Bias Selection Using Task-Targeted Random Subspaces for Robust Application of Graph-Based Semi-Supervised Learning

Christopher T. Symons, Ranga Raju Vatsavai
Computational Sciences and Engineering
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
Email: {symonsct, vatsavairr}@ornl.gov

Goo Jun
Center for Statistical Genetics
University of Michigan
Ann Arbor, Michigan 48109
Email: gjun@umich.edu

Itamar Arel
Electrical and Computer Engineering
University of Tennessee
Knoxville, TN 37996
Email: itamar@ieee.org

***Abstract*—Graphs play a role in many semi-supervised learning algorithms, where unlabeled samples are used to find useful structural properties in the data. Dimensionality reduction and regularization based on preserving smoothness over a graph are common in these settings, and they perform particularly well if proximity in the original feature space closely reflects similarity in the classification problem of interest. However, many real-world problem spaces are overwhelmed by noise in the form of features that have no useful relevance to the concept that is being learned. This leads to a lack of robustness in these methods that limits their applicability to new domains. We present a graph-construction method that uses a collection of task-specific random subspaces to promote smoothness with respect to the problem of interest. Application of this method in a graph-based semi-supervised setting demonstrates improvements in both the effectiveness and robustness of the learning algorithms in noisy problem domains.**

***Keywords*-applications; graph Laplacian; semi-supervised;**

I. Introduction

The areas of major emphasis in machine learning research are continually shifting based on requirements identified through new applications. At the same time, the potential of computational learning is being recognized by a more diverse set of users with broader and more complex sets of issues to analyze. This leaves machine learning practitioners in a position of constantly adapting algorithms to the peculiarities of new domains. The more labor involved in applying algorithms to new problems, the less likely that the advantages they offer will ever be realized. Thus, the historical trend has been one in which classification algorithms that are less sophisticated, but better understood, are applied despite their limitations. In essence, the customer with an understanding of the data is unaware of the peculiarities of more advanced algorithms, so that unless new methods are made extremely robust to variations in application, they often remain a niche product used by only a few, regardless of any potential they might have to enhance our data analysis capabilities across a broad set of applications.

A prominent example of this trend can be seen in the subfield of semi-supervised learning [1]. Despite its direct relevance to some major issues that span many application spaces, we have yet to see semi-supervised learning establish a strong presence in many real systems. Successful application of semi-supervised learning requires both very specific domain knowledge of the application and detailed understanding of the assumptions underlying particular learning algorithms. In the benchmark analysis in [1], the authors lament this point, emphasizing that while generic supervised learning algorithms can perform well, application of semi-supervised algorithms requires deeper understanding of the data, such that "black box" solutions are unlikely to succeed.

Many popular forms of semi-supervised learning rely on graph-based methods of spectral dimensionality reduction. Some common forms of this type of nonlinear dimensionality reduction include Isomap [2], Locally-Linear Embedding (LLE) [3], Laplacian Eigenmaps (LEM) [4], Diffusion Maps (DM) [5], Semidefinite Embedding [6], etc. A more comprehensive list can be found in [7]. A central construct in many methods is the graph Laplacian. A graph is constructed to represent a manifold (or densely populated region of interest in the ambient space), and the graph Laplacian facilitates the discovery of a low-dimensional space that is smooth with respect to this graph. A typical approach is to use unlabeled data to find a low-dimensional space on which learning via the labels can be more effective. Manifold regularization approaches to semi-supervised learning (e.g. see [8]) use the graph Laplacian as well by penalizing model choices that would disagree with the graph.

As emphasized in [9], normalized-output algorithms, such as LLE, LEM, and DM, do not handle noise well, and should not be applied arbitrarily, and there is a need for improvements that are robust. Goldberg, et al. [10] recognize the potential problem of having data that resides on multiple manifolds and offer some methods for applying semi-supervised manifold learning in such cases. However, while the methodology adds some robustness in such multi-manifold cases, the manifold description is made without the use of the labels so that high levels of noise can still hide the manifolds. Furthermore, it is quite possible that many of these discoverable manifolds are not relevant to the target problem.

