Building Acceptable Classification Models for Financial Engineering Applications *Thesis Summary*

David Martens
Department of Decision Sciences and Information Management
Katholieke Universiteit Leuven, Belgium

Classification is a popular data mining task, where the value of a discrete (dependent) variable is predicted, based on the values of several independent variables. In this research, we investigate how predictive classification models can be inferred from the available data. The classification models are required to make good predictions, and be comprehensible and intuitive. The aspect of humanly understandable and intuitive models is of crucial importance in any domain where the model needs to be validated before it can be implemented, such as in the medical diagnosis and credit scoring domain. A classification model that is accurate, comprehensible and intuitive is defined in this thesis as acceptable for implementation. Building such acceptable models is the goal of this text.

We examine how rule based classifiers can be built that satisfy these requirements. In a first approach, we use rule extraction from Support Vector Machines (SVMs) to extract rules that are accurate, comprehensible, and mimic the SVM model as much as possible. Next, the use of artificial ant colonies for classification is studied, attempting to induce acceptable classification models from data. In a final part, we discuss the application of the investigated algorithms for real-life case studies, such as the prediction of defaults, going concern opinions, software faults, and business/ICT alignment.

SVM Rule Extraction

We examine how rule based classifiers can be built that satisfy the aforementioned prerequisites. An initial exploratory benchmarking study examines the current opportunities for SVM rule extraction [2]. With these lessons learnt, a new methodology is developed for SVM rule extraction: active learning based approach (ALBA) [6]. ALBA extracts rules from the trained SVM model by explicitly making use of key concepts of the SVM: the support vectors, and the observation that these are typically close to the decision boundary. Active learning implies the focus on apparent problem areas, which for rule induction techniques are in the regions close to the SVM decision boundary where most of the noise is found. By generating extra data close to these support vectors, that are provided with a class label by the trained SVM model, rule induction techniques are better able to discover suitable discrimination rules.

Classification with Ant Colony Optimization

The use of Ant Colony Optimization (ACO) for classification is investigated in depth, with the development of the AntMiner+ algorithm [5]. AntMiner+ builds rule based classifiers, with a focus on the predictive accuracy and comprehensibility of the final models. The key differences between the proposed AntMiner+ and previous AntMiner versions are the usage of the better performing $\mathcal{MAX}$-$\mathcal{MIN}$ ant system, a clearly defined and augmented environment for the ants to walk through, with the inclusion of the class variable to handle multiclass problems, and the ability to include interval rules in the rule list. Furthermore, the commonly encountered problem in ACO of setting system parameters is dealt with in an automated, dynamic manner. In a benchmarking study, AntMiner+ is compared with several state-of-the-art classification techniques, such as C4.5, RIPPER and support vector machines. These experiments show an AntMiner+ accuracy that is superior to that obtained by the other AntMiner versions, and competitive or better than the results achieved by the included classification techniques.

Incorporating Domain Knowledge

Incorporating domain knowledge is of great importance, as whenever comprehensibility is demanded, justifiability is of crucial importance as well. Models are required to be comprehensible as to validate the intuitiveness of the model. Being able to incorporate domain knowledge will therefore be of major importance for success of any data mining application. This performance criterion is often overlooked in the research, even by those who do acknowledge the importance of the comprehensibility aspect. As our case studies reveal, the intuitiveness of the models turns out to be an absolute necessity. This option to incorporate domain knowledge is also of much help when a limited amount of data is available.

Next to a motivation of this essential requirement, it is shown how the AntMiner+ technique can be extended to incorporate such domain knowledge [4]. By changing the environment and influencing the heuristic values, we can respectively limit and direct the search of the ants to those regions of the solution space that the expert believes to be logical and intuitive.

We also make an attempt to come to a measurement for this evaluation criterion. Although measures for predictive performance (and to some degree comprehensibility) have been well researched, measuring justifiability has not yet been handled. The proposed measurement can be applied to all rule based and linear classifiers.

