# Special Issue on FLAIRS 2004
# Feature Selection and Extraction for Classification

INGRID RUSSELL, University of Hartford, USA
ZDRAVKO MARKOV, Central Connecticut State University, USA
BRIAN CARSE, University of West of England, UK
ANTHONY G. PIPE, University of West of England, UK
LAWRENCE B. HOLDER, University of Texas at Arlington, USA

## Preface

Welcome to this special issue of the International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). This issue contains articles that expand on papers published in the proceedings of the 17th International Florida AI Research Society (FLAIRS) in May 2004. Approximately 15% of the papers submitted to the special tracks on Neural Network Applications and Machine Learning held during the 17th International FLAIRS conference were accepted for publication in the special issue.

In the lead article, the issue's editors set the scene and context of current & future trends in feature selection and extraction for classification problems. Below, we briefly summarize the other articles in this issue.

Text classification is an intensively studied learning task, both because it is difficult and because it is an important application for AI. Generally the difficulties arise from the classical vector-space model where text documents are represented in a very high dimensional space determined by the large number of terms (words) in the corpora. Small documents in this space that are semantically similar may not share any terms, and thus appear very distant in the vector space. Conversely, documents that occasionally share terms, but are not related to each other may appear closer than they actually are. A solution to this problem is the so-called Latent Semantic Indexing (LSI) approach. By using the Singular Value Decomposition (SVD) technique, the original term-to-document matrix is transformed into a smaller one where, instead of the original terms, orthogonal factors are used to represent documents. Zelikovitz and Marquez further explore this approach by applying it in a semi-supervised setting. In particular, they use and incorporate the so-called transductive learning scheme into the SVD process. The basic idea is that the test examples are also used in the learning process, which improves the accuracy of the classification process. Experiments with the nearest neighbor classifier and a number of short-text popular data sets are performed. The results show that using the test examples in the SVD process can be more useful than extending the data set with more labeled examples. Another set of experiments shows that incorporating the transductive approach into the SVD process provides additional vocabulary that has not been seen in the training set and thus allows small documents to be classified properly.

Using unlabeled data for classification is becoming a popular alternative to the classical training and testing supervised learning framework in domains where labeling requires substantial human effort. Typical examples are text categorization and sequence analysis (e.g. labeling genes and EEG time-series). While Zelikovitz & Marquez's article investigates the former application area, Zhong focuses on sequence classification. This article's main contribution is an empirical study of semi-supervised learning of hidden Markov models (HMM). The article first introduces the semi-supervised learning paradigms and presents the basics of the hidden Markov models used in the study. Further, the semi-supervised setting is introduced by modifying the HMM-based k-means clustering algorithm. The basic idea is to use the labeled data to initialize HMMs and to constrain the parameter estimation of

HMMs. Three different versions of this algorithm based on different strategies to combine labeled and unlabeled data are presented and empirically evaluated. The experiments are carried out with a synthetic HMM-generated dataset and a real EEG data set. The results generally show that introducing unlabeled sequences improves the classification accuracy especially when labeled data are used in the training process.

The experiments also provide some evidence for comparing two semi-supervised learning paradigms; regular semi-supervised learning, where the training set contains both labeled and unlabeled data and the classifier is used to classify separate set of unlabeled data (used also as test data to evaluate the performance of the classifier), and transductive learning, which does not separate unlabeled training data from unlabeled test data.

Zhang presents a theoretical analysis of one of the most popular classification algorithms, Naïve Bayes, and suggests a deeper explanation for the algorithm's surprisingly good performance in cases when its basic assumption of conditional independence is not met. The well-known explanation for the good performance of Naïve Bayes' is based on the fact that its error is evaluated with the zero-one loss function, which does not penalize inaccurate probability estimation as long as the maximum probability is assigned to the correct class.  Zhang argues, however, that this explanation does not uncover why the strong dependencies among attributes could not change the classification and provides a new theoretical explanation based on the distribution of the local attribute dependencies in each class. The basic idea is to investigate the interaction between local dependencies across different classes, i.e., whether they work together consistently (supporting certain classification) or inconsistently (cancel each other). This new approach allows definition of formal conditions for the optimality of Naïve Bayes, which are presented and proven in the paper. Further, the optimality of Naïve Bayes under Gaussian distribution, when dependency among attributes exists, is investigated and a sufficient condition for optimality in this case is proven. The paper also presents an application of the theoretical analysis of Naïve Bayes' conditional dependence to learning a class of Bayesian networks. An extension of an existing algorithm for this purpose is presented and empirically evaluated.

Neural networks are among the most popular learning algorithms. Bisant describes an application of a neural network combined with embedded feature selection to perform document genre identification. With the explosion of email traffic, email filtering is becoming increasingly important. Current methods can filter based on information including sender recognition, date information, and topical combination of keywords. Such methods have achieved limited success. Genre identification seeks to identify the family of a given email. The article presents work on the use of a technique borrowed from molecular sequence analysis, applied to sequence analysis and genre identification in email filtering. Neural networks are employed to improve the performance in automated genre identification.  Results presented show that the neural network approach performs significantly better than the next best approach, that of decision trees. In addition, neural networks are further used to identify the most significant features in the identification process. The analysis of the features indicates that second order information is being exploited by the networks for better performance, meaning that neural networks will outperform statistical models or other methods that only utilize first order information.

Multi-layer perceptron (MLP) neural networks have been widely used in various pattern recognition applications. It is well known that MLP training algorithms are data dependent and, as a result, several pre-processing techniques have been developed to improve training performance. Yu et al present an analysis of the impact on training dynamics of pre-processing techniques. To set the scene, the article first summarizes pre-processing techniques and reviews commonly used MLP training algorithms. A theoretical treatment is then developed, using a concept of “equivalent states”, which analyzes the conditions under which pre-processing of data leads to different dynamics during training. Yu et al demonstrate that, under certain conditions, there is no significant impact on improvement of the training process using pre-processed data compared with the original data on a suitably transformed initial set of network parameters (weights and thresholds). This analysis is given for conjugate descent, back propagation and Newton methods. A new training algorithm is then proposed which extends Chen’s ‘output weight optimization – hidden weight optimization’ (OWO-HWO) method by applying a Newton-like learning technique to the HWO part of the algorithm to exploit second order information to accelerate network learning. Experimental results are presented which confirm the earlier theoretical analysis, and which demonstrate the improved performance of the proposed training algorithm on a variety of real-world data sets.

While fully connected neural networks, and other networks with monolithic structure, have produced good results when applied to certain classification problems, research has shown that an increase in complexity associated with large input spaces causes a significant reduction in network performance. Modular neural networks have shown potential in overcoming such problems particularly in areas associated with applications that introduce scalability and data complexity issues such as large databases with complex interrelationships between data. Modular neural networks’ independent modules allow for training to be carried out in parallel, thus significantly reducing training effort and making it easier to add modules without the need for retraining others. Ferguson et al take such an approach, presenting a modularized neural network model for solving a complex problem of character recognition. Experimental results illustrate its effectiveness, yielding enhanced learning and generalization performance. This is accomplished by first identifying network modules through task decomposition using expert knowledge and clustering approaches. The function of these modules are then improved with decision tree learning algorithms. This combination of neural network modules and decision tree learning results in an improved design and a significant increase in the overall system performance.

In conclusion, we wish to thank the reviewers of these papers for their timely and detailed contributions. These efforts ensured the high quality of this issue. We are also thankful to Lakshmi Narayan and her staff at World Scientific Publishing for providing great support during the process. We hope the reader will enjoy this special issue.