A consistent theme among most graph-based semi-supervised learning approaches is that the unlabeled data is used to discover the manifold and the labeled data is used to learn a model. We focus on the fact that these methods can be used to find functions that are smooth *with respect to the graph that is constructed*. From this perspective, it makes sense to use the labeled data to guide the graph construction process in order to avoid close connections due to noise in the feature set. However, in order to make good use of the labeled data, it is necessary to go beyond using them as constraints. It is also necessary to be cognizant of the fact that there likely will be within-class clusters, so that we must avoid strongly linking members of the same class if they are from very different subsets of that class. Doing so would violate the assumptions of local-distance preservation (whether geodesic or other) on which the methods are based. In practice, using the labels to guide the graph construction can confound the search for a relevant manifold if the labels are applied carelessly, while conservative methods of using the labels have little effect.

In this paper, we present a novel method for biasing the graph-construction process from the viewpoint of local, task-relevant measures based on discriminative learning of small random subspaces that are specific to the target problem. We evaluate the use of the graph construction method on noisy domains for both regularization and dimensionality reduction. When the only variant in the learning process is the graph construction method, we demonstrate that we can greatly improve classification accuracy and stability in noisy real-world domains.

II. Background

A. Graph-Based Semi-Supervised Learning

Semi-supervised learning [1] is generally defined as any learning method that uses both labeled and unlabeled data during the model discovery process. A broad subset of successful semi-supervised algorithms is built around graph-based methods. A succinct description of many of these methods can be found in [7]. The graph Laplacian from spectral graph theory [11] is an important device in many of these algorithms. When used for dimensionality reduction, the eigenvalues and eigenvectors of the Laplacian matrix provides a means for discovering dimensions that are smooth with respect to the graph that defines it. If the graph varies smoothly with respect to the target problem (i.e. examples from different classes or clusters are rarely linked, similar examples from the perspective of the target problem are linked, etc.), then it can be used to represent a manifold that is useful for learning a classifier. The normalized graph Laplacian is a matrix defined as follows:

$$\mathcal{L}(u,v) = \begin{cases} 1, & \text{if } u = v \text{ and } d_v \neq 0 \\ \frac{-1}{\sqrt{d_u d_v}}, & \text{if } u \text{ and } v \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where d is the degree of a vertex. However, since it works well in practice, we use the unnormalized form given below:

$$L(u,v) = \begin{cases} d_v, & \text{if } u = v \\ -1, & \text{if } u \text{ and } v \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Typically, semi-supervised learning methods based on the graph Laplacian separate the manifold discovery process from the learning process in such a way that manifold discovery is completely unsupervised. There is a general recognition that the labels can be used as constraints when building the graph, but in semi-supervised learning, such an approach would touch only a small portion of the graph. In [12], a dissimilarity measure is used to alter the graph. However, the method either requires ground truth dissimilarity, in which the only parts of the graph that are affected are those for which labels are available, or domain specific features that are manually constructed to enforce disparity.

B. Random Subspace Methods

Random subspace selection [13], [14], or attribute bagging, is a method of ensemble learning where multiple classifiers are constructed using random subsets of the original feature set. In addition to improving accuracy, these methods are particularly good for providing stability in the learning process. They are well suited for use with small labeled sets due to the natural dimensionality reduction that occurs. In other words, while many standard ensemble methods use subsets of the training data while keeping the feature set constant, these methods use subsets of the features while keeping the training data constant, thus they are less likely to be affected by the *curse of dimensionality*.

III. Methodology

In this section, we will describe our methodology after building up the intuition behind the approach. Our goal is to use the small number of labels available for semi-supervised learning to influence the edge construction in a manner that allows concept-specific structure to be encoded in the graph. We imagine that any approach to doing so can be less than optimal for extremely clean feature spaces that already meet the assumption that nearby points are in the same class. However, this assumption is rarely valid in practice, unless the feature space has already been heavily hand engineered. We are primarily concerned with finding good semi-supervised solutions for newer domains that lack a good understanding of the noise and how to eliminate it in preprocessing. Moreover, we focus on problems where labeled data is expensive and unlabeled data is abundant.

