Analyze Students Performance of a National Exam Using Feature Selection Methods

Hanieh Zehtab Hashemi*, Parvane Parvasideh†, Zahra Hasan Larijani† and Fatemeh Moradi†
*The faculty member of Virtual University of Medical Science, Tehran, Iran
Email: zehtab@behdasht.gov.ir
†Center of Medical Education Evaluation, Ministry of Health and Medical Education, Tehran, Iran

***Abstract*—Recently, educational institutions are generating the mass of data and interesting to analyze these data for their applications. This purpose is achieved by data mining methods to extract knowledge required by the systems. This kind of dataset is usually huge and include many samples and unnecessary features. The nature of dataset implies that the analysis of data leads to inaccurate results without preprocessing. In this study, we want to find and evaluate the most important features by different feature selection methods. These methods give different results based on their nature. Therefore in the following, we evaluate obtained feature subsets with applying some machine learning methods. Here we use one educational dataset of an exam and want to construct a reliable model to predict the final outcome of this exam. We survey different feature selection and machine learning algorithms and find out the Information Gain and Gain Ratio yield better performance.**

keywords: Educational data mining, Feature selection, Prediction accuracy, Exam, Kappa statistic, F-measure.

1. Introduction

The educational institutes play an important role in the future of the world's scientific fields. The students must pass the exams to reach the higher level in education. These systems generated a huge amount of data that involve students information categorized as demographic, academic, and lifestyle data. Educational data mining (EDM) becomes a new growing area of study that focuses on the analysis of a large educational dataset to develop models which explore hidden patterns of knowledge behind their complex structure [1]. EDM applications comprised of assessing students' learning performance, guide students' learning, adapt learning recommendations, and evaluate learning materials and so on. The researchers have been developed data mining methods to address such problems [2].

The prediction of student performance is one of the more interest work in this area which is useful for the educational institute to identify students' level. These could be beneficial to students that focus on important points and improve weakness during their study [3]. These could be performed with classification and regression models of data mining. But sometimes seen the educational data are huge and contains a large number of features. Applying data mining algorithms on such a large data would reach to a model with high computational complexity. Usually, some of the features are redundant and have no effect on modeling, also the performance of the prediction model depends on the choice of relevant features. This issue can be resolved by applying feature selection techniques on the dataset in order to discover important features. It is shown that using feature selection influence the prediction accuracy [4].

Our focus is on a national exam that which is as a bridge to pass students to the higher level. This exam is a two-step process which students participate in the exam then based on given score of this exam and acceptance law in the exam, they have accepted to spend the higher level. There is an intense competition among students to admission due to the low capacity of universities relative to the number of students. The volunteers study hard during the current year and thus it is important for them to achieve the desired result. Therefore planner of this exam must be aware of the total aspects of this exam. Using the predictive model reveal the characteristic of students that reach each outcome. Thus, this need is felt to design and develop a forecasting system to analyze students learning behavior and discover probability issue in educational systems [5]. Thus our objectives are identifying the most important features and then evaluate them using some classifiers using the Kappa statistic and F-measure criterions more efficiently. We investigate some feature selection algorithms on this dataset and evaluate the quality of these feature subset by applying some machine learning algorithms. The constructed model can reveal new patterns of knowledge of dataset to help students improve their situations to passing the exam.

The remainder of this paper is organized as follows. In section 2 we review the feature selection methods. Section 3 presents the details of the dataset and the methods utilized. Section 4 explains results and survey them in the viewpoint of prediction accuracy.

2. Feature Selection

The feature selection is one type of dimension reduction techniques that transfer data to lower space with more knowledge. In fact, it is the process of choosing the subset

Corresponding author: Parvane Parvasideh (email: parvane.parvasideh@gmail.com).

978-1-5386-9569-2/18/$31.00

of features by eliminating irrelevant ones to use in model construction. Often available features might overlap each other and/or are considered extra. The aim of dimension reduction is to find a new subset of features from original features which having most information from input data. In this approach, it is possible to lose some information that might have positive or negative influence, however in some cases, dimension reduction techniques eliminate noise [6].

