# A Unified Approach to the Extraction of Rules from Artificial Neural Networks and Support Vector Machines

João Guerreiro[1] and Duarte Trigueiros[2]

[1] Department of Information Science and Technology, ISCTE-IUL, Portugal
[2] Faculty of Economics, University of Algarve, Portugal

**Abstract.** Support Vector Machines (SVM) are believed to be as powerful as Artificial Neural Networks (ANN) in modeling complex problems while avoiding some of the drawbacks of the latter such as local minimæ or reliance on architecture. However, a question that remains to be answered is whether SVM users may expect improvements in the interpretability of their models, namely by using rule extraction methods already available to ANN users. This study successfully applies the Orthogonal Search-based Rule Extraction algorithm (OSRE) to Support Vector Machines. The study evidences the portability of rules extracted using OSRE, showing that, in the case of SVM, extracted rules are as accurate and consistent as those from equivalent ANN models. Importantly, the study also shows that the OSRE method benefits from SVM specific characteristics, being able to extract less rules from SVM than from equivalent ANN models.

**Keywords:** Data Mining, Support Vector Machines, Artificial Neural Networks, Orthogonal Search-based Algorithm, OSRE, Pedagogical, Decompositional, Rule Extraction.

## 1 Introduction

Support Vector Machines (SVM) proved to be accurate analytical tools, quite able to predict complex relations in various application fields. Similarly to Artificial Neural Networks (ANN), such accuracy stems from their ability to represent any given function [1] [2] [3] or to define complex decision boundaries. Despite their predictive ability, ANN and SVM have well-known drawbacks such as their black-box approach to modeling and the ensuing lack of transparency. What can be learnt from their systemic underlying knowledge representation is little more than a set of weights, activation functions and optimal parameters, discovered during the Neural Network training, or the kernel function and the optimized parameters of the Support Vector Machine. Hidden inside such complexity is an eventually meaningful relationship between inputs and predicted values.

Given the obvious need to understand the underlying learning mechanisms of ANN, in recent years authors have proposed varied techniques to overcome this missing transparency. However, there is still a demand for a unified rule extraction method that encompasses ANN and SVM, ensuring a compromise between accuracy and fidelity to the original models, together with consistency and comprehensibility, while being highly portable between different algorithms.

L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part II, LNCS 6441, pp. 34–42, 2010.

The paper demonstrates a quite successful application of the Orthogonal Search-based Rule Extraction algorithm (OSRE) [4] to Support Vector Machines modeling. Results obtained with Artificial Neural Networks and the same rule extraction methods are used as benchmarks, showing that the use of OSRE with SVM is capable of maintaining the accuracy of the original classifiers while extracting, in both cases, a consistent set of rules.

## 2 Rule Extraction

Figure 1 displays chronologically the most relevant algorithms hitherto proposed to extract rules from ANN or other modeling tools. In Figure 1 algorithms are further organized according to the translucency of the rule extraction algorithm [5].

The translucency criterion considers the techniques perception of the learning method, thus creating three broad families of rule extraction approaches: the decompositional approach, the pedagogical approach and the eclectic approach.

