Experimental Evaluation of Multilayer Perceptrons with Entropic Risk Functionals on Real-World Datasets

Luís A. Alexandre, Luís M. Silva, Jorge Santos, J.P. Marques de Sá

Abstract

We investigate the performance of MLPs with four risk functionals: the classical mean square error (MSE), the cross-entropy (CE), a generalized exponential risk (EXP), and the Shannon entropy of the classifier's output error (HS). The performance is compared with an SVM with RBF kernel in terms of average balanced and unbalanced error rates, and their generalization, on practical classification tasks. For this purpose we carried out experiments on 35 public real-world datasets.

A battery of statistical tests applied to the experimental results showed no significant difference among the classifiers in terms of unbalanced error rates. However, in terms of balanced error rates SVM-RBF performed significantly worse than MLP-CE and MLP-EXP. Regarding generalization, SVM-RBF and MLP-EXP scored as the classification methods with significantly better generalization, both in terms of balanced and unbalanced error rates.

1 Introduction

Support Vector Machines were proposed by Vapnik (see namely the 1999 seminal monograph [20]) as a classifier type endowed with optimal generalization ability, given the constraint on the norm of the weight vector and consequent constraint on the Vapnik-Chervonenkis dimension. In a 2004 work, Collobert and Bengio [5] elucidated the links between SVMs and MLPs. After showing that under simple conditions a perceptron is equivalent to an SVM, they have also shown that the early stopping rule used in stochastic gradient descent training of MLPs is a regularization method that constraints the norm of the weight vector, and therefore improves its generalization ability.

In the present work we present experimental evidence on the compared classification performance of SVMs and MLPs, in real world datasets. We use four different types of MLPs characterized by using different empirical risks

$$\hat{R}_L(X) = \sum_{t \in T} P(t) \sum_{x \in X} L(t, y(x)) \ , \qquad (1)$$

where X denotes the input set, t and $y(x)$ represent the desired output and the classifier output, respectively, $P(t)$ are the priors, and $L(\cdot)$ is a loss function (also known as error function). For instance, the square error loss $L(t, y(x)) = (t(x) - y(x))^2$ is a popular choice in machine learning algorithms; for this choice of loss function, $\hat{R}_L(X) = \hat{R}_{MSE}(X) =$ is the well-known Mean Square Error (MSE) risk.

Traditionally, the role played by different risk functionals in classifier performance has been somewhat overlooked. There has been a persistent belief that the choice of loss function is more a computational issue than an influencing factor in classifier performance [4, 20]. This way of thinking has been shown by Rosasco and co-workers [14] to be incorrect for the loss functions used by support vector machines. These authors have namely shown that the choice of loss function influences the convergence rate of the empirical risk towards the true (theoretical) risk. In what concerns classifiers, our main concern is their probability of error, not the minimum risk. To this respect one must note that a minimum of the risk does not necessarily imply a minimum of the probability of error [10, 12]. The work of Silva et al., [19] also presents an example clearly showing that different risk functionals may behave quite differently in what concerns the attainment of the minimum probability of error allowed by the classifier architecture.

The paper is organized as follows: section 2 presents the SVM and risk functionals used in our experiments with MLPs; section 3 presents the datasets; section 4 describes the experimental settings, performance measures, and statistical methods used to draw inferences from the experiments; section 5 presents the results which are finally discussed in the concluding section 6.

2 Classifiers

2.1 SVM

Since its introduction, SVMs have been increasingly used as the standard classifier for classification problems given their asymptotical convergence properties.

Among the SVMs with general purpose kernels (such as linear kernel, polynomial kernel, etc.), the SVM with an RBF (Radial Basis Function) kernel is widely used in many different applications due to its excelent performance and to the fact that it has only two parameters to be adjusted, thus reducing the time to develop a usable classifier.

In this paper we used the SVM-RBF classifier, with inputs normalized in the $[-1, 1]$ interval and the following kernel

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2),\ \gamma > 0$$

where x_i and x_j are data points and γ is a parameter inversely proportional to the kernel bandwidth. Hence, the two free parameters to be set are $C > 0$ that corresponds to the penalty error parameter and γ.

2.2 Risk Functionals

Four different risk functionals were used by the MLPs in the experiments. Besides the classic Mean Square Error (MSE) and Cross-Entropy (CE) risks, we used two other ones, recently developed by the authors. One of the risks is the information-theoretic Shannon's entropy of the error (HS), which uses a Parzen window estimate of the error probability density function (PDF). Details on this procedure can be found in [19]. HS was studied both theoretically and experimentally by Silva et al. [17, 19]. The second unconventional risk is the generalized exponential (EXP) risk. This is in fact a sort of meta-risk capable of emulating a whole series of risk functionals (MSE included). It was described and studied in Silva et al. [18]. The following list presents the formulas of the empirical risks ($\hat{R}_L$), which are directly plugged in the gradient descent formulas of MLP training:

1. Mean square error (MSE):

$$\hat{R}_{MSE}(X) = \sum_{t\in T} P(t) \sum_{x\in X_t} (t - y(x))^2$$

2. Cross-entropy (CE):

$$\hat{R}_{CE}(X) = \sum_{t\in T} P(t) \sum_{x\in X_t} (t \ln y(x) + (1-t)\ln(1-y(x)))$$

3. Generalized exponential (EXP):

$$\hat{R}_{CE}(X) = \sum_{t\in T} P(t) \sum_{x\in X_t} (\tau e^{t-y(x))^2)/\tau}$$

4. Shannon's entropy of the error (HS):

$$\hat{R}_{HS}(x) = -\frac{1}{n}\sum_{i=1}^{n} \ln \sum_{j=1}^{n} K_h(x_i - x_j) \ . \tag{2}$$

3 Datasets

The SVM and MLP classification algorithms were applied to 35 public real-world datasets presented in Table 1, which are quite diverse in terms of number of instances, features, and classes. They are from the well-known UCI repository [1], except the following ones: Cloud5f which is the same as the Cloud dataset from the statlib archive, after removing a single nominal feature; Olive is from [8]; Pb12 is from [11]; Telugu is from [13]. The LRS100f dataset is the same as the Low Resolution Spectrometer dataset from [1] after removal of its single nominal feature (from the original 101 features). We removed classes *omL*, *imL* and *imS* from the E-coli dataset because of their unreasonably low number of instances (respectively, 5, 2, 2).

