FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 16, Nº 2, 2017, pp. 95 - 116
DOI: 10.22190/FUACR1702095D

COMPARATIVE STUDY: FEATURE SELECTION METHODS IN THE BLENDED LEARNING ENVIRONMENT

UDC 372.881.111.1:(004.4+004.65)

Gabrijela Dimić[1], Dejan Rančić[2], Ivan Milentijević[2], Petar Spalević[3], Katarina Plećić[4]

[1] High School of Electrical Engineering and Computer Science, Belgrade, Republic of Serbia
[2]University of Niš, Faculty of Electronic Engineering, Department of Computer Science, Niš, Republic of Serbia
[3]University of Priština, Faculty of Technical Science, Republic of Serbia
[4]University of Singidunum, Belgrade, Republic of Serbia

Abstract. *Research presented in this paper deals with the unknown behavior pattern of students in the blended learning environment. In order to improve prediction accuracy it was necessary to determine the methodology for students` activities assessments. The Training set was created by combining distributed sources – Moodle database and traditional learning process. The methodology emphasizes data mining preprocessing phase: transformation and features selection. Information gain, Symmetrical Uncert Feature Eval, RelieF, Correlation based Feature Selection, Wrapper Subset Evaluation, Classifier Subset Evaluator features selection methods were implemented to find the most relevant subset. Statistical dependence was determined by calculating mutual information measure. Naïve Bayes, Aggregating One-Dependence Estimators, Decision tree and Support Vector Machines classifiers have been trained for subsets with different cardinality. Models were evaluated with comparative analysis of statistical parameters and time required to build them. We have concluded that the RelieF, Wrapper Subset Evaluation and mutual information present the most convenient features selection methods for blended learning environment. The major contribution of the presented research is selecting the optimal low-cardinal subset of students' activities and a significant prediction accuracy improvement in blended learning environment.*

Key words: *Blended Learning, Educational Data Mining, Features Selection, Mutual information*

Received May 11, 2017
Corresponding author: Gabrijela Dimić
High School of Electrical Engineering and Computer Science,Belgrade, Republic of Serbia
E-mail: gdimic@gmail.com

1. INTRODUCTION

Application of learning management systems in the educational process enabled storage and created a great vault of education data. The Learning Management System (LMS) represents an education platform which contains learning resources, software that enables creation, administration of one or multiple courses for learning in the classroom or in virtual environment. Three-layered architecture of such systems includes a database where activities and actions of the users are stored. On the basis of records in the database, the LMS systems can generate various reports. Sets of features reports are generated by system itself, but the extraction of useful and usable patterns of students' behavior and monitoring of the learning process for a particular course is a quite time-consuming task. Over the past few years, researches in the domain of education use data mining methodology and machine learning for analysis, findings of interesting recurrences, patterns and concepts in educational data. DM (sometimes called knowledge discovery) is defined as the process of extraction of new, potentially useful, interesting, understandable information contained in large databases, with an aim to facilitate proper business decision making [1],[2]. The analytical methods used in most of the cases are mathematical techniques and algorythms derived from statistics, machine learning and databases [3]. Implementation of the DM methodology in education caused new exploration field known as Educational Data Mining (EDM) that deals with the issues in developing methods for extrication of knowledge from data in educational environment [4]. The EDM process transforms row data from educational systems into useful information that could have great influence to educational research and practice. EDM uses the typical DM techniques of classification, clustering, association rules, sequential patterns, text mining as well as new methodologies such as discover of knowledge with models and integration with psychometrically modeled environment [5]. Classification is one of the most studied issues in the field of data mining. The main task of classification is prediction of class variable on the basis of the value of input features. It is considered as the task of supervised learning meaning that the values of function are set in the training set. Dataset is divided into a training and a test set. In the training set each instance is labeled with a class label that identifies the class it belongs to. Algorithms of supervised learning are used to encourage classifier from the set of correctly classified instances which represents a training set. Testing set is used for measurement of classifier quality acquired after applied training process.

Decision trees [3]. Represent a classifier organized in a hierarchical structure containing zero or more internal nodes and one or more leaves of nodes. All internal nodes have two or child nodes, the criteria that test feature values and execute further division, tree branching respectively. Tree branches, links between internal nodes and their child nods, are marked with the test results. Each node leave has a corresponding class label. A decision tree generates a predictive model where instances are classified by following the requirements path from a tree root up to the leave of a relevant class. The advantages are: simplicity and ease of understanding, the ability to work with numerical and categorical values, prompt assortment of new instances, flexibility, as well as a possibility of visual representation.

Support Vector Machines (SVM) [6]. represents a linear classifier that detects optimal hyper surface in order to separate two different data classes. In the case of linearly separable classes, number of hyper surfaces that classifies training set instances is

infinite. The most effective assortment of new testing set instances will be executed by optimal hyper surface with maximum separation margin. Vectors that limit the margins width are borderline examples of support vectors. Linear combination of support vector is the model solution; other data points are ignored. The number of the selected support vector by the SVM algorithm is usually small, so these classifiers are more suitable for smaller sets with a higher number of feature data. For most data sets, the SVM classifier cannot find the optimal separating hyper surface due to misclassified instances within data. This problem is solved by applying soft margins that accept misclassification of training data set samples. In the case of the inherent data, the SVM classifier performs the data mapping into multi-dimensional feature space and defines an integral hyper surface. Subsequently to hyper surface creation, a special kernel function to map new points in multi-dimensional space classification is used. Taking into consideration that kernel function defines a multi-dimensional feature space where training set instances are to be classified; its choice selection is of extreme importance. In practice, SVMs are generally considered better classifiers because they generate better results.

Bayesian classifiers assume that the knowledge on some event in the world is described by probability of occurrence of such event. For each event in the model there is probability of occurrence that is either assigned or extracted from data. The Statistical dependence of these classifiers is represented by a visual graph structure [7]. Each node matches one feature, income borderlines to the node are determined by certain features, and strength of the dependence is defined by conditional probabilities. When the Bayesian classification network is used, first thing that should be considered is network on dependence structure between features A1, .., Ak and class feature C. After selection of the structure, parameters learn from data and define conditional class distribution for all possible data points and all values of the class C. Classification probability of t data point into class c is calculated as per Bayes rule:

$$P(C = c \mid t) = \frac{P(C = c)P(t \mid C = c)}{P(t)} \tag{1}$$

Bayes network is a very attractive method for educational domain; nevertheless general Bayes network is too complex for sets with small number of data. Solution of such problems is utilization of Naïve Bayes [8] classifiers that generate the model by limited strong assumption independency. Structure on Naïve Bayes network consists of two layers only, the layer of basic node with class variable and the layer of leaves nodes where all others are variable. For conditionally dependent features A_1, ...,A_k, probability for class feature C is calculated according to the following equation:

$$P(A_1, ...A_k \mid C) = \prod_{i=1}^{k} P(A_i \mid C) \tag{2}$$

Simplicity, efficiency, easy interpretation, adjustability to smaller data set are basic advantages of the created Naïve Bayes classification model. In the educational domain an assumption of a conditional independency is often ignored and disturbed. Considering that variables are inter-connected, Naïve Bayes classifiers can tolerate strong surprising dependence between independent variables. It is considered that Naïve Bayes classifiers exceed more sophisticated classifiers such as decision trees and general Bayes classifiers, especially in case of dataset with a smaller number of instances [9].

AODE (Aggregating One-Dependence Estimators) is Bayes method that accomplishes very precise classification by calculating the average over the space of alternative Bayes models that have weaker, and less damaging independent assumptions than Naïve Bayes [10]. The resulting algorithm is more efficient comparing to Naïve Bayes method in case dataset with features that are independent.

