This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

Large Margin Local Estimate with Applications to Medical Image Classification

Yang Song, *Member, IEEE,* Weidong Cai, *Member, IEEE,* Heng Huang, Yun Zhou, David Dagan Feng, *Fellow, IEEE,* Yue Wang, Michael J. Fulham, and Mei Chen, *Member, IEEE*

***Abstract*—Medical images usually exhibit large intra-class variation and inter-class ambiguity in the feature space, which could affect classification accuracy. To tackle this issue, we propose a new Large Margin Local Estimate (LMLE) classification model with sub-categorization based sparse representation. We first sub-categorize the reference sets of different classes into multiple clusters, to reduce feature variation within each subcategory compared to the entire reference set. Local estimates are generated for the test image using sparse representation with reference subcategories as the dictionaries. The similarity between the test image and each class is then computed by fusing the distances with the local estimates in a learning-based large margin aggregation construct to alleviate the problem of inter-class ambiguity. The derived similarities are finally used to determine the class label. We demonstrate that our LMLE model is generally applicable to different imaging modalities, and applied it to three tasks: interstitial lung disease (ILD) classification on high-resolution computed tomography (HRCT) images, phenotype binary classification and continuous regression on brain magnetic resonance (MR) imaging. Our experimental results show statistically significant performance improvements over existing popular classifiers.**

***Index Terms*—Medical image classification, sparse representation, sub-categorization, large margin fusion.**

This work was supported in part by Australian Research Council (ARC) grants. H. Huang was partially supported by US NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, and NSF-DBI 1356628. *Asterisk indicates corresponding author*.

Y. Song* is with Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia (e-mail: yson1723@uni.sydney.edu.au).

W. Cai is with BMIT Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia.

H. Huang is with Department of Computer Science and Engineering, University of Texas, Arlington, TX 76019, USA.

Y. Zhou is with the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

D. D. Feng is with BMIT Research Group, School of IT, University of Sydney, NSW 2006, Australia, and also with Med-X Research Institute, Shanghai Jiaotong University, Shanghai 200030, China.

Y. Wang is with the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA.

M. J. Fulham is with the Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, NSW 2050, Australia, and also with Sydney Medical School, University of Sydney, NSW 2006, Australia.

M. Chen is with the Department of Informatics at the University of Albany State University of New York, Albany, NY 12222, USA, and also with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

I. Introduction

Many medical imaging problems involve image classification as a major component. For example, in lesion detection, image classification is typically performed to differentiate the lesion from normal tissues. Similar processing is also applicable to segmentation tasks, by classifying different types of tissues to obtain the object boundaries. Image classification can also be the ultimate goal of medical imaging analysis to distinguish between different disease types or biomarkers based on images of various modalities.

One of the main challenges of image classification is the large intra-class variation and inter-class ambiguity. Desirably, images of the same class would have similar appearances and images of different classes would show different visual characteristics. With these properties, images would be easily and effectively classified. However, in real life, images of the same class can exhibit quite disparate features due to a class containing multiple types of visual patterns [1], especially in the presence of inter-subject variability. Images of different classes could also be difficult to distinguish due to low contrast between different tissues and structures. This confusion is commonly seen in medical imaging and can cause difficulties in image classification.

The pipeline of image classification typically comprises two stages: feature extraction and classification. Current research in feature extraction mainly focus on enhancing the descriptiveness and discriminative power of features by designing new feature descriptors [2]–[7], incorporating dictionary learning techniques [8]–[12], performing optimized feature selection [13]–[15], and conducting automated feature learning [16]–[20]. Given the intrinsic problems of inter-subject variability and low contrast between structures in medical imaging, even with the advanced feature extraction methods, there are often considerable intra-class variation and inter-class ambiguity. Therefore, in our study, we have focused on the second stage and designed an effective feature classification scheme to accommodate this feature space complexity.

A. Related Work

Classifiers are essential in medical image classification, to determine the class label of test images based on prior knowledge gathered from the training data. The most widely used classifiers are typically monolithic, meaning one classifier is used to classify all images. A set of binary classifiers of the same type trained in one-versus-all or one-versus-one manners for multi-class classification is also considered monolithic.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

Monolithic classifiers are often built on parametric models to represent the feature space separation. Popular choices used in medical imaging include the linear discriminant analysis (LDA) [21], [22] and support vector machine (SVM) [2], [3], [9], [13], [16], [17], [22], [23]. These classifiers can be very effective in generating clear feature space separation with highly descriptive and discriminative feature descriptors. Misclassification would however occur in cases with large intra-class variation and inter-class ambiguity.

An alternative to the monolithic classifiers is the data driven non-parametric model, which makes predictions based on the similarities between the test and reference images and does not rely on the parametric modeling of feature space separation. In particular, the k-nearest neighbor (kNN) classifier has been widely applied in medical imaging applications [21], [22], [24]–[27]. The classification performance of kNN is heavily dependent on the distance metric. Learning-based algorithms have thus been proposed to improve the discriminative power of distance metrics [28]–[33]. Distance metric learning has also been incorporated to bridge the semantic gap in bag-of-words representations [34]. Among these approaches, the large margin nearest neighbor (LMNN) [33] model has been successful in the general imaging domain. It learns a Mahalanobis distance metric with large margin constraints that the k-nearest neighbors from the correct class would be more similar to the test image in a linearly transformed feature space than those from the wrong classes. The learned distance metric thus produces superior classification performance over the standard kNN classifier. However, since the distance metric is monolithic and parametric, its effectiveness could be restricted by the feature space complexity.

Besides the learning-based distance metrics, the sparse representation classifier [35] can be considered as another enhanced kNN model with a different way of choosing the k-nearest neighbors. While kNN selects the nearest neighbors by similarity ranking using a distance metric, the sparse representation classifier finds the nearest neighbors by computing a weighted linear combination of the reference dictionary with sparse representation. Classification is then performed based on the distances between the test image and its sparse representations of various classes. Such sparse representation classification has become increasingly popular in the medical imaging domain [11], [36], [37]. Spatial information [38], [39] and multi-modality data [39]–[43] are often incorporated as additional constraints to improve the classification accuracy. Note that we focus our review on sparse representation classifiers with sparsity constraints on the combination of reference data, rather than sparse selection of feature variables.

It is commonly acknowledged that the effectiveness of sparse representation classifier is highly dependent on the quality of the reference data [35]. With large intra-class variation and inter-class ambiguity, the reference images of the wrong class could exhibit similar feature to the correct class. In such cases, it would be more likely to obtain close sparse representation for the wrong class than good feature space separation. This issue can be remedied by adapting the reference dictionary to the test images. For example, the locally-constrained linear coding (LLC) algorithm [44] has been adopted in many imaging applications including medical imaging [45]. Rather than using the entire set of reference images, sparse representation is computed with reference images that are locally similar to the test image. In previous work [6], the reference dictionary is non-linearly rescaled based on the feature distance between the test and reference images. In latter work [39], additional constraints are added so that the reference images that are more similar to the test image would have higher weights in the sparse representation. While these approaches demonstrate higher performance over the standard sparse representation classifier, these adaptation schemes can be heuristic.

Different classifiers often provide similar overall classification accuracies but different performance for individual data samples. Ensemble learning, such as bagging and boosting, is thus an intuitive option to fuse the classification outputs from different weak classifiers to obtain more accurate results. The weak classifiers can be of any type, including SVM [14], decision tree [5], [15], [22], [46]–[48], and logistic regression [21]. Sparse representation classifiers have also been integrated into the boosting model as weak classifiers based on random subsets of reference images [37], [49]. The combination of weak classifiers is normally based on a predefined weighting scheme, such as the majority voting or averaging of probabilities in bagging [5], [14], [15], [22], [47], [48], or choosing the best performing weak classifier at each training iteration with error-based weight computation in boosting [21], [37], [46], [49]. While these weighting schemes are often effective, they are however predefined, greedy, and and might not reflect the best adaptation to the dataset.

In view of feature space complexity, recently sub-categorization classification models [50]–[54] have been proposed to tackle this issue. They are similar to ensemble learning in that reference images are divided into subsets. However, in these sub-categorization models, the reference images of a certain class are divided into clusters based on some optimization criteria, not randomly. The sub-categorization models can be generally grouped into two types. The first type [50]–[52] designs a discriminative classifier with integrated clustering objective, so that the classification is optimized based on the separation between the subcategories. The second type [53], [54] clusters each class into subcategories, generates an individual classifier for each subcategory and then fuses the subcategory-level results to obtain the final classification. In the first type the clustering and classification objectives are effectively integrated, but the multi-stage design of the second type is clearer and each stage can be customized in a modular manner. In addition, these existing models have been built based on discriminative classifiers including LDA and SVM, but have not been extended to the other classifiers such as the sparse representation classifier.

B. Our Contribution

In this work, we propose a *large margin local estimate* (LMLE) classification model. Our hypothesis is that by embedding the sparse representation classifier in a sub-categorization construct, our classifier can effectively address the issues

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

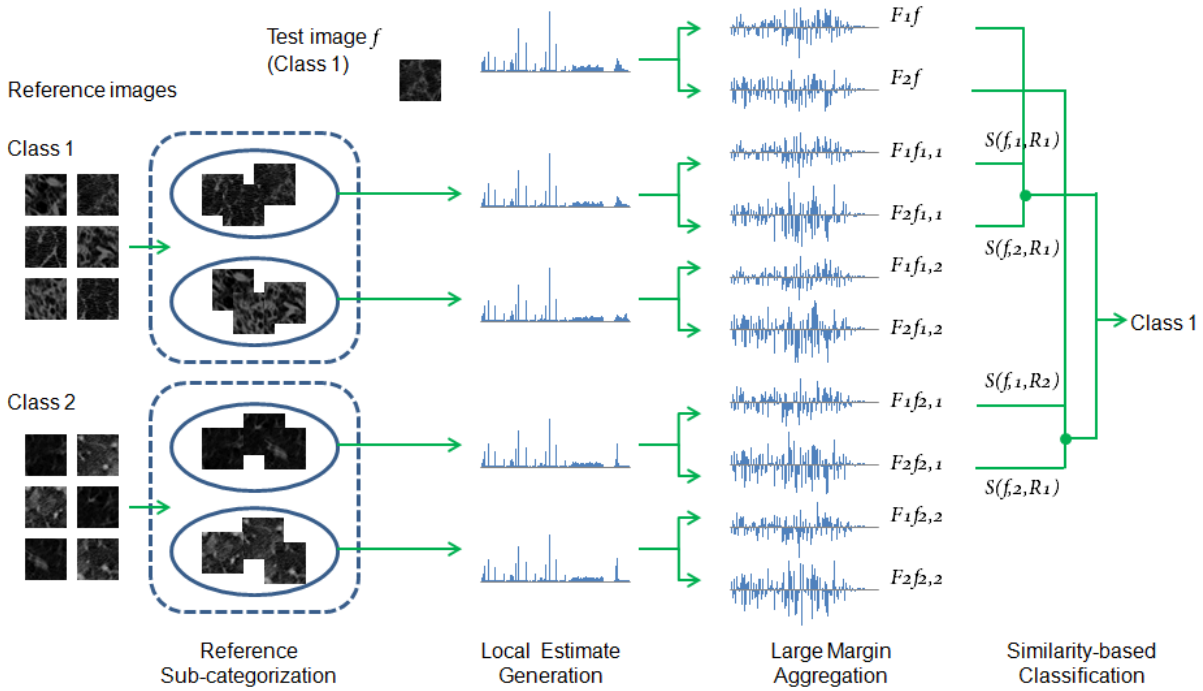

Fig. 1. Overview of our LMLE model. In this simplified example, there are two classes of images and the test image f is of class 1. Each reference set is clustered into two subcategories, and the top local estimate from each reference set is used in the large margin aggregation. The feature vector of f, its local estimates and the vectors transformed with learned matrices F_1 and F_2 are shown with histograms.

of intra-class variation and inter-class ambiguity. A brief overview of our LMLE model is as follows. First, the reference dictionary of each class is clustered into subcategories using a weighted clustering method. Then, using each subcategory as a reference dictionary, a local estimate is generated for the test image with sparse representation. Next, the similarity between the test image and each class is computed by fusing the local estimates with a learning-based large margin aggregation method. Finally, the test image is classified according to its similarities with the various classes.

