Semi-Supervised Weighted Maximum Variance Dimensionality Reduction

Pranitha Surya Andalam and Usman Roshan
Department of Computer Science
New Jersey Institute of Technology
University Heights, Newark, NJ 07102
Email: usman@cs.njit.edu

***Abstract*—The weighted maximum variance is a general procedure for dimensionality reduction of which principal component analysis and the maximum margin criterion discriminant are special cases. In previous work we studied a simple two parameter version of this that we call 2P-WMV. There we show that with our extracted features we obtain a lower average classification error given by 1-nearest neighbor compared to other dimensionality reduction methods and the raw features. Here we extend our method to work in a semi-supervised setting. We present our method with experimental results on several real datasets and show that it yields the lowest error particularly when only 50% of the data is available for training.**

Keywords: **dimensionality reduction, weighted maximum variance, semi-supervised**

I. Introduction

Semi-supervised learning has produced mixed results often performing comparable to supervised learning [1]. Here we extend our recent work on dimensionality reduction to the semi supervised case. We consider the weighted maximum variance [2] for semi-supervised learning using nearest neighbor graphs.

Suppose we are given the vectors $x_i \in R^d$ for $i = 0...n-1$ and a real matrix $C \in R^{n \times n}$. Let X be the matrix containing x_i as its columns (ordered x_0 through x_{n-1}). Now consider the optimization problem

$$\arg\max_w \frac{1}{2n} \sum_{i,j} C_{ij}(w^T(x_i - x_j))^2 \tag{1}$$

where $w \in R^d$ and C_{ij} is the entry in C corresponding to the i^{th} row and j^{th} column. We call this the weighted maximum variance (WMV) and have previously shown it to be a more general representation of principal component analysis [3] and the maximum margin criterion [2], [4].

We briefly review our previous two parameter version of this and then present the semi-supervised extension. We compare the two versions on real data with 90% and 50% available training data. On 90% and 50% we find the semi-supervised to perform better than the supervised cases.

II. Methods

We briefly review the two parameter weighted maximum variance (2P-WMV) before extending it to the semi-supervised case.

A. Two Parameter Weighted Maximum Variance Discriminant

In Equation 1 if we let $C_{ij} = G_{ij} - 2L_{ij}$ then we obtain the following form of WMV.

$$\arg\max_w \frac{1}{2n}(\sum_{i,j} G_{ij}(w^T(x_i - x_j))^2 - \sum_{i,j} 2L_{ij}(w^T(x_i - x_j))^2) \tag{2}$$

Now define the matrix $G \in R^{n \times n}$ as $G_{ij} = \frac{1}{n}$ for all i and j and $L \in R^{n \times n}$ as

$$L_{ij} = \begin{cases} \alpha & \text{if } y_i = y_j \\ \beta & \text{if } y_i \neq y_j \\ 0 & \text{if } y_i \text{ or } y_j \text{ is undefined} \end{cases}$$

This gives us the discriminant $w^T(S_t - 2(\alpha S'_w + \beta S'_b))w$ where

$$S'_w = \frac{1}{n}\sum_{k=1}^{c} n_k \sum_{cl(x_j)=k}(x_j - m_k)(x_j - m_k)^T$$
$$S'_b = \frac{1}{2n}\sum_{c=1}^{k}\sum_{d=c+1}^{k}\sum_{cl(x_i)=c, cl(x_j)=d}(x_i - x_j)(x_i - x_j)^T$$

The discriminant yielded by 2P-WMV is given by the standard total scatter matrix, a modified within-class matrix, and a pairwise inter-class scatter matrix. We can obtain the maximum margin criterion from this by setting $\alpha = \frac{1}{n_k}$ if $y_i = k, y_j = k$ and $\beta = 0$. This discards the inter-class scatter matrix and makes $S'_w = S_w$.

B. Semi-Supervised Weighted Maximum Variance

1) k nearest neighbors: In the semi-supervised case we define the matrix L_{ij} as

$$L_{ij} = \begin{cases} \alpha & \text{if } y_i = y_j \\ \beta & \text{if } y_i \neq y_j \\ \alpha & \text{if } i \text{ is among the } k \text{ nearest neighbors of } j \text{ (or vice-versa)} \\ 0 & \text{otherwise} \end{cases}$$

After defining L and G compute L_g the Laplacian of G, L_l the Laplacian of L, and the matrix $\frac{1}{n}X(L_g - L_l)X^T$ (the SSWMV discriminant). The solution to 2P-WMV is w that maximizes $\frac{1}{n}w^T X(L_g - L_l)X^T w$ which is in turn is given by the largest eigenvector of $\frac{1}{n}X(L_g - L_l)X^T$ [5].

