Short Papers

A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity

Shaohong Zhang, Hau-San Wong, Ying Shen, and Dongqing Xie

Abstract—Feature selection is widely established as one of the fundamental computational techniques in mining microarray data. Due to the lack of categorized information in practice, unsupervised feature selection is more practically important but correspondingly more difficult. Motivated by the cluster ensemble techniques, which combine multiple clustering solutions into a consensus solution of higher accuracy and stability, recent efforts in unsupervised feature selection proposed to use these consensus solutions as oracles. However, these methods are dependent on both the particular cluster ensemble algorithm used and the knowledge of the true cluster number. These methods will be unsuitable when the true cluster number is not available, which is common in practice. In view of the above problems, a new unsupervised feature ranking method is proposed to evaluate the importance of the features based on consensus affinity. Different from previous works, our method compares the corresponding affinity of each feature between a pair of instances based on the consensus matrix of clustering solutions. As a result, our method alleviates the need to know the true number of clusters and the dependence on particular cluster ensemble approaches as in previous works. Experiments on real gene expression data sets demonstrate significant improvement of the feature ranking results when compared to several state-of-the-art techniques.

Index Terms—Unsupervised feature ranking, gene selection, cluster ensembles.

1 Introduction

In recent decades, high-throughout microarray techniques enable biologists to acquire the expression profile of large number of genes through a single experiment [1], [2], [3]. In general, these acquired microarray expression data can be usually viewed as matrices with large numbers of features (genes), and relatively small numbers of instances (samples). In many bioinformatics applications, feature selection is widely established as an important technique to analyze these kinds of microarray data [4], [5], [6], [7], [8], [9]. Specifically, feature selection techniques aim to search for the most important feature subsets of a given data set or to order the features according to their relevance to the current task. As a result, these techniques are fundamentally important for improved analysis. Recently, a large number of studies have been performed on supervised feature selection for classification problems [4], [10], [11], [12], [13], i.e., learning with a training set in which there are data instances with known class labels. For traditional supervised feature selection techniques, there are two main categories: 1) filtering methods or ranking methods [10], [11], [12], in which the features are evaluated based on the intrinsic structure of the data without involving any learning algorithms. 2) Wrapper methods [4], [13], in which the features are evaluated based on an actual learning algorithm.

However, in many bioinformatics applications, we are only given data without any class label information, and it is usually too expensive to perform the labeling through experts. In view of this limitation, it is important to develop unsupervised approaches which can perform the feature selection task with only the unlabeled data. Compared to the supervised case, unsupervised feature selection is more challenging and thus related methods are relatively scarce. Without the label information, previous studies on unsupervised feature selection tend to adopt a variety of filtering methods. These methods perform ranking on each feature based on different statistical criteria, such as variance or range [5], [14], [15], [16], spectral properties of the Laplacian Score (Laplacian) [17], [18], [19], principal component analysis [15], and Ranking by SVD-Entropy [6]. The reasons of adopting filtering methods include the followings 1) for unsupervised learning, there are no available labels for learning a classification model. In addition, many popular clustering algorithms tend to produce different results with different initializations, as in the case of the well-known clustering algorithm Kmeans [20]; 2) filtering methods outperform wrapper methods in terms of efficiency since the latter usually involves an iterative search process. This advantage becomes more important when the number of features increases, especially for gene expression data which might contain thousands of genes.

Recent studies on unsupervised feature selection attempt to adopt multiple clustering results on the same data set to improve results [21], [22], [23]. A recent approach in this category, known as Unsupervised Feature Ranking from Multiple Views (FRMV) proposed in [21], combines the individual rankings of the features with respect to different clustering solutions of the same data set into a consensus ranking. Another important progress in unsupervised feature selection follows from similar improvements of clustering results based on cluster ensembles. Specifically, cluster ensemble [24] provides an alternative framework to obtain a more accurate solution from combining a number of individual clustering solutions. In general, cluster ensemble techniques include two main phases: 1) generation of the individual clustering solutions, and 2) combination of these individual solutions into a consensus result. Representative examples of cluster ensemble methods include the Cluster-Based Similarity Partitioning Algorithm (CSPA) [24], the Hybrid Bipartite Graph Formulation algorithm (HBGF) [25], the HyperGraph Partition Algorithm (HGPA) [24], and the Evidence Accumulation (EAC) method based on the hierarchical agglomerative clustering algorithm [26]. Consensus solutions generated from cluster ensemble techniques are usually more accurate and stable when compared with the individual clustering solutions [24], [26], [27]. In view of these characteristics, the works described in [22] and [23] use the consensus solutions from cluster ensembles to provide estimated class labels for supervised wrapper methods.

Although promising performances are reported for various newly proposed methods [21], [22], [23], they all require the strong assumption that the true cluster number be known, which is usually unavailable in practice [28]. Also, all of these methods depend on the consensus solution generated by a specific cluster ensemble method. However, how to select the best set of features in an unsupervised manner is still an open problem. In view of this, we propose a new feature ranking method, which is referred to as Feature Ranking based on the Consensus Matrix (FRCM). FRCM includes three main phases: 1) computation of the consensus matrix from a number of individual clustering solutions; 2) computation of the associated affinity matrix for each feature between pairwise entries of all instances; and 3) ranking of each feature with respect to the consistency between their corresponding affinity matrix and the consensus matrix. To our best knowledge, the proposed FRCM method represents a first attempt to use cluster ensemble in unsupervised feature ranking based on matrix comparison. Compared with previous works [21], [22], [23], our proposed

- *S. Zhang is with the Department of Computer Science, Guangzhou University, Guangzhou, P.R. China and the Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China. E-mail: zimzsh@gmail.com.*
- *H.-S. Wong and Y. Shen are with the Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China. E-mail: cshswong@cityu.edu.hk.*
- *D. Xie is with the Department of Computer Science, Guangzhou University, Guangzhou, P.R. China. E-mail: dongqing_xie@hotmail.com.*

Manuscript received 23 July 2011; revised 2 Feb. 2012; accepted 4 Feb. 2012; published online 21 Feb. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-07-0194. Digital Object Identifier no. 10.1109/TCBB.2012.34.

method has several differences: 1) FRCM performs feature ranking based on pairwise similarity matrices, which is more general than previous methods which use class labels; 2) FRCM removes the requirement of knowing the true cluster number; and 3) FRCM does not depend on any particular cluster ensemble method to generate the final consensus solution as in previous related works.

