EXPERIMENTS ON AUTOMATIC DRUG ACTIVITY CHARACTERIZATION USING SUPPORT VECTOR CLASSIFICATION

Francesc J. Ferri and Wladimiro Díaz
Departament d'Informàtica
Universitat de València
Dr. Moliner, 50 Burjassot (València) Spain
email: francesc.ferri@uv.es

Maria J. Castro
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camì de Vera s/n, València, Spain
email: mcastro@dsic.upv.es

ABSTRACT
The characterization of pharmacological properties from their chemical structure has become a challenging and promising technique in computer aided drug design. The idea consists of finding appropriate representations of candidate compounds in terms of their chemical formulae and try to apply a particular machine learning method able to appropriately characterize certain desired properties or kinds of pharmacological activity. In this particular work antibacterial activity has been considered. Several classic pattern classification methods have already been applied to this problem with promising results. In this work, the support vector machine model is considered and compared to multilayer perceptrons in this particular context. The natural and unpredictable imbalance and the fact that only relatively small samples can be used for learning make this a challenging and interesting problem.

KEY WORDS
Support Vector Machines, Pattern Classification, Multilayer Perceptron, Pharmacological drug selection.

1 Introduction

The design of new medical drugs with desired chemical properties is a challenging and very important problem in the pharmaceutical industry. The traditional approach for formulating new compounds requires the designer to test a very large number of molecular compounds, to select them in a blind way, and to look for the desired pharmacological property. Therefore, it is very useful to have tools to discriminate the pharmacological activity of a given molecular compound so that the laboratory experiments can be directed to those molecular groups in which there is a high probability of finding new compounds with the desired properties.

All methods developed for this purpose are based on the fact that the activity of a molecule derives from its structure and therefore it is possible to find a relationship between this structure and the properties that the molecule exhibits [14]. Thus, the way the molecular structure is represented has special relevance.

[-1]This work has been partially funded by Spanish MEC projects TIC2003-08496 and DPI2006-15542.

In Chemical Graph Theory, molecular structures are represented as doubly labeled graphs which can be conveniently characterized by a number of specific topological indices [8]. In this work, a reduced set of 62 topological indices [9] are considered.

These or similar representations have already been applied to different discrimination problems in drug design (analgesic, antidiabetic, antibacterial, etc.). In the particular case of antibacterial activity, very good classification results have been reported using multilayer perceptrons (MLP) [4, 11].

Also important in the above mentioned application is the cost/benefit problem and the corresponding discrimination thresholds that have to be used to maximize the outcomes of the learned classifiers from the point of view of the pharmacological problem. The use of Receiver Operating Characteristic (ROC) curves [5] has been shown to be a valuable tool in this particular context to evaluate the classifier in a wide range of practical situations.

This paper introduces Support Vector Machines (SVM) in this particular problem and studies the differences and particularities of the corresponding solutions as compared with the state-of-the-art solution based on multilayer perceptrons [11]. A detailed analysis has been performed in order to assess the suitability and adaptability of these methods for the particular task using ROC analysis.

2 The molecular structure-activity relationship discovery

The so-called quantitative structure-activity relationship (QSAR) models are currently used in the computer aided design of new medical drugs with desired chemical properties. As an alternative to the methods based on the "exact" description of the electronic properties of a molecule calculated by mechanical-quantum methods, the molecular topology describes the molecule as a set of indices. These topological indices are numerical descriptors that encode information about the number of atoms and their structural environment. This representation is derived from the hydrogen-suppressed molecular formula seen as a graph and it requires a relatively low calculation effort [1, 2, 8].

The molecular topology considers a molecule as a

planar graph where atoms are represented by vertices and chemical bonds are represented by edges. The chosen set of molecular descriptors should adequately capture the phenomena underlying the properties of the compound. In this and other related works, a set of 62 indices has been selected [9, 15, 12]. Fourteen of these indices are related to the molecular attributes of the compound; for example, the total number of atoms of a certain element (carbon, nitrogen, oxygen, sulphur, fluorine, chlorine, ...), the total number of bonds of a certain type (simple, double or triple), the number of atoms with a specific vertex degree, distance between the bonds, etc...

