Extending Data Reliability Measure to a Filter Approach for Soft Subspace Clustering

Boongoen, Tossapon; Shang, Changjing; Iam-On, Natthakan; Shen, Qiang

Published in:
IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)

DOI:
10.1109/TSMCB.2011.2160341

Publication date:
2011

Citation for published version (APA):
Boongoen, T., Shang, C., Iam-On, N., & Shen, Q. (2011). Extending Data Reliability Measure to a Filter Approach for Soft Subspace Clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *41*(6), 1705-1714. https://doi.org/10.1109/TSMCB.2011.2160341

Extending Data Reliability Measure to a Filter Approach for Soft Subspace Clustering

Tossapon Boongoen, Changjing Shang, Natthakan Iam-On and Qiang Shen

***Abstract*—The measure of data reliability has recently proven useful for a number of data analysis tasks. This paper extends the underlying metric to a new problem of soft subspace clustering. The concept of subspace clustering has been increasingly recognized as an effective alternative to conventional algorithms (which search for clusters without differentiating the significance of different data attributes). While a large number of crisp subspace approaches have been proposed, only a handful of soft counterparts are developed with the common goal of acquiring the optimal cluster-specific dimension weights. Most soft subspace clustering methods work based on the exploitation of *k*-means and greatly rely on the iteratively disclosed cluster centres for the determination of local weights. Unlike such wrapper techniques, this paper presents a filter approach which is efficient and generally applicable to different types of clustering. Systematical experimental evaluations have been carried out over a collection of published gene expression datasets. The results demonstrate that the reliability-based methods generally enhance their corresponding baseline models and outperform several well-known subspace clustering algorithms.**

***Index Terms*—Data reliability, soft subspace clustering, wrapper and filter, attribute weight, gene expression analysis.**

I. Introduction

The concept of data reliability was initially introduced for the task of information aggregation, with the preliminary measure being formulated using the proximity to a 'local cluster' [4]. Despite its inefficiency, the underlying measure has proven effective for classification and feature selection problems. Recently, an enhanced variation has been proposed in [5], where a hierarchical clustering process required by the original model is replaced by a search of nearest neighbors. The resulting metric has been successfully used to establish a data fusion method for detecting possible terrorists' alias names [6]. In addition, an unsupervised feature selection technique was also built on top of the data reliability measure, with the performance being superior to alike algorithms found in the literature. Inspired by such achievement, this research extends the application of data reliability measure to the problem of 'soft subspace clustering', which has attracted a great deal of interests amongst data analysts and researchers in the past decade (e.g. [12], [13], [15], [18], [25]).

The practice of subspace clustering or bi-clustering has recently emerged in response to the challenges of high-dimensional data, especially in gene expression analysis [8], [20], [24], [34], [41], [42]. With the revolution of microarray technology, gene expression data obtained from microarray experiments has inspired several novel applications, including the identification of differentially expressed genes for further molecular studies [35], [43], and the creation of classification systems for improved cancer diagnosis [9], [38]. Another typical application is to reveal natural structures and identify interesting patterns in expression data [24], [37]. In particular, traditional algorithms such as *k*-means and agglomerative hierarchical clusterings have proven useful for identifying biologically relevant clusters of tissue samples and genes. The present research focuses on the work where samples with similar profiles of gene expression values are grouped together [11].

Generally, cluster detection is based on a distance/proximity measure between objects of interest. However, with high-dimensional data, meaningful clusters cannot be easily identified as distances are increasingly indifferent as dimensionality increases [3], [23]. To disclose patterns obscured by irrelevant dimensions, a global feature selection/reduction method, e.g. principle components analysis (PCA) [26], is effective only to some extent. Particularly, it fails to detect in each dimension, locally varying relevance for distinct object groups. In order to overcome such limitations, many different subspace clustering algorithms have been proposed with the common objective of discovering locally relevant dimensions per cluster (see [29] for a survey). In Fig.1, for example, which represents different clusters of n objects $(x_1, x_2, \ldots, x_n)$ in d dimensions $(f_1, f_2, \ldots, f_d)$, Cluster 1 corresponds to a traditional cluster in a full data space, whilst the other clusters associate with specific dimensional subsets.

T. Boongoen, C. Shang, N. Iam-On and Q. Shen are with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK e-mail:{tsb,cns,nii07,qqs}@aber.ac.uk

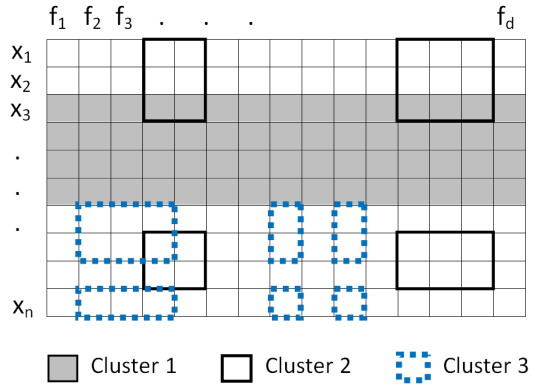

Fig. 1. Illustration of clusters in different subspaces defined by subsets of the original data attributes.

Regarding the techniques introduced for the determination of cluster-specific subspaces, subspace clustering approaches can be characterized in two categories: crisp and soft. The former finds an exact subspace for each cluster (see [1], [7], [29] for examples). The latter, a soft subspace clustering method detects clusters in a full data space. For each cluster, different dimensions are assigned with dissimilar weights in accordance with their relevance in identifying the underlying cluster. In practice, an optimal subspace can be obtained using either wrapper or filter approach [14]. The former wraps the search around a specific clustering algorithm (e.g. *k*-means), whilst the latter selects the feature subspaces, prior to the actual unsupervised learning process.

Existing soft clustering techniques (e.g. [13], [15], [18], [25]) rely on a specific clustering method, typically k-means to search for the optimal set of weights. Unfortunately, this implementation of a wrapper nature cannot be extended beyond the underlying basic clustering mechanism. Such algorithms repeatedly update dimension weights from intermediate cluster centers (or centroids) which are iteratively modified such that the overall intra-cluster variance is minimized. In so doing, the accuracy of cluster-specific weights may not be retained and the quality of discovered centroids is usually arbitrary.

In order to overcome these shortcomings, this paper presents a 'filter' approach to soft subspace clustering. It makes use of the data reliability measure [5] to construct the object-dimension association matrix, which can be employed to guide the weight configuration within a clustering process. Note that such a method is generally applicable to a wide range of clustering algorithms: k-means, spectral [30] and hierarchical clusterings [22], for instance. As for the *k*-means alike techniques where the object-dimension information remains unchanged overtime, the underlying weight modification is partially dependent of disclosed centroids, thereby improving the likelihood of weight accuracy being maintained. This intuitive implementation has shown to be effective on a number of published gene expression data and is robust to parameter perturbation.

