Intelligent Data Analysis 16 (2012) 741–743
DOI 10.3233/IDA-2012-0548
IOS Press

Preface

Combined learning methods and mining complex data

Machine Learning has shown tremendous progress in the last decades. Started as a sub-domain within Artificial Intelligence it has grown into an independent field of study. New research problems have been identified, many new methods have been introduced and the number of their applications in various areas is still increasing. Machine Learning together with the related, younger field of Data Mining have become powerful tools for knowledge discovery in databases and intelligent data analysis.

One of the most popular tasks in learning is supervised learning, in particular discovering classification knowledge from data, which usually leads to creation of classifiers. The classifiers could be constructed in different ways, the most common ones being decision trees, decision rules, Bayesian approaches, instance based learning, artificial neural networks, discriminant function or support vector machines. Nevertheless, most of both past and present research concerns developing single learning algorithms. However, according to both theoretical (e.g. "no free lunch theorem") considerations as well as experimental experiences one cannot expect of one single algorithm to be simply the best. Each reasonable algorithm has its own area of superiority and it cannot outperform others in all possible learning problems. Overcoming limitations of single algorithms and improving predictive accuracy could be achieved by integrating several diversified classifiers into a combined system. Such systems, known under the names *ensembles* or *multiple classifiers*, have received a noticeable research and application interest at least since the late 80's. As a result, several different approaches for generating component classifiers and aggregating their predictions have been proposed. The most influential works include Bagging and Boosting, which manipulate learning examples. However, solving difficult problems involves also changing feature representations and integrating classifiers learned from different feature subsets. Multiple classifiers are also used in the context of learning from dynamic data streams, where the target classes may change over time. Furthermore, for some problems hybrid classifiers are suitable, which use multi-strategic learning that fits the data with different representations.

Another critical issue concerns observation that most of learning or data mining algorithms have been developed for stable, tabular representations of data. Such tabular data are present in many software systems and they are often easily obtained from relational databases – commonly met in any application domain. Moreover, data coming from other sources are often transformed into this format; see, e.g., processing text documents with natural language techniques or extracting numerical features from images. However, in some domains this data model appears to be too restrictive. Many modern automatic systems in science, engineering, medical or social fields are able to collect larger data with increasing their structure. This growing complexity comes both from the need for getting richer and more precise descriptions of real world objects and from development of new technologies for their measuring or collecting. For example, such *complex data* may appear in biology, bio-informatics, analysis of heterogeneous representations/files inside the patient electronic record, mining graph structure, industrial

or business workflows as well as dynamic networks. In some of these problems it is also necessary to incorporate domain knowledge. Mining more complex, larger and, generally speaking, "more difficult" data sets, poses new challenges for researchers and asks for novel and dedicated approaches.

Following the above-mentioned motivations Jerzy Stefanowski organized the special session called "Combined Learning Methods and Mining Complex Data" which was located within the 7th Rough Sets and New Trends in Computing Conference (RSCTC 2010) that was held in Warsaw, Poland, June 28-30, 2010.

Besides the general goals covering multiple classifiers and mining difficult, complex data, the session was also focused on the data characteristics referring to the following challenges:

1. *Processing data streams* and learning in changing environments. Many sources generate data continuously with the distributions and target concepts changing over time. Mining such data streams and adapting to *concept drifts* is of great interest from theoretical and practical perspective
2. *Semi-supervised learning*. In many domains only a limited number of labelled examples is available, therefore, learning approaches should take as much advantage of unlabelled data as it is possible to produce an efficient solution. In case of classification, it results in the development of such approaches as *co-training* or *active learning*.
3. *Learning from imbalanced data*, where one class contains much smaller number of examples than the remaining classes. The imbalanced distribution of classes constitutes a difficulty for standard algorithms for learning classifiers and calls for specialized approaches.

Furthermore, application papers devoted to these topics were expected. Over 20 papers were submitted. After detailed reviewing and considering the conference limits, 10 papers were accepted and finally presented during double session slots. More information about the session, Program Committee and the final program can be found at http://www.cs.put.poznan.pl/jstefanowski/cld.html.

As the session received such a high interest both from authors and conference audience, the session organizer decided to prepare a journal special issue devoted to these topics.

The authors of invited presentation significantly extended their contribution, added new methodological elements as well as carried out more experimental validations. To sum up, the present issue of the "Intelligent Data Analysis" journal contains five papers, which are briefly characterized below:

The first paper concerns semi-supervised learning algorithms based on co-training ideas, where multiple different sets of features, called views, are used to train classifiers. The crucial issue in this kind of learning is the identification of unlabeled examples to be asked for labeling and to be added to the training set. The main authors' contribution is to propose a new agreement based sampling procedure, which replaces the random sampling inside the strategy for selecting unlabeled examples. This procedure is applied together with multi-viewed semi-supervised algorithms and successfully evaluated in a large series of experiments.

The second paper describes the study of applying multiple classifiers to data within the Australasian Data Mining 2009 Analytic Challenge. The author's approach was awarded with the Grand champion prize for achieving the best overall result. In this paper, the purpose of the challenge is described and details of the winning approach are given. The main idea of, so called, Genetic Meta Blender, is to use a version of a genetic algorithm to optimize proportions of contributions of component classifiers into the final decision.

The third paper contains a proposal of a new ensemble IIvotes for learning from imbalanced data. This is an adaptive ensemble where a special focused pre-processing method, called SPIDER, is combined with importance sampling while constructing training subsets in each of subsequent iterations. Such integration changes the bias of typical classifiers and leads to better balance between the sensitivity

and specificity measures for the minority class. Additional contribution includes showing advantages of abstaining component rule classifiers, so refraining from making their predictions when they are uncertain as to a classified instance.

Learning concept drift in data streams is a topic of the fourth paper. Authors paid attention to special case of recurring concepts. They proposed and studied extensions and combination of some existing approaches, which include drift detection method and reusing previously learned classifiers in situations where some older concept reappear. The key point of their proposal is retrieving the most appropriate concepts for particular contexts and dealing with removing stored classifiers when the available memory is limited.

Finally, the fifth paper is devoted to mining difficult biomedical data. A novel data representation for learning from gene expression data is introduced where gene-gene interactions play important role. Another benefit of this representation is possibility of incorporating external knowledge in the form of semantic similarity based on the Gene Ontology. The novel representation together with this semantic similarity are applied to guide feature selection and their weighting. These proposals are evaluated in experiments on genetic data sets, where ensembles like Random Forest are also applied.

I take this opportunity to thank all contributors for agreeing to write their papers to this special issue. Moreover, I wish to express my gratitude to all colleagues who helped me in reviewing these papers. Finally, I owe a vote of thanks to Dr A. Fazel Famili, Editor in Chief of Intelligent Data Analysis – an International Journal, for his positive decision about inviting these papers to the journal and his support for my efforts.

Jerzy Stefanowski
Poznań, June 201