Multi-Source Feature Selection via Geometry-Dependent Covariance Analysis

Zheng Zhao and Huan Liu

Department of Computer Science and Engineering, Arizona State University

Abstract. Feature selection is an effective approach to reducing dimensionality by selecting relevant original features. In this work, we studied a novel problem of *multi-source feature selection* for unlabeled data: given multiple heterogeneous data sources (or data sets), select features from one source of interest by integrating information from various data sources. In essence, we investigate how we can employ the information contained in multiple data sources to effectively derive intrinsic relationships that can help select more meaningful (or domain relevant) features. We studied how to adjust the covariance matrix of a data set using the geometric structure obtained from multiple data sources, and how to select features of the target source using geometry-dependent covariance. We designed and conducted experiments to systematically compare the proposed approach with representative methods in our attempt to solve the novel problem of multi-source feature selection. The empirical study demonstrated the efficacy and potential of multi-source feature selection.

1 Introduction

Much progress has been made over the last decade in developing effective feature selection algorithms [1]. Many feature selection algorithms have been developed and proven to be effective in handling data of single source. One area in which feature selection is intensively used is bioinformatics where high-throughput techniques generate data (e.g., genomics and proteomics) with thousands of dimensions but only hundreds of samples. Recent development in bioinformatics has made various data sources available. For example, recent work has revealed the existence of a class of small non-coding RNA species in addition to messenger RNA, known as microRNAs (miRNAs), which have critical function cross various biological processes. The new availability of multiple data sources presents unprecedented opportunities to advance research enabling us solve previously unsolvable problems. This is because each data source has singleton strengths that may help find intrinsic relationships that can be found by using multi-source data. For instance, the miRNA profiles are surprisingly informative, reflecting the developmental lineage and differentiation state of the tumors [2]. In their work, Lu, et al. successfully classified poorly differentiated tumors using miRNA expression profiles, whereas messenger RNA profiles were highly inaccurate. The manifestation of their work is that judiciously using additional data sources can help achieve better learning performance.

In this work, we studied a novel problem of selecting features from a data set of interest with additional data sources. It is an unsupervised feature selection problem arising from a study of bioinformatical cancer research in which cancerous samples are collected with both messenger RNA (mRNA) and miRNA (i.e., two data sources), and we need to find pertinent genes from the mRNA data (or the target data) for biological investigation. Since late 90's, the demand for unsupervised feature selection increases as data evolve with the rapid growth of computer generated data, text/Web data, and high-throughput data in genomics and proteomics [3]. Many unsupervised feature selection algorithms have been developed [4–6]. Most data collected are without class labels since labeling data can incur huge costs. Without any label information, one idea for unsupervised feature selection is to find features that can promote the data separability. The goal of unsupervised feature selection can be defined as finding the smallest feature subset that best uncovers "interesting natural" clusters from data according a chosen criterion [7]. Although various approaches to unsupervised feature selection have been proposed, to the best of our knowledge, the potential of using additional data sources in unsupervised feature selection has not been explored. This work presented our approach to addressing the novel problem of unsupervised feature selection with multi-source data. Given multiple data sources, a global geometric pattern can be extracted to reflect the intrinsic relationships among instances [8]. We propose to use the obtained global geometric pattern in covariance analysis for multi-source feature selection.

2 Notations and Definitions

Assume we have H data sources, $\mathcal{D}_1$,..., $\mathcal{D}_H$ jointly depicting the same set of N objects $\mathbf{o}_1, \ldots, \mathbf{o}_N$. In a data source $\mathcal{D}$, the N objects are represented by N instances as $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, $\mathbf{x}_i \in R^M$. We use $F_1, F_2, \ldots, F_M$ to denote the M features, and $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ are the corresponding feature vectors[1]. In this work, we utilize spectral graph theory [9] as a tool to represent and explore the geometric structure of a data. Let $\mathbb{G}$ denote the graph represents relationships among instances, its *similarity matrix* is denoted by $W(i,j) = w_{ij}$. Let $\mathbf{d}$ denote the vector: $\mathbf{d} = \{d_1,\ d_2,...,\ d_N\}$, where $d_i = \sum_{k=1}^{N} w_{ik}$, the *degree matrix* D of graph $\mathbb{G}$ is defined by: $D(i,j) = d_i$ if $i = j$, and 0 otherwise. Given D and W, the *Laplacian matrix* L and *normalized Laplacian matrix* $\mathcal{L}$ are defined as:

$$L = D - W; \quad \mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{1}$$

It is well known that the leading eigenvectors of L forms the soft cluster indicators of the data [10]. Also in this work we denote $\mathbf{S}_+$, the space of symmetric positive semidefinite matrices; K, the kernel matrix; $\mathbf{C}$, the covariance matrix, $\mathbf{1}$, the vector with all its elements equal to 1 and I, the identity matrix.