Some of the objectives of classifying are: anticipation of students' successfulness on the basis of data extracted from educational system web log files [11]; identification of students with low level of motivation and detection of appropriate activities in order to decrease ratio of giving up [12]; providing feedback to the students on achieved accomplishments [13]; possibility of directing and recommending learning processes in order to achieve the best results possible [14]; detecting a group of students with similar characteristics and reaction to special educational strategies [15]. There are several studies that dealt with the issue of comparison of classifiers accuracy to the data in the field of education. Authors separate the best classifiers for predicting students that give up on the course in e-learning environment by comparison of accuracy of six classifying algorithms[16]. The set of 350 records was extracted which encompassed demographic data, results of the first written task and participation to the group meetings. The data were of a numerical and a categorical type. Two suggested classifiers, Naïve Bayes and Neural network, were able to predict with 80% of accuracy the cases of drop out. Comparison of six classifiers was used to predict the final results of the course realized in the web system for learning [11]. The data extracted from the log files were related to the results of the solved tasks and students activity on the system (participation in communication, reading of educational material). The dataset contained records of 250 students. K-nearest neighbor classifier accomplished over 80% of accuracy when final results were divided into two classes, pass and fail. The authors [17] compare accuracy of five classifying methods for prediction of students' success (pass/fail) on the course in Intelligent Tutor System. Extracted data set had 125 records only. For numerical data, they used linear regression and Support Vector Machines classifier, and for categorical data three variations of Naïve Baies classifiers. They recommended Naïve Baies classifiers that can work with various types of variables and generate probability classes. In the paper [18] the authors test eight classifiers for prediction of students' engagement in virtual courses using data from log files. Applied techniques provide sufficient level of prediction, but IBK algorithm provided the most accurate results. Extracted data sets have 341 and 450 records. All features, except for class variable were numerical. The authors [19] present comparison of various classification algorithms for prediction of the final grade of students on the bases of data from LMS system. The authors [20] created applications that use two algorithms and tested them on the same corpus of documents. For both algorithms, they are presenting improvements that provide a faster search and better results.

This paper uses a concept of classification techniques in order to predict students' grades on the basis of activities realized in the LMS system educational environment and traditional classroom teaching method. Research was conducted for the needs of the case study at the High School of Electrical Engineering and Computer Science of Applied Studies in Belgrade. Dataset was created by integrating Moodle database [21], data about students' participation in the classic teaching and final grades taken from the school information system database. Transformation of numerical data into nominal ones was performed using the method of equal-width interval binning unsupervised discretization which was expanded by analyzing the values of standard deviation aiming to determine

the optimal number of bins for feature domain division. Determination of the most effective method for the selection of optimal features vector is based on comparative analysis of statistical procedures for evaluating classifying models during the phase of predictive modeling. Algorithms providing better results on small samples and supporting work with categorical data were implemented in order to create classification models. The identified methods for ranking and extraction of the optimal set of features were used with the purpose to suggest environment of conceptual model of predictive system. Having in mind that the study uses data from both e-learning and traditional learning environments, the output of the system based on the established methodology should enable a teacher to see the student's activities from distributed sources in the form of a ranking list on the basis of their impact to the final grade. Depending on the data sources that are identified as less significant materials and activities, the teacher can make modification and improvement according to the environment.

The rest of the paper is organized as follows: section 2 describes the feature selection, algorithms used in ranking and extraction of subsets with features of importance and provides an overview of the papers dealing with the similar issues. Section 3 describes the procedure of set extraction for analysis from distributed sources; presents procedure for preparation and discretization of extracted data set; method for ranking and selection of the optimal feature set and calculation of mutual information measure. Section 4 provides procedure for predictive modeling and overview of the statistical measure for results validation. Comparative validation study of implemented methods for feature selection is described in section 5. Section 6 concludes and states the aims of further research.

2. FEATURES SELECTION

Features selection is an important and commonly used technique in the phase of pre-processing data [20]. Optimal features of analyzed dataset are determined by applying an algorithm for selection whereas redundant data without importance and with a hint of noise are removed. In that manner, the effect of acceleration of classification algorithms is achieved and the overall performance is improved, such as prediction accuracy and comprehension of the generated model [23]. Selection of the optimal feature vector allows us to increase the accuracy, comprehensibility of the prediction model and reduces the time needed to generate these models. Implementing such a model removes unnecessary and excess features, extracts optimal subset accelerating data mining algorithms accordingly and achieves higher accuracy and comprehension of the results.

Methods of features selection are used in various research fields for static shape recognition. [24], [25] machine learning [26], [27], text categorization [28], [29], image retrieval [30], client relationship management [31], intrusion detection [32], genome analysis [33]. Yang et al [34] and Forman [35] showed comparative studies in order to find the most effective method of features selection for classification issue. Authors [36] use a smaller dataset to explore advantages and weaknesses of the various methods for selection based on various data types, presence of the noise in data, dataset with multiple values of the class feature. They also use methods of features selection for enhancing predictive accuracy. Application of these methods for removing irrelevant features is presented in the paper [20]

In the domain of education, selection of the best features subset is of the outmost importance especially in the case of predictive modeling since the models are more complex and it is impossible to test all potential ones. However, accuracy prediction decreases in case of the dataset containing irrelevant features and model of prediction becomes unnecessarily complex. In paper [37] authors used a method for features selection for determining an optimal set of input variables that influence the efficiency of grade prediction for the set of 772 students. Total accuracy of the model was 44.69% and the discovered results were satisfying comparing to other model of prediction. In paper [38] the application was aimed at finding the importance of extracted variables for prediction of students who drop out and identification on the most important factors that have influence on students' success. Analysis was conducted on the set of 450 records. For creating a model of prediction algorithms of Regression Trees (CART) and CHAID classification were used. Authors [39] use set of method of features selection for extraction of the subset of input variables. On the extracted set, the following classification algorithms were used: Decision tree, Perception-based Learning, Bayesian Nets, Instance-Based Learning and Rule learning for prediction of student' performances on e-learning module. The dataset included 365 records, and the class grade feature was of binary type with values pass/fail. The conclusion is the Naïve-Bayes generates the most of predictive accuracy (74%) in the case of a class with two values (pass/fail). In paper [40] authors present the effectiveness of the greedy forward feature selection method in classification of student performance prediction. The results demonstrate that neural network classifier generates models with 91.16% accuracy. Authors [41] implemented Information Gain, Gain Ratio, Sequential Backward Selection, Sequential Forward Selection for determining the importance of students' demographic features. They used Naïve Bayes and Support Vector Machines classifiers for evaluating the impact of ranking features to the quality of prediction.

The task of feature selection is to select subsets that are needed and sufficient to describe the target concept [42]. Optimality of the extracted subset is measured by the criterion for assessment. As dimensionality of the domain expands, the number of features increases. Acquiring an optimal subset is mostly complex and complicated procedure related to a great number of issues [43]. In case of supervised learning, the main objective of feature selection is extraction of the subset of model accuracy eliminating the features that have no influence on the model accuracy[44]. This task had shown how to increase accuracy and decrease complexity of the results model [45]. Finding the optimal features subset is a process consisting of four basic steps: subset generating, subset evaluation, stopping criteria and results validation. Subset generating is based on an appropriate search strategy [46] which provides candidates for evaluation. Each subset of candidates is assessed and compared to the best previously corresponding evaluation methods. In case that the new subset is more appropriate it replaces the previously generated one. The process is repeated until it meets the stopping criterion. Evaluation of the best selected subset determines features' significance, disregarding their possible mutual interaction. Valuation methods are based on statistics, information theory, or classification function output [47]. According to the description of the feature value evaluation, features selection algorithms could be divided into three model categories: filter models [48], [49], [50], wrapping models [43] and hybrid models [51].