As in [15] we approach the problem from the viewpoint that there should be a notion of compatibility, χ, between the hypothesis and the data distribution (as estimated using the unlabeled examples). Graph-based manifold learning methods assume that the target concept passes through a dense

portion of the ambient space, and that this manifold can be discovered using the unlabeled data. In data sets where the feature space has been refined (e.g. image recognition data that has been preprocessed, centered, scaled, etc.), the ability to find the manifold is well-demonstrated experimentally. Unfortunately, many applications do not have such a nice, clean ambient space.

In [15], various sample complexity bounds are provided for semi-supervised learning. In some cases an assumption is made that the target function's unlabeled error rate is low (essentially zero), and in other cases the bounds depend on the unlabeled error of the true target function. A graph constructed over noisy samples is likely to have many "errors." Therefore, the first assumption is too simplistic for many real-world situations. Using unlabeled data alone, the target function's unlabeled error cannot be bound at all. In other words, the number of mistakes in the notion of compatibility itself (the graph) cannot be bound while ignoring all information concerning the target concept. The use of label information in the construction of the graph has the potential to bound this error with respect to the target, allowing bounds to be applied to the overall sample complexity.

In semi-supervised learning, the amount of labeled data is typically small. This raises several issues when trying to incorporate useful information from the labels at an early stage. For example, feature selection is made that much more difficult, and therefore, attempts at using feature selection to eliminate noise are not particularly helpful. Furthermore, applying the labels directly through the use of a classifier is not advisable, since many classes contain subclasses or clusters, such that enforcing links between points based on labels alone is unwise. For example, creating an edge between two very dissimilar points can create a graph that violates the Riemannian assumptions upon which the methods were designed. We would like to have some sense that the end result will preserve some semblance of geodesic distance.

Random subspace selection is sometimes used as a method for finding different views for ensemble learning. In such cases the hope is that the views will be independent. We are under no such inclination here, which is important, because it could be difficult to ensure independence due to label sparsity. In ensemble learning, the independence is a necessary condition to ensure improved accuracy. However, we care very little about accuracy in the traditional sense, because we are not looking to make a real classification. We want to place examples together in such a way as to provide some sense of local smoothness according to our target, and therefore, if all of the classifiers were completely accurate, we would expect to end up ignoring any in-class clusters, and thus we would still have difficulty finding a good manifold.

We need to link together examples that share subsets of features that are useful in determining a classification in our target problem. Using the classifications of models that represent *good* hypotheses from different subspaces allows us to find a more delicate similarity that represents our target concept. If two examples share feature subsets that are useful according to our labeled data, then we want them to be classified together. If they both lack good predictive features in particular subspaces, then they will often be misclassified together, which is desirable in our case.

During graph construction, we perform random feature splits to obtain task-relevant subspace classifiers that are trained on the labeled data. In essence, each trained classifier represents a subspace in which local proximity is relevant to the target problem. We then construct a nearest neighbor graph based on cosine similarity while weighting the similarity scores based on the percentage of shared classifications using the random subspaces. We then build the unnormalized graph Laplacian. Using an unweighted graph not only performs well in practice, but we also expect it to be more robust to noise.

There are several parameters here that can be chosen rather arbitrarily. While model selection can be be applied to choose these parameters, our focus in this paper is on the generic approach. Therefore, for our experiments we chose these parameters based on suggestions from the literature whenever possible.

Algorithm 1 Graph Construction

Input: k, s, data $\{(x_i, y_i)\}_{i=1}^{l}$, $\{x_i\}_{i=l+1}^{l+u}$
for $index = 1$ to s **do**
  Randomly split feature set into two equal parts
  Train linear classifier (SVM) on each split
  Classify each sample point
end for
Assign edge weights = (cosine similarity) $\times$ (percent of time classified together)
Retain k nearest neighbors

Once the graph has been constructed any graph-based learning approach can be applied. Based on recommendations from the various related literature, we set the number of nearest neighbors, k, equal to 8, and the number of feature splits, s, equal to 100.