Our main methodological contributions are three-fold. First, we incorporated sub-categorization into the sparse representation classifier. Due to the lower intra-class variation within a subcategory, the subcategory-level local estimates would exhibit a better representation of the similarity between the test and reference images in the localized feature space, compared to the global estimate using the entire reference dictionary. Then, we designed a large margin aggregation method to fuse the local estimates for similarity computation while tackling the issue of inter-class ambiguity. Transformation matrices are learned in a large margin construct and applied to the feature vector of the test image and the local estimates, so that the test image comes closer to the local estimates from the correct class and farther away from the wrong classes in the transformed feature space. Third, we designed a sub-categorization method based on weighted k-means clustering to group reference images with similar features vectors and feature separations into one subcategory.

Our LMLE model is closely related to the approaches that integrate sparse representation with AdaBoost [37], [49]. The main distinctions of our model are that the reference images are sub-categorized into clusters rather than randomly divided into subsets, and the sub-level results are fused with large margin aggregation rather than boosting. Our LMLE model is also related to the sub-categorization approach [53], [54]. Compared to these other approaches, which focus on data sub-categorization and contain a simple fusion step, our model involves the learning-based aggregation method for fusion, which leads to a less complicated sub-categorization step. In addition, our model differs from the other approaches in that it is built on the sparse representation classifier.

Preliminary data from this work were reported in our conference paper [55]. In this work, we have enhanced the sub-categorization method with automatically computed view and variable weights, simplified the final classification component based on similarities, elaborated our method design with further details, and performed more thorough performance comparison with related approaches. We have also extended the evaluation to a brain MR database for phenotype classification and regression tasks, in addition to the original database used for ILD classification in lung HRCT images, to demonstrate the general applicability of our LMLE model.

The rest of the paper is organized as follows. Section II gives the detailed description of our LMLE model. Section III describes the three databases used in our evaluation and the application-specific settings. Section IV presents the evaluation results and discussion. Section V concludes the paper.

II. Large Margin Local Estimate

We define the classification objective as finding the class label $\mathcal{L}(f)$ of a test image f, with $\mathcal{L}(f) \in \{1, ..., L\}$ and L

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

is the number of classes. In a supervised setting, L sets of reference images $\{R_l : l = 1, ..., L\}$ are given. Each reference set R_l represents one class l and contains N_l reference images, $R_l = \{r_l^i : i = 1, ..., N_l\}$. Assume feature vectors are precomputed for all images. The classification problem is to derive $\mathcal{L}(f)$ based on $\{R_l\}$. For notation simplicity, throughout this section, f and r_l^i denote both the test/reference images and the corresponding H-dimensional feature vectors.

Consider that a reference set of a certain class usually exhibits large intra-class variation and different reference sets often have large inter-class ambiguity. A close sparse representation of the test image could be achieved for the wrong class by combining reference images with diverse features, especially if the test image does appear similar to the reference images of the wrong class. In such cases, the representation output would not accurately reflect the similarity relationships between the test image and the various reference sets, causing misclassification. To better capture the similarity information, we propose the LMLE method that generates local estimates with reference sub-categorization and derives the similarity with large margin aggregation.

Fig. 1 gives an overview of our LMLE method. First, each reference set R_l is sub-categorized into a number of clusters. The reference images within a certain cluster will exhibit lower intra-class variation compared to that of the entire reference set. Then, with each cluster as a reference dictionary, a local estimate is computed for the test image f with sparse representation. The set of local estimates of a certain class represents the similarity between f and that class. The next step is to compute the similarity between f and each class by fusing the distances between f and the local estimates from that class. During this fusion step, to obtain higher similarity with the correct rather than incorrect classes, the local estimates are transformed with a large margin aggregation approach. Lastly, the test image f is classified based on the similarities.

A. Reference Sub-categorization

The first step of our method is to sub-categorize a reference set R_l into K_l clusters/subcategories. Denote one subcategory as $R_l^k \subset R_l : k = 1, ..., K_l$. The problem is to assign each reference image $r_l^i \in R_l$ to one of the subcategories. We design a sub-categorization method based on the k-means clustering technique.

We expect that reference images with similar feature vectors and similar feature separations from the other classes to be grouped into the same subcategory. Formally, given two reference images r_l^i and r_l^j, we define two difference terms: (i) $\|r_l^i - r_l^j\|$, which is the Euclidean distance between the two feature vectors; and (ii) $\|d_l^i - d_l^j\|$, which is the Euclidean distance between the two feature separations. The feature separation d_l^i is an $L-1$ dimensional vector, in which each element $d_l^i(l')$ is the mean Euclidean distance between r_l^i and the reference set $R_{l'}$, with $l' = 1, ..., L$ and $l' \neq l$.

Then, we formulate the sub-categorization objective to minimize the within-subcategory distances:

$$\underset{\{R_l^k\}_k}{\operatorname{argmin}} \sum_{k=1}^{K_l} \sum_{r_l^i \in R_l^k} (r_l^i - \mu_l^k)^T W_l (r_l^i - \mu_l^k) + (d_l^i - \theta_l^k)^T U_l (d_l^i - \theta_l^k), \tag{1}$$

which can also be written as:

$$\underset{\{R_l^k\}_k}{\operatorname{argmin}} \sum_{k=1}^{K_l} \sum_{r_l^i \in R_l^k} (x_l^i - \phi_l^k)^T Q_l (x_l^i - \phi_l^k), \tag{2}$$

where μ_l^k and θ_l^k represent the mean r_l^i and d_l^i of the subcategory R_l^k, and W_l and U_l are diagonal matrices containing weight factors. In the combined form, $x_l^i \in \mathbb{R}^{H+L-1}$ is the concatenation of r_l^i and d_l^i, $\phi_l^k \in \mathbb{R}^{H+L-1}$ is the concatenation of μ_l^k and θ_l^k, and $Q_l \in \mathbb{R}^{(H+L-1)\times(H+L-1)}$ represents the diagonal matrix combining W_l and U_l.

Our design of the objective function is based on the following considerations. First, the feature vector r_l^i is typically multiview, i.e. comprising multiple views of features with one view representing one type of feature. A feature vector r_l^i can thus be divided into a number of views, and each view contains multiple feature elements, i.e. variables. We hypothesize that different views and different variables should carry different weights to the distance computation. The weight matrix W_l thus represents the view and variable weights and it is class-specific. Second, each element of the feature separation d_l^i should also have a certain weight, and the weight matrix U_l also controls the balance between the two distance terms (feature vector and separation). Third, rather than predefining the weights W_l and U_l, we would like to compute them automatically to help achieve better sub-categorization. We adopt the two-level weighting k-means (TW-k-means) algorithm [56] to solve this minimization problem.

Specifically, we reformulate the objective function as:

$$\begin{aligned}
\underset{\{R_l^k\}_k}{\operatorname{argmin}} \quad & \sum_{k=1}^{K_l} \sum_{r_l^i \in R_l^k} \sum_{v=1}^{V+1} \sum_{h \in G_v} \alpha_l(v)\beta_l(h)\{x_l^i(h) - \phi_l^k(h)\}^2 \\
& + C_1 \sum_{v=1}^{V+1} \alpha_l(v) \log\{\alpha_l(v)\} \\
& + C_2 \sum_{h=1}^{H+L-1} \beta_l(h) \log\{\beta_l(h)\}, \\
s.t. \quad & \sum_{v=1}^{V+1} \alpha_l(v) = 1, \;\; \alpha_l(v) \geq 0, \\
& \sum_{h \in G_v} \beta_l(h) = 1, \;\; \beta_l(h) \geq 0, \;\; \forall v = 1, ..., V+1.
\end{aligned} \tag{3}$$

Here the first term is indeed the same as Eq. (2). The vectors $\alpha_l \in \mathbb{R}^{V+1}$ and $\beta_l \in \mathbb{R}^{H+L-1}$ represent the view and variable weights, with $V+1$ and $H+L-1$ as the numbers of views and variables in x_l^i. V is the number of views in r_l^i, and the plus one is to include d_l^i as an additional view. G_v denotes the set of variable indices belonging to view v. Here for a variable index h, assuming it belongs to view v, the element-wise weight would be $\alpha_l(v)\beta_l(h)$, and this value would equal to $Q_l(h,h)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

The difference between Eq. (3) and (2) is the additional second and third terms to minimize the negative entropies of the view and variable weights. C_1 and C_2 are two parameters to control the contributions of the two terms. In addition, we also require that the sum of all view weights is one and the sum of variable weights of a certain view is one. In this way, the view weights determine the relative importance of the multiple views and the variable weights are localized at the view-level.

To solve Eq. (3), we apply an iterative approach [56] with the following steps.

Step 1. With fixed $\{\phi_l^k\}$, α_l and β_l, $\{R_l^k\}$ is obtained by assigning each reference image r_l^i to the nearest subcategory center with the distance computed as $\sum_{v=1}^{V+1}\sum_{h\in G_v}\alpha_l(v)\beta_l(h)\{x_l^i(h)-\phi_l^k(h)\}^2$.

Step 2. With fixed $\{R_l^k\}$, α_l and β_l, update $\{\phi_l^k\}$ that $\phi_l^k(h)=N_l^{-1}\sum_{r_l^i\in R_l^k}x_l^i(h)$.

Step 3. With fixed $\{\phi_l^k\}$, $\{R_l^k\}$ and β_l, α_l is computed via:

$$\alpha_l(v)=\frac{\exp\{-A_l(v)C_1^{-1}\}}{\sum_{n=1}^{V+1}\exp\{-A_l(n)C_1^{-1}\}}, \tag{4}$$

where

$$A_l(v)=\sum_{k=1}^{K_l}\sum_{r_l^i\in R_l^k}\sum_{h\in G_v}\beta_l(h)\{x_l^i(h)-\phi_l^k(h)\}^2. \tag{5}$$

Step 4. With fixed $\{\phi_l^k\}$, $\{R_l^k\}$ and α_l, β_l is computed via:

$$\beta_l(h)=\frac{\exp\{-B_l(h)C_2^{-1}\}}{\sum_{n\in G_v}\exp\{-B_l(n)C_2^{-1}\}}, \tag{6}$$

where

$$B_l(h)=\sum_{k=1}^{K_l}\sum_{r_l^i\in R_l^k}\alpha_l(v)\{x_l^i(h)-\phi_l^k(h)\}^2, \tag{7}$$

with v indicating the view that the variable h belongs to. We refer the readers to [56] for the proofs of Steps 3 and 4.