Based on whether they assess the importance of each feature or subset of features, filter models can be categorized into two groups, the algorithms that estimate rank, feature weighting algorithms and the algorithms that estimate subset of features, subset

search algorithms. The feature weighting algorithms assign specific values for each feature and select an optimal set based on importance to the target concept. A huge number of different definitions of feature importance is stated in the literature mentioned in the papers [20], [27]. A feature is good and has to be selected if the importance of its value is higher than threshold value. Subset search algorithms perform spatial search of features based on applicable measure for evaluation of the subset candidates. An optimal subset is selected once the search is completed. Some of the existing evaluation measures that are considered as efficient in disregarding irrelevant and excessive features include consistency measure [52] and correlation measure [48]. Consistency measure attempts to locate minimal number of features that consistently divide classes as a complete set. Inconsistency is defined for two instances that have feature values but different class labels. Different algorithms are created by combining different search strategies and aforementioned evaluation measures.

Information gain feature ranking is one of the simplest and fastest feature ranking methods, based on entropy measure used in information theory. Let A and C denote the feature and the class, respectively. Entropy class prior to A feature observation is expressed with the following equation

$$\mathrm{H(C)} = -\sum_{c \in C} \mathrm{p(c)} \cdot \log_2 \mathrm{p(c)} \tag{3}$$

where $\mathrm{p(c)}$ is marginal function of density probability for class C. Values of class C, set S is divided on the basis of feature A value. C class entropy monitored in regards to division induced by division of feature A is lower than the entropy prior to division considering relation between class C and feature A. C class entropy after feature A monitoring is expressed by the equation

$$\mathrm{H(C|A)} = -\sum_{a \in A} \mathrm{p(a)} \sum_{c \epsilon C} \mathrm{p(c|a)} \log_2(\mathrm{p(c|a)}) \tag{4}$$

where p(c|a) is conditional probability c for given a. Taking into consideration the fact that entropy can be observed as criteria of impurity within the S set, the measure that reflects additional information on C class could be defined as acquired from the feature *A* that represents the amount *C* class entropy is decreasing for. This measure is known as information gain [53] and is given by the equation 5.

$$\mathrm{IG(Class, Feature)} = \mathrm{H(Class)} - \mathrm{H(Class|Feature)} =$$

$$\mathrm{H(Feature)} - \mathrm{H(Feature|Class)} \tag{5}$$

Relief is the method for features ranking based on instances presented in the paper [42]. Relief algorithm works on the random sample instance principle from the data, subsequently identifying the nearest neighboring ones with the same or the opposite class. Feature values of the nearest neighbors are compared to the instance taken as a sample and used to update the evaluation of each feature's relevance. This procedure is repeated for user specified number of m instances. An important feature should make a difference in case of instances with different class, to have equal values for instances of the same class, respectively. Primarily, this method was defined for class problem with two values ReliefF is an expanded method with the possibility of application for the data set of multidimensional class and occurring noise in the data [54]. ReliefF decreases the noise influence in data on the basis of average income from K- nearest neighbors of the same and opposite class of each sampled instance instead of one nearest neighbor. For

data with the problem of multidimensional class, ReliefF detects nearest neighbors for each class that differ from the currently sampled instance and evaluates contributions of the previous probability of each class. In the case of higher number of sampled instances the reliability of ReliefF ranking algorithm is higher, and at the same time the execution time is increasing.

Symmetrical Uncertainty Feature Evaluation (SymmU) is the method of feature ranking by measuring symmetrical uncertainty in comparison with the class. Estimated values by applying this filter can be in the range from 0 to 1, where one marks that the feature is relevant to the class, and zero that it is class irrelevant [55]. Implementing SymmU method for ranking of the S set features awards with each A feature on the basis of (6).

$$SummU(\text{Class}, \text{Feature}) = 2\left[\frac{(\text{H(Class)} - \text{H(Class} \mid \text{Feature}))}{\text{H(Class)} + \text{H(Feature)}}\right] \tag{6}$$

where H(Class) is a class entropy measure, and H(Class | Feature) is a conditional entropy which qualifies remaining uncertainty of class comparing to the feature.

CFS (Correlation-based Feature Selection) is one of the methods for extraction of an optimal set of features. This method considers predictive ability of each feature, finding out their features and verifies existence of the redundancy among the selected features. A subset of those features that are highly correlated with the class is extracted. Heuristic assessment allocates high value to extracted subset of features. Redundant features are discriminated since they will be significantly co-related to one or more other features. MeritS represents heuristic value of extracted feature subset for the set S and it is calculated by equation (7)[56].

$$\text{Merit}_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{7}$$

In equation (5), *S* is a data set and *k* is the number of features within set *S*. $\overline{r_{cf}}$ represents average correlation of feature and class, and $\overline{r_{ff}}$ average inter-correlation between features. Exponent $k\overline{r_{cf}}$ can be considered as a hint of predictive feature set, and exponent $\sqrt{k + k(k-1)\overline{r_{ff}}}$ marks how many redundant features are among them.

3. CASE STUDY

In research presented in this paper, the conducted methodology consists of the following steps.

1. Extraction of dataset, preparation and transformation
2. Preparation and transformation: applying unsupervised equal-width method by analyzing standard deviation measure.
3. Ranking and optimal feature subset selection: Information-Gain (IG), Symmetrical Uncert Attribute Eval (SUAE) and RelieF (RF) filtering methods are implemented in order to rank all the features of the dataset. Statistical dependence between input features and class label is determined by calculating Mutual information measure. Correlation based Feature Selection (CFS), Wrapper Subset Evaluation (WSE) and Classifier Subset Evaluator (CSE) methods are used to extract optimal subsets from input features.

4. Predictive modeling of datasets with different cardinality. Naïve Bayes classifier is applied to the generated subsets. The evaluation process detected the best performance for subsets of different cardinality. Results are validated by applying Decision trees, Support Vector Machines and AODE classifiers on the best subsets and evaluating the performance (accuracy, time required to build model) of the classifiers.

3.1.Dataset extraction

Educational data are very transparent since they are collected automatically from log files and databases. Missing values and wrong feature values are a common occurrence, especially in case when educational data set created from multiple sources. Cleaning the "raw" data involves identifying missing values, incorrect values, hint of noise and then applying appropriate procedures to perform irregularities correction. The extracted set is organized in a form of a table whereas rows represent instances, and column of features, data features, respectively. At the High School of Electrical Engineering and Computer Science of Applied studies in Belgrade LMS Moodle system is used not only as a platform for realization of distance learning module but for blended learning program of support to classical teaching as well. The research described in this paper is aimed to analyze the course Computer graphics for the school year 2014/15. We have extracted from LMS Moodle system database information about activities of students in e-learning course. Data about student's activities in classroom and laboratory exercises were taken from the application for participation in the classical teaching. Final grades were taken from the information system of the educational institution. Learning materials in the form of lessons and video tutorials, exercises and tests were available to students within the Moodle course. Also, students had an opportunity to participate in discussion forums and electronic consultation through private text exchange. Features FD, MM, LVT, PDF, LESS¸ DZ1, DZ2, DZ3, DZ4, DZ5 P1, P2 T1, T2¸ and FT are extracted from Moodle database [57]. Features LAB and BB represent score points of the laboratory exercises and classroom activity, respectively. Final grades are taken from the VISER' information system database.

