Hybrid Rule-Extraction from Support Vector Machines

Joachim Diederich
Faculty of Applied Sciences
Sohar University
Sohar, PC311, Oman

Nahla Barakat
Faculty of Applied Sciences
Sohar University
Sohar, PC311, Oman

School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane Q 4072, Australia

***Abstract* - Rule-extraction from artificial neural networks (ANNs) as well as support vector machines (SVMs) provide explanations for the decisions made by these systems. This explanation capability is very important in applications such as medical diagnosis. Over the last decade, a multitude of algorithms for rule-extraction from ANNs have been developed. However, rule-extraction from SVMs is not widely available yet. In this paper, a hybrid approach for rule-extraction from SVMs is outlined. This approach has two basic components: (1) data reduction using a logistic regression model and (2) learning based rule-extraction. The quality of the extracted rules is then evaluated in terms of fidelity, accuracy, consistency and comprehensibility. The rules are also verified against the available knowledge from the domain problem (diabetes) to assure correctness and validity.**

Keywords **– data mining, hybrid computational intelligence algorithms, rule-extraction and explanation.**

I. INTRODUCTION

In recent years, support vector machines (SVMs) have shown good performance in a number of application areas including text classification [1]. However the learning capability of SVMs comes at a cost: an inherent inability to explain the process by which a learning result was reached. Hence, the situation is similar to artificial neural networks (ANNs)[2] where the apparent lack of an explanation capability has led to various approaches aiming at extracting symbolic rules from neural networks. For SVMs to gain acceptance in areas such as medical diagnosis, it is desirable to offer an "explanation" capability.

A. Rule-Extraction and Machine Learning

One potential method of classifying rule-extraction techniques is in terms of the "translucency" of the view taken within the rule-extraction method of the underlying classifier. This motif yields two basic categories of rule-extraction techniques: "transparent" and "pedagogical" [2].

The distinguishing characteristic of the "transparent" (or decompositional) approach is that the focus is on extracting rules at the level of individual components of the underlying machine learning method. In neural networks, these are hidden and output units.

The classification -"pedagogical" or “learning-based” is given to those rule-extraction techniques that treat the underlying classifier as a "black box". Such techniques typically are used in conjunction with a learning algorithm that provides rule-based explanations and the basic motif is to use the trained classifier to generate examples for a second learning algorithm that generates rules as output. A third group in this classification scheme are composites that incorporate elements of both the "transparent" and "pedagogical" rule-extraction techniques. This is the "hybrid" or “eclectic” group [2] [3].

In case of SVMs, decompositional approaches can be based on the analysis of support vectors generated by the SVM while learning-based approaches learn what the SVM has learned. An example for learning-based rule-extraction from SVMs is Mitsdorffer [4].

The first part of this paper highlights the importance of rule-extraction algorithms and reviews some rule-extraction from ANN and SVMs techniques. The second part of the paper focuses on hybrid rule-extraction from SVMs loosely modeled after the DEDEC algorithm for neural networks. The application area is medical diagnosis.

B. The Importance of Rule-Extraction

1) Provision of an Explanation Component: The ability of symbolic artificial intelligence (AI) systems to provide a declarative representation of knowledge about the problem domain offers natural explanation for the decisions made by the system. Davis et al. [5] argue that even limited explanation can positively influence user acceptance. An

explanation capability can also provide a check on the internal logic of the system as well as being able to give a novice insight into the problem [6]. ANN's and SVMs have no such declarative knowledge structures, and hence, are limited in providing an explanation component.

2) Knowledge Acquisition for Symbolic Artificial Intelligence Systems: Constructing and debugging knowledge bases is the most difficult and time consuming task in building an expert system [7]. One of many motivations for introducing machine learning algorithms over the past decades was to overcome this 'knowledge acquisition' problem [8] [9]. Rule-extraction algorithms allow a trained ANN or SVM to be used as the basis for the construction of a knowledge/rule base.