By initializing $\{\phi_l^k\}$ randomly and α_l and β_l as uniform weights, steps 1 to 4 are iterated until the objective function Eq. (3) reaches the local minimum. The subcategories $\{R_l^k\}$ are then obtained.

In the above reference sub-categorization process, the parameters C_1 and C_2 are important. To choose suitable C_1 and C_2, we design a sub-categorization criterion (SC) to measure the compactness of subcategories and the separation between subcategories of different classes. The design is based on the Dunn index [57], but extended to accommodate subcategories of multiple classes. Specifically, we define SC as follows:

$$\begin{aligned}&SC(C_1,C_2)=(\sum_{l=1}^{L}N_l)^{-1}\cdot\\&\sum_{l=1}^{L}N_l\{\frac{\min_{k_1=1}^{K_l}\min_{k_2=1,k_2\neq k_1}^{K_l}\|\phi_l^{k_1}-\phi_l^{k_2}\|^2}{\max_{k=1}^{K_l}\Delta R_l^k}+\\&\sum_{l'=1,l'\neq l}^{L}\frac{\min_{k=1}^{K_l}\min_{k'=1}^{K_{l'}}\|\phi_l^{k}-\phi_{l'}^{k'}\|^2}{\max_{k_1=1}^{K_l}\max_{k_2=1,k_2\neq k_1}^{K_l}\|\phi_l^{k_1}-\phi_l^{k_2}\|^2}\},\end{aligned} \tag{8}$$

where $\cdot$ at the end of the first line denotes the multiplication operation, and ΔR_l^k denotes the maximum Euclidean distance between pairs of x_l^i in the subcategory R_l^k. The parameters C_1 and C_2 are implicitly incorporated into this equation that the subcategory assignments R_l^k and ϕ vectors are computed with certain values of C_1 and C_2 in Eq. (3). The two terms in the sum operation represent (i) the ratio between the minimum between-subcategory distance and the maximum within-subcategory distance of class l, and (ii) the total of the ratios between the minimum between-subcategory distances with the other classes $l'\neq l$ and the maximum between-subcategory distance within class l. The two terms are computed for each class $l=1,...,L$. The average over all classes, which is weighted by the number of reference images in each class, generates the SC. A higher SC implies that the subcategories are more compact and more separated within the same class and between different classes. Then by varying C_1 and C_2 from 10 to 25 with an increment of 5, we perform the sub-categorization with varying settings of C_1 and C_2 and select the values producing the highest SC. This range of 10 to 25 is used according to the performance analysis in [56], which shows balanced distributions of view and variable weights.

Besides C_1 and C_2, the sub-categorization method also involves the parameter K_l. Similar to k-means clustering, we set it manually based on our empirical studies. We will provide more details in Section III.

B. Local Estimate Generation

With the subcategories generated for each reference set, our second step is to compute the local estimates with each reference subcategory R_l^k as the reference dictionary. Formally, given a test image f, we find the local estimate $f_{l,k}$ by linear combination of a sparse set of reference data in R_l^k. The local estimate $f_{l,k}$ represents a sparse representation of the test image f. A small representation error $\|f-f_{l,k}\|^2$ means high similarity between f and the reference dictionary R_l^k, and hence a high probability of f belonging to class l.

We suggest that any sparse coding algorithm can be used to generate $f_{l,k}$. There are a number of existing sparse coding formulations with various L1 [58] or L0 [59] regularizations. In our study, we choose to use the L0 formulation for its simplicity. The local estimate is thus generated by:

$$\begin{aligned}y_{l,k}=&\ \operatorname{argmin}_{y_{l,k}}\|f-R_l^k y_{l,k}\|_2^2,\ \ s.t.\ \|y_{l,k}\|_0\leq C,\\f_{l,k}=&\ R_l^k y_{l,k},\end{aligned} \tag{9}$$

where $R_l^k\in\mathbb{R}^{H\times N_l^k}$ represents the concatenated matrix of feature vectors of all reference images in R_l^k with N_l^k denoting the number of reference images, $y_{l,k}\in\mathbb{R}^{N_l^k}$ is the sparse coefficient vector and C is a constant. The orthogonal matching pursuit (OMP) [59] algorithm[1] is used to derive $y_{l,k}$. Note that this component is not the focus of our methodology and hence we choose to use the standard algorithms rather than proposing our own solution.

C. Large Margin Aggregation

After the local estimates are obtained, the next step is to compute the similarity between the test image f and each

[1]The OMP package is downloaded from http://www.cs.technion.ac.il/ronrubin/software.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

reference set R_l. The degree of similarity corresponds to the probability of classifying f to class l. Formally, the problem is to compute the similarity $S(f, l, R_l)$ based on the set of local estimates $\{f_{l,k} : k = 1, ..., K_l\}$ generated for class l.

There are a number of ways to do this. For example, a simple max pooling technique can be applied, by selecting the local estimate that is the closest to the test image to derive similarity. Mean pooling can also be used, by computing the mean local estimate and calculating the distance between the mean and the test image. A variation of the mean pooling is to incorporate the concept of kNN, by computing the mean local estimate based on only the top few local estimates that are the closest to the test image.

The main design consideration for this similarity measure is that we would like a large similarity from the correct class and small similarities from the wrong classes, based on which the test image would then be accurately classified. The above mentioned techniques would not necessarily satisfy this expectation, mainly due to the large inter-class ambiguity. With inter-class ambiguity, some local estimates from the incorrect class would inevitably appear close to the test image; and these similarities could possibly be larger than those computed from the correct class.

To overcome this issue, we designed a large margin aggregation algorithm with a learning-based transformation matrix. By transforming the test image and local estimates in a large margin construct, the test image would become closer to the local estimates of the correct class and more distant from those of the wrong classes. The similarities derived in this way would then lead to better classification.

Specifically, we define the similarity between the test image f and reference set R_l as:

$$S(f, l, R_l) = \{1 + D(f, l, R_l)\}^{-1}, \tag{10}$$

where $D(f, l, R_l)$ is the distance between f and R_l, and computed as the accumulated transformed Euclidean distance between f and M local estimates that are the closest to f:

$$D(f, l, R_l) = \sum_{m=1}^{M} \|F_l f - F_l f_{l,m}\|^2. \tag{11}$$

Here $F_l \in \mathbb{R}^{H \times H}$ is the learned transformation matrix and is specific to class l. The selection of M closest local estimates $\{f_{l,m} : m = 1, ..., M\}$ is based on the Euclidean distance between the transformed vectors $F_l f$ and $F_l f_{l,k}$. Note that $S(f, l, R_l)$ and $D(f, l, R_l)$ involve the l factor, which means that f is assumed to belong to class l (not necessarily true) and the transformation matrix F_l of class l is to be used to compute the distance between f and R_l. In other words, $S(f, l, R_l) \neq S(f, l', R_l)$ with $l' \neq l$ and likewise $D(f, l, R_l) \neq D(f, l', R_l)$, since a different $F_{l'}$ would be used to compute $D(f, l', R_l)$.

With this distance function, assuming the class label of f is l, we would expect that $D(f, l, R_l) < D(f, l, \{R_{l'} : l' = 1, ..., L, l' \neq l\})$. This is equivalent to:

$$\sum_{m=1}^{M} \|F_l f - F_l f_{l,m}\|^2 < \sum_{m'=1}^{M} \|F_l f - F_l f_{l',m'}\|^2, \tag{12}$$

where m' indexes the M closest local estimates $\{f_{l',m'}\}$ from all the wrong classes $l' \neq l$. Note that $\{f_{l',m'}\}$ are pooled from all classes $l' \neq l$ rather than selected for each wrong class, in order to minimize the number of constraints and hence the training complexity.

To obtain the transformation matrix F_l, we gather a set of I training samples from the reference set R_l, which are denoted as $\{r_l^i : i = 1, ..., I\}$. For a certain training sample r_l^i, the mth closest local estimates from the correct class is denoted as $r_{l,m}^i$, and likewise, the m'th closest local estimate from the wrong class is denoted as $r_{l',m'}^i$. The degree of closeness is determined by the Euclidean distance between r_l^i and the local estimates.

We then formulate our goals of training as: (i) to minimize the total distance between the transformed feature $F_l r_l^i$ and M closest local estimates $\{F_l r_{l,m}^i\}$ from the correct class:

$$\sum_{m=1}^{M} \|F_l(r_l^i - r_{l,m}^i)\|^2, \tag{13}$$

and (ii) to impose a large margin difference so that the transformed local estimate from the wrong classes is one unit further away than that from the correct class, for all pairs of $F_l r_{l,m}^i$ and $F_l r_{l',m'}^i$:

$$\begin{aligned} &\|F_l(r_l^i - r_{l,m}^i)\|^2 + 1 \leq \|F_l(r_l^i - r_{l',m'}^i)\|^2, \\ &\forall m = 1, ..., M,\ m' = 1, ..., M. \end{aligned} \tag{14}$$

The overall training objective thus becomes:

$$\begin{aligned} &\underset{F_l}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{m=1}^{M} \|F_l(r_l^i - r_{l,m}^i)\|^2 + \\ &\sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{m'=1}^{M} [1 + \|F_l(r_l^i - r_{l,m}^i)\|^2 - \|F_l(r_l^i - r_{l',m'}^i)\|^2]_+, \end{aligned} \tag{15}$$

where $[z]_+ = \max(0, z)$ is the standard hinge loss. By solving this objective function, the transformed feature $F_l r_l^i$ would become more similar to $\{F_l r_{l,m}^i : m = 1, ..., M\}$ than $\{F_l r_{l',m'}^i : m' = 1, ..., M\}$, and hence would be classified accurately as class l.

To solve the objective function Eq. (15), we note that $\|F_l(r_l^i - r_{l,m}^i)\|^2$ can be rewritten as $(r_l^i - r_{l,m}^i)^T X_l (r_l^i - r_{l,m}^i)$ where $X_l = (F_l)^T F_l$. We then reformulate Eq. (15) following the semidefinite programming model [60] as:

$$\begin{aligned} \underset{X_l}{\operatorname{argmin}} \quad & \sum_{i=1}^{I} \sum_{m=1}^{M} (r_l^i - r_{l,m}^i)^T X_l (r_l^i - r_{l,m}^i) + \\ & \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{m'=1}^{M} \xi_{imm'}, \\ s.t. \quad & (r_l^i - r_{l',m'}^i)^T X_l (r_l^i - r_{l',m'}^i) - \\ & (r_l^i - r_{l,m}^i)^T X_l (r_l^i - r_{l,m}^i) \geq 1 - \xi_{imm'}, \\ & \xi_{imm'} \geq 0, \quad X_l \succeq 0, \end{aligned} \tag{16}$$

where X_l is required to be positive semidefinite. The slack variable $\xi_{imm'}$ is introduced to represent the hinge loss.

Overall, our formulation of the optimization goal for large margin aggregation, especially Eq. (16), is mathematically

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

similar to the LMNN algorithm [33]. However, the underlying concepts are different. In particular, if applying LMNN to classify the images, the distances would be computed between the reference image r_l^i and its neighboring images r_l^m and $r_{l'}^{m'}$ from the same and different classes. In our model, the distances are computed between the reference image r_l^i and its local estimates $r_{l,m}^i$ and $r_{l',m'}^i$, and there is no nearest-neighbor relationship between the reference images. On the other hand, considering the LMNN solver[2] is optimized for efficiency, we choose to modify it to adapt to our formulation and solve Eq. (16) to obtain X_l and then F_l. Briefly, the solver is modified to identify similar local estimates from the various classes for each feature vector and construct this information as the input to the optimization routine.

