BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

New approaches in Data Mining. Applications in Educational Data Mining

PhD thesis summary

PhD student: Liana-Maria Crivei
Scientific supervisor: Prof. Dr. Czibula Gabriela

September 2020

Keywords: Educational data mining, students performance prediction, relational association rule mining, random forests, deep neural networks, self-organizing maps.

Contents

Thesis table of contents

List of publications

All rankings are listed according to the 2014 classification of journals[1] and conferences[2] in Computer Science.

Publications in Web of Science - Science Citation Index Expanded

[CCMC19] Czibula, G., Czibula, I.G., Miholca, D.L., **Crivei, M.L.**. A novel concurrent relational association rule mining approach. *Expert systems with Applications* 125(6), 2019, pp. 142–156. (**indexed Web of Science, IF=3.768**)

Rank A, 4 points.

Publications in Web of Science - Conference Proceedings Citation Index

[MCC18a] Diana-Lucia Miholca, Gabriela Czibula, **Liana Maria Crivei**. *A new incremental relational association rules mining approach.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES), 2018, Procedia Computer Science, Volume 126, 2018, pp. 126-135. (**indexed Web of Science**)

Rank B, 4 points.

[CMC19] Gabriela Czibula, Andrei Mihai, **Liana Maria Crivei**. *SPRAR: A novel relational association rule mining classification model applied for academic performance prediction.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES), Procedia Computer Science, Volume 159, 2019, pp. 20-29. (**indexed Web of Science**)

Rank B, 4 points.

[CAG19] **Liana Maria Crivei**, Andrei Mihai, Gabriela Czibula. *A study on applying relational association rule mining based classification for predicting the academic performance of students.* The 12th International Conference on Knowledge Science, Engineering and Management (KSEM), LNAI 11775, 2019, pp. 287-300. (**indexed Web of Science**)

Rank B, 4 points.

[CCCD19] **Liana Maria Crivei**, Gabriela Czibula, George Ciubotariu, Mariana Dindelegan. *Unsupervised learning based mining of academic data sets for students' performance analysis.* IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2020, pp. 11-16.

[1] hfpop.ro/standarde/doctorat/2014-jurnale.pdf
[2] https://hfpop.ro/standarde/doctorat/2014-conferinte.pdf

Rank C, 1 point.

Publications in journals and conference proceedings indexed international databases

[CIC19] **Liana Maria Crivei**, Vlad-Sebastian Ionescu, Gabriela Czibula. An analysis of supervised learning methods for predicting students' performance in academic environments. *ICIC Express Letters*, 13(3), 2019, pp. 181-189 (**indexed Scopus**)

Rank C, 2 points.

[Cri18a] **Liana Maria Crivei**. Incremental relational association rule mining of educational data sets. *Studia Universitatis Babes-Bolyai Series Informatica*, Vol. 63(2), 2018, pp. 102-117. (**indexed Mathematical Reviews**)

Rank D, 1 point.

[CC19] George Ciubotariu, **Liana-Maria Crivei**. *Analysing the academic performance of students using unsupervised data mining. Studia Universitatis Babes-Bolyai Series Informatica*, Vol. 64(2), 2019, pp. 34–48. (**indexed Mathematical Reviews**)

Rank D, 1 point.

[Cri18b] **Liana-Maria Crivei**. *Using unsupervised learning models for analyzing students' academic performance.* PhD Colloquium at 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2018, pp. 1-4.

Publications score: 21 points.

Introduction

The main research domain of our doctoral thesis is Educational Data Mining. *Educational Data Mining* (EDM) is an emerging research field focusing on the analysis of the *educational environments* from the *data mining* perspective. The major goal is to better comprehend the educational related phenomena and to uncover, through data mining techniques, relevant hidden patterns within educational data. *Data mining* (DM) techniques are extensively applied nowadays in various domains including medicine, bioinformatics, telecommunications, software engineering, financial data analysis to discover relevant patterns in large databases, especially due to their potential of uncovering hidden information from data.

Applying *machine learning* techniques in education [BCR18] is continuously attracting researchers from the *educational data mining* domain. The main focus in EDM is that of uncovering meaningful patterns from data that comes from various educational environments. One purpose of EDM is to offer additional insights into the students' learning process and thus to offer a better comprehension of the educational related phenomena. Extracting relevant patterns from the educational processes would also be useful for understanding students and how they learn, as well as improving the educational outcomes (e.g. learning outcomes). EDM has received lately considerable attention from the research community since extracting hidden knowledge from educational data is of particular interest for the academic institutions and also useful for improving their teaching methodologies and learning processes [MT13a].

Association rule (AR) mining represents an important data analysis and mining technique [CLCH16] useful in multiple *machine learning* tasks for revealing meaningful rule based patterns within various data sets. *Ordinal association rules* (OARs) [CcTM06] were proposed as a particular class of ARs which express ordinal relationships between the attributes characterizing a data set. *Relational association rules* (RARs) [CcM06, cSCC06] have been introduced as an extension of OARs able to express different types of non-ordinal relations between data attributes. Both ordinal and relational association rule mining have been successfully applied to address problems that range from data mining tasks (e.g. detecting software design defects [CMC14] or data cleaning [CcM06]) to supervised classification (e.g. software defect prediction [MCC18b]).

Motivation

Due to their ability of expressing relevant relationships in data, applying RARs on academic data sets could offer new and effective approaches for solving various EDM tasks. A concrete example in which RARs would be much more meaningful than classical ARs is related to the *Educational Data Mining* (EDM) domain which refers to mining data sets from educational environments. In a scenario where the data set to be mined contains information regarding the academic performance of students (e.g. each student is characterized by his grades), RARs would be able to express relevant relationships between the students' grades.

Thus, RARs may provide additional insights into the students' performance prediction task.

The EDM literature reveals that DM is very useful in the educational field particularly when exploring the online learning environment [MT13a]. In this respect, due to its practical relevance for *online* data mining, the problem of *incremental relational association rules mining* in the context of a dynamic data set, to which new instances are added is attractive for DM researchers. Since the learning processes within educational environments are by nature online processes, the idea of investigating the incremental RAR perspective on EDM comes naturally.