The rest of this paper is organized as follows. Section II introduces the concepts of soft subspace clustering upon which the current research is established. In Section III, the filter method for soft subspace clustering and the theoretical ideas underlying this approach are presented. The performance evaluation of the proposed algorithms against other comparable techniques, over several gene expression datasets, is reported and discussed in Section IV. The paper is concluded in Section V, with a discussion of future work.

II. Soft Subspace Clustering: Concepts and Existing Techniques

The idea of soft subspace clustering was originally observed in the study of [15], where the 'Clustering Objects on Subsets of Attributes (COSA)' algorithm was introduced to determine a weight to every dimension in each cluster. Fig.2 illustrates this concept, where the weighted dimensional space allows a cluster to be visualized and identified more easily. Specific to 'Cluster1' in this example, the weights w_x^1 and w_y^1 of the dimensions f_x and f_y, respectively, are equal in the original setting, shown in Fig. 2(a). With respect to Fig. 2(b) in which these weights are adjusted such that w_x^1 remains unchanged and $w_y^1 = \frac{1}{3}w_x^1$, the cluster becomes more structurally rigid and clearly identifiable. Similarly, 'Cluster2' is separable from the former in Fig. 2(b), where $w_x^2 = \frac{1}{4}w_y^2$. Despite its promising performance, this method has been heavily criticized for its inefficiency [29].

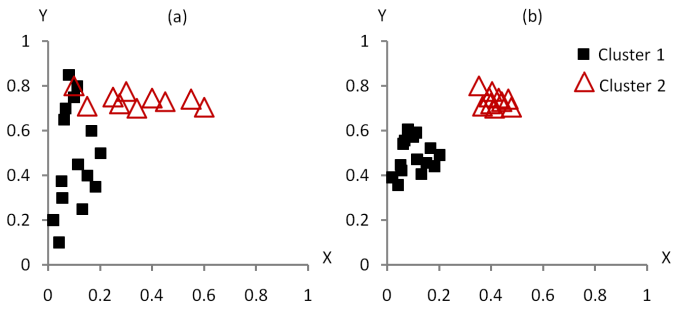

Fig. 2. Illustration of clusters in (a) original data-attribute space and (b) weighted space, where clusters are clearly separable.

Following this initial approach, a few well-known extensions of k-means have been proposed for a less

expensive soft subspace clustering [13], [18], [25]. With these wrapper methods, cluster-specific dimension weights are repeatedly updated, along the iterative minimization of intra-cluster variances in k-means clustering. Let $X = \{x_1, \ldots, x_n\}$ be a set of objects and each object $x_i = (x_{i1}, \ldots, x_{id}), i = 1 \ldots n$ is a vector of values characterized by a set of features or attributes $F = \{f_1, \ldots, f_d\}$. The k-means searches for the partition $C = \{C_1, \ldots, C_k\}$ of X into k clusters, minimizeing the following objective function:

$$J_0(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{d} u_{il} \, (x_{ij} - z_{lj})^2 \tag{1}$$

where $\sum_{l=1}^{k} u_{il} = 1$, $U \in \mathbb{R}^{n \times k}$ is a matrix in which each entry u_{il} represents a membership degree that object i has with regard to cluster l ($u_{il} \in \{0, 1\}$ and $u_{il} \in [0, 1]$ for crisp and soft clustering, respectively). In addition, $Z = \{z_1, \ldots, z_k\}$ denotes a vector representing the centroids of k clusters, i.e. $z_l = (z_{l1}, \ldots, z_{ld}), l = 1 \ldots k$.

Specifically to the wrapper method of [25], called Entropy Weight k-Means (EWKM), the objective function $J_1(U, Z, W)$ modified from that of the classical k-means, is defined as

$$\sum_{l=1}^{k} \Big[\sum_{i=1}^{n} \sum_{j=1}^{d} u_{il} \, w_{lj} (x_{ij} - z_{lj})^2 + \gamma \sum_{j=1}^{d} w_{lj} \log w_{lj} \Big] \tag{2}$$

Here, $\gamma \in \mathbb{R}$ denotes a constant that controls the incentive of weight changes, $W \in \mathbb{R}^{k \times d}$ is a matrix in which each entry w_{lj} represents a weight of dimension j in cluster l, $w_{lj} \in [0, 1], \forall l = 1 \ldots k, j = 1 \ldots d$ and $\sum_{j=1}^{d} w_{lj} = 1$. In each iteration of the k-means alike process, W is updated by

$$w_{lj} = \frac{\exp\left(\frac{-D_{lj}}{\gamma}\right)}{\sum_{t=1}^{d} \exp\left(\frac{-D_{lt}}{\gamma}\right)} \tag{3}$$

and

$$D_{lj} = \sum_{i=1}^{n} u_{il} (x_{ij} - z_{lj})^2 \tag{4}$$

In addition to this technique, a similar wrapper model has been introduced in [13], namely the Locally Adaptive Clustering (LAC) algorithm. The corresponding objective function $J_2(U, Z, W)$ is specified as follows, where $|C_l|$ is the cardinality of the cluster $C_l, l = 1 \ldots k$:

$$\sum_{l=1}^{k} \sum_{j=1}^{d} [w_{lj} O_{lj} + h \, w_{lj} \log w_{lj}] \tag{5}$$

$h \geq 0$ which is the constant that controls the relative differences between dimension weights, and O_{lj} is calculated by

$$O_{lj} = \frac{1}{|C_l|} \sum_{\forall x_i \in C_l} (x_{ij} - z_{lj})^2 \tag{6}$$

As with EWKM, LAC works also by repeatedly updating W using the following:

$$w_{lj} = \frac{\exp\left(\frac{-O_{lj}}{h}\right)}{\sum_{t=1}^{d} \exp\left(\frac{-O_{lt}}{h}\right)} \tag{7}$$

Finally, another wrapper technique called Fuzzy Subspace Clustering (FSC) has also been proposed in [18] with the following objective function, $J_3(U, Z, W)$:

$$\sum_{l=1}^{\beta} \sum_{j=1}^{d} w_{lj}^{\delta} \left[\sum_{i=1}^{n} u_{il} (x_{ij} - z_{lj})^2 + \varepsilon \right] \tag{8}$$

where $\delta \in (1, \infty)$ denotes a weight component (or fuzzy index) and ε is a very small positive number. FSC iteratively updates W by:

$$w_{lj} = \frac{1}{\sum_{t=1}^{d} \left[\frac{\sum_{i=1}^{n} u_{il} (x_{ij} - z_{lj})^2 + \varepsilon}{\sum_{i=1}^{n} u_{il} (x_{it} - z_{lt})^2 + \varepsilon} \right]^{1/(\delta - 1)}} \tag{9}$$

With these wrapper methods, cluster-specific dimension weights are repeatedly updated, along the iterative minimization of intra-cluster variances in k-means clustering. The modification process is based typically on the distances between object members to the disclosed cluster centers, which can be sub-optimal. Hence, the accuracy of weights cannot always be maintained. To address this important shortcoming, a new filter approach is introduced in the next section to extend k-means, amongst other basic clustering techniques, such that the resulting soft subspace clustering algorithm becomes less dependent of recovered centroids, enabling the update of cluster-specific weights to be more accurate.