Previous works using the Breast dataset (see, e.g., the work cited in [1]) have shown that its *fad*, *mas*, and *gla* classes have a large overlap and cannot be discriminated in any reasonable way; this led us to merge them and set up the Breast4 dataset, specially aimed at a more reliable detection of the relevant carcinoma (*car*) class. A similar reason led us to set up the E-coli4 dataset by merging classes *im* and *imU* of E-coli.

4 Experiments

4.1 Experimental Settings

The SVM implementation used the libSVM [3] library. In order to determine adequate SVM parameters a joint search of the best C ($C > 0$ is the SVM regularization parameter) and γ (the inverse bandwidth parameter of the RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2))$ was performed in a grid with $\log_2 C$ in $\{-2, -1, 0, 1, 2\}$ and $\log_2 \gamma$ in $\{-10, -9, \ldots, -1, 0\}$. This search was performed with stratified 2-fold CV with one repetition.

All MLPs had the same architecture, namely with one-hidden layer and, for each dataset, a fixed number of hidden nodes. The MLPs were all trained with the well-known back-propagation algorithmic procedure (based on gradient descent), the only available training procedure for the HS and EXP risk functionals. The four MLP algorithms, each plugging into the back-propagation procedure the gradient formulas of the risk functionals, were implemented in MATLAB. In order to determine an adequate number of hidden neurons, n_h, to be used for each dataset, several preliminary experiments were performed with the MSE risk. For this purpose, we repeated 10 times (for each dataset) an error rate evaluation using stratified ten-fold cross-validation (CV10), or two-fold cross-validation (CV2) for datasets with less than 50 instances per class. We also took into account the well-known rule of thumb $n_h = w/\hat{P}_e$ (based on a formula given in [2]), where w is the number of weights and $\hat{P}_e$ is the expected error rate. Once the MLP architecture was defined we proceeded to the selection of the number of epochs of MLP training (early stopping rule) as well as the choice of specific algorithm parameters, namely the value of the bandwidth h in the Parzen window estimation of the error PDF needed for HS, and the τ parameter of the EXP risk. This parameter selection task was based on series of 10 experiments for each algorithm and dataset.

In all experiments the input data was normalized with zero mean and unit variance.

4.2 Performance Measures

For each dataset, 20 repetitions of the cross-validation procedure (CV10 or CV2 for datasets with less than 50 instances per class) were carried out. Each of the $i = 1, \ldots, 20$ repetitions provided average training (design) set and test set classification matrices (also known as confusion matrices), from which the respective error rates were computed. Since the error rates are estimates of probability of error they are denoted, respectively, $\hat{P}_{ed}(i)$ and $\hat{P}_{et}(i)$.

For each classification matrix the balanced error rates (BER), defined as $\hat{P}_b = \sum_{k=1}^{c} \hat{P}_{e(k)}$, where the $\hat{P}_{e(k)}$ represent error rates for each of the c classes, were also computed. BER is considered to be more appropriate as a performance index for unbalanced datasets, since it assumes equal priors. As for the error rates, training (design) set and test set estimates were computed for the BER quantities: $\hat{P}_{bd}(i)$ and $\hat{P}_{bt}(i)$.

From the 20, $\hat{P}_{ed}(i)$, $\hat{P}_{et}(i)$, $\hat{P}_{bd}(i)$ and $\hat{P}_{bt}(i)$, the following performance measures were computed:

1. Sample means, denoted as $\bar{P}_{ed}$, $\bar{P}_{et}$, $\bar{P}_{bd}$ and $\bar{P}_{bt}$, using the bar as sample mean operator. (In rigor, $\bar{P}_{ed}$ is $\bar{\hat{P}}_{ed}$, and likewise for the other quantities; we use $\bar{P}_{ed}$ for notational simplicity reasons.)

2. Sample standard deviations, denoted as sP_{ed}, sP_{et}, sP_{bd}, and sP_{bt}.

3. The pooled means: $\bar{P}_e = (\bar{P}_{ed} + \bar{P}_{et})/2$, $\bar{P}_b = (\bar{P}_{bd} + \bar{P}_{bt})/2$.

4. The pooled standard deviations: $sP_e = (sP_{ed}^2/2 + sP_{et}^2/2)^{1/2}$, $sP_b = (sP_{bd}^2/2 + sP_{bt}^2/2)^{1/2}$.

5. The generalization measures: $D_e = \bar{P}_{et} - \bar{P}_{ed}$, $D_b = \bar{P}_{bt} - \bar{P}_{bd}$.

We will focus our attention on performance measures 3, 4, and 5.

