested with multiple parameter values (normally in the range [5,20] for R and [50,300] for α) for a certain data set. For each parameter value, an algorithm is run 100 times with random initialization and the averaged values for VI and ARI are recorded. The algorithm instance that gives highest averaged VI value is then reported.
- For the hard clustering algorithms where no parameter fine tuning is needed, the algorithms are simply run 100 times with random initialization and the averaged values for VI and ARI are recorded.
- Data sets with external knowledge are used, i.e., the number of clusters and the class label of each gene or sample.

IV. Generalized KL-divergence for the gene clustering problem

The experiment was performed on 3 real microarray data sets for which external knowledge is available.

A. Data sets

1) Yeast cell cycle data: The yeast cell cycle data studied by Cho *et al.* [5] showed the fluctuation of expression level of more than 6000 genes over two cell cycles (17 time points). Following [25] we use two different subsets of this data with independent external criteria (known class label for each gene):

- *Set 1 - 384 genes:* This set consists of 384 genes whose expression level peak at different time points corresponding to the five phases of cell cycle.

- *Set 2 - 237 genes:* This set consists of 237 genes corresponding to four categories in the MIPS database. The four categories (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism and ribosomal proteins) were shown to be reflected in clusters from the yeast cell cycle data. These four functional categories form the four classes in the external criterion for this data set.

2) Yeast galactose data: A subset of 205 genes from the yeast galactose data set of Ideker *et al*. [12] has been used by Yeung *et al*. to assess the performance of various clustering algorithms [24]. The expression patterns reflect four functional categories in the Gene Ontology (GO) listings. We use the same subset in this study.

B. Results

To assess the ability of the gKL-divergence to group genes with similar profile patterns we first visually inspect the

clustering results. Fig. 2, 3 and 4 are typical clustering results for the above three data sets using the soft and hard gKL-clustering algorithms starting with random initialization. It can be observed that the clusters created are quite coherent and of distinct temporal patterns, confirming that the gKL-divergence is able to effectively detect the dissimilarity in *shape* of gene profiles. To qualitatively assess the algorithms,

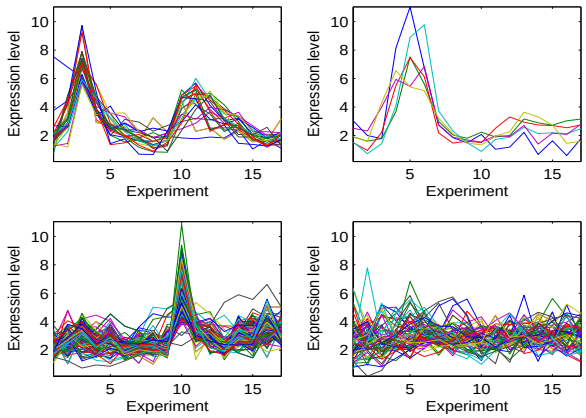

Fig. 2. Cho's yeast data set 1: clusters created with the gKL-divergence soft clustering algorithm with $k = 4$ and $\alpha = 50$.

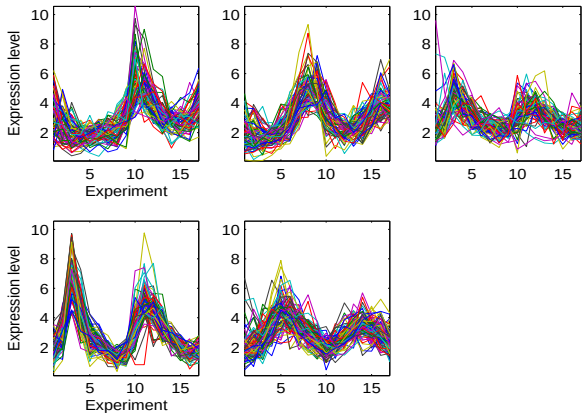

Fig. 3. Cho's yeast data set 2: clusters created with the gKL-divergence hard clustering algorithm with $k = 5$.

experiment method as described in section III-C has been carried out. Results for the 3 data sets are summarized in Table I with the best values of VI(the lowest) and ARI(the highest) in bold. The VI and ARI are quite concordant with each other, albeit not perfectly. It can be observed that all four algorithms produced clusterings with quite comparable average quality while the soft gKL-clustering algorithm seems to create clusterings with slightly better quality. We thus conclude that the gKL-divergence is efficient for the gene clustering problem. The soft and hard gKL-divergence based clustering algorithms thus can be added to the current repertoire of gene clustering algorithms.