Note that this large margin aggregation method is used to learn the transformation matrix F_l for each class. The similarity between the test image f and reference set R_l could then be derived using Eq. (10). The only parameter involved is M, which we found based on our empirical studies that its value is best to be application specific. The number of training samples I depends on the experimental design of training and testing. We provide more details in Section III.

D. Similarity-based Classification

The last step is to classify the test image f based on its similarities with the reference sets $\{R_l\}$. This procedure is defined as:

$$\mathcal{L}(f) = \underset{l}{\operatorname{argmax}}\{\max_{l'=1,...,L} S(f, l', R_l)\}, \tag{17}$$

with $l = 1, ..., L$. The class label $\mathcal{L}(f)$ thus corresponds to the highest similarity value among the various reference sets computed with the various transformation matrices.

III. Applications to Medical Image Classification

The proposed LMLE method was applied to three problems: ILD classification in lung HRCT images, phenotype classification and regression in brain MR. We describe the problem domains, imaging datasets, and application-specific processing in the following sections. The evaluation metrics are also described.

A. ILD Classification in Lung HRCT Images

Interstitial lung disease (ILD) represents a group of lung diseases that affect the interstitium, and cause progressive scarring of lung tissue and progressive dyspnea [61]. HRCT is typically used to visualize the tissue patterns to identify the specific type of ILD. Manual interpretation is challenging and time-consuming especially with large inter-subject variation even for the same type of ILD. Large inter-class ambiguity and intra-class variation also cause difficulties in designing automated methods. As shown in Fig. 2, the images of different types can be similar while those of the same type can appear different. For example, the two ground glass images look different, with extensive ground glass opacity in the top image but minimal in the bottom image. The fibrotic types include a mixture of tissue patterns, such as reticulation, architectural distortion and honeycombing, hence encompassing a large degree of visual variation. In the emphysema images, there are low attenuation patterns and the mild or moderate cases can exhibit similar appearances to normal cases. Identifying the correct tissue type is thus challenging with such intra-class variation and inter-class ambiguity.

[2]The LMNN package is downloaded from http://www.cse.wustl.edu/ kilian/code/code.html

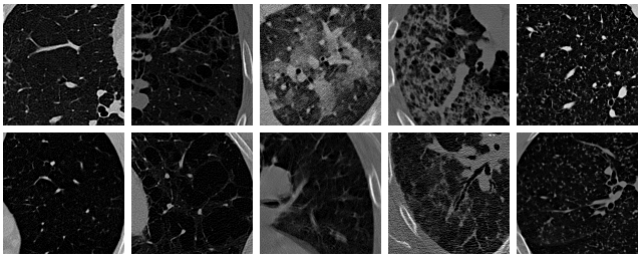

Fig. 2. Two example images (segments of HRCT axial slices) for each of the five lung tissue types. From left to right columns: normal (NM), emphysema (EM), ground glass (GG), fibrosis (FB) and micronodules (MN).

TABLE I
Number of image patches and subjects of the five tissue classes in the ILD database.

Class	NM	EM	GG	FB	MN
# image patches	6438	1474	2974	4396	7849
# subjects	12	5	35	35	16

The ILD database [62] has been developed by Depeursinge et al. to provide a common platform for evaluating automated ILD classification methods. The database contains 113 HRCT images, with 2062 2D regions each annotated as one of the seventeen types of lung tissues. Following the setup of the existing classification methods for this ILD database [2], [6], [49], we classify five classes of lung tissue ($L = 5$): normal (NM), emphysema (EM), ground glass (GG), fibrosis (FB) and micronodules (MN). The classification was performed on 2D image patches of 31×31 pixels. The patches are half-overlapping with centroids residing inside the annotated regions. The number of image patches belonging to each tissue class is listed in Table I, selected from 93 HRCT images. Note that some subjects contain more than one tissue type. For simplicity, we refer to the image patches as images in the following description.

To apply our LMLE method, each image was first represented by a 176 dimensional ($H = 176$) texture-intensity-gradient (TIG) feature vector [6]. Specifically, the feature vector included three views ($V = 3$): rotation-invariant Gabor-local binary patterns (RGLBP) texture feature, intensity histogram, and multi-coordinate histogram of oriented gradients (MCHOG) gradient feature. We chose to use the TIG feature vector because it showed good descriptiveness and discriminative power on the ILD database [6], [49].

We found empirically that if using the entire reference set as the reference dictionary (without sub-categorization), good sparse representations could be obtained by combining

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

a minimum of 10 reference images. We thus decided that the number of reference images in a subcategory should be large enough that at least 10 of them could be sparsely selected. Consequently, we chose to set the number of subcategories $K_l = \lfloor N_l/40 \rfloor$, so that a subcategory would contain around 40 images; and the constant C for local estimate generation was set to 10. We preferred not to use a number larger than 40 since we also required a sufficient number of subcategories to fuse the local estimates. In addition, based on our parameter selection defined in Eq. (8), the parameters C_1 and C_2 were set to 15 and 10, respectively. For large margin aggregation, various settings of the number of closest local estimates M were experimented (details shown in Section IV), and $M = 5$ was found to provide the best classification results.

For consistency of experimental setup and convenience of performance comparison with our previous work [49], [55], we divided the database into four subsets of similar number of subjects. Within each subset, a leave-one-subject-out scheme was used by having each subject as the test data and the other subjects as the reference data to generate subcategories and local estimates. For large margin aggregation, training was conducted by combining every three subsets and randomly selecting 10% of the data to learn one set of transformation matrices; hence a total of four sets of transformation matrices were learned. During testing, the transformation matrices learned from the subsets not containing the test subject were applied to classify the test data.

B. BP Classification and Regression in Brain MR Images

The MICCAI 2014 Machine Learning Challenge (MLC) [63] is aimed at evaluating different machine learning methods in predicting clinically relevant brain phenotype (BP) using MR scans. MLC includes two specific objectives: binary two-class ($L = 2$) classification to predict a binary class label, and continuous regression to predict a continuous numerical label, for a certain subject. The database comprises 150 subjects with annotated binary labels (with 75 from each of the two classes), and 315 subjects with annotated continuous labels (ranged from 1.8 to 9.4). Note that the clinical context of the dataset is not released to the public hence the exact meanings of the binary and continuous labels are unknown. Nevertheless, we chose to employ this database in our experimentation, since it provided a standardized platform for performance comparison with pre-computed feature vectors; and it helped to demonstrate the generalizability of our method with a different imaging modality (MR) and organ of interest (brain) from the HRCT lung database and including a regression task.

For the continuous regression problem, with our LMLE model, the continuous labels were first quantized into four discrete labels (divided equally between 1.6 and 9.6) to convert the regression problem into a four-class ($L = 4$) classification problem. The number of images belonging to each class is listed in Table II. After the classification using LMLE, a continuous label was then generated as the regression result based on the reference images. Specifically, assume the test image f was classified to class $l \in \{1, ..., 4\}$. Its continuous label was computed by averaging the sparse represented continuous

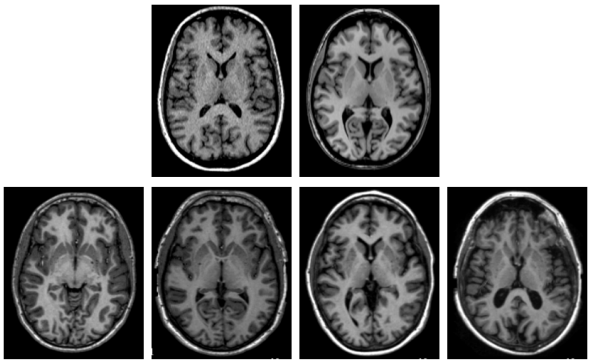

Fig. 3. Example brain MR images, showing one axial slice from one subject of each class. Top row: two classes of the binary classification task. Bottom row: four classes of the continuous regression task.

TABLE II
NUMBER OF IMAGES OF DIFFERENT CLASSES IN THE MLC DATABASE.

	Binary		Continuous			
Class	1	2	1	2	3	4
# images	75	75	149	58	65	43

labels from the top M subcategories of class l:

$$M^{-1} \sum_{m=1}^{M} \rho_l^m y_{l,m}, \tag{18}$$

in which $y_{l,m}$ is obtained using Eq. (9), and $\rho_l^m \in \mathbb{R}^{1 \times N_l^m}$ contains the continuous labels of the reference images in R_l^m.

We directly used the feature vectors provided in the MLC database. A 184-dimensional feature vector ($H = 184$) is provided for each subject, which is a set of standard morphological features computed with FreeSurfer [64], including volumes of cortical and sub-cortical structures and average thickness measurements within cortical regions. The detailed description of the features can be found in the database. Example images of the two tasks are shown in Fig. 3. To obtain the number of views, we analyzed the distributions of each variable and grouped the variables of similar value ranges into one view. We thus identified three feature views ($V = 3$), with the first 115 variables as view 1, variables 115-183 as view 2 and the last variable 184 as view 3. Such a separation of views was also semantically verified based on the list of feature names.

Similar to our processing for the ILD dataset, we found that 10 reference images could provide good sparse representation, and hence set $C = 10$ for local estimate generation. However, considering that the number of subjects is limited in the MLC dataset, we set the number of subcategories $K_l = \lfloor N_l/20 \rfloor$ so that there would be a sufficient number of subcategories. The parameters C_1 and C_2 were found to be both 10, according to Eq. (8). All reference images were used to learn the transformation matrices $\{F_l\}$. The parameter M was set to 2, based on our experimental evaluation (details in Section IV). The process of training and testing was similar to that for the ILD database, based on leave-one-subject-out but without subdivision of subsets for reference data construction, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

TABLE III
RECALL, PRECISION, F-SCORE AND OVERALL ACCURACY OF ILD CLASSIFICATION.

	NM	EM	GG	FB	MN
Recall	0.861	0.801	0.830	0.874	0.877
Precision	0.893	0.738	0.763	0.869	0.900
F-score	0.877	0.768	0.795	0.872	0.888
Accuracy			0.861		

TABLE IV
CONFUSION MATRIX OF ILD CLASSIFICATION.

Ground Truth	Prediction NM	EM	GG	FB	MN
NM	**0.861**	0.045	0.028	0.013	0.053
EM	0.129	**0.801**	0.005	0.065	0.000
GG	0.076	0.000	**0.830**	0.043	0.052
FB	0.005	0.014	0.059	**0.874**	0.048
MN	0.030	0.000	0.058	0.036	**0.877**

with four-fold cross validation for learning the large margin aggregation. The above settings were common to the binary and four-class classification problems.

C. Metrics of Performance Evaluation

In the ILD classification and BP binary classification tasks, we evaluated the recall, precision and F-score measures for each class. Accuracy was also measured for the overall database, defined as the number of correct classifications divided by the total number of test samples in the database. The receiver operating characteristic (ROC) analysis was performed to show the true positive rates versus false positive rates. To obtain the ROC curves, the derived similarities $S(f, l, R_l)$ were linearly normalized so that for each test image f, the similarity value corresponding to the classification output was rescaled to score of 1; and the classification threshold varied based on these scores. For multi-class ILD classification, an ROC curve was generated for each class using the one-versus-all technique, i.e. choosing one positive class and the others as the negative classes. The area under the curve (AUC) was then computed to quantify the classification performance. We compared with the standard classifiers that are related to our LMLE model, including the kNN, LMNN, SVM, and sparse representation classifiers. McNemar's test was used to analyze the statistical significance of performance improvement over these compared classifiers.