III. A Novel Filter Approach to Soft Subspace Clustering

The existing methods [13], [25] for soft subspace clustering are based on the wrapper framework, which inherently limits their applications only to a single type of clustering algorithm – k-means, for instance.

To overcome this limitation, a new filter approach is introduced here, with its extensions to several different basic clustering techniques. The new method exploits the data reliability measure of [5] to preliminarily construct an object-dimension association matrix that represents locally relevance degree of each dimension for every data object.

Let $\alpha \in \{1 \dots (n-1)\}$ be the number of nearest neighbors of any object under examination. The object-dimension association matrix $AS^{\alpha} \in \mathbb{R}^{n \times d}$ is a collection of informative entries $AS^{\alpha}_{ij} \in [0,1]$ representing the strengths that an object $x_i \in X$ is similar to (or associated with) a set $N^{\alpha}_{ij} \subset X$ of its α nearest-neighboring objects in a given dimension $f_j \in F$. Formally, the underlying measure can be defined as

$$AS^{\alpha}_{ij} = 1 - \left(\frac{D^{\alpha}_{ij}}{D^{\alpha}_{*}} \right) \tag{10}$$

where

$$D^{\alpha}_{*} = \max_{\forall i,j} D^{\alpha}_{ij} \tag{11}$$

with D^{α}_{ij} being

$$D^{\alpha}_{ij} = \frac{1}{\alpha} \sum_{\forall q \in N^{\alpha}_{ij}} \sqrt{(x_{ij} - q_j)^2} \tag{12}$$

Note that the estimation of data reliability relies on the search for α nearest neighbors of any object in question. Particularly, the following 'NN' algorithm is employed to find $N^{\alpha}_{ij}, \forall i = 1 \dots n, j = 1 \dots d$. The 'SORT' function exploited here can be any efficient algorithm in the literature, e.g. the 'pancake' sort [10] whose time complexity is $O(z)$ where z is the number of values to be sorted. The resulting NN is computationally less expensive than the previous algorithm presented in [5], with the complexity being reduced from $O(n^2)$ to approximately $O(nd)$.

ALGORITHM: **NN**(X, α)

N^{α}_{ij}, a set of α nearest neighbors of $x_i \in X$ in dimension $f_j \in F$;
$dist(x_{ij}, x_{i'j}) = \sqrt{(x_{ij} - x_{i'j})^2}$, a distance between x_{ij} and $x_{i'j}$;

(1) **For each** $f_j \in F$
(2) $T = \{(x_{1j}, 1), \dots, (x_{nj}, n)\}$
(3) $ST = SORT(T | x_{ij}, i = 1 \dots n)$
(4) **For each** $st_q \in ST, st_q = (x_{ij}, i)$
(5) $ST_{ij} = \{st_{q-\alpha}, \dots, st_{q-1}, st_{q+1}, \dots, st_{q+\alpha}\}$
(6) $D_{ij} = \emptyset$
(7) **For each** $\sigma \in ST_{ij}, \sigma = (x_{i'j}, i')$
(8) $D_{ij} = D_{ij} \cup (dist(x_{ij}, x_{i'j}), x_{i'j}, i')$
(9) $N_{ij} = SORT(D_{ij} | dist(x_{ij}, x_{i'j}), \forall i')$
(10) $N^{\alpha}_{ij} = \{g_1, \dots, g_{\alpha}\}, g_{\beta} \in N_{ij}$

The measure AS^{α}_{ij} has an intuitive interpretation towards the problem of subspace clustering. When it approaches 1, the dimension f_j is highly relevant to the local cluster which object x_i is an element in. If however, the underlying measure is close to 0, the dimension is irrelevant to the clustering of x_i. Conceptually, the resulting AS^{α} matrix can be used to configure the dimensional weighting scheme of disclosed clusters. To illustrate the effectiveness and generality of this measure, it is applied to several basic clustering algorithms, each of which is discussed below.

Reliability-based k-means (R-KM) extends the conventional k-means algorithm such that the association values in AS^{α} are automatically employed in formulating object clusters. Its objective function is defined as

$$J_R(U, Z, W) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{d} u_{il}\, w_{lj} (x_{ij} - z_{lj})^2 \tag{13}$$

For the initial stage where centroids $Z = \{z_1, \dots, z_k\}$ correspond to a set of randomly selected objects, the weight w_{lj} of the l-th cluster is estimated by

$$w_{lj} = \frac{AS^{\alpha}_{ij}}{\sum_{t=1}^{d} AS^{\alpha}_{it}} \ , j = 1 \dots d \tag{14}$$

given that $x_i = z_l$. In the following iterations, the dimension weight w_{lj} of each cluster C_l is updated by

$$w_{lj} = \frac{MA^{\alpha}_{lj}}{\sum_{t=1}^{d} MA^{\alpha}_{lt}} \ , j = 1 \dots d \tag{15}$$

where MA^{α}_{lj} is the association measure to the j-th dimension which is minimally shared by all members in C_l:

$$MA^{\alpha}_{lj} = \min_{\forall x_i \in C_l} AS^{\alpha}_{ij} \tag{16}$$

With Z and W being fixed, the cluster membership $u_{il} \in U, i = 1 \dots n, l = 1 \dots k$ can be specified such that

$$u_{il} = \begin{cases} 1 & if \ \ l = \arg\min\limits_{s=1 \dots k} \sum\limits_{j=1}^{d} w_{sj} (x_{ij} - z_{sj})^2 \\ 0 & otherwise \end{cases} \tag{17}$$

Similar to the typical k-means method, the set of centroids Z is updated using the following equation:

$$z_{lj} = \frac{\sum_{i=1}^{n} u_{il}\, x_{ij}}{\sum_{i=1}^{n} u_{il}} \tag{18}$$

The R-KM algorithm that minimizes the objective function defined in Eq. 13 is summarized as follows:

ALGORITHM: **R-KM**(k, AS^{α})
(1) Randomly initialize Z
(2) Calculate initial weights using Eq. 14
(3) **Repeat**
(4) Update U by Eq. 17
(5) Update Z by Eq. 18
(6) Update W by Eq. 15
(7) **Until** the objective function obtains its local minimum