2 Proposed Framework

We first introduce some related background knowledge of clustering and cluster ensemble used in this paper. In our work, we use Kmeans [20] as our clustering algorithm since it is one of the most well-known algorithms. Consider a data set with N instances, $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$, in which each instance $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$ has D features. In an unsupervised manner, Kmeans groups these instances into a partition P with a predefined number K of clusters, such that the following cost function is minimized:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2, \tag{1}$$

where $\boldsymbol{\mu}_k$ is the associated centroid of partition P_k, which can be computed as follows:

$$\boldsymbol{\mu}_k = \frac{\sum_{\boldsymbol{x}_i \in P_k} \boldsymbol{x}_i}{|P_k|}, \tag{2}$$

in which $|P_k|$ is the cardinality of the cluster X_k.

The partition P generated by Kmeans consists of K mutually disjoint clusters $\{P_k\}_{k=1}^{K}$, i.e.,

$$P = \{P_k\}_{k=1}^{K}, \text{ where } P_{k_1} \bigcap P_{k_2} = \emptyset \text{ and } \bigcup_{k=1}^{K} P_k = X. \tag{3}$$

In addition, the $N \times N$ coassociation matrix generated from the partition P is defined as follows:

$$M_{ij} = \begin{cases} 1, & \text{if } \exists\ k,\ \boldsymbol{x}_i \in P_k \text{ and } \boldsymbol{x}_j \in P_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For a set of clustering solutions $\{P^{(t)}\}_{t=1}^{T}$, the $N \times N$ consensus matrix can be constructed as the average of their co-association matrices as follows:

$$\mathcal{M} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{M}^{(t)}. \tag{5}$$

A number of well-known cluster ensemble algorithms [24], [26] use this consensus matrix to generate the final partition.

Motivated by the above cluster ensemble framework and previous related works [21], [22], [23], [29], we propose the following Feature Ranking based on the Consensus Matrix method. Given a data set X, FRCM first generates a consensus matrix $\mathcal{M}$ from a number of individual clustering partitions using Kmeans on a set of data instances with their features sampled at random. Then, an affinity matrix $\mathcal{A}$ associated with each feature between all instance pairs is computed. Finally, each feature is ranked with respect to the consistency between their corresponding affinity matrix and the consensus matrix. The rank of each feature is regarded as a measure of its importance.

2.1 Generation of the Consensus Matrix from the Cluster Ensemble

In the first step, FRCM generates a consensus matrix based on the cluster ensemble constructed from a number of individual clustering solutions. Specifically, for a data set X with N instances and D features, a reduced subspace with a set of randomly sampled $D/2$ features is constructed. The N instances in this reduced subspace are then clustered using Kmeans with the cluster number k randomly selected from the range $\{2, \ldots, K_{max}\}$, where

$$K_{max} = min\left(\sqrt{N}, K_\nu\right), \tag{6}$$

in which $min(a, b)$ returns the smaller value between a and b, and K_ν is a parameter. The clustering process is repeated for T times to generate the corresponding coassociation matrices defined in (4), and these matrices are used to construct the consensus matrix defined in (5).

2.2 Affinity Matrix Associated with Each Feature

We introduce the following affinity matrix for characterizing the contribution of a particular feature to the overall pairwise similarity between the data instances, given all the other features. For a data set X with N instances and D features, the entry of the affinity matrix corresponding to two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with respect to the feature d is defined as follows:

$$\mathcal{A}_{ij}^{(d)} = \sqrt{1 - \frac{(x_{id} - x_{jd})^2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}}. \tag{7}$$

As can be seen above, if the term $(x_{id} - x_{jd})^2$ from feature d dominates the distance between the two instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, i.e., $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, the contribution of feature d in the affinity matrix $\mathcal{A}_{ij}^{(d)}$ is close to zero. On the other hand, if the term $(x_{id} - x_{jd})^2$ is small in comparison to the overall distance, the corresponding matrix entry will be close to one. Thus, the affinity matrix $\mathcal{A}^{(d)}$ can be viewed as the similarity between the data instances from the perspective of a particular feature d. As a result, through a comparison between the affinity matrices and the consensus matrix, we can rank the features according to their importance.