3) Data Exploration and the Induction of Scientific Theories: Over time AI systems like ANNs and SVMs have proven to be extremely powerful tools for discovering previously unknown dependencies and relationships in data sets. As Craven & Shavlik [10] observe, *`a (learning) system may discover salient features in the input data whose importance was not previously recognised.'*

4) Improving the Generalisation of AI Solutions: In spite of the good generalization ability of SVMs, in some cases, when a limited or unrepresentative data set is used in the training process, the generalisation may fail, even with evaluation methods such as cross-validation. A rule-extraction process is essential to anticipate or predict a set of circumstances under which generalisation failure can occur.

C. Rule-Extraction from ANNs

Andrews et al. [3] proposed a classification schema for rule-extraction algorithms. The schema defines two basic categories of rule-extraction techniques *viz – ' decompositional'* and *`pedagogical'* and a third - labelled as *`eclectic'* - which combines elements of the two basic categories.

The *`decompositional'* approach focuses on extracting rules at the level of individual (hidden and output) units within the trained artificial neural network. The rules extracted at the individual unit level are then aggregated to form the composite rule base for the ANN as a whole. Some examples of decompositional methods are described in [11],[12],[13], and [14].

Methods classified as *'pedagogical'* are those rule-extraction techniques which treat the trained ANN as a *`black box*. Such techniques are typically used in conjunction with a symbolic learning algorithm. The basic motif is to use the trained artificial neural network to both classify examples and to generate examples for the learning algorithm. Some examples of pedagogical methods presented in [8],[10],[15], and [16]. The proposed third category in this classification scheme are composites which incorporate elements of both the *`decompositional'* and *`pedagogical'* (or*`black-box'*) rule-extraction techniques. This is the *`eclectic'* group [17].

D. DEDEC – A Hybrid Technique

DEDEC [17] uses the basic pedagogical motif and utilises the knowledge embedded in the architecture and weight vectors of the trained network to rank the inputs in order of their relative significance. This additional information is then used to direct the strategy for generating a minimal set of cases for the rule-extraction-as-learning phase. The following is a schematic outline of the DEDEC algorithm [17].

TABLE I
THE DEDEC ALGORITHM

(i) Assign each propositional variable from the problem domain (ie, a rule antecedent) to an input unit for an ANN. The network output unit corresponds to the decision outcome of the ANN applied to an input vector from the problem domain, ie, in essence the decision outcome corresponds to the rule consequent.
(ii) Train the ANN
(iii) Rank the input units in order of the relative share of their contributions to the output prediction using a weight partitioning algorithm based on Garson, (1991).
(iv) Initialise the set of symbolic rules as the attribute value(s) for the minimum set of input units required to form one valid case/example in the problem domain.
Repeat
select the next input unit from the ranked list (Created at step iii)
use the trained ANN to give the decision/classification for the set of cases/examples from the problem domain involving this input unit and all previously selected input units
for each case/example generated at the previous step
(a) update the rules for classifying each existing case/example in the presence of this new case;
(b) determine the rules for classifying the new case/example in the presence of the existing cases;
remove all rules which are not minimal
Until stopping criteria met

E. The Quality of the Extracted Rules

A framework is presented by Andrews et al. [3] for measuring the quality of rules extracted from neural networks. The criteria for the quality of rule-extraction include *fidelity, accuracy, consistency and comprehensibility*. Fidelity indicates the extent to which the rules mimic the behaviour of the ANN. Accuracy measures the correctness of classification of previously unseen examples. Consistency is the extent to which the generated rule sets produce the same classification of unseen examples, even if they are extracted from different ANNs trained on the same problem. Finally, comprehensibility is the number of rules plus the number of antecedents per rule.

F. *Rule-Extraction from SVMs*

Núñez et al. [18] introduced an approach for rule-extraction from SVMs (the SVM+ prototype method). The basic idea of this method is to use the output decision function from an SVM and then use K-means clustering to determine prototype vectors for each class. These vectors are combined with support vectors to define an ellipsoid in the input space for a mapping to if-then rules.

II. HYBRID SVM RULE-EXTRACTION

A. *The Problem Domain*

Diabetes mellitus is a chronic disease which is associated with increased concentration of glucose in the blood which in turn damages many of the body's blood vessels and nerves. People with IGT[1] are at a substantially higher risk of developing diabetes than those with normal glucose tolerance. The other risk factors include obesity, family history of diabetes[2] etc.