The feature subset must be evaluated to describe the target of the problem, carefully and without loss of information. There are several advantages to use feature selection such as reduce computational complexity, and interpret model easily. The importance of features become obvious when using these to construct a predictive model [7]. The feature selection method is combinations of a search technique for proposing new feature subsets, along with an evaluation measure which provides weight to features based on a certain criterion [8].

These algorithms assign a weight to features by specific evaluation measure. The choice of evaluation metric make three categories namely Wrapper, Filters and Embedded. Wrapper method uses a predictive model to give scores to features. Each new subset is used to train model and the number of the mistakes determine the score to the subset of features [9]. Filter method uses some measure on the property of the data to score the feature subset. Filter methods have also been used as a preprocessing step for wrapper methods. The most known Filter based feature selection algorithms are Chi-Square attribute evaluation, Gain Ratio attribute evaluation, Information Gain attribute evaluation, Relief attribute evaluation, ... [10]. Embedded method perform feature selection as part of the model construction process. For example, LASSO method used a 1_1-norm as penalty function on the coefficients of the linear model. This cause to some of the coefficient becomes zero and features which correspond to the non zero coefficients are selected. [11]. It is reported that the use of Filter methods are faster than Wrappers, but Wrapper achieves better results than Filter.

The researchers use this methodology to resolve proposed problems in this area. For example, Thangaraju uses Genetic algorithm, Particle Swarm Optimization techniques, and Correlation feature evaluation for evaluation and then used Naive Bayes classifier to evaluate them [12]. Nguyn et al. used Information Gain technique to predict the performance of students using decision tree and Bayesian algorithm [13]. Ramaswami and Bhaskaran have used Correlation feature evaluation to determine important features for passing an examination [14]. Acharya and et al. used different feature selection algorithms on a student data from graduate students in Computer Science of University of Calcutta then classification algorithms applied on this feature subset to predict student grades [15]. Ramaswami and Bhaskaran perform a comparative study of six Filter feature selection algorithms to find which reach to the optimal dimensionality of the feature subset. The results show that predictive accuracy increased by reduced features [16].

3. Dataset and Algorithms

This study is based on an educational dataset of a national exam which is held annually in Iran and about 14,000 students compete together. This dataset includes enrolment information of students and examination result. The volunteers of this exam whom passed an elementary step in university and will be to spend higher level. The students must take admission to a university to can pass higher level in education. These people have different properties as academic information, demographic, and lifestyle data. These difference lead to their distinct performance in the examination.

This exam is a two-step procedure: at first, the exam is taken to produce a raw score which is one of the features. This exam contains 200 four-choice question. It will belong 3 positive scores to each correct answer and one negative score to each incorrect answer. The volunteers who gain at least 150 grade of the total raw score, can perform selections of fields of study in favorite order. The absence people are removed from next step of the exam. Also, the people who have not acceptable score(at least 150) can not participate in next step. Here, we remove the people who absence in the exam and also who gets the score lower than 150. Therefore the dataset become a smaller dataset contain information of 7723 of permissible volunteers, each includes 20 features contains Score, Average, Pre-internship grade and so on. The label class is a binomial variable indicate that students passed the second step of the exam or failed. The features of this dataset are listed in table 1.

Then in the second step, the students have the acceptable score, can select the favorite fields in some universities respectively up to a hundred choice. Therefore our goal is to develop a prediction model to simulate the acceptance process of students to identify which students can obtain apply of their choices. The acceptance process will be performed by considering acceptance law and volunteers' score, quota, and then the priority of selection of study fields.

