Discriminative Regularization: A New Classifier Learning Method

Hui Xue[1] Songcan Chen[1*] Qiang Yang[2]

[1] *Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, 210016, Nanjing, P.R. China*

[2] *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong*

Abstract:

Regularization involves a large family of the state-of-the-art techniques in classifier learning. However, since traditional regularization methods essentially derive from ill-posed multivariate functional fitting problems which can be viewed as a kind of regression, in classifier design, they usually give more concerns to the smoothness of the classifier, and do not sufficiently use the prior knowledge of given samples. Actually, due to the characteristics of classification, the classifier is not always necessarily smooth anywhere, especially near the discriminant boundaries between classes. Radial Basis Function Networks (RBFNs) and Support vector machines (SVMs), as two most famous ones in the regularization family, have been aware of the importance of the prior information to some extent. They focus on either the *intra*-class or the *inter*-class information respectively. In this paper, we present a novel regularization method – Discriminative Regularization (DR), which provides a general way to incorporate the prior knowledge for classification. Through introducing the prior information into the regularization term, DR aims to minimize the empirical loss between the desired and actual outputs, as well as maximize the *inter*-class separability and minimize the *intra*-class compactness in the output space simultaneously. Furthermore, by embedding equality constraints in the formulation, the solution of DR can follow from solving a set of linear equations. The

[*] Corresponding author: Tel: +86-25-84896481 Ext. 12106; Fax: +86-25-84498069; E-mail: s.chen@nuaa.edu.cn (S. Chen), xuehui@nuaa.edu.cn (H. Xue) and qyang@cse.ust.hk (Q. Yang)

classification experiments show the superiority of our proposed DR.

Keywords: Classifier design and evaluation, Computing methodology, Design methodology, Pattern recognition

1. Introduction

Regularization has a rich history which can date back to the theory of ill-posed problem[1, 2, 3]. By incorporating the right amount of prior information into the formulation, regularization techniques have been shown to be powerful in making the solution stable[4, 5]. In the past decades, regularization theory was introduced to the machine learning community on the premise that the learning can be viewed as a multivariate functional fitting problem[5, 6, 7, 8] and has been successfully applied to the classifier learning, deducing a large family of the state-of-the-art techniques. However, due to the original derivation, most of traditional regularization methods actually deal with classification as a special regression, typically Regularization Networks (RNs). Consequently, in classifier design, these methods usually give more concerns to the smoothness of the classifier, in the sense that similar inputs correspond to similar outputs. But, for classification, this assumption is sometimes too general. In fact, some similar samples near the discriminant boundaries more likely belong to different classes. Therefore, it is such characteristics of classification that the classifier is not always necessarily smooth anywhere, especially near the boundaries between classes. This means that traditional regularization methods do not sufficiently use the prior knowledge of given samples for classification. The famous “No Free Lunch” theorem states formally, that prior knowledge or assumptions of a problem at hand must be incorporated into the solution[9]. Without prior knowledge, no best classification systems or best pattern representation exist[10].

Radial Basis Function Networks (RBFNs) and Support vector machines (SVMs) are two most famous techniques in the regularization. They have applied some prior structural information to some extent. However, they emphasize on either the *intra*-class information or the *inter*-class information respectively, which are still

insufficient for classification.

In this paper, we focus on a traditional type of regularization with specific prior knowledge for classification, termed as Discriminative Regularization (DR). In view of the large family of regularization methods, it is valuable to ask for a general way to incorporate the prior information into the formulation, thus extends regularization for classification.

1.1 Goals and Paper Organization

We briefly list some desired properties of a general method for regularization to incorporate prior knowledge:

1. *Further Incorporation*: The method should incorporate further prior information, including the *inter*-class separability and the *intra*-class compactness simultaneously, compared to RBFNs and SVMs.
2. *Easy Incorporation*: The method should incorporate the prior information easily, but not destroy the traditional regression framework and increase more computational complexity.
3. *Easy Solution*: The method should keep the easily analytic solution framework just as regularization networks.
4. *Good Applicability*: The applicability on real world problems should be possible with respect to both good classification and generalization performances. The method should match or outperform the state-of-the-art regularization methods.

These points will be addressed and satisfied by the proposed method DR. In the following subsection, we briefly introduce the related works in regularization. Section 2 presents the proposed DR. In Section 3, we discuss the analytic solution to DR. Section 4 gives the experimental analysis. Some conclusions are drawn in Section 5.

The following is just given a result on the toy problem.

4.1 Toy Problem

In the toy problems, three two-moon datasets (I), (II) and (III) with different complexity are discussed. Each dataset contains one hundred samples in each class.

As shown in Fig. 1, '·' denotes the training samples and '+' denotes the testing samples. We compare RN ((a), (e), (i)), RBFN ((b), (f), (j)), SVM ((c), (g), (k)) with DR ((d), (h), (l)). The twelve subfigures show the discriminant boundaries of the four methods in each dataset. Furthermore, the respective training and testing accuracies are labeled in Table 2, where the first row in each grid shows the training accuracy, and the second row denotes the corresponding testing accuracy.