2.3 Consistency between the Affinity Matrices and the Consensus Matrix

As mentioned above, once we have obtained the affinity matrices and the consensus matrix, the main task is how to evaluate the consistency between them. One important observation is that, in a consensus matrix, there are usually more entries close to zero than those close to one. This is in agreement with corresponding observations which motivate the formulation of the well-known clustering comparison measure, the Rand Index, based on pairwise agreement counting [29], [30], [31]. Specifically, for a data set X with N instances, let N_{same} be the count of instance pairs belonging to the same cluster in both partitions and N_{diff} be the count of those belonging to different clusters in the two partitions, the Rand index can be computed as follows [32], [33]:

$$R = \frac{N_{same} + N_{diff}}{N(N-1)/2}. \tag{8}$$

When the number of clusters K increases, the probability of a particular instance pair belonging to the same cluster becomes much smaller than that of the pair assigned to two different clusters, i.e., the ratio N_{diff}/N_{same} increases with K. Thus, the Rand index tends to give an inflated score in the case of a large number of clusters [34]. Also, the Rand index values tend to be different for different pairs of random clustering partitions, which is another well-known defect [30], [31]. Different improved measures are proposed to overcome these problems of the Rand index, and the most popular of these is the Adjusted Rand Index (ARI) [30]. Specifically, suppose $P = \{P_1, P_2, \ldots, P_{K^{(P)}}\}$ and $Q = \{Q_1, Q_2, \ldots, Q_{K^{(Q)}}\}$ are two partitions on a data set X with N instances, N_{ij} is the number of instances in cluster P_i in partition P and in cluster Q_j in partition Q, $N_{i\cdot}$ is the number of instances in cluster P_i in partition P, and $N_{\cdot j}$ is the number of instances in cluster Q_j in partition Q. ARI is defined as follows:

$$r_0 = \sum_{i=1}^{K^{(P)}} \sum_{j=1}^{K^{(Q)}} \binom{N_{ij}}{2}, r_1 = \sum_{i=1}^{K^{(P)}} \binom{N_{i\cdot}}{2}, \\ r_2 = \sum_{j=1}^{K^{(Q)}} \binom{N_{\cdot j}}{2}, r_3 = \frac{2r_1 r_2}{N(N-1)}, \tag{9}$$

$$ARI(P,Q) = \frac{r_0 - r_3}{0.5(r_1 + r_2) - r_3}, \tag{10}$$

where $\binom{n}{k}$ is the binomial coefficient.

Motivated by the fuzzy extension of ARI [35] and the newly proposed measure ARI between similarity matrix and cluster partitions (ARImp) [29], we generalize ARI to a new type of matrix comparison measure, which is referred to as Adjusted Rand Index between matrices (ARImm), and use it as a measure to evaluate the consistency between the affinity matrices and the consensus matrix. Specifically, given an $N \times N$ consensus matrix $\mathcal{M}$ and an $N \times N$ affinity matrix $\mathcal{A}$, ARImm is defined as follows:

$$s_0 = \sum_{i,j,i\neq j} \frac{\mathcal{M}_{ij}\mathcal{A}_{ij}}{2}, s_1 = \sum_{i,j,i\neq j} \frac{\mathcal{M}_{ij}}{2}, \\ s_2 = \sum_{i,j,i\neq j} \frac{\mathcal{A}_{ij}}{2}, s_3 = \frac{2s_1 s_2}{N(N-1)}, \tag{11}$$

$$ARImm(\mathcal{M},\mathcal{A}) = \frac{s_0 - s_3}{0.5(s_1 + s_2) - s_3}. \tag{12}$$

Similar to ARI, ARImm has the same desirable properties of ARI. Specifically, for two consensus matrices constructed from two independent sets of different partitions, the ARImm score between them is very close to zero. When the two matrices have higher consistency than cases arising from pure chance, ARImm between them turns out to be a positive value. Thus, ARImm can be used as the measure to evaluate the consistency between the affinity matrices and the consensus matrix.

2.4 Summary

The complete FRCM approach is summarized in Algorithm 1.

Algorithm 1: FRCM

input : $N \times D$ microarray data $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i = \{x_{id}\}_{d=1}^{D}$;
input : maximum number of clusters K_ν;
input : number of individual clustering solutions T;
output: gene ranking list L.

$K_{max} \leftarrow \min(\sqrt{N}, K_\nu)$;
for *each individual clustering solution* $t \leftarrow 1$ **to** T **do**
  generate a reduced subset X_t with $D/2$ genes;
  sample a cluster number K_t from $\{2, \cdots, K_{max}\}$;
  cluster X_t with `kmeans`;
  compute the co-association matrix $\boldsymbol{M}^{(t)}$ using (4);
end
compute the consensus matrix $\mathcal{M}$ using (5);
for *each gene* $d \leftarrow 1$ **to** D **do**
  compute the affinity matrix $\mathcal{A}^{(d)}$ using (7);
  compute $\boldsymbol{z}^{(d)} \leftarrow ARImm(\mathcal{M}, \mathcal{A}^{(d)})$ using (12);
end
rank the genes according to $\boldsymbol{z}$ in descending order;
return the gene ranking list L;

2.5 Computational Complexity

Based on Algorithm 1, we estimate the computational complexity of FRCM as follows:

Time complexity. The main steps include.

- Generation of T individual clusterings using Kmeans on $D/2$ genes with the number of clusters not more than K_{max}. Assume the Kmeans clustering algorithm converges in at most C iterations (C is usually smaller than 100), this step takes $O(TK_{max}CND)$.
- Generation of T coassociation matrices, which takes $O(TN^2)$.
- Generation of the consensus matrix, which takes $O(TN^2)$.
- Generation of the affinity matrix for all the D genes, which takes $O(N^2D^2)$.
- Computation of the ARImm scores for all the D genes, which takes $O(N^2D)$.
- Quick sorting of the ARImm scores to generate the gene ranking list L, which takes $O(D \log D)$.

For gene expression data, the number of genes D is the dominant factor. Thus the overall FRCM time complexity is $O(N^2D^2)$.

Space complexity. The main steps include.