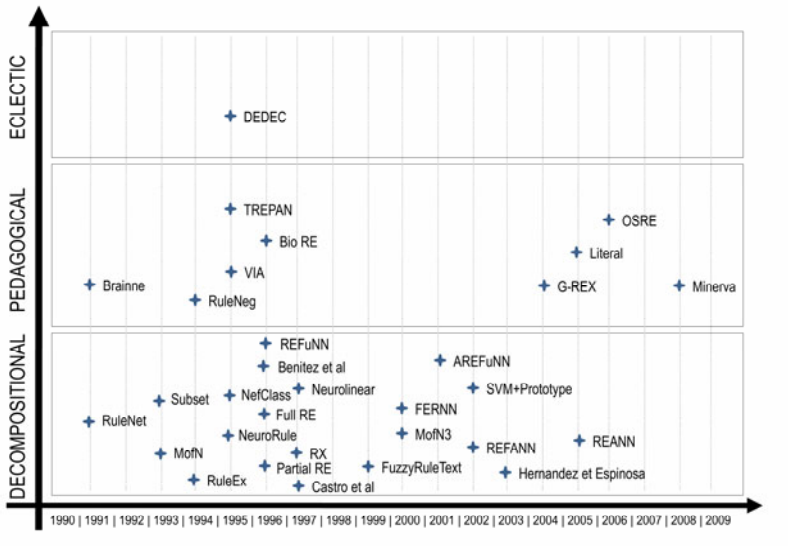


**Fig. 1.** Rule Extraction Methods organized by broad families and by year

The decompositional approach extracts rules at the level of the individual units by analyzing the activation values, weights and biases of the Neural Networks and the kernel function, vectors and optimization parameters of the Support Vector Machines [5]. Its disadvantage lies in its dependency on the learning mechanism coupled with the inability to accurately derive the logic of the underlying decision surface [4] [5].

The pedagogical approach used considers the trained ANN or SVM as a black box and using the classifier algorithm as an oracle through which it tests its predicted responses [5]. While changing the input values, rules are extracted which express the

relationship between inputs and outputs of the Neural Network or Support Vector Machine. The main issue with most pedagogical approaches is that they are exponential in their complexity [4]. The number of rules grows at a rate of $k^n$, with n being the number of input variables with k possible values. Nevertheless, they are highly portable due to their ability to operate with all types of classifiers.

Finally, the eclectic approach incorporates elements of both decompositional and pedagogical rule extraction techniques. The algorithms of the eclectic type use the internal architecture of the trained ANN or SVM to complement a symbolic learning algorithm [4] [5].

## 3 The Use of OSRE with ANN

The Orthogonal Search-based Rule Extraction algorithm (OSRE) from Etchells and Lisboa [4] is a successful pedagogical methodology often applied in biomedicine (see [6] [7] [8] [9] and others). OSRE possesses the attractive characteristic of reducing the problem from exponential to linear in terms of the number of inputs [4] and is based on a formalism proposed by Tsukimoto [11]. OSRE extends the algorithm proposed by Ruleneg [10] to ordinal and continuous variables using trained data to perform a 1-from-N coding, while searching, in orthogonal directions, where the decision surface crosses a decision boundary.

An illustrative example of the use of OSRE (with an ANN) follows: given 3 input variables, each with the following values:

$a_1 = [1,2]$, $a_2 = [1,2,3]$, $a_3 = [1,2,3,4]$.

In one case in the dataset where $[a_1,a_2,a_3] = [1,2,2]$, OSRE codes the original values into 1-from-N form, converting them into, respectively, [0,1|0,1,0|0,0,1,0].

Considering that the original case had a neural network activation response > 0.5, then while stepping through all values of $a_1$ leaving $a_2$ and $a_3$ fixed, OSRE uses the classifier as an oracle inspecting its response as shown in Table 1. This method is called stepwise negation [10] [4]. Since, in this case, there is no change in $a_1$, this input variable will not be included in the rule. Indeed, this input does not change the classifiers response. Stepping through $a_2$ and putting aside $a_1$ and $a_3$, the algorithm again inspects the classifiers response, and as can be seen, there has been a change at $a_2$ =2.

**Table 1.** Artificial Neural Network activation response

| Stepping through $a_1$ | Stepping through $a_2$ | Stepping through $a_3$ |
|---|---|---|
| Net (01\|010\|0010) > 0.5 | Net (01\|001\|0010) > 0.5 | Net (01\|010\|0001) < 0.5 |
| Net (10\|010\|0010) > 0.5 | Net (01\|010\|0010) > 0.5 | Net (01\|010\|0010) > 0.5 |
| | Net (01\|100\|0010) < 0.5 | Net (01\|010\|0100) > 0.5 |
| | | Net (01\|010\|1000) > 0.5 |

Therefore, a first rule is extracted: $a_2 \leq 2$. Finally, stepping through $a_3$, the algorithm finds a change at $a_3 = 2$. A second rule will be $a_3 \geq 2$. In brief, the network states that when $a_2 \leq 2$ and $a_3 \geq 2$ the response of the network is > 0.5 i.e. in class. So the rule for this data point is $(a_2 \leq 2) \wedge (a_3 \geq 2)$.

The resulting set of rules can then be refined by deleting repeated rules and those which fall below a predetermined specificity. The last refinement considers reducing the conjunctions and determining whether there has been a drop in specificity. Where specificity remains the same, conjunctions are eliminated thus defining the final set.

