Transduction with Confidence and Credibility

C. Saunders, A. Gammerman, V. Vovk
Royal Holloway, University of London
Egham, Surrey, England.
{craig,alex,vovk}@dcs.rhbnc.ac.uk

Abstract

In this paper we follow the same general ideology as in [Gammerman *et al.*, 1998], and describe a new transductive learning algorithm using Support Vector Machines. The algorithm presented provides confidence values for its predicted classifications of new examples. We also obtain a measure of "credibility" which serves as an indicator of the reliability of the data upon which we make our prediction. Experiments compare the new algorithm to a standard Support Vector Machine and other transductive methods which use Support Vector Machines, such as Vapnik's margin transduction. Empirical results show that the new algorithm not only produces confidence and credibility measures, but is comparable to, and sometimes exceeds the performance of the other algorithms.

1 Introduction

In this paper, we describe a new method of *transductive inference* using Support Vector machines [Vapnik, 1995]. Whereas induction tries to learn a general rule (e.g. of classification) from a given training set, transduction reasons from particular to particular. That is, instead of trying to obtain a general rule, the learning process is focussed on obtaining the classification of a single new example, or given set of new examples. In section 3 we introduce a method of transduction based on a Support Vector (SV) machine which uses the statistical measure of p-values. By measuring p-values the algorithm gives confidence values for each of its predictions. The method also provides a credibility measure based on the p-values for different predictions. These measures can be interpreted as an indication of the quality of our prediction. The performance of the new algorithm is then compared to two other techniques (which simply give flat predictions, and no measure of confidence or credibility), viz. a standard Support Vector Machine, and Vapnik's margin transduction.

Results show that the transductive method presented here is comparable to, and sometimes exceeds the performance of the other two methods, whilst providing the additional information of confidence and credibility values for its prediction. Our trasnductive algorithm therefore gives us the best of both worlds: as in [Gammerman *et al.*, 1998; Gammerman, 1997] it provides confidence and credibility values (the predictive performance however, of the algorithm described in [Gammerman *et al.*, 1998] is poor; this is probably explained by the "distortion phenomenon" : see [Gammerman *et al.*, 1998], section 8.2 for details); as in Support Vector Machines, it achieves good predictive performance.

2 SV Implementation

In this section we describe the method upon which the transduction algorithms used in this paper are based. The method involves adding k examples to a training set and then training a separate SV machine for every possible classification of the k examples. Although the two transduction algorithms discussed here (our new algorithm and Vapnik's margin technique) both use this as a basis, the method of prediction, and any additional information (such as confidence and possibility) about the test examples which they produce, is different. The details of the algorithms will be presented in the next section, for now though we present the general ideas which are common to both.

Suppose we have some training data

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l) \;\; y_i \in \{-1, 1\} \;\; \mathbf{x}_i \in \mathbb{R}^n, \qquad (1)$$

and a set of test data,

$$\mathbf{x}_1^*, \dots, \mathbf{x}_k^*. \qquad (2)$$

(Note: as with a Support Vector Machine, we assume that both the training and test data are generated independently from the same distribution.) For a fixed set of classifications of the test data

$$y_1^*, \dots, y_k^*, \qquad (3)$$

we construct a Support Vector Machine on the combined sequence

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l), (\mathbf{x}_1^*, y_1^*), \dots, (\mathbf{x}_k^*, y_k^*). \qquad (4)$$

For simplicity we will only consider the separable case, however the following can easily be generalised to the non-separable case (for details see e.g. [Vapnik, 1995]). We therefore want to find an optimal hyperplane $\mathbf{w}$ such that

$$\frac{1}{2}||\mathbf{w}||^2 \tag{5}$$

is minimised, subject to the constraints

$$y_i[(\mathbf{x}_i \cdot \mathbf{w}) + b] \geq 1, \ i = 1, \ldots, l, \tag{6}$$

$$y_j^*[(\mathbf{x}_j^* \cdot \mathbf{w}) + b] \geq 1, \ j = 1, \ldots, k. \tag{7}$$

