A Preliminary MML Linear Classifier using Principal Components for Multiple Classes

Authors

Lara Kornienko, David W. Albrecht and David L. Dowe

School of Computer Science and Software Engineering, Monash University, Clayton, Victoria, 3800, Australia.

Details of Contact Author

Lara Kornienko: **Email:** Lara.Kornienko@csse.monash.edu.au **Phone:** (03)98425302
David Albrecht: **Email:** David.Albrecht@csse.monash.edu.au **Phone:** +61399055526

Abstract

In this paper we improve on the supervised classification method developed in Kornienko et al. (2002) by the introduction of Principal Components Analysis to the inference process. We also extend the classifier from dealing with binomial (two-class) problems only to multinomial (multi-class) problems.

The application to which the MML criterion has been applied in this paper is the classification of objects via a linear hyperplane, where the objects are able to come from any multi-class distribution. The inclusion of Principal Component Analysis to the original inference scheme reduces the bias present in the classifier's search technique. Such improvements lead to a method which, when compared against three commercial Support Vector Machine (SVM) classifiers on Binary data, was found to be as good as the most successful SVM tested. Furthermore, the new scheme is able to classify objects of a multiclass distribution with just one hyperplane, whereas SVMs require several hyperplanes.

Content Areas

Machine Learning, Knowledge discovery and data mining.

A Preliminary MML Linear Classifier using Principal Components for Multiple Classes

Lara Kornienko[1], David W. Albrecht[1], and David L. Dowe[1]

School of Computer Science and Software Engineering,
Monash University, Clayton, Victoria, 3800, Australia
{Lara.Kornienko,David.Albrecht}@csse.monash.edu.au

Abstract. In this paper we improve on the supervised classification method developed in Kornienko et al. (2002) by the introduction of Principal Components Analysis to the inference process. We also extend the classifier from dealing with binomial (two-class) problems only to multinomial (multi-class) problems.
The application to which the MML criterion has been applied in this paper is the classification of objects via a linear hyperplane, where the objects are able to come from any multi-class distribution. The inclusion of Principal Component Analysis to the original inference scheme reduces the bias present in the classifier's search technique. Such improvements lead to a method which, when compared against three commercial Support Vector Machine (SVM) classifiers on Binary data, was found to be as good as the most successful SVM tested. Furthermore, the new scheme is able to classify objects of a multiclass distribution with just one hyperplane, whereas SVMs require several hyperplanes.

Key words: Machine Learning, Knowledge discovery and data mining

1 Introduction

This paper extends the binary linear classification method presented by the authors in [9] by introducing Principal Component Analysis to the inference process. Furthermore, the capability of dealing with multinomial distributions is also developed.

The original method, named the Spikey method [10, Chapter 5] [9, 11], used a Linear classifier in conjunction with the Bayesian 'Minimum Message Length' (MML) principle [18–21] as an objective function to infer the correct distributions of a given binary-labelled data set. We have developed a new scheme named PCA-Spikey, which takes advantage of any biases in the spread of the data by using the Principal Components as an initial set of axes on which to begin the search for the separating hyperplane. Furthermore, rather than just discriminate between two classes of data, as most linear classifiers do, PCA-Spikey allows for any kind of multinomial distribution either side of the hyperplane.

As a benchmark, we have used an implementation in Statistical Learning Theory, namely the Support Vector Machine (SVM) [17].

2 The PCA-Spikey Codes

The primary aim of introducing Principal Components Analysis (PCA) to the Spikey program was to improve the inference technique by obtaining a set of axes on which to perform the inference that were more representative of the natural spread of the data. Doing this would, in theory, enable inherent biases in the data to be recognised - thus producing hyperplanes that were a better fit at a cheaper MML cost. One reason for the belief that introducing PCA would produce cheaper hyperplanes is due to the search technique used in the Spikey scheme: hyperplanes in the direction of or perpendicular to the major axes are given the cheapest encoding costs. Thus, by transforming the original axes to the directions and scales of greatest spread in the data, the cheapest hyperplanes are going to be amongst those which split the data perpendicular to these directions. This suggests the following process to obtain a Linear classifier:

1. The Principal Components are found for the skewed data and the points projected into the Principal Component space.
2. The data in the Principal Component space is then normalised so that it falls within a hyper-cubed region (a square in two dimensions).
3. Inference is performed in the normalised Principal Component space via the Spikey program [10, Chapter 5] [9].
4. The hyperplane found was transformed back into the original coordinates.

