Efficient l_p-Norm Multiple Feature Metric Learning for Image Categorization

Shuhui Wang[1] Qingming Huang[1, 2] Shuqiang Jiang[1] Qi Tian[3]

[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
{shwang,sqjiang}@jdl.ac.cn

[2]Graduate University, Chinese Academy of Sciences, Beijing, 100049, China
qmhuang@jdl.ac.cn

[3]Dept. of Computer Science, Univ. of Texas at San Antonio, TX78249, U.S.A.
qitian@cs.utsa.edu

ABSTRACT

Previous metric learning approaches are only able to learn the metric based on single concatenated multivariate feature representation. However, for many real world problems with multiple feature representation such as image categorization, the model trained by previous approaches will degrade because of sparsity brought by significant dimension growth and uncontrolled influence from each feature channel. In this paper, we propose an efficient distance metric learning model which adapts Distance Metric Learning on multiple feature representations. The aim is to learn the *Mahalanobis* matrices for each independent feature and their non-sparse l_p-norm weight coefficients simultaneously by maximizing the margin of the overall learned distance metric among the pairs from the same class and the distance of pairs from different classes. We further extend this method to nonlinear kernel learning and category specific metric learning, which demonstrate the applicability of using many existing kernels for image data and exploring the hierarchical semantic structures for large scale image datasets. Experiments on various datasets demonstrate the promising power of our method.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology — *Classifier design and evaluation*

General Terms: Algorithms, Performance, Experimentation

1. INTRODUCTION

A good distance metric stands in the core for many learning methods, for example, the kernel methods and nearest neighbor method. A commonly used distance metric is the *Euclidean* distance, a choice which has both the advantages of simplicity and generality. Despite of these advantages, for many real world classification tasks, the *Euclidean* distance is not well adapted. Metric learning is an emerging area of statistical learning in which the goal is to learn a more powerful distance metric from a set of labeled samples or weakly labeled samples. With a learned distance metric, even the simplest lazy learning method such as *k*-NN can achieve good generalization power. In fact, significant improvements have been observed within several distinguished solutions for this problem, such as neighborhood components analysis [6], large margin *k*-NN classification [15], and information theoretic metric learning [3].

For many real world applications, there will be multiple descriptions for a single data item. For example, Web image can be represented by shape feature, texture feature, color feature and textual feature extracted from the surrounding text. These features may have heterogeneous statistical characteristics. Previously, to make use of multiple features, a common way is concatenating all the features into one single feature vector. Then a classification model is obtained based on this concatenated representation. However, this will lead to explosive dimension increases, and the contribution of each original feature channel cannot be controlled. Moreover, the correlation among features usually contains noisy information that may lead to unpredictable risk of model degradation and over-fitting. For metric learning, the *Mahalanobis* matrix usually have $O(k^2)$ coefficients compared with $O(k)$ for linear SVM, where k denotes the average dimension of each feature representation. Therefore, the significant increase of computation as well as over-fitting brought by metric learning with concatenated feature are two major intrinsic weaknesses for applying previous metric learning approaches to application with multiple feature representation.

Recently, Multiple Kernel Learning [1][10][11][12] has been proposed to solve the problem of learning with multiple features and kernel representations. The aim is to maximize the margin on the training data and approximately find sparse weights to calculate the similarity using several features and kernels. Kloft *et al.* [8] and Vishwanathan *et al*. [14] proposed l_p-MKL, which imposes arbitrary non-sparse l_p norm on the kernel weight coefficients. Promising results and good generalization power have been reported by their studies. In this paper, borrowing the idea from l_p-MKL [14], we try to learn an overall combined distance metric by simultaneously optimizing a set of distance metrics on each independent feature channel. Our model requires only $O(Mk^2)$ coefficients to optimize instead of $O(M^2k^2)$ for metric learning with concatenated features, so the *curse-of-dimensionality* problem will be alleviated. We formulated our problem into the max-margin framework. The *Frobenius* norm and l_p squared norm are imposed on each *Mahalanobis* matrix and the weight coefficients of feature channels respectively.

