Knowledge Initialisation for Support Vector Machines

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

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
j.diederich@soharuni.edu.om, n.barakat@soharuni.edu.om

Keywords: Knowledge initialisation, machine learning, support vector machines

Abstract

Since their introduction more than a decade ago, support vector machines (SVMs) have shown good performance in a number of application areas, including text classification, pattern recognition and bioinformatics. However, the success of SVMs comes at a cost – there is no way to utilise prior knowledge. SVMs are purely inductive learning machines. In this paper, a novel approach for rule-initialisation for support vector machines is presented. The application domain is medical diagnosis.

The approach presented here uses domain knowledge in the form of propositional rules to create a virtual data set to bias an SVM. The virtual data set is combined with real data for SVM learning. Knowledge-initialisation results in better classification accuracy and enhanced rule quality compared with purely inductive learning.

1. Introduction

The approach here is not unlike rule-refinement by use of neural networks [2,4]. The rule-refinement process starts with an initial knowledge base in the form of symbolic rules. The goal is to use learning technology to produce a "better" (i.e. a "refined") set of symbolic rules, which can then be applied in the original problem domain:

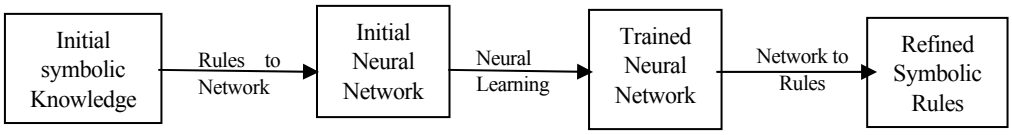

Figure 1: Rule Refinement using KBANN

In the context of support vector machine, rule-initialisation or similar approaches have been considered in the fields of image processing, text classification and natural language processing [1,3]. These studies selected or tailored appropriate kernel functions which leads to enhanced classification/prediction accuracy.

However, by design, SVMs are inductive learning machines. SVMs realise learning-from-scratch and therefore, any modification that allows the direct insertion of knowledge violates design principles. It is much more natural to encode prior knowledge in data and to use this data for SVM training.

In first experiments with the UCI “Diabetes” data set, we have used the following approach:

- Translate rules that represent available domain knowledge to virtual data which is used in addition to empirical data for SVM learning.
- Additional input data is used to explore the generalisation behaviour of the SVM. The SVM output is recorded.
- The resulting data set is used to train various machine learning techniques that produce propositional rules.
- The quality of the extracted rules is assessed in terms of fidelity, accuracy, comprehensibility, and consistency.

Results demonstrate that the extracted rules are correct and valid from a medical point of view. In addition, knowledge-initialisation improves the accuracy of SVM learning.

References

[1] Fung G., Mangasarian O., Shavlik J., "Knowledge-Based Support Vector Machine Classifiers", Advances in Neural Information Processing Systems, 2003,Vol. 15, pp: 521-528.

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

[3] Schölkopf, B., Simard P.Y., Smola A.J., Vapnik V.N., "Prior knowledge in support vector kernels" Advances in neural information processing systems, 1998, Vol. 10, pp 640-646.

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

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.