For the BP continuous regression task, we evaluated the root-mean-square error (RMSE) between the predicted and annotated continuous labels. We also measured the per-class recall, precision and F-score, and overall accuracy for the four-class classification. We compared with the standard approaches, including the linear regression, kNN, LMNN, and sparse representation. The statistical significance of performance improvement was evaluated with the paired t-test.

IV. RESULTS AND DISCUSSION

A. Overall Performance

The recall, precision, F-score and accuracy of ILD classification using our LMLE model are shown in Table III. The confusion matrix is shown in Table IV. The results show good classification performance with relatively balanced rates among different tissue types. All tissue types exhibited recall levels above 80%. The EM tissue type obtained the lowest precision, although in fact there were not many other types of images misclassified as EM. Its precision was largely affected by the small number of EM images compared to the other tissue types. The results also show very good differentiation between GG and EM, and EM and MN, with close to 0 misclassification rates. A special case was that 12.9% of EM images were misclassified as NM. We observed that a majority of these EM images displayed high visual similarities with the NM images, and these images were difficult to differentiate even by visual analysis. The low discriminative power of feature descriptors thus affected the classification performance between EM and NM.

TABLE V
WITHIN-CLASS FEATURE REPRESENTATIVENESS AND BETWEEN-CLASS FEATURE DISCRIMINATION IN THE ILD DATABASE.

	NM	EM	GG	FB	MN	Multi
NM	0.834	0.843	0.803	0.914	0.949	0.567
EM	0.638	0.866	0.953	0.843	0.798	0.637
GG	0.853	0.860	0.902	0.662	0.903	0.625
FB	0.971	0.919	0.937	0.901	0.967	0.816
MN	0.773	0.869	0.705	0.577	0.847	0.832

To better demonstrate the effectiveness of our LMLE model, the results in Table III and IV need to be analyzed relative to the difficulty of the problem. Specifically, we wanted to measure the intra-class variation and inter-class ambiguity in the feature space, and then gauge the classification performance based on the measured data. To do this, we conducted one-class classification for each tissue type, one-versus-one pairwise classification between each pair of tissue types, and one-versus-all multi-class classification for all tissue types, all using SVM. Four-fold cross validation was applied and the polynomial kernel was found to perform the best. The LIBSVM [65] package was used. The results are shown in Table V, in which the diagonal shows the one-class classification rates, the last column shows the one-versus-all multi-class classification rates, and the other off-diagonal elements show the one-versus-one pairwise classification rates. With ideally representative and discriminative features, we would expect perfect rates. In this database, however, there were on average 15% one-class and pairwise misclassifications, and 30% multi-class misclassifications, which indicated the influence of intra-class variation and inter-class ambiguity.

Compared to the multi-class classification results in Table V, our recall rates showed a large improvement, suggesting the effectiveness of our LMLE model in accommodating the feature space complexity. The other results in Table V are not directly comparable with Table IV, since our results were obtained from multi-class classification. Nevertheless, it can be seen that only 63.8% of EM images were correctly classified with pairwise classification of EM and NM images. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

TABLE VI
RECALL, PRECISION, F-SCORE AND OVERALL ACCURACY OF BP BINARY CLASSIFICATION.

	Class-1	Class-2
Recall	0.929	0.960
Precision	0.972	0.916
F-score	0.944	0.945
Accuracy	0.945	

TABLE VII
CONFUSION MATRIX OF BP BINARY CLASSIFICATION.

Ground Truth	Prediction Class-1	Prediction Class-2
Class-1	**0.929**	0.071
Class-2	0.040	**0.960**

illustrated that EM and NM types were not well separated in the feature space with high inter-class ambiguity. This also explained the high misclassification rates between EM and NM in Table IV. The poorest separation was between MN and FB images at 0.577, while our result show relatively low misclassification rate (0.036) between these two types. While there were considerable misclassification between GG and EM, and EM and MN (Table V), our results in Table IV show very accurate differentiation between these pairs. These results demonstrate that by minimizing the intra-class variation at the subcategory-level and tackling the inter-class ambiguity with large margin aggregation, our LMLE model could better discover the intrinsic feature space separation and enhance accurate classification.

The BP binary classification results are shown in Table VI and VII. The feature space analysis using one-class and pairwise SVM classifiers are shown in Table VIII. The linear kernel was found to perform the best with SVM and four-fold cross validation was used. The low rates in Table VIII suggest that this classification task was particularly challenging with large intra-class variation and inter-class ambiguity. In this case, since the task is binary classification, we can compare the classification recall in Table VI directly with the off-diagonal rates in Table VIII, 0.929 vs. 0.533 and 0.960 vs. 0.640. The large improvement demonstrates that our LMLE model was very effective in tackling the feature space complexity.

For the BP continuous regression task, we obtained an RMSE of 0.845. The results of four-class classification are shown in Talbe IX and X. The feature space characteristics evaluated based on one-class, one-versus-one pairwise, and one-versus-all multi-class linear-kernel SVM classification are shown in Table XI. It can be seen from Table XI that the

TABLE VIII
WITHIN-CLASS FEATURE REPRESENTATIVENESS AND BETWEEN-CLASS FEATURE DISCRIMINATION IN THE BP BINARY CLASSIFICATION DATABASE.

	Class-1	Class-2
Class-1	0.829	0.533
Class-2	0.640	0.772

TABLE IX
RECALL, PRECISION, F-SCORE AND OVERALL ACCURACY OF BP FOUR-CLASS CLASSIFICATION.

	Class-1	Class-2	Class-3	Class-4
Recall	0.993	0.983	0.969	0.721
Precision	0.980	0.966	0.851	1.000
F-score	0.987	0.974	0.907	0.838
Accuracy	0.949			

TABLE X
CONFUSION MATRIX OF BP FOUR-CLASS CLASSIFICATION.

Ground Truth	Prediction Class-1	Prediction Class-2	Prediction Class-3	Prediction Class-4
Class-1	**0.993**	0.007	0.000	0.000
Class-2	0.017	**0.983**	0.000	0.000
Class-3	0.031	0.000	**0.969**	0.000
Class-4	0.000	0.023	0.256	**0.721**

majority of misclassifications were between nearby classes, which had close continuous labels, and the further classes were well separated. A similar trend was shown in our results also (Table X). Furthermore, in our results, misclassifications were largely reduced, compared to the multi-class classification rates with SVM. It can also been seen that in the pairwise classification there was misclassification between classes 1 and 4 and between classes 2 and 3, while our results show perfect differentiation between the class pairs. These results thus demonstrate the effectiveness of our LMLE model against the intra-class variation and inter-class ambiguity. Between classes 3 and 4, LMLE classified all class 3 images accurately but misclassified 25.6% class 4 images as class 3. This tendency of classifying more images to class 3 was due to the smaller intra-class variation in class 3 compared to class 4 (0.939 vs. 0.850) that it was easier to learn an effective large margin aggregation model for class 3.

Our method was implemented in Matlab, running on a PC with a 2.66-GHz dual core CPU. The majority of execution time was allocated to reference sub-categorization, local estimate generation and learning of transformation matrix. For ILD classification, reference sub-categorization needed about 29 seconds, and learning of transformation matrix needed about 22 minutes. Note that these operations were run offline at the subject-level with four-fold cross validation for matrix learning, hence the total number of runs was small. With additional new test images, the reference subcategories and transformation matrices need not be regenerated either.

TABLE XI
WITHIN-CLASS FEATURE REPRESENTATIVENESS AND BETWEEN-CLASS FEATURE DISCRIMINATION IN THE BP FOUR-CLASS CLASSIFICATION DATABASE.

	Class-1	Class-2	Class-3	Class-4	Multi
Class-1	0.881	0.910	0.960	0.995	0.920
Class-2	0.604	0.824	0.611	0.955	0.517
Class-3	0.832	0.726	0.939	0.797	0.631
Class-4	0.953	0.892	0.663	0.850	0.581

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

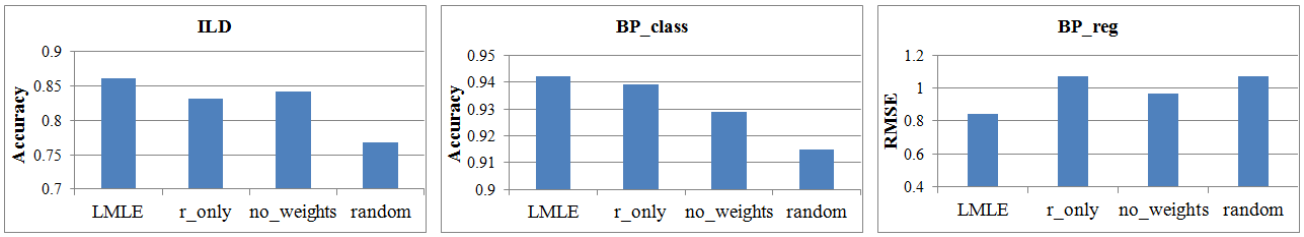

Fig. 4. Accuracy of ILD classification and BP binary classification, and RMSE of BP continuous regression, with the proposed reference sub-categorization (LMLE), without the feature separation factor (r_only), without view and variable weights (no_weights), and random partitioning (random).

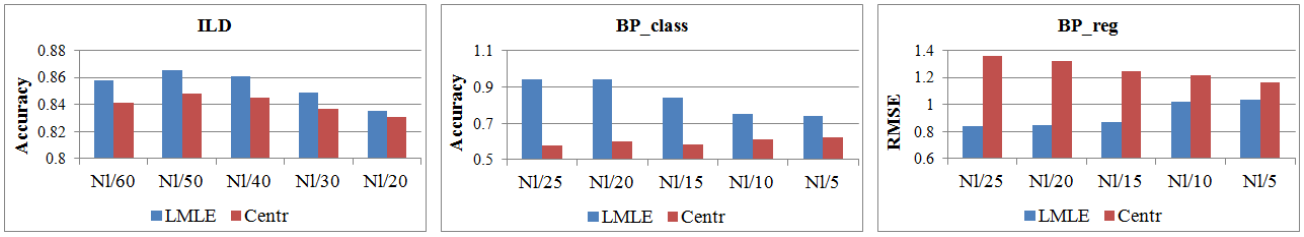

Fig. 5. Accuracy of ILD classification and BP binary classification, and RMSE of BP continuous regression, with the proposed reference sub-categorization (LMLE), and cluster centroids as the local estimates (Centr), with various numbers of subcategories $K_l = \lfloor N_l/60 \rfloor$ to $\lfloor N_l/20 \rfloor$ for the ILD database, and $K_l = \lfloor N_l/25 \rfloor$ to $\lfloor N_l/5 \rfloor$ for the MLC database.

The image-wise classification needed on average 0.4 second, which was mainly spent on local estimate generation. For BP binary classification, reference sub-categorization, learning of transform matrix and image-wise classification required 0.12 second, 1 minute and 0.02 second, respectively. For BP continuous regression, these numbers were 0.27 second, 2 minutes and 0.04 second, respectively. Much less time was required for the two BP tasks, due to the smaller numbers of images compared to the ILD database.