The pooled mean $\bar{P}_e$ is a better estimate of the true probability of error of the classifier than either the average training set error, $\bar{P}_{ed}$, or the average test set error, $\bar{P}_{et}$. As a matter of fact, the true probability of error, P_e, would only be computable (in principle) if one knew the data distributions. An arbitrarily close estimate of Pe would also be determinable, if one disposed of an arbitrarily large number of data instances, n, profiting from the convergence of the empirical error to the true error with $n \to \infty$. For consistent learning algorithms as the ones we use, it is a known fact that $\bar{P}_{ed}$ converges from below to P_e with $n \to \infty$ (the training set estimate is optimistic on average), and $\bar{P}_{et}$ converges from above (the test set estimate is pessimistic on average). Figure 1 illustrates this asymptotic property of the average error rates for the Ozone dataset and the SVM and CE classifiers (see also [20]), by plotting the $\bar{P}_{ed}$ and $\bar{P}_{et}$ values with the standard deviations obtained in

20 CV10 experiments for a grid of n values in $[50, 1800]$ with increments of 50 instances (learning curves). This convergence property justifies the use of $\bar{P}_e$ as a more reliable estimate (see e.g. [15]). Similar considerations apply to the use of the $\bar{P}_b$ performance measure. For a given dataset, in order to assess the statistical significance of $\bar{P}_e$ ($\bar{P}_b$) for the various classification methods one needs the information conveyed by sP_e (sP_b).

The concept of classifier generalization implies empirical risk estimates of the probability of error that are close to its true value. The use of D_e and D_b as generalization measures is then appropriate (see also [21]).

4.3 Statistical Methods

The performance measures obtained for the six risk functionals were statistically evaluated following recommendations in [6, 9, 22, 16], by namely applying the following multiple comparison tests:

1. The Friedman test: The Friedman test is the non-parametric equivalent of the two-way Anova. It is clearly adequate to multiple comparison of scores depending on two influencing factors; in our case, these are the classifier method and the dataset. The Friedman test is recommended by several authors in applications such as ours (see, namely, [9]). When the test produces a statistically significant result (we will always set the significance level at $p = 0.05$), one may then proceed to apply post-hoc tests, namely the Dunn-Sidak test for multiple comparison and the Finner test for comparison of a chosen reference method against any of the other ones. The Finner test for post-hoc comparisons of a proposed method against another was analyzed in [9] and found to be more powerful than competing tests.

2. The multiple sign test: This test is described in [9] and is specially suited to the comparison of each method against a fixed one, in a multiple comparison context. It doesn't need any previous application of another test, as the post-hoc tests mentioned above do.

3. Counts of wins and losses: This is a traditional and simple multiple comparison method. For each method, the number of datasets where it produces the best (win) and worst (loss) results is computed. For the $\hat{P}_e$, $\hat{P}_b$, D_e and D_b scores, "best" means smallest. One then proceeds to apply an adequate statistical method to the wins and losses; in the

following we apply the chi-square test of goodness of fit to the uniform distribution (the null hypothesis is that there are no differences among the methods). Note that in all statistical methods described so far the information of the standard deviation has not been used. This means, for "counts of wins and losses" as it is traditionally used, that for instance a method with error rate 10.00% will win over another method with error rate 10.01% even if the pooled standard deviation is 1%, or whatever value for that matter. This is clearly inadequate. We apply, however, a variant of "counts of wins and losses" that takes this information into account: for each dataset, instead of determining the absolute winning and losing methods, we determine the statistically significant winning and losing methods. For this purpose, we apply the one-way Anova test to each dataset and the post-hoc Tukey's least significant difference criterion if the test produces a significant p; otherwise, the more strict Tukey's honestly significant difference criterion is used [22]. Note that it is known that the results of the one-way Anova change very little by moderate violations of the assumption of normal distribution and equal variance especially for not too small sample size, as in our case (the sample size is 20; see e.g., [7]).

We now present two examples of the "counts of wins and losses" method. Consider the error rate scores (percentages) with the corresponding standard deviations inside parentheses presented in Table 2 (they correspond to the third dataset of Table 4). In this case the absolute win is CE (smallest error rate) and the absolute loss is MSE. The one-way Anova, however, finds no significant difference among the methods. This clearly seems a more reasonable conclusion, taking into account the standard deviations.

The second example is in Table 3 (second dataset of Table 4). The absolute win is HS and the absolute loss is SVM. The one-way Anova finds the methods significantly different. The post-hoc test assigns the methods to three groups: the wining group EXP, HS, the losing group SVM, and the intermediary group MSE, CE. Again, this seems a more reasonable decision.

Besides the above multiple comparison tests we also applied the Wilcoxon paired rank-sum test for comparisons of pairs of algorithms.

5 Results

5.1 Error Rates

Table 4 presents the values of $\hat{P}_e$ $(s\hat{P}_e)$ for all algorithms and datasets, with the statistically significant wins and losses (as always, at $p = 0.05$).

The Friedman test did not detect significant differences ($p = 0.11$) of the scores for the 35 datasets. The mean ranks of the five methods (following from now on Table 4 order) are: 3.23, 3.21, 2.57, 2.64, and 3.34.

The multiple sign test did not detect any significant difference from SVM considered as the reference algorithm, versus any of the MLPs, in the context of multiple comparison. Table 6 shows the computations. The critical value of the sum of minuses at $p = 0.05$ is 10; since all sums of minuses are above the critical value (10), the test does not reject the null hypothesis of equality of the methods. The same conclusion was arrived at in the paired comparison context with the Wilcoxon test.

The counts of significant wins and losses are 7, 5, 10, 11, 8 and 9, 8, 6, 5, 8, respectively, with the chi-square p well above 0.05: $p = 0.595$ for the wins and $p = 0.827$ for the losses. In 16 datasets the algorithms are tied.