[1] We do not assume all data sources have this feature-instance representation. Representations can be heterogenous. For example a data source may only provide instance similarity via kernel matrix.

3 Multi-Source Feature Selection

Given H data sources, $\mathcal{D}_1, \ldots, \mathcal{D}_H$ that jointly depict a set of N objects, let $\mathcal{D}_t$ $(1 \leq t \leq H)$ be the target source for feature selection. The task of Multi-Source Feature Selection is to identify relevant features from $\mathcal{D}_t$ according to the information extracted from the H data sources[2]. With multiple data sources, the key issue is how to integrate information for selecting features. An intuitive way for information integration is to learn a global geometric pattern from all sources to reflect the intrinsic relationships among instances [8]. With a global geometric pattern, the issue now becomes how to use it effectively for feature selection. Below we introduce the concept of geometry-dependent covariance, which enables us to use the global geometric pattern in covariance analysis to select features. Based on geometry-dependent covariance we propose GDCOV, a framework for multi-source feature selection.

3.1 Geometry Dependent Covariance

Let $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be a data set containing n instances. The sample covariance matrix of X is given by:

$$\mathbf{C} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\mathbf{x}_k - \bar{\mathbf{x}}\right)\left(\mathbf{x}_k - \bar{\mathbf{x}}\right)^T, \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k. \tag{2}$$

$\mathbf{C}$ is an unbiased estimator of the covariance matrix. It captures the correlation between all possible feature pairs, and is symmetric positive semidefinite. Let $\mathbf{C}_{i,j}$ be the ij-th element of $\mathbf{C}$. It measures the covariance between features F_i and F_j, and is calculated by:

$$\mathbf{C}_{i,j} = \frac{1}{n-1}\left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1}\right)^T\left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1}\right), \quad \bar{f}_i = \sum_{k=1}^{n} f_{ik}. \tag{3}$$

In this work we propose to adjust the covariance measure between features according to the geometric structure of the objects. We give the definition of geometry-dependent sample covariance.

Definition 1. *Geometry-Dependent Sample Covariance with L. Given $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, the feature vectors of features F_i and F_j, and L, a Laplacian matrix, the Geometry-Dependent Sample Covariance between F_i and F_j is given by:*

$$\widehat{\mathbf{C}}_{i,j} = \frac{1}{n-1}\left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1}\right)^T \Gamma(L)\left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1}\right), \tag{4}$$

where $\Gamma : \mathbf{S}_+ \rightarrow \mathbf{S}_+$ is a prescribed spectral matrix function induced from an non-increasing real function of positive input, $\gamma : (0, \infty) \rightarrow (0, \infty)$, $\gamma(0) = 0$.

[2] We allow feature selection on more than one sources, by specifying multiple targets.

Let $A \in \mathbf{S}_+$, $A = U \Sigma U^T$ be the singular value decomposition (SVD) [11] of A, where $U^T U = I$ and $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\Gamma(A)$ is given by:

$$\Gamma(A) = U \hat{\Sigma} U^T, \hat{\Sigma} = \operatorname{diag}(\gamma(\lambda_1), ..., \gamma(\lambda_n)) \tag{5}$$

An example of such functions is: $\gamma(\lambda) = 0$ if $\lambda = 0$, and $\frac{1}{\lambda}$ otherwise. This gives $\Gamma(L) = L^+$. And $\widehat{\mathbf{C}}_{i,j} = \frac{1}{n-1} \left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1} \right)^T L^+ \left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1} \right)$. Computing $\mathbf{C}_{i,j}$, the sample covariance, requires centralizing each feature vector to have zero mean. However we can show that the step is redundant for calculating, $\widehat{\mathbf{C}}_{i,j}$, the geometry-dependant sample covariance.