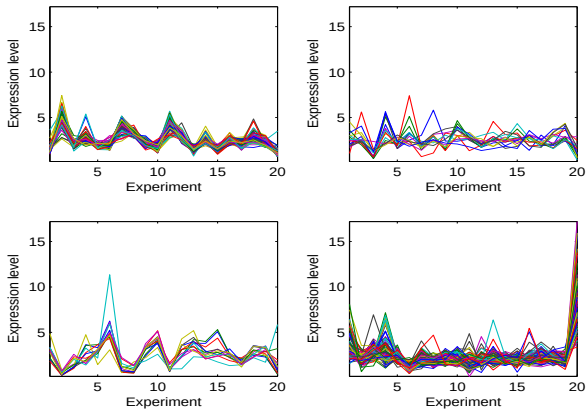

Fig. 4. Galactose yeast data set: clusters created with the gKL-divergence hard clustering algorithm with $k = 4$.

TABLE I

GENE CLUSTERING RESULTS

Data set	**Algorithm**	**VI**	**ARI**
Cho's yeast data set 1	Spherical K-means	1.49	**0.51**
	nSE soft clustering, $R = 5$	1.47	**0.51**
	gKL hard clustering	1.54	0.40
	gKL soft clustering, $\alpha = 50$	**1.38**	0.49
Cho's yeast data set 2	Spherical K-means	1.52	**0.43**
	nSE soft clustering, $R = 5$	1.52	**0.43**
	gKL hard clustering	1.55	0.41
	gKL soft clustering, $\alpha = 60$	**1.49**	**0.43**
Yeast galactose	Spherical K-means	0.38	0.86
	nSE soft clustering, $R = 8$	0.36	0.86
	gKL hard clustering	0.42	0.80
	gKL soft clustering, $\alpha = 80$	**0.34**	**0.87**

V. GENERALIZED KL-DIVERGENCE FOR THE SAMPLE CLUSTERING PROBLEM

A. *Connection with the multinomial distribution*

In [3], Banerjee *et al.* have shown that there is a bijection between regular exponential families and Bregman divergences. The corresponding exponential family for the generalized KL-divergence is the multinomial distribution. The Multinomial mixture model has been successfully applied in text clustering [19] where documents are represented using a bag-of-word model: each document is represented as a high-dimensional vector which merely stores the counts of each word in the document. The documents will be clustered together based on the relatively higher frequencies of certain keywords distinct for each group. In the microarray sample clustering problem we notice a similar criterion for grouping samples: each microarray experiment (sample) can be considered similar to a document and the mRNA corresponding to each gene can be considered as a word. The number of mRNA will be counted(measured) and reported as the gene expression level. The samples will be then grouped based on the relative expression level of the genes just like documents are grouped based on the relative frequency of the words. Therefore we can expect a reasonably good performance of the hard and soft gKL-divergence based algorithms when applied to the sample clustering problem.

B. Results

We tested the algorithms on four real microarray data sets with known sample class labels:

Leukemia data: we use the training data set published by Golub *et al.* [10]. The data set consists of 38 bone marrow samples with 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML) samples. Each sample contains the expression level of 6817 human genes. We used all these 6817 genes in our experiments. Since the original data set contains some negative expression values (due to the normalization/background substraction process), we shifted the whole data set so that the smallest expression value is 1 before performing data normalization for the gKL-divergence based algorithms.

Colon data: this data set contains expression level of more than 6500 genes with samples of 40 tumors and 22 normal colon tissues. We used the publicly available subset of 2000 genes with highest minimal intensity across samples [1].

NCI60 data: published by Ross *et al*. (2000) [20], the data set contains gene expression for nearly 8000 genes in 60 human cell lines obtained from various tumor sites: 7 breast, 5 Central Nervous System (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 NSCLC, 6 ovarian, 2 prostate, 9 renal and 1 unknown. We excluded the 2 prostate and 1 unknown samples to form 8 classes with sufficient size. Genes with missing values were filtered out, resulting in a reduced data set of 57 samples $\times$ 6189 gene expression profiles.

Pediatric acute leukemia data: the original data set consisted of 327 samples of pediatric acute lymphoblastic leukemia [23]. We use a subset of 248 samples formed by 6 well defined subclasses namely BCR-ABL (15 samples), E2A-PBX1 (27 samples), Hyperdiploid > 50 chromosomes (64 samples), MLL (20 sampels), T-ALL (43 samples) and TEL-AML1 (79 samples). A simple variation filter as described in [22] was used to filter out genes that do not show a relative variation of 5 and an absolute variation of 500. The reduced data set contains 248 samples $\times$ 1347 genes.