The R-KM algorithm converges to a local minimum of the objective function defined in Eq. 13. Formally, let Z^{π}, W^{π} and U^{π} respectively denote the centroids, weights, and cluster assignments derived in the π^{th} iteration. In U^{π}, each data object x_i is assigned to its closest cluster according to the weights and centroids in the previous iteration, i.e. $Z^{\pi-1}$ and $W^{\pi-1}$. From this, the following relation results:

$$J_R(U^{\pi}, Z^{\pi-1}, W^{\pi-1}) \leq J_R(U^{\pi-1}, Z^{\pi-1}, W^{\pi-1})$$

For the given U^{π}, the optimal Z^{π} and W^{π} are computed using Eqs. 18 and 15. Hence, the following also holds:

$$J_R(U^{\pi}, Z^{\pi}, W^{\pi}) \leq J_R(U^{\pi}, Z^{\pi-1}, W^{\pi-1})$$

Overall, it can be concluded that $J_R(U^{\pi},Z^{\pi-1},W^{\pi-1})$ is no greater than $J_R(U^{\pi-1},Z^{\pi-1},W^{\pi-1})$. It is guaranteed that R-KM reduces the objective value in iterations. The clustering problem is to group n objects into k disjoint sets and there are only a finite number of data partitions. For a given U, the minimal objective value is determined for the corresponding optimal centroids and weights. Therefore, the objective value for a given assignment is lower-bounded. The objective value in the R-KM algorithm decreases gradually until the value reaches a fixed point. This fixed point is a local minimal of $J_R(U, Z, W)$.

Reliability-based hierarchical clustering (R-CL) extends the well-known agglomerative hierarchical clustering technique [22]. It generates a tree (called 'dendogram') as nested groups of data organized hierarchically. The algorithm begins by considering each data sample as a cluster, and then gradually merges similar clusters until all the clusters are combined into one big group (i.e. the top node of the resulting dendogram). The hierarchical dendogram reveals cluster-subcluster relations, and the order in which the clusters were merged or split. Particularly to the 'complete linkage (CL)' approach [28] that is of interest in the present work, this is obtained by defining the distance $DS(C, C')$ between two clusters $C \subset X$ and $C' \subset X$ such that

$$DS(C, C') = \max_{\forall x_i \in C, x_{i'} \in C'} \sum_{j=1}^{d} (x_{ij} - x_{i'j})^2 \qquad (19)$$

Note that the CL technique requires an adjacency matrix $A \in \mathbb{R}^{n \times n}$ that represents pairwise-proximity measures amongst objects as an input. The original A is based on a uniform dimensional weight setting, which may be enhanced using the information of local relevance in AS^{α}. Effectively, each entry $A(x_i, x_{i'}) \in A$ which corresponds to the weighted distance between objects $x_i, x_{i'} \in X$, can be defined as

$$A(x_i, x_{i'}) = \sum_{j=1}^{d} w_{ii'j} (x_{ij} - x_{i'j})^2 \qquad (20)$$

where $w_{ii'j}$ is estimated by

$$w_{ii'j} = \frac{\min(AS_{ij}^{\alpha}, AS_{i'j}^{\alpha})}{\sum_{t=1}^{d} \min(AS_{it}^{\alpha}, AS_{i't}^{\alpha})}, j = 1 \ldots d \qquad (21)$$

Following these definitions, Eq. 19 is simplified as

$$DS(C, C') = \max_{\forall x_i \in C, x_{i'} \in C'} A(x_i, x_{i'}) \qquad (22)$$

The R-CL algorithm is summarized as follows:

ALGORITHM: **R-CL**(k, AS^{α})
(1) Initialize a set of clusters $\mathbb{C} = \{C_1, \ldots, C_n\}, C_i = \{x_i\}$, $i = 1 \ldots n$
(2) Create an adjacency matrix A using Eq. 20
(3) **Repeat**
(4) Find $C_p, C_q \in \mathbb{C}$; $DS(C_p, C_q) = \min_{\forall C, C' \in \mathbb{C}, C \neq C'} DS_{C,C'}$, using Eq. 22
(5) Combine C_p and C_q; $C_o = C_p \cup C_q$
(6) Update $\mathbb{C}$; $\mathbb{C} = ((\mathbb{C} - C_p) - C_q) \cup C_o$
(7) **Until** $|\mathbb{C}| = 1$

Reliability-based spectral clustering (R-SPT) extends the spectral clustering technique of [30], which operates on the pairwise similarity matrix $S \in \mathbb{R}^{n \times n}$, given that $S = 1 - A$. Similar to the conventional hierarchical clustering method, the original S is estimated from an unweighted dimensional space. For this purpose, a new similarity matrix is constructed from the adjacency matrix created by Eq 20. Then, the k largest eigenvectors of S, $v_1, \ldots, v_k$, are found (chosen to be orthogonal to each other in the case of repeated eigenvalues), forming the matrix $V = [v_1, \ldots, v_k]$ by stacking the eigenvectors in columns. Another matrix V^* is subsequently constructed

from V by normalizing each row of V to have a unit length. By considering each row of V^* as k-dimensional embedding of an object in X, k-means is used to divide objects (i.e. rows of V^*) into a partition of k clusters.

The R-SPT algorithm is summarized as follows:

ALGORITHM: **R-SPT**(k, AS^α)

(1) Create a similarity matrix $S \in \mathbb{R}^{n\times n}$, $S = 1 - A$, using Eq. 20

(2) Find k largest eigenvectors $v_1, \ldots, v_k$ of S

(3) Form a transformed data matrix $V \in \mathbb{R}^{n\times k}$, $V = [v_1, \ldots, v_k]$

(4) Create a normalized matrix $V^* \in \mathbb{R}^{n\times k}$:

$$V^*_{ij} = \frac{V_{ij}}{\sqrt{\sum_{\forall j} V^2_{ij}}}, i = 1 \ldots n, j = 1 \ldots k$$

(5) Apply k-means to V^*

IV. Performance Evaluation

In order to investigate the performance of the filter-based algorithms, experimental studies are set out in comparison with standard and existing soft subspace clustering methods, over real data. Examined datasets and experiment design are outlined below, followed by a discussion of experimental results, including parameter analysis.

A. Investigated Gene Expression Datasets

This evaluation is conducted on gene expression data obtained from six published microarray studies. Each of the investigated datasets is briefly described below, with its statistics summarized in Table I. Note that, to resolve the problems with missing and extreme values, the pre-processing procedure of [21] is applied to these datasets (but the appreciation of the present results does not require the understanding of this procedure). In addition to the expert-directed number of sample classes (k), a set of possible class numbers (C) is specified for each dataset and used to assess the robustness of a given clustering method.

TABLE I
DESCRIPTION OF INVESTIGATED DATASETS: NUMBERS OF SAMPLES (n), GENES (d), GIVEN CLASSE NUMBER (k) AND A SET OF POSSIBLE CLASS NUMBERS (C, WHERE $k \in C$).