Theorem 1. *For $\forall \boldsymbol{f}_i, \boldsymbol{f}_j \in \mathbb{R}^n$, we have the following equation:*

$$\widehat{\mathbf{C}}_{i,j} = \frac{1}{n-1} \left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1} \right)^T \Gamma(L) \left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1} \right) = \frac{1}{n-1} \boldsymbol{f}_i^T \Gamma(L) \boldsymbol{f}_j \tag{6}$$

Proof: Let $L = U_1 \Sigma_1 U_1^T$ be the truncated SVD of L. Since $\gamma(0) = 0$, to prove the theorem it is sufficient to show that $U_1^T \mathbf{1} = \mathbf{0}$. Because of $L\mathbf{1} = 0$, we have $\mathbf{1}^T U_1 \Sigma_1 U_1^T \mathbf{1} = 0$, which means $\Sigma_1^{\frac{1}{2}} U_1^T \mathbf{1} = 0$. Since the diagonal elements of Σ_1 are all larger than zero, we have $U_1^T \mathbf{1} = 0$. ■

In Equations (4) and (6), $\boldsymbol{f}$ and L are used to calculate the covariance measure. However the scale of the two factors can affect the measure arbitrarily. An effective way to handle this issue is to apply normalization on both $\boldsymbol{f}$ and L. Let

$$\widetilde{\boldsymbol{f}} = \left\| D^{\frac{1}{2}} \boldsymbol{f} \right\|^{-1} \cdot D^{\frac{1}{2}} \boldsymbol{f} \tag{7}$$

be the normalized feature vector [3] of $\boldsymbol{f}$. Based on Equation (6) of Theorem 1, we can define a geometry-dependent covariance measure using the normalized Laplacian matrix $\mathcal{L}$:

Definition 2. *Geometry-Dependent Sample Covariance with $\mathcal{L}$. Given $\widetilde{\boldsymbol{f}}_i$ and $\widetilde{\boldsymbol{f}}_j$, the normalized feature vectors, and $\mathcal{L}$, the normalized Laplacian matrix, the Geometry-Dependent Sample Covariance between F_i and F_j is given by:*

$$\widetilde{\mathbf{C}}_{i,j} = \frac{1}{n-1} \widetilde{\boldsymbol{f}}_i^T \Gamma(\mathcal{L}) \widetilde{\boldsymbol{f}}_j, \tag{8}$$

For geometry-dependent covariance matrices $\hat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$, we have the following theorems hold:

Theorem 2. *Assume all features have unit norm, let $W = \frac{1}{n} \cdot \mathbf{1}\mathbf{1}^T$, we have $\widehat{\mathbf{C}} = \widetilde{\mathbf{C}} = \gamma(1) \cdot \mathbf{C}$, where $\mathbf{C}$ is the standard covariance matrix.*

[3] The form of normalization is commonly used in spectral dimension reduction, e.g., Laplacian Eigenmap [12]. Basically, we adjust the feature vector according to data density before normalize it.

Theorem 3. $\hat{\mathbf{C}}$ *and* $\widetilde{\mathbf{C}}$ *are both symmetric and positive semidefinite.*

The proof of Theorems 2 and 3 are straightforward, we therefore ignore them due to the space limit. Theorem 2 says that by setting W in a special form, the geometry-dependent covariance matrix is equivalent to a scaled standard covariance matrix. Theorem 3 tells that since any symmetric positive semidefinite matrix is a valid covariance matrix, $\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$ are valid as covariance matrices.

Intuition of Geometry-Dependent Covariance Given two feature vectors, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, the geometry-dependent covariance measures their correlation by considering geometric structure captured by the Laplacian matrix L. To see this, letting $L = U\Sigma U^T$ be the SVD of L, we can decompose $(n-1)\hat{\mathbf{C}}_{i,j}$ as $\left(\left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1}\right)^T U\Gamma\left(\Sigma\right)^{\frac{1}{2}}\right)\left(\Gamma\left(\Sigma\right)^{\frac{1}{2}} U^T \left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1}\right)\right)$. Instead of calculating the covariance by directly applying inner product on $\left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1}\right)$ and $\left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1}\right)$, geometry-dependent covariance first projects the two vectors into the space spanned by the eigenvectors of L, then reweight the transformed vectors using the eigenvalues of the corresponding coordinates. According to spectral cluster theory [10, 13], the eigenvectors of L are the soft cluster indicators and the corresponding eigenvalues indicate their consistency with the geometric structure of the data. Let $U = (\mathbf{u}_1, \ldots, \mathbf{u_n})$, $a_i = \mathbf{u}_i^T\left(\boldsymbol{f} - \bar{f} \cdot \mathbf{1}\right)$. We have:

$$\left(\Gamma\left(\Sigma\right)^{\frac{1}{2}} U^T \left(\boldsymbol{f} - \bar{f} \cdot \mathbf{1}\right)\right) = \left(a_1\gamma(\lambda_1)^{\frac{1}{2}}, \ldots, a_n\gamma(\lambda_n)^{\frac{1}{2}}\right)^T \tag{9}$$

