ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

DIMENSIONALITY REDUCTION USING BAYESIAN LEARNING PREDICTIVE SUBSPACES FOR SUPERVISED AND SEMI-SUPERVISED MULTI-LABEL LEARNING

T.Seeniselvi[1], M.Manjula[2], R.Deepa[3]

Associate Professor, PG & Research Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore, India [1]

Research Scholar, PG & Research Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore, India [2,3]

Abstract: For supervised learning problems, dimensionality reduction is generally applied as a pre-processing step. However, Coupled training of dimensionality reduction and classification is proposed previously to improve the prediction performance for single-label problems. In this paper, we first introduce a novel Bayesian method that combines linear dimensionality reduction with linear binary classification for supervised multi-label learning and present a deterministic variational approximation algorithm to learn the proposed probabilistic model. The proposed method is to find intrinsic dimensionality of the projected subspace using automatic relevance determination and to handle semi-supervised learning using a low-density assumption. Our proposed method significantly outperforms combined Bayesian with multiple kernel Fisher discriminate analysis followed by a standard kernel-based learner, especially on low dimensions.

Keywords: Dimensionality reduction, multi-label learning, subspace learning, Bayesian, supervised learning, semi-supervised learning.

I. INTRODUCTION

Dimensionality reduction algorithms try to find low dimensional representations of the input data for three main goals: 1) removing the inherent noise to improve the prediction performance, 2) obtaining 2D or 3D visualizations to do exploratory data analysis, and 3) reducing space and time complexities for testing phase. Principal component analysis (PCA) is the most basic dimensionality reduction algorithm; it performs a linear projection on the input data [1]. PCA basically tries to maximize the explained variance of the input data in the projected subspace. The projection matrix is found by computing the eigenvalue decomposition of the data covariance matrix. It generally performs badly for classificationproblems due to its linear and unsupervised nature. Kernel principal component analysis (KPCA) is an extension of PCA obtained by introducing nonlinearity using kernel functions [2].

Fisher discriminant analysis (FDA) is another well-known linear dimensionality reduction algorithm [3]. FDA tries simultaneously to minimize the within-class variance and to maximize the between-class variance. The main limitation of FDA is that the dimensionality of the projected subspace can be at most K 1 where K is the number of classes. Kernel Fisher discriminant analysis (KFDA) is a nonlinear extension of FDA formulated using kernel functions [4]. For supervised learning problems, dimensionality reduction and prediction steps are generally performed separately in a serial manner.

Supervised dimensionality reduction algorithms use output values (e.g., labels) to find a better subspace for the prediction step but they generally have their own target functions different from the one that the learner trained on the projected subspace uses, leading to low prediction performance. Hence, coupled training of these two steps may improve the overall system performance.

II. RELATED WORK

A.Coupled training model

Coupled training of dimensionality reduction and classification has been studied previously. For example, Globerson and Roweis [5] and Weinberger and Saul [6] try to learn a Mahalanobis distance metric while considering the nearest neighbour classification performance. More similar to our approach, Pereira and Gordon [7] propose to optimize the projection matrix and the parameters of a linear classifier with an alternating optimization method but the objective function has an additional regularization term for reconstruction error. Ji and Ye [8] follow the same idea of joint learning of dimensionality reduction and classification parameters for multi-label learning.

B.MKL

Recently, dimensionality reduction algorithms that use MKL in the inner loop to combine different kernels are proposed. Liang and Li [9] give an alternating optimizationmethod for generalized discriminant analysis

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

with multiple kernels. At each iteration, two different eigenvalue decompositions are calculated for updating the projectionmatrix and the kernel weights separately. Lin et al. [10] extend the graph embedding framework of [11], which is a unified framework for different dimensionality reductionalgorithms, by introducing multiple kernels. Different supervised (e.g., FDA), unsupervised (e.g., locality preserving projections [12]), and semi-supervised (e.g., semi-superviseddiscriminant analysis [13]) dimensionality reduction algorithms can be modified to include MKL. Similarly, Hou et al. [14] give a unified framework for local dimensionality reduction methods using kernel matrices and it can also be extended to MKL setup.