The features of the dataset are now discussed. The 'Quota' is an important factor because volunteers compete with each other according to quota type in the acceptance process. In fact, it is a privilege that belongs to some people that have certain conditions. This option facilitates acceptance process of these students. This is a multi-value variable due to the existence of some type of quota. The students don't have any quota, can be identified by a neutral value that denoted as 1. This feature has some subtype such as 'Fighter quota', 'Elite quota', 'Deprived state', and 'Extra Deprive'. For example, 'Deprived state' is an option of quota which gives privilege to qualified students to facilitate accept process based on specified limitations. This is possible that 'Sex' can be effective in modeling. The University of the previous level of students determines educational background. The next feature, 'Student-Exam', reflects the condition of the exam site. The male students must identify the 'Military' status. Also, all of the students must be spend training commitment under 'Internship' sta-

tus which is a service term that students must spend it after graduating in the elementary stage. 'Booklet Code' represents the type of question paper of students. 'Score' is the primary outcome of this exam. 'Pre-internship' and 'Average' are given grades from the previous level during the study that indicated the quality of students' academic performance. 'Military staff' is a code represents military staff status of students to can use another privilege named scholarship that facilitates acceptance of students, also. The scholarship has two type: 'Military staff scholarship' and 'Army scholarship' which students have certain codes of Military staff, can benefit these options. The living place of students may be effective in admission which reviewed as 'Reside'. The work conditions of students as 'Work state' and 'Workplace' must be affected in the acceptance process.

All these features are used to predict admission of students. This model estimate outcome of the problem as a binomial variable: 1 indicate admit and 0 denote reject in the acceptance process. We will guess the students that get the acceptable score of this exam, have accepted in the second stage of exam based on their features and which features has the most effect in the acceptance process.

In these study six feature selection methods include four Filter method and two Wrapper method are used to find out the better feature subset. In following four Filter methods are described then two Wrapper methods are stated. The Chi-Square attribute evaluation, $\chi(F, k)$, evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. Information Gain attribute evaluation which known as Mutual Information, measure how much information a feature gives us about the class based on the reduction in entropy.

In Gain Ratio attribute evaluation each attribute is assigned a score where the score is delineated by means of the difference of attributes entropy and its class conditional entropy. Relief attribute evaluation is a Filter based feature selection method which is based on the identification of feature value differences between the nearest instance of the same and different class.

We used here two Wrapper feature selection algorithm: C4.5 and Naive Bayes. C4.5 is a decision tree algorithm which use information gain to give a score to features [17]. Naive Bayes is a classification algorithm for calculating the probability of each input for each class.

4. Results

The experimental methodology that implements here is as follows. The optimal set of features are extracted for Filter and Wrapper feature selection methods. As expected, each of these methods gives different features due to the underlying algorithms. These algorithms with the search methods are shown in Table 2. Then the machine learning algorithms are used to identify the effectiveness of these subsets of features.

We used these methods to obtain a subset of features in descending order of their importance. The accuracy of this set is evaluated by a classifier. We use C4.5 as the classifier and present the performance of these methods by two measurements: The Kappa and F-measure. The kappa statistic is a measure of how closely the instances classified by the machine learning methods matched with the actual class. F-measure that is a combination of Precision and Recall, means how many instances it classifies correctly, as well as how many instances classifies incorrectly. This creates a balance between Precision and Recall. For each method, we considered a number of features based on their order of importance, then calculate the Kappa and F-measure.

We perform these experiments in WEKA 3.6, which is a free software to data analysis. We record the total of process of finding maximum Kappa statistic and F-measure for the Chi-square algorithms in Table 3. This process is continued for other Filter methods and the results are shown in Table 4. From these results, we found out that the Kappa and F-measure have the maximum value in same feature subset. Also, two of these algorithms reach to same feature subset. Based on Kappa statistic and F-measure, we found that Chi-square and Information Gain algorithms obtain the better result than other. However, we use all of them to build a predictive model.

Now we perform the same methodology that follows for Filter methods for Wrapper methods. Here, we use two machine learning algorithm. C4.5 is a Decision tree algorithm. Also, Naive Bayes belong to the Bayesian learning algorithm. We extract feature subset that maximizing the Kappa and F-measure from these algorithms and record results in Table 5. From these results, we found that C4.5 and Naive Bayes obtain same results and also this result has been obtained by Gain Ratio algorithm.