Equation (9) answers two questions: 1) Why should $\gamma(\cdot)$ be an non-increasing function of positive input? and 2) why is the covariance measure geometry dependent? First, the eigenvalues of L are related to the cut values associated with the corresponding eigenvectors, and measure the consistency of eigenvectors with the structure of the graph. The smaller the value, the more consistent the eigenvector. If we use these eigenvectors as the coordinates for projection, a non-increasing function $\gamma(\cdot)$ ensures to assign larger weights to the coordinates, which are more consistent with the geometric structure of the data. The reweightting effect of $\gamma(\cdot)$ helps answer the second question. Assume the centralized feature vectors have been normalized to have unit norm. A non-increasing $\gamma(\cdot)$ ensures to assign small weights to eigenvectors which are inconsistent to the given geometric pattern. If a feature vector $\boldsymbol{f}$ only aligns to inconsistent eigenvectors, the reweightting ensures that all elements in the weighted vector be small. Thus, in calculating covariance, the inner product between it and another weighted vector will be small, even if the two vectors align well. Similar analysis holds for the geometry-dependent covariance defined with $\mathcal{L}$. Note in Equation (9), all eigen-pairs are used to evaluate features, however if $\gamma(\cdot)$ returns zero on the bottom eigenvalues, we can also achieve the effect of removing the bottom eigen-pairs from consideration, which is a mechanism commonly used in spectral dimension reduction and clustering for denoising [12].

Efficient Calculation of Geometry-Dependent Covariance The calculation of geometry-dependent covariance involves a spectral matrix function induced from a non-increasing real function, which may require a full eigen decomposition on L or $\mathcal{L}$ that has a time complexity of $O(N^3)$, where N is the number of instances. We show that for the geometry-dependant covariance defined with $\mathcal{L}$, assuming W is positive semidefinite[4], we are able to obtain $\Gamma(\mathcal{L})$ in $O(N^2)$ with a special definition of $\gamma(\cdot)$.

Theorem 4. *Given $\gamma(\cdot)$ defined as: $\gamma(\lambda) = 0$, if $\lambda = 0$, and $1 - \lambda$, otherwise. We have: $\Gamma(\mathcal{L}) = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} - \left(D^{\frac{1}{2}} 1 1^T D^{\frac{1}{2}}\right) / \left(1^T D 1\right)$ Under this $\gamma(\cdot)$,*

$$\widetilde{\mathbf{C}} = \frac{1}{n-1} \, \Pi X \left(W - \frac{W \mathbf{1} \mathbf{1}^T W}{\mathbf{1}^T W \mathbf{1}} \right) X^T \Pi, \tag{10}$$

where Π is the diagonal matrix with $\Pi_{i,i} = \left\| D^{\frac{1}{2}} \boldsymbol{f}_i \right\|^{-1}$

3.2 Feature Selection with Multiple Data Sources

Given multiple data sources jointly depicting a set of objects, we can extract a global geometric pattern that reflects the intrinsic relationships among instances. The obtained global geometric pattern can be then used in the proposed geometry-dependent covariance for feature selection.

Global Geometric Pattern Extraction Given multiple local geometric patterns, a global pattern can be obtained by linearly combining local patterns [8, 14] as follows.

$$W_{global} = \sum_{i=1}^{H} \alpha_i W_i \tag{11}$$

In the equation, W_is are the geometric patterns extracted from each data source, and α_is are the combination coefficients, which can be learnt automatically from data in either supervised [8] or unsupervised [15, 16] ways. The combination coefficients can also be assigned by domain experts according to their domain knowledge[5] [14]. We refer readers to literature for comprehensive study on the research issues of kernel combination.

Feature Selection with Global Geometric Pattern Given multiple heterogenous data sources, we now use the global geometric pattern extracted from all data sources to select features for the target data source. With the global geometric pattern obtained from multiple data sources, one can build a geometry-dependent sample covariance matrix based on which features can be selected using two methods: 1) GPCOV$_{var}$: sorting the diagonal of the covariance matrix and choosing the features that have the biggest variances; 2) GPCOV$_{spca}$: applying sparse principle component analysis (SPCA) to select a set of features

[4] This ensures the eigenvalues of $\mathcal{L}$ is bounded by 1 from above, so that $\gamma(\cdot)$ is valid.