In order to find the optimal hyperplane we have to solve the following quadratic optimisation problem: maximise

$$\begin{aligned} W(\alpha, \alpha^*) &= \sum_{i=1}^{l} \alpha_i + \sum_{j=1}^{k} \alpha_j^* \\ &-\frac{1}{2} \sum_{i,r=1}^{l} \alpha_i \alpha_r y_i y_r K(\mathbf{x}_i, \mathbf{x}_r) \\ &-\frac{1}{2} \sum_{j,r=1}^{k} \alpha_j^* \alpha_r^* y_j^* y_r^* K(\mathbf{x}_j^*, \mathbf{x}_r^*) \\ &-\sum_{i=1}^{l} \sum_{r=1}^{k} y_i y_r^* \alpha_i \alpha_r^* K(\mathbf{x}_i, \mathbf{x}_r^*), \end{aligned} \tag{8}$$

subject to the constraints

$$\alpha_i \geq 0, \ i = 1, \ldots, l, \tag{9}$$

$$\alpha_j^* \geq 0, \ j = 1, \ldots, k, \tag{10}$$

$$\sum_{i=1}^{l} y_i \alpha_i + \sum_{j=1}^{k} y_j^* \alpha_j^* = 0. \tag{11}$$

Here $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function - a general expression for the inner product in Hilbert space. According to Hilbert-Schmidt theory $K(\mathbf{x}_i, \mathbf{x}_j)$ can be any symmetric function that satisfies Mercer's conditions (for details see [Vapnik, 1995]).

The process above is repeated for all possible classifications $y_1^*, \ldots, y_k^*$ of the test set. In the following sections we shall describe how the two algorithms use this process in different ways.

3 Transduction Algorithms

The method of transduction which we introduce here uses the statistical measure of p-values to determine the significance of the α_i-value(s) associated with the test example(s), once the quadratic optimisation problem (8)-(11) has been solved for all possible classifications of the test set. This method not only gives predicted classifications, but also provides valid measures of confidence and credibility for its predictions [Vovk and Gammerman, 1999]. First of all we shall consider the case when the test set to be classified only contains one example and there are two possible classifications. This will then be extended to a multi-class classification. Finally we shall consider problems involving multiple test examples and binary classes.

3.1 Confidence values for a single test example

If we are only interested in the binary classification of a single example, then the quadratic optimisation problem (8) has to be solved twice (once where the new example is classified as +1, the other −1), and therefore two hyperplanes are obtained. For each hyperplane, we obtain a value of "strangeness" for the test example. This is defined as follows.

Consider the training set, with the new test example included,

$$\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{new}.$$

Once (8) has been solved, each of the examples in this set has an associated Lagrange multiplier

$$\alpha_1, \ldots, \alpha_l, \alpha_{new}.$$

Suppose we are interested in the probability that the α_{new} is actually the largest Lagrange multiplier in this set. Since the training examples and the one new example are exchangeable, then this probability is

$$P\{\alpha_{new} > \max_{1 \leq i \leq l} \alpha_i\} \leq \frac{1}{l+1},$$

(the randomisation is over all permutations of the examples). The value of the Lagrange multiplier can be interpreted as a measure of "supportiveness" of the example, and therefore high values indicate that this example is "strange" and unlikely to occur. In order to determine how unlikely a certain α-value is, we can use the statistical measure of p-values.

Simply examining whether or not the α-value associated with the new test example is the highest or not and accepting or rejecting it as the correct classification based on this alone would not produce a reliable classifier. We therefore look at the p-value associated with the α-value and make a decision based on this value. If the rank of α_{new} is n (i.e. α_{new} is the nth highest α-value), the p-value is defined as

$$\text{p-value} = P\{\text{rank}(\alpha_{new}) \geq n\}, \tag{12}$$

(once again, the randomisation is over all permutations of the examples), which is equivalent to

$$\text{p-value} = \frac{\#\{i : \alpha_i \geq \alpha_{new}\}}{l+1}. \tag{13}$$

The p-value is a measure of how "strange" our test example is when given a certain classification. That is, the p-value tells us the probability of observing this particular ordering of the alpha values under the assumption that y_{new} is the correct classification. The classification y_{new} which yields the highest corresponding p-value, determines the classification predicted by the algorithm. The confidence in prediction can then be defined as

$$\text{Confidence } = 1 - P_2, \tag{14}$$

where P_2 is the p-value obtained when the example was given the classification which we did not predict.