5.2 Balanced Error Rates

Table 5 presents the values of $\hat{P}_b$ $(s\hat{P}_b)$ for all algorithms and datasets, with the statistically significant wins and losses.

The Friedman test found significant differences of the $\hat{P}_b$ scores for the 35 datasets ($p = 0.01$). The mean ranks of the five methods are: 3.54, 3.41, 2.45, 2.60, and 2.99. Although the post-hoc Finner test didn't find a significant difference of SVM compared against every other method (taking into account the multiple comparison setting), the post-hoc Dunn-Sidak test found the SVM score significantly worse than the CE score (see Figure 2).

The multiple sign test did not detect a significant difference of SVM compared against every other method (taking into account the multiple comparison setting). The counts of significant wins and losses are 7, 9, 14, 16, 12 ($p = 0.332$) and 14, 6, 5, 3, 5 ($p = 0.026$), with SVM scoring significantly more losses. In 12 datasets the algorithms are tied.

In the paired comparison context, the Wilcoxon test found SVM performing significantly worse than CE and EXP and with no significant difference from MSE and HS.

5.3 Error Rate Generalization

The Friedman test found a significant difference ($p = 0.003$) of the De scores for the 35 datasets. The mean ranks for the five methods are: 2.09, 3.16, 3.24, 3.09, and 3.43. The post-hoc Finner test found SVM responsible of this difference; SVM having significantly better generalization than any of the other algorithms. The Dunn-Sidak test only found a significant difference with respect to MSE, CE, and HS, i.e., according to the Dunn-Sidak test SVM doesn't generalize significantly better than EXP. On the other hand, the multiple sign test found that SVM doesn't generalize significantly better than HS.

The counts of significant wins and losses are 15, 4, 3, 2, 1 ($p \approx 0$) and 3, 8, 8, 11, 13 ($p = 0.156$). SVM scored, therefore, significantly more wins.

In the paired comparison context, the Wilcoxon test also confirmed SVM as generalizing significantly better than any of the other competitors.

5.4 Balanced Error Rate Generalization

The Friedman test found a significant difference ($p = 0.035$) of the D_b scores for the 35 datasets. The mean ranks for the five methods are: 2.29, 3.16, 3.14, 3.00, and 3.41. The post-hoc Finner test found SVM responsible of this difference; SVM having significantly better generalization than any of the other algorithms. The Dunn-Sidak test only found a significantly better generalization of SVM with respect to HS. On the other hand, the multiple sign test only found a significantly better generalization of SVM with respect to CE.

The counts of significant wins and losses are 10, 5, 3, 3, 2 ($p = 0.062$) and 4, 7, 8, 10, 9 ($p = 0.594$). SVM scored, therefore, significantly more wins.

In the paired comparison context, the Wilcoxon test did not find SVM generalizing significantly better than any of the other competitors.

6 Conclusions

Regarding the error performance $\hat{P}_e$, the Friedman test did not detect any significant difference among the algorithms. The Friedman test, however, "sees" the performance score table as a whole; by summing the ranks along the columns (algorithms) it is in some way gauging an average tendency.

Therefore, what the Friedman test results say of our experiments is that "on average" there is not a marked difference among the classifiers in terms of $\hat{P}_e$.

The multiple sign test, where an algorithm is compared against any of the remaining ones in the context of a multiple comparison, provides a different picture. Whereas in the Friedman test, ranks are assigned taking all methods together, in the multiple sign test a reference algorithm is ranked against a single algorithm; the "averaging" effect we mentioned above is less pronounced. Counts of significant wins and losses also help us to make a finer categorization of the algorithms.

In our experiments the $\hat{P}_e$ scores didn't show any significant differences both in terms of the multiple sign test and in the counts of significant wins and losses.

For the balanced error rate scores, $\hat{P}_b$, a different conclusion emerged. The Friedman test found a significant difference with SVM scoring significantly worse than CE (Dunn-Sidak test). Even though the multiple sign test did not detect a significant difference of SVM compared against every other method, the counts of significant wins and losses confirmed SVM performing significantly worse than CE; moreover, the Wilcoxon test also found SVM performing significantly worse than EXP.

In what concerns generalization, a significant difference among the algorithms was found by the Friedman test for both De and Db, with SVM scoring at least as one of the algorithms with significantly better generalization. Counts of significant wins and losses also confirmed this finding. The multiple sign test, however, didn't find a significant difference of SVM with respect to any of the other algorithms in terms of De; in terms of Db the superiorness of SVM was only found with respect to CE.

The post-hoc Friedman tests also point to a superiorness of SVM relative to CE and possibly to MSE (De) and HS (Db).

In the paired comparison context, the Wilcoxon test provided a remarkably different picture of the generalization of balanced and unbalanced error rates: whereas it found SVM generalizing significantly better than any of the other competitors for the unbalanced error rates, no difference was found for the balanced error rates. Taking into account all the results provided by the statistical tests, EXP appears therefore as a good competitor of SVM in terms of generalization for both balanced and unbalanced error rates. For the unbalanced error rates this conclusion can be extended to MSE.

References

[1] A. Asuncion and D.J. Newman. UCI machine learning repository. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, University of California, School of Information and Computer Science, 2010.

[2] E.B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] V. Cherkassky and F. Mulier. *Learning from data: concepts, theory and methods*. John Wiley & Sons, 1998.

[5] Ronan Collobert and Samy Bengio. Links between perceptrons, mlps and svms. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8,*, volume 69, 2004.