- Generation of T individual clusterings using Kmeans on $D/2$ genes with the number of clusters not more than K_{max}. Note that in addition to the data, there are additional memory requirements for storing the cluster information, which together takes $S((N + K_{max})D)$.
- Generation of T coassociation matrices, which takes $S(TN^2)$.
- Generation of the consensus matrix, which takes $S(N^2)$.
- Generation of the affinity matrix for all the D genes, which takes $S(N^2D)$.
- Computation of the ARImm scores for all the D genes, which takes $S(D)$.
- Quick sorting of the ARImm scores to generate the gene ranking list L, which takes $S(\log D)$.

For gene expression data, the number of genes D is the dominant factor. Thus the overall FRCM space complexity is $S(N^2D)$.

3 Results

3.1 Experiment Settings

In this section, we conduct experiments on a number of data sets to evaluate the performance of our FRCM approach for unsupervised feature selection. We first introduce our experimental settings, including the data sets used, the parameter selections, the previous algorithms to be compared with, and the evaluation metrics.

Data sets. We conduct our experiments on four data sets from the well-known UCI machine learning repository,[1] and four public gene expression data sets. Details of the gene expression data are listed as follows:

- **armstrong-2002-v2** [36], which consists of three classes of mixed-lineage leukemia genes: 24 conventional acute lymphoblastic (ALL), 20 lymphoblastic leukemias with MLL translocations (MLL), and 28 acute myelogenous leukemias (AML). This data set is preprocessed as described in [37].
- **nutt-2003-v1** [38], which consists of four classes: 14 classic glioblastomas (CG), 7 classic anaplastic oligodendrogliomas (CO), 14 nonclassic glioblastomas (NG) and 15 nonclassic anaplastic oligodendrogliomas (NO). This data set is preprocessed as described in [37].

1. http://archive.ics.uci.edu/ml/.

TABLE 1
Summary of the Data Sets After Preprocessing

Data set	Source	K	N	D
Dermatology	UCI	6	366	34
Image-Segmentation-test	UCI	7	2100	19
Pendigits-test	UCI	10	3498	16
Wine	UCI	3	178	13
armstrong-2002-v2	[36]	3	72	2194
nutt-2003-v1	[38]	4	50	1377
tomlins-2006	[39]	5	104	2315
bredel-2005	[40]	3	50	1739

- **tomlins-2006** [39], which consists of five classes: 27 benign epithelium (EPI), 20 metastatic disease (MET), 32 prostate cancer (PCA), 13 putative precursor lesions prostatic intraepithelial neoplasia (PIN) and 12 stromal (STROMA). This data set is preprocessed as described in [37].
- **bredel-2005** [40], which consists of three classes: 31 pure glioblastomas (GBM), 14 tumors with enrichment for oligodendroglial morphology (OG) and five grade 1-3 astrocytomas (A). This data set is preprocessed as described in [37].

Table 1 provides information about these data sets, in which K is the number of classes, N is the number of points of the data set and D is the number of features.

Parameter selection. We adopt Kmeans, based on the euclidean distance, as our clustering algorithm. There are only two parameters to be considered: 1) K_ν, the maximum number of clusters in (6), and 2) T, the number of different clustering solutions in (5). We use the same parameter values as in [21] for all the data sets, i.e., $K_\nu = 20$ and $T = 100$.

Previous algorithms to be compared with. The following three feature selection methods are used for comparison with the proposed approach:

- Laplacian Score [17], which ranks the features according to their roles in the preservation of the local manifold structure. The number of nearest neighbors is set to five as used in [17].
- Unsupervised Feature Filtering (UFF) [6], which ranks the features according to their contribution to the SVD-based entropy.
- Unsupervised Feature Ranking from Multiple Views [21], which combines a number of feature ranking results into a consensus one. The number of clustering solutions is set to 100 as used in the paper [21]. Note that these clustering solutions are also used in our proposed FRCM method.

Evaluation metrics. We use the Normalized Mutual Information (NMI) measure [24] and Adjusted Rand Index [30] as evaluation measures in our experiments. For all the experiments, the mean results of 40 independent trials are reported.

3.2 Experimental Results: Comparison with Different Feature Ranking Methods

Clustering performances of different feature subsets of the UCI data sets and the gene expression data sets are shown in Figs. 1 and 2, respectively. For the UCI data sets which only have a small number of features, we perform clustering on all possible d best feature subsets, i.e., $d = 1, 2, 3, \ldots, D$, where D is the total number of features. Since the gene expression data sets contain thousands of features, we perform clustering on those d best feature subsets where $d = 20, 40, 60, \ldots, D$. Mean results of 40 independent trials are presented. We can observe from Figs. 1 and 2 that FRCM has the best performance among all the methods. Specifically, for the data sets of armstrong-2002-v2 and bredel-2005, while the performances of the feature subsets selected by FRMV, Laplacian, and UFF improve when the number of included features increases, FRCM attains the best performance for the initially selected small subsets. These curves converge with each other when the number of included features is close to D. For the data sets of nutt-2003-v1 and tomlins-2006, FRCM has a significantly higher NMI performance for the initial small subsets than any other method, and the performance of all the methods further improves as the number of selected features increases.