## 4 The Use of OSRE with SVM

OSRE is now applied to the extraction of rules from Support Vector Machines, and we will highlight here the successful results obtained. Following the steps used above to demonstrate the use of OSRE with Artificial Neural Networks, the first step consists of searching for a change in the classification result of the SVM, while stepwise negating the Boolean space of each input in the training dataset as we did in Table 1.

In order to discuss the ability of OSRE to extract rules from SVM, the original benchmark datasets from the Etchells and Lisboa study [4] are used here, namely the three Monks [12] datasets, the Wisconsin Breast Cancer [13] dataset and the Iris [14] dataset. Given that the relationships underlying these datasets have been widely studied and debated, namely using ANN, it is now possible to compare those published results with the SVM case highlighted in this paper, thereby evaluating the set of rules extracted by such method against those obtained from ANN. The training of SVM was carried out using SMO (Sequential Minimal Optimization) [15] and the kernel parameters were chosen to attain or surpass the accuracy of the original Artificial Neural Networks used in the OSRE literature. Results by dataset are as follows:

**Monks.** The Monks data consists of three artificial generated datasets each having 2 classes and 6 input variables:

$$a_1 = [1,2,3],\ a_2 = [1,2,3],\ a_3 = [1,2],\ a_4 = [1,2,3],\ a_5 = [1,2,3,4],\ a_6 = [1,2].$$

**Monks-1.** The known rule for these dataset is:

$$(a_5=1)\ (a_1=a_2)\ . \tag{1}$$

Using a SVM with a polynomial kernel we were able to get a trained model with 99.8% accuracy and the set of rules represented in Table 2.

The results show that the rules extracted from the SVM are consistent with the known rule and identical to the ones obtained with the ANN [4].

**Table 2.** OSRE SVM Extracted Rules - Monks-1

| Specificity | Sensitivity | Rules |
|---|---|---|
| 1 | 0.4928 | $[a_5=1]$ |
| 1 | 0.2338 | $[a_1=3, a_2=3]$ |
| 1 | 0.2266 | $[a_1=2, a_2=2]$ |
| 1 | 0.2050 | $[a_1=1, a_2=1]$ |

**Monks-2.** For the second dataset, the known rule is:
Exactly two of:

$$\{a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1, a_5 = 1, a_6 = 1\}. \quad (2)$$

Table 3 shows that it is possible to extract from a SVM, trained with a polynomial kernel, identical rules to the ANN OSRE results, which represent all the permutations of the known rule.

**Table 3.** OSRE SVM Extracted Rules - Monks-2

| Specificity | Sensitivity | Rules |
|---|---|---|
| 1 | 0.1601 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 1, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 1]$ |
| 1 | 0.1019 | $[a_1 = 1, 2 \le a_2 \le 3, a_3 = 1, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0970 | $[2 \le a_1 \le 3, a_2 = 1, a_3 = 1, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0825 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 1, a_4 = 1, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0825 | $[a_1 = 1, 2 \le a_2 \le 3, a_3 = 2, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 1]$ |
| 1 | 0.0776 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 2, a_4 = 1, 2 \le a_5 \le 4, a_6 = 1]$ |
| 1 | 0.0728 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 2, 2 \le a_4 \le 3, a_5 = 1, a_6 = 1]$ |
| 1 | 0.0728 | $[2 \le a_1 \le 3, a_2 = 1, a_3 = 2, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 1]$ |
| 1 | 0.0485 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 1, 2 \le a_4 \le 3, a_5 = 1, a_6 = 2]$ |
| 1 | 0.0485 | $[2 \le a_1 \le 3, a_2 = 1, a_3 = 2, a_4 = 1, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0388 | $[a_1 = 1, a_2 = 1, a_3 = 2, 2 \le a_4 \le 3, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0339 | $[a_1 = 1, 2 \le a_2 \le 3, a_3 = 2, a_4 = 1, 2 \le a_5 \le 4, a_6 = 2]$ |
| 1 | 0.0291 | $[a_1 = 1, 2 \le a_2 \le 3, a_3 = 2, 2 \le a_4 \le 3, a_5 = 1, a_6 = 2]$ |
| 1 | 0.0291 | $[2 \le a_1 \le 3, a_2 = 1, a_3 = 2, 2 \le a_4 \le 3, a_5 = 1, a_6 = 2]$ |
| 1 | 0.0242 | $[2 \le a_1 \le 3, 2 \le a_2 \le 3, a_3 = 2, a_4 = 1, a_5 = 1, a_6 = 2]$ |

Results thus confirm that the use of a different smooth classifier such as the SVM, does not change the rules that define the problem boundaries, as extracted by OSRE.