By considering the cluttered distribution in feature space of the image data, we make two extensions of our method so that it is capable of real world image categorization. Firstly, the model can be easily kernelized as the inner product of each feature channel is preserved. Therefore, the existing image kernels are applicable to the specific features, such as spatial pyramid kernel and χ^2 kernel for visual bag-of-word feature. Secondly, we adapt our method to learn categorical specific metrics. In this paper, we

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10...$10.00.

study the capability of our localized metric learning method so as to provide a guidance to reach a good balance of computational complexity and performance for large scale image applications.

In this paper, we propose an efficient multiple feature distance metric learning method by considering all the mentioned issues. We will introduce our work as follows: We describe our distance metric model in Section 2. Experiments are conducted and discussed in Section 3. In Section 4 we summarize the paper.

2. METHOD

2.1 Multiple Feature Metric Learning

Suppose we are given a set of N training samples with class labels, denoted by $\{(\mathbf{x}_i, c_i) \mid i = 1,...,N\}$, where $\mathbf{x}_i$ represents a training sample and c_i denotes its class label. For each $\mathbf{x}_i$, we calculate M different types of features, denoted by $x_i^{(m)}, m = 1,...,M$. For each data pair $(\mathbf{x}_i, \mathbf{x}_j)$, their original distance and learned distance metric with respect to different features are represented by $\bar{d}_{i,l}^{(m)}$ and $\tilde{d}_{i,j}^{(m)}$. In this paper, we define the overall distance metric and similarity between any data pairs as:

$$\begin{aligned} \tilde{d}_{i,j} &= \sum_{m=1}^{M} \tilde{d}_{i,j}^{(m)},\ \tilde{d}_{i,j}^{(m)} = (x_i^{(m)} - x_j^{(m)})^t A^{(m)} (x_i^{(m)} - x_j^{(m)}) \\ \tilde{K}_{i,j} &= \sum_{m=1}^{M} \tilde{K}_{i,j}^{(m)},\ \tilde{K}_{i,j}^{(m)} = (x_i^{(m)})^t A^{(m)} x_j^{(m)} \end{aligned} \tag{1}$$

Where $A^{(m)}$ denotes the *Mahalanobis* matrix for the m^{th} each feature. The training data is represented by a set of triplets as:

$$P^3 : (i, j, l),\ s.t.\ c(i) = c(j),\ c(i) \neq c(l) \tag{2}$$

We introduce a non-negative weight coefficient b_m on each feature representation as well as a set of initial *Mahalanobis* matrices $A_0^{(1)},...,A_0^{(M)}$. Borrowing the ideas from structural risk minimization and [14], we introduce the following convex objective function as:

$$\begin{aligned} &\min_{\mathbf{b},\mathbf{A}} \ \frac{1}{2}\sum_{m=1}^{M} \frac{1}{b_m} \| A^{(m)} - A_0^{(m)} \|_{Fro}^2 + C\sum_{ijl} \xi_{ijl} + \frac{\eta}{2} \| \mathbf{b} \|_p^2 \\ &s.t. \quad \tilde{d}_{i,l} - \tilde{d}_{i,j} \geq 1 - \xi_{ijl},\ \xi_{ijl} \geq 0, b_m \geq 0, p > 1,\ A^{(m)} \succeq 0 \end{aligned} \tag{3}$$

Where the original *Mahalanobis* matrices $A_0^{(1)},...,A_0^{(M)}$ can be a set of identity matrix, or it can be any *Mahalanobis* matrices learned on some other dataset. Following the similar derivation of [14], we obtain the dual problem as:

$$\min_{\boldsymbol{\alpha}} R(\boldsymbol{\alpha}) = \frac{1}{8\eta}\left(\sum_{m=1}^{M} \left(\boldsymbol{\alpha}^t \mathbf{Q}^{(m)} \boldsymbol{\alpha}\right)^q\right)^{2/q} - (\mathbf{1} - \mathbf{s})^t \boldsymbol{\alpha},\ s.t.\ 0 \leq \alpha_{ijl} \leq C \tag{4}$$