The remaining forty-eight indices include different topological information, such as the number of double bonds at distance 1 or 2, and the minimum distance between pairs of atoms, which are counted as the number of bonds between atoms. These indices are classified into six groups which are associated to the most frequent elements that constitute the molecules with pharmacological activity: nitrogen, oxygen, sulphur, fluorine, chlorine, bromine, and a general group in which the distances between pairs of atoms are considered without identifying the type of atom.

This molecular representation has shown its ability for discriminating and predicting different kinds of pharmacological properties. Nevertheless, it is known that certain indices are more important that others for detecting particular cases. For example, it has been shown that only eight out of the above topological indices are enough to predict antibacterial activity with about 80% accuracy (and about 90% inactivity accuracy or 10 % false alarm rate) using linear discrimination models [12].

Obviously, the QSAR studies rely on the key fact that the activity of a molecule directly derives from its structure or, more precisely, from certain aspects of it. The better the chosen set of indices captures these particular aspects, the better the (blind) machine learning methods will characterize the activity of the molecule. As the molecular descriptors or indices have to be general in order to be applied in a wide range of drug design contexts, the ability of the particular learning methods used to capture non linear relations and high order dependencies among them becomes a key fact in the whole process.

3 Minimizing the risk in automated drug design

It is important to note that we are interested not only in achieving a high accuracy in classification but also a convenient compromise between true positive and false alarm rates. The high economical costs due to the pharmacological tests on each candidate molecule in drugs research makes an important issue to keep the number of false positives as low as possible, even if this implies to reject some true positives.

Given a particular classifier whose output consists of a continuous value in a specified interval (as in the cases considered in this work), the Receiver Operating Characteristic (ROC) curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form (FP,TP). The performance of any classifier (with the corresponding threshold included) can be represented by a point in the ROC space. ROC curves move from the "all-inactive" point (0,0) which corresponds to the highest value of the threshold to the "all-active" point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A number techniques to obtain different measures from ROC curves have been also developed [7]

4 Machine learning techniques for antibacterial activity discrimination

The particular discrimination problem was to determine whether a molecule has antibacterial activity or not. To this end, four different classification approaches have been considered or are referred to in this work: Linear Discrimination Analysis [12], Multilayer Perceptrons [4], Support Vector Machines [3] and Gaussian Naive Bayes [6] as a reference.

4.1 Linear discriminant analysis

Linear discriminant analysis (LDA) has been widely used for this and similar problems in the specific literature [12]. The method consists of finding the optimal separation hyperplane. It is well known that the LDA solution can be very far from the optimal one in the case of highly nonlinear relations among data. This kind of limitation is shared by all linear classification methods.

The results presented in this work using LDA have been directly taken from a previous work in the specialized literature [12] which used a different (and slightly more representative) dataset and can be considered as the best results to date for this particular problem. Both the data and the results are explicitly shown in this reference.

4.2 Naive Bayes Classifier

The Naive Bayes (NB) classifier consists of applying the optimal Bayes classifier under the assumption that all features are statistically independent. In this way, the tough task of estimating the posterior probabilities for each class becomes feasible as it consists of a unidimensional problem.

The Naive Bayes implementation used was taken from the data mining and machine learning package Weka [16]. This classifier is not well suited for this problem in which very high dependencies are present among the corresponding features. The only reason to consider this here is as a baseline reference.

4.3 Multilayer Perceptron

Multilayer Perceptrons (MLP) have been extensively used in a wide range of applications that require a nonlinear approach. This classifier consists of a variable number of layers composed by neurons each of which is in fact a linear classifier. Neurons are nonlinearly connected by using a nonlinear sigmoid-like activation function. MLPs can be adaptively trained by suboptimal gradient descent methods.

The training of the MLPs was carried out here by using the neural software package "SNNS: Stuttgart Neural Network Simulator" [17]. The network topology (layers and connections among them), training algorithm and parameter settings were chosen from a previous work [4] which was not particularly aimed at looking for antibacterial activity.