The problem to be solved here is the prediction of the onset of diabetes mellitus within 5 years by use of 8 variables present in the Pima Indian diabetic benchmark database[3]. Specifically, the approach here provides an explanation on how those predictions are reached.

For a medical application, it is crucial to understand the systems' decision in order to be confident that future prognoses are correct. In following section, we outline an approach for hybrid rule-extraction from SVMs applied to a medical domain

B. *The Approach*.

The basic idea is to reduce the data set before training and to select more representative patterns for training and testing purposes. Knowledge of the probability that a certain pattern belongs to the target class helps towards this end. Hence, our approach can be summarized in the following steps:

- A modified Pima Indians data set **M** is used to train a logistic regression model which in turn identifies the most significant attributes, and the probability by which each pattern of M belongs to the target class.
- The resulting model is used to independently select representative patterns for training and testing purposes from data set **M**. This is done by sorting data set M with regard to the probability of patterns to belong to the target class (obtained above). Hence, data set **A** is created.
- Data set **A** is used to train SVMs, i.e. to build a final model with acceptable accuracy, precision and recall.
- Synthetic data sets are generated with the same attributes but modified/different values to explore the generalisation behaviour of the SVM. That is, the SVM is used to predict the class labels for these data sets. Thereby, data set B, C, and D are obtained [19].
- Data sets B, C, and D are used to train a machine learning technique with explanation capability. Thereby, rules are generated that represent the generalisation behaviour of the SVM.

C. *The Experiment*

The Pima Indians diabetic database originally has 786 patterns. The risk factors are:

Inputs: 1- Number of times pregnant, 2- 2-hour OGTT plasma glucose, 3- Diastolic blood pressure, 4- Triceps skin fold thickness, 5- 2 hour serum insulin, 6- Body Mass Index (BMI), 7- Diabetes pedigree function**,** 8- Age - All attributes are numeric.

Output: Diagnosis (Diabetes onset within 5 years).

1) Training Examples: We have selected a random sample of 496 patterns from the original data set, after removing all patterns with zero value for attributes 2-4 plus patterns which probably include noise based on medical expertise. This leaves a data set **M** which has a considerable number of patterns with zero value for attribute 5. Data set M is used to train logistic regression model.

A data set **A** of 60 representative patterns for each probability range (from 0.999 to 0.53) is selected from data set M (see section B). Data set A has 40 negative, and 20 positive patterns which preservers the distribution of original data set.

2) SVM Training: Data set **A** is used for SVM training. A variety of learning parameters have been tested to reach an SVM model with an acceptable degree of accuracy, precision and recall. A linear SVM is sufficient and the results are shown in Table II.

TABLE II
SVM LEARNING (DATA SET A)

	Accuracy %	Precision %	Recall %
Training	95	90	95
Leave-one-out cross-validation	86	73	70

[1] IGT refer to levels of blood glucose concentration above normal range (OGTT 140-199), but below those which are diagnostic for diabetes (OGTT >=200)

[2] Diabetes pedigree function

[3] Available at the UCI Machine Learning Repository

3) SVM Generalisation: Data Sets B, C and D: The SVM model is used to predict the classes in data sets **B, C** and **D**, which have different distributions of positive and negative examples compared to the training set (Table III). The eight risk factors are used as input to the SVM which in turn predicts a target class for each pattern. This results in data sets B, C, and D. These data sets are used to extract the rules learned by the SVM, and then, test the quality of one of the extracted rules.

TABLE III
PATTERNS DISTRIBUTIONS

Data set	Positive	Negative	Rule set
A(training)	38%	62%	A
B (testing)	30%	70%	B
C (testing)	49%	51%	C
D (testing)	60%	40%	D

D. Results.

1) Decision Tree Learning: The C5 algorithm [20] is used to generate decision trees and rule sets from data sets A, B, C, and D. To ensure the quality of extracted rules, leave-one-out cross-validation is used.

Comparing the rule sets which are indirectly generated from the SVM classifier by use data sets B, C, and D with the rules generated from the training data set A (Table IV; coverage 10% and confidence 0.6), it can be shown that the most significant risk factors are present (plasma glucose and BMI). Yet, the total number of rules is slightly different. The difference in the number of rules is partially due to the difference in the distribution of positive and negative examples in data sets.

