مجلّة جامعة عدن للعلوم الانسانيّة والاجتماعية

EJUA-HS Vol. 5 No. 2 (2024)

https://doi.org/10.47372/ejua-hs.2024.2.360

ISSN: 2708-6275

RESEARCH ARTICLE

DISCOVER THE ATTRIBUTES AFFECTING GRADE OF UNIVERSITY STUDENT THROUGH FEATURE SELECTION METHODS

Roiss Mohammed Salem Alhutaish[1,*], Anees Abdullah Shafal Ali[2], and Khalid Kaied Shafal Ali[3]

[1]Dept. of Information Technology, Aden Community Aden, Aden- Yemen.
[2]Dept. of computer Science, Faculty of Computer & Information Technology, Abyan University, Abyan- Yemen
[3]Dept. of computer, Faculty of Education, Aden University, Aden- Yemen

***Corresponding author: Roiss Mohammed Salem Alhutaish; E-mail: roiss2000@hotmail.com**

Received: 29 March 2024 / Accepted: 28 May 2024 / Published online: 30 June 2024

Abstract

There are many techniques to choose subset features. This stage is the preprocessing of the data mining task. The process of choosing a subset of features from the original collection of attributes is known as feature selection. Obtaining relevant information to predict characteristics affecting the estimation of the CGPA of graduates of private universities is a hard task. The experiments were conducted on four datasets. Two methods are used in order to choose the features. They are CfsSubsetEval and Correlation Ranking. Correlation Ranking uses 0.1 as the threshold for feature selection. The results indicate that there is a relationship between a student's grade in the previous qualification and his grade when he graduates from the university.

Keywords: Feature Selection, CfsSubsetEva, Correlation Ranking, Cumulative GPA.

1. Introduction

Feature selection is a technique used to find a minimum subset of features. It is the process of identifying and removing as much irrelevant and redundant information as possible. This method makes a small dataset, which speeds up and improves the efficiency of learning algorithms. Finding a feature subset that can describe the data for a learning task as well as, or better than, the original dataset is the goal of feature selection. Feature extraction and feature selection are the two techniques for reducing the dimension. Without utilizing previous knowledge, feature extraction projects or transforms original features into fewer dimensions. However, it is not comprehensible and makes use of all the original features, which might be problematic in feature spaces with lots of features. Conversely, feature selection removes redundant and irrelevant features from the original features in order to identify the best feature subsets. It has the ability to reduce complexity, increase classification accuracy, speed computation, and improve comprehensibility by preserving the original semantics of datasets. Feature selection can be divided into four categories: filter, wrapper, hybrid, and embedded methods (Duangsoithong & Windeatt, 2009; Liu & Yu, 2005; Saeys, Inza, & Larranaga, 2007).

The educational process is cumulative over the years. In the last decade, there has been skepticism about students' grades in the pre-college degree, particularly the high school diploma. This research took a real data set from the Ministry of Higher Education. It consists of student information at the time of high school graduation and student information during university studies. Feature selection was used in order to prove the hypothesis that the outcomes of secondary education are still the basis of university education. In addition, discovering other characteristics that influence the student’s assessment when he graduates from the university. This paper suggests using feature selection in order to know what features are most closely related to the class. It is organized into the following sections: In the next section, related works are described. Section III contains methodology, where the feature selection methods used in the experiments were described. Section IV presents the experiments and discusses the results, while Section V offers a summary of the conclusions.

2. Related works

Data mining is a technique used to search for useful information in large data sets, with techniques such as association rules, classification, clustering, prediction, and sequential models. There are many strategies that be developed and implemented in order to transform a wealth of information into a wealth of knowledge predictability (Phyu & Oo, 2016). The feature selection method is used in educational institutions to categorize student performance. This method offers a number of advantages as removing redundant features and also eliminating the irrelevant features. J48 and prism with feature selection methods are increasing of accuracy. Moreover, feature selection can help students forecast their failure and dropout rates more accurately. (Duangsoithong & Windeatt, 2009) conducted a comparative analysis of six filter feature section algorithms in order to determine the best approach and the ideal feature subset's dimensionality. many classifier models use to conduct benchmarking of the filter feature selection method. The present study's results effectively supported the widely recognized idea that the presence of fewer features increases predictive accuracy. the results show decrease in computing time and construction costs in the student performance model's training and classification stages. (Zaffar, Hashmani, Savita, & Rizvi, 2018) analyze of performance f various feature selection algorithms through using two different datasets. The findings showed that feature selection algorithms with datasets containing varying numbers of features performed significantly differently; accuracy percentages varied by 10 to 20 percent. The performance of the filter feature selection method reduces when number of feature increases. this study predicts the academic performance of the student based on wrappers feature selection techniques. they plan to assess the feature selection outcomes using confusion in the future. they also can't ignore the benefits of filter feature selection methods. In order to predict student performance, a few hybrid feature selection techniques could be applied to student datasets in the future to improve the study. they presented a study that revealed students attribute their academic success to both internal and external factors, some of which can be controlled and used intentionally to improve their performance. These factors can be dysfunctional, such as luck or difficulty with homework. The research suggests that understanding students' self-attributions can help develop educational intervention strategies that improve motivation and achievement. The study also highlights the importance of motivational and affective aspects in the teaching-learning process. Future research should compare findings with students at public universities, considering social and cultural contexts and visions of education. Other research paths include extrinsic-relational attributions like teacher support, family, friends, or school organizational structure. Convergence validity assessments with multi-item instruments of scholastic causality attributions are recommended to reinforce the use of single-item measurements.