Dataset is created by integrating data from multiple sources into a form applicable for data mining technique application. Table columns represent features and rows record for each student. Extracted dataset had 225 instances. The data were numerical and categorical. Feature names and descriptions are given in Table 1.

Table 1 Features for extracted dataset

Feature	Description
FD	Participation in discussions in forums
MM	Usage of private Moodle messages for e-consultation
LVT	Usage of e-tutorials
PDF	Usage of PDF materials
LESS	Activities in accessed lessons
DZ1, DZ2, DZ3, DZ4, DZ5	Points won in solving first, second, third, fourth, fifth exercise
P1, P2, P3	Average points of all attempts in solving preparatory test for selftesting
T1, T2	Points won in first and second test
FT	Points won on the final exam
LAB	Points won at the lab exercises
BB	Points won for activities during the class
Grade	Final grade

3.2. Dataset preparation

Educational data sets contain different types of features (numerical and categorical, ordinal and nominal), which in most cases are characterized by uneven interval values. Within a preparatory phase, a detailed analysis of extracted data set domain values was conducted. For the purposes of electronic consultations, students were using a forum where they were able to ask questions, participate in discussion, inter-communicate, and provide answers and responses together with teachers. Such students' activity was labeled as an FD feature. Another form of consultation was enabled by private text messaging utilization within the course that students could use between themselves as well as to consult the teacher. This activity was extracted as a MM feature. Value domain of these two features, after extraction from Moodle database was of a numeric type in terms of the number of posts and sent texts by students and included the integer numerical values from 0 to 10. For preparation of lab exercises, video tutorials and PDF databases were at students' disposal and for taking tests and exams, the Moodle lessons in the form of HTML pages with questions at the end of the lesson for knowledge check up. Utilization of PDF material and number of views of video tutorials was extracted as PDF and LVT features. It marked the number of utilized PDF documents. Value domain of these features encompassed the set of integer numerical values from 0 to 10. The method of utilization of Moodle lessons was extracted as a LESS feature. By extraction from the Moodle database, the set of nominal values was acquired as the result {NoAction, LessonView, LessonViewAndAnswerQuestion} that represent a domain of this feature. Features DZ1, DZ 2, DZ3, DZ4, DZ5 represented points won for homework created as Moodle tests. Each test included appropriate teaching material from lectures and exercises. DZ1, DZ 2, DZ3 included encompassed areas tested at the first colloquium, whereas DZ4, DZ5 areas tested at the second one. The tests were adjusted for two attempts of solving, and the attempt with the higher number of gained points was assessed and they were available until the beginning of the exam term. Homework was not mandatory, and students could choose time and place of testing. Value domain of such features included the numerical values from 0 to 2. Features P1, P2, P3 represented the maximum of gained points of all attempts of solving preparatory tests. Number, type and questions, temporal duration is adjusted in a manner that it simulates eliminatory tests. Students could take such tests for an unlimited number of times during the period of availability. Features T1, T2 represented the points gained on first and second test within one semester. Feature value domain P1, P2, T1, T2 included numerical values from 0 to 20. The points gained at the exam test were extracted as the feature FT with value domain from 0 to 30 that was also a domain for the feature P3 that represented the points from preparatory test from the exam. The points gained on exercises were extracted as the feature LAB with values from the set from 0 to 15. Activities and presence at the lectures were marked as the feature BB and encompassed value set from 0 to 5. Feature Grade marked final grade of students extracted from the information system. For this feature, values belonged to the numerical domain {5,6,7,8,9,10}.

3.3. Dataset transformation

Variations in size and type of the feature domain were determined by analyzing values of extracted dataset. For numerical features statistical measures of standard deviation and arithmetic mean were calculated and presented in Table 2.

Table 2 Value type, mean, standard deviation of numerical features

Feature	Value type	Mean	StDev	Value range
LAB	real	12.453	+/- 1.483	[8.000 ; 14.900]
DZ1	real	1.587	+/- 0.438	[0.160 ; 2.000]
DZ2	real	1.726	+/- 0.401	[0.640 ; 2.000]
DZ3	real	1.727	+/- 0.380	[0.400 ; 2.000]
DZ4	real	1.430	+/- 0.486	[0.220 ; 2.000]
DZ5	real	1.560	+/- 0.540	[0.000 ; 2.000]
T1	real	12.441	+/- 5.348	[0.000 ; 20.000]
T2	real	13.116	+/- 4.405	[1.000 ; 20.000]
FT	real	20.601	+/- 6.810	[0.630 ; 30.000]
PDF	integer	4.420	+/- 2.694	[0.000 ; 8.000]
LVT	integer	6.630	+/- 3.739	[0.000 ; 12.000]
Grade	real	7.556	+/- 1.733	[5.000 ; 10.000]

Discretization is an important special case of reduction that transforms numerical continual values into smaller sets of discrete (numerical of categorical) values. It was enforced by applying unsupervised method of domain value division into intervals of equal size. In the manner the scope of data was reduced and numerical values transformed into appropriate more understandable classes. Number of intervals of domain division of numerical features is determined on the basis of statistical measure valued of standard deviation *(StDev).* Feature value domain *DZ1, DZ2, DZ3, DZ4, DZ5* is divided into two intervals. *StDev* of such features had the following values +/- 0.438, +/- 0.401, +/- 0.380, +/- 0.486 respectively. Feature value domain LAB is divided into three intervals. *StDev* of such feature had value of +/- 1.483. Feature value domain *PDF, LVT* is divided into three intervals. *StDev* of *PDF* feature had value of +/- 2.694, and of *LVT* feature +/- 3.739. Feature value domain *T1, T2* is divided into four intervals. *StDev* of such features had value of +/- 5.348 and +/- 4.405. Feature value domain FT is divided into four intervals. *StDev* of such features had value of +/- 6.81. Discretization for those features was executed according to following pseudo code:

set max(Att_i)
set min(Att_i)
md (Att_i)= $\frac{max(Atti)-min(Atti)}{2}$
set max($StDev_i$)
set min($StDev_i$)

if -0.5< $StDev_i$ <0.5
then
    if values(Att_i) < *md* (Att_i)
    values(Att_i) = *bad*
    if values(Att_i) >= *md* (Att_i)
    values(Att_i) = *good*
end

if -1.5< StDev$_i$ <1.5
set th(Att$_i$) $= \frac{max(Atti)-min(Atti)}{3}$
then
if min(Att$_i$) <values(Att$_i$) < th(Att$_i$)
values(Att$_i$) = insufficient
if th(Att$_i$)<=values(Att$_i$) < md (Att$_i$)
values(Att$_i$) = sufficient
if md (Att$_i$)<=values(Att$_i$) <= max(Att$_i$)
values(Att$_i$) = most
end
if -5.5< StDev$_i$ <5.5
set th(Att$_i$) $= \frac{max(Atti)-min(Atti)}{3}$
set qr(Att$_i$) $= \frac{max(Atti)-min(Atti)}{4}$
then
if min(Att$_i$) <values(Att$_i$) < qr(Att$_i$)
values(Att$_i$) = bad
if qr(Att$_i$)<=values(Att$_i$) < th(Att$_i$)
values(Att$_i$) = good
if th(Att$_i$)<=values(Att$_i$) < md (Att$_i$)
values(Att$_i$) = very good
if md (Att$_i$)<=values(Att$_i$) <= max(Att$_i$)
values(Att$_i$) = excellent
end

Grade feature is labeled as a class and based on points that they gained during the semester. Discretization was performed at six intervals therefore instances were allocated as per intervals on the basis student's final grade.

min(Grade) =5, max(Grade) = 10
Grade = fail, if grade is = 5
Grade =six, if grade is = 6
Grade = seven, if grade is = 7
Grade =eight if grade is = 8
Grade =nine, if grade is = 9
Grade =ten if grade is = 10

Feature value domain BB, P1, P2, P3, FD, MM showed a great data dispersion that conditioned its transformation into binominal features. Transformation and discretization of value domain of such features were executed as per the following rules:

- P1, P2, P3: student complete / not complete preparatory test
- BB : student participate / not participate in the classic teaching lectures
- MM: student utilize/not utilize Moodle private text messaging for consultations
- FD: student participate /not participate on Moodle forum

The polynominal LESS feature included the set of three possible values. Discredited dataset with transformed feature values is presented in Table 3.