[5] This provides a way to incorporate domain knowledge

that can maximally retain the total variance. As discussed in the intuition of the geometry-dependent covariance, for $\widetilde{\mathbf{C}}_{i,i}$ to achieve a big value, $\widetilde{\boldsymbol{f}}_i$ must align well to the eigenvectors that are consistent with the geometric structure depicted by $\mathcal{L}$. In other words, a bigger $\widetilde{\mathbf{C}}_{i,i}$ indicates $\widetilde{\boldsymbol{f}}_i$ is more consistent with the global geometry pattern. Therefore selecting features according to the first method is equivalent to selecting features which are consistent to the structure of the global geometry pattern. Similar analysis holds for $\widehat{\mathbf{C}}$, assuming all features have a similar scale. Since the first method measures feature relevance individually for each feature, it may select a feature set with redundant features. The second method, applying sparse principle component analysis such as the one proposed in [17], considers the interacting effects among features, and is able to select a feature set containing less redundancy.

3.3 GDCOV - the Framework

The proposed framework for multi-source feature selection is based on analysis of geometry-dependent covariance, and is realized in Algorithm 1. We give a time complexity analysis for the proposed multi-source feature selection framework. Since the H data sources are of heterogeneous representation, the time complexity of constructing the local pattern on each data source can vary greatly. Assuming each data sources has M features to depict the N objects, and we use RBF kernel to record the local geometric pattern. Then the time complexity of the first step is $O(HMN^2)$. Assume we linearly combine W_is with a set of prescribed combination coefficients, the cost is of the second step is $O(HN^2)$. Using the method specified in Theorem 4 to form the geometry-dependent covariance matrix, the cost is $O(MN^2 + M^2)$. Then selecting features based on the covariance matrix requires $O(M \log M)$, if we use features variance to select features; or $O(M^3)$, if we use the SPCA approach proposed in [17]. Hence the overall time complexity of GDCOV with the above specification is $O(M^2 + HMN^2)$ if we use feature variance to select features, otherwise it is $O(M^3 + HMN^2)$.

Algorithm 1: GDCOV - a framework for multi-source feature selection with geometry-dependent covariance

```
Input: $\mathcal{D}_1, \ldots, \mathcal{D}_H,\ X,\ \gamma(\cdot)$
Output: $SF$ - the selected features list
for each $\mathcal{D}_i$ do
    Construct $W_i$, the local geometric pattern;
end
Obtain global pattern $W$ from $W_1, \ldots, W_H$;
Form the geometry-dependent covariance matrix $\mathbf{C}$ using the global pattern $W$;
Select features according to $\mathbf{C}$ and form $SF$;
Return $SF$;
```

4 Experimental Study

We empirically evaluate the performance of GDCOV and compare multi-source with single source feature selection on a biological data set: miRNA-mRNA. In our experiments, we found that $\widetilde{\mathbf{C}}$ provides more robust performance than $\widehat{\mathbf{C}}$, i.e., the normalization step does lead to better performance. The results we presented and analyzed in this section are all produced from $\widetilde{\mathbf{C}}$, the geometry-dependent sample covariance matrix defined with $\mathcal{L}$. In order to objectively and systematically evaluate the performance of GDCOV, we use accuracy as well as the feature biological relevance as performance measures. Recall that our algorithm is for unsupervised feature selection using multi-source data, and no class label information is used in calculation of geometry-dependent covariance. For miRNA-mRNA Data, the biological relevance of genes is evaluated by checking whether the genes are cancer related by using gene function annotation information from the Ingenuity Pathways Analysis (IPA) system [18].

4.1 Experiment Setup

In the experiment, we compare multi-source feature selection with single source feature selection. For feature selection using single data source we choose 2 unsupervised algorithms as baseline: SEPER [19] and pathSPCA [17]. To obtain local geometric patterns, we build RBF kernel for local data sources. To extract the global geometric pattern, we use two unsupervised kernel learning algorithms: A-KPCA [15] and NAML [16]. In GDCOV, we use the real function defined in Theorem 4 as $\gamma(\cdot)$. We apply one-nearest-neighbor classifier with selected features, and use its accuracy to measure feature quality. All reported results are based on averaging the accuracy from 10 trials of experiments.