Dataset	n	d	k	C
Leukemia1 [2]	72	2,194	3	{2, 3, 4, 5, 6}
Leukemia2 [19]	72	1,877	2	{2, 3, 4, 5, 6}
Brain-Tumor [32]	50	1,377	4	{2, 3, 4, 5, 6}
CNS [33]	42	1,379	5	{3, 4, 5, 6, 7}
Muti-Tissues [40]	174	1,571	10	{8, 9, 10, 11, 12}
SRBCT [27]	83	1,069	4	{2, 3, 4, 5, 6}

- *Leukemia1* was originally obtained from the peripheral blood or bone marrow of affected individuals at diagnosis or relapse [2]. In particular, three sample classes are established: 20 cases of lymphoblastic leukemia with MLL translocations (MLL), 24 and 28 conventional acute lymphoblastic (ALL) and acute myelogenous leukemias (AML), respectively.
- *Leukemia2* includes 72 bone marrow samples that were obtained from acute leukemia patients at the time of diagnosis [19]: 47 ALL and 25 AML.
- *Brain-Tumor* contains a collection of 50 gliomas that were exploited in the investigation of [32]: 14 classic glioblastomas (CG), 14 non-classic glioblastomas (NG), 7 classic anaplastic oligodendrogliomas (CO) and 15 non-classic anaplastic oligodendrogliomas (NO).
- *CNS* includes embryonal tumors of the central nervous system studied in [33]: 10 cases of medulloblastomas (MD), 8 primitive neuroectodermal tumors (PNET), 10 atypical teratoid/rhabdoid tumors (Rhab), 10 malignant gliomas (Mglio) and 4 normal tissues.
- *Multi-Tissues* presents the collection of samples used in the study of [40] that determines the categorization of human tumors according to their primary anatomical site of original. A large-scale RNA profiling was used to create a molecular classification scheme, collectively accounting for approximately 70% of all cancer-related deaths in the United States.
- *SRBCT* contains small, round blue-cell tumors that were investigated and classified to diagnostic categories [27]: neuroblastomas (NB), Burkitt lymphoma (BL), rhabdomyosarcoma (RMS) and Ewing (EWS) tumors.

B. Experiment Design

The main focus of this experiment is to investigate the performance of R-KM in comparison with other k-means alike algorithms. This is motivated by the observation that a number of such techniques have been introduced in the literature, providing a good evaluation platform. In addition, results in comparison with other filter-based method, i.e. R-CL and R-SPT, are also obtained to demonstrate the effectiveness and general applicability of the proposed approach. The experiment settings are given below.

- To investigate the robustness of the filter approach, two models of each reliability-based clustering are examined – for instance, R2-KM and R3-KM, with $\alpha = 2$ and $\alpha = 3$, respectively.
- Compared methods include three baseline clustering algorithms of KM, CL and SPT. In addition, three

TABLE II
CA AND NMI MEASURES OF KM ALIKE CLUSTERING METHODS, GIVEN CLASS NUMBER k.

Dataset	**Method**						
	R3-KM	R2-KM	KM	LAC	EWKM	FSC	ProClus
CA Measure							
Leukemia1	0.748	0.738	0.664	0.681	0.551	0.702	0.625
	(0.051)	(0.046)	(0.086)	(0.051)	(0.105)	(0.087)	(0.063)
Leukemia2	0.718	0.713	0.672	0.711	0.672	0.698	0.653
	(0.023)	(0.022)	(0.044)	(0.119)	(0.049)	(0.124)	(0.054)
Brain-Tumor	0.608	0.596	0.566	0.536	0.458	0.534	0.460
	(0.042)	(0.024)	(0.065)	(0.066)	(0.097)	(0.071)	(0.030)
CNS	0.519	0.457	0.410	0.571	0.451	0.441	0.548
	(0.037)	(0.050)	(0.091)	(0.060)	(0.070)	(0.093)	(0.034)
Muti-Tissues	0.655	0.634	0.595	0.615	0.473	0.619	0.316
	(0.061)	(0.055)	(0.075)	(0.068)	(0.077)	(0.077)	(0.025)
SRBCT	0.434	0.463	0.399	0.428	0.464	0.430	0.458
	(0.034)	(0.057)	(0.064)	(0.069)	(0.076)	(0.113)	(0.047)
NMI Measure							
Leukemia1	0.574	0.585	0.447	0.359	0.278	0.484	0.274
	(0.087)	(0.087)	(0.146)	(0.084)	(0.101)	(0.136)	(0.059)
Leukemia2	0.187	0.171	0.092	0.154	0.064	0.202	0.261
	(0.024)	(0.017)	(0.098)	(0.105)	(0.089)	(0.184)	(0.057)
Brain-Tumor	0.493	0.512	0.472	0.353	0.238	0.349	0.184
	(0.072)	(0.076)	(0.122)	(0.068)	(0.108)	(0.107)	(0.041)
CNS	0.423	0.364	0.285	0.491	0.334	0.356	0.376
	(0.034)	(0.025)	(0.113)	(0.059)	(0.096)	(0.094)	(0.028)
Muti-tissues	0.627	0.611	0.573	0.546	0.536	0.561	0.202
	(0.032)	(0.043)	(0.074)	(0.062)	(0.060)	(0.051)	(0.032)
SRBCT	0.153	0.153	0.116	0.108	0.209	0.149	0.183
	(0.040)	(0.034)	(0.110)	(0.104)	(0.090)	(0.092)	(0.050)

soft subspace clustering methods are also employed: EWKM [25], LAC [13] and FSC [18]. In particular,, for each trial, the parameter γ of EWKM is randomly selected from $[0.25, 1]$ and the parameter h of LAC is randomly selected from $\{1, 2, \ldots, 5\}$. Similarly, the parameters δ and ε of FSC are arbitrarily chosen from $(1, 5]$ and $[0.01, 0.1]$, respectively. See Section II for details.

To consolidate the evaluation, the performance of ProClus [1], one of the best known crisp subspace algorithms, is also investigated. Principally, ProClus is a k-medoid-like clustering method. It first randomly selects a set of k potential cluster centers (or medoids), $M = \{m_1, \ldots, m_k\}$, from the object set X. Then, in its iterative cluster refinement phase, the subspace of each medoid $m_g \in M$ is determined by minimizing the standard deviation of distances between m_g and its neighboring objects along each dimension. Following that, objects are assigned to the closest medoid considering the relevant subspace of each medoid. The clusters are refined by replacing low-quality medoids with new medoids from M. This continues as long as the clustering quality (the average similarity between objects and the nearest medoid) increases. In its post-processing step, ProClus identifies outliers, i.e. objects that are excessively far away from their closest medoids. Since M is randomly identified, different runs with the same parametrization usually result in dissimilar clusterings. In this work, the minimum subspace size per cluster which is a mandatory parameter of ProClus, is manually adjusted for each dataset, such that the number of outliers is minimized (i.e. all objects are assigned to one of the disclosed clusters).