2) Measuring the Quality of Extracted Rules: Comparing the extracted rule sets in Tables V-VII, it can be noticed that rule set C has only two rules, and only one antecedent attribute. Yet rule set B has more coverage, wider range of attribute values, and is more consistent with rule set D and the clinical knowledge of the problem domain in terms of upper and lower bounds for the plasma glucose attribute. Hence, rule set B (specifically rule # 1 for 0, & rule #1 for 1) is selected for testing fidelity, accuracy and consistency. That is, rule set B is used to classify data sets B, C, and D. The resulting class for each pattern "Rclass" is compared with the target class for this pattern. This measures the accuracy of the rule set.

To test the fidelity, the "Rclass" for each pattern is compared with the output class from the SVMs "SVMclass" for the same pattern.

The algorithms used for measuring fidelity of rule set B is outlined in Table VIII. Results are shown in Table IX.

TABLE IV
RULE SET A

Rules for 0:
Rule #1 for 0:
if BMI =< 43.5
and Plasmaglo =< 107
then -> 0 (17, 1.0)
Rule #2 for 0:
if BMI =< 43.5
and Plasmaglo > 107
and Age =< 36
and 2hourserum =< 165
then -> 0 (12, 1.0)
Rules for 1:
Rule #1 for 1:
if BMI =< 43.5
and Plasmaglo > 129
and Age =< 36
and 2hourserum > 165
then -> 1 (6, 0.833)
Rule #2 for 1:
if BMI =< 43.5
and Plasmaglo > 107
and Age > 36
then -> 1 (14, 0.786)
Rule #3 for 1:
if BMI > 43.5
then -> 1 (6, 1.0)

TABLE V
RULE SET B

Rules for 0:
Rule #1 for 0:
if Plasmaglo =< 124
then -> 0 (36, 1.0)
Rule #2 for 0:
if Plasmaglo > 124
and Plasmaglo =< 143
and BMI =< 36.4
then -> 0 (11, 0.909)
Rules for 1:
Rule #1 for 1:
if Plasmaglo > 143
then -> 1 (14, 1.0)

TABLE VI
RULE SET C

Plasmaglo =< 106 [Mode: 0] (14, 1.0) -> 0
Plasmaglo > 106 [Mode: 1] (13, 1.0) -> 1

TABLE VII
RULE SET D

Rules for 0:
Rule #1 for 0:
if Plasmaglo =< 143
and Age =< 43
and BMI =< 36.4
then -> 0 (34, 0.971)
Rules for 1:
Rule #1 for 1:
if Plasmaglo =< 143
and Age > 43
then -> 1 (10, 0.8)
Rule #2 for 1:
if Plasmaglo > 143
then -> 1 (44, 0.909)

TABLE VIII
FIDELITY MEASUREMENT ALGORITHM

```
f = 0
    For each pattern i in a data set of size I
        Find the class SVMclass for i by use of SVM
        Find class Rclass for i by use of Rule set
        if Rclass = SVMclass
        then f++
fidelity= f/I
```

TABLE IX
FIDELITY & ACCURACY FOR RULE SET B

Data set	rule Fidelity	rule Accuracy	SVM classification accuracy
B (testing)	92	83	87
C (testing)	92	88	92
D(testing)	80	78	86

The best fidelity and accuracy for rule set B is obtained by applying it to data sets B & C. This can be attributed to the smaller number of patterns in data set C, and for data set B it is expected as it is the data set used for the rule-extraction. It can also be noticed that the rule accuracy is slightly lower than the SVM classification accuracy, which is also expected as the fidelity is lower than 100%.

Finally, the worst accuracy result is obtained when applying rule set B to data set D, which is the extended data set, with noise and modified patterns. This can be also attributed to the lower accuracy of SVM classification, and lower fidelity of the rule set. However, the set still has better fidelity.

In terms of consistency, changing the training parameters for the linear SVMs did not result in different rule sets for the same data set.

However, rule set B is consistent with rule set D, especially for the rule that classify positive patterns. The slight difference in rules for classifying negative patterns can be attributed to the distribution of positive and negative examples.