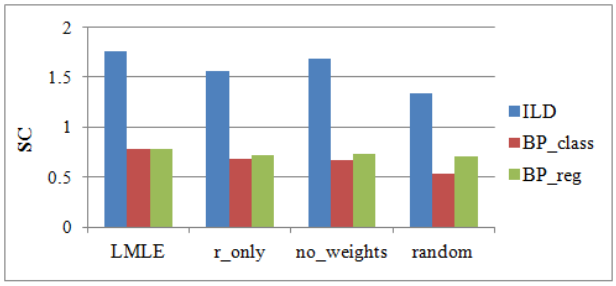

Fig. 6. The SC measure on the three tasks computed with the proposed reference sub-categorization (LMLE), without the feature separation factor (r_only), without view and variable weights (no_weights), and random partitioning (random).

B. Component Analysis

We present our evaluations of the various important components in our LMLE model in the following sentences. First, we evaluated our design of the reference sub-categorization component. In particular, we were interested in analyzing the effects of (i) including feature separation d_l^i into the distance computation Eq. (1); and (ii) including view and variable weights α_l and β_l into the sub-categorization objective Eq. (3). As shown in Fig. 4, our method obtained higher classification recall and precision and lower RMSE, when compared to the alternatives r_only (without feature separation d_l^i) and no_weights (without view and variable weights α_l and β_l). This suggests the benefit of including the feature separation factor and view/variable weights into the sub-categorization method. Fig. 6 shows the SC values computed using Eq. (8). For all three tasks, our reference sub-categorization method provided the highest SC, meaning that the generated subcategories were more compact and better separated than using the alternative methods r_only, no_weights and random.

In addition, to assess if the clustering-based method is necessary for reference sub-categorization, we compared it with random partitioning of the reference data into subcategories. For a thorough evaluation of the random approach, different numbers of subcategories (varied between $\pm$ 50% of the numbers of subcategories used in our LMLE model) were evaluated. We obtained minimal differences in classification performance between these different numbers of subcategories. Multiple runs of random partitioning were conducted to obtain the average performance. As shown in Fig. 4 and 6, our clustering-based approach offered large improvement over the random approach. It was also observed that across the different runs, the classification performance remained relatively consistent for the ILD classification task, but varied to a larger extent for the BP classification and regression tasks. These observations suggest that the resultant subcategory assignments had large impact on the classification performance. With the small MLC database, the intrinsic subcategory structure in the feature space could be under-represented by the limited data, and the random partitioning could sometimes produce similar effects to our clustering-based approach. For the large ILD database, the feature space would form subcategories more naturally; and with the large number of images, it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

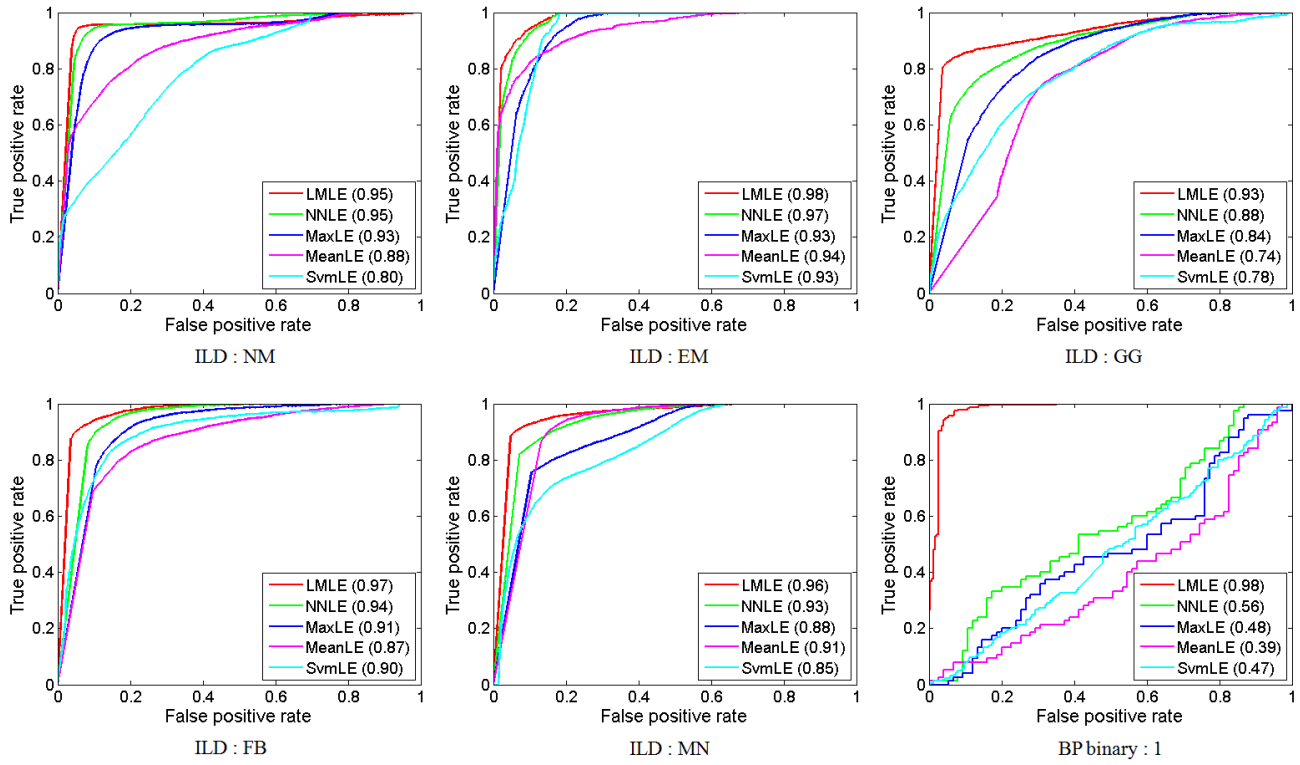

Fig. 7. ROC curves of ILD classification and BP binary classification, comparing our large margin aggregation method (LMLE) with NNLE, MaxLE, MeanLE and SvmLE. The notation after ":" indicates the label of the positive class. The numbers in the brackets indicate the AUC values.

was impractical for the random approach to obtain similar subcategories to the clustering-based approach.

Then, we evaluated our design of using sparse representation to generate the local estimates. In particular, we compared with the more intuitive way of using the cluster/subcategory centroids as the local estimates (Centr), and analyzed the effect of different numbers of subcategories. The experimental results are shown in Fig. 5, where our model with sparse representation for local estimates outperformed the approach using cluster centroids in all three tasks. A trend common to the three tasks was that the performance difference between LMLE and Centr gradually decreased as the number of subcategories increased. This was mainly because with larger numbers of subcategories, each subcategory becomes more homogeneous, and the resultant sparse representation moves towards the cluster centroid (with a scaling factor) leading to similar performance to Centr. We also found that with a smaller number of subcategories, the local estimates generated with sparse representation were indeed more adapted to the test data than the cluster centroid, and the large margin aggregation component worked better with such local estimates. Consequently, LMLE achieved higher performance with relatively small number of subcategories. In addition, for BP classification and regression, the performance of Centr improved with larger number of subcategories, mainly because the cluster centroids became more representative for the more homogeneous subcategories. However for ILD classification, the classification accuracy reduced with larger number of subcategories. We suggest that this could be due to over subcategorization and that data in the overlapping areas of the feature space were clustered into individual subcategories and caused noise in the top M most similar local estimates.

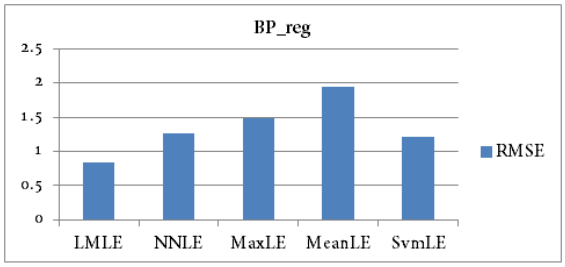

Fig. 8. RMSE of BP continuous regression, comparing our large margin aggregation method (LMLE) with NNLE, MaxLE, MeanLE and SvmLE.

Next, we evaluated our design of the large margin aggregation. Specifically, we compared our LMLE with (i) using kNN to compute the similarity based on the top M similar local estimates (NNLE); (ii) max pooling based on the most similar local estimates from each class (MaxLE); (iii) mean pooling based on the mean distances of the local estimates (MeanLE); and (iv) classification with SVM with the concatenated distances between the test image and local estimates as the feature vector (SvmLE). For NNLE, the parameter M followed the same settings as our LMLE model. For SvmLE, the polynomial kernel performed the best, and the training and testing procedure was the same as ours. We show the various ROC curves for ILD classification and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

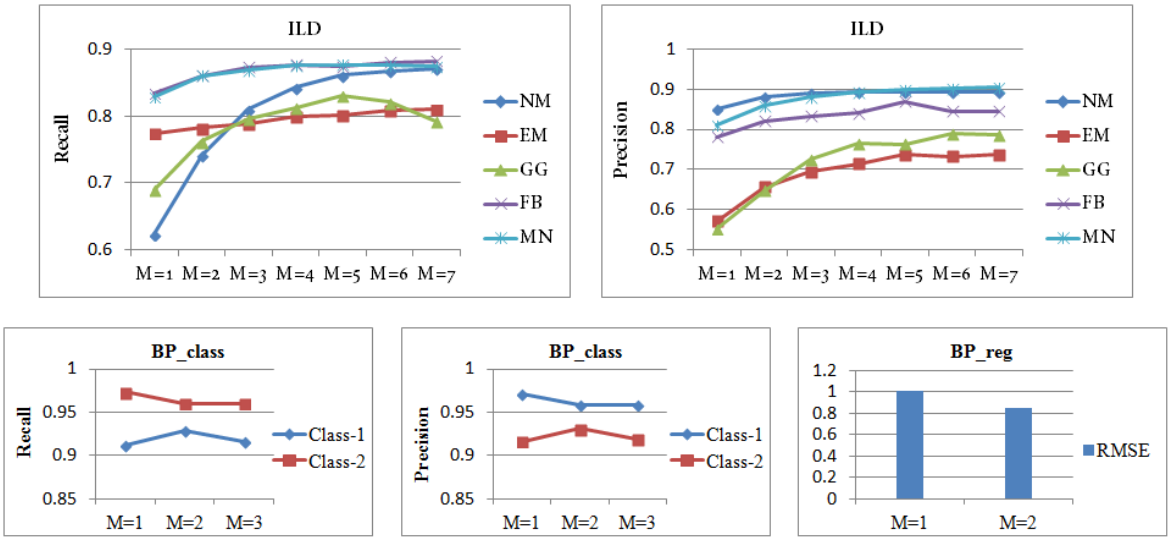

Fig. 9. Recall and precision of ILD classification and BP binary classification, and RMSE of BP continuous regression, with different settings of M.

BP binary classification in Fig. 7, and different RMSEs for BP continuous regression in Fig. 8. Note that since the BP binary classification task involves only two classes, we only show the ROC curves with class 1 as the positive class. Our method achieved the highest AUCs and lowest RMSE. Among the compared approaches, our method can be considered an extension of the NNLE with top similar local estimates identified in the transformed feature space. The performance improvement over NNLE thus demonstrated the benefit of the proposed learning-based large margin aggregation method. In addition, the NNLE approach performed better than the pooling approaches MaxLE and MeanLE. While SvmLE also involved supervised learning based on the local estimates, its performance was unsatisfactory for the two classification tasks. This implies the benefit of the similarity-based classification in our large margin aggregation method. The SvmLE approach was however more suitable than NNLE for the BP continuous regression task with a lower RMSE.