2) Majority among k nearest neighbors: Sometimes Semi-Supervised data can be wrongly classified by simply relying on k-nearest neighbors with the above definition for the matrix L_{ij}. In order to avoid those scenarios, we leveraged the labels of labeled points and determined the class of unlabeled point by finding the majority class among the k-nearest neighbor labeled points.

$$L_{ij} = \begin{cases} \alpha & \text{if } y_i = y_j \\ \beta & \text{if } y_i \neq y_j \\ \alpha & \text{if } i \text{ and } j \text{ belong to same class} \\ \beta & \text{if } i \text{ and } j \text{ belong to different class} \end{cases}$$

3) Clustering: In this case, we have used k-means clustering to determine the classes of unlabeled points. We define the matrix L_{ij} as

$$L_{ij} = \begin{cases} \alpha & \text{if } y_i = y_j \\ \beta & \text{if } y_i \neq y_j \\ \alpha & \text{if } i \text{ and } j \text{ belong to same cluster} \\ \beta & \text{if } i \text{ and } j \text{ belong to different clusters} \end{cases}$$

4) Relative Clustering Validity Criterion: Relative clustering validity criteria is used to quantitatively measure the quality of data partitions formed using clustering. One important validation criterion is the silhouette width criterion [8] . Silhouette width criterion coefficient is calculated using the mean intra-cluster distance and the mean nearest cluster distance for each sample.

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if a(i) < b(i)} \\ 0 & \text{if a(i) = b(i)} \\ \frac{b(i)}{a(i)} - 1 & \text{if a(i) > b(i)} \end{cases}$$

Where a(i) the measure of how dissimilar is i to its own cluster and b(i) is the lowest average dissimilarity of i to any other cluster. Thus an S(i) close to one means that the datum is appropriately clustered and if S(i) is close to negative one, then it is more appropriate if it was clustered in its neighboring cluster. An S(i) near zero means that the datum is on the border of two natural clusters.

III. Results

To evaluate the classification ability of our extracted features we use the simple and popular 1-nearest neighbor (1NN) algorithm. In previous work we [2] found 2P-WMV extracted features to have lower average error (with statistical significance) than other dimensionality reduction programs such as the weighted maximum margin discriminant (WMMC), PCA, and the features as they are. Here we consider training validation splits of 90% and 50% to evaluate the effect of training data size on our method and compare it to just 2P-WMV. We apply the 1-nearest neighbor classification algorithm to features extracted from our new semi-supervised method SSWMV and the previous 2P-WMV [2]. We calculate average error rates across 20 randomly selected datasets shown in Table I from the UCI Machine Learning Repository [6].

TABLE I
DATASETS FROM THE UCI MACHINE LEARNING REPOSITORY THAT WE USED IN OUR EMPIRICAL STUDY[6]

Code	Dataset	Classes	Dimensions	Instances
1	Liver Disorders	2	6	345
2	Wine	3	13	178
3	Heart	2	13	270
4	Australian Credit Approval	2	14	690
5	Climate	2	18	540
6	Diabetic Retinopathy	2	20	1150
7	Statlog German Credit Card	2	24	1000
8	Breast Cancer	2	30	569
9	Dermatology	6	34	366
10	Ionosphere	2	34	351
11	Qsar	2	41	1055
12	SPECTF Heart	2	44	267
13	Sonar	2	60	208
14	Ozone	2	72	1847
15	Hill Valley	2	100	606

A. Experimental Methodology

In both 2P-WMV and SSWMV we let β range from {-2,-1.9,-1.8,-1.7,-1.6,-1.5,-1.4,-1.3,-1.2,-1.1,-1,-.9,-.8,-.7,-.6,-.5,-.4,-.3,-.2,-.1,-.01} and α fixed to 1. For each parameter we reduce dimensionality to 20 and then pick the top $1 <= k <= 20$ features that give the lowest 1NN error on the training. Thus the cross-validation on the training set gives us the best values of α and the reduced number of features which we then apply to the validation set.

We wrote our code in C and R and make it freely available at http://www.cs.njit.edu/usman/sswmv/. Our C programs use CLAPACK libraries for performing the eigenvector and singular value decompositions.

B. Experimental Results on 15 Datasets

We compute the misclassification rate ($\frac{numberofmisclassifications}{numberoftest}$) for each training-validation split during cross-validation and take the mean to be the average cross-validation error. In Table II we show the average cross-validation error on each dataset. Across the 15 datasets 2P-WMV+1NN achieves the lowest average error of 18.45% and has the lowest error in 4 out of the 15 datasets. 1NN have average errors at 18.13%. 1NN have the lowest error in 8 out of the 15 datasets respectively.

We measure the statistical significance with the Wilcoxon rank test [7]. This is a standard test to measure the difference between two methods across a number of datasets. Roughly speaking it shows statistical significance between two methods when one outperforms the other each time on a large number of datasets. In Table II, the p-values show that 2P-WMV+1NN statistically significantly outperforms the other three method across all 15 datasets.

IV. Discussion

Both 2PWMV+1NN and WMMC+1NN reduce dimensionality by determining optimal parameters specific to the given dataset. This approach is better than the unsupervised PCA and the non-parametric MMC (results not shown here). In

fact 1NN applied to the raw data can be better than non-parameteric MMC most of the time.