Table 3 Dataset with transformed feature values

Feature	Categorical feature domain
LAB, PDF, LVT	insufficient, sufficient, most
BB, P1, P2, P3, FD, MM	yes, no
DZ1, DZ2, DZ3, DZ4, DZ5	bad, good
T1, T2, FT	bad, good, very good, excellent
LESS	NoAction, LessView, LessViewAndAnswerQuestion
Grade	fail, six, seven, eight, nine, ten

3.4. Ranking and optimal feature subset selection

Selection of the most important features from the input variables of analyzed dataset has the great influence on the model performances. Selection of the appropriate method depends on several factors aimed at achieving as realistic results as possible. In the procedure of creating the optimal vector two processes are anticipated, such as: ranking of the features on the basis of influence on the class and extraction of the optimal feature set. Besides that, the following factors were taken into consideration: time of execution, representation of the resulting output, correlation between expected and total number of features, dimensionality of the class feature, type of the feature, data quality and correlation between total number of features and total number of instances in the dataset [58]. In this case study, original feature vector contained 18 input features LAB, DZ1, DZ2, DZ3, DZ4, DZ5, T1, T2, FT, PDF, LVT, LESS, P1, P2, P3, MM, FD, BB. For output, multidimensional feature Grade with domain of six categorical values (fail, six, seven, eight, nine, ten) was defined as the response feature. Input data type as well as response variable was nominal. In the research filter algorithms were used, such as InformationGain (IG), SymmetricalUncertFeatureEval (SUAE), ReliefF (RF) combined with Ranker search method for feature ranking. For extraction of the optimal feature subset following algorithms were applied CFS (Correlation-based Feature Selection) and Wrapper Subset Evaluation (WSE) with GreedyStepwise search method. Table 4 presents feature ranking.

Table 4 Feature weighting methods

Rang	Information Gain		Symmetrical Uncertainty		ReliefF	
	Weight	Ranked features	Weight	Ranked features	Weight	Ranked features
1	1.3206	FT	0.56947	FT	0.62976	FT
2	1.0253	T2	0.43656	T2	0.53215	T2
3	0.9287	T1	0.39549	T1	0.42601	T1
4	0.5982	DZ5	0.32004	DZ5	0.34691	DZ5
5	0.4543	DZ3	0.24498	DZ3	0.27098	BB
6	0.4437	DZ1	0.24173	DZ1	0.25796	DZ4
7	0.4023	DZ4	0.22464	LAB	0.23767	DZ2
8	0.3926	DZ2	0.21523	DZ4	0.23664	DZ3
9	0.3926	LAB	0.21114	DZ2	0.23368	DZ1
10	0.3632	BB	0.19718	BB	0.12749	LAB
11	0.0899	P2	0.04815	P2	0.10392	P2
12	0.0453	LVT	0.021	LVT	0.07065	P3
13	0.0442	LESS	0.02055	LESS	0.04745	P1
14	0.027	PDF	0.01336	P3	0.02418	PDF
15	0.0249	P3	0.01273	PDF	0.00305	MM
16	0.0175	P1	0.01079	MM	-0.012	FD
17	0.0163	MM	0.00968	P1	-0.019	LESS
18	0.0148	FD	0.00794	FD	-0.019	LVT

FT, T1, T2, DZ5, P2 were ranked so that the features FT, T1, T2, DZ5 were assigned rank 1,2,3,4 respectively, and feature the P2 rank 1. Feature DZ3 was assigned rank 5 by algorithms IG and SUAE, but rank 8 by RF algorithm. The same incongruence occurred in the case of DZ1 feature (Rang (IG, SUAE)=6, Rang(RF)=9), BB (Rang (IG, SUAE)=10, Rang(RF)=5), LVT (Rang (IG, SUAE)=12, Rang(RF)=18), LESS (Rang (IG, SUAE)=13, Rang(RF)=17), PDF (Rang (IG, RF)=14, Rang(SUAE)=15), FD (Rang (IG, SUAE)=18, Rang(RF)=16). Features LAB, DZ2, DZ4, P1, P2, P3, MM were ranked differently in case of all three algorithms.

For extraction of the optimal feature vector following methods were implemented Correlation based Feature Selection (CFS), Wrapper Subset Evaluation (WSE) and Classifier Subset Evaluator (CSE) with GreedyStepwise search method. In case of application of WSE and CSE methods, NaïveBayes classifier was used. Results of implemented subset search methods are given in Table 5.

Table 5 Subset search methods

Method	Subset
Correlation based Feature Selection	FT, T2, T1, DZ5, BB, P3,FD
Wrapper Subset Evaluator- NB	FT, T2, T1, DZ5, LAB
Classifier Subset Evaluator - NB	LAB, DZ5, T1, T2, FT, FD

CFS, WSE and CSE method extracted feature subsets with features FT, T2, T1, DZ5 within, the best ranked features by previously applied filter methods. Subset with the smaller number of features was extracted by WSE method (5 features), and subset with the highest number of features by CFS method (7 features). Calculation of mutual information measure MI (Mutual information) was executed for combination of input variables with response variable – Grade. Mutual information [59] is information theory measure for detecting statistical dependence between two features that determines the quantity of information two variables share X and Y. If X and Y are independent, then variable X does not contain information on Y and vice versa, while their mutual information is zero. In other extreme when X and Y are identical, then they share all information. In table 6 are given results of calculated statistical dependence for which information measure is equal or greater than set threshold value (MI >=0.10).

Measure of mutual information extracted features that were marked in previous measures with the measure of higher importance than the others. For features FT, T2, T1, DZ5, DZ4, DZ1, BB, LAB, DZ3, DZ2 higher value of mutual measure MI was noticed that indicated higher independence of listed features comparing to the class one. Important values of mutual information measure were within the range from 0.16 to 1.67. For features MM, FD calculated MI value was lower than importance threshold 0.1, indicating the fact that these features have no influence on the class one, therefore they were not taken into further consideration. Features FT, T2, T1, DZ5 were marked with the highest values of dependence with the class feature. As previously applied methods extracted aforementioned features, it is determined that FT, T2, T1, DZ5 represent part of optimal feature vector. Although the primary objective of a method for feature selection is extraction of minimal cardinality subset, the incorrectly ignored features are also of great importance. Incorrectly disregarded or selected features in the educational domain could cause generation of complex or unreliable prediction models.

Therefore, it is necessary to determine importance of other features that resulted in Noncompliant results in the further process of research.