Another observation is that FRMV based on the same set of multiple clustering solutions only significantly outperforms the other two competitors (i.e., Laplacian and UFF) on three of the data sets, armstrong-2002-v2, tomlins-2006 and bredel-2005. For the data set nutt-2003-v1, the FRMV result is not satisfactory. A possible reason is that, in this case, the class information is less correlated with the consensus affinity information of the data, and the results based on unsupervised feature ranking and clustering are thus less indicative of the original class assignment. In addition, while in some cases an ensemble of multiple ranking lists of features can achieve better results, this is not always the case when the quality of individual clustering solutions is weak. The weak quality may be due to the following reasons: 1) without supervised information, the instability of Kmeans might lead to inferior clustering results; and 2) without the knowledge of the true number of clusters, those clustering solutions associated with an incorrect number of clusters, might be very different from the

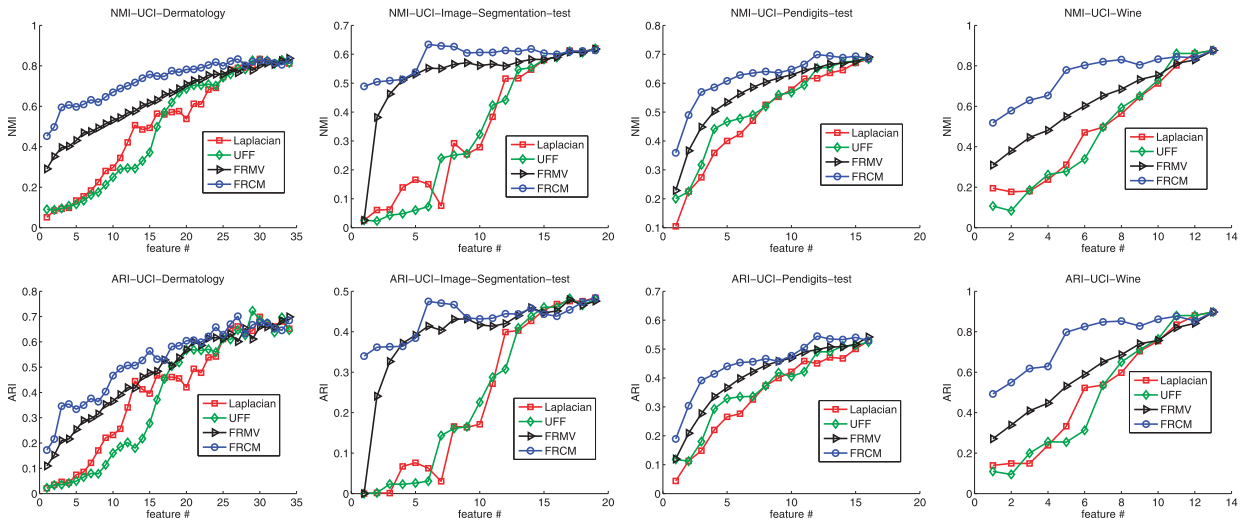

Fig. 1. UCI data sets: clustering performance as a function of different selected feature subset sizes by unsupervised feature filtering methods: Laplacian Score, Unsupervised Feature Filtering, FRMV and Feature Ranking based on the Consensus Matrix. FRCM achieves the best performance.

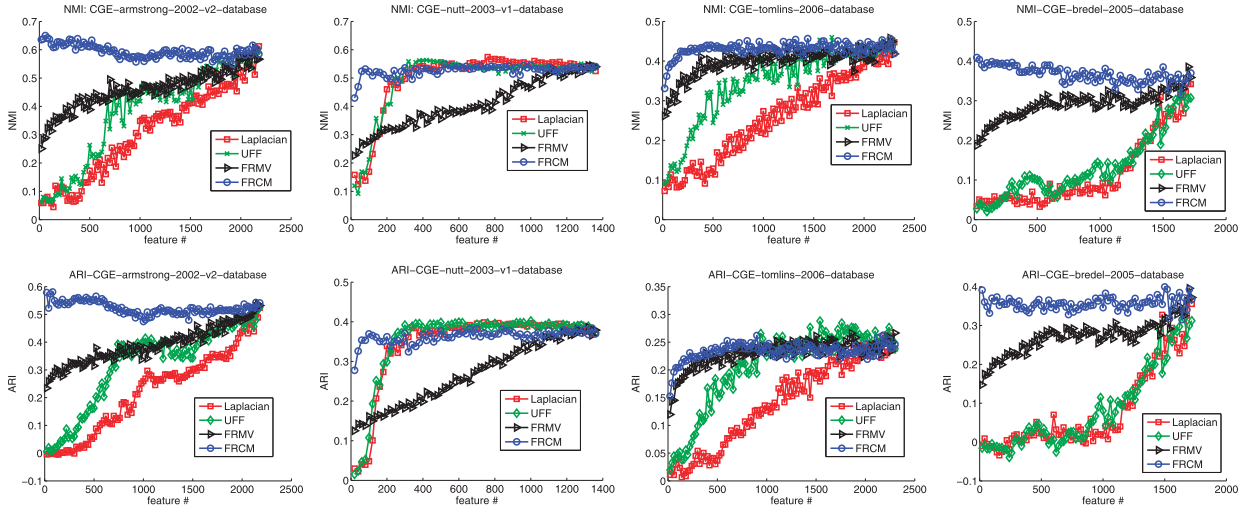

Fig. 2. Gene expression data sets: clustering performance as a function of different selected feature subset sizes by unsupervised feature filtering methods: Laplacian Score, Unsupervised Feature Filtering, FRMV and Feature Ranking based on the Consensus Matrix. FRCM achieves the best performance.

ground truth and thus mislead the feature ranking results. In contrast, we can observe that FRCM based on the consensus matrix and the affinity matrices is consistently better than all the competitors including FRMV. This is in accordance with the well-known observation that the consensus combination of weak partitions can usually help to identify the true underlying clusters with complex structure [26], [41]. This desirable property are also exploited in recent works on unsupervised wrapper feature selection [22], [23], which use the consensus matrix to obtain a final clustering solution.

