Learning-based rule-extraction from support vector machines: Performance on benchmark data sets

Nahla Barakat[1] **Joachim Diederich** [1,2]

Faculty of Applied Sciences [1]
Sohar University, Sohar, PC311, Oman

School of Information Technology and Electrical Engineering [2]
The University of Queensland, Brisbane Q 4072, Australia
n.barakat@soharuni.edu.om , j.diederich@soharuni.edu.om,

Keywords: Rule-extraction, explanation, support vector machines

Abstract

Over the last decade, rule-extraction from neural networks (ANN) techniques have been developed to explain how classification and regression are realised by the ANN. Yet, this is not the case for support vector machines (SVMs) which also demonstrate an inability to explain the process by which a learning result was reached and why a decision is being made. Rule-extraction from SVMs is important, especially for applications such as medical diagnosis. In this paper, an approach for learning-based rule-extraction from support vector machines is outlined, including an evaluation of the quality of the extracted rules in terms of fidelity, accuracy, consistency and comprehensibility. In addition, the rules are verified by use of knowledge from the problem domains as well as other classification techniques to assure correctness and validity.

The approach can be summarised as follows:

- Subsets of benchmark data sets (Data set A) are used for SVM learning purposes, i.e. to build a model with acceptable accuracy, precision and recall.
- The resulting SVM model (classifier) is used to predict the class labels for two different subsets from each benchmark data sets. Thereby, the synthetic data sets B and C are generated.
- Data set B is used to train various machine learning techniques *with explanation capability* (decision tree learners). As a result, rules are extracted that represent the concepts learned by the SVMs as well as its generalization behaviour.
- The quality and validity of the extracted rules are assessed by applying them to data set C.

Four different benchmark data sets (available from the UCI repository) have been used to validate the approach.

The key question is how closely approximate the extracted rules the generalization behaviour of the SVM? This is evaluated by the use of data sets C. Results are shown in Table 1 (LOO = leave-one-out cross-validation).

Data Set (UCI)	SVM training "LOO" accuracy (data set A)	SVM model classification accuracy (data set C)	Extracted rules classification accuracy (data set C)	Extracted rules fidelity (data set C)	Extracted rules comprehensibility (#rules / #antecedents)
Diabetes (Pima Indian)	85 %	92%	93%	93%	2/1
Breast cancer	95%	87%	85%	88%	2/1
Heart Disease	82%	72%	74%	79%	3/2
Hepatitis	85%	83%	83%	100%	2/1

Table 1: Experimental results.

Summary of results

- The rule sets offer comprehensible explanation of the concepts learned by the SVM.

- The extracted rules are correct and valid from the medical point of view and consistent with rules represented by the training data sets (A).
- The rules extracted by the learning-based approach demonstrate high degree of fidelity, accuracy and consistency.

References

[1] N. Kasabov, "On–line learning, reasoning, rule extraction and aggregation in locally optimized evolving fuzzy neural networks", *Neurocomputing*, Vol. 1207, pp 1-21, 2001.

[2] A.B. Tickle, R.Andrews, M.Golea, and J.Diederich, "The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural network", *IEEE Transactions on Neural Networks*, Vol. 9(6), pp. 1057-1068, 1998.

Professor Joachim Diederich is Dean, Faculty of Applied Sciences, Sohar University and Honorary Professor in the School of Information Technology and Electrical Engineering as well as the Centre for Online Health at the University of Queensland. Prof Diederich's qualifications include a Habilitation in Computer Science from the University of Hamburg (Germany), a Doctorate in Computational Linguistics (summa cum laude) from the University of Bielefeld (Germany), and a Masters degree in Psychology from the University of Münster (Germany). Prof Diederich's research interests are in the area of machine learning and natural language processing.

Nahla Barakat is a Lecturer in the Faculty of Applied Sciences, Sohar University. Prior to her current appointment she was General Manager of the Central IS Department at Philips International - Alexandria, Egypt. Ms Barakat received a Bachelor degree in Electrical and Electronic Engineering and an MBA (Major IT) from Alexandria University, Egypt. Her current research interests are in machine learning and medical data mining.