From Fig. 1 and Table 2, it can be seen that: (1) Due to the characteristics of traditional regularization, the boundaries of RN in the three datasets always keep smooth ((a), (e), (i)). When the two classes are far from each other, the training and testing accuracies of RN are comparable to RBFN, SVM and DR ((a)). However, when the classes get nearer and the complexity of classification increases, RN performs much worse than SVM and DR. And it is more likely (locally) over-smooth in the other two datasets ((e), (i)). It means that only emphasis on the smoothness of the classifier in the traditional regularization is too general for classification. (2) As the approximation to RN, RBFN retains the smoothness of the classifier ((b), (f), (j)). Owing to the partial incorporation of *intra*-class information generated from clustering, the accuracies of RBFN are better than RN in the Dataset (II). However, in the Dataset (III), the accuracies of RBFN are the same as RN, but much worse than DR. It also seems to be over-smooth just as RN, which justifies that only consideration of *intra*-class information is not sufficient in the complex classification problem. On the contrary, thanks to more emphasizing the *inter*-class information, the boundaries of SVM do not always keep smooth anywhere (relative to RN) just as DR. On the one hand, in the first dataset, the boundaries of DR and SVM are adequately smooth as well as RN and RBFN ((c), (d), (g), (h)). On the other hand, the boundaries become no longer smooth with the increase of the complexity ((k), (l)), the classification performance of DR and SVM is yet still much better than RN and RBFN. However, for only considering the *inter*-class information, the boundaries of SVM seem always be in the middle of the classes. Consequently, when the samples belonging to different classes overlap more heavily, SVM more likely can not effectively distinguish the samples near the boundaries, which leads to the

classification accuracies of SVM are worse than DR in the Dataset (III). It validates that only emphasis on the *inter*-class information is also not sufficient for classification. (3) Due to the introduction of the *intra*-class compactness as well as the *inter*-class separability into the regularization term, the boundaries derived from DR actually more accord with the total distribution of the samples ((d), (h), (l)). Hence, it always has the best training and testing accuracies in the three Two-moon datasets.