4.2 Data Sets

miRNA-mRNA Data The data set consists of two sets of gene expression profiles from a mixture of **88** normal and cancerous tissue samples: a miRNA expression profile for **151** human miRNAs and a mRNA expression profiles for **16,063** human mRNAs (**the target**) [2, 20]. The 11 involved tissues are: Colon, Pancreas, Kidney, Bladder, Prostate, Ovary, Uterus, Lung, Mesothelioma, Melanoma and Breast. Among the 11 different tissues, 4 of them, Mesothelioma, Uterus, Colon and Pancreas, have at least 7 cancerous samples, and totally contribute **33** cancerous samples to the whole data. For each sample we have both miRNA and mRNA expression profiles. It is observed in [2] that comparing with mRNA, miRNA profiles is of more power for discriminating cancer from noncancer samples as well as cancerous samples of different types of tissues.

4.3 Results on miRNA-mRNA Data

For miRNA-mRNA data, we run each feature selection algorithm to obtained a ranked gene list and evaluate the quality of the top genes in each list via checking their power on distinguishing cancer and noncancer samples as well as the 33 cancerous samples from 4 different types of tissues. We also evaluate the biological relevance of the top genes in each list by checking how many of them are cancer related. In the experiment, A-KPCA returns a combination coefficient of (0.447, 0.553), that is miRNA: 0.447 and mRNA: 0.553. And NAML returns a combination coefficient of (0.005, 0.995). We also tried three prescribed combination coefficients to linearly combine the two local geometric patterns: (1, 0), (0.7, 0.3), and (0, 1). The first and third coefficients correspond to using only miRNA and mRNA data, respectively. The second one corresponds to use both miRNA (0.7) and mRNA (0.3). We assign miRNA data more weight, since exiting findings in the literature suggest that miRNA has better discriminative power. Here multi-source feature selection corresponds to the cases of using GDCOV with miRNA, mi&mRNA, A-KPCA and NAML. GDCOV+miRNA is multi-source feature selection, since we learn geometric pattern from miRNA profile, while select features for mRNA profile.

Cancer vs. Noncancer Table 1 compares two unsupervised baseline feature selection algorithms using mRNA profiles (the target data source) with GDCOV using miRNA profiles, mRNA profiles and both profiles to select features on mRNA profiles. From the table we can see that by using miRNA profiles or both profiles, GDCOV select features that provide better accuracy. The observation suggests that, (1) the geometric pattern obtained from miRNA profiles indeed possesses better discriminative power, which is consistent with the findings in [2]. (2) By combining multiple data sources we are able to achieve better performance than using any individual data source. This is consistent with the observations in [8]. And (3) given a geometric pattern with higher quality, GDCOV selects better features. This supports the use of multiple data sources in feature selection, and shows GDCOV is effective. We also notice that using only mRNA profiles, GDCOV outperforms SPCA and is comparable with SEPER.

Algorithm	2	5	10	15	20	30	50	Ave
SPCA	0.31	0.31	0.38	0.41	0.66	0.60	0.64	0.47
SEPER	0.62	0.70	0.77	0.83	0.84	0.81	0.86	0.77
GDCOV	0.68	**0.78**	0.72	0.72	0.74	0.78	0.73	0.73
mRNA	0.68	**0.78**	0.72	0.78	0.75	0.77	0.69	0.74
GDCOV	0.67	0.74	0.79	0.90	0.82	0.86	0.93	0.82
miRNA	**0.69**	0.64	0.73	0.84	0.89	0.84	0.85	0.78
GDCOV	0.67	0.73	**0.88**	0.86	0.88	0.86	**0.94**	**0.83**
mi&mRNA	**0.69**	0.68	0.72	0.89	0.88	0.85	0.85	0.79
GDCOV	0.62	0.68	**0.88**	**0.91**	**0.91**	**0.90**	0.93	**0.83**
A-KPCA	**0.69**	0.59	0.72	0.89	0.88	0.84	0.88	0.78
GDCOV	0.68	**0.78**	0.70	0.76	0.76	0.77	0.78	0.75
NAML	0.68	**0.78**	0.71	0.84	0.80	0.74	0.73	0.75

Table 1. Cancer vs. Non-Cancer: accuracy achieved by algorithms. **mi&mRNA** stands for combining with the prescribed coefficient (0.7, 0.3). Numbers with boldface indicate the highest number achieved in each case. **For GDCOV, the first and the second row stand for GDCOV$_{var}$ and GDCOV$_{SPCA}$ respectively**.