[6] J. Demzar. Statistical comparisons of classifiers over multiple data sets. *J. of Machine Learning Research*, 7:1–30, 2006.

[7] W.J. Dixon and F.J. Massey. *Introduction to Statistical Analysis*. McGraw-Hill Companies, 4th edition, 1983.

[8] M. Forina and C. Armanino. Eigenvector projection and simplified non-linear mapping of fatty acid content of italian olive oils. *Ann. Chem.*, 72:125–127, 1981.

[9] S. García, A. Fernández, J. Luengo, and F. Herrera. Advanced non-parametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180:2044–2064, May 2010.

[10] J. B. Hampshire and A. H. Waibel. A novel objective function for improved phoneme recognition using time-delay neural networks. *IEEE Tran. on Neural Networks*, 1(2):216–228, 1990.

[11] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, (3):79–87, 1991.

[12] M. Möller. *Efficient Training of Feed-Forward Neural Networks.* PhD thesis, Computer Science Department, Aarhus University, 1993.

[13] S. K. Pal and D. D. Majumder. Fuzzy sets and decision making approaches in vowel and speaker recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 7:625–629, 1977.

[14] L. Rosasco, E. De Vito, A. Caponnetto, and M. Piana. Are loss functions all the same ? *Neural Computation*, 16(5):1063–1076, 2004.

[15] Dobbins R.W. and R.C. Eberhart. *Neural Network PC Tools: A Practical Guide.* Academic Press, 1990.

[16] S.L. Salzberg. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, (1):317–327, 1997.

[17] L.M. Silva, J. Marques de Sá, and Luís A. Alexandre. Neural network classification using Shannon's entropy. In *13th European Symposium on Artificial Neural Networks - ESANN 2005*, pages 217–222, Bruges, Belgium, 2005.

[18] L.M. Silva, J.P. Marques de Sá, and Luís A. Alexandre. Data classification with multilayer perceptrons using a generalized error function. *Neural Networks*, 21(9):1302–1310, November 2008.

[19] Luís M. Silva, J. Marques de Sá, and Luís A. Alexandre. The MEE principle in data classification: A perceptron-based analysis. *Neural Computation*, 22(10):2698–2728, 2010.

[20] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, second edition, 1999.

[21] U. von Luxburg and B. Schölkopf. Statistical learning theory: Models, concepts, and results. In S. Hartmann D. Gabbay and J. Woods, editors, *Handbook for the History of Logic, Vol. 10: Inductive Logic.* Elsevier, 2011.

[22] A.C. Tamhane Y. Hochberg. *Multiple Comparison Procedures.* John Wiley& Sons, 1997.

Table 1: The datasets.

	Blood	Breast	Breast4	Cloud5f	Cork stop.
No. cases	748	106	106	108	150
No. features	4	9	9	2	10
No. classes	2	6	4	1	3
	E-coli	E-coli4	Glass	H. surv	Inflamm.
No. cases	327	327	214	306	120
No. features	5	5	9	3	6
No. classes	5	4	6	2	4
	Ionosphere	Iris	Jap-vowels	Libras	LRS100f
No. cases	351	150	196	360	531
No. features	34	4	12	90	100
No. classes	2	3	9	15	6
	Lung Cancer	Olive	Ozone	Parkinson	Pb12
No. cases	31	572	1847	195	608
No. features	55	8	72	22	2
No. classes	3	9	2	2	4
	P. Diabetes	Robot-1	Robot-4	Robot-5	SC-chart
No. cases	768	88	88	164	600
No. features	8	90	90	90	60
No. classes	2	3	4	5	6
Sonar	Spectf-Heart	Telugu	Thyroid	Vehicle	
No. cases	208	267	871	215	846
No. features	60	44	3	5	17
No. classes	2	2	6	3	4
	Wdbc	Wdbc-org	Wine	Wpbc	Yeast
No. cases	569	683	178	194	1479
No. features	30	9	13	32	6
No. classes	2	2	3	2	9

Table 2: First example of the counts of wins and losses method.

SVM	MSE	CE	EXP	HS
7.79 (1.79)	8.23 (3.88)	7.29 (3.34)	7.34 (3.26)	7.71 (3.52)

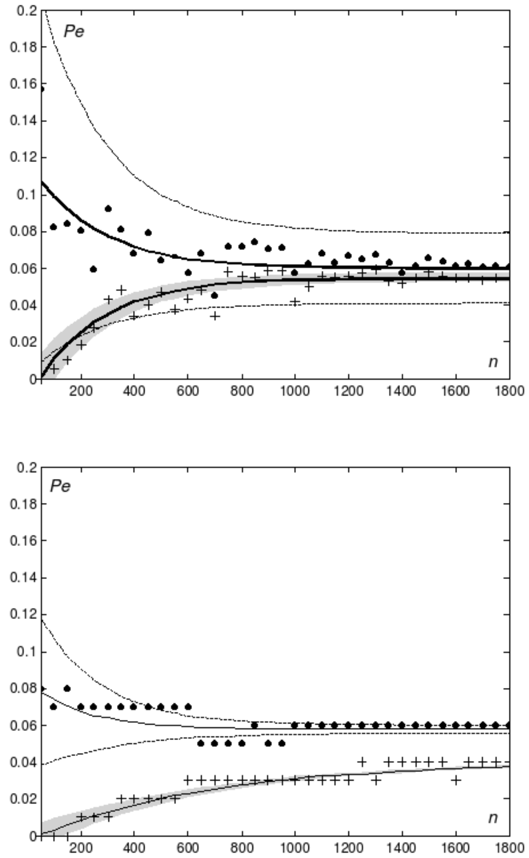

Figure 1: Learning curves for the Ozone dataset with the MLP-CE (left) and SVM-RBF (right) classifiers. The learning curves were obtained by exponential fits to the $\bar{P}_{ed}(n)$ ('+') and $\bar{P}_{et}(n)$ ('.') values. The shadowed region represents $\bar{P}_{ed} \pm sP_{ed}$); the dotted lines represent $\bar{P}_{et} \pm sP_{et}$).