- For any clustering technique that is non-deterministic, its quality measure is the average of 20 trials.
- This evaluation compares the quality of partitions generated by the proposed clustering model and other comparable methods, over six gene expression datasets. Two validity indices of CA (Classification Accuracy) [31] and NMI (Normalized Mutual Information) [39] are employed here to gauge the goodness of a data partition.

C. Experiment Results

With an expert-directed cluster numbers k, Table II presents both CA $\in [0, 1]$ and NMI $\in [0, 1]$ measures

TABLE III
CA AND NMI MEASURES OF OTHER RELIABILITY-BASED CLUSTERING METHODS, GIVEN CLASS NUMBER k.

Dataset	**Method**					
	R3-SPT	R2-SPT	SPT	R3-CL	R2-CL	CL
CA Measure						
Leukemia1	0.732	0.732	0.722	0.750	0.750	0.472
	(0.000)	(0.000)	(0.000)	(n/a)	(n/a)	(n/a)
Leukemia2	0.653	0.653	0.653	0.653	0.653	0.653
	(0.000)	(0.000)	(0.000)	(n/a)	(n/a)	(n/a)
Brain-Tumor	0.556	0.554	0.544	0.540	0.540	0.480
	(0.009)	(0.006)	(0.013)	(n/a)	(n/a)	(n/a)
CNS	0.488	0.490	0.381	0.500	0.500	0.405
	(0.029)	(0.043)	(0.061)	(n/a)	(n/a)	(n/a)
Muti-tissues	0.801	0.796	0.770	0.546	0.546	0.523
	(0.014)	(0.014)	(0.042)	(n/a)	(n/a)	(n/a)
SRBCT	0.439	0.435	0.428	0.386	0.373	0.349
	(0.006)	(0.006)	(0.006)	(n/a)	(n/a)	(n/a)
NMI Measure						
Leukemia1	0.653	0.653	0.632	0.694	0.694	0.213
	(0.000)	(0.000)	(0.000)	(n/a)	(n/a)	(n/a)
Leukemia2	0.067	0.067	0.043	0.042	0.042	0.042
	(0.040)	(0.040)	(0.039)	(n/a)	(n/a)	(n/a)
Brain-Tumor	0.394	0.395	0.390	0.369	0.369	0.349
	(0.011)	(0.013)	(0.030)	(n/a)	(n/a)	(n/a)
CNS	0.434	0.437	0.356	0.435	0.435	0.380
	(0.038)	(0.031)	(0.053)	(n/a)	(n/a)	(n/a)
Muti-tissues	0.766	0.764	0.740	0.548	0.540	0.564
	(0.017)	(0.014)	(0.034)	(n/a)	(n/a)	(n/a)
SRBCT	0.122	0.122	0.116	0.100	0.085	0.063
	(0.007)	(0.006)	(0.008)	(n/a)	(n/a)	(n/a)

obtained by KM and its extensions for soft subspace clustering. It is shown that both R2-KM and R3-KM perform consistently better than the baseline, i.e. KM. Furthermore, they are usually more effective than EWKM, LAC, FSC and ProClus. In addition to this finding, Figs. 3-4 illustrate the performance of R3-KM that is robust to a set of possible cluster numbers (C). Note that the measures of R2-KM are similar to those of R3-KM, thus they are not included in these figures for clear presentation.

ProClus, which is a crisp subspace clustering technique, appears to be less effective than its soft subspace clustering counterparts. This scenario occurs due to the fact that ProClus attempts to find a crisp subspace in which several relevant features may be unfortunately dropped. EWKM, LAC and FSC are highly sensitive to their input parameters (γ, h, δ and ε, respectively) where a uniform setting is not obtainable for dissimilar data. For instance, a particular parameter value might cause an extremely drastic change of weights in one dataset, and a constant pace in another. In addition, cluster-specific weights are similarly updated with respect to the distances between object members to the disclosed cluster centroids, which can be sub-optimal. Hence, the accuracy of weight modification cannot always be maintained. This is reflected by their results which are good only with few specific datasets, but generally worse than those of R3-KM and R2-KM. Note that, with the reliability-based KM models, weights are updated using object-specific reliability measures, which represent true characteristics of local relevance and remain unchanged over time.

The results shown in Table III reinforce the observation that the proposed filter approach is effective and generally applicable to different clustering algorithms. In particular, both R3-SPT and R2-SPT improve their baseline model (i.e. SPT), while R3-CL and R2-CL also enhance the performance of the conventional CL technique. In this table, standard deviations of deterministic techniques (CL, R3-CL and R2-CL) are marked as 'n/a', since their performance measures are obtained from a single trial. Based on these findings, the reliability based framework presented here has proven useful for refining the underlying distance measures employed by KM and CL in an original data space, and for that done by SPT in the reduced space (via transformation).

D. Parameter Analysis

To maximize the potential of a soft subspace clustering algorithm, the major obstacle typically encountered is

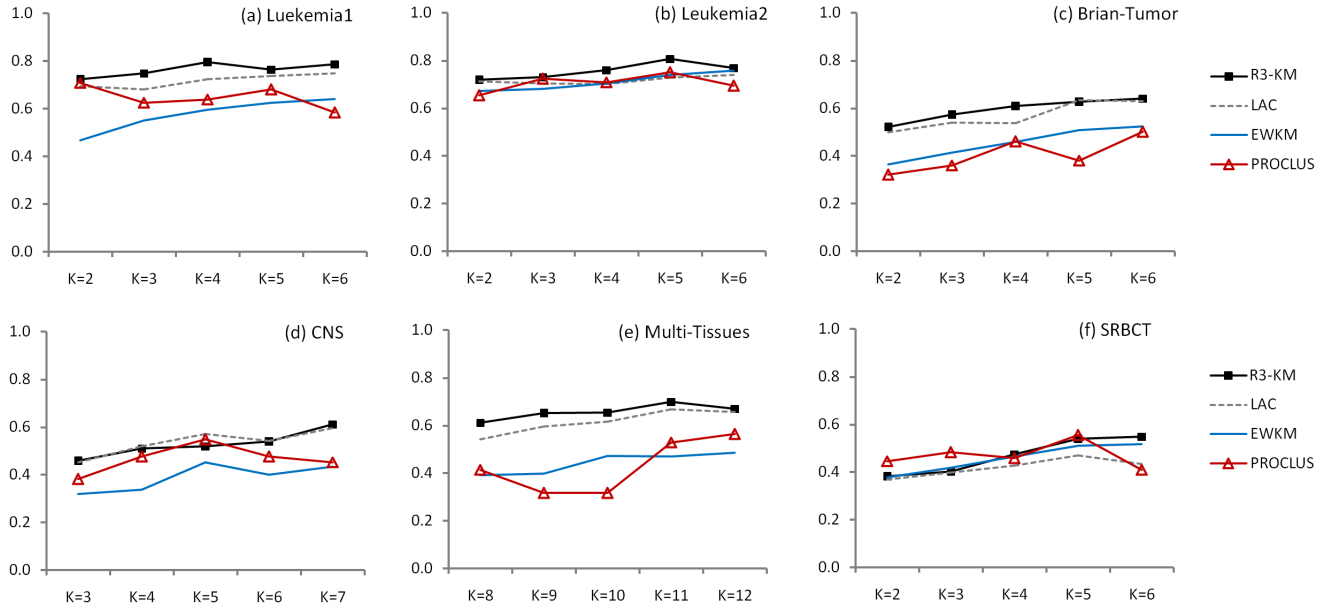

Fig. 3. CA measures obtained by examined clustering methods, displayed in accordance with each dataset and different cluster numbers ($k \in C$).