3.2 Classifying multiple new examples

If we are to consider the case when multiple new examples $\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*$ are added to the existing training set, then the QP problem (8)-(11) has to be solved a total of 2^k times (in a two-class scenario). Unfortunately this is impractical for large values of k (e.g. $k \geq 7$) Each solution of this problem yields Lagrange multipliers corresponding to each of the test examples. The p-value associated with a particular assignment of classifications is then defined as

$$\text{p-value} = P\{(\text{rank}(\alpha_1)+\ldots+\text{rank}(\alpha_k)) \geq (n_1+\ldots+n_k)\},$$

where $n_1, \ldots, n_k$ are the actual measured ranks of the corresponding Lagrange multipliers $\alpha_1, \ldots, \alpha_k$, and the randomisation is over all permutations of the examples. As in the single example case, the classifications predicted for the test examples are those which yield the highest p-value. Confidence is defined to be $1 - P_2$ where P_2 is the second highest p-value, and as in the single example case, P_1 (the highest p-value) corresponds to our credibility.

3.3 Measure of Credibility

Not only do we obtain a confidence value for our prediction, but we also consider a measure of credibility which indicates the quality of the data on which we base our decision. We define credibility as the value P_1, i.e. the p-value obtained when the test data are given the predicted classification(s).

In order to see how this can be interpreted as a measure of the quality of the data, first consider an "ideal" case. Suppose we are adding a single test example. Also suppose that when the correct classification is given to our test example it is not a support vector and therefore will have a p-value of 1. Assume that when given the incorrect classification, the p-value obtained from the example is at most 0.05. In this situation, our confidence would be 95% or greater and the value of credibility would be 1 (100%). This would mean we have high confidence in our prediction from a good set of data. Now consider a similar case where the highest p-value still corresponds to the correct classification, but is much lower, say 0.3. If the other p-value obtained was the same as before, then we would still have a high confidence of 95%. Our measure of credibility however would be much lower (30%). This would convey the meaning that although we confidently rejected all other classifications of this test example, the test example is actually "strange" in both scenarios and therefore the data is not sufficient to give us a totally secure prediction. Section 4.1 introduces empirical evidence which supports this line of reasoning.

The measure of credibility provides us with a filter mechanism with which we can "reject" certain predictions. That is, if for any task the consequences of making a wrong prediction are quite severe, we can choose to reject those predictions which have a low credibility value associated with them. The more severe the consequences for making an incorrect prediction are, the higher we can set the rejection threshold.

3.4 Vapnik's Margin Transduction

As a point of comparison for our technique we shall use a method of transduction suggested by Vapnik [Vapnik, 1998]. This method also uses the basic ideas described in section 2. The predicted classifications are the ones which separate the joint sequence

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), (\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_k^*, y_k^*), \qquad (15)$$

with maximal margin.

The predicted classifications are therefore given by

$$\arg \min_{y_1^*, \ldots, y_k^*} \min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2, \qquad (16)$$

(subject to constraints (6) and (7)), over all possible classifications $y_1^*, \ldots, y_k^*$.

In the dual representation, this is equivalent to maximising (8)-(11) for all possible classifications of the test set, and predicting the classifications which achieve the overall minimum.

4 Experiments and Results

First of all we shall present some empirical evidence of the quality of the confidence and credibility values obtained by the new transductive algorithm, based on a two-class digit recognition problem. A performance comparison is then made between our new algorithm, Vapnik's Margin algorithm, and a standard SV machine on the same data set.

Unless stated otherwise, the experiments in this section were performed on the US Postal Service database of handwritten digits (see e.g. [LeCun *et al.*, 1990]).

The kernel function used in these experiments was a polynomial of the form

$$K(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} \cdot \mathbf{y})^d}{256},$$

for which the best performance is achieved with $d = 3$.

4.1 Confidence and Credibility Values

Table 1 shows an example of the confidence and credibility values obtained on a digit-recognition task of separating the digit '8' from all other digits. The training set used in these experiments consisted of 49 examples of the digit '8' and 451 examples of other digits ('0'...'9'). The test set consisted of 100 other digits from the database. Both the new transduction algorithm and an SV machine were run on the data. Out of 100 test examples both methods classified all but three examples correctly (they both misclassified the same three examples). The table shows the confidence and credibility values for these three misclassified examples (along with the examples themselves). For all of the misclassified examples the credibility of the prediction is very low (no more than 5%). This suggests that for all of these examples, the quality of the data is not sufficient on which to base a prediction.