Where $1/p + 1/q = 1$, $\mathbf{1}$ denotes a vector with all ones, and:

$$s_{ijl} = \bar{d}_{i,l} - \bar{d}_{i,j} = \sum_{m=1}^{M} \left(\bar{d}_{i,l}^{(m)} - \bar{d}_{i,j}^{(m)}\right),\ Q_{ijl,i'j'l'}^{(m)} = tr\left(\left(\hat{x}_{ijl}^{(m)}\right)^t * \hat{x}_{i'j'l'}^{(m)}\right)$$

$$\begin{aligned} \hat{x}_{ijl}^{(m)} &= \left(x_j^{(m)} - x_l^{(m)}\right)\left(x_i^{(m)}\right)^T + x_l^{(m)}\left(x_l^{(m)}\right)^T \\ &\quad + x_i^{(m)}\left(x_j^{(m)} - x_l^{(m)}\right)^T - x_j^{(m)}\left(x_j^{(m)}\right)^T \end{aligned}$$

For p>1, b_m has the following form:

$$b_m = \frac{1}{2\eta} o_m^{q-1} \left(\sum_{m'=1}^{M} \left(o_{m'}\right)^q\right)^{2/q-1},\ o_m = \boldsymbol{\alpha}^t \mathbf{Q}^{(m)} \boldsymbol{\alpha} \tag{5}$$

In this paper, $A^{(m)}$ is not guaranteed to be positive semi-definite. Although in most case we observe that there is no negative eigenvalues for $A^{(m)}$, one can set the negative eigenvalues to be 0 to make $A^{(m)}$ to be positive semi-definite. In future study, we will investigate the positive semi-definiteness of the learned metric.

2.2 Metric Learning with Kernel

The previous description can be easily kernelized by replacing all $\mathbf{x}$ by $\phi(x)$, where ϕ is the feature map corresponding to any given kernel. We denote the original kernel for each feature channel as $K^{(m)}$, then the learned kernel is given by:

$$\begin{aligned} &\tilde{K}(x_a, x_b) = \sum_{m=1}^{M} K^{(m)}(x_a, x_b) + \sum_{m=1}^{M}\sum_{t'=1}^{T} \rho_{ij} b^{(m)} \bullet \\ &\left(K^{(m)}(x_a, x_i) - K^{(m)}(x_a, x_j)\right)\left(K^{(m)}(x_b, x_i) - K^{(m)}(x_b, x_j)\right) \end{aligned} \tag{6}$$

Where ρ_{ij} is calculated by reorganizing the support vector representation into the weighted combination of training pairs.

2.3 Category Specific Metric Learning

Since a single metric is not robust to deal with real world image categorization with hundreds to thousands of classes, a better way is to learn multiple metrics where a subset of classes share a single metric. We use an automatic hierarchical grouping method according to the overall average similarity on multiple feature representations among different classes. We calculate the average multiple feature representations on the images from each class, so that each semantic class will have an average image feature on the concatenated feature representation. We conduct hierarchical clustering on these feature representations. Then a hierarchical visual semantic structure can be obtained. We use different level of the hierarchical structure from this hierarchy so that each metric corresponds to different category structures and different average number of classes.

2.4 Solver

Since the dual problem (4) is differentiable with respect to $\boldsymbol{\alpha}$, there are many possible solutions that can be used to solve this problem. We propose a special purpose solver based on the recent proposed coordinate gradient descent method [7], which applies Newton-Raphson method to solve a one-variable sub-problem each time until convergence. Moreover, we implement two acceleration schemes proposed by [7], namely, random permutation of sub-problems and shrinking with gradient thresholding. Due to limited space, we omit the details of the optimization process. Readers may refer to our technical report.

2.5 Training Triplet Generation

Our method requires that the data should be organized as pairs, it leads to O(N^2) scale in terms of training data size. However, as has been discussed in literatures such as [6], the neighborhood of each data is most influential to the model. Thus using this heuristic reduces the training data size to O(N) scale. For the linear case in our method, we adopt Locality Sensitive Hashing [4] for approximated nearest neighbor search. For the kernel version, we adopt the KLSH [9]. We build KLSH for each feature channel, and the overall hash code for each image is the concatenation of the hash code vector from all feature channels.