Table 6 MI values

Feature	Mutual Information
FT	1.67
T2	1.33
T1	1.28
DZ5	0.97
DZ4	0.83
DZ1	0.83
BB	0.81
LAB	0.79
DZ3	0.76
DZ2	0.69
P2	0.38
P1	0.34
P3	0.34
PDF	0.14
LVT	0.16
LESS	0.16

3.5. Feature subset with different cardinality used for comparison based on assessment of classification models

Noncompliant results of applied methods indicated the need for predictive modeling of extracted subset with different cardinality. For each extracted feature subset, NB classifier model was generated and calculation of statistical measure correctly classified instance (CCI), Kappa, Precision, Recall, F-Measure. Accuracy classification represents statistical measure of positively and negatively classified instance; the probability that instance shall be correctly classified respectively. The average of classified instances where estimated values of class feature are identical to real values represents True Positive Rate (TPR). False Positive Rate (FPR) shows average of instances where estimated values of class feature are not identical with real values. True Negative Rate (TNR) shows percentage of really negatively classified instances, whereas False Negative Rate (FNR) percentage of falsely classified negative instances. F-measure represents harmonical middle of measure values and response.

Precision is defined as a ratio of relevant samples in the total number of detected samples compared to the total number of relevant samples and it is calculated by equation (8). Recall and F-Measure are defined by equations (9) and (10) respectively:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{9}$$

$$\text{F} - \text{Measure} = \frac{\text{2xRecallxPrecision}}{\text{Recall+Precision}} \tag{10}$$

Kappa measure was used as the indicator of importance, which measures estimation compliance to the real class. Basically, Kappa is appropriate to test whether predictive and actual classes are correlated. Values of Kappa parameter between 0.4 and 0.5 represents a moderate correlation, values from 0.6 to 0.79 significant, a value of about 0.8 remarkable correlation [60]. It is considered that Kappa value of at least 0.6, very often and more than 0.7 indicated good level of agreement. High level of agreement is accomplished in case when Kappa value is above 0, 5, a low level of agreement if Kappa is less than 0.3.

For implementation of analyzed selection methods open-source Weka 3.6 version was used [61]. IG, SUAE, RF algorithm filters ranked 18 input features. Same features in all three ranking methods were ranked by 1, 2, 3, 4, 11. Rank 5, 6, 12, 13, 14, 18 was assign to the same features but for two methods only. All three methods assigned weights of ranked features that were heading to following directions: Weight (IG) from 0.0148 to 1.3206, Weight (SUAE) from 0.00794 to 0.56947, Weight (RF) from -0.019 to 0.62976 . To determine the subset of optimum ranked features the weight threshold was set at 0.1. In case of calculation of mutual measure information feature subsets were created on the basis of various range of measure value MI; S1(MI from 0.69 do 1.67), S2(MI from 0.34 do 1.67), S3(MI from 0.16 do 1.67). For extracted subsets of different cardinality Naïve Bayes classifying model was generated.

From presented results in Table 8, we observed that performances of created NB models differ slightly. For subset IG(10) and SUAE(10), NB classifier produced same model. The models with the best performances were created for subset RF(11), WSE(5) and S2 subset with mutual information measure MI from 0.34 to 1.67. Cardinality of extracted optimal subsets was different: S(WSE)=5, S(RF)=11, S(MI)=13.

Table 8 NB models for subsets of different cardinality:

Feature subsets	Kappa	P	R	FM
IG (10) FT, T1, T2, DZ5, DZ3, DZ1,DZ4, DZ2, LAB, BB	0.6	0.7	0.7	0.7
SUAE (10) FT, T1, T2, DZ5, DZ3, DZ1,LAB, DZ4, DZ2, BB	0.6	0.7	0.7	0.7
RF (11) FT, T2, T1, DZ5, BB, DZ4, DZ2, DZ3, DZ1, LAB, P2	0.7	0.74	0.74	0.73
CFS (7) FT, T2, T1, DZ5, BB, P3,FD	0.6	0.7	0.7	0.7
WSE(5) FT, T2, T1, DZ5, LAB	0.7	0.73	0.73	0.7
CSE(6) LAB, DZ5, T1, T2, FT, FD	0.6	0.7	0.7	0.7
S_1(MI from 0.69 do 1.67) FT, T2,T1,DZ5,DZ4, DZ1,BB, LAB, DZ3, DZ2	0.6	0.7	0.7	0.7
S_2(MI from 0.34 do 1.67) FT, T2,T1,DZ5,DZ4, DZ1,BB, LAB, DZ3, DZ2, P2,P1,P3	0.7	0.74	0.74	0.73
S_3(MI from 0.16 do 1.67) FT, T2,T1,DZ5,DZ4, DZ1,BB, LAB, DZ3, DZ2, P2,P1,P3,PDF,LVT,LESS	0.6	0.71	0.71	0.7

Final determination of the method that extracts the most relevant optimal feature vector subset is based on comparison of prediction accuracy decision trees, support vector and AODE classificatory model. Table 11 shows percentage of correctly classified instance and time necessary to generate classifying models.

Table 11. Correctly classified instance and time necessary to generate models

Feature selection	Decision Tree		Support Vector		AODE	
	CCI (%)	Time (s)	CCI (%)	Time (s)	CCI (%)	Time (s)
MI (13) (FT, T2,T1,DZ5,DZ4, DZ1,BB, LAB, DZ3, DZ2, P2,P1,P3)	74	0.1	73	0.4	74	0.01
RF (11) (FT, T1, T2, DZ5, BB, DZ4, DZ2, DZ3, DZ1, LAB, P2)	72	0.1	72	0.3	74	0.01
WSE(5) (FT, T2, T1, DZ5, LAB)	72	0.06	72	0.1	72	0.01

From the obtained results, we can conclude that the models with the highest percent of correctly classified instances for all three classifiers were created for subset MI(13) For the implemented methods, the best ranked features are FT, T2, T1 and DZ5. On the basis of the above mentioned it can be concluded that the first, second test, final exam and the last given homework are the most important students' activities that have great influence on the final grade. Optimal feature vector includes testing and self-testing activities within Moodle environment as well as students' participation in the classical lectures and laboratory exercises.

4. Results and Discussion

Research described in this paper deals with application of various methods for feature selection within data preprocessing phase in data mining analysis process. Dataset is created by combining data from distributed sources: LMS Moodle System, application for recording students' activity during classical education and information system of educational institution VISER. The extracted set was characterized by data of numerical and categorical type. In preprocessing phase, raw data were cleared from presence of the noise and missing values. Transformation of numerical data domain was executed by equal-width interval binning method that can be categorized as more simple direct method of unsupervised discretization. This method executed division of domain of observed features on k bins of the same size (δ) with *k+1* division point. On the basis of the domain value distribution of each numerical feature, taking into consideration standard deviation measure *(StDev),* a division to discrete values was executed. Each discrete value was assigned an applicable categorical label for better understanding. The discretized dataset is defined by 18 input features that were nominal and binominal data type. Multi-dimensional class feature Grade with category values for which class labels were introduced {fail, six, seven, eight, nine, and ten} was determined. The procedure of feature selection was implemented by applying filter methods Information-Gain (IG), SymmetricalUncertFeatureEval (SUAE), Relief (RF) combined with Ranker search method. All three methods ranked same features