TABLE II
SAMPLE CLUSTERING RESULTS

Data set	Algorithm	VI	ARI
Leukemia data	Spherical K-means	0.84	0.16
	nSE soft clustering, $R = 15$	0.79	0.21
	gKL hard clustering	0.80	0.24
	gKL soft clustering, $\alpha = 100000$	**0.74**	**0.34**
Colon data	Spherical K-means	1.04	0.21
	nSE soft clustering, $R = 7$	**0.82**	**0.40**
	gKL hard clustering	1.02	0.20
	gKL soft clustering, $\alpha = 80$	0.92	0.36
NCI60 data	Spherical K-means	2.05	0.24
	nSE soft clustering, $R = 8$	1.96	**0.26**
	gKL hard clustering	1.94	0.21
	gKL soft clustering, $\alpha = 100$	**1.81**	0.21
Pediatric leukemia	Spherical K-means	2.41	0.17
	nSE soft clustering, $R = 8$	2.40	0.16
	gKL hard clustering	1.45	0.48
	gKL soft clustering, $\alpha = 300$	**1.22**	**0.57**

Experimental results for the 4 data sets are summarized in Table II. It can be observed that for the first 2 data sets where the number of classes is small, all the algorithms produce clusterings with quite comparable quality on average. For the Leukemia data set, all the algorithms can produce clusterings as good as without any or only 1 misclassified sample. This result is quite interesting as we did not use any unsupervised feature selection procedure but worked directly with the original data set. As the soft gKL clustering algorithm did not produce an acceptable result with $\mu \in [50, 300]$, we tried some larger value for μ until the algorithm produced quite good result at $\mu = 10000$. On the Colon data set, all the algorithms can produce similarly effective clusterings with a minimum of 5 misclassified samples. We noticed that those 5 samples had also been previously often misclassified by clustering/classification algorithms as reported in [1], [16], [17].

TABLE III
PEDIATRIC LEUKEMIA DATA SET: GKL-DIVERGENCE HARD CLUSTERING MATCHING MATRIX, VI=1.09

Class \ Cluster		1(55)	2(39)	3(47)	4(17)	5(52)	6(38)
BCR-ABL	(15)	1	0	2	**11**	0	1
E2A-PBX1	(27)	0	0	6	0	0	**21**
Hyperdiploid	(64)	**54**	0	7	2	0	1
MLL	(20)	0	0	3	2	0	15
T-ALL	(43)	0	**39**	3	1	0	0
TEL-AML1	(79)	0	0	**26**	1	**52**	0

TABLE IV
PEDIATRIC LEUKEMIA DATA SET: NSE-DISTANCE HARD CLUSTERING MATCHING MATRIX, VI=2.09

Class \ Cluster		1(30)	2(7)	3(83)	4(54)	5(14)	6(60)
BCR-ABL	(15)	0	1	8	2	0	4
E2A-PBX1	(27)	4	0	0	8	0	15
Hyperdiploid	(64)	5	2	**47**	10	0	0
MLL	(20)	2	2	0	2	2	12
T-ALL	(43)	0	1	0	2	**11**	**29**
TEL-AML1	(79)	**19**	1	28	**30**	1	0

For the last 2 data sets where the number of classes is larger, it can be observed that the gKL-divergence based algorithms gradually provide better clustering quality over the nSE-distance based algorithms. The difference is quite noticeable in the Pediatric leukemia data set. Closer inspection of the result reveals that for the NCI60 data set, most of the clusterings created by the four algorithms are not of high quality. This is due to the inherent difficulty in clustering this data set which contains a rather limited number of samples (57) but with a relatively large number of classes (8). The Pediatric Leukemia data set on the other hand contains a relatively large number of samples (248) with a relatively smaller number of classes (6) and therefore we should expect a better performance of all the clustering algorithms on this data set. The result however shows that the gKL-divergence based algorithms produce on average relatively better clusterings compared to the ones obtained by the nSE-distance based algorithms. Tables III and IV show the Matching matrices (or contingency tables) for the best clusterings created by the hard gKL-divergence and nSE-distance clustering algorithms respectively. It can be observed that the clusters created by the hard gKL-divergence

clustering algorithm are considerably more coherent than the ones created by the nSE-distance clustering algorithms. The experimental result suggests that the gKL-divergence and the underlying mixture of multinomial distributions might be a good choice for the sample clustering problem as discussed earlier in this section. Again it is interesting to note here that we did not use any sophisticated unsupervised feature selection scheme but only the quite basic variation filter. Nevertheless the results obtained with the gKL-divergence based algorithms are very promising.