A. Graph-Based Classifiers

1) Laplacian Eigenmaps (LEM): We are focusing on the effects of altering the graph, so we construct a simple linear classifier in the new space in the same manner as the approach in [16], in which the coefficients for the new dimensions are set by minimizing the sum of squared error on the labeled data. In other words, the weights of our new dimensions are given by the vector **a** in the following:

$$a = (EE^T)^{-1}Ec \tag{3}$$

where c is a vector representing the class labels, the entries of E are $\lambda_k v_{i,k}$, i is the index of the labeled point in the matrix, and k is the index in the new low-dimensional space; i.e. the k-th eigenvalue and eigenvector provide the mapping into the new space for point i. As recommended in [16], we set the final dimensionality, or basis size, equal to twenty percent of the size of the labeled set.

The LEM method is inherently transductive, meaning that it only creates a mapping for an unlabeled example if it was part of the set used for graph construction. For an out-of-sample extension that allows efficient application to new points, we utilize the Nystrom Formula as described in [17]. The method uses the Laplacian matrix as a data-dependent Kernel function K_D in the following formula in order to map a new point into each dimension k of the new decision space:

$$f_k(x) = \frac{\sqrt{n}}{\lambda_k} \sum_{i=1}^{n} v_{ik} K_D(x, x_i) \tag{4}$$

where n is the size of the original dataset, and (λ_k, v_k) are the k-th eigenvalue and eigenvector.

2) Laplacian Regularized Least Squares: In order to evaluate our graph modification method using a very different form of learning, we use the manifold regularization technique known as the Laplacian Regularized Least Squares (LapRLS) [8]. This algorithm uses the graph Laplacian for regularization. In our experiments, we use the unnormalized form of the graph Laplacian here as well.

The output function that is learned is the following:

$$f(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x), \tag{5}$$

where K is the $(l+u) \times (l+u)$ Gram matrix over labeled and unlabeled points, and α is the following learned coefficient vector:

$$\alpha = (JK + \gamma_A l I + \frac{\gamma_I l}{(l+u)^2} LK)^{-1} Y, \tag{6}$$

with L being the Laplacian matrix described above, I being the $(l+u) \times (l+u)$ identity matrix, J being the $(l+u) \times (l+u)$ diagonal matrix with the first l diagonal entries equal to 1 and the rest of the entries equal to 0, and Y being the $(l+u)$ label vector, $Y = [y_1, ..., y_l, 0, ..., 0]$. See [8] for details.

This method has two parameters that control the amount of regularization. For all of our experiments, we use the following settings, as suggested in [8]: $\gamma_A l = 0.005$, $\frac{\gamma_I l}{(l+u)^2} = 0.045$.

IV. Experiments

Our experiments were constructed to demonstrate the effect of altering the graph in the manner proposed on noisy real-world problems. For comparison purposes, we apply the Laplacian Eigenmap approach [16] (LEM) and the LapRLS [8], where the graph construction process is varied between experiments. We denote the approaches using the label-biased graph as LB-LEM and LB-LapRLS in the experimental results.

Table I
Average error of Laplacian-Eigenmap-based classifiers on BCI data.

Graph construction method	Average Error
Unlabeled data graph (LEM)	0.4856
Random subspace label-biased graph (LB-LEM)	**0.3086**
Top 10 feature label-biased graph	0.4596
Top 20 feature label-biased graph	0.4673
Top 30 feature label-biased graph	0.4746

A. Brain-Computer Interface

The Brain-Computer Interface (BCI) problem comes from a collection of electroencephalography (EEG) recordings using 39 probes. The goal is to determine whether the human subject was concentrating on moving their right or left hand during the monitoring process. This is an extremely noisy dataset, and one in which the use of the unlabeled data alone is very unsuccessful in discovering a good predictive space. More details can be found in [1]. Note that the BCI dataset was identified in [1] as one in which it is very difficult to improve over the supervised baseline obtained using an SVM (error: 34.31%; AUC: 71.17%).