by rank 1, 2, 3, 4, 11. Rank 5, 6, 12, 13, 14, 18 was assigned to the same features but by two methods only. For selection of optimal vector feature set the following methods were used: Correlation-based Feature Selection method, Wrapper Subset Evaluation and Classifier Subset Evaluator with GreedyStepwise search method. For implementation of Wrapper Subset Evaluation and Classifier Subset Evaluator, the selected classifier was NaïveBayes. Correlation-based Feature Selection method extracted vectors of seven features. Classifier Subset Evaluator vector with six features and Wrapper Subset Evaluation vector with five features. All three methods, among others, selected four identical features FT, T2, T1, DZ5 which filter method ranked the best. The statistical dependence of input features in comparison with the class one was determined by calculating the rates of mutual information. Value of lower significance threshold was set at 0.1. Calculated values of MI rate were divided into three ranges: 0.69 <= MI = <1.67, 0.34 <= MI = <1.67), 0.16 <= MI <= 1.67 and in that manner three features subsets S1, S2, S3 with different cardinalities were extracted. Procedure of determining the method that extracted most of input feature optimal vector with significant influence on class variable was based on predictive modeling. Naïve Bayes classifying models were created for sets of various cardinalities. Models of the best performances were created for subset of feature cardinality 11 extracted by ReliefF method, for subset of feature 5 cardinality extracted by Wrapper Subset Evaluation method and for S_2 cardinality subset 13 with measure of mutual information MI in the range from 0.34 to 1.67. Starting from the fact that the analyzed subset of educational data was extracted from distributed sources the goal of this research was aimed at determining, as accurately as possible, a feature subset that has a positive influence on classification performances.

Over potentially optimal sets MI(13), RF(11) and WSE(5) decision tree classifying model was created together with support vector and AODE Bayesian. Comparative model analysis was based on comparison of percentage of correctly classified instances and time necessary to create models. Models with the highest percent of correctly classified instances were created for subset MI(13), for all three classifiers. Figure 1 shows comparative model analysis based on comparison of percentage of correctly classified instances for all three classifiers.

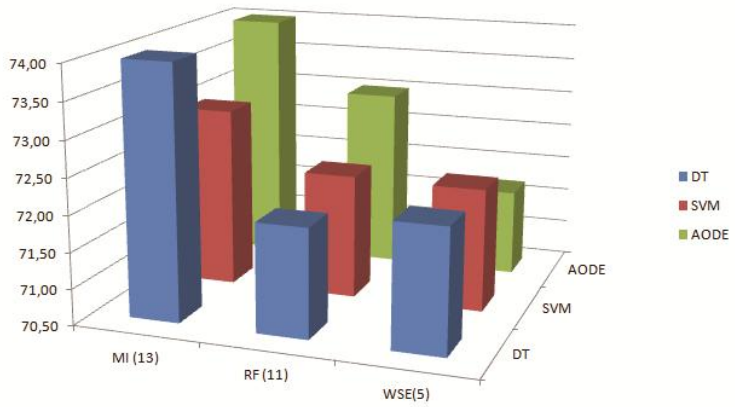

Fig. 1 Percentage of correctly classified instances for DT, SVM, AODE classifier

Figure 2 shows time necessary to create models for all three classifiers. Time needed to create a model on analyzed subsets was approximate. In case of decision tree DT and support vector smallest amount of time was recorded for subset WSE (5). Time to create AODE, was the same for all three sets and represented general minimum time in case of generating enlisted models. Case study showed that calculation of mutual information measure between input and class nominal feature for set extracted by combining multiple sources of educational data, enables creation of optimal vector of features A_{opt} (a_1, a_2, ..a_k).

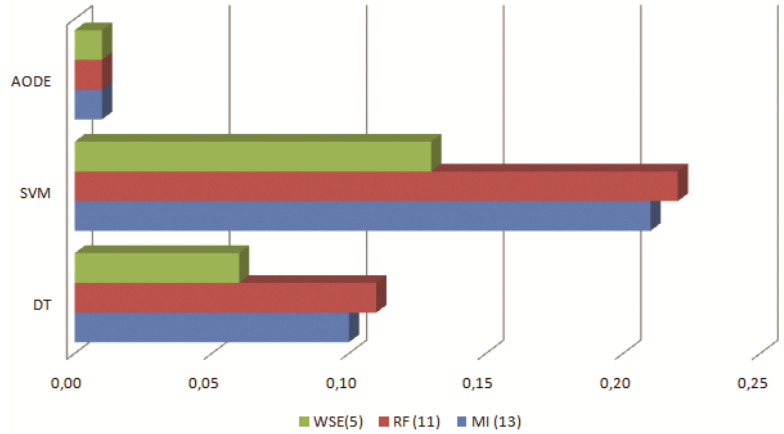

Fig. 2 Time necessary to create DT, SVM, AODE model

Decision Tree, SVM, AODE classificatory models generated over the dataset with A_{opt} vector input features, achieved approximate prediction accuracy higher than 70% during timeframe from 0.01s to 0.21s. On the basis of described research, we have come to the conclusion that selection of optimal feature vector subset of educational data requests implementation of appropriate filter and wrapped methods combined with calculation of mutual information measure. The consideration of which technique is best depends upon the nature of data used for experiment.

5. Conclusion

This paper conducted a comparative study of the methods for ranking, extracting of optimal subset and mutual information measure of dependability between input and class features. Analyzed set was created by combining distributed sources of educational data. The aim was to determine the methodology process that will define the subset of optimal dimensionality features. Determination of optimal dimensionality was carried out by testing the extracted subset using predictive modeling procedures. Naïve Bayes, AODE, decision tree, support vector machine classifiers were implemented. Selection of the most efficient models was conducted by comparative analysis of following measures of created predictive models: CCI, Kappa, Precision, Recall, F-Measure. The case study described combined application of ReliefF filter method, WSE wrapping method and MI mutual information measure which extracted optimal features vector of the educational dataset.

Research pointed out the fact that the optimal feature vector does not also imply minimum cardinality. The drawn conclusion is of great importance since incorrectly selected or rejected features of data set cause weaker predictive models of poor performance. Further research will be focused on testing the influence of supervised and unsupervised discretization methods in preprocessing phase on proceeding of extracting the most relevant feature vector in order to increase predictive accuracy in blended learning environment.