Due to its NP-completeness, the problem of mining all the interesting RARs within a data set is computationally difficult. As the dimensionality of the data set to be mined increases, the classical algorithm *Discovery of Relational Association Rules* (*DRAR*) for RARs mining fails in providing the set of rules in reasonable time. With the aim of reducing the overall mining time, we will investigate a *concurrent relational association rule* mining perspective using concurrency for the RARs discovery process.

Within the EDM research, developing machine learning models for analyzing and predicting the students' performance is of great interest. As a natural stage preceding the development of new machine learning models for students' performance prediction, the usefulness of applying *unsupervised machine learning* methods in analyzing students' academic performance data is investigated. As a subsequent step, due to their ability to uncover meaningful knowledge from data, RARs will be explored in the context of predicting the academic performance of students. More specifically, supervised *machine learning models* based on RARs are envisaged.

To summarize, the research conducted in the PhD thesis will be focused on three main directions: (1) conceptual contributions in the field of RAR mining by introducing new approaches to *incremental* and *concurrent* RAR mining; (2) investigating and highlighting the relevance and effectiveness of *relational association rule mining* on educational data sets; and (3) developing supervised and unsupervised *machine learning models* for students' academic performance prediction based on RARs and rule based systems. More specifically, the major goals of the research conducted in this thesis are: using relational association rule mining in the EDM field; developing new machine learning models (mainly based on RAR mining) for predicting students' academic performance; developing a concurrent version of the RAR mining algorithm in order to reduce the computational time of the mining process and investigating an incremental version of the RAR mining algorithm which would be appropriate in online mining scenarios where new instances are added to the data set to be mined.

Thesis structure

The rest of the thesis is structured as follows.

Chapter 1 presents the major background concepts regarding EDM and association rule mining, along with the problems we intend to approach from a computational perspective. The importance of EDM in the context of machine learning is emphasised in Section 1.1. We approach the concepts of supervised and unsupervised learning and present a comprehensive literature review on the existing research papers related to this field. A literature review on predicting instructors' performance and other EDM studies is also included. Section 1.2 presents the fundamental concepts of frequent itemsets and classical association rules, as well as a comprehensive description on relational association rules, its definition and a practical example. An overview of the existing literature on incremental and concurrent association

rule mining is also given in this section. The last part of this chapter, Section 1.3 introduces three fundamental machine learning models: artificial neural networks, random forests and self-organizing maps that will be subsequently used for the students' performance prediction task.

Chapter 2 introduces the original research on the incremental relational association rule mining in the context of dynamic data sets, as well as the concurrent RAR mining. In Section 2.1 a new incremental method named $IRARM$ (*Incremental Relational Association Rule Mining*) is introduced. Through $IRARM$ we intend to progressively adjust the interesting RARs discovered in an original data set which is subsequently enlarged with new incoming instances. Our objective is to acquire a computation time that is superior compared to that obtained through the $DRAR$ method applied from scratch on the extended data set. Furthermore the methodology used for $IRARM$ is presented, with the computation algorithm and a breakdown of the two stages named $filtering$ and *extending*. Experiments are performed on two case studies in order to evaluate and highlight $IRARM$ mining time efficiency. Section 2.2 introduces a new concurrent mining approach named $CRAR$ *Concurrent Relational Association Rule mining*. The new method employs concurrency for RARs discovery and considerably decreases the mining time. $CRAR$ performance from a computation time standpoint is evaluated and compared with respect to the sequential $DRAR$ algorithm. An exhaustive presentation of $CRAR$ algorithm is included together with the functions for *candidate generation*, *partitioning* and *rule generation* on multiple threads. The experimental evaluations are conducted upon nine open source data sets from various domains: software engineering, educational data mining and bioinformatics. The implementation of $CRAR$ on the previously mentioned data sets demonstrates that the algorithm achieves a significantly lower running time in comparison with $DRAR$ method applied from scratch.

Chapter 3 presents the original contributions and results in developing unsupervised learning models in the EDM context. Section 3.1 explores two unsupervised machine learning models, specifically $self\text{-}organizing\ maps$ and *relational association rule mining* in order to evaluate students performance in academic settings. The analysis is conducted on a real academic data set where the attributes of each instance represent the measures of student performance at a certain academic course. Our study reveals the potential of the two unsupervised ML techniques to identify relevant patterns and relationships within the academic data that are valuable for predicting students' academic achievements. In Section 3.2 the classical RAR mining and the $IRARM$ method formerly introduced in Section 2.1 are tested and evaluated in the context of educational data. Experiments are conducted considering two case studies and a total of four academic data sets. The study emphasises the importance and effectiveness of incremental RAR mining, but also the classical RAR mining in discovering meaningful knowledge and patterns within educational data.

Chapter 4 introduces our contributions in developing supervised learning models for students' performance prediction. Section 4.1 analyses supervised methods useful for predicting students performance in academic environments. Two regression models *Random forests*(RF) and *Artificial neural networks* (ANN) are investigated in terms of effectiveness with regard to the prediction task. The experiments are performed on three real academic data sets that contain students' grades obtained during the semester at various undergraduate courses. The prediction performance is measured using *root mean squared error* (RMSE) and *F-score*. A discussion upon the results obtained as well as a comparison to the related work is also presented. Section 4.2 introduces a new classification model, $SPRAR$ (*Students Performance prediction using Relational Association Rules*) for predicting, based on the grades received

during the semester, the final result of a student at a certain academic discipline using *relational association rules* (RARs). Experiments are performed on three real academic data sets. The performance of the $SPRAR$ classifier on the considered case studies is compared against the existing related work, being superior to previously proposed students' performance predictors. A comparative study on the previously introduced classification model $SPRAR$ is further performed in Section 4.3. Three new classification scores are introduced in this section and used in the classification stage of $SPRAR$. Their performance is analysed on the same real academic data sets used in the original research from Section 4.2 and compared to similar existing results from the literature.

We have to note the generality of the models introduced in this PhD thesis that are not only specific to the students performance prediction task. For instance $IRARM$ or $CRAR$ methods do not depend on the nature of the data set mined, nor on the type of relations defined and used in the mining process. Also, $SPRAR$ classifier can be easily extended and further adapted to other classication problems (binary or multi-class).