3 Results

Four PCA-Spikey methods were developed in [10, Chapters 6,7,8]. These were PCA-MML$_{NU}^{WT}$, PCA-MML$_{NUR}^{WT}$, PCA-MML$_{SR}^{WT}$ and PCA-MML^{ANG}. The SVMs used to test the PCA-Spikey methods against are SVM^{light} [8], the Lagrangian SVM [13] and SMOBR [15].

As with the Spikey methods, all the PCA-Spikey methods were compared to a true hyperplane, if known, using the Kullback-Leibler distance [12] or if the true hyperplane was not known (as for real data), using 10-fold cross-validation in conjunction with Probabilistic Scoring [5, 14, 4] [16, Section 3.1] [3, Section 11.4.2] and the Right/Wrong Predictive Accuracy scoring metric [10, Section 4.5]. We tested the PCA-Spikey methods on both real and artificial data having both binomial and trinomial distributions.

The boldface entries in the table columns highlight those methods that performed the best for those data sets with a 95% significance level using Student's t distribution on the population mean.

Data The artificial binary data sets were simply the original uniformly distributed data sets as input to the Spikey methods, but skewed according to two linear translation matrices (TM). In this paper, we present the Kullback-Leibler distances from the second Translation Matrix [10, Section 7.6] [11].

The data presented here is distributed relative to a true hyperplane, $y = 1.6x + 10.0$, where points were generated randomly having 95% of the points positive on one side of the hyperplane and 5% positive on the other side (see [10, Section 7.7] for elaboration). 'N' refers to the size of the data sets. All Kullback-Leibler scores presented in the tables are of the form 'Mean ± Standard Deviation', or $\mu \pm \sigma$. Table 1 shows the results for this data.

	N = 10	**N = 100**	**N = 1000**
PCA-MML$_{NU}^{WT}$	0.2081 ± 0.0527	**0.0844 ± 0.0462**	0.0303 ± 0.0264
PCA-MML$_{SR}^{WT}$	0.2038 ± 0.0535	**0.0547 ± 0.0469**	**0.0056 ± 0.0050**
PCA-MML$_{NUR}^{WT}$	0.2110 ± 0.0464	**0.0837 ± 0.0441**	0.0679 ± 0.0124
PCA-MML^{ANG}	0.1801 ± 0.0735	**0.0767 ± 0.0502**	0.0122 ± 0.0122
MML_{NU}^{WT}	0.1805 ± 0.0450	0.1139 ± 0.0258	0.1055 ± 0.0029
MML_{SR}^{WT}	0.2167 ± 0.0409	0.3679 ± 0.0109	0.3902 ± 0.0010
SVM^{light}	0.1884 ± 0.0280	0.2609 ± 0.0432	0.1597 ±0.0136 ^{*2}
Lagrangian	**0.0573 ± 0.0434**	**0.0690 ± 0.0349**	0.0296 ± 0.0156
SMOBR	0.1411 ± 0.0673	0.1093 ± 0.0501	0.1282 ± 0.0236

Table 1. Kullback-Leibler distances ($\mu \pm \sigma$) between the true hyperplane ($y = 1.6x + 10.0$) and inferred hyperplanes for TM2 - on **95/05** data, N = 10, 100, 1000 points. The 'n' in $\mu \pm \sigma$ ^{*n} denotes the number of data sets on which SVM^{light} did not converge.

The real data sets used are the Wisconsin Prognostic Breast Cancer Database, January 8, 1991 [1] and the trinomial Iris data set [7].