More specifically, the results presented have been obtained by using the standard Backpropagation algorithm with a learning rate equal to 0.05. The first (input) layer consists of 62 input units. Only one hidden layer with a variable number of neurons has been considered. In the third (output) layer only one neuron is needed that gives a value in the range $[-1, 1]$ that indicates whether the feature vector at the input is inactive or active, respectively.

The hyperbolic tangent function was used as activation function for all neurons in order to keep outputs in the interval $[-1, 1]$ as in the original LDA experiments [12]. The number of neurons in the hidden layer has been set to 0 (linear), 2, 6 and 8. In the forthcoming results only some of these are shown due to space limitations.

4.4 Support vector machines

Support vector machines (SVM) are well-founded and widely employed classification and regression methods. From a technical point of view, SVM are linear classifiers that operate in an appropriately transformed space that allows a wide range of nonlinear discrimination possibilities as in the case of MLP. The mentioned transformation is made by using the so-called kernel trick [13].

The training of SVM is carried out by the structural risk minimization principle that in this particular case leads to a margin maximization problem that is solved by quadratic programming. The margin can be visualized as the "space" between samples of a particular class and the separation hypersurface. It can be shown that, trained in this way, the SVM constitutes the optimal classifier from the point of view of generalization abilities with the training information available (and once a transformation/kernel has been set).

In order to cope with noise and natural overlapping between classes, the concept of soft-margin can be introduced. In this way, a parameter, C, controls what proportion of samples of a particular class can be wrongly placed with regard to the separation hypersurface and its margin. In other words, the soft margin parameter controls the overtraining/generalization trade-off when learning a particular classifier from data.

From all possible families of kernels, the radial basis function (RBF) or proximity kernel parameterized by an influence parameter, γ, has been considered for the experiments. The experiments have been carried out by using the SVM$^{\text{light}}$ [10] software package. The particular kernel function is

$$K(a, b) = e^{\gamma \cdot ||a-b||^2}$$

5 Data preparation, normalization and experimental design

For the experiments presented in this work, a dataset of 434 samples with potential pharmacological activity has been considered. Out of these, 218 molecules are known to exhibit antibacterial activity while the other 216 compounds do not have this activity at all. These balanced proportions are by no means representative of the a priori probability of antibacterial activity in real pharmacological design trials.

In order to obtain results as significant as possible, 10-fold stratified cross-validation has been used to compute all accuracies. When shown, 95% confidence intervals are computed considering a standard distribution.

All 62 topological indices were used to obtain feature vectors in which values were linearly normalized to the interval $[-1, 1]$ in an independent way. As in previous works [12, 5], each feature vector was labeled with 1 (indicating that the molecule has antibacterial properties) and -1 (the molecule is inactive).

5.1 Validation and parameter search

Extensive experimentation has been carried out both with multilayer perceptrons and with support vector machines. In the first case, the use of a validation set taken from each training set in the cross validation procedure has been used to select some of the parameters of the MLP and in particular, to select the final network and prevent overtraining.

In the case of support vector machines, only two parameters have to be selected: the soft margin (C) and the

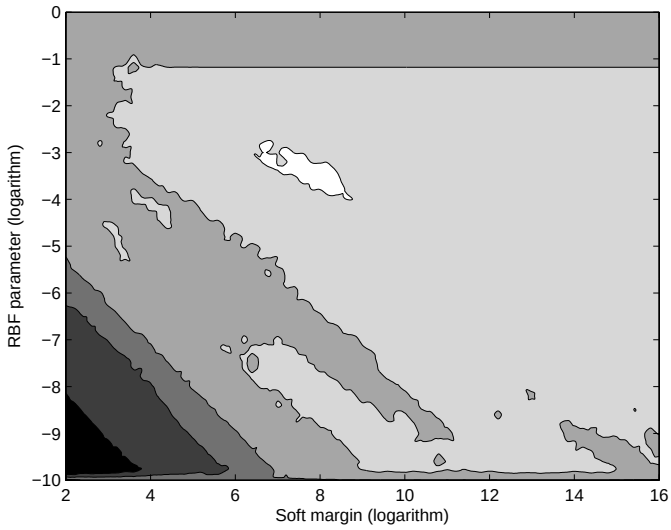

Figure 1. Performance results of the grid-search carried out in the parameter space. The contour lines shown correspond to the accuracy values 80, 84, 87, 90, 93, and 95.4. The white region corresponds to higher values.