Original contributions

There are three major contributions of the thesis contained in Chapters 2, 3 and 4 as follows:

1. Conceptual contributions in the field of DM, specifically in the direction of *incremental* and *concurrent* RAR mining, introduced in Chapter 2 and published in the original papers [MCC18a] and [CCMC19].

    - An incremental mining method, called $IRARM$ *Incremental Relational Association Rule Mining*, is introduced in Section 2.1 for efficiently mining all the interesting RARs from a data set which has been increased (i.e. to which new entities have been added) [MCC18a]. $IRARM$ adapts the set of all interesting RARs previously discovered in the initial data set to compute the new set of all RARs which are interesting in the extended data set. $IRARM$ algorithm is experimentally evaluated on two open source data sets. The performed experiments emphasize that $IRARM$ is capable of providing the complete set of interesting RARs more efficiently than applying the mining algorithm from the start on the extended data set. The original work was published in [MCC18a].
    - A concurrent approach $CRAR$ *Concurrent Relational Association Rule mining* to RAR mining is introduced in Section 2.2. The effectiveness of mining is empirically validated on nine open source data sets. The reduction in the mining time when using $CRAR$ against $DRAR$ emphasizes that it can be successfully applied in various practical data mining scenarios. The original work was published in [CCMC19].

2. New unsupervised learning models in the EDM field, more specifically for *Students Performance Prediction* (SPP) are introduced in Chapter 3 and published in the original papers [Cri18a], [CC19] and [Cri18b].

    - *Unsupervised learning* models for analyzing students' academic performance data are proposed in Section 3.1. Experiments performed on a real academic data set

highlighted the potential of *unsupervised learning* models for uncovering meaningful knowledge within educational data. The original work was published in [Cri18b] and [CC19].

- We have investigated and highlighted the relevance and effectiveness of *relational association rule mining* on educational data sets in Section 3.2 and we applied the *incremental relational association rule mining* approach $IRARM$ previously introduced in Chapter 2 in the context of EDM, emphasizing its relevance and effectiveness. The original work was published in [Cri18a].

3. New supervised *machine learning models* for SPP, built on rule based systems and RARs are developed in Chapter 4 and were published in the original papers [CIC19], [CMC19] and [CAG19].

   - We conduct in Section 4.1 a study upon applying two regression models, *Random forests*(RF) and *Artificial neural networks* (ANN), for predicting the academic performance of students. For each regressor, two computational models (a regression one and a classification one) are introduced for predicting the final examination grade for a student based on his/her grades received during an academic semester. The original work was published in [CIC19].
   - We are introducing in Section 4.2 a new classification model, $SPRAR$ (*Students Performance prediction using Relational Association Rules*) for predicting, based on the grades received during the semester, the final result of a student at a certain academic discipline using *relational association rules* (RARs). Experiments are performed on three real academic data sets collected from Babeş-Bolyai University from Romania. The performance of the $SPRAR$ classifier on the considered case studies is compared against existing related work, being superior to previously proposed students' performance predictors. The original work was published in [CMC19].
   - A comparative study on the previously introduced classification model $SPRAR$ is further performed in Section 4.3. Three new classification scores are introduced in this section and used in the classification stage of $SPRAR$. Their performance is analyzed on the same real academic data sets used in the original paper [CMC19] and compared to similar existing results from the literature. The original work was published in [CAG19].

Chapter 1

Background

In this chapter we present the fundamental background concepts related to the *Educational data mining* (EDM) field as well as the main concepts related to the computational models we target in our research.

1.1 Educational data mining (EDM)

Data mining (DM) techniques are considerably applied nowadays in multiple fields including medicine, software engineering, bioinformatics, to uncover relevant patterns in large databases, particularly due to their potential of uncovering hidden information within data sets.

Applying DM techniques in education [BCR18] has attracted researchers from both DM and educational research and thus a new interdisciplinary research discipline known as *educational data mining* (EDM) emerged. The main focus in EDM is to develop methods for extracting knowledge from data that come from various academic environments and educational information systems. Through mining educational data sets, EDM's purpose is to better understand the students' learning process and thus to provide additional insights into educational related phenomena.

EDM is a challenging research field where the underlying concept is that of revealing the *data mining* perspective within the educational context. The main goal is to better comprehend the educational related processes by extracting meaningful hidden patterns from educational data using data mining techniques.

Extracting relevant patterns from educational data proves to be effective for understanding students and their learning practices, as well as improving their learning outcomes. EDM has recently received significant attention from the scientific community since extracting hidden knowledge from educational data is particularly valuable for the academic institutions and also beneficial for improving their teaching techniques and methodologies or the learning processes [MT13a].

1.2 Background on association rule mining

In this section we present the fundamental concepts related to *relational association rule* (RAR) mining [cSCC06]. *Association rule* (AR) mining represents an important data analysis and mining technique [CLCH16] applied in various *machine learning* tasks for uncovering relevant rule based patterns in data sets. *Ordinal association rules* (OARs) [CcTM06] were

proposed as a particular class of ARs which express ordinal relationships between the attributes characterizing a data set. *Relational association rules* (RARs) [CcM06, cSCC06] have been introduced as an extension of OARs able to express different types of non-ordinal relations between data attributes.

Classical *association rules* (ARs) [SDL+15] do not consider the relations that may occur between the attribute values, but only their co-occurrences. Unlike these, *ordinal association rules* (OARs) [MML01] express ordinal relations that commonly appear in data. But informative relations that are not ordinal may also exist between the attribute values.

The *Relational Association Rules* (*RARs*) notion is defined in the following paragraphs.

We consider $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ a set of *instances* or *records*. Let $\Omega = (a_1, \ldots, a_m)$ be a sequence of m attributes characterizing each instance from the data set $\mathcal{D}$. Each attribute a_i takes values from a non-empty and non-fuzzy domain Δ_i, which also contains a *null* (*empty*) value. We denote by $\Psi(d_j, a_i)$ the value of attribute a_i for an instance d_j.

We denote by $\mathcal{T}$ the set of all possible relations that are not necessarily ordinal which can be defined between two domains Δ_i and Δ_j.