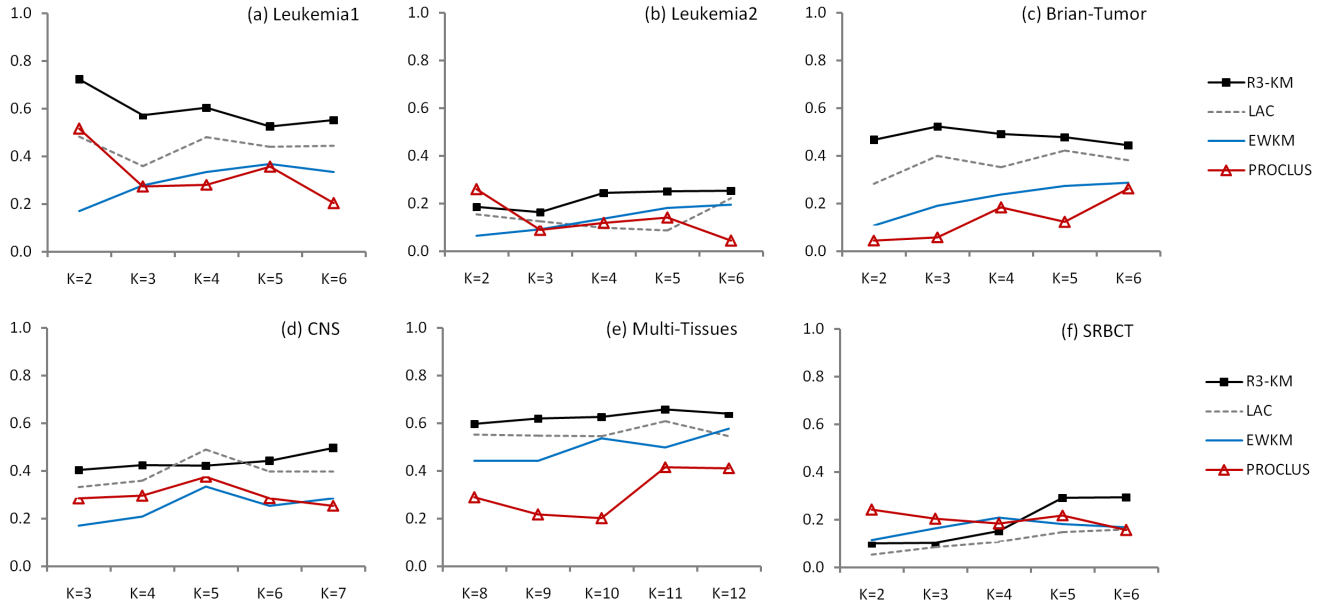

Fig. 4. NMI measures obtained by examined clustering methods, displayed in accordance with each dataset and different cluster numbers ($k \in C$).

the appropriate selection of input parameters. This might also be the case with the use of R-KM, R-CL and R-SPT techniques which require the size of nearest neighbours (α) to be identified before hand. In addition to the above performance comparison, it is therefore important to demonstrate that the effectiveness of filter-based methods is obtainable, with respect to the perturbation of α.

Fig. 5 shows the NMI measures obtained by R-KM using different values of α and a given number of clusters (k). These measures are summarized across all investigated datasets. The results show that the performance of R-KM is not strongly dependent of α and is consistently better than LAC, EWKM and FSC. Similar results are also observed using the CA measure and both R-CL and R-SPT. This empirical evaluation indicates that the new algorithms proposed in this paper are reliable in support of cluster analysis.

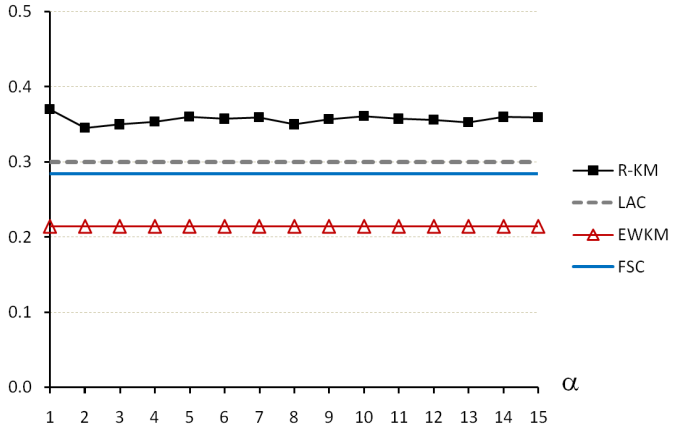

Fig. 5. NMI measures obtained by the R-KM algorithm across all investigated datasets, with different sizes of nearest neighbours (α). Note that the performance measures of other soft subspace clustering techniques (LAC, EWKM and FSC) are included for comparison.

V. Conclusion

This paper has presented a novel filter approach to soft subspace clustering, which is, unlike the existing wrappers, applicable to different clustering algorithms. The underlying measure has also been made more efficient and feasible with large datasets. Based on the evaluation over gene expression data, different reliability-based models improve their corresponding baseline techniques and outperform important soft and crisp subspace clustering methods.

While the popular minimum operator currently employed is effective to summarize cluster-specific feature weights, it has the bias on the smallest and ignores the rest. It is therefore interesting to observe the behavior of reliability-based methods with respect to other types of aggregation operator, e.g. OWA (Ordered Weighted Averaging) [44] and its data-dependant variants [4], [5]. Work is also ongoing to apply, and to further evaluate the potential of, this filter methodology to completely different high-dimensional data, e.g. large-scale true-colors Mars images [36]. Its utilization in the context of cluster ensembles for gene expression data [21] is another significant future research. Additionally, the current performance evaluation is carried out numerically. To support user understanding and interpretation of the results, it may be beneficial to investigate how advanced techniques for fuzzy compositional modeling [16] may be utilized to obtain linguistically valued performance measures [17].

Acknowledgments

The authors would like to thank Dr. Carlotta Domeniconi for the implementation of the LAC algorithm. Thanks also go to the Associate Editor and the reviewers for their invaluable comments which have helped significantly to improve this work.