Therefore, we have three different feature subset that obtained by discussed feature selection algorithms. We renamed them to Information Gain-6, Gain Ratio-14, and Relief-9. We will find the best feature subset by finding the predictive accuracy of these subset. We use four machine learning algorithms to evaluate of these feature subset: CART that is decision tree algorithm [18], Naive Bayes from Bayesian learning algorithm, MLP that belong to Neural Network and 1NN is a lazy learning. The each of these algorithms performs as different as other algorithms, thus we use different algorithms to finally compute the average these results. we used four-fold cross-validation to partition data to train and test set. The results of using four machine learning algorithm without feature selection are reported in Table 6.

The results show that Information Gain and Gain Ratio gives the best results with 6 and 14 features respectively. The best features based on Information Gain are 10,11,6,12,5,3 in importance order, i.e. Score, Pre-Internship, Internship, Average, Military, and Student-Uni. The Gain Ratio obtains better performance in 10,19,11,6,16,12,18,5,13,14,8,20,3,2, i.e. Score, Elite quota, Pre-Internship, Internship, Work state, Average, Workplace, Military, Military staff, Military staff scholarship, Deprived state, Army scholarship, Student-Uni, and Sex. The 'Score' is the most important feature and it is natural to find it as the first feature. The Pre-Internship and Average are important due to relying on

TABLE 1. FEATURES OF DATASET

Feature Number	Feature Name	Description	Domain
1	Quota	student's quota type	$\{1, 4, 5, 6\}$
2	Sex	student's sex	$\{0 : man, 1 : woman\}$
3	Student-Uni	student's university	categorical
4	Student-Exam	student's test site	categorical
5	Military	student's military service status	categorical
6	Internship	student's internship status	categorical
7	Fighter quota	student's fighter quota type	$\{1, 11\}$
8	Deprived state	student's deprived state code	categorical
9	Booklet code	student's exam booklet code	$\{1 : A, 2 : B, 3 : C, 4 : D\}$
10	Score	student's raw score	numeric
11	Pre-internship	student's pre-internship grade	numeric
12	Average	student's average grade	numeric
13	Military staff	student's military staff status	categorical
14	Military staff scholarship	student's military staff scholarship status	$\{1, 11\}$
15	Reside	city of reside	categorical
16	Work state	student's working condition	$\{1, 11\}$
17	Extra Deprive	Extra Deprive capacity	$\{10, 11\}$
18	Workplace	student's working place	categorical
19	Elite quota	student's elite quota type	$\{1, 11\}$
20	Army scholarship	student's army scholarship status	$\{1, 11\}$
Class Label			
	Accept	pass or fail	$\{1 : pass, 0 : fail\}$

TABLE 2. FEATURE SELECTION ALGORITHMS

Type of feature selection algorithms	Algorithms	Search methods
Filter Based Feature Selection Algorithms	Chi-Square feature evaluation	Rank search
	Information Gain feature evaluation	Ranker
	Gain Ratio feature evaluation	Ranker
	Relief feature evaluation	Ranker
Wrapper Based Feature Selection Algorithms	C4.5	Rank search
	Naive Bayes	Rank Search

TABLE 3. FEATURE SUBSET OF CHI-SQUARE

Step	Feature subset	Kappa statistic	F-measure
Initialization	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20,9,1,14,7,17	0.6505	0.846
1	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20,9,1,14,7	0.6505	0.846
2	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20,9,1,14	0.6505	0.846
3	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20,9,1	0.6505	0.846
4	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20,9	0.6573	0.848
5	10,11,6,12,5,3,18,8,4,15,16,19,13,2,20	0.657	0.848
6	10,11,6,12,5,3,18,8,4,15,16,19,13,2	0.657	0.848
7	10,11,6,12,5,3,18,8,4,15,16,19,13	0.6445	0.843
8	10,11,6,12,5,3,18,8,4,15,16,19	0.6722	0.853
9	10,11,6,12,5,3,18,8,4,15,16	0.6722	0.853
10	10,11,6,12,5,3,18,8,4,15	0.6722	0.853
11	10,11,6,12,5,3,18,8,4	0.6734	0.854
12	10,11,6,12,5,3,18,8	0.6745	0.855
13	10,11,6,12,5,3,18	0.6802	0.857
14	10,11,6,12,5,3	0.6837	0.859
15	10,11,6,12,5	0.6628	0.848
16	10,11,6,12	0.6628	0.848
17	10,11,6	0.6594	0.847