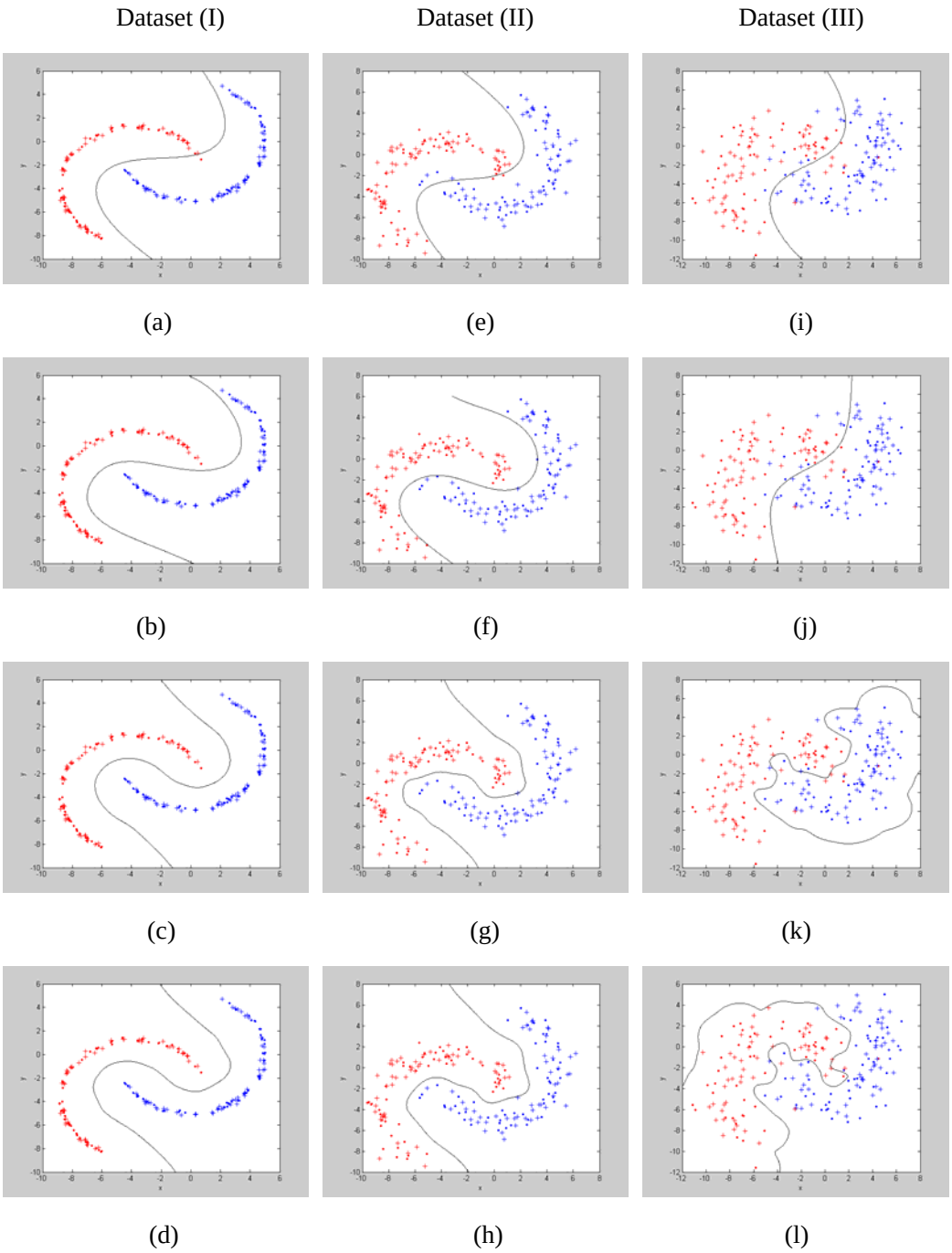

Fig. 1. The discriminant boundaries in three Two-Moon datasets: RN ((a), (e), (i)), RBFN ((b),

(f), (j)), SVM ((c), (g), (k)) and DR ((d), (h), (l))

Table2. Training and testing accuracies (%) compared between RN, RBFN, SVM and DR in the three Two-Moon datasets

	RN	RBFN	SVM	DR
Dataset (I)	99.00	*100.00*	*100.00*	*100.00*
	100.00	*100.00*	*100.00*	*100.00*
Dataset (II)	95.00	99.00	*100.00*	*100.00*
	98.00	100.00	*100.00*	*100.00*
Dataset (III)	92.00	92.00	97.00	*99.00*
	90.00	90.00	92.00	*95.00*

References

[1] A.N. Tikhonov, On solving incorrectly posed problems and method of regularization. Doklady Akademii Nauk USSR, vol.151, 501-504, 1963.

[2] A.N. Tikhonov and V.Y. Aresnin, Solutions of Ill-posed Problems. Washington, DC:W.H. Winston, 1977.

[3] V.A. Morozov, Methods for Solving Incorrectly Posed Problems, Springer-Verlag, 1984.

[4] S. Haykin, Neural Networks: A Comprehensive Foundation, Tsinghua University Press, 2001.

[5] Z. Chen and S. Haykin, On different facets of regularization theory. Neural Computation, vol.14(12), 2791-2846, 2002.

[6] T. Poggio and F. Girosi, Networks for approximation and learning. Proc. of the IEEE. vol.78, 1481-1497, 1990a.

[7] T. Poggio and F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks. Science, vol.247, 978-982, 1990b.

[8] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), Nonparametric functional estimation and related topics, 561-576, 1991.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Wiley, 2001.

[10] B. Haasdonk and H. Burkhardt, Invariant kernel functions for pattern analysis and machine learning. Machine Learning, vol.68, 35-61, 2007.

[11] A.V. Balakrishnan, Applied Functional Analysis, New York: Springer-Verlag, 1976.

[12] V. Vapnik, Statistical Learning Theory, Wiley, 1998.

[13] N. Cristianini and J.S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[14] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines. Advances in Computational Mathematics, vol.13(1), 1-50, 2000.

[15] H. Li, T. Jiang, and K. Zhang, Efficient and robust feature extraction by maximum margin criterion. IEEE Trans. on Neural Networks, vol.17(1), 157-165, 2006.

[16] A. Martinez and A. Kak, PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.23(2), 228-233, 2001.

[17] H. Xiong, M.N.S. Swamy, and M.O. Ahmad, Optimizing the kernel in the empirical feature space. IEEE Trans. on Neural Networks, vol.16(2), 460-474, 2005.

[18] J.A.K. Suykens and J. Vandewalle, Least squares support vector machine classifiers. Neural Processing Letters, vol.9, 293-300, 1999.

[19] T. Evgeniou, C.A. Micchelli, and M. Pontil, Learning multiple tasks with kernel methods. J. Machine Learning Research, vol.6, 615-637, 2005.
[20] C.A. Micchelli and M. Pontil. Kernels for multi-task learning. NIPS, 2004.
[21] C.A. Micchelli and M. Pontil, On learning vector-valued functions. Neural Computation, vol.17, 177-204, 2005.
[22] S. Szedmak and J. Shawe-Taylor, Muticlass learning at one-class complexity. Technical Report No: 1508, School of Electronics and Computer Science, Southampton, UK, 2005.
[23] E. Pekalska, P. Paclik, and R.P.W. Duin, A generalized kernel approach to dissimilarity-based classification. J. Machine Learning Research, vol.2, 175-211, 2001.