In theseapproaches, classification has to be performed after dimensionality reduction and there is no guarantee that the projected subspace will be predictive.

When thedifferent feature representations and/or similarity measures, it may be a good idea to use a weighted combination of those for dimensionality reduction inspiredby MKL algorithms that are generally formulated for supervised learning setups. In dimensionality reduction coupled with MKL, we definitely need to optimize thekernel combination parameters to obtain the most predictive subspace. The target function of the prediction step should also be combined with MKL and dimensionalityreduction to improve the prediction performance.

III. PROPOSED SYSTEM

The proposed method is to find intrinsic dimensionality of the projected subspace using automatic relevance determination and to handle semi-supervised learning using a low-density assumption. Our proposed method significantly outperforms combined Bayesian with multiple kernel Fisher discriminate analysis followed by a standard kernel-based learner, especially on low dimensions. A novel supervised and semi-supervised multi-label learning method where the linear projection matrix and the binary classification parameters are learned together tomaximize the prediction performance in the projected subspace.

A.Coupled dimensionality reduction and classification

To Performing the dimensionality reduction and classification successively (with two different objective functions) may not result in a predictive subspace and may have low generalization performance. In order to find a better subspace, coupling dimensionalityreduction and single-output supervised learning. To consider the predictive performance of the target subspace while learning the projection matrix.

In order to benefitfrom the correlation between the class labels in a multi-label learning scenario, we assume a common subspace and perform classification for all labels in that subspace using different classifiers for each label separately. The predictive quality of the subspace nowdepends on the prediction performances for multiple labels instead of a single one.

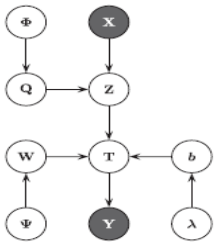


Fig: 1. Coupled dimensionality reduction

Fig. 1 illustrates the probabilistic model for multi-label binary classification with a graphical model and its distributional assumptions. The data matrix X is used to project data points into a lowdimensional space using the projection matrix Q. The low-dimensionalrepresentations of data points Z and the classification parameters fb;Wg are used to calculate the classification scores. Finally, the given class labels Y are generated from the auxiliary matrix T, which is introduced to make the inference proceduresefficient. We formulate a variational approximation procedure for inference in order to have a computationally efficient algorithm.

B.Inference variational approximation method

The variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors to find the joint parameter distribution. Assuming independence between the approximate posteriors in the factorable ensemble canbe justified because there is not a strong coupling between our model parameters. We can write the factorable ensemble approximation of the required posterior as

$$p(\Phi,\sqcap\,|X,Y) \approx q(\Phi,\sqcap) = q(\Phi)\,q(Q)\,q(Z)\,q(\lambda)\,q(\psi)\,q(b,W)\,q(T) \quad (1)$$

To choose a model projected data instances explicitly (i.e., not marginalizing out them) and independently (i.e., assuming a distribution independent of other variables) in the factorable ensemble approximation in order to decouple the dimensionality reductionand classification parts. By doing this, we achieve to obtain update equations for Q and fb;Wg independent of each other.

C.Bayesian Multi-label Learning Algorithm

The complete Bayesian Multi-label Learning Algorithm inference algorithm is listed in Algorithm 1. The inference mechanism sequentially updates the approximate posterior distributions of the model parameters and the latent variables until convergence, which can be checked by monitoring the lower bound in (1). Exact form of the variational lower bound can be found in Section II B. The first term of the lower bound corresponds to the sum of exponential forms of the distributions inthe joint likelihood.