Definition 1. A *relational association rule [cSCC06]* is an expression $(a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}) \Rightarrow (a_{i_1}\tau_1 a_{i_2}\tau_2 a_{i_3} \ldots \tau_{h-1} a_{i_h})$, where $\{a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}\} \subseteq \Omega$, $a_{i_k} \neq a_{i_p}$, $k, p = 1, \ldots, h$, $k \neq p$ and $\tau_k \in \mathcal{T}$ is a relation over $\Delta_{i_k} \times \Delta_{i_{k+1}}$, Δ_{i_k} being the domain of the attribute a_{i_k}.

a) If $a_{i_1}, a_{i_2}, \ldots, a_{i_h}$ are non-missing in m instances from the data set then we call $s = \frac{m}{n}$ the *support* of the rule.

b) If we denote by $D' \subseteq \mathcal{D}$ the set of instances where $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}$ are non-missing and all the relations

$$\Psi(d_j, a_{i_1})\tau_1\Psi(d_j, a_{i_2}), \Psi(d_j, a_{i_2})\tau_2\Psi(d_j, a_{i_3}), \ldots, \Psi(d_j, a_{i_{h-1}})\tau_{h-1}\Psi(d_j, a_{i_h})$$

hold for each instance d from D' then we call $c = \frac{|D'|}{n}$ the *confidence* of the rule.

Interesting RAR's were defined in [cSCC06] as those rules which have both their *support* and *confidence* greater than or equal to specified minimum thresholds. For mining interesting RARs an Apriori-like algorithm named *DRAR* (Discovery of Relational Association Rules) was proposed in [CBC12] as an extension of the *DOAR* algorithm introduced in [CcTM06] for uncovering OARs.

The *DRAR* algorithm consists of length-level generation of RARs, starting with the rules of length 2. The set of 2-length rules is filtered for determining the interesting rules (i.e those rules which verify the minimum support and confidence requirements). At a given step, for determining the RARs of a certain length l, we start from the set of interesting RARs of length l-1 generated at the previous step. This set is used to generate by join new possible interesting RARs, called candidate rules which will be filtered to preserve only the rules which are interesting. After the set of l-length interesting RARs is generated, the iterative process continues with generating the rules of length l+1. The process stops when no new interesting RARs are discovered.

1.3 Machine learning models used

This sections presents the supervised and unsupervised learning models used in the thesis for predicting the students' academic performance: *artificial neural networks*, *random forests* and *self-organizing maps*.

1.3.1 Artificial neural networks

Artificial neural networks are widely used as supervised learning models for various applications such as pattern recognition, prediction [LB05], speech recognition [SH07], system identification and control. Similarly to the biological neural systems, the artificial neural networks [Mit97] consist of a densely interconnected set of computational units, called *neurons*.

An *artificial neural network* (ANN) [YAF+17] is considered an adaptive system which learns a mapping (that is an input/output function) from the training data, by autonomously adjusting within the *training phase* the parameters of the system. The ANNs parameters obtained after the training was completed are further used in the *testing phase* to solve the problem in question.

In a supervised learning setting, an input instance is given to the ANN including the corresponding target output [MCC18c]. An error is represented by the difference between the target output (which is the desired output) and the system output. The error data are reintroduced into the network and in this way the system parameters are modified and adjusted through the *learning rule*. The ANN is built using an iterative procedure whose goal is to minimize the obtained error, which is commonly referred to as the *learning rule*. This process is repeated until an admissible performance is attained. The algorithm for building the ANN model is called the *backpropagation* algorithm.

1.3.2 Random forests

Random forests (RFs) [Bre01] are an ensemble learning method consisting of combinations of several tree predictors using bootstrap aggregation. During the building process of each of the individual tree, only a random subset of features and an arbitrary subset of training examples are considered for analysis. In this way, overfitting is avoided and better stability is achieved for generalization. As the trees' number within the forest increases, the generalization error for forests will converge to a limit, thanks to the law of large numbers. This error is influenced by the correlation between the individual trees within the forest as well as the strength of the individual trees. Using an arbitrary feature selection in order to split each node returns error rates which are similar to Adaboost [Wan12]. Being built on decision trees, random forests can be used in classification and regression problems.

1.3.3 Self-organizing maps

Self-organizing maps (SOMs) are unsupervised learning models, also known in the literature as *Kohonen maps*, were introduced by Teuvo Kohonen and are widely used as tools for visualizing high-dimensional data. SOMs are connected to the *artificial neural networks* (ANNs) in literature and to *competitive learning*. In *competitive learning*, the output neurons compete themselves to be activated. A *self-organizing map* [SK99] is trained using an unsupervised learning algorithm (Kohonen's algorithm) to map, using a non-linear mapping, the continuous input space of high-dimensional instances into a discrete (usually two-dimensional) output space called a *map* [EDB08]. The *topology preserving mapping* is the main characteristic of the mapping provided by a SOM. This means that the input samples which are nearby in the input space will be also close to each other on the map (output space). Thus, a SOM is able to provide clusters of similar data instances [LO92].

Self organizing maps are a class of ANNs that employ a neighborhood function in order to maintain the topological characteristics of the input space. SOMs are using unsupervised

learning for training and they are one of the most innovative types of artificial neural networks in the current specialized literature. SOMs algorithm uses unsupervised, competitive learning where nodes compete against each other in order to respond to a subset of incoming data. It maintains the topology of the high-dimensional space which is mapped into a low dimension, usually a two-dimensional space referred to as a map. The dimensionality reduction through topology preservation means that neighboring or similar input data points will be mapped in the nearby units of the output space [KOS09]. Therefore SOMs can be viewed as an instrument to visualize high-dimensional data as well as a clustering tool.

Chapter 2

New approaches in Data Mining

This chapter introduces the contributions of our PhD thesis in the *data mining* domain, more specifically in developing new *incremental* and *concurrent* relational association rule mining approaches.