It can also be observed that FRCM performs especially well with small number of selected features in gene expression data sets (e.g., approximately 100 features, i.e., $\leq 10\%$ of the complete feature set). This is particularly important for gene expression data sets which may include thousands of features. It also suggests that, while the selected feature subsets based on FRCM are compact, their representativeness is not compromised.

Finally, we compare the selected informative subsets based on the principle proposed in [6]. The selected feature subset described in [6] includes features with high contribution to the SVD Entropy (CE). Specifically, a feature with $CE > c + \sigma$ is selected where c is the average of all the CE values, and σ is the standard deviation. We select the informative subsets for the other methods in a similar way as UFF. Sizes of these selected subsets and corresponding clustering performances on these subsets from 40 independent trials are shown in Table 2. From this table, we can observe that the selected subsets of the four methods have fairly similar sizes in three data sets, except in the case of the nutt-2003-v1 data set where FRCM selects a smaller subset. For all the data sets, the standard deviation values of these sizes in 40 independent trials are rather small when compared to the mean size values. For the clustering performance of these subsets, we find that those selected by FRCM have much better results than the other three methods. It is also notable that the standard deviation values of the results by FRCM are relatively small. These observations suggest that the selected subsets by FRCM are more discriminative, and the clustering results are more accurate and more robust. We also compare FRCM to another approach, Principal Component Analysis (PCA), using the same feature dimensionality as FRCM. In addition to achieving a similar performance as PCA, the importance of FRCM is in its capability, with its associated ranked gene list, to provide biological insights for further downstream analysis of the gene expression data. On the other hand, it will be difficult to obtain suitable biological interpretation of the PCA dimensionality reduction results.

3.3 Experimental Results: Comparison with Features Selected Based on Cluster Ensemble

We also compare the performance of FRCM with those based on the selected features using different cluster ensemble approaches. Specifically, the consensus partition of a cluster ensemble of 100 different individual clustering solutions generated as before is obtained based on three well-known methods: Cluster-Based Similarity Partitioning Algorithm [24], Hybrid Bipartite Graph Formulation algorithm [25], and cluster ensemble based on the Average-Link hierarchical clustering algorithm (AL) [26], with the cluster number ranging from 2 to 10. Features are selected based on the linear correlation coefficient approach with reference to each consensus solution. Clustering results of the 10 percent best

TABLE 2
Experimental Results Using Kmeans: The Number of Selected Features, and the Clustering Performance on the Selected Subsets

	armstrong-2002-v2			nutt-2003-v1		
Method	#features	NMI	ARI	#features	NMI	ARI
Laplacian	335	0.056 ± 0.049	0.006 ± 0.034	247	0.474 ± 0.045	0.328 ± 0.036
UFF	342	0.143 ± 0.077	0.105 ± 0.089	228	0.450 ± 0.065	0.333 ± 0.065
FRMV	395 ± 38	0.436 ± 0.110	0.355 ± 0.131	247 ± 20	0.309 ± 0.094	0.176 ± 0.080
FRCM	337 ± 8	0.626 ± 0.035	0.552 ± 0.063	113 ± 9	0.531 ± 0.057	0.373 ± 0.056
PCA	set to FRCM dimensionality	0.593 ± 0.055	0.542 ± 0.056	set to FRCM dimensionality	0.533 ± 0.029	0.374 ± 0.032
	tomlins-2006			bredel-2005		
Method	#features	NMI	ARI	#features	NMI	ARI
Laplacian	389	0.124 ± 0.056	0.044 ± 0.039	270	0.049 ± 0.023	0.002 ± 0.059
UFF	380	0.246 ± 0.074	0.107 ± 0.057	254	0.062 ± 0.028	-0.026 ± 0.021
FRMV	411 ± 21	0.380 ± 0.066	0.218 ± 0.062	305 ± 24	0.260 ± 0.138	0.227 ± 0.168
FRCM	374 ± 6	0.427 ± 0.033	0.225 ± 0.029	254 ± 6	0.394 ± 0.034	0.361 ± 0.047
PCA	set to FRCM dimensionality	0.443 ± 0.051	0.250 ± 0.064	set to FRCM dimensionality	0.346 ± 0.096	0.361 ± 0.129

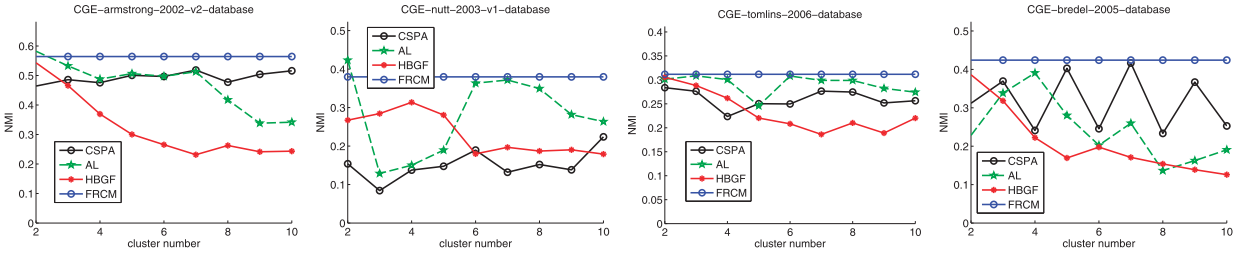

Fig. 3. Clustering performance of the 10 percent best features selected by FRCM, and those based on consensus solutions obtained from the same cluster ensemble using three well-known methods: CSPA, AL, and HBGF. FRCM achieves the best performance.

features are evaluated with reference to the ground truth using NMI, and the mean results of 40 independent trials are shown in Fig. 3. Note that the cluster number is not used in FRCM, and thus the corresponding curve appears horizontal. From Fig. 3, we can observe that FRCM outperforms the other approaches in most cases, except on the data sets armstrong-2002-v2 and nutt-2003-v1 with cluster number equal to two, where the performance of AL is slightly better. We can also observe that, except for the case of HBGF where the performance tends to drop with an increase of the cluster number, no clear relationship between the final performance and the adopted cluster number can be observed. Of the three previous cluster ensemble methods used, it is difficult to determine which one is to be chosen. It is also notable that these three methods do not necessarily achieve the best results when the true cluster number is adopted in constructing the cluster ensemble. These observations highlight the advantage of FRCM in adopting the consensus matrix as the reference oracle, which avoids the problems of selecting the optimal number of clusters and a suitable cluster ensemble method.