III. CONCLUSION

The decision trees and rule sets produced by C5 offer an explanation of the concepts learned by the SVM. The extracted rules are correct and valid from the medical point of view (confirmed by a domain expert) and consistent with clinical knowledge of diabetes risk factors. It can also be shown that the rules extracted by this approach demonstrate a high degree of fidelity and accuracy.

In summary, hybrid rule-extraction from support vector machines offers opportunities for clinical applications.

REFERENCES

[1] T. Joachims, "Transductive Inference for Text Classification using Support Vector Machines", International Conference on Machine Learning (ICML), 1999.

[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.

[3] R. Andrews, J. Diederich, and A.B. Tickle, "A Survey and Critique of Techniques For Extracting Rules From Trained Artificial Neural Networks", Knowledge Based Systems, 1995, vol. 8, pp. 373-389.

[4] R. Mitsdorffer, J. Diederich, and C. Tan, "Rule-extraction from Technology IPOs in the US Stock Market", ICONIP02, Singapore, 2002.

[5] R. Davis, B.G. Buchanan, and E. Shortcliff, "Production Rules as a Representation for a Knowledge Based Consultation Progra", Artificial Intelligence, 1977, vol. 8(1), pp.15-45.

[6] S. Gallant, "Connectionist Expert System", Communications of the ACM, 1988, vol. 31 (2), pp. 152-169.

[7] S. Sestito, and T. Dillon, "Automated Knowledge Acquisition", Prentice Hall, Australia, 1994.

[8] K. Saito and R. Nakano, "Medical Diagnostic Expert System Based on PDP Model", IEEE International Conference on Neural Networks (San Diego CA), 1988, vol. 1, pp. 255-262.

[9] S. Sestito and T. Dillon, "Automated Knowledge Acquisition of Rules With Continuously Valued Attributes", 12th International Conference on Expert Systems and their Applications (AVIGNON'92), Avignon -France, 1992, pp. 645-656.

[10] M.W. Craven, and J.W. Shavlik, "Using Sampling and Queries to Extract Rules From Trained Neural Networks", Proceedings of the 11th International Conference on Machine learning, NJ, 1994, pp.37-45.

[11] L.M Fu. "Rule Learning by Searching on Adapted Net" Proceedings of the Ninth National Conference on Artificial Intelligence, Anaheim CA,1991, pp 590-595.

[12] G. Towell, and J. Shavlik. "The Extraction of Refined Rules From Knowledge Based Neural Networks", Machine Learning, 1993, vol. 131, pp.71-101.

[13] M.C. Mozer, C. McMillan, and P. Smolensky, "The Connectionist Scientist Game: Rule Extraction and Refinement in a Neural Network", Proc of the 13th Annual Conference of the Cognitive Science Society, Hillsdale NJ, 1991.

[14] V. Tresp, J. Hollatz, and S. Ahmad, "Network Structuring and Training Using Rule-Based Knowledge", Advances in NEural Information Processing, 1993, vol. 5, pp. 871-878.

[15] S. Thrun, "Extracting Provably Correct Rules From Artificial Neural Networks", Technical Report IAI-TR-93-5, Institut fur Informatik III Universitaet Bonn, 1994

[16] M.W. Craven, and J.W. Shavlik, "Extracting Tree– Structured Representation of Trained Networks", Advances in Neural Information Processing Systems, MIT Press, Cambridge, 1996, vol. 8, pp.24-30.

[17] A. Tickle, A, M. Orlowski, M, J. Diederich, "DEDEC: A Methodology for Extracting Rules from Trained Artificial Neural Networks. "In: Andrews, R.; Diederich, J. (Eds.): Rules and Networks. Brisbane, Qld.: QUT Publication 1996, 90-102.

[18] H. Núñez, C. Angulo, and A.Catala, "Rule-extraction from Support Vector Machines", Proceedings of European Symposium on Artificial Neural Networks, Burges, 2002, ISBN 2- 930307-02-1, pp.107-112 .

[19] N. Barakat, and J. Diederich, "Learning-based rule-extraction from support vector machines", The 14th International Conference on Computer Theory and applications ICCTA'2004, Alexandria, Egypt, Sept, 28-30, 2004, in press.

[20] http://www.rulequest.com