3. EXPERIMENTS

We conduct experiments on 3 different image dataset, the Yale Face B face recognition dataset, the NUS-WIDE-OBJ [2] and the ImageNet-250 dataset [16] which is composed of 250 classes of images covering many common visual concepts including various kinds of categories. Each image category in ImageNet-250 contains more than 500 images. For Yale Face B, We extract 4 types of image features on this dataset, the original gray level feature, the LBP feature on the original image, the Gabor feature from 40 filters with different scales and orientations, and the Gabor LBP feature. For NUS-WIDE-OBJ, five image features are provided by [2] including color features (CM, CH, CORR), edge

histogram (EDH) and wavelet features (WT). We calculate 9 types of features and image kernels for ImageNet-250. Note that in linear cases, we reduce the dimensionality on each single feature with dimensionality higher than 300 by using PCA. The ratios of random training/testing split for the three datasets are 0.7/0.3, *default* and 0.6/0.4, respectively. We conduct evaluation on 10 different random splits for each dataset, and all the reported results are the averaging of all the data splits as done in [15].

We denote our linear and kernelized multiple feature metric learning method as MFDML-L and MFDML-K. We compare our methods with 6 baseline approaches using concatenated feature representation on the three datasets: (1) EUC: the Euclidean metric on the concatenated feature. (2) EUC-PCA: the Euclidean metric on the concatenated feature representation after dimensional reduction using PCA. (3) ITML [3]. (4) LFDA [13]: localized FDA. (5) LMNN: Large Margin Nearest Neighbor method [15]. (6) NCA [6]. We implement our code using Matlab. The computing devices include a desktop computer with Intel i5 760 CPU and 4G RAM and another desktop with Intel Core2 E7500 and 4G RAM.

Another important issue is the setting of model parameters. Although the optimal setting can be found by cross validation, we notice that the performance is not very sensitive to different settings. For all the experiments, we set $A_0^{(m)} = I$, C=5, and $\eta = 1$. We heuristically set p=2 for all the experiments as [14] shows that this setting provides good solution

3.1 Evaluation Criteria

We adopt *Mean Accuracy* (*MA*) which records the average percent of correct predictions among all the classes. We use *k*-NN as the learning and predicting model. Suppose the number of returned nearest neighbors for each query is N_R , we denote the number of images from each category in the retrieved nearest neighbors as a vector $[N_1,...,N_C]$ and $N_R = \sum N_q$. Since multiple category specific metrics are learned for different subset of classes, we denote the task of learning each metric as $\mathbf{T}_t$. The decision output of query $\boldsymbol{x}$ is calculated as follows:

$$Q = \arg\max_q \frac{1}{|N_q|} \sum_{j=1}^{|N_q|} \widetilde{K}_t(\mathbf{x}, \mathbf{x}_t^j)|_{j\in q,\ q\in \mathbf{T}_t} \quad q \in [1, C] \tag{7}$$

Where $\widetilde{K}_t(\mathbf{x}, \mathbf{x}_j)|_{j\in q,\ q\in \mathbf{T}_t}$ denotes the learned similarities between $\boldsymbol{x}$ and the q^{th} class samples $\mathbf{x}_j$. Because class q belongs to the t^{th} task, we use the task specific similarity $\widetilde{K}_t$.

3.2 Yale Face Data B

In this part, we conduct single metric MFDML-L and MFDML-K since there is no semantic structure existing in the face data. The results of the performance on test data using with different number of nearest neighbors are demonstrated in Figure 1. We observe that for almost all the best recognition rate was achieved when the number of nearest neighbors is set to be 3. When the nearest number increases, the performance will degrades possibly because the training data only contains less than 2K training images. A slightly different result is observed in our approach MFDML-L, where the highest recognition rate is achieved when the neighborhood size is 7. The best performance of our method MFDML-K slightly outperforms the state-of-the-art LMNN. Another observation for Figure 1 is that the performance of our method decreases much slowly compared with other approaches, which demonstrates the robustness of our method on the setting of nearest neighbors. Finally, the highest performance for the baselines is 98.6% by LMNN, which outperforms the recognition rate 95.95% reported in [15] which uses the pixel values as the feature. The highest performance for our method MFDML-K reaches 98.98%. The results strongly support the claim that distance metric learning with multiple features usually achieves better generalization than single feature for image classification.