Table 3: Second example of the counts of wins and losses method.

SVM	MSE	CE	EXP	HS
36.89 (2.98)	25.07 (4.73)	25.17 (4.85)	21.63 (4.81)	21.25 (4.65)

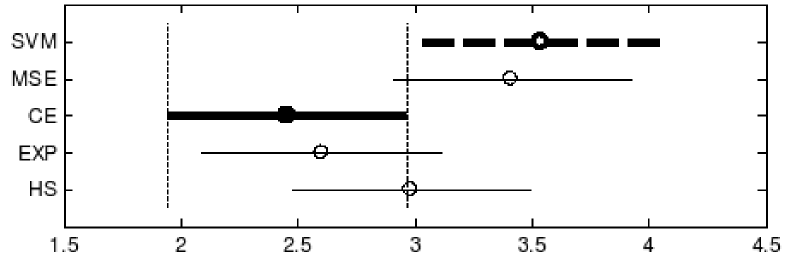

Figure 2: Dunn-Sidak comparison intervals (column means ±2× standard deviations) of the $\bar{P}_b$ scores. Only the SVM and CE intervals are clearly separated, with the SVM score significantly worse than the CE score.

Table 4: CV estimates (% values) of $\bar{P}_e$ (sP_e) with significant wins (bold) and losses (italic).

	Blood	Breast	Breast4	Cloud5f	Cork stop.
SVM	*22.58* (0.18)	*36.89* (2.98)	7.79 (1.79)	39.91 (2.56)	9.95 (0.36)
MSE	**19.83** (3.06)	25.07 (4.73)	8.23 (3.88)	34.63 (10.76)	9.74 (5.49)
CE	**19.88** (3.22)	25.17 (4.85)	7.29 (3.34)	32.18 (11.02)	9.55 (5.44)
EXP	**19.92** (3.21)	**21.63** (4.81)	7.34 (3.26)	33.20 (9.98)	10.01 (5.71)
HS	**19.95** (3.36)	**21.25** (4.65)	7.71 (3.52)	37.80 (12.64)	10.36 (5.93)
	E-coli	E-coli4	Glass	H. surv.	Inflamm.
SVM	11.00 (1.27)	4.27 (0.61)	25.54 (1.47)	24.53 (0.46)	*0.73* (1.55)
MSE	11.76 (2.28)	4.82 (1.49)	*27.72* (3.95)	24.74 (5.32)	**0.00** (0.00)
CE	11.45 (1.77)	4.86 (1.42)	**18.88** (3.83)	24.52 (4.86)	**0.00** (0.00)
EXP	12.31 (3.07)	5.07 (1.23)	25.53 (3.13)	24.31 (4.92)	**0.00** (0.00)
HS	11.58 (2.79)	4.69 (1.32)	21.46 (3.50)	27.63 (6.28)	**0.00** (0.00)
	Ionosphere	Iris	Jap vowels	Libras	LRS100f
SVM	**3.66** (0.33)	2.99 (0.53)	4.67 (2.46)	10.57 (1.76)	*7.60* (0.66)
MSE	*6.52* (3.75)	2.30 (3.31)	6.14 (3.04)	*11.63* (2.62)	6.62 (1.06)
CE	*6.12* (3.58)	2.05 (3.13)	5.50 (2.35)	**9.42** (2.20)	**5.10** (0.97)
EXP	*6.19* (3.92)	2.07 (3.22)	4.63 (2.24)	**9.80** (2.61)	**5.41** (1.28)
HS	*6.63* (4.02)	2.06 (3.06)	4.86 (2.74)	**9.62** (2.63)	**5.69** (1.40)
	Lung Cancer	Olive	Ozone	Parkinson	Pb12
SVM	*40.24* (7.61)	**2.92** (0.45)	4.94 (0.13)	**4.54** (1.36)	6.91 (0.22)
MSE	**24.92** (6.69)	*3.57* (0.80)	5.77 (1.29)	6.69 (3.27)	6.58 (2.14)
CE	**26.37** (7.95)	*3.77* (0.87)	5.83 (1.17)	6.04 (3.17)	6.60 (2.17)
EXP	**24.79** (7.96)	**3.03** (0.88)	5.82 (1.40)	5.63 (1.91)	6.32 (2.35)
HS	**27.36** (9.64)	*3.65* (0.84)	5.63 (1.17)	*7.61* (3.20)	7.07 (4.28)
	P. Diabetes	Robot-1	Robot-4	Robot-5	SC-chart
SVM	21.96 (0.19)	**13.61** (2.19)	18.76 (6.61)	*32.67* (2.85)	0.24 (0.10)
MSE	22.38 (3.57)	*19.23* (4.40)	16.69 (6.05)	28.69 (4.90)	0.48 (0.83)
CE	22.80 (3.61)	*18.21* (4.52)	16.36 (4.49)	**24.48** (3.26)	0.54 (0.84)
EXP	22.42 (3.40)	*18.33* (6.05)	15.31 (3.72)	**25.26** (3.61)	0.51 (1.12)
HS	22.68 (3.18)	*21.00* (5.74)	16.93 (4.93)	**26.30** (3.84)	0.51 (1.03)
	Sonar	Spectf-Heart	Telugu	Thyroid	Vehicle
SVM	**4.86** (0.78)	**11.54** (0.70)	*14.62* (0.29)	*3.35* (0.82)	14.70 (0.49)
MSE	*9.29* (5.62)	*16.64* (5.27)	11.05 (2.45)	**2.07** (1.08)	12.77 (2.87)
CE	*9.27* (5.65)	*16.01* (5.00)	**10.23** (2.39)	**2.13** (1.31)	13.54 (2.88)
EXP	*9.32* (5.70)	*16.36* (5.10)	10.73 (2.58)	**2.08** (1.15)	12.46 (2.90)
HS	*9.31* (6.20)	*14.80* (5.14)	11.79 (2.38)	2.62 (2.51)	12.98 (3.02)
	Wdbc	Wdbc-org	Wine	Wpbc	Yeast
SVM	2.02 (0.13)	2.81 (0.10)	*2.11* (0.83)	**13.16** (0.96)	39.19 (0.63)
MSE	1.74 (1.34)	2.83 (1.30)	**1.34** (1.24)	*23.67* (2.45)	38.76 (2.15)
CE	1.76 (1.33)	2.76 (1.39)	**1.07** (1.05)	*23.71* (3.33)	39.79 (2.44)
EXP	1.81 (1.36)	2.79 (1.49)	**1.32** (1.13)	*23.71* (3.13)	**37.77** (1.51)
HS	1.82 (1.26)	2.63 (1.30)	**1.27** (1.05)	*21.83* (5.19)	*47.78* (5.16)