VI. Discussion and conclusion

In this work we have assessed the usefulness of the generalized KL-divergence for microarray data clustering including both the genes and samples clustering problems. Two novel hard and soft gKL-divergence based clustering algorithms have been introduced. For the gene clustering problem, experimental results showed that gKL-divergence based algorithms produce clusterings of very comparable quality in terms of the Variation of Information metric and the Adjusted Rand Index to those obtained by similar algorithms based on the more popular normalized Squared Euclidean distance. For the sample clustering problem, gKL-divergence based algorithms produce better clustering results on data sets with large number of classes. The two validation indices used in this research, namely the Adjusted Rand Index and Variation of Information metric, are quite concordant with each other albeit not perfectly. It is left for future work to study the suitable value for the normalization constant α as well as the effect of unsupervised feature selection on the gKL-divergence based algorithms. Also, the study of gKL-divergence in conjunction with the other popular algorithms for microarray data clustering is an interesting research direction.

VII. Acknowledgement

The authors are grateful to Dr. Julien Epps, School of Electrical Engineering and Telecommunications, the University of New South Wales, for his valuable comments. This work is supported in part by the Australian Asia Endeavour Award and the University of New South Wales International Support Scholarship.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc Natl Acad Sci U S A*, vol. 96, no. 12, pp. 6745–50, 1999.

[2] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2859–2865, 2007.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[4] D. P. Berrar, W. Dubitzky, and M. Granzow, *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers Boston, 2003.

[5] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[6] D. Dembele and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 3, pp. 973–80, 2003.

[7] I. S. Dhillon, E. M. Marcotte, and U. Roshan, "Diametrical clustering for identifying anti-correlated gene clusters," *Bioinformatics*, vol. 19, no. 13, pp. 1612–9, 2003.

[8] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143 – 175, January-February 2001.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci USA*, vol. 95, no. 25, pp. 14 863–8, 1998.

[10] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–7, 1999.

[11] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, pp. 193–218, 1985.

[12] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood, "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network," *Science*, vol. 292, no. 5518, pp. 929–934, 2001.

[13] J. Kasturi, R. Acharya, and M. Ramanathan, "An information theoretic approach for analyzing temporal patterns of gene expression," *Bioinformatics*, vol. 19, no. 4, pp. 449–458, 2003.

[14] S. Y. Kim, J. W. Lee, and J. S. Bae, "Effect of data normalization on fuzzy clustering of dna microarray data," *BMC Bioinformatics*, vol. 7, no. 134, p. 134, 2006.

[15] M. Meilă, "Comparing clusterings: an axiomatic view," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*. New York, NY, USA: ACM, 2005, pp. 577–584.

[16] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[17] P. Nguyen and T. Hoang, "Clustering in a fixed manifold to detect groups of genes with similar expression patterns," in *2nd International Conference on Bioinformatics Research and Development*, Vienna, Austria, 2008.

[18] A. L. Olex, D. J. John, E. M. Hiltbold, and J. S. Fetrow, "Additional limitations of the clustering validation method figure of merit," in *ACM-SE 45: Proceedings of the 45th annual southeast regional conference*. New York, NY, USA: ACM, 2007, pp. 238–243.

[19] L. Rigouste, O. Cappé, and F. Yvon, "Inference and evaluation of the multinomial mixture model for text clustering," *Inf. Process. Manage.*, vol. 43, no. 5, pp. 1260–1280, 2007.

[20] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat Genet*, vol. 24, no. 3, pp. 227–35, 2000.

[21] W. Shannon, R. Culverhouse, and J. Duncan, "Analyzing microarray data using cluster analysis," *Pharmacogenomics*, vol. 4, no. 1, pp. 41–52, 2003.

[22] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc Natl Acad Sci USA*, vol. 96, no. 6, pp. 2907–12, 1999.

[23] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–43, 2002.

[24] K. Yeung, M. Medvedovic, and R. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biology*, vol. 4, no. 5, p. R34, 2003.

[25] K. Y. Yeung, "Cluster analysis of gene expression data," Ph.D. dissertation, University of Washington, Seattle, WA, 2001.