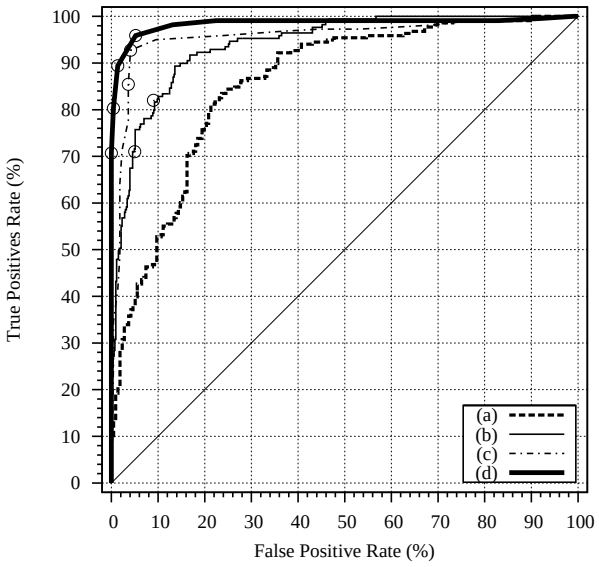

Figure 2. Averaged ROC curves for a) Gaussian Naive Bayes, b) Linear Discriminant Analysis, c) Multilayer Perceptron with 8 hidden neurons, and d) Support Vector Machine with $C = 194$ and $\gamma = 0.095$. Circles indicate some of the particular classifiers obtained that are mentioned in the text.

influence parameter (γ). This two parameters are quite well-behaved and allow for a systematic search in the corresponding 2D parameter space. For the experiments in this work, a kind of grid-search has been implemented by using part of the training data. In particular, several thousands of SVMs have been trained by uniformly sampling the (logarithmic) parameter space considering the ranges $[0, 60000]$ for C and $[0.001, 1]$ for γ. For each SVM, the overall accuracy by taking zero threshold is computed. The corresponding surface for one of the experiments is shown in Figure 1 conveniently interpolated and smoothed.

6 Experiments, results and discussion

The above mentioned classification methods have been applied to the training sets taken from the available data set and the corresponding (continuous) outputs have been obtained for the test data. For each partition into train and test, a ROC curve is obtained.

Figure 2 shows the corresponding averaged ROC curves for the four classification schemes considered. In the particular case of LDA, the curve corresponds to a unique partition into train and test data as explained in [12].

The ROC curve corresponding to LDA shows that, apart from the classifier $(8, 82)$ mentioned in [12] with discrimination threshold set to zero, there are other possible convenient classifiers as the one with point $(5, 71)$ in ROC space which uses a higher discrimination threshold, namely 0.61.

The results obtained with the Gaussian Naive Bayes are clearly and significantly worse than the others showing that this classification scheme is not well suited for this problem.

On the other hand, the Multilayer Perceptron with 8 units in the hidden layer, significantly outperforms the LDA results along the whole range of the curve. A particular (averaged) result that can be mentioned is $(10, 95)$ that involves discrimination thresholds very close to 1. For the particular application in drug design, the most convenient MLP classifiers are $(4, 85)$ and $(4, 93)$ with discrimination thresholds around zero. It is worth mentioning that MLPs with different number of hidden units gave in our experiments results that were not significantly different than the ones with 8 hidden units.