Four Different Types of Cancer Table 2 contains the results of using genes selected by each algorithm to distinguish cancerous samples from 4 different types of tissues. Since the number of instances becomes fewer, while the number of classes becomes larger, we observe that the performance of the algorithms using single data source degenerate significantly. However, on the other hand, we also observe that by using multiple data sources GDCOV's performance is consistently good. This indicates that (1) GDCOV is able to select good features according to the intrinsic global pattern. (2) The geometric patterns obtained from miRNA profile as well as both profiles are relatively stable - genes selected by GDCOV are discriminative for identifying cancer and noncancer samples as well as cancerous samples from different tissues, indicating that they may be consistent with the intrinsic structure of the underlying model.

Algorithm	2	5	10	15	20	30	50	Ave
SPCA	0.22	0.25	0.26	0.26	0.34	0.36	0.45	0.31
SEPER	0.15	0.21	0.39	0.39	0.42	0.65	0.69	0.41
GDCOV	0.24	0.15	0.25	0.21	0.15	0.09	0.18	0.18
mRNA	0.24	0.15	0.21	0.30	0.09	0.19	0.24	0.20
GDCOV	**0.65**	0.79	0.82	0.82	0.85	0.88	**0.94**	0.82
miRNA	0.61	0.72	0.76	0.84	0.81	0.84	0.84	0.78
GDCOV	**0.65**	**0.82**	0.88	0.85	0.85	0.85	0.91	**0.83**
mi&mRNA	0.61	0.76	0.82	0.84	**0.88**	**0.91**	0.81	0.81
GDCOV	0.49	0.79	**0.91**	**0.88**	**0.88**	0.82	0.91	0.81
A-KPCA	0.61	0.74	0.82	0.84	0.79	0.85	**0.94**	0.80
GDCOV	0.24	0.15	0.42	0.24	0.15	0.31	0.27	0.25
NAML	0.24	0.15	0.27	0.15	0.18	0.27	0.23	0.21

Table 2. Four Different Types of Cancer: accuracy achieved by algorithms. **mi&mRNA** stands for combining with the prescribed coefficient (0.7, 0.3). Numbers with boldface indicate the highest number achieved in each case. The advantage of using multiple data sources are significant in this case. **For GDCOV, the first and the second row stands for GDCOV$_{var}$ and GDCOV$_{SPCA}$ respectively**.

Although GDCOV+NAML uses multiple data sources for feature selection, we can observe from Table 2 that it does not perform well. This is understandable, since NAML assigns too few weight to miRNA profile and make GDCOV+NAML almost equivalent to single source feature selection.

Study of Biological Relevance To evaluate the biological relevance of the top genes in each list we check how many of them are cancer related. The results can be found in Table 3. From the table we can observe that by using multiple sources, averagely, GDCOV selects the most cancer related genes. Interestingly, although SPCA selects many cancer related genes, accuracy results suggest that these genes are actually of poor discriminative power, which says that these genes may not be really relevant to the underlying process. On the other hand, the genes selected by GDCOV using multiple data sources are cancer related and of strong discriminative power as well. This suggests that these genes are possible to be relevant to the underlying process from which the studied samples are generated. In order to closely examine the biological relevance of the selected genes, we performed a further study in which our biologist collaborators examined the top 10 genes selected by GDCOV$_{var}$+mi&mRNA. It turned out that all 10 genes are actually cancer related. Detailed information of the selected genes are shown in Table 4. Among them, seven were already annotated to be related to cancer in the IPA system. The other genes are found to be differentially expressed in cancer cell lines and are supported by literature.

Algorithm	2	5	10	15	20	30	50	Ave
SPCA	1	**4**	5	7	9	11	16	7.57
SEPER	0	1	3	4	4	6	8	3.71
GDCOV	0	0	2	3	3	6	8	3.14
mRNA	0	0	2	3	4	6	8	3.29
GDCOV	**2**	**4**	5	8	**10**	14	**22**	9.29
miRNA	**2**	3	5	7	**10**	13	21	8.71
GDCOV	**2**	3	**7**	**9**	**10**	**15**	21	**9.57**
mi&mRNA	**2**	**4**	5	6	7	14	21	8.43
GDCOV	1	3	**7**	**9**	**10**	14	18	8.86
A-KPCA	**2**	2	5	6	9	14	**22**	8.57
GDCOV	0	0	2	4	5	6	11	4.00
NAML	0	0	2	5	5	6	12	4.29