**Monks–3.** In this case, both ANN and SVM classified cases with a 98.91% accuracy and again, rules extracted using OSRE converge towards the known rule.
The known rule is:

$$(a_2 \neq 3)\ (a_5 \neq 4) \vee (a_4 = 1)\ (a_5 = 3)\ . \quad (3)$$

The rules from the SVM are shown in Table 4. They mimic the known rule.

**Table 4.** OSRE-SVM Extracted Rules – Monks-3

| Specificity | Sensitivity | Rules |
|---|---|---|
| 0.9812 | 0.9479 | $[1 \le a_2 \le 2, 1 \le a_5 \le 3]$ |
| 0.9962 | 0.1493 | $[a_4 = 1, a_5 = 3]$ |

**Wisconsin Breast Cancer.** This data set contains nine variables with discrete values between 1 and 10 and two outcome classes showing whether the patient's cancer is benign or malign. An ANN correctly classifies 96% of cases in the training phase and 96.9% in the validation phase. SVM obtained identical results. Rules extracted from ANN and SVM using OSRE were discarded where specificity fell below 90%. Rules extracted from ANN are represented in Table 5.

**Table 5.** OSRE – ANN Extracted Rules – Wisconsin Breast Cancer

| Specificity | Sensitivity | Rules |
|---|---|---|
| 0.9913 | 0.7261 | $[2 \le a_2 \le 10,\ 2 \le a_3 \le 10,\ 3 \le a_5 \le 6 \vee 8 \le a_5 \le 10,\ a_8 = 1 \vee 3 \le a_8 \le 6 \vee 8 \le a_8 \le 10]$ |
| 0.9137 | 1 | $[2 \le a_2 \le 10,\ 2 \le a_3 \le 10]$ |
| 0.9568 | 0.8690 | $[2 \le a_1 \le 10,\ 2 \le a_2 \le 10,\ 2 \le a_3 \le 10,\ 2 \le a_6 \le 10]$ |

Figure 2 shows the boundaries of the extracted rule with higher sensitivity:

$$(a_2 \ge 2)(a_3 \ge 2)\ . \tag{4}$$

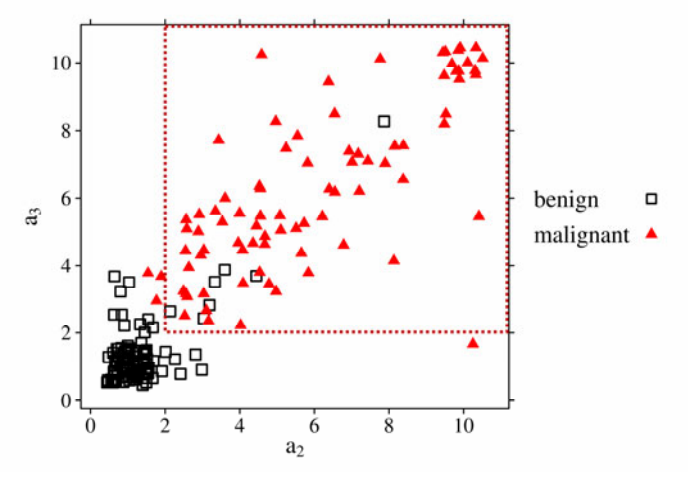


**Fig. 2.** Decision Boundaries (*dotted border*) found by the ANN rule with higher sensitivity

Using a polynomial kernel, OSRE extracted from SVM rules as in Table 6.