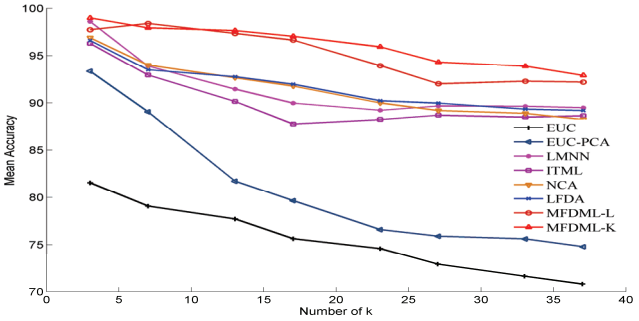

Figure 1: The recognition accuracies on Yale Face B dataset with respect to different numbers of nearest neighbors.

3.3 Number of Metrics and Kernels

We study how our multiple metrics improves the performances with different number of kernels. To this end, we conduct experiment on ImageNet-250 datasets. We test MFDML-K for this experiment. We use the data driven hierarchical clustering method as introduced in Section 2.4 to form multi-level category group structures. The performance with different number of metrics and kernels are shown in Table 1. We demonstrate the results using different number of metrics (row) and number of features (column). We see that when using more features, the performance is improved. This observation is consistent with many previous MKL studies. For the semantic categorization on ImageNet-250 dataset, when the number of metrics increases, the performance is improved. When the number of metrics is small, increasing the number of metrics will lead to significant improvements. When the number of metrics is more than 62, increasing the number of metrics lead to slight improvement. In fact, for large scale image applications, one can seek a better tradeoff between performance and efficiency by using different number of metrics.

Table 1: The performance of MFDML-K with different number of metrics and features on ImageNet-250

	2	4	6	9
1	0.334	0.354	0.361	0.372
4	0.365	0.368	0.387	0.410
16	0.378	0.381	0.410	0.433
62	0.384	0.398	0.431	0.453
250	0.387	0.412	0.441	0.464

3.4 NUS-WIDE-OBJ

In this experiment, the recognition rates of the best single feature (wavelet feature) and using all the features are shown in Table 3. We observe an interesting phenomenon that for some feature such as EDH, the distance metric learned by LMNN or ITML underperform the simplest Euclidean metric. The model trained by LMNN and ITML on the concatenated feature representation even underperform the model trained by wavelet feature (WT), which shows the inflexibility of previous DML approaches on learning with heterogeneous feature. However, our method achieves remarkable improvement at almost all the situations except that when the neighborhood size is 3, LMNN achieves the highest performance by using WT. This result proves the

effectiveness of our distance metric learning when facing with multiple features, especially on the classification problems on Web image. When the number of nearest neighbors k increases, the performance is likely to be enhanced in the experiment because of the complicated neighborhood structure. We observe a decrease in the improvement rate for larger neighborhood size. Therefore, there will be a neighborhood size k_0 around 20 that our methods and other baseline approaches achieve the best results.

3.5 ImageNet-250

In this section, we conduct experiments to compare our methods with several state-of-the-art similarity learning and distance metric learning approaches mentioned in previous sections on visual semantic categorization. We denote learning category specific metrics for each class using LMNN by st-LMNN, and learning one unified metric using LMNN as u-LMNN. We conduct our category specific linear metric learning and kernel learning method for each visual category, which corresponds to 250 learned metrics. The results are shown in Table 2. We see that MFDML-K achieves the highest performance by using all the features. We see from both the performance of st-LMNN and our approaches that multiple metrics are indeed helpful for enhancing the performance. MFDML-L does not significantly outperform st-LMNN because we conduct PCA to avoid dimension explosion, which is an important pre-processing technique required by many existing linear distance metric learning method. In general, our methods achieve promising results on real world image semantic categorization, especially when using image kernels.