We also evaluated the effect of parameter M, which is the number of closest local estimates for large margin aggregation. Its impact on the classification and regression performance is shown in Fig. 9. For ILD classification, as M increased, the classification recall and precision gradually improved and became relatively stable at $M = 5$. From $M = 5$ onwards, while the recall of EM continued to improve, the recall of GG dropped. Therefore, we chose $M = 5$ for the ILD database. For BP binary classification, there were only three subcategories in each class, hence we tested $M = 1$ to 3 only. At $M = 2$, we obtained more balanced results compared to $M = 1$, and higher recall and precision compared to $M = 3$. Therefore, M was set to 2 for the BP binary classification task. For the BP continuous regression task, the fourth class contained only two subcategories, we thus tested $M = 1$ and $M = 2$ only. It was found that $M = 2$ gave a smaller RMSE.

Finally, to assess the risk of overfitting during learning of the transformation matrices for large margin aggregation, we evaluated the results separately on the training and testing datasets. If the results were excellent on the training sets but poor on the testing sets, overfitting was indicated. We found that compared to the testing set, the performance on the training set was about 5% and 3% higher in accuracy for ILD classification and BP binary classification, and 0.01 lower in RMSE for BP continuous regression. These differences were relatively small, implying that overfitting was not an important concern for these tasks.

C. Performance Comparison

We present the performance comparison with the existing techniques in this section. First, for the ILD classification, we compared our method to the most recent approaches reported for the same ILD database that we used, including the approach based on localized features (LF) [2], our previous patch-adaptive sparse approximation (PASA) method that proposed the TIG feature descriptor [6], and the boosted multifold sparse representation (BMSR) model based on boosted sparse representation [49]. For LF and BMSR, we directly used the results reported in [2] and [49]. However, we note that those results for LF were obtained based on a slightly different selection of images from our work. We re-ran the PASA method on our current database. In addition, we compared our approach to a number of standard and related classification models, including the standard sparse representation classifier (SRC) with L0 regularization, kNN based on Euclidean distances, LMNN with learning-based distance computation based on kNN, and the polynomial-kernel multi-class SVM. These approaches all had the same TIG feature descriptors and training and testing procedures that were used with LMLE.

The differences in recall and precision between our LMLE model and the other approaches are shown in Fig. 10. Positive numbers suggest performance improvements. The results show that our LMLE model achieved large improvements over the standard SRC, kNN, LMNN and SVM classifiers. It was interesting to see that the approaches based on discriminative learning, i.e. LMNN and SVM, did not gain advantage over

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

the simple kNN approach. Such behavior could be attributed to the feature space complexity that the performance of discriminative learning in LMNN and SVM was limited by the large number of contradicting constraints defined by the training samples. The standard SRC model did not perform well for this task either. This was mainly due to the occurrences of better sparse representation for the wrong class, which contained highly similar reference images to the test image. Compared to the approaches specifically designed for ILD classification, i.e. LF, PASA and BMSR, our LMLE model achieved overall higher recall and precision, and more balanced results for different tissue types. The LF method was especially effective in handling the EM images, but were not as effective for the MN images. We obtained considerable improvements in recall levels over PASA and BMSR, but lower precision in several cases. This was due to a different distribution in the misclassification results. For example, compared to PASA, much fewer images were misclassified as EM but more were misclassified as MN. The performance gain over PASA indicates the benefit of our model when compared to the data-driven dictionary adaptation. The overall improvement over BMSR demonstrates the advantage of reference subcategorization over random subdivision and fusion of sub-level results using large margin aggregation over boosting.

We further tested the statistical significance of performance improvement between our model and each approach (excluding LF) using McNemar's test. The test was performed by formulating a 2×2 matrix containing (i) the number of images classified accurately by LMLE and the compared approach, (ii) the number of images classified accurately by LMLE but inaccurately by the compared approach; (iii) the number of images classified inaccurately by LMLE but accurately by the compared approach; and (iv) the number of images classified inaccurately by both approaches. The null hypothesis was that the LMLE model and the compared approach provided equal classification accuracies. We obtained p-value $< 10^{-15}$ in all pairwise comparisons. This thus suggests that our model was significantly better than the other approaches. The LF method was not included in this test since we did not have the image-wise label results.

For the BP binary classification task, we compared our model to BMSR, SRC, kNN, LMNN and SVM. The compared approaches were applied based on the same set of feature descriptors and testing procedures as our model. As shown in Fig. 11, our model obtained large improvements in classification recall and precision over the other approaches. The McNemar's test derived p-values $< 10^{-15}$ suggesting statistical significance as well. The results also show that kNN performed better than the sparse representation-based approaches BMSR and SRC. This could be due to the objective of sparse representation, which aimed at finding optimal combination of reference images to represent the test image. With large intra-class variation and inter-class ambiguity, such an objective could lead to close representation from the wrong classes and possibly more misclassifications than simply combining the most similar reference images. The discriminative LMNN and SVM approaches did not perform well. The intra-class variation, inter-class ambiguity and small number of samples

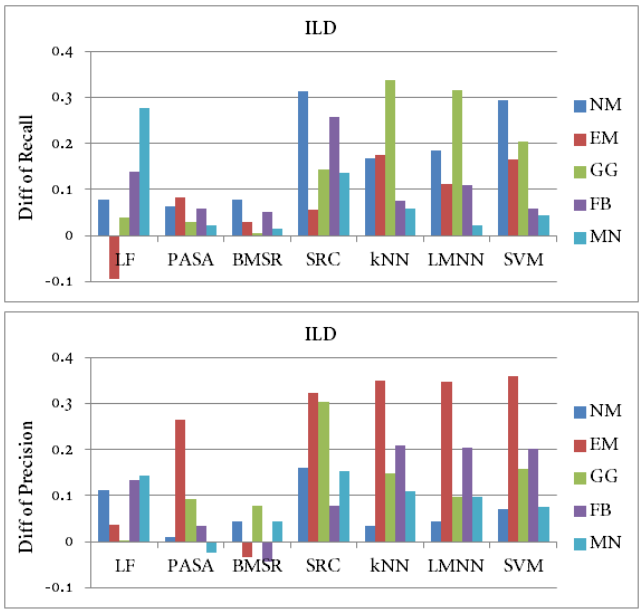

Fig. 10. Differences in recall and precision between our LMLE model and the other various classification methods, for ILD classification.

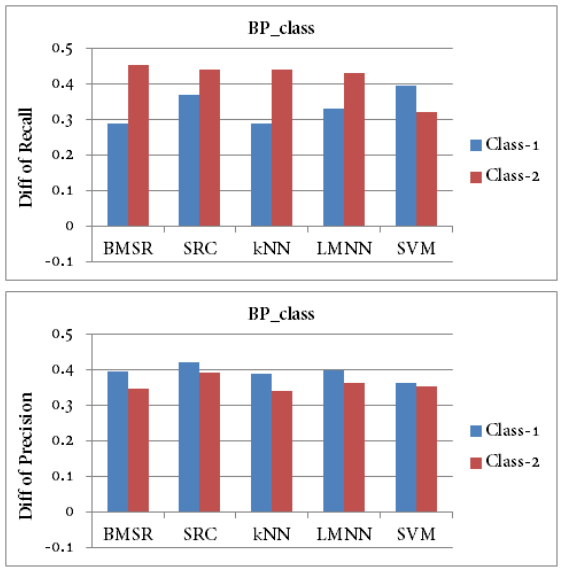

Fig. 11. Differences in recall and precision between our LMLE model and the other various classification methods, for BP binary classification.

relative to the feature length could have restricted the learning capabilities of these algorithms. Our results suggest that the subcategory-based sparse representation and large margin aggregation were particularly effective with this database.

The decrease in RMSE using our model compared to the other approaches for BP continuous regression is shown in Fig. 12. To compute the continuous label using the BMSR, SRC, kNN and LMNN classifiers, four-class classification was first conducted; the continuous labels corresponding to the sparsely selected (for BMSR and SRC) or nearest neighbor (for kNN and LMNN) reference images from the identified class were then combined to generate the continuous label of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2393954, IEEE Transactions on Medical Imaging

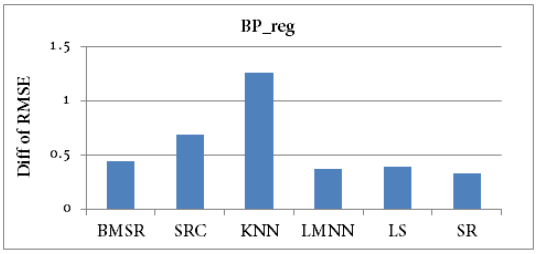

Fig. 12. Differences in RMSE between the other various regression methods and our LMLE model, for BP continuous regression.

the test image. The standard least squares (LS) and sparse representation (SR) were also applied for regression to compute the continuous labels without involving the four-class classification. The results show that our model performed the best. The LMNN approach improved considerably over kNN. When comparing Table XI with Table VIII, it can be seen that the intra-class variation and inter-class ambiguity were smaller in the four-class classification. This could explain why LMNN obtained more obvious performance improvement over kNN for this BP continuous regression task than the binary classification task. As expected, by including the L0 regularization, SR performed better than LS. The classification-based sparse representation approaches BMSR and SRC were not as effective as direct regression using SR. This suggests that sparse representation using the entire reference set was more suitable for this regression task than using subsets of reference images. Such behavior can be explained as follows. In the four-class construct, images with close continuous labels near the class boundaries were separated into different classes; and this separation would limit the flexibility and effectiveness of the class-specific sparse representation (BMSR and SRC) compared to using the entire reference set (SR). Our LMLE model was affected by this issue as well. However, with our subcategory-based sparse representation and large margin fusion, the issue was effectively tackled and highly accurate classification was achieved.

The statistical significance of performance improvement of our model over each approach we compared it to for BP continuous regression was evaluated using the paired t-test. The error vector computed based on the LMLE results was paired with the error vector from the compared approach. The null hypothesis was that the predictions using LMLE had the same mean error as the predictions using the compared approach. A one-tailed test was conducted to determine if our mean error was smaller than the mean error from the compared approach. We obtained a p-value $< 10^{-10}$ in all pairwise comparisons. This indicated that the LMLE model achieved statistically significant improvement over the compared approaches.

V. Conclusions

We present a Large Margin Local Estimate (LMLE) model for medical image classification. By first sub-categorizing the reference sets, the derived reference subcategories would exhibit lower intra-class variation compared to the overall reference sets. Sparse representation is then used to generate the local estimates at the subcategory-level. The distances between the test image and the local estimates are then fused using large margin aggregation to minimize the influence of inter-class ambiguity. The test image is finally classified based on its similarities with the various classes. The LMLE model is independent of the feature design and was applied to ILD classification in lung HRCT images, phenotype classification and regression in brain MR images. Our extensive performance evaluation showed that our model outperformed other often-used classifiers.

References

[1] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 854–869, 2007.

[2] A. Depeursinge, D. V. de Ville, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Muller, "Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 665–675, 2012.

[3] Y. Song, W. Cai, J. Kim, and D. Feng, "A multi-stage discriminative model for tumor and lymph node detection in thoracic images," *IEEE Trans. Med. Imag.*, vol. 31, no. 5, pp. 1061–1075, 2012.

[4] F. Zhang, Y. Song, W. Cai, Y. Zhou, S. Shan, and D. Feng, "Context curves for classification of lung nodule images," *in DICTA*, pp. 1–7, 2013.