Table 5: CV estimates (% values) of $\bar{P}_b$ (sP_b) with significant wins (bold) and losses (italic).

	Blood	Breast	Breast4	Cloud5f	Cork stop.
SVM	*46.73* (0.22)	*39.48* (3.21)	10.40 (3.03)	*41.54* (2.72)	9.95 (0.36)
MSE	**34.82** (1.00)	26.30 (5.13)	9.97 (5.67)	**34.48** (11.14)	9.74 (5.49)
CE	**34.73** (4.06)	26.53 (5.19)	8.48 (4.24)	**31.77** (11.56)	9.55 (5.44)
EXP	**35.04** (4.18)	**22.57** (5.04)	8.43 (3.95)	**32.71** (10.47)	10.01 (5.71)
HS	**36.14** (4.07)	**22.25** (4.81)	9.11 (4.54)	37.61 (11.51)	10.36 (5.93)
	E-coli	E-coli4	Glass	H. surv.	Inflamm.
SVM	16.01 (2.32)	7.03 (2.06)	*33.69* (2.99)	*43.15* (0.73)	*0.75* (1.59)
MSE	16.74 (4.54)	7.45 (2.71)	23.89 (4.66)	40.77 (5.89)	**0.00** (0.00)
CE	15.57 (2.89)	7.35 (2.74)	**20.14** (5.22)	38.06 (6.08)	**0.00** (0.00)
EXP	16.46 (4.25)	7.46 (2.52)	30.22 (4.66)	38.68 (5.73)	**0.00** (0.00)
HS	15.30 (4.34)	7.03 (2.52)	24.11 (4.32)	**33.77** (7.04)	**0.00** (0.00)
	Ionosphere	Iris	Jap vowels	Libras	LRS100f
SVM	**4.58** (0.43)	2.99 (0.53)	4.57 (2.38)	10.57 (1.76)	*20.07* (1.72)
MSE	*8.47* (4.69)	2.30 (3.31)	5.92 (2.92)	*11.63* (2.62)	*21.37* (2.99)
CE	*7.87* (4.57)	2.05 (3.13)	5.26 (2.24)	**9.42** (2.20)	**13.42** (2.65)
EXP	*7.59* (4.52)	2.07 (3.22)	4.47 (2.17)	**9.80** (2.61)	**13.27** (2.88)
HS	*8.70* (5.17)	2.06 (3.06)	4.70 (2.74)	**9.62** (2.63)	17.65 (3.42)
	Lung Cancer	Olive	Ozone	Parkinson	Pb12
SVM	*43.01* (6.81)	4.27 (0.71)	**31.92** (0.76)	**8.11** (2.72)	6.95 (0.23)
MSE	**23.71** (7.32)	4.92 (1.19)	34.11 (5.63)	9.66 (5.69)	6.60 (2.22)
CE	**25.30** (8.79)	*5.07* (1.36)	36.83 (4.33)	**7.75** (4.62)	6.59 (2.22)
EXP	**23.65** (7.88)	**4.08** (1.06)	**33.32** (6.76)	**6.87** (2.73)	6.30 (2.34)
HS	**26.39** (9.44)	*5.06* (1.37)	**32.47** (4.54)	*11.32* (5.69)	6.97 (4.02)
	P. Diabetes	Robot-1	Robot-4	Robot-5	SC-chart
SVM	28.81 (0.23)	17.96 (2.92)	*23.82* (10.33)	*34.58* (3.34)	0.24 (0.10)
MSE	26.67 (3.88)	20.20 (4.74)	**17.42** (5.77)	31.03 (4.79)	0.48 (0.83)
CE	26.05 (4.24)	19.91 (4.83)	**17.17** (6.28)	**26.29** (3.55)	0.54 (0.84)
EXP	26.61 (3.97)	19.94 (5.50)	**16.55** (4.93)	**27.15** (4.23)	0.51 (1.12)
HS	26.39 (3.58)	22.02 (5.23)	**16.66** (4.89)	28.69 (4.07)	0.51 (1.03)
	Sonar	Spectf-Heart	Telugu	Thyroid	Vehicle
SVM	**5.19** (0.85)	**20.15** (1.42)	*18.14* (0.32)	*6.83* (1.67)	*14.60* (0.49)
MSE	*9.46* (5.72)	*34.35* (8.87)	13.10 (2.90)	**2.59** (1.83)	**12.57** (2.58)
CE	*9.41* (5.70)	29.00 (8.11)	**11.65** (3.02)	**2.87** (2.46)	13.41 (2.85)
EXP	*9.46* (5.76)	29.65 (7.87)	12.55 (3.03)	**3.46** (2.53)	**12.31** (2.78)
HS	*9.45* (6.29)	**23.08** (8.12)	14.18 (2.90)	**4.14** (5.99)	**12.85** (2.78)
	Wdbc	Wdbc-org	Wine	Wpbc	Yeast
SVM	2.48 (0.13)	2.75 (0.12)	*1.90* (0.84)	**25.63** (1.14)	**51.48** (0.86)
MSE	2.07 (1.59)	2.88 (1.47)	**1.20** (1.10)	*49.93* (0.62)	**51.30** (2.04)
CE	2.04 (1.54)	2.90 (1.58)	**0.98** (0.98)	*50.00* (0.00)	**51.52** (1.92)
EXP	2.10 (1.60)	2.98 (1.75)	**1.20** (0.99)	*50.00* (0.00)	**50.28** (1.93)
HS	2.16 (1.58)	2.58 (1.33)	**1.16** (1.01)	29.76 (5.19)	*59.68* (4.04)