The second term is the sum of negative entropies of the approximate posteriors in the ensemble. The only nonstandard distribution in the second term is the truncated normal distributions of the auxiliary variables; nevertheless, the truncatednormal distribution has a closed-form formula also for its entropy.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

Algorithm1: Bayesian Multi-label Learning

Require: **X,Y,R,**αλ,βλ,αΦ,βΦ,αΨ and βΨ
1: Initialize q(Q), q(Z), q(b, W), and q(T) randomly
2: repeat
3: Update q(Φ) and q(Q)
4: Update q(Z) using (1)
5: Update q(λ), q(Ψ), and q(b, W) using (1)
6: Update q(T)
7: until convergence
8: return q(Q) and q(b, W)

D.semi-supervised multi-label learning

Labelling large data collections may not be possible due to extensive labour required. In such cases, we should efficiently use a large number of unlabeled data points in addition to a few labelled data points (i.e., semi-supervised learning). Semi-supervisedlearning is not well-studied in the context of multi-label learning. There are a few attempts that formulate the problem as a matrix factorization problem.The only consider dimensionality reduction and classification together for multi-label learning in a semi-supervised setup.We modify our probabilistic model described above for semi-supervised learning assuming a low-density region between the classes. We basically need to make the class labels partially observed and to introduce a new set of observedauxiliary variables denoted by L. The distributional assumptions for Y and L are defined as follows:

$$y_i^o | t_i^o \sim \begin{cases} \delta(y_i^o t_i^o > 1/2) y_i^o \in \{\pm 1\} \\ 1 - \delta 1/2 \geq t_i^o \geq -1/2 \; otherwise \end{cases} \quad (2)$$

The first distributional assumption has two main implications: (i) Alow-density region is placed between the classes similar to the margin in support vector machines. (ii) Unlabeled data points are forced to be outside of this low-density region.

IV. EXPERIMENTAL RESULTS

We test our proposed Bayesian Multi-label Learning Algorithm on one digit recognition and two bioinformatics data sets by comparing it with a baseline algorithm discussed in previous studies[16], [17]. BMLA is our proposedalgorithm that combines MKL, dimensionality reduction, and supervised learning into a joint optimization problem, outlined in Algorithm 1.

We implement these algorithms in MATLAB1 and solve SVM problems with LIBSVM software [22]. In the experiments, we use the one-versus-all decomposition strategy and select the regularization parameter C of SVM problemsfrom f1; 10; 100; 1,000g using cross-validation.
To perform experiments on the two proteins sub-cellular localization data sets PLANT and PSORT+3 from [23]. Table 1 lists the average accuracies with their standard deviations obtained by MKFDA+SVM, SMKE and BMLA on the PLANT and PSORT+ data sets. We see that BMLA is better than SMKE,MKFDA+SVM on both data sets.

Table 1: The Average Test Accuracies with Their Standard Deviations on the PLANT and PSORT+ Data Sets

Algorithm	Plant	Psort +
MFDA+SVM	73.42	79.86
SMKE	81.71	82.64
BMLA	89.97	90.65

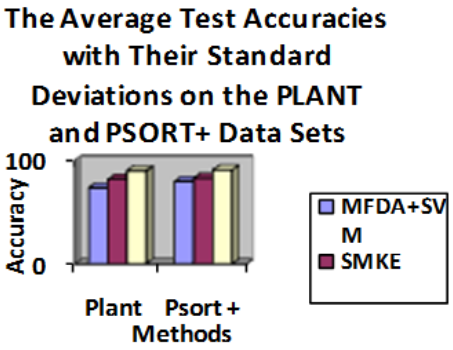
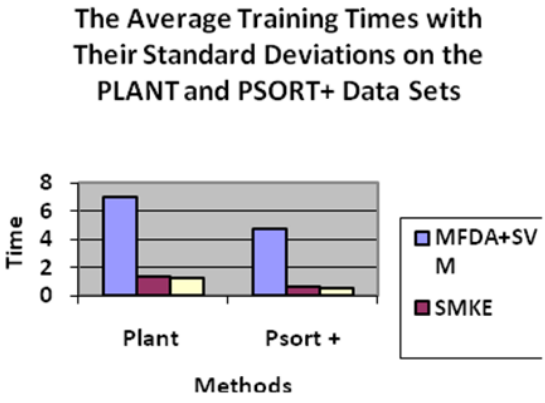


Table 2 provides the running times of the two algorithms with their standard deviations on the PLANT and PSORT+ data sets. Different from the results on the MULTIFEAT dataset, we see that BMLA is significantly faster than SMKE when there are a large number of kernels (69 in our experiments).