**Table 6.** OSRE – SVM Extracted Rules – Wisconsin Breast Cancer

| Specificity | Sensitivity | Rules |
|---|---|---|
| 0.9568 | 0.8690 | $[2 \le a_1 \le 10,\ 2 \le a_2 \le 10,\ 2 \le a_3 \le 10,\ 2 \le a_6 \le 10]$ |
| 0.9568 | 0.8928 | $[2 \le a_2 \le 10,\ 2 \le a_3 \le 10,\ 2 \le a_6 \le 10]$ |

OSRE extracted from SVM less rules than with ANN and again, the rule with higher sensitivity is similar to the rule extracted from ANN but with one more variable ($a_6$) to explain the problem.

$$(a_2 \geq 2)(a_3 \geq 2)(a_6 \geq 2)\ . \tag{5}$$

Replacing the input variables with their description, the extracted rule can be represented by decision boundaries shown in Figure 3.

$$(\text{Uniformity of Cell Size} \geq 2)(\text{Uniformity of Cell Shape} \geq 2)(\text{Bare Nuclei} \geq 2)\ . \tag{6}$$

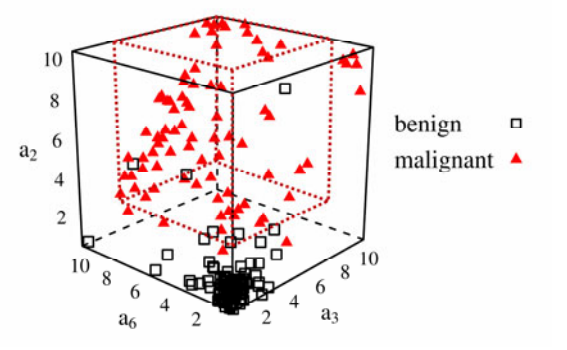


**Fig. 3.** Decision Boundaries (*dotted border*) found by the SVM rule with higher sensitivity

Rules extracted from ANN are thus consistent with those from SVM, albeit minor differences are also observed, as might be expected from using techniques that employ distinct types of decision surfaces.

**Iris.** The Iris dataset, widely viewed as a classification benchmark since Ronald Fisher introduced it in 1936, was also used in the OSRE introductory document [16] to test the ability of OSRE to represent ANN models using orthogonal rules. The Iris data has 150 items with 4 attributes, namely sepal length, sepal width, petal length and petal width, and 3 classes: Iris Setosa, Iris Virginica and Iris Versicolor. We extracted rules from ANN and SVM trained to represent the Iris Versicolor. Both ANN and SVM correctly classify 97.3% of cases in the training and in the validation phase. Rules extracted from ANN are shown in Table 7.

**Table 7.** OSRE – ANN Extracted Rules – Iris Dataset

| Specificity | Sensitivity | Rules |
|---|---|---|
| 0.9803 | 0.96 | $[2 \leq a_3 \leq 8,\ 2 \leq a_4 \leq 8]$ |
| 0.9803 | 0.96 | $[3 \leq a_4 \leq 5,\ 7 \leq a_4 \leq 8]$ |
| 0.9803 | 0.96 | $[4 \leq a_4 \leq 5]$ |
| 0.9411 | 0.89 | $[3 \leq a_3 \leq 8,\ 1 \leq a_4 \leq 5 \vee 7 \leq a_4 \leq 8]$ |

Using a polynomial kernel, OSRE extracted from SVM the rules shown in Table 8.

**Table 8.** OSRE – SVM Extracted Rules – Iris Dataset

| Specificity | Sensitivity | Rules |
|---|---|---|
| 0.9803 | 0.96 | $[3 \leq a_4 \leq 5, 7 \leq a_4 \leq 8]$ |

Once again, experimental analysis shows that OSRE extracted less rules from SVM than from ANN, thus improving the interpretability of the final solution.

## 5 Conclusions and Future Research

The study shows that rules can indeed be extracted from SVM models. Moreover, such rules are identical or similar to those extracted from ANN models. It is thus concluded that users of SVM, when trying to improve the interpretability of their models, can expect the support already available to users of ANN. Orthogonal Search-based Rule Extraction (OSRE) can be used to extract consistent and accurate rules from both SVM and ANN, adding to Data Mining tasks the interpretability and comprehensibility these tools lack.

Moreover, since SVM are capable of unique solutions [17] and unique solutions may mean simpler decision surfaces, it follows that OSRE may be able to extract less rules from SVM than from equivalent ANN models, as is indeed the case of two of the instances in the paper. Less rules, in turn, mean improved pattern interpretability. Indeed, the paper opens up an interesting line of research, namely that aimed at ascertaining whether there exists some type of intrinsic adequacy between model interpretation algorithms such as OSRE and model building algorithms such as SVM.