The experiments in Table I were conducted using the method described in [16], with the only difference being in the construction of the graph. For each of the methods described, we ran 10 experiments in which 100 labeled samples were randomly selected and the remaining 300 samples were used as unlabeled data. The unlabeled data were then used as the transductive test set. In addition, in order to demonstrate that some simple methods for using the labeled data in the graph construction can have significant robustness problems, we ran a third version of experiments with a feature-selection-based graph construction process. In this approach, we select the top features (out of 117 total) as ranked using a combination of mutual information and fisher criterion, as in [18]. Table II shows a comparison of results across the 12 data splits from the benchmark set in [1], where the LapRLS performed the best out of all methods in the benchmark. We see that our simple approach can improve even these results.

B. Dimensionality Reduction in Hyperspectral Image Analysis

Land cover classification by hyperspectral image (HSI) data analysis has become an important part of remote sensing research in recent years [19]. Compared to conventional multi-spectral images where each pixel usually contains tens of bands, pixels in hyperspectral images usually consist

Table II
AVERAGE ERROR OF LAPLACIAN RLS CLASSIFIERS ON BCI DATA.

Model Building Condtions	# Labeled Data	Average Error	AUC
Unlabeled data graph (LapRLS)	100	0.3244	0.7431
Unlabeled data graph (LapRLS*)	100	0.3136	0.7483
Random subspace label-biased graph (LB-LapRLS)	100	**0.2697**	**0.8083**

*LapRLS results in [1], obtained using model selection, a normalized graph Laplacian, and an RBF base kernel; which was the best result among all 11 semi-supervised algorithms tested in the benchmark.

of more than a hundred spectral bands, providing fine-resolution spectral information. Classification techniques for this multiclass application need to handle high-dimensional, high-resolution data. Obtaining ground truth is another challenge, since HSI can cover very large areas and it is not usually possible to obtain highly accurate class labels for all locations in the image. In this domain, dimensionality reduction is very important due to the well-known Hughes effect.

A Hyperion hyperspectral image taken from Okavango Delta, Botswana in May 2001 is used for the experiments. The acquired data originally consisted of 242 bands, but only 145 bands are retained after preprocessing. The area used for the experiments has 1476×256 pixels with 30m spatial resolution. We used two spatially disjoint class maps from the same geographical region, and there are 9 classes in total. The training set consists of 1580 labeled instances, and the test set has 1434 instances.

For these experiments, we treat the problem as one of purely transductive learning. In other words, we use the test points to influence the dimensionality reduction. This is practical in this case since we want to classify unlabeled portions from the same image or geographic region. For classification based on the dimensionality reduction, we employed a maximum-likelihood classifier (MLC) with Gaussian distribution, where the class-conditional distribution of each class is modeled as a multi-variate Gaussian: $p(\mathbf{x}|y_i) \sim \mathcal{N}(\mu, \Sigma_i)$. The mean and covariance of each distribution are measured by maximum-likelihood estimation. Given a test data point, the MLC outputs the class label with maximum posterior probability, $y = \arg\max_i P(y_i|\mathbf{x}) \propto \arg\max_i p(\mathbf{x}|y_i)P(y_i)$. The prior probability distribution $P(y_i)$ is measured emipirically from the training set.

In this case, we compare effects of the dimensionality reduction obtained by our method (LB-LEM) to that of the standard LEM [4], as well as dimensionality reduction performed via an alternative non-linear dimensionality reduction method, ISOMAP [2], and the standard linear principle component analysis (PCA).

As shown in Figure 1, the modifications in the graph have a very positive effect on the classifier. In addition, it appears that the use of the domain knowledge through the labels incorporates some robustness into the selection of the appropriate dimensionality, as shown in Figure 2.

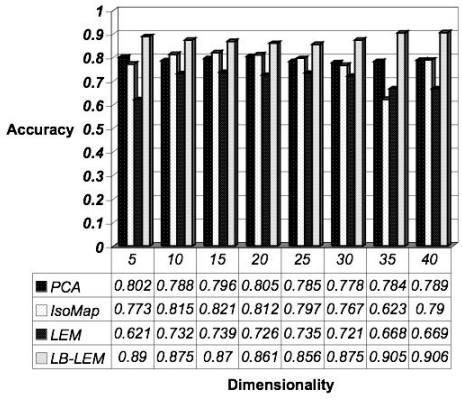

Figure 1. Classification accuracy on Botswana data using dimensionality reduction with ML classifier.