The Wisconsin Prognostic Breast Cancer data set consists of 699 data, each having 10 input attributes plus a binary class attribute. The first attribute is the sample code number, which was ignored. The Iris data set contains 150 data points, where each of the three classes contains 50 points and each point consists of four numeric attributes and a class specification. However, we just report here the test on the last two attributes. Tables 2 and 3 refer to the results for the Breast Cancer and Iris data respectively. The first column of each table refers to the results obtained using Probabilistic Scoring and the second column refers to the Right/Wrong Predictive Accuracy score [5, 14, 4] [16, Section 3.1] [3, Section 11.4.2] [10, Section 4.5]. No SVMs have been tested on the Iris data set as how SVM classifiers deal with multinomial data sets is still an open question.

4 Discussion and Conclusion

It has been shown that the original Spikey encoding scheme described in [10, Chapter 5] [9] could be improved by the introduction of Principal Component Analysis, as results indicate that the PCA-Spikey methods out-performed the original Spikey methods on skewed data. It was also found that the PCA-Spikey methods performed significantly better than the SVMs on the larger artificial Binomial data sets tested, while the SVMs tended to dominate the smaller data sets. On the real Binomial data sets, due to the fact that the data was not highly skewed, the PCA-Spikey methods performed similarly to the original Spikey

Wisconsin Breast Cancer	**Prob. (bit score) error**	**Right/Wrong Acc'y**
PCA-MML$_{NU}^{WT}$	14.0531 ± 13.6865	0.905714 ± 0.160809
PCA-MML$_{SR}^{WT}$	10.1341 ± 7.52156	0.957143 ± 0.0349927
PCA-MML$_{NUR}^{WT}$	9.97667 ± 7.52561	0.957143 ± 0.0368856
Iris 2D	**Prob. (bit score) error**	**Right/Wrong Acc'y**
$PCA - MML_{NU}^{WT}$	8.2333 ± 1.5605	0.6567 ± 0.0568
$PCA - MML_{SR}^{WT}$	7.8656 ± 1.1779	0.6667 ± 0.0544
$PCA - MML_{NUR}^{WT}$	7.3257 ± 1.6426	0.6667 ± 0.0685
$PCA - MML^{ANG}$	8.0853 ± 1.6950	0.6600 ± 0.0717

Table 2. Real Data Set - Wisconsin Prognostic Breast Cancer Database. Probabilistic prediction bit score error results and "right/wrong" predictive accuracy results using 10-fold cross-validation ($\mu \pm \sigma$).

Table 3. Real Data Set - Iris, 2D. Probabilistic prediction bit score error results and "right/wrong" predictive accuracy results using 10-fold cross-validation (mean ± standard deviation).

methods, and they were as good as the best SVM on that data. The PCA-Spikey methods are also flexible enough to deal with multinomial data without any major changes to their implementation. It was found that the PCA-Spikey methods, when run on the trinomial 'Iris' data set, were able to separate the one separable class from the remaining two inseparable classes.

Overall, several improvements can be made to the PCA-Spikey methods in terms of the search procedures and coding schemes used, particularly for smaller data sets. A possible alternative is the encoding of data points to geometrically define a hyperplane, similar to the Support Vectors in SVMs (for an alternative MML approach to a related problem, which does not use Principal Components, see [16]). Furthermore, the type of distribution used for the input data may be varied to non-uniform distributions. Another recent development in MML that may be looked into has been in Comley and Dowe [2, 3], where a concrete application of Dowe's abstract notion of inverse learning [6] to generalised Bayesian networks including a mix of both continuous and discrete variables is given.