Finally, the SVM with the appropriately selected parameters gave much more stable results which were significantly better than the previous ones. It is remarkable that the variability of the results among the 10 partitions considered was much smaller than for the other classifiers. As can be seen, the ROC curve obtained was significantly smoother. Two of the overall best SVM classifiers that can be mentioned are $(5, 95)$ and $(1, 89)$ with thresholds -0.2 and 0.3, respectively. But as before, if one considers the particular application for real drug design, the classifier $(0.5, 80)$ or even $(0, 71)$ can be considered as the ones with more applicability. The two discrimination thresholds for these classifiers are 0.8 and 1.0, respectively.

The threshold averaged ROC curves have been computed as explained in [7]. The corresponding intervals for a 95% confidence level have also been computed and are

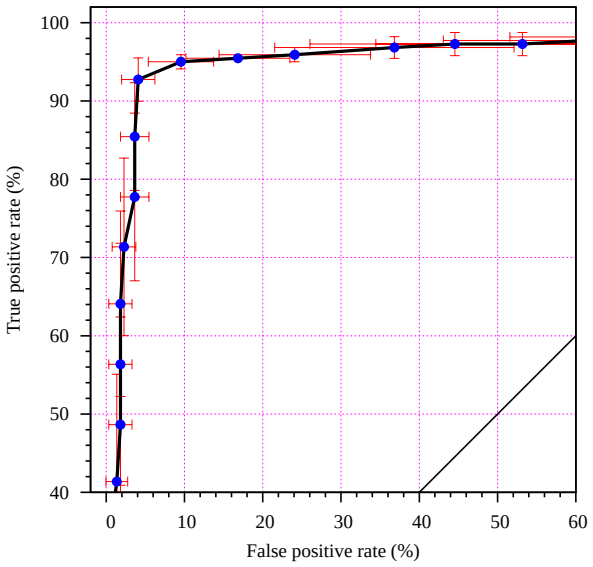

Figure 3. Averaged ROC curve corresponding to the multilayer perceptron approach along with 95% confidence intervals.

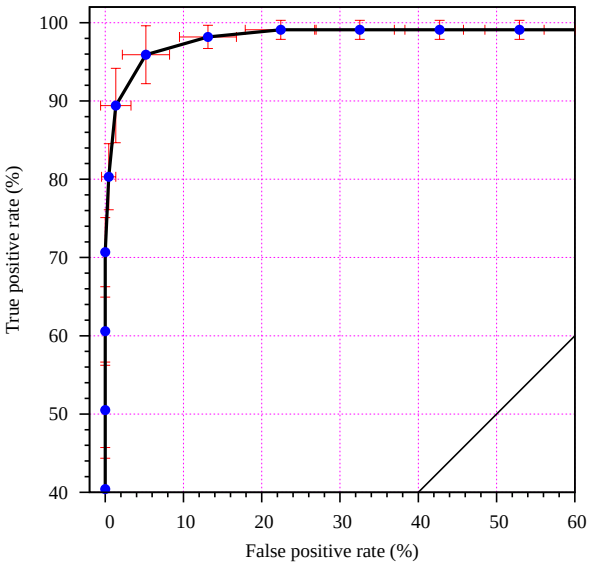

Figure 4. Averaged ROC curve corresponding to the SVM approach along with 95% confidence intervals.

shown in Figure 3 for the case of Multilayer Perceptron and in Figure 4 for the case of Support Vector Machine. In both cases, only the most interesting part of the whole ROC space is shown. It can be observed the differences in the variability of both curves.

As an overall measure of accuracy, the Area Under the Curve (AUC) has been computed for all methods considered. The AUC is a common method used in ROC analysis to give global measure of classifier goodness. When normalized, this is an scalar value in the range $[0, 1]$. The corresponding AUC values for the four classifiers considered here are shown in Table 1.

Table 1. Area under the curve (AUC) for all classifiers considered.

	Classifiers			
	NB	LDA	MLP	SVM
AUC	0.857	0.941	0.958	0.986

7 Concluding remarks and considerations for further work

In this work a classical ROC analysis has been performed on a particular drug activity discrimination problem. Preliminary results show that this kind of analysis is very interesting and can significantly improve the overall costs in the whole drug design methodology. Multilayer Perceptron has been shown to significantly improve previously used approaches in a wide range of situations. Moreover, Support Vector Machines as used in this work improve upon MLP's results. This improvement is not only in absolute numbers (which is not very dramatic) but also in the robustness and stability and generalization ability of the classifiers obtained.