Table 2: The accuracy on ImageNet-250 data

EUC	EUC-PCA	st-LMNN	u-LMNN	NCA
0.192	0.264	0.367	0.324	0.315
ITML	LFDA	MFDML-L	MFDML-K	
0.298	0.305	0.373	**0.464**	

4. CONCLUSION

In this paper, we address the problem of metric learning with multiple features. We propose a new distance metric learning method which learns an overall distance metric by optimizing a set of *Mahalanobis* matrix for several feature representations at one time. We made two extensions to facilitate our method with learning with real world image data. The first is kernelizing the model so that it can take advantage of many existing image kernels, and the second is to learn multiple metrics according to a pre-computed semantic structure. The experiments on three image datasets show remarkable performance on the effectiveness and efficiency of our method compared with several state-of-the-art DML approaches. For future study, we will study how to combine our method with the existing LSH technique for efficient metric learning and the applications of retrieval.

5. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China: 61025011, 60833006 and 61070108, in part by Beijing Natural Science Foundation: 4092042 and 4111003, and in part by NSF IIS 1052851, Google Faculty Research Award, gift grants from FXPAL, and NEC Labs of America, Cupertino, CA, to Dr. Qi Tian, respectively.

6. REFERENCES

[1] F. Bach, G. Lanckriet, and M. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. *ICML*, 2004.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. "NUS-WIDE: A Real-World Web Image Database from National University of Singapore", *ACM International Conference on Image and Video Retrieval*. Greece. Jul. 8-10, 2009

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. *ICML*, 2007.

[4] M. Charikar. Similarity Estimation Techniques from Rounding Algorithms. In *ACM Symposium on Theory of Computing*, 2002.

[5] A. Globerson, S. Roweis. Metric learning by collapsing classes. *NIPS*, 2006.

[6] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov. Neighborhood Component Analysis. *NIPS*, 2005.

[7] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi and S. Sundararajan. A dual coordinate descent method for large scale linear SVM. *ICML*, 2008.

[8] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K. R. Muller, A. Zien. Efficient and Accurate l_p norm Multiple Kernel Learning. *NIPS*, 2009.

[9] B. Kulis, and K. Grauman. Kernelized Locality Sensitive Hashing for Scalable Image Search. *ICCV*, 2009.

[10] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui and M. Jordan. Learning the Kernel Matrix with Semi-definite Programming. *JMLR*, 5:27–72, 2004.

[11] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

[12] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large Scale Multiple Kernel Learning. *JMLR*, 7:1531-1565, 2006.

[13] M. Sugiyama. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. *JMLR*, 8, pp.1027-1061, 2007.

[14] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, M. Varma. Multiple Kernel Learning and the SMO Algorithm. *NIPS*, 2010.

[15] K. Q. Weinberger, L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10: 207-244, 2009.

[16] http://www.image-net.org.

Table 3: The recognition accuracy (%) with different k using the best single feature and combined features

WT	3	5	7	9	11	13	15	17	19
EUC-PCA	20.9	23.42	24.85	25.49	26.41	26.9	27.1	27.65	27.76
LMNN	**21.56**	24.31	25.47	26.61	27.12	27.56	28.05	28.09	28.55
ITML	20.56	23.44	25.46	26.41	26.94	27.45	27.92	28.24	28.32

Combined	3	5	7	9	11	13	15	17	19
EUC-PCA	20.26	22.53	23.67	24.20	24.69	25.13	25.47	25.67	25.78
ITML	19.82	22.87	24.11	25.04	25.98	26.17	26.64	26.88	27.51
LMNN	14.12	15.97	16.76	17.33	18.04	18.23	18.70	18.96	19.17
MFDML-L	19.54	26.31	28.03	29.36	30.10	30.92	31.34	31.59	31.76
MFDML-K	21.31	**28.79**	**31.94**	**33.57**	**34.68**	**35.03**	**35.82**	**36.05**	**36.19**