References

1. C.L. Blake and C.J. Merz (1998), *UCI Repository of machine learning databases*. http://www.ics.uci.edu/∼mlearn/MLRepository.html, Irvine, CA: University of California, Dep. of Information and Computer Science.
2. J.W. Comley and D.L. Dowe (2003). General Bayesian Networks and Asymmetric Languages, in Proc. Hawaii Int. Conf. on Stats. and Related Fields, 5-8 June, 2003.
3. J.W. Comley and D.L. Dowe (2005). Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages, Chap. 11 in P. Grunwald, I. J. Myung and M. A. Pitt (Eds.), Advances in Minimum Description Length: Theory and Applications, M.I.T. Press, April 2005, ISBN 0-262-07262-9. Final Camera Ready copy submitted October 2003.

4. D.L. Dowe, G.E. Farr, A.J. Hurst and K.L. Lentin (1996). Information-theoretic football tipping, in N. de Mestre (ed.), Third Australian Conference on Mathematics and Computers in Sport, Bond University, Qld, 233-241, 1996.
5. Dowe, D.L, & Krusel N. (1993). A decision tree model of bushfire activity, (Technical report 93/190) Dept Computer Science, Monash University, Melbourne, 7pp, 1993
6. Dowe, D.L. and Wallace, C.S. (1998). Kolmogorov complexity, minimum message length and inverse learning, abstract, page 144, 14th Australian Statistical Conference (ASC-14), Gold Coast, Qld, 6-10 July, 1998.
7. R.A. Fisher (1936), The use of multiple measurements in taxonomic problems, in Annals of Eugenics, Volume 7, pp. 179-188.
8. Joachims, T. (1998). Making Large-Scale SVM Learning Practical. In B. Scholkopf, C. J. C. Burges & A. J. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, USA, 1998.
9. L. Kornienko and D.L. Dowe and D.W. Albrecht (2002), Message Length Formulation of Support Vector Machines for Binary Classification - A Preliminary Scheme in Lecture Notes in Artificial Intelligence (LNAI) 2557, 15th Aust. Joint Conf. on A.I., Canberra, Australia. Springer-Verlag. pp 119-130.
10. L. Kornienko (2005), Implementing a Support Vector Machine in a Message Length Framework. Masters Thesis, School of Information Technology, Monash University, Clayton, Australia.
11. L. Kornienko and D.L. Dowe and D.W. Albrecht (2005), A Preliminary MML Linear Classifier using Principal Components for Multiple Classes. Tech.Report .School of Information Technology, Monash University, Clayton, Australia.
12. Kullback, S. (1959) *Information Theory and Statistics*. John Riley and Sons, Inc.
13. Mangasarian, O. L., & Musicant, D. R. (2000). *Lagrangian Support Vector Machines*. (Tech. Report 00-06). Data Mining Institute.
14. Needham, Scott L., & Dowe, David L. (2001). Message Length as an Effective Ockham's Razor in Decision Tree Induction. Proc. 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001), pp 253-260, Key West, Florida, U.S.A., Jan. 2001.
15. Platt, J. (1999). *Sequential minimal optimization: A fast algorithm for training support vector machines* in *Advances in Kernel Methods - Support Vector Learning*, Bernhard Scholkopf, Christopher J. C. Burges and Alexander J. Smola, Eds. 1999, pp. 185-208, MIT Press.
16. P.J. Tan and D.L. Dowe (2004), MML Inference of Oblique Decision Trees in Lecture Notes in Artificial Intelligence, G.I.Webb and X.Yu, Eds., 17th Australian Joint Conf. on Advances in A.I., Cairns, Australia. pp. 1082-1088, Springer-Verlag. ISSN:0302-9743, ISBN:3-540-24059-4, Vol 3339.
17. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York.
18. C.S. Wallace (2005), Statistical and Inductive Inference by Minimum Message Length, Springer. ISBN: 0-387-23795-X
19. Wallace, C. S., & Boulton, D.M. (1968). An information measure for classification. *Computer Journal*, 11, 185-194.
20. Wallace, C .S., & Dowe, D. L. (1999). Minimum Message Length and Kolmogorov Complexity. *Computer Journal*. 42(4), 270-283.
21. Wallace, C. S., & Freeman, P.R. (1987). Estimation and Inference by Compact Coding, *J Royal Stat. Soc. B*. 49, 240-252.