3.4 Experimental Results: Correlation with the Class Information

We have also performed a detailed study on the correlation between the ranked gene subsets by FRCM with the ground truth class information. Specifically, the genes are first assigned Borda scores [42] according to their importance. The best feature will be scored D, and the worst one will be scored one. The other features are scored accordingly, e.g., the dth best feature will be scored $D - d + 1$. For each data set, mean Borda values of selected gene subsets of different sizes expressed as a percentage of the complete feature set are presented as curves in Fig. 4. If the ranking is effective, the initial small-sized subsets should have higher mean Borda values than larger subsets. However, since the feature ranking process and the clustering process are both unsupervised, there will be some unavoidable gaps between the ranking results and the class information, especially for the gene expression data sets, which may be affected by different factors such as high dimensionality, outliers, and noise. In general, we can observe that the Borda values tend to decrease when the sizes of the gene subsets increase. This observation suggests that the ranking by FRCM has the expected level of consistency with the class information associated with the different data sets.

4 Conclusion

In this paper, we have proposed a new principle for unsupervised feature ranking based on pairwise similarity matrices. Specifically, we use the consensus matrix associated with a cluster ensemble as the primary reference information, a set of affinity matrices as representatives for the features, and the ARImm measure as a comparison metric between these two types of similarity matrices. Although wrapper-based feature selection using a consensus partition from a cluster ensemble has been proposed in recent studies, all of them require the generation of a final consensus partition and the knowledge of the true number of clusters in advance, which is usually unavailable in practice. The main contribution of our work is thus to reduce the dependence of the ranking results on the adoption of particular cluster ensemble methods and the knowledge of the true number of clusters.

We compare our approach with several previous unsupervised feature ranking methods on different UCI and gene expression data sets. The results show that FRCM significantly outperforms the other methods, which is in accordance with the observation that the consensus clustering of weak partitions can usually help in identifying the true underlying clusters with complex structure. A particular notable observation is that for the gene expression data sets, the initial subsets with small sizes (e.g., approximately 100 features, $\leq 10\%$ of the complete set of features) selected by FRCM result in very good performances. This is particularly important for these data set types which may include thousands of features. We also find that FRCM even outperforms those selected feature sets based on consensus partitions generated using different approaches from the same cluster ensemble. These results suggest that, FRCM can select parsimonious feature subsets which perform well on different types of data. Finally, we also observe that the ranking results by FRCM correlate well with the ground truth class distribution of most of the data sets. This indicates that we can make use of information associated with a cluster ensemble in a different way for unsupervised feature selection, without requiring the generation of a final consensus partition. This desirable property also enables FRCM to be applicable in more general circumstances.

Acknowledgments

The work described in this paper was supported by a grant from the City University of Hong Kong [Project No. 7008044] and a research grant from the Joint Funds of NSFC-Guangdong of China [Project No. U1135002].

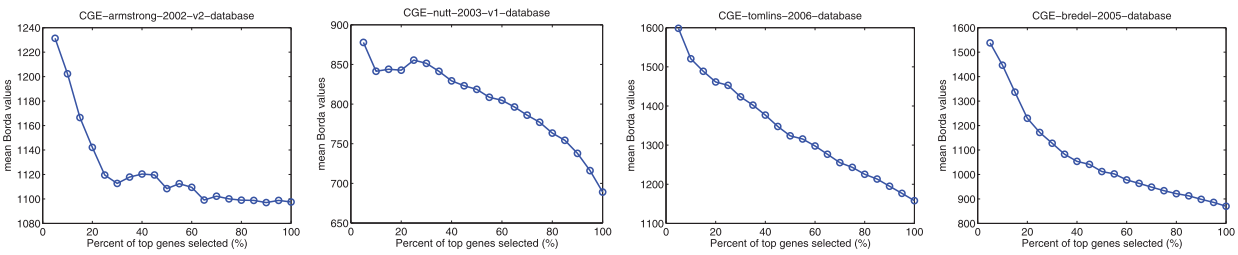

Fig. 4. Mean Borda values which measure the extent of correlation of selected gene subsets of different sizes by FRCM with the ground truth class information. It can be observed that the mean Borda values corresponding to all the four data sets in general decrease when the sizes of the selected gene subsets increase.

REFERENCES

[1] J. Quackenbush, "Computational Analysis of Microarray Data," *Nature Rev. Genetics,* vol. 2, no. 6, pp. 418-427, 2001.

[2] P. Baldi and G. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge Univ. Press, 2002.

[3] N. Armstrong and M. van de Wiel, "Microarray Data Analysis: From Hypotheses to Conclusions Using Gene Expression Data," *Cellular Oncology,* vol. 26, nos. 5/6, pp. 279-290, 2004.