It would also be interesting to extend the application of OSRE to larger and more challenging datasets trained using other smooth classifiers, in order to test its unifying abilities beyond ANN and SVM. Extracted rules that stand out among different classifiers may represent a more generic description of the problem's solution.

## References

1. Fisher, D.H., McKusick, K.B.: An Empirical Comparison of ID3 and Back-Propagation. In: 11th International Joint Conference on Artificial Intelligence, vol. 1, pp. 788–793. Morgan Kaufmann, Michigan (1989)
2. Shavlik, J.W., Mooney, R.J., Towell, G.G.: Symbolic and Neural Learning Algorithms: An Experimental Comparison. Machine Learning 6, 111–143 (1991)
3. Weiss, S.M., Kapouleas, I.: An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods. In: 11th International Joint Conference on Artificial Intelligence, vol. 1, pp. 781–787. Morgan Kaufmann, Michigan (1989)
4. Etchells, T.A., Lisboa, P.J.G.: Orthogonal Search-based Rule Extraction (OSRE) for Trained Neural Networks: A Practical and Efficient Approach. IEEE Transactions on Neural Networks 17, 374–384 (2006)

5. Andrews, R., Diederich, J., Tickle, A.B.: Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. Knowledge-Based Systems 8, 373–389 (1995)
6. Aung, M.S., Lisboa, P.J., Etchells, T.A., Testa, A.C., Calster, B., Huffel, S., Valentin, L., Timmerman, D.: Comparing Analytical Decision Support Models Through Boolean Rule Extraction: A Case Study of Ovarian Tumour Malignancy. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007. LNCS, vol. 4492, pp. 1177–1186. Springer, Heidelberg (2007)
7. Lisboa, P.J.G., Jarman, I.H., Etchells, T.A., Ramsey, P.: A Prototype Integrated Decision Support System for Breast Cancer Oncology. In: 9th International Work Conference on Artificial Neural Networks, pp. 996–1003. Springer, San Sebastián (2007)
8. Jarman, I.H., Etchells, T.A., Martín, J.D., Lisboa, P.J.G.: An Integrated Framework for Risk Profiling of Breast Cancer Patients Following Surgery. Artificial Intelligence in Medicine 42, 165–188 (2008)
9. Lisboa, P.J.G., Etchells, T.A., Jarman, I.H., Aung, M.S.H., Chabaud, S., Bachelot, T., Perol, D., Gargi, T., Bourdès, V., Bonnevay, S., Négrier, S.: Time-to-event Analysis with Artificial Neural Networks: An Integrated Analytical and Rule-Based Study for Breast Cancer. Neural Networks 21, 414–426 (2008)
10. Pop, E., Hayward, R., Diederich, J.: RULENEG: Extracting Rules from a Trained ANN by Stepwise Negation. Queensland University of Technology, Australia (1994)
11. Tsukimoto, H.: Extracting Rules from Trained Neural Networks. IEEE Transactions on Neural Networks 11, 377–389 (2000)
12. Thrun, S.B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., De Jong, K., Deroski, S., Fahlman, S.E., Fisher, D., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Michalski, R.S., Mitchell, T., Pachowicz, P., Reich, Y., Vafaie, H., Van de Welde, W., Wenzel, W., Wnek, J., Zhang, J.: The MONK's Problems: A Performance Comparison of Different Learning Algorithms. Technical Report CS-91-197, Computer Science Department, Carnegie Mellon University, Pittsburgh (1991)
13. Wisconsin Breast Cancer, `http://archive.ics.uci.edu/ml/datasets.html`
14. Iris Dataset, `http://archive.ics.uci.edu/ml/datasets.html`
15. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Platt, J.C. (ed.) Advances in Kernel Methods: Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
16. Etchells, T.A.: Rule Extraction from Neural Networks: A Practical and Efficient Approach. Ph.D Dissertation. John Moores University, Liverpool (2003)
17. Burges, C., Crisp, D.: Uniqueness of the SVM Solution. In: Advances in Neural Information Processing Systems, vol. 12, pp. 223–229. MIT Press, Cambridge (2000)