the educational background of students. 'Internship' is the situation of internship that these students must spend. The Elite quota, Deprived state, Military staff scholarship and Army scholarship are some beneficial features that facilitate the acceptance process and therefore appear as important features.

5. Conclusion

In this paper, we performed an analysis on the educational dataset of a national exam by different feature selection methods to find out the important feature subsets. These methods assign own weights to features and specify features with major roles. Then we evaluate these subsets

TABLE 4. FEATURE SUBSET MAXIMIZING KAPPA STATISTIC AND F-MEASURE FOR FILTER METHODS

Algorithms	Chi-Square	Information Gain	Gain Ratio	Relief
Number of features	6	6	14	9
Feature subset	10,11,6,12,5,3	10,11,6,12,5,3	10,19,11,6,16,12,18,5,13,14,8,20,3,2	10,6,18,3,8,16,4,2,11
Kappa statistic	0.6837	0.6837	0.682	0.6734
F-measure	0.859	0.859	0.858	0.854

TABLE 5. FEATURE SUBSET MAXIMIZING KAPPA AND F-MEASURE FOR WRAPPER METHODS

Algorithms	C4.5	Naive Bayes
Number of features	14	14
Feature subset	10,19,11,6,16,12,18,5,13,14,8,20,3,2	10,19,11,6,16,12,18,5,13,14,8,20,3,2
Kappa statistic	0.682	0.682
F-measure	0.858	0.858

TABLE 6. EVALUATE THE FEATURE SUBSET

Feature subset	F-measure				Average F-measure
	CART	NB	MLP	1NN	
Information Gain-6	0.834	0.821	0.823	0.76	0.8095
Gain Ratio-14	0.835	0.819	0.832	0.752	0.8095
Relief-9	0.836	0.822	0.833	0.74	0.80775
Without FS	0.838	0.818	0.836	0.69	0.7955

with some machine learning methods. The results show the effective features to accept in this exam. This guide volunteers and planners of this exam to can better decision to improve their performance. Also, using feature selection increase accuracy of prediction of such problem.

Acknowledgments

This project was funded by the National Agency for Strategic Research in Medical Education. Tehran. Iran. Grant No.961787. This research supported by the Center of Medical Education Evaluation, Ministry of Health and Medical Education, Iran. We appreciate Mr. Tourani and the reviewers of this article.

References

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM— Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[2] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker, *Handbook of educational data mining*. CRC press, 2010.

[3] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in Human Behavior*, vol. 47, pp. 168–181, 2015.

[4] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15 991–16 005, 2017.

[5] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[6] M. B. Christopher, *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.

[7] B. Ratner, *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data*. CRC Press, 2011.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[9] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[10] J. Stefanowski, "An experimental study of methods combining multiple classifiers-diversified both by feature selection and bootstrap sampling," *Issues in the Representation and Processing of Uncertain and Imprecise Information*, pp. 337–354, 2005.

[11] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.

[12] T. Karthikeyan and P. Thangaraju, "Genetic algorithm based cfs and naive bayes algorithm to enhance the predictive accuracy," *Indian Journal of Science and Technology*, vol. 8, no. 26, 2015.

[13] N. T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*. IEEE, 2007, pp. T2G–7.

[14] M. Ramaswami and R. Bhaskaran, "A chaid based performance prediction model in educational data mining," *arXiv preprint arXiv:1002.1144*, 2010.

[15] A. Acharya and D. Sinha, "Application of feature selection methods in educational data mining," *International Journal of Computer Applications*, vol. 103, no. 2, 2014.

[16] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *arXiv preprint arXiv:0912.3924*, 2009.

[17] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.