In order to obtain more confident results significant also from a pharmacological point of view, the whole experimentation in this work needs to be repeated with a larger and more representative data set. Also, ROC analysis including a reject option as in [5] is under consideration. In this case, by considering true positive rate, false alarm rate and reject rate it would be possible to achieve a full characterization of the discrimination problem in drug design applications.

Acknowledgments

We are grateful to Dr. Mr. Facundo Pérez and Dra. Ms. María Teresa Salabert, from the Chemistry and Physics Department of the Pharmacy Faculty of the Universitat de València, for their help in supplying the datasets and specially for their supervision in getting the topological indices and the elaboration of the samples used in the experimentation.

References

[1] A.T. Balaban, I. Motoc, D. Bonchev, and O. Makenyan. Topological indices for structure-activity correlations. *Top. Curr. Chem.*, 114:21–55, 1983.

[2] S.C. Basak, S. Bertelsen, and G. Grunwald. Application of graph theoretical parameters in quantifying

molecular similarity and structure-activty studies. *J. Chem. Inf. Comput. Sci.*, 34:270–276, 1994.

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.

[4] M. J. Castro, W. Díaz, P. Aibar, and J. L. Domínguez. Prediction and Discrimination of Pharmacological Activity by Using Artificial Neural Networks. In F. J. Perales, A. J. C. Campilho, N. Pez de-la Blanca, and A. Sanfeliu, editors, *Pattern Recognition and Image Analysis*, volume 2652 of *LNCS*, pages 184–192. Springer-Verlag, 2003. ISSN 0302-9743.

[5] W. Díaz, M.J. Castro, F.J. Ferri, F. Pérez, and M. Murcia. Improving pattern recognition based pharmacological drug selection through ROC analysis. In A. Sanfeliu and J. Ruíz-Shulcloper, editors, *Progress in Pattern Recognition, Image Analysis and Applications. Lecture Notes in Computer Science. Vol. 3287*, pages 621–628. Springer-Verlag, 2004.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, second edition, 2001.

[7] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, submitted(`http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ROC101.pdf`), 2004.

[8] J. Gálvez, R. García-Domenech, J.V. de Julián-Ortiz, and R. Soler. Topological approach to drug design. *J. Chem. Inf. Comput. Sci.*, 35:272–284, 1995.

[9] J. Jaén-Oltra, M.T. Salabert-Salvador, F.J. García-March, F. Pérez-Giménez, and F. Tomás-Vert. Artificial neural network applied to prediction of fluorquinolone antibacterial activity by topological methods. *J. Med. Chem.*, 43:1143–1148, 2000.

[10] T. Joachims. Making large-scale svm learning practical. In B. Schkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

[11] M. Murcia-Soler, F. Pérez-Giménez, F.J. García-March, M.T. Salabert-Salvador, W. Díaz-Villanueva, M.J. Castro-Bleda, and A. Villanueva-Pareja. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. and Comp. Sciences*, 3:1031–1041, 2004.

[12] M. Murcia-Soler, F. Pérez-Giménez, F.J. García-March, M.T. Salabert-Salvador, W. Díaz-Villanueva, and P. Medina-Casamayor. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. *J. Mol. Graph. Model.*, 21:375–390, 2003.

[13] B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[14] P.G. Seybold, M. May, and U.A. Bagal. Molecular structure-propertiy relationships. *J. Chem. Educ.*, 64:575–581, 1987.

[15] F. Tomás-Vert, F. Pérez-Giménez, M.T. Salabert-Salvador, F.J. García-March, and J. Jaén-Oltra. Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *J. Mol. Struct. (THEOCHEM)*, 504:272–276, 2000.

[16] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[17] A. Zell et al. *SNNS: Stuttgart Neural Network Simulator. User Manual, Version 4.2*. Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Germany, 1998.