REFERENCES

[1] U. Fayyad,G. Shapiro, P. and P.Smyth, "From data mining to knowledge discovery in databases", AI Magazine, Vol.17, No.3, pp.37–54, 1996.
[2] W. Frawley, J.Shapiro, G. P. and C.J. Matheus, "Knowledge discovery in databases: An overview", AI Magazine, Vol.13, No.3, pp.57–70, 1992
[3] I.H.Witten and E.Frank, "Data mining – Practical Machine Learning tools and Techniques (3rd Edition.)," Book, Morgan Kaufmann Publisher, 2011.
[4] C. Romero,S Ventura,"Educational Data Mining: a Survey from 1995 to 2005", Expert Systems with Applications, Vol.33, No.1, pp.135-146, 2007.
[5] R. Baker, K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, No.1, Vol.1, pp.3-17, 2009.
[6] P.-N. Tan, M. Steinbach, V.Kumar, "Introduction to Data Mining", Book, Publisher:Addison-Wesley, 2006.
[7] J.Pearl,"Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Publishers Morgan Kaufmann, San Mateo, CA, 1988.
[8] H. Zhang,"The optimality of Naive Bayes",Proc. *17th Int.FLAIRS conference*, AAAI Press, 2004.
[9] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", Machine Learning, Vol.29, Issue 2, pp.103–130, 1997.
[10] G. Webb, J. Boughton, Z. Wang, "Not So Naive Bayes: Aggregating One-Dependence Estimators", Machine Learning, Vol.58, Issue 1, pp.5-24, 2005.
[11] B. Minaei-Bidgoli, Kashy, D. A. Kortemeyer, G. and Punch, W. F."Predicting student performance: an application of data mining methods with an educational web-based system", 33rd Annual Conference on Frontiers in Education (FIE 2003), Vol. 1, pp. 13-18, 2003.
[12] M. Cocea, S Weibelzah, "Can Log Files Analysis Estimate Learners' Level of Motivation?", Workshop on Adaptivity and User Modeling in Interactive Systems, pp.32-35, 2006.
[13] A. J. Angel, T. Daradoumis, J. Faulin and F. Xhafa, "A data analysis model based on control charts to monitor online learning processes", International Journal Business Intelligence and Data Mining, Vol.4, No.2, pp.159 – 174, 2009.
[14] X.Li, Q.Luo and J.Yuan, "Personalized recommendation service system in e-learning using web intelligence", *In Proc. 7th Int. conf. Computational Science*, Part III, pages 531–538, ICCS 2007.
[15] G. Chen at all. "Discovering Decision Knowledge from Web Log Portfolio for Managing Classroom Processes by Applying Decision Tree and Data Cube Technology", Journal of Educational Computing Research, Vol.23, No.3, pp.305–332, 2000.
[16] S B. Kotsiantis, C.Pierrakeas, P.E. Pintelas, " Preventing student dropout in distance learning using machine learning techniques" *In proc. 7th Int.Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, pp.267-274, 2003.
[17] W. Hämäläinen, M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems", *In proc. of the 8th Int. Conf. on Intelligent Tutoring Systems*, pp. 525-534, June 2006.
[18] M.Cocea, S.Weibelzahl, "Cross-System Validation of Engagement Prediction from Log Files." *Second European Conference on Technology Enhanced Learning*, EC-TEL 2007, pp. 14–25, 2007.
[19] C. Romero, S. Ventura, P.G. Espejo and C. Hervas. "Data mining algorithms to classify students", *In Educational data mining 2008: Proceedings of the 1st international conference on educational data mining*, 8–17, 2008.
[20] M.Ilić, D. Rančić, P. Spalević, "Comparison of data mining algorithms, inverted index search and suffix tree clustering search", FACTA UNIVERSITATIS Series: Automatic Control and Robotics Vol. 15, No 3, 2016, pp. 171 - 185 DOI: 10.22190/FUACR1603171I
[21] Moodle, a free open source course management system for online learning, http://moodle.org/ (2006)
[22] A.L. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, vol. 97, pp. 245-271, 1997.

[23] H. Liu and H. Motoda, "Feature Extraction, Construction and Selection: A Data Mining Perspective", Springer-Verlag New York Inc 2013.
[24] M. Ben-Bassat, "Pattern Recognition and Reduction of Dimensionality," Handbook of Statistics-II, pp. 773-791, North Holland, 1982.
[25] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 3, pp. 301-312, March, 2002.
[26] K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," *In Proc.10th Nat'l Conf. Artificial Intelligence*, pp. 129-134, 1992.
[27] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, No. 1-2, pp. 273-324, 1997.
[28] E. Leopold and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?", Machine Learning, Vol. 46, pp. 423-444, 2002.
[29] K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM," Machine Learning, Vol. 39, pp.103-134, 2000.
[30] Y. Rui, T.S. Huang, and S. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues", Journal of Visual Communication and Image Representation, Vol. 10, No. 1, pp. 39-62, 1999.
[31] K.S. Ng and H. Liu, "Customer Retention via Data Mining," AI Review, Vol. 14, No.6, pp. 569-590, 2000.
[32] W. Lee, S.J. Stolfo, and K.W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach," AI Review, Vol. 14, No. 6, pp. 533-567, 2000.
[33] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data,"Proc. 18th Int'l Conf. Machine Learning,pp. 601-608, 2001.
[34] Y.Yang and J Pederson, "A comparative study on feature selection in text categorization*", Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 412–420, 1997.
[35] G. Forman, "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, Vol.3, pp. 1289-1305, 2003.
[36] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, Vol. 1, No. 1-4, pp.131–156, 1997.
[37] M. Ramaswami, R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining," International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010.
[38] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data*", Informing Science & IT Education Conference*, pp.647–665, 2010.
[39] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, No. 5, pp. 411-426, 2004.
[40] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining", *7th International Conference on Information Technology and Electrical Engineering* , pp. 420-424, DOI:10.1109/ICITEED.2015.7408983, 2015.
[41] B. Trstenjak and D. Đonko, "Determining the impact of demographic features in predicting student success in Croatia," *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention*, pp. 1222-1227, DOI: 10.1109/MIPRO.2014.6859754, 2014.
[42] K. Kira and L. Rendell,"A practical approach to feature selection", *ML92 Proceedings of the 9th International Workshop on Machine Learning*, pp. 249-256, Morgan Kaufmann Publishers Inc. San Francisco, CA, 1992.
[43] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection", Artificial Intelligence, Vol. 97, Issues 1-2, pp. 273-324, 1997.
[44] S. Piramuthu "Evaluating feature selection methods for learning in data mining applications" European Journal of Operational Research, Vol. 156, Issue 2, pp.483–494, 2004.
[45] D. Koller and M. Sahami, "Toward optimal feature selection", *Machine Learning: Proc. of the 13th International Conference*, Morgan Kaufmann, 1996.
[46] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining" Book, Kluwer Academic Publishers Norwell, MA, 1998.
[47] W. Duch at all, "Feature Ranking, Selection and Discretization*", International Conference on Artificial Neural Networks (ICANN) and International Conference on Neural Information Processing* , pp. 251-254, 2003.
[48] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning", *Proc. 17th Int'l Conf. Machine Learning*,pp. 359-366, 2000.
[49] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection-A Filter Solution", *Proc. 13th Int'l Conf. Machine Learning*,pp. 319-327, 1996.
[50] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", *Proc. 20th Int'l Conf.Machine Learning*, pp. 856-863, 2003.

[51] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection", *Proc. 18th Int'l Conf. Machine Learning*, pp. 74-81, 2001.
[52] M. Dash, H. Liu, "Consistency-based feature selection", Journal Artificial Intelligence, Vol.151, Issues 1-2 , pp.155-176, 2003.
[53] J. R. Quinlan, "C4.5: Programs for Machine Learning", Book, Morgan Kaufmann Publishers Inc. San Francisco, CA, 1993 .
[54] I. Kononenko, "Estimating features:Analysis and extensions of RELIEF", *In: Proceedings of the 17th European Conference on Machine Learning*, pp. 171-182, 1994.
[55] L.Hu and L. Zhang, "Real-time internet traffic identification based on decision tree", *World Automation Congress (WAC)*, pp.1-3, 2012.
[56] M. A. Hall, "Correlation–based Feature Selection for Machine Learning" PhD thesis, University of Waikato, 1999.
[57] G. Dimić, D. Prokin, K. Kuk, P.Spalević,"The use of data mining methods for analyzing and evaluating course quality in the Moodle system", *Международна научна конференция УНИТЕХ'10, Габрово*, pp. 309-315,2010.
[58] H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, pp.491-502, 2005.
[59] R.O. Duda, P.E. Hart, and D.G. Stork,"Pattern classification", Book,Wiley-Interscience Publication, 2nd edition, 2000.
[60] J. R. Landis and G.G. Koch, "The Measurement of Observer Agreement for Categorical Data", Biometrics, Vol. 33, No. 1, pp. 159-174, 1977.
[61] Weka 3: Data Mining Software in Java, Available: http://www.cs.waikato.ac.nz/ml/weka/