Table 2: The Average Training Times with Their Standard Deviations on the PLANT and PSORT+ Data Sets

Algorithm	Plant	Psort +
MFDA+SVM	7.0	4.69
SMKE	1.30	0.64
BMLA	1.22	0.54



VI. CONCLUSION

In this paper, the Bayesian supervised multi-label learning method that couples linear dimensionality reduction and linear binary classification. We then provide detailed derivations for supervised learning using a deterministic variational approximation approach.To formulate two variants: (i) an automatic relevance determination variant to find intrinsic dimensionality of the projected subspace

and (ii) a semi-supervised learning variant with a low-density region between the classes to make use of unlabeled data.

The proposed models can be extended in different directions: First, we can use a nonlinear dimensionality reduction step before multi-label classification step using kernels instead of data matrix. Second, we can use a nonlinear classification algorithm such asGaussian processes instead of probit model in our formulation to increase the prediction performance. Lastly, we can learn a unified subspace for multiple input representations (i.e., multitask learning) by exploiting the correlations between different tasks definedon different input features.

REFERENCES

[1] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," Philosophical Magazine, vol. 2, no. 6, pp. 559-572, 1901.
[2] B. Scholkopf, A. Smola, and K.-R.Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," Neural Computation, vol. 10, no. 5, pp. 1299-1319, 1998.
[3] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.
[4] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K.-R. Muller, "Constructing Descriptive and Discriminative Nonlinear Features: Rayleigh Coefficients in Kernel Feature Space," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 5, pp. 623-628, May 2003.
[5] A. Globerson and S. Roweis, "Metric Learning by Collapsing Classes," Proc. Conf. Advances in Neural Information Processing Systems, vol. 18, 2006.
[6] K.Q. Weinberger and L.K. Saul, "Distance Metric Learning for Large Margin NearestNeighbor Classification," The J. Machine Learning Research, vol. 10, pp. 207-244, 2009.
[7] F. Pereira and G. Gordon, "The Support Vector Decomposition Machine," Proc. 23rd Int'l Conf. Machine Learning, 2006
[8] S. Ji and J. Ye, "Linear Dimensionality Reduction for Multi-Label Classification," Proc. 21st Int'l Joint Conf. Artifical Intelligence, 2009.
[9] Z. Liang and Y. Li, "Multiple Kernels for Generalised Discriminant Analysis," IET Computer Vision, vol. 4, pp. 117-128, 2010.
[10] Y.-Y. Lin, T.-Y.Liu, and C.-S. Fuh, "Multiple Kernel Learning for Dimensionality Reduction," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 6, pp. 1147-1160, June 2011.
[11] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 40-51, Jan. 2007.
[12] X. He and P. Niyogi, "Locality Preserving Projections," Proc. Conf. Advances in Neural Information Processing Systems, vol. 16, 2004.
[13] D. Cai, X. He, and J. Han, "Semi-Supervised Discriminant Analysis," Proc. 11th Int'l Conf. Computer Vision, 2007.
[14] C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable Local Dimensionality Reduction Approaches," Pattern Recognition, vol. 42, no. 9, pp. 2054-2066, 2009.

BIOGRAPHIES

Mrs.K.T.Seeniselvi Pursuing Ph.D in Bharathiar University. Currently she is working as an Associate Professor of PG & Research Department of Computer Science in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree M.sc in Madurai Kamaraj university and also her UG Degree B.Sc in Madurai Kamaraj University. Totally she has 13 years and 8 Months of Experience in Teaching Field.

Ms.M.Manjula Pursuing M.Phil., Research Degree in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree MCA in Info Institute of Engineering at Coimbatore and also her UG Degree BCA in Bharathiar University. She had 2 years of Experience as a Lecturer in Sasurie College of Engineering.

Ms.R.Deepa Pursuing M.Phil., Research Degree in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree MCA in Navarasam Arts & Science College for Women at Erode and also her UG Degree B.Sc (Mathematics) in Bharathidasan University. She had 2 years of Experience as an Assistant Professor in Sasurie College of Arts & Science.