2.1 A new incremental relational association rule mining approach

First, a new incremental method named *IRARM* (*Incremental Relational Association Rule Mining*) is introduced and evaluated in terms of running time efficiency. *IRARM* aims to identify the interesting RARs in a dataset which is dynamic by nature, meaning that new instances are incrementally added to the initial dataset. The characteristic of *IRARM* is that it adapts the relevant RARs previously discovered without requiring a re-scan of the entire updated dataset, as oposed to *DRAR* method. The effectiveness of *IRARM* is evaluated through experiments conducted on two open source data sets and the results highlight *IRARM*'s capacity to significantly improve the mining time as compared to *DRAR*.

For future work, more data sets will be considered, in order to further extend the evaluation of *IRARM*. We also aim to investigate the exact conditions in which it is more efficient to apply the incremental *IRARM* method than to run the offline *DRAR* algorithm from scratch. Furthermore, we aim to use IRARM in concrete dynamic data mining tasks, such as incremental *software defect prediction* and incremental mining of educational data (*educational data mining* [MT13b]). In addition, we plan to investigate an adaptive incremental approach to *relational association rule* mining, in which both new features and new objects are added to the mined data set. For improving the efficiency of the mining process, a distributed approach will be also envisioned.

2.2 A novel concurrent relational association rule mining approach

The second section from this chapter introduces a new concurrent mining approach named *CRAR* (*Concurrent Relational Association Rule mining*). Concurrent programming is frequently used in order to enhance performance, resource exploitation and also to develop systems which are fault-tolerant and scalable. *CRAR* algorithm was developed on these principles with the aim of acquiring a superior mining time in the RAR discovery process.

We perform a comprehensive evaluation on $CRAR$ by conducting experiments on nine open source data sets from various domains: educational data mining, software defect prediction, bioinformatics. $CRAR$ provides a significant time improvement by reducing the mining time with an average of 52,3% in comparison to the classical $DRAR$ algorithm. The approaches presented in the current chapter are original research works published in [MCC18a] and [CCMC19].

For future work, we aim to investigate optimizations of $CRAR$ such as increasing $CRAR$'s amount of parallel work, particularly the computation of binary RARs, which would increase the algorithms' *scaling efficiency* and *parallel speedup*. For better highlighting the usefulness of $CRAR$, we plan to use relational association rule mining for concrete supervised classification problems from *search-based software engineering* (e.g. *software defect prediction* [MCC18c]) and *educational data mining* [Cri18a]) (such as *students' performance prediction*). Further work will also be focused on extending the $CRAR$ approach introduced in this paper for mining *gradual relational association rules* [CCM17] instead of the classical (non-fuzzy or non-gradual) RARs.

Chapter 3

New unsupervised learning models in Educational Data mining

We introduce in this chapter our contributions in developing unsupervised learning models for analyzing students' academic performance data. Section 3.1 evaluates self-organizing maps and relational association rule mining models with regard to their potential of detecting relevant patterns in students' academic data sets, more precisely students performance reflected in their grades obtained during a semester. Section 3.2 investigates *IRARM* (*Incremental relational association rule mining*) algorithm in the context of educational data. Experiments are performed on four data sets and highlight the potential and relevance of incremental RAR mining in uncovering meaningful knowledge in academic data.
The approaches presented in the current chapter are original research works published in [Cri18a] [CC19] and [Cri18b].

3.1 Using unsupervised learning models for analyzing students' academic performance

In this section we investigate the usefulness of two *unsupervised machine learning* techniques (*self-organizing maps* and *relational association rule mining*) in analysing students' academic performance, with the broader goal of developing new machine learning models for students' performance prediction. The experimental results obtained by applying the aforementioned unsupervised learning models on a real data set collected from Babeş-Bolyai University emphasize their effectiveness in mining relevant relationships and rules from educational data which may be useful for predicting the academic performance of students.

The approach from this section was introduced in our original papers [Cri18b] and [CC19].

RARs have not been used in the literature, so far, for analyzing academic data sets. To the best of our knowledge, a similar study has not been performed in the literature. The study initiated in this paper represented the starting point for our further research conducted for developing new *machine learning models* for students' academic performance prediction.

This section examined two unsupervised learning models, *self-organizing maps* and *relational association rule mining*, in the context of analysing data sets related to students' academic performance. Experiments performed on a real data set collected from Babeş-Bolyai University, Romania highlighted the potential of unsupervised learning based data mining tools to detect meaningful patterns regarding the academic performance of students. The current study represents the starting point of the research further conducted in order

to obtain a better comprehension of the students' learning processes and to develop new machine learning methods for predicting students' academic performance.

Future work will be carried out in order to extend the experimental evaluation on other academic data sets and to interpret the mined interesting RARs. For increasing the performance of the unsupervised learning process, methods for detecting anomalies and outliers in data will be further investigated. In addition, a post-processing phase for filtering the set of mined RARs will be analysed for removing rules which overlap with multiple classes. In addition, a classification model based on RARs will be envisioned for predicting students' academic performance. Other *classification* and *regression* machine learning techniques for students performance prediction (e.g. random forests) will be investigated, as well.

3.2 Incremental relational association rule mining of educational data sets

Incremental Relational Association Rule Mining (*IRARM*) has been previously introduced as an effective *online data mining* method for dynamically mining interesting *relational association rules* (RARs) in a dynamic data set which is extended with new data instances. The study conducted in this section was introduced in the original paper [Cri18a] and is aimed to emphasize the effectiveness of both RAR and *IRARM* mining methods in *educational data mining* settings. Experiments performed on various academic data sets highlight the potential of using *relational association rules* for uncovering relevant knowledge from educational related data.

We are approaching in this section the problem of *incremental relational association rule mining* (*IRARM*) in the context of EDM. The process of incremental RAR mining is appropriate specifically for *online* DM scenarios, where the data set to be mined is dynamic and thus continuously extended with real-time arriving data streams. In such situations, *IRARM* approach aims to progressively adapt the interesting RARs identified in a data set, when it is enlarged with new instances. Since the learning processes within educational environments are by nature online processes, the idea of investigating the *IRARM* perspective in EDM comes naturally. The EDM literature also reveals that DM is very useful in the educational field particularly when exploring the online learning environment [MT13a].