[4] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[5] J. Herrero, R. Dlaz-Uriarte, and J. Dopazo, "Gene Expression Data Preprocessing," *Bioinformatics,* vol. 19, no. 5, pp. 655-656, http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics19.html#HerreroDD03, 2003.

[6] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *Bioinformatics,* vol. 22, no. 14, pp. 507-513, 2006.

[7] Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[8] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature Selection for Gene Expression Using Model-Based Entropy," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 1, pp. 25-36, Jan.-Mar. 2010.

[9] F. Yang and K. Mao, "Robust Feature Selection for Microarray Data Based on Multi-Criterion Fusion," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 8, no. 4, pp. 1080-1092, July/Aug. 2011.

[10] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection—A Filter Solution," *Proc. 13th Int'l Conf. Machine Learning,* pp. 319-327, 1996.

[11] M. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," *Proc. 17th Int'l Conf. Machine Learning,* pp. 359-366, 2000.

[12] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proc. 20th Int'l Conf. Machine Learning,* vol. 20, no. 2, pp. 856-863, 2003.

[13] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, nos. 1/2, pp. 273-324, 1997.

[14] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.

[15] C. Ding, "Unsupervised Feature Selection via Two-Way Ordering in Gene Expression Analysis," *Bioinformatics,* vol. 19, no. 10, pp. 1259-1266, 2003.

[16] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research,* vol. 3, pp. 1157-1182, http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf, 2003.

[17] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," *Proc. Advances in Neural Information Processing Systems (NIPS),* http://dblp.uni-trier.de/db/conf/nips/nips2005.html#HeCN05, 2005.

[18] L. Wolf and A. Shashua, "Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach," *J. Machine Learning Research,* vol. 6, pp. 1855-1887, http://dblp.uni-trier.de/db/journals/jmlr/jmlr6.html#WolfS05, 2005.

[19] Z. Zhao and H. Liu, "Spectral Feature Selection for Supervised and Unsupervised Learning," *Proc. 24th Int'l Conf. Machine Learning,* pp. 1151-1157, 2007.

[20] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Math. Statistics and Probability,* vol. 1, pp. 281-297, 1967.

[21] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, "Consensus Unsupervised Feature Ranking from Multiple Views," *Pattern Recognition Letters,* vol. 29, no. 5, pp. 595-602, 2008.

[22] H. Elghazel and A. Aussem, "Feature Selection for Unsupervised Learning Using Random Cluster Ensembles," *Proc. IEEE 10th Int'l Conf. Data Mining,* pp. 168-175, 2010.

[23] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, "Unsupervised Feature Selection Using Clustering Ensembles and Population-Based Incremental Learning Algorithm," *Pattern Recognition,* vol. 41, no. 9, pp. 2742-2756, 2008.

[24] A. Strehl and J. Ghosh, "Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions," *J. Machine Learning Research,* vol. 3, pp. 583-617, 2002.

[25] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," *Proc. 21st Int'l Conf. Machine Learning,* 2004.

[26] A.L.N. Fred and A.K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 835-850, June 2005.

[27] L.I. Kuncheva and D. Vetrov, "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 28, no. 11, pp. 1798-1808, Nov. 2006.

[28] J.G. Dy and C.E. Brodley, "Feature Selection for Unsupervised Learning," *J. Machine Learning Research,* vol. 5, pp. 845-889, http://dblp.uni-trier.de/db/journals/jmlr/jmlr5.html#DyB04, 2004.

[29] S. Zhang and H.-S. Wong, "Arimp: A Generalized Adjusted Rand Index for Cluster Ensembles," *Proc. 20th Int'l Conf. Pattern Recognition (ICPR '10),* 2010.

[30] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, pp. 193-218, 1985.

[31] N. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" *Proc. 26th Ann. Int'l Conf. Machine Learning,* 2009.

[32] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Statistical Assoc.,* vol. 66, pp. 846-850, 1971.

[33] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroed, "Constrained k-Means Clustering with Background Knowledge," *Proc. 18th Int'l Conf. Machine Learning,* 2001.

[34] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance Metric Learning, with Application to Clustering with Side-Information," *Proc. Advances in Neural Information Processing Systems 15,* pp. 505-512, 2003.

[35] R.J.G.B. Campello, "A Fuzzy Extension of the Rand Index and Other Related Indexes for Clustering and Classification Assessment," *Pattern Recognition Letters,* vol. 28, no. 7, pp. 833-841, 2007.

[36] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den 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," *Nature Genetics,* vol. 30, no. 1, pp. 41-47, 2002.

[37] M. De Souto, I. Costa, D. De Araujo, T. Ludermir, and A. Schliep, "Clustering Cancer Gene Expression Data: A Comparative Study," *BMC Bioinformatics,* vol. 9, no. 1, article 497, 2008.

[38] C. Nutt et al., "Gene Expression-Based Classification of Malignant Gliomas Correlates Better with Survival Than Histological Classification," *Cancer Research,* vol. 63, no. 7, pp. 1602-1607, 2003.

[39] S. Tomlins et al., "Integrative Molecular Concept Modeling of Prostate Cancer Progression," *Nature Genetics,* vol. 39, no. 1, pp. 41-51, 2006.

[40] M. Bredel, C. Bredel, D. Juric, G. Harsh, H. Vogel, L. Recht, and B. Sikic, "Functional Network Analysis Reveals Extended Gliomagenesis Pathway Maps and Three Novel Myc-Interacting Genes in Human Gliomas," *Cancer Research,* vol. 65, no. 19, pp. 8679-8689, 2005.

[41] A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[42] J. Aslam and M. Montague, "Models for Metasearch," *Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 276-284, 2001.