Real-Life Case Studies

The AntMiner+ technique is validated as a potential data mining tool to be used in business practices with a number of real-life case studies. The applications can be divided in financial engineering and information systems applications. As the aim is the final acceptance of the model, all these cases are carried out in close collaboration with the respective domain experts.

The first case study concerns the assessment of credit risk, and more specifically determining whether a counterpart will default on his/her financial obligation. Since the credit risk models will be subject to supervisory review and evaluation, they must be easy to understand and transparent. Hence, techniques such as neural networks or support vector machines are less suitable due to their black box nature. Building upon previous research, AntMiner+ is used to build internal rating systems for credit risk. Experiments are conducted using various types of credit data sets: retail, small- and medium-sized enterprises (SMEs) and banks. It will be shown that the extracted rule sets are both powerful in terms of discriminatory power, and comprehensibility. Furthermore, a framework is presented describing how AntMiner+ fits into a global Basel II credit risk management system.

The second case concerns the prediction of the going concern opinion [3], as issued by an auditor (such as KPMG, Deloitte, PWC and Ernst & Young). The auditor is required to evaluate whether substantial doubt exists about the client entity's ability to continue as a going concern. Accounting debacles in recent years have shown the importance of proper and thorough audit analysis. To provide specific audit guidelines, we infer rules that are subsequently converted into a decision table allowing for truly easy and user-friendly consultation in every day audit business practices.

Alignment between business and ICT (Information and Communication Technology) in organisations is still high on the management agenda of many a Chief Information Officer (CIO), and constitutes our third practical case [1]. Most often, making sure that investments in ICT are in harmony with the organisation's business objectives proves to be more challenging than initially expected, especially in today's fast-changing, dynamic environment. A lot has been written on B/ICT alignment, yet there are few studies that come up with actionable results that can be used by practitioners. In this study practical guidelines for managers are described on how to strive for better alignment of ICT investments with business requirements based on the inferred patterns.

Finally, AntMiner+ is used for software fault prediction [7]. Software managers are routinely confronted with software projects that contain errors or inconsistencies and exceed budget and time limits. By mining software repositories with comprehensible data mining techniques, predictive models can be induced that offer software managers the insights they need to tackle these quality and budgeting problems in an efficient way. This case study focuses on the prediction of errors in software modules by the use of data mining techniques. This enables software managers to focus their testing activities on those software modules that are classified as fault-prone.

PhD Committee

- Bart Baesens (*supervisor* - K.U.Leuven)
- Jan Vanthienen (*supervisor* - K.U.Leuven)
- Tony Van Gestel (Dexia Group)
- Thomas Stützle (Universite Libre de Bruxelles)
- Foster Provost (New York University)
- Dolores Romero-Morales (University of Oxford)

REFERENCES

[1] B. Cumps, D. Martens, M. De Backer, S. Viaene, G. Dedene, R. Haesen, M. Snoeck, and B. Baesens. Inferring rules for business/ict alignment using ants. *Information and Management*, Forthcoming.

[2] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007.

[3] D. Martens, L. Bruynseels, B. Baesens, M. Willekens, and J. Vanthienen. Predicting going concern opinion with data mining. *Decision Support Systems*, Forthcoming.

[4] D. Martens, M. De Backer, R. Haesen, B. Baesens, C. Mues, and J. Vanthienen. Ant-based approach to the knowledge fusion problem. In *Proceedings of the Fifth International Workshop on Ant Colony Optimization and Swarm Intelligence*, Lecture Notes in Computer Science, pages 85–96. Springer, 2006.

[5] D. Martens, M. De Backer, R. Haesen, M. Snoeck, J. Vanthienen, and B. Baesens. Classification with ant colony optimization. *IEEE Transaction on Evolutionary Computation*, 11(5):651–665, 2007.

[6] D. Martens, T. Van Gestel, and B. Baesens. Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering*, Forthcoming.

[7] O. Vandecruys, D. Martens, B. Baesens, C. Mues, M. De Backer, and R. Haesen. Mining software repositories for comprehensible software fault prediction models. *Journal of Systems and Software*, 81(5):823–839, 2008.