References

[1] C. Aggarwal, C. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park. Fast algorithms for projected clustering. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 61–72, 1999.

[2] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30:41–47, 2002.

[3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *Proceedings of IDBT*, pages 217–235, 1999.

[4] T. Boongoen and Q. Shen. Clus-DOWA: A New Dependent OWA Operator. In *Proceedings of IEEE International Conference on Fuzzy Sets and Systems*, pages 1057–1063, 2008.

[5] T. Boongoen and Q. Shen. Nearest-neighbor guided evaluation of data reliability and its applications. *IEEE Transactions on Systems, Man and Cybernetics – Part B*, 40(6):1622–1633, 2010.

[6] T. Boongoen, Q. Shen, and C. Price. Disclosing false identity through hybrid link analysis. *AI and Law*, 18(1):77–102, 2010.

[7] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *Proceedings of International Conference on VLDB*, pages 89–100, 2000.

[8] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[9] S. Cleator and A. Ashworth. Molecular profiling of breast cancer: clinical implications. *Br J Cancer*, 90:1120–1124, 2004.

[10] D. S. Cohen and M. Blum. On the problem of sorting burnt pancakes. *Discrete Applied Mathematics*, 61:105–120, 1995.

[11] M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9:497, 2008.

[12] Z. Deng, K. Choi, F. Chung, and S. Wang. Enhanced soft subspace clustering integrating within-cluster and between-cluster information. *Pattern Recognition*, 43(3):767–781, 2010.

[13] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos. Locally adaptive metrics for clustering high dimensional data. *Data Minining and Knowledge Discovery*, 14(1):63–97, 2007.

[14] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[15] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *J. of Royal Statistics Society - B*, 66(4):825–849, 2004.

[16] X. Fu and Q. Shen. Fuzzy compositional modeling. *IEEE Transactions on Fuzzy Systems*, 18(4):823–840, 2010.

[17] X. Fu and Q. Shen. Fuzzy complex numbers and their application for classifiers performance evaluation. *Pattern Recognition*, 44(7):1403–1417, 2011.

[18] G. J. Gan and J. H. Wu. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm. *Pattern Recognition*, 41:1939–1947, 2008.

[19] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[20] J. Gu and J. S. Liu. Bayesian biclustering of gene expression data. *BMC Genomics*, 9:S4, 2008.

[21] N. Iam-on, T. Boongoen, and S. Garrett. LCE: A link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, 26(12):1513–1519, 2010.

[22] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

[23] R. Jensen and Q. Shen. Are more features better? *IEEE Transactions on Fuzzy Systems*, 17(6):1456–1458, 2009.

[24] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[25] L. Jing, M. K. Ng, and J. Z. Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. on Knowledge and Data Engineering*, 19(8):1026–1041, 2007.

[26] I. Joliffe. *Principal Component Analysis*. Springer: New York, 1986.

[27] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, 7(6):673–679, 2001.

[28] B. King. Step-wise clustering procedures. *J. Am. Stat. Assoc.*, 69:86–101, 1967.

[29] H. P. Kriegel, P. Kroger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on KDD*, 3(1):1–ex, 2009.

[30] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in NIPS*, 14, 2001.

[31] N. Nguyen and R. Caruana. Consensus clusterings. In *Proceedings of IEEE International Conference on Data Mining*, pages 607–612, 2007.

[32] C. Nutt, D. Mani, R. Betensky, P. Tamayo, J. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. McLaughlin, T. Batchelor, P. Black, A. Deimling, S. Pomeroy, T. Golub, and D. Louis. Gene expressionbased classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, 2003.

[33] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. Louis, J. Mesirov, E. Lander, and T. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[34] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[35] S. Ramaswamy, K. Ross, E. Lander, and T. Golub. A molecular signature of metastasis in primary solid tumors. *Nat Genet*, 33:49–54, 2003.

[36] C. Shang, D. Barnes, and Q. Shen. Facilitating efficient Mars terrain image classification with fuzzy-rough feature selection. *International Journal of Hybrid Intelligent Systems*, 8(1):3–13, 2011.

[37] C. Shang and Q. Shen. Aiding classification of gene expression data with feature selection: a comparative study. *Computaional Intelligence Research*, 1(1):68–76, 2006.

[38] R. Spang. Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. *BIOSILICO*, 1:264–268, 2003.

[39] A. Strehl and J. Ghosh. Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[40] A. Su, J. Welsh, L. Sapinoso, S. Kern, P. Dimitrov, H. Lapp, P. Schultz, S. Powell, C. Moskaluk, H. Frierson, and G. Hampton. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*, 61(20):7388–7393, 2001.

[41] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(1):136–144, 2002.

[42] G. Tseng and W. Wong. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61:10–16, 2005.

[43] A. Wallqvist, A. Rabow, R. Shoemaker, E. Sausville, and D. Covell. Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. *Mol Cancer Ther*, 1:311–320, 2002.

[44] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18:183–190, 1988.

Tossapon Boongoen is a Postdoctoral Research Associate with the Department of Computer Science at Aberystwyth University, UK. Prior to this appointment, he obtained a PhD in artificial intelligence from Cranfield University, UK and worked as a lecturer at the Royal Thai Air Force Academy, Thailand. His research interests include data mining, link analysis, data clustering, fuzzy aggregation and classification systems.

Changjing Shang is a Research Fellow with the Department of Computer Science at Aberystwyth University, UK. She obtained a PhD in Computing and Electrical Engineering from Heriot-Watt University, UK and worked for Heriot-Watt, Loughborough, Glasgow and Edinburgh Universities prior to joining Aberystwyth. Her research interests include pattern recognition, data mining and analysis, feature extraction and selection, image processing and classification.

Natthakan Iam-On is a PhD candidate with the Department of Computer Science at Aberystwyth University, UK. Her research focuses on cluster ensembles and applications to biomedical data analysis. Prior to this study which is funded by Royal Thai Government, she obtained a BSc and an MSc in computer science from Chiangmai University, Thailand and worked as a lecturer at Mae Fah Luang University, Thailand.

Qiang Shen holds the established Chair in Computer Science and is the Head of the Department of Computer Science at Aberystwyth University, UK. Prof. Shens research interests include: computational intelligence, fuzzy and qualitative modeling, reasoning under uncertainty, pattern recognition, data mining, and real-world applications of such techniques for intelligent decision support (e.g. crime detection, consumer profiling, systems monitoring, and medical diagnosis). Prof. Shen is currently an associate editor of two premier IEEE Transactions (Systems, Man and Cybernetics - Part B, and Fuzzy Systems), and an editorial board member of several other leading international journals. He has authored 2 research monographs, and over 280 peer-reviewed papers, including one which received an Outstanding Transactions Paper Award from IEEE.