In this study we fixed α for 2PWMV and varied only β. If we cross-validated α we could potentially obtain lower error but at the cost of increased running time. In the current experiments 2PWMV+1NN and WMMC+1NN are the slowest methods yet still tractable for large datasets.

We chose 1NN as the classification method for this study due to its simplicity and its popularity with dimensionality reduction programs. Other classifiers such as the support vector machine [3] may perform better when replaced with 1NN. However, in that case the regularization parameter would also need to be optimized via cross-validation which increases the total runtime.

V. Conclusion

We introduce a two parameter variant of the weighted maximum variance discriminant and optimize it with cross-validation followed by 1-nearest neighbor for classification. Compared to existing approaches our method obtains the lowest average error with statistical significance across several real datasets from the UCI machine learning repository.

References

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

[2] Turki, T., Roshan, U., *Weighted maximum variance dimensionality reduction*. In Martnez-Trinidad, J., Carrasco-Ochoa, J., Olvera-Lopez, J., Salas-Rodrguez, J., Suen, C., eds., *Pattern Recognition. Volume 8495 of Lecture Notes in Computer Science*, Springer International Publishing, 11-20, 2014..

[3] Alpaydin, E., *Machine Learning. MIT Press*, 2004.

[4] Li, H., Jiang, T., Zhang, K., *Efficient and robust feature extraction by maximum margin criterion*. In Thurn, S., Saul, L., Scholkopf, B., eds., *Advances in Neural Information Processing Systems 16* MIT Press, Cambridge, MA, 2004.

[5] Niijima, S., Okuno, Y., *Laplacian linear discriminant analysis approach to unsupervised feature selection. Computational Biology and Bioinformatics*, IEEE/ACM Transactions on 6(4):605- 614, 2009.

[6] Lichman, M., *UCI Machine Learning Repository*, 2013.

[7] Kanji, G.K., *100 Statistical Tests*, Sage Publications Ltd, 1999.

[8] Vendramin, L., Campbello, R.J.G.B., Hruschka, E.R., *Relative clustering validity criteria: A comparative overview*, Statistical Analysis and Data Mining, 3(4):209-235, 2010.

TABLE II

AVERAGE CROSS-VALIDATION ERROR OF DIFFERENT ALGORITHMS ON EACH OF THE 15 REAL DATASETS FROM THE UCI MACHINE LEARNING REPOSITORY. SHOWN IN BOLD IS THE METHOD WITH THE LOWEST UNIQUE ERROR.

	Dataset	2P-WMV+1NN		SSWMV+1NN		SSWMV + Majority 15NN		SSWMV + clustering		SSWMV + clustering relative validity criteria	
		90%	50%	90%	50%	90%	50%	90%	50%	90%	50%
1	Liver Disorders	0.38	0.382	0.377	**0.371**	**0.317**	0.377	0.368	0.386	0.38	0.395
2	Wine	0.078	0.084	**0.072**	**0.0752**	0.272	0.498	0.267	0.309	0.267	0.309
3	Heart	0.244	0.236	**0.241**	**0.227**	0.263	0.288	0.267	0.237	0.270	0.244
4	Australian Credit Approval	0.189	**0.201**	0.189	**0.201**	0.207	0.212	**0.187**	0.214	**0.187**	0.214
5	Climate	0.067	0.094	0.067	0.094	**0.065**	**0.082**	0.085	0.087	0.061	0.088
6	Diabetic Retinopathy	**0.318**	**0.373**	0.319	0.374	0.396	0.388	0.395	0.389	0.406	0.387
7	Statlog German Credit Card	0.347	0.336	0.343	**0.334**	0.344	0.368	0.342	0.377	**0.341**	0.38
8	Breast Cancer	0.095	0.066	0.094	**0.064**	**0.089**	0.094	0.096	0.092	0.095	0.092
9	Dermatology	**0.044**	**0.067**	0.045	**0.067**	0.092	0.527	0.092	0.157	0.092	0.157
10	Ionosphere	0.092	0.123	**0.086**	**0.112**	0.117	0.129	0.119	0.131	0.105	0.135
11	Qsar	0.22	**0.222**	0.212	0.231	**0.206**	0.251	0.215	0.246	0.211	0.244
12	SPECTF Heart	0.237	**0.238**	0.241	0.245	0.255	0.278	**0.204**	0.249	0.222	0.277
13	Sonar	0.219	0.244	0.219	0.235	**0.195**	0.267	0.2	0.228	0.214	**0.222**
14	Ozone	**0.112**	0.117	0.113	0.122	0.114	0.132	0.119	**0.115**	0.114	0.12
15	Hill Valley	0.042	0.069	**0.034**	**0.035**	0.035	0.296	0.302	0.367	0.300	0.364
	Average Error	0.178933	0.190133	0.1768	0.18581	0.1978	0.27913	0.2172	0.23893	0.2176	0.24186