(Saleh, Saedi, al-Aqbi, & Salman, 2020) evaluates some technique a significant role in disease detection in the healthcare industry, particularly in predicting and classifying cardiovascular diseases like heart. Data mining algorithms like K-star, J48, SMO, Naïve Bayes, MLP, Random Forest, Bayes Net, and REPTREE are used to predict heart problems. However, previous studies have limitations in accuracy and feature number. This paper surveys recent data mining techniques for predicting heart diseases and identifying major risk factors. In addition, a lot of research has been done to predict diseases, and one of the deadliest diseases is cancer (Rao, Gladence, & Lakshmi, 2019). This paper discusses the application of various feature selection and classification methods to early-stage breast cancer prediction. They use the WEKA tool to process breast cancer data from the UCI repository, revealing SVM as the best classification algorithm for early disease prediction.

(Phyu & Oo, 2016) proposed new mutual information for the select subset feature. The algorithm's effectiveness is evaluated using standard datasets from UC Iravine and WEKA, considering classification accuracy and the number of selected features. They compare the new method with three methods of feature selection. These methods are Information Gain (IG), Symmetrical Uncertainty (SU) and Relief-F. The new method of selecting of features with Naïve Bayes and J48 classifiers, shows that it is effective to select features. Moreover, its accuracy results are better.

For years, researchers in the fields of data mining, machine learning, statistics, and neural networks have given feature selection and ensemble classification a great deal of attention (Duangsoithong & Windeatt, 2010). The correlation-based and causal feature selection for ensemble classifiers using MLP and SVM, Naive Bayes, and Decision Tree are proposed by Duangsoithong and Windeatt (2010). They compare these methods. Results show correlation-based algorithm improves accuracy and reduces complexity, while ensemble using Bagging algorithm enhances accuracy.

3. Methodology

In our previous work, we proposed the article in order to discover the relationship between features through association rules (Alhutaish, Ali, & Ali, 2023). This article uses feature selection methods to discover the relationship between features and labels. Filter-based feature selection is used. Using the Waikato Environment for Knowledge Analysis, features were chosen. Weka uses its Select attributes module to carry out feature selection. There are many methods in this field. FsSubsetEval, or Correlation-based feature selection method (CFS) is concerned with the hypothesis that features are related to class, but there is no correlation between the features (Hall, 1999).

Two stages are used in this article (Figure 1 shows all phases): Data Preprocessing and Feature Selection. In stage data preprocessing, we obtain many files of many years in Excel format, where the data processing process are as follows: first step is data cleaning where remove incomplete, incorrect, inaccurate and duplicate data from the dataset. This step done on all files separately. Second step is Data integration where each university's files were grouped together to form an independent dataset. Third step, all these files are transformed through data transformation. In this step, data convert from text data into Nominal data. The second Stage use weka tools in order to selection attributes. In this stage, many steps use in order to know the features related to each label. First, cumulative grade point average (CGPA) is used as label. The second step is determining the components Attribute Evaluator. The third step is to choose the search method that is compatible with Attribute Evaluator. The third step is to determine the criterion value for research methods that require a selection criterion.