The contribution of the section is summarized as follows. First, we are emphasizing the relevance of RAR mining in the field of *educational data mining* (EDM) with the goal of uncovering meaningful patterns within educational data sets. Secondly, we extend the experimental evaluation of our previously proposed *incremental relational association rule* mining approach (*IRARM*) [MCC18a] on several EDM case studies. The effectiveness of *IRARM* is emphasized through the reduction in mining time achieved when using *IRARM* against RAR mining from scratch when a data set is extended with new instances. The study conducted in this paper is novel in the EDM literature, since neither the classical nor the incremental RAR mining approaches have been applied on academic data sets, so far.

We investigated in this section the application of classical and incremental RAR mining for knowledge discovery in data sets from educational environments, with the goal of uncovering meaningful patterns within educational data sets. The relevance of uncovering RARs in academic data sets has been emphasized in the context of the students' learning process, as offering additional insights into educational related phenomena. Additionally, the effectiveness of *incremental* RAR mining in online EDM scenarios was highlighted through

several case studies.

Future work will be done in order to extend the experimental evaluation of *IRARM* on other EDM tasks, to further test its performance. An *incremental adaptive* RAR mining will be also investigated for academic data sets, when both new instances and new features are added to the data set. Furthermore, we plan to apply RAR, gradual RARs [CCM17] and *IRARM* mining algorithm in supervised learning EDM scenarios, such as predicting students' academic performance.

Chapter 4

New supervised learning models for students' performance prediction

In this chapter we introduce our contributions in developing supervised learning models for students' performance prediction.

We perform in Section 4.1, research work [CIC19], an analysis of supervised learning methods for predicting students' performance in academic environments. The experiments are conducted on three real academic data sets and investigate the effectiveness of two regression models *Random forests* (RF) and *Artificial neural networks* (ANN) in predicting the academic performance of students. A new classification model named $SPRAR$ (*Students Performance prediction using Relational Association Rules*) is introduced in research paper [CMC19] and presented in Section 4.2. $SPRAR$ is used for predicting the final results obtained by students at a certain academic course, based on their grades received during the semester. The evaluation is conducted on three real academic data sets collected from Babeş-Bolyai University from Romania. $SPRAR$ classifier is compared in terms of performance with the existing supervised classifiers from the related work. The experiments reveal that $SPRAR$ outperforms the previously proposed students' performance predictors in a vast majority of cases. Section 4.3 introduces three new classification scores for the $SPRAR$ model that was formerly proposed. Their performance is evaluated against the same real academic data sets selected for the original paper [CMC19]. The best performing score is compared to the original $SPRAR$ and also to related results from the current literature. The original work was published in [CAG19].

The approaches presented in the current chapter are original research works published in [CIC19], [CMC19] and [CAG19].

4.1 An analysis of supervised learning methods for predicting students' performance in academic environments

This section analyzes the performance of supervised classification and regression models for predicting the academic performance of students. Experiments are conducted on three real data sets collected from Babeş-Bolyai University, Romania. We investigate the effectiveness of different supervised machine learning models and aim to predict the students' final grade at a certain academic discipline based on their performance during the semester. The presentation from this section is based on the original paper [CIC19].

We conduct in this section a study upon applying two regression models, *Random forests*

(RF) and *Artificial neural networks* (ANN), for predicting the academic performance of students. For each regressor, two computational models (a regression one and a classification one) are introduced for predicting the final examination grade for a student based on his/her grades received during an academic semester. The results obtained on three real data sets collected from Babeş-Bolyai University, Romania reveal that supervised regressors are helpful for identifying means to increase the quality of the educational processes. To the best of our knowledge, a study similar to ours has not been conducted in the EDM literature so far. In addition, RF classifiers have been applied in the literature, but in other tasks than the one considered in this paper.

To summarize, the purpose of the study conducted in this section is to answer the following research questions:

RQ1 What is the potential of supervised regression models to predict the final examination grade of students based on the grades they received during the semester?

RQ2 To what extent can feature selection improve the accuracy of predicting the academic performance of students?

RQ3 How do the supervised learning models used in this paper compare to other related work from the literature in terms of performance prediction?

The study from this section was conducted with the aim to highlight the effectiveness of supervised regression models in predicting the academic performance of students.

The experiments conducted on three data sets containing real academic data collected from a Romanian University highlighted that generally RF are the best regression model for predicting the final examination grade of students, based on the grades received during an academic semester. For experiment E1 we observed that the SVM regressor provides slightly better results than RF. The study also revealed the difficulty of the students' performance prediction task and the importance of increasing the number of features used in the learning process.

Future work will be made in order to add more features to our learning tasks as well as to apply preprocessing techniques for detecting outliers in data. We also aim to investigate the use of other learning models for predicting students' final performances, such as a classifier based on *relational association rule mining*.

4.2 *SPRAR*: A relational association rule mining classification model applied for academic performance prediction

This section approaches the problem of predicting the academic performance of students, a problem intensively investigated within the Educational Data Mining (EDM) literature. For a better understanding of the educational related phenomena, there is a continuous interest in applying *supervised* and *unsupervised* learning methods for obtaining additional insights into the students' learning process. The problem of predicting if a student will pass or fail at a certain academic discipline based on the students' grades received during the semester is a difficult one, highly dependent on various factors such as the course, the number of examinations during the semester, the instructors and their exigences. We propose a new classification model, *SPRAR* (*Students Performance prediction using Relational Association Rules*) for predicting the final result of a student at a certain academic discipline using

relational association rules (RARs). *SPRAR* approach was introduced in the original paper [CMC19]. Experiments are performed on three real academic data sets collected from Babeş-Bolyai University from Romania. The performance of the *SPRAR* classifier on the considered case studies is compared against existing related work, being superior to previously proposed students' performance predictors.

This section approaches the problem of predicting if a student will pass or fail at a certain academic discipline based on the students' grades received during the semester. The problem has a major practical relevance in educational environments, since it could provide relevant feedback to the students which are likely to fail at a certain academic course. Having such an advice during the semester, the students will have the possibility to prevent a possible academic unsuccess. Despite its obvious pragmatic importance, the problem is a complex and difficult one, highly dependent on various conditions. We propose a new classification model, *SPRAR* (*Students Performance prediction using Relational Association Rules*) for predicting the final result of a student at a certain academic discipline using *relational association rules* (RARs). We have to note *SPRAR* classifier's generality, that is not specific to the students' performance prediction task. *SPRAR* can be easily adapted to other classification problems.