Table 6: Comparison of $\bar{P}_e$ (% values) between SVM (control) vs. any of the MLPs.

Dataset	SVM	MSE		CE		EXP		HS	
Blood	22.49	19.83	(−)	19.88	(−)	19.92	(−)	19.95	(−)
Breast	36.96	25.07	(−)	25.17	(−)	21.63	(−)	21.25	(−)
Breast4	7.62	8.23	(+)	7.29	(−)	7.34	(−)	7.71	(−)
Cloud5f	39.81	34.63	(−)	32.18	(−)	33.2	(−)	37.8	(−)
Cork stop.	9.87	9.74	(−)	9.55	(−)	10.01	(+)	10.36	(+)
E-coli	10.88	11.76	(+)	11.45	(+)	12.31	(+)	11.58	(+)
E-coli4	4.19	4.82	(+)	4.86	(+)	5.07	(+)	4.69	(+)
Glass	26.08	27.72	(+)	18.88	(−)	25.53	(−)	21.46	(−)
Hsurv	24.56	24.74	(+)	24.52	(−)	24.31	(−)	27.63	(+)
Inflamm.	0.15	0	(−)	0	(−)	0	(−)	0	(−)
Ionosph.	3.65	6.52	(+)	6.12	(+)	6.19	(+)	6.63	(+)
Iris	3.07	2.3	(−)	2.05	(−)	2.07	(−)	2.06	(−)
Jap vowels	4.43	6.14	(+)	5.5	(+)	4.63	(−)	4.86	(+)
Libras	10.02	11.63	(+)	9.42	(−)	9.8	(−)	9.62	(−)
Lrs100f	7.46	6.62	(−)	5.1	(−)	5.41	(−)	5.69	(−)
Lung cancer	37.58	24.92	(−)	26.37	(−)	24.79	(−)	27.36	(−)
Olive	2.95	3.57	(+)	3.77	(+)	3.03	(+)	3.65	(+)
Ozone	4.89	5.77	(+)	5.83	(+)	5.82	(+)	5.63	(+)
Parkinson	4.89	6.69	(+)	6.04	(+)	5.63	(+)	7.61	(+)
Pb12	6.97	6.58	(−)	6.6	(−)	6.32	(−)	7.07	(+)
P. Diabetes	22.09	22.38	(+)	22.8	(+)	22.42	(+)	22.68	(+)
Robot 1	12.56	19.23	(+)	18.21	(+)	18.33	(+)	21	(+)
Robot 4	17.72	16.69	(−)	16.36	(−)	15.31	(−)	16.93	(−)
Robot 5	32.29	28.69	(−)	24.48	(−)	25.26	(−)	26.3	(−)
SC-Chart	0.24	0.48	(+)	0.54	(+)	0.51	(+)	0.51	(+)
Sonar	5.11	9.29	(+)	9.27	(+)	9.32	(+)	9.31	(+)
Spectf-Heart	11.8	16.64	(+)	16.01	(+)	16.36	(+)	14.8	(+)
Telugu	14.47	11.05	(−)	10.23	(−)	10.73	(−)	11.79	(−)
Thyroid	3.54	2.07	(−)	2.13	(−)	2.08	(−)	2.62	(−)
Vehicle	13.4	12.77	(−)	13.54	(−)	12.48	(−)	12.98	(−)
Wdbc	1.97	1.74	(−)	1.76	(−)	1.81	(−)	1.82	(−)
Wdbc-org	2.82	2.83	(+)	2.76	(−)	2.79	(−)	2.63	(−)
Wine	2.23	1.34	(−)	1.07	(−)	1.32	(−)	1.27	(−)
Wpbc	13.09	23.67	(+)	23.71	(+)	23.71	(+)	21.83	(+)
Yeast	39.21	38.76	(−)	39.79	(+)	37.77	(−)	47.78	(+)
# minuses			17		21		22		18