	Misclassified Examples		
Example No.	1	2	3
Example			
True Class	8	8	8
Confidence	95%	96%	99%
Credibility	4.6%	4.5%	0.8%

Table 1: Confidence and credibility values for miss-classified examples.

Method	$n = 20$	$n = 40$	$n = 100$	$n = 200$
p-value Trans	640	390	**262**	186
Margin Trans	526	**355**	268	**180**
Standard SVM	**522**	**355**	272	191

Table 2: Incorrect classifications over a total of 20000 runs. In addition to providing confidence and credibility values, the p-value transductive algorithm has good generalisation ability.

4.2 Relative Performance of the Algorithms

In this section we compare the predictive performance of the new algorithm, alongside the margin transduction technique and a Support Vector Machine. For this experiment, we again used a subset of the digit database. All of the examples of the digits 2 and 7 were extracted from the database, giving a total set of 1721 examples. In each of these experiments a subset of n examples were randomly chosen and used as a training set. A single further example was then randomly picked as a test example. All three algorithms were trained on the same training set and gave their predictions for the test example. This process was then repeated for a total of 20000 runs, and for different values of n. Table 2 summarises these results. It is clear from the table that the new algorithm does not suffer in performance despite providing the extra information of confidence and credibility values.

5 Discussion

In this section we briefly discuss related work and highlight possible directions for further research based on the results presented in this paper.

5.1 Adding multiple examples

At the present time, adding k examples to our original training set in order for them to be classified is impractical for large values of k. Recent developments in the training of Support Vector machines, however, such as those presented in [Platt, 1998] may yield improvements in the application of this algorithm. Another transductive algorithm has recently been proposed in [Bennett and Demiriz, 1998] which is based on the margin transduction technique. This technique minimises the $\mathbf{w}$ vector in the L_1-norm rather than the L_2-norm and uses integer programming to rapidly find hyperplanes which separate the training data, even if the number of examples is large. This method however, does not provide confidences or credibility values for its predictions.

5.2 Extension to Regression

An important direction of this research is to extend it to the case of regression, i.e. where the classifications y_i are no longer required to be binary values, but can be real numbers. Statistically valid p-values may be obtainable from Support Vector Machines for regression estimation (see e.g. [Vapnik, 1998]), or other related methods such as those in [Saunders *et al.*, 1998].

6 Conclusion

In this paper we have presented a new transduction algorithm which is based on the Support Vector technique. It has been shown that the algorithm produces confidence values for its predictions, and also gives a measure of credibility which indicates the quality of data upon which the prediction is based, and therefore serves as a guideline of how reliable the prediction actually is. Empirical results have been presented which show that values of confidence and credibility produced by the algorithm do correctly reflect the reliability of the predictions given. This method has been shown not only to produce these values, but also to have good generalisation ability on a test set, comparable to and sometimes exceeding the results achievable by a Support Vector Machine.

7 Acknowledgements

This work was partially supported by EPSRC GR/L35812 and GR/M15972, and EU INTAS-93-725-ext grants. In addition we are indebted to the support provided by IFR Ltd. We would also like to an anonymous referee whose helpful comments prompted many changes and improvements in the paper.

References

[Bennett and Demiriz, 1998] K. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Proceedings of Neural Information Processing Systems*, 1998.

[Gammerman *et al.*, 1998] A. Gammerman, V. Vapnik, and V. Vovk. Learning by transduction. In *Uncertainty in Artificial Intelligence*, pages 148–155, 1998.

[Gammerman, 1997] A. Gammerman. Machine learning : Progress and prospects. RHUL, ISBN 0900145935, 1997.

[LeCun *et al.*, 1990] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel. "Handwritten digit recognition with back-propagation network,". *Advances in Neural Information Processing Systems*, pages 396–404, 1990.

[Platt, 1998] J. Platt. Sequential minimal optimisation: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

[Saunders *et al.*, 1998] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *ICML '98. Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann, 1998.

[Vapnik, 1995] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Vapnik, 1998] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Vovk and Gammerman, 1999] V. Vovk and A. Gammerman. Algorithmic theory of randomness. Forthcoming Book, 1999.