Table 3. Numbers of known cancer related genes according to literature in the top 2, 5, 10, 15, 20, 30 and 50 genes provided by algorithms. Numbers with boldface indicate the highest number achieved in each case. **For GDCOV, the first row stands for GDCOV$_{var}$ and the second row stands for GDCOV$_{SPCA}$.**

Gene Name	Functions and Biological Process	Disease
LGALS4	sugar binding, cell adhesion	colon cancer
LTF	ubiquitin ligase complex	prostate cancer
FUT6	integral to membrane,L-fucose catabolism	colon cancer
FABP1	fatty acid metabolism,GABA-A receptor	prostate cancer
GPX2	oxidoreductase,response to oxidative stress	breast cancer
CNN1	calmodulin binding,actin filament binding	melanoma
CDH17	transporter activity, calcium ion binding	colon cancer
TM4SF3	signal transducer activity	esophageal cancer
MYH11	calmodulin binding,actin binding	lung cancer
KRT15	structural constituent of cytoskeleton	breast cancer

Table 4. The top 10 genes selected by GDCOV+mi&mRNA. Genes with boldface names are the ones directly detected by the IPA system as cancer related. In the table genes are ordered according to their relevance scores from highest to lowest.

5 Conclusions

In this work, we investigated a novel problem arising from the need to select features on one data source given multiple sources but without class labels. We first proposed the concept of geometry-dependent covariance, we studied its properties and showed how to employ this covariance measure for multi-source feature selection. We designed and conducted extensive experiments to objectively and systematically evaluate the proposed approaches in comparison with existing representative single source feature selection methods. The affirmative results demonstrate that using multi-source data can help improve feature selection of the target source. As multi-source data become more common, learning and feature selection using multi-source data will be in high demand in many real applications. We can show that the approaches proposed in [6] are actually special cases of the framework proposed in this paper, we are studying the general properties of the algorithms generated from this framework. Besides feature selection, it is also possible to use geometry-dependent covariance in discriminant analysis and regression, which forms one line of our ongoing work.

6 Acknowledgments

This work is in part by a grant from NSF.

References

1. Liu, H., Motoda, H., eds.: Computational Methods of Feature Selection. Chapman and Hall/CRC Press (2008)
2. Lu, J., etal.: Microrna expression profiles classify human cancers. Nature **435**
3. Smyth, P., etal.: Data-driven evolution of data mining algorithms. Communications of the Association for Computing Machinery **45**(8) (August 2002) 33 – 37
4. He, X., etal.: Laplacian score for feature selection. In: NIPS 18. (2005)
5. Varshavsky, R., Gottlieb, A., Linial, M., Horn, D.: Novel unsupervised feature filtering of biological data. Bioinformatics **22** (2006) e507–e513
6. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: ICML. (2007)
7. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. J. Mach. Learn. Res. **5** (2004) 845–889
8. Lanckriet, G.R.G., etal.: Learning the kernel matrix with semidefinite programming. J. Mach. Learn. Res. **5** (2004) 27–72
9. Chung, F.: Spectral graph theory. AMS (1997)
10. von Luxburg, U.: A tutorial on spectral clustering. Technical report, Max Planck Institute for Biological Cybernetics (2007)
11. Golub, G.H., Van Loan, C.F.: Matrix Computations. Third edn. The Johns Hopkins University Press (1996)
12. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. NIPS **15** (2003)
13. Yu, S.X., Shi, J.B.: Multiclass spectral clustering. In: ICCV. (2003)
14. Zhou, D., Burges, C.: Spectral clustering and transductive learning with multiple views. In: ICML. (2007)
15. Zhang, D., Zhou, Z.H., Chen, S.: Adaptive kernel principal component analysis with unsupervised learning of kernels. In: ICDM. (2006)
16. Chen, J., Zhao, Z., Ye, J., Liu, H.: Nonlinear adaptive distance metric learning for clustering. In: SIGKDD. (2007)
17. d'Aspremont, A., Bach, F., Ghaoui, L.E.: Optimal solutions for sparse principal component analysis. Technical report, Princeton University (2007)
18. Ingenuity-Systems: Ingenuity pathways analysis. http://www.ingenuity.com
19. Dash, M., Choi, K., Scheuermann, P., Liu, H.: Feature selection for clustering – a filter solution. In: ICDM. (2002)
20. Huang, J.C., etal.: Using expression profiling data to identify human microrna targets. NATURE METHODS **4** (2007) 1045–1049