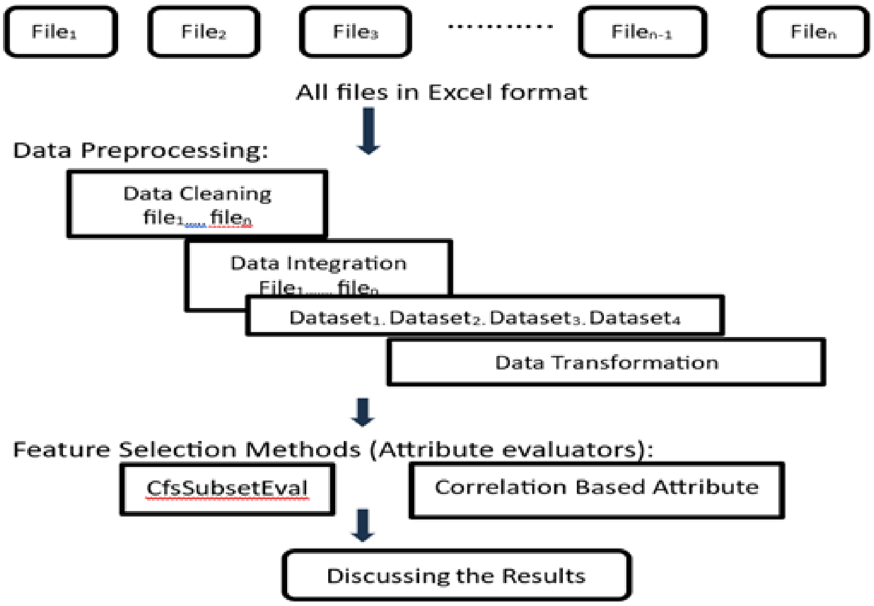

Figure (1): The Methodology

3.1. Attribute evaluators

Basically, an attribute evaluator is used to rank each feature based on a specific metric. In WEKA, there are different attribute evaluators available. This work uses two types of attribute evaluators: CfsSubsetEval and Correlation Based Attribute evaluation.

a. CfsSubsetEval:

Considers the degree of redundancy between features and each feature's individual ability to determine the value of a subset of attributes. Preferred feature subsets are those with a low intercorrelation with other attributes and a strong correlation with the class.

b. Correlation Based Attribute Evaluation:

The correlation between each attribute and the target class attribute is measured using Pearson's correlation method. It bases its consideration of nominal attributes on value, with each value serving as an indicator.

3.2. Search Methods:

To find the optimal set of features, these methods search the entire set of possible features. In this work, two search techniques—BestFirst and Ranker, which are available in Weka—are compared.

a. BestFirst:

It looks for the state of a subset of the BestFirst attribute with a backtrack by greedyhillclimbing. Specifying the number of consecutive non-healing nodes allowed controls the level of rollback performed. The BestFirst can start with an empty set of attributes and search forward, or start with a full set of attributes and search backward, or start anywhere and search both ways (considering all possible additions and deletions of individual attributes at a given position).

b. Ranker:

The Ranker approach ranks attributes based on their individual assessments, sets a threshold, or indicates how many attributes to keep. This paper uses 0.1 as the threshold for choice of attributes where the attribute with a correlation coefficient higher than the threshold is selected.

4. Experiments and Results

The experiments use the dataset obtained from the Private Education Administration. Four private universities provided to High Education information about graduate students. The experiments have gone through several stages. One of these stages is the stage of preparation, where the features were coded as follows: Faculty (F), Specialty (S), Graduate Year (GY), Previous Certificate (PC), Year of Previous Certificate (YPC), GPA of Previous Certificate (GPA). In addition, there are four cases of class that represent cumulative grade point average (CGPA). They are Class 1 (Excellent), Class 2 (very good), Class 3 (good) and Class 4 (Acceptable).

According to Table 1, CorrelationAttributeEval with ranker reduced the number of features to 80–90%, while CfsSubsetEval with BestFirst got up to 46% reduction.

Table (1): List of features selected by CfsSubsetEval + BestFirst & CorrelationAttributeEval + Ranker

Datasets	Instances	Total Features	Number of selected Features by CfsSubsetEval + BestFirst	Number of selected Features by CorrelationAttributeEval + Ranker
First	3413	59	27	6
Second	1103	49	21	6
Third	770	47	19	6
Fourth	514	38	13	8

Table 2 shows the results of the data set for graduates of the First University, showing that there is a correlation between the grades of previous qualification and the grade at graduation. Students who obtained excellent grades in the previous stage also obtained excellent grades when they completed their graduate studies. Moreover, most students graduating in 2017 and 2021 have excellent grades. It has also been observed that students whose grades were acceptable in the previous certificate had excellent, very good, and good grades at the university.