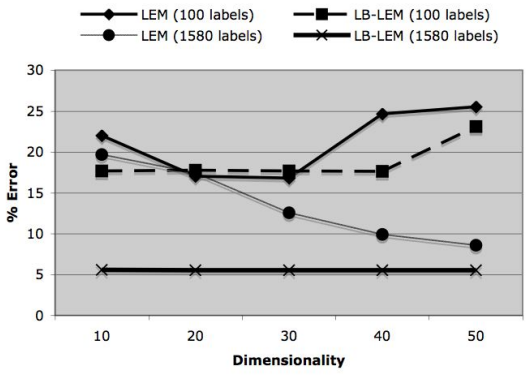

Figure 2. Change in classification error on Botswana data based on number of dimensions retained.

Since it is difficult to identify an optimal dimensionality when using methods like the Laplacian Eigenmap, a method that provides consistency across a variety of basis sizes is appealing.

V. Conclusion

We have outlined a robust method for biasing the graph construction in relevant semi-supervised learning methods and demonstrated its effectiveness. There are many obvious ways in which to optimize this general approach above and beyond its basic description, but at this juncture the focus has been on determining a general approach to providing robustness that does not require a lot of effort for manually incorporating domain knowledge. Since we utilize the labels more effectively as a source of domain knowledge, we can notice clear improvements when the feature set was not extensively hand engineered and is therefore noisy with respect to the target problem. These improvements are realized without requiring any significant increase in effort for domain adaptation.

Future work will include combining graph-altering methods with other forms of automated model selection suitable for small labeled training sets. In addition, it makes sense to ensure coverage of the space is essentially equivalent given that we want to alter a distance metric, so we have begun exploring modifications to the Stochastic Discrimination (SD) [20] approach to ensemble learning as a way of enforcing uniformity. In addition, while random subspace methods are very closely related to SD, the latter has nice theoretical error bounds that can potentially be applied to the resulting graph, thereby placing a bound on the unlabeled error rate.

Acknowledgment

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] O. Chapelle, B. Scholkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[2] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, 2000.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[5] B. Nadler, S. Lafon, and R. R. Coifman, "Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators," in *Advances in Neural Information Processing (NIPS)*, 2005.

[6] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[7] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 796–809, 2008.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[9] Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov, "Manifold learning: The price of normalization," *Journal of Machine Learning Research*, vol. 9, pp. 1909–1939, 2008.

[10] A. Goldberg, X. Zhu, A. Singh, Z. Xu, and R. Nowak, "Multi-manifold semi-supervised learning," in *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida, 2009.

[11] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI: American Mathematical Society, 1997.

[12] A. B. Goldberg, X. Zhu, and S. Wright, "Dissimilarity in graph-based semi-supervised classification," in *the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[13] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, pp. 1291–1302, 2003.

[14] M. Skurichina and R. P. W. Duin, "Bagging and the random subspace method for redundant feature spaces," in *Multiple Classifier Systems (MCS) 2001*, J. Kittler and F. Roli, Eds., vol. 2096. LNCS, 2001, pp. 1–10.

[15] M.-F. Balcan, "New theoretical frameworks for machine learning," PhD Thesis, 2008.

[16] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine Learning*, vol. 56, pp. 209–239, 2004.

[17] M. Ouimet and Y. Bengio, "Greedy spectral embedding," in *the 10th Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[18] C. S. Dhir and S. Y. Lee, "Hybrid feature selection: Combining fisher criterion and mutual information for efficient feature selection," in *International Conference on Neural Information Processing (ICONIP)*, 2009.

[19] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17–28, Jan 2002.

[20] A. Irle and J. Kauschke, "On kleinberg's stochastic discrimination procedure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 7, pp. 1482–1486, 2011.