The classification model introduced in the following is a binary one (i.e. there are only two classes to predict: *pass* or *fail*), but the proposed model is a general one, it can be extended for a multi-class classification problem (i.e. to predict the final grade of the student). From a supervised learning perspective, the problem of predicting the successful completion of a course during an academic semester is a hard problem, particularly due to the imbalanced nature of the training data (i.e. the number of students which *passed* the exam is generally much higher than the number of students which *failed* the exam). Experiments conducted on three real data sets from Babeş-Bolyai University, Romania highlight the effectiveness of the classification using *SPRAR*. The literature regarding students' performance prediction reveals that the use of RARs in predicting the academic performance of students is a novel approach.

In summary, the aim of the research further conducted is that of proposing a classification model *SPARAR* based on RAR mining for predicting the students' academic performance, as a proof of concept. With this aim the proof of concept considers only three medium sized data sets for highlighting that *SPRAR* is suitable for the approached problem. If this stands, the study of applying *SPRAR* for students' performance prediction can be further extended on a larger scale.

Future work will be carried out in order to extend the experimental evaluation of *SPRAR* on other case studies. In addition, we aim to generalize the *SPRAR* binary classifier to a multi-class classifier such that to predict the students' final examination grade at a certain academic course. Regarding the relational association rules discovery process, we plan to extend our model considering *gradual relational association* rules [CCM17]. Alternative measures for defining the probabilities p_+ and p_- will be further investigated, as well as the idea of using an ensemble of *SPRAR* classifier for improving the predictive performance. *SPRAR* may also be extended for handling n-ary relations between the attributes' values.

4.3 A study on students' performance prediction using *SPRAR*

This section analyses the classification model SPRAR (Students Performance prediction using Relational Association Rules) introduced in [CMC19], Section 4.2, for predicting the academic results of students using relational association rules. Three new classification scores

are introduced in this chapter and used in the classification stage of SPRAR. Their performance is analyzed using three real academic data sets from Babes-Bolyai University, Romania and compared to similar existing results from the literature.

The material presented in this section is original work pusblished in [CAG19]. In this paper we extended the $SPRAR$ classification model, by introducing three alternative scores which can be used in the classification stage of $SPRAR$ for deciding if a student will *pass* or *fail* a certain course. Three real case studies were used for experimenting the performance of $SPRAR$ considering the newly proposed classification scores. With the goal of empirically determining which is the the most appropriate classification score, the proposed scores are comparatively analyzed and evaluated against the original variant introduced in [CMC19]. The study performed in this section is new in the *educational data mining* literature.

This section extended the classification model SPRAR (Students Performance prediction using Relational Association Rules) [CMC19] for predicting students achievement to a particular course based on *relational association rule* (RAR) mining [cSCC06]. The membership of a query instance to a pass or fail class is determined using academic data sets obtained from Babes-Bolyai University, Romania. Three new classification scores are introduced and used in the classification stage of SPRAR for deciding if a student will *pass* or *fail*. SPRAR performance is evaluated using several evaluation measures: *sensitivity*, *specificity* and *Area under the ROC curve*. The best prediction score is identified and compared to existing results from the current literature.

Through several experiments performed on real academic data, we obtained an empirical evidence that relational association rules are relevant for distinguishing between the academic performance of students. We also observed that the classification methodology (identifying the best score for discriminating between students' performance) is dependent on the data sets, therefore we could further investigate the possibility to automatically learn it using supervised methods. As future research directions we also aim to extend our experiments on other data sets, to add additional attributes and relationships in the RAR mining process for improving SPRAR's predictive performance.

Conclusions

The current thesis emphasized our main contributions in developing and applying data mining models for solving problems related to the educational field. Educational Data Mining (EDM) is an emerging research field focusing on developing models to identify and extract valuable information from educational related data. Research in EDM aims to reveal patterns in students' learning and enhance their academic achievements. Students academic success and performance can be predicted through data mining techniques [Hue13] using *supervised* or *unsupervised* learning.

Relational association rules (RARs) [CcM06, cSCC06] are able to express different types of non-ordinal relations between data attributes. Both ordinal and relational association rule mining have been successfully applied to address problems that vary from data mining tasks (e.g. detecting software design defects [CMC14] or data cleaning [CcM06]) to supervised classification (e.g. software defect prediction [MCC18b]).

We introduced a new Incremental Relational Association Rules mining method, named $IRARM$, that efficiently mines all interesting RARs from a dynamic data set. Since the RAR mining process is computationally expensive, we have also focused on developing a concurrent version of the RAR mining algorithm, $CRAR$, with the major goal of reducing the computational time of the mining process.

Unsupervised and supervised learning models were also investigated in the thesis in the context of EDM: incremental RAR mining and self organizing maps for students performance data analysis. The experiments performed on several real educational data sets highlighted the potential of unsupervised learning methods in uncovering meaningful patterns in students academic data. As a proof of concept, we introduced a novel classification model $SPRAR$ [CMC19] based on *relational association rule* discovery for predicting the successful completion of an academic course, based on the grades received by students during the academic semester. Three classification scores were introduced and used in the classification stage of SPRAR for deciding if a student will *pass* or *fail* [CAG19].

Regarding our conceptual contributions in the DM field, we aim to investigate the exact conditions in which it is more efficient to apply the incremental $IRARM$ method than to run the offline $DRAR$ algorithm from scratch. In addition, we plan to investigate an adaptive incremental approach to *relational association rule* mining, in which both new features and new objects are added to the mined data set. Further optimizations of the concurrent RAR mining method ($CRAR$) are envisioned, such as: increasing $CRAR$'s amount of parallel work, particularly the computation of binary RARs, which would increase the algorithms' *scaling efficiency* and *parallel speedup*.

Regarding the $SPRAR$ classifier for *students' performance prediction*, future work will be carried out in order to extend the experimental evaluation of $SPRAR$ on other case studies. In addition, we aim to generalize the $SPRAR$ binary classifier to a multi-class classifier such that to predict the students' final examination grade at a certain academic course. Regarding

the relational association rules discovery process, we plan to extend our model considering *gradual relational association* rules [CCM17]. $SPRAR$ may also be extended for handling n-ary relations between the attributes' values.