Table (2): Results of first dataset

Class	Features selection by CfsSubsetEval + BestFirst	Features selection by CorrelationAttributeEval + Ranker
Class 1 (Excellent)	S5, S24, S25, S12, S11, S22, S28, GY17, GY21, YPC1, YPC16, GPA4, GPA1	GY21, GY17, GPA1, GPA4, GPA2
Class 2 (very good)	F1, S4, S21, S25, S26, S16, S8, S28, GY17, PC0, YPC6, YPC5, YPC15, YPC20, GPA4	GY17, GPA4
Class 3 (good)	S25, S26, S12, S11, S8, S28, GY17, GY21, PC0, YPC13, GPA4, GPA1	GY21, GY17, GPA1, GPA4, GPA2, F1
Class 4 (Acceptable)	S25 S26, S16, S20, GY17, PC0, YPC7, YPC1, YPC17, GPA4	GY17, GPA4

Experiments on the data set of students at the second university are shown in Table 3. Excellence students in the previous stage received various grades in university, with the exception of very good, which only a small number of university students obtained. In addition, there is a gap in the results of graduates of the Faculty of Science and Engineering, as experiments showed that graduates obtained distinction or good.

Table 4 shows the results for the third university graduates' dataset. Most of the students who had excellent grades in their previous degree also had excellent grades when they graduated from university. Experiments also showed that most of the students whose grades were acceptable also had their grades accepted at the university.

Table (3): Results of second dataset

Class	Features selection by CfsSubsetEval + BestFirst	Features selection by CorrelationAttributeEval + Ranker
Class 1 (Excellent)	F3, YPC4, YPC9, YPC16, YPC17, GPA1	GPA1, GPA3
Class 2 (very good)	S9, S12, S14, GY11, GY16, GY18, YPC1, GPA2, GPA3, GPA4	GPA2, GPA3, GPA4, GY16
Class 3 (good)	F3, S11, YPC4, YPC2, YPC17, GPA3, GPA1	GPA1, GPA2, GPA3
Class 4 (excepted)	S7, S12, GY19, YPC12, YPC2, GPA4, GPA1	GPA1, GPA2, GPA4, S7

Table (4): Results of third dataset

Class	Features selection by CfsSubsetEval + BestFirst	Features selection by CorrelationAttributeEval + Ranker
Class 1 (Excellent)	S10, YPC9, YPC2, YPC18, GPA4, GPA1	GPA1, GPA3, GPA4, S9, S10
Class 2 (very good)	S10, S19, YPC7, YPC0, YPC4, YPC16, YPC15, YPC18, GPA2, GPA4	GPA2, GPA4
Class 3 (good)	S9, S10, YPC7, YPC3, YPC4, YPC18, GPA3, GPA1	GPA1, GPA3, S9, S10
Class 4 (excepted)	S17, PC1, GPA4	GPA2, GPA4

The results of graduates in the faculty of Sharia showed a logical distribution of graduates' estimates, as shown in the graduates' dataset in Table 5 for the fourth university. And also, the results of graduates of the Faculty of Engineering and Information Technology have a logical distribution. Most of the graduates' grades were very good grad and good grad. This emphasizes that the level of study in the Faculty of Engineering and Computer Science is very high and it is not easy for a student to get an Excellent.

Table (5): Results of fourth dataset

Class	Features selection by CfsSubsetEval + BestFirst	Features selection by CorrelationAttributeEval + Ranker
Class 1 (Excellent)	F1, S2, GY18, GY20, YPC8, YPC0, YPC17, GPA1	GY20, F2, S2
Class 2 (very good)	F1, F3YPC7, YPC8, YPC6, YPC17, YPC4, GPA1	F1, S1
Class 3 (good)	F1, F3, S1, GY21, YPC9, YPC7, YPC8, YPC0, YPC6, YPC12, YPC4	F3, S3
Class 4 (excepted)	GY21, YPC7, YPC0, YPC6, YPC4, GPA1	GY20, F2, S2

The experiments result of the most dataset show that high school results are still reliable. there are many cases of the students that obtain excellent in high school and also obtain excellent in the university. Furthermore, students who obtained Acceptable grade in high school can earn an Excellent, very good, or good in university, which is another indication of the reliability of high school results. In addition, there are students who received high grades in high school and whose grades were divided into high, medium, and low grades. This is due to the student's circumstances and interest and cannot be projected onto the level of high school results because many results show positive high school results. In all cases, there was a correlation between the distinction grade in the previous certificate and the distinction grade upon graduation from the university, with the exception of the Ranker method with the fourth dataset.