[5] G. Lee, S. Ali, R. Veltri, J. I. Epstein, C. Christudass, and A. Madabhushi, "Cell orientation entropy (core): predicting biochemical recurrence from prostate cancer tissue microarrays," *in MICCAI*, pp. 396–403, 2013.

[6] Y. Song, W. Cai, Y. Zhou, and D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 797–808, 2013.

[7] F. Zhang, Y. Song, W. Cai, M. Lee, Y. Zhou, H. Huang, S. Shan, M. Fulham, and D. Feng, "Lung nodule classification with multi-level path-based context analysis," *IEEE Trans. Bio-Med. Eng.*, vol. 61, no. 4, pp. 1155–1166, 2014.

[8] S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," *Medical Image Analysis*, vol. 16, no. 7, pp. 1385–1396, 2012.

[9] Y. Song, W. Cai, H. Huang, Y. Wang, and D. Feng, "Object localization in medical images based on graphical model with contrast and interest-region terms," *in Proc. CVPRW*, pp. 1–7, 2012.

[10] S. Zhang, X. Li, J. Lv, X. Jiang, D. Zhu, H. Chen, T. Zhang, L. Guo, and T. Liu, "Sparse representation of higher-order functional interaction patterns in task-based fMRI data," *in MICCAI*, pp. 626–634, 2013.

[11] N. Weiss, D. Rueckert, and A. Rao, "Multiple sclerosis lesion segmentation using dictionary learning and sparse coding," *in MICCAI*, pp. 735–742, 2013.

[12] S. Liu, W. Cai, Y. Song, S. Pujol, R. Kikinis, L. Wen, and D. Feng, "Localized sparse code gradient in alzheimer's disease staging," *in Proc. EMBC*, pp. 5398–5401, 2013.

[13] F. Liu, H. Suk, C. Wee, H. Chen, and D. Shen, "High-order graph matching based feature selection for alzheimer's disease identification," *in MICCAI*, pp. 311–318, 2013.

[14] E. Parrado-Hernandez, V. Gomez-Verdejo, M. Martinez-Ramon, J. Shawe-Tylor, P. Alonso, J. Pujol, J. M. Menchon, N. Cardoner, and C. Soriano-Mas, "Discovering brain regions relevant to obsessive-compulsive disorder identification through bagging and transduction," *Medical Image Analysis*, vol. 18, no. 3, pp. 435–448, 2014.

[15] M. Yaqub, M. K. Javaid, C. Cooper, and J. A. Noble, "Investigation of the role of feature selection and weighted voting in random forests for 3-D volumetric segmentation," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 258–271, 2014.

[16] H. Chang, Y. Zhou, P. Spellman, and B. Parvin, "Stacked predictive sparse coding for classification of distinct regions in tumor histopathology," *in Proc. ICCV*, pp. 169–176, 2013.

[17] H. Suk and D. Shen, "Deep learning-based feature representation for AD/MCI classification," *in MICCAI*, pp. 583–590, 2013.

[18] C. Becker, R. Rigamonti, V. Lepetit, and P. Fua, "Supervised feature learning for curvilinear structure segmentation," *in MICCAI*, pp. 526–533, 2013.

[19] Y. Song, W. Cai, S. Huh, M. Chen, T. Kanade, Y. Zhou, and D. Feng, "Discriminative data transform for image feature extraction and classification," *in MICCAI*, pp. 452–459, 2013.

[20] Q. Li, W. Cai, and D. Feng, "Lung image patch classification with automatic feature learning," *in Proc. EMBC*, pp. 6079–6082, 2013.

[21] C. Jacobs, C. I. Sanchez, S. C. Saur, T. Twellmann, P. A. de Jong, and B. van Ginneken, "Computer-aided detection of ground glass nodules in thoracic CT images using shape, intensity and context features," *in MICCAI*, pp. 207–214, 2011.

[22] Q. Zhao, K. Okada, K. Rosenbaum, L. Kehoe, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru, "Digital facial dysmorphology for genetic screening: hierarchical constrained local model using ICA," *Medical Image Analysis*, vol. 18, no. 5, pp. 699–710, 2013.

[23] Y. Song, W. Cai, H. Huang, Y. Wang, D. Feng, and M. Chen, "Region-based progressive localization of cell nuclei in microscopic images with data adaptive modeling," *BMC Bioinformatics*, vol. 14, no. 173, pp. 1–16, 2013.

[24] L. Sorensen, M. Nielsen, P. Lo, H. Ashraf, J. H. Pedersen, and M. de Bruijne, "Texture-based analysis of COPD: a data-driven approach," *IEEE Trans. Med. Imag.*, vol. 31, no. 1, pp. 70–78, 2012.

[25] D. Liu and S. K. Zhou, "Anatomical landmark detection using nearest neighbor matching and submodular optimization," *in MICCAI*, pp. 393–401, 2012.

[26] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "Non-melanoma skin lesion classification using colour image data in a hierarchical k-NN classifier," *in Proc. ISBI*, pp. 358–361, 2013.

[27] Y. Song, W. Cai, Y. Zhou, L. Wen, and D. Feng, "Pathology-centric medical image retrieval with hierarchical contextual spatial descriptor," *in Proc. ISBI*, pp. 202–205, 2013.

[28] G. Xiao and A. Madabhushi, "Aggregated distance metric learning (ADM) for image classification in presence of limited training data," *in MICCAI*, pp. 33–40, 2011.

[29] M. Liu, L. Lu, J. Bi, V. Raykar, M. Wolf, and M. Salganicoff, "Robust large scale prone-supine polyp matching using local features: a metric learning approach," *in MICCAI*, pp. 75–82, 2011.

[30] B. Andre, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1276–1288, 2012.

[31] T. Syeda-Mahmood, F. Wang, R. Kumar, D. Beymer, Y. Zhang, R. Lundstrom, and E. McNulty, "Finding similar 2D x-ray coronary angiograms," *in MICCAI*, pp. 501–508, 2012.

[32] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi, "Neighbourhood approximation using randomized forests," *Medical Image Analysis*, vol. 17, no. 7, pp. 790–804, 2013.

[33] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[34] L. Wu and S. C. H. Hoi, "Enhancing bag-of-words models with semantics-preserving metric learning," *IEEE MultiMedia*, vol. 18, no. 1, pp. 24–37, 2011.

[35] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[36] M. Liu, L. Lu, X. Ye, S. Yu, and M. Salganicoff, "Sparse classification for computer aided diagnosis using learned dictionaries," *in MICCAI*, pp. 41–48, 2011.

[37] X. Huang, D. P. Dione, C. B. Compas, X. Papademetris, B. A. Lin, A. Bregasi, A. J. Sinusas, L. H. Staib, and J. S. Duncan, "Contour tracking in echocardiographic sequences via sparse representation and dictionary learning," *Medical Image Analysis*, vol. 18, no. 2, pp. 253–271, 2014.

[38] Y. Xu, X. Gao, S. Lin, D. W. K. Wong, J. Liu, D. Xu, C. Cheng, C. Y. Cheung, and T. Y. Wong, "Automatic grading of nuclear cataracts from slit-lamp lens images using group sparsity regression," *in MICCAI*, pp. 468–475, 2013.

[39] Y. Song, W. Cai, H. Huang, X. Wang, Y. Zhou, M. Fulham, and D. Feng, "Lesion detection and characterization with context driven approximation in thoracic FDG PET-CT images of NSCLC studies," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 408–421, 2014.

[40] Y. Song, W. Cai, Y. Zhou, and D. Feng, "Thoracic abnormality detection with data adaptive structure estimation," *in MICCAI*, pp. 74–81, 2012.

[41] T. Tong, R.Wolz, P. Coupe, J. V. Hajnal, D. Rueckert, and ANDI, "Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling," *NeuroImage*, vol. 76, pp. 11–23, 2013.

[42] U. Srinivas, H. S. Mousavi, V. Monga, A. Hattel, and B. Jayarao, "Simultaneous sparsity model for histopathological image representation and classification," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1163–1179, 2014.

[43] L. Wang, F. Shi, Y. Gao, G. Li, J. H. Gilmore, W. Lin, and D. Shen, "Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation," *NeuroImage*, vol. 89, pp. 152–164, 2014.

[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *in Proc. CVPR*, pp. 3360–3367, 2010.

[45] P. Zhang, C. Wee, M. Nieghammer, D. Shen, and P. Yap, "Large deformation image classification using generalized locality-constrained linear coding," *in MICCAI*, pp. 292–299, 2013.

[46] L. Gorelick, O. Veksler, M. Gaed, J. A. Gomez, M. Moussa, G. Bauman, A. Fenster, and A. D. Ward, "Prostate histopathology: learning tissue component histograms for cancer detection and classification," *IEEE Trans. Med. Imag.*, vol. 32, no. 10, pp. 1804–1818, 2013.

[47] P. Chatelain, O. Pauly, L. Peter, S. Ahmadi, A. Plate, K. Botzel, and N. Navab, "Learning from multiple experts with random forests: application to the segmentation of the midbrain in 3D ultrasound," *in MICCAI*, pp. 230–237, 2013.

[48] D. Allen, L. Lu, J. Yao, J. Liu, E. Turkbey, and R. M. Summers, "Robust automated lymph node segmentation with random forests," *SPIE Med. Imaging*, p. 90343X, 2014.

[49] Y. Song, W. Cai, H. Huang, Y. Zhou, Y. Wang, and D. Feng, "Boosted multifold sparse representation with application to ILD classification," *in Proc. ISBI*, pp. 1023–1026, 2014.

[50] M. Zhu and A. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, 2006.

[51] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Linear subclass support vector machines," *IEEE Signal Process. Lett.*, vol. 19, no. 9, pp. 575–578, 2012.

[52] M. Hoai and A. Zisserman, "Discriminative sub-categorization," *in Proc. CVPR*, pp. 1666–1673, 2013.

[53] S. Escalera, D. M. J. Tax, O. Pujol, P. Radeva, and R. P. W. Duin, "Subclass problem-dependent design for error-correcting output codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1041–1054, 2008.

[54] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan, "Subcategory-aware object classification," *in Proc. CVPR*, pp. 827–834, 2013.

[55] Y. Song, W. Cai, H. Huang, Y. Zhou, D. Feng, and M. Chen, "Large margin aggregation of local estimates for medical image classification," *in MICCAI*, pp. 196–203, 2014.

[56] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, "TW-k-Means: automated two-level variable weighting clustering algorithm for multiview data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 932–944, 2013.

[57] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[58] J. Liu, S. Ji, and J. Ye, "SLEP: sparse learning with efficient projections," *Arizona State University*, 2009. [Online]. Available: http://www.public.asu.edu/ jye02/Software/SLEP/

[59] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, 2004.

[60] L. Vandenberghe and S. P. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[61] W. R. Webb, N. L. Muller, and D. P. Naidich, *High-resolution CT of the lung*. Lippincott Williams Wilkins, 2008.

[62] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Muller, "Building a reference multimedia database for interstitial lung diseases," *Comput. Med. Imaging Graph.*, vol. 36, no. 3, pp. 227–238, 2012.

[63] MICCAI 2014 machine learning challenge. [Online]. Available: https://www.nmr.mgh.harvard.edu/lab/laboratory-computational-imaging-biomarkers/miccai-2014-machine-learning-challenge

[64] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, pp. 774–781, 2012.

[65] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelli. Syst. Technol.*, vol. 2, pp. 1–27, 2011.