Bibliography

[BCR18] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[Bre01] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[CAG19] Liana-Maria Crivei, Mihai Andrei, and Czibula Gabriela. A study on applying relational association rule mining based classification for predicting the academic performance of students. In *KSEM 2019 : The 12th International Conference on Knowledge Science, Engineering and Management, LNAI 11775*, pages 287–300, 2019.

[CBC12] Gabriela Czibula, Maria-Iuliana Bocicor, and Istvan Gergely Czibula. Promoter sequences prediction using relational association rule mining. *Evolutionary Bioinformatics*, 8:181–196, 04 2012.

[CC19] George Ciubotariu and Liana Maria Crivei. Analysing the academic performance of students using unsupervised data mining. *Studia Universitatis Babes-Bolyai Series Informatica*, pages 1–14, 2019.

[CCCD19] Liana Maria Crivei, Gabriela Czibula, George Ciubotariu, and Mariana Dindelegan. Unsupervised learning based mining of academic data sets for students' performance analysis. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2019)*, pages 1–6. IEEE Hungary Section, 2019.

[CcM06] Alina Câmpan, Gabriela Şerban, and Andrian Marcus. Relational association rules and error detection. *Studia Universitatis Babes-Bolyai Informatica*, LI(1):31–36, 2006.

[CCM17] Gabriela Czibula, Istvan Gergely Czibula, and Diana-Lucia Miholca. Enhancing relational association rules with gradualness. *International Journal of Innovative Computing, Communication and Control*, 13:289–305, 2017.

[CCMC19] G. Czibula, I.G. Czibula, D.-L. Miholca, and L.M. Crivei. A novel concurrent relational association rule mining approach. *Expert Systems with Applications*, 125(6):142–156, 2019.

[CcTM06] Alina Campan, Gabriela Şerban, Traian Marius Truta, and Andrian Marcus. An algorithm for the discovery of arbitrary length ordinal association rules. In *DMIN*, pages 107–113, 2006.

[CIC19] L.M. Crivei, V.-S. Ionescu, and G. Czibula. An analysis of supervised learning methods for predicting students' performance in academic environments. *ICIC Express Letters*, 13(3):181–190, 2019.

[CLCH16] H. Y. Chang, J. C. Lin, M. L. Cheng, and S. C. Huang. A novel incremental data mining algorithm based on fp-growth for big data. *2016 International Conference on Networking and Network Applications (NaNA)*, pages 375–378, 2016.

[CMC14] Gabriela Czibula, Zsuzsanna Marian, and István Gergely Czibula. Detecting software design defects using relational association rule mining. *Knowledge and Information Systems*, 2014.

[CMC19] G. Czibula, A. Mihai, and L.M. Crivei. A novel relational association rule mining classification model applied for academic performance prediction. volume 159, pages 20–29, 2019.

[Cri18a] Liana Maria Crivei. Incremental relational association rule mining of educational data sets. *Studia Universitatis Babes-Bolyai Series Informatica*, 63(2):102–117, 2018.

[Cri18b] Liana Maria Crivei. Using unsupervised learning models for analyzing students' academic performance. In *PhD Coloquium at SYNASC'18, Symbolic and Numeric Algorithms for Scientific Computing*, pages 1–4, 2018.

[cSCC06] Gabriela ¸Serban, Alina Câmpan, and Istvan Gergely Czibula. A programming interface for finding relational association rules. *International Journal of Computers, Communications & Control*, I(S.):439–444, June 2006.

[EDB08] N. Elfelly, J.-Y. Dieulot, and P. Borne. A neural approach of multimodel representation of complex processes. *International Journal of Computers, Communications & Control*, III(2):149–160, 2008.

[Hue13] Richard A Huebner. A survey of educational data-mining research. *Research in Higher Education Journal*, 19:1–13, 2013.

[KOS09] Andreas Khler, Matthias Ohrnberger, and Frank Scherbaum. Unsupervised feature selection and general pattern discovery using self-organizing maps for gaining insights into the nature of seismic wavefields. *Computers Geosciences*, 35(9):1757 – 1767, 2009.

[LB05] Kristopher R. Linstrom and A. John Boye. A neural network prediction model for a psychiatric application. *International Conference on Computational Intelligence and Multimedia Applications*, 0:36–40, 2005.

[LO92] J. Lampinen and E. Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2(3):261–272, 1992.

[MCC18a] Diana-Lucia Miholca, Gabriela Czibula, and Liana Maria Crivei. A new incremental relational association rules mining approach. In *22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, volume 126 of *KES2018*, pages 19–28. Procedia Computer Science, 2018.

[MCC18b] Diana-Lucia Miholca, Gabriela Czibula, and István Gergely Czibula. A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. *Inf. Sci.*, 441:152–170, 2018.

[MCC18c] Diana-Lucia Miholca, Gabriela Czibula, and Istvan Gergely Czibula. A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. *Information Sciences*, 441:152 – 170, 2018.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[MML01] Andrian Marcus, Jonathan I. Maletic, and King-Ip Lin. Ordinal association rules for error identification in data sets. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 589–591, New York, NY, USA, 2001. ACM.

[MT13a] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, 2013. The 9th International Conference on Cognitive Science.

[MT13b] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97(Supplement C):320 – 324, 2013. The 9th International Conference on Cognitive Science.

[SDL+15] Anping Song, Xuehai Ding, Mingbo Li, Wei Cao, and Ke Pu. A novel binary bat algorithm for association rules mining. *ICIC Express Letters*, 9(9):2387–2394, September 2015.

[SH07] Mark D. Skowronski and John G. Harris. Automatic speech recognition using a predictive echo state network classifier. *Neural Networks*, 20(3):414–423, April 2007.

[SK99] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10:151–159, 1999.

[Wan12] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800 – 807, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

[YAF+17] Tiantian Yang, Ata Akabri Asanjan, Mohammad Faridzad, Negin Hayatbini, Xiaogang Gao, and Soroosh Sorooshian. An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. *Information Sciences*, 418-419(Supplement C):302 – 316, 2017.