5. Conclusion

This article suggests a feature selection technique to determine the characteristics that affect the estimate of the cumulative GPA of graduates from private universities. Features include information about a student while studying at the previous qualification level as well as information about a student while studying at the university level. In addition, the class represents the cumulative grade point average (CGPA) when graduating from the university level. Experiments have shown that there is a correlation between a student's grade at graduation and the grade in the prior qualification. Moreover, some years of graduation as well as some faculties and specializations have shown excellence in the cumulative grade point average (CGPA) at graduation.

References

[1] Alhutaish, R., Ali, A., & Ali, K. (2023). The relationship between previous qualification characteristics and university graduates using association rules. Electronic Journal of University of Aden for Humanity and Social Sciences, 4, 673-678. doi:10.47372/ejua-hs.2023.4.316

[2] Duangsoithong, R., & Windeatt, T. (2009). Relevance and redundancy analysis for ensemble classifiers. Paper presented at the Machine Learning and Data Mining in Pattern Recognition: 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009. Proceedings 6.

[3] Duangsoithong, R., & Windeatt, T. (2010). Correlation-based and causal feature selection analysis for ensemble classifiers. Paper presented at the Artificial Neural Networks in Pattern Recognition: 4th IAPR TC3 Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings 4.

[4] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(4), 491-502 .

[5] Phyu, T. Z & ,.Oo, N. N. (2016). Performance comparison of feature selection methods. Paper presented at the MATEC web of conferences.

[6] Rao, K., Gladence, L., & Lakshmi, V. (2019). Research of feature selection methods to predict breast cancer. International Journal of Recent Technology and Engineering (IJRTE), 8, 2353-2355 .

[7] Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517 .

[8] Saleh, B. J., Saedi, A. Y. F., Al-Aqbi, A. T. Q., & Salman, L. a. (2020). A review paper: Analysis of weka data mining techniques for heart disease prediction system.

[9] Zaffar, M., Hashmani, M. A., Savita, K., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. International Journal of Advanced Computer Science and Applications, 9 .(5)

مقالة بحثية

اكتشاف السمات المؤثرة في تقدير الطالب الجامعي من خلال طرق اختيار الميزات

رويس محمد سالم الحتيش[1]،*، انيس عبدالله شعفل علي[2] و خالد قائد شعفل علي[3]

[1] قسم تكنولوجيا المعلومات، كلية المجتمع عدن، عدن، الجمهورية اليمنية.
[2] قسم علوم الحاسوب، كلية الحاسوب وتقنية المعلومات، جامعة أبين، أبين، الجمهورية اليمنية.
[3] قسم الحاسوب- كلية التربية، جامعة عدن، عدن، الجمهورية اليمنية.

* الباحث الممثل: رويس محمد سالم الحتيش؛ البريد الالكتروني: roiss2000@hotmail.com

استلم في: 29 مارس 2024 / قبل في: 82 مايو 2024 / نشر في: 30 يونيو 2024

المُلخّص

هناك العديد من التقنيات لاختيار ميزات المجموعة الفرعية. هذه المرحلة هي المعالجة المسبقة لمهمة التنقيب عن البيانات. تُعرف عملية اختيار مجموعة فرعية من السمات من المجموعة الأصلية من السمات باسم اختيار السمات. يعد الحصول على المعلومات ذات الصلة للتنبؤ بالخصائص التي تؤثر على تقدير المعدل التراكمي لخريجي الجامعات الخاصة مهمة صعبة. أُجريت التجارب على أربع مجموعات بيانات. تم استخدام طريقتين لاختيار الميزات. وهما CfsSubsetEval وترتيب الارتباط. يستخدم تصنيف الارتباط 0.1 كعتبة لاختيار الميزة. تشير النتائج إلى وجود علاقة بين درجة الطالب في المؤهل السابق ودرجته عند تخرجه من الجامعة ومع ذلك، أدت هذه الاستراتيجيات إلى ترجمات غير مرضية.

الكلمات المفتاحية: اختيار الميزة، CfsSubsetEval، ترتيب الارتباط، المعدل التراكمي.

How to cite this article:

Alhutaish, R. M. S., Ali, A. A. Sh., Ali, Kh. K. Sh., (2024). DISCOVER THE ATTRIBUTES AFFECTING GRADE OF UNIVERSITY STUDENT THROUGH FEATURE SELECTION METHODS. *Electronic Journal of University of Aden for Humanity and Social Sciences*, 5(2), p160-167. https://doi.org/10.47372/ejua-hs.2024.2.360