Handling High Dimensional Educational Data using Feature Selection Techniques

Amirah Mohamed Shahiri[1], Wahidah Husain[1], Nur'Aini Abd Rashid[2]
[1]*School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia.*
[2]*Department of Computer Sciences, College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University, KSA*
ams14_com044@student.usm.my

***Abstract*— Huge amounts of data in educational datasets may cause the problem in producing quality data. Recently, data mining approach are increasingly used by educational data mining researchers for analyzing the data patterns. However, many research studies have concentrated on selecting suitable learning algorithms instead of performing feature selection process. As a result, these data has problem with computational complexity and spend longer computational time for classification. The main objective of this research is to provide an overview of feature selection techniques that have been used to analyze the most significant features. Then, this research will propose a framework to improve the quality of students' dataset. The proposed framework uses filter and wrapper based technique to support prediction process in future study.**

***Index Terms*—Educational Data Mining (EDM); Feature Selection; Filter; High Dimensional Data;, Wrapper.**

I. Introduction

In recent years, the data grows larger and faster that may cause difficulties for data scientists to analyze and interpret the complexity of data for getting new knowledge. Similar problem is also faced by the educational system where its database contains a large amount of educational data [1]. These data are very useful for analytical purpose to assist educational institutions in solving various educational research issues. Generally, data mining techniques have used to analyze and predict the data pattern and trend of student's performance. This data mining technique in educational field, is referred as Educational Data Mining (EDM) [2].

However, most of EDM communities are overlooked on how to improve the data quality that may provide better results in prediction accuracy. Instead, most of them only concentrated on selecting suitable learning algorithms. Educational dataset contain many features that can influence the performance of prediction. Accordingly, problem of large amount data is likely to be seen in high dimensional of data and high computational complexity [3, 5]. Feature selections, as proposed by Singh et. al., refers to the process involving the selection of subset features from original features [5]. However, most predictive models are avoiding the appropriate techniques for selecting the best feature. As a result, the probability density function of the feature vector space is ineffective during the classification operation [6].

Realising the importance of producing the quality sets of data, this study will propose a framework that can enhance the prediction model of students' performance for gaining a better classification result. The objectives of this study are as follows:

a. To analyze, investigate and propose feature selection technique to reduce dimension data problem in educational database.
b. To propose a framework on the hybrid of feature selection techniques that can improve the quality of students' datasets in order to improve the prediction model.

II. Background Study

Educational Data Mining (EDM) is the process of extracting useful information and patterns from the huge educational database [2]. This study will be focused on the processes of producing quality data. The process of producing quality data is very important because the data quality will affect the classification process that may lead to decrease the accuracy prediction model [7]. Data quality can be produced during the pre-processing phase in data mining process. Therefore, the feature selection technique can be used to solve high dimensional data problem.

A. Feature Selection

Feature selection is one of dimensionality reduction processes in data mining that only select appropriate subset features from the original features [7]. The purpose for this feature selection is to remove the irrelevant, redundant and noisy data. When the numbers of features are reduced, the data mining performance can be improved by increasing predictive accuracy and speeding up data mining algorithm [8]. Feature selection consists of four basic steps which are subset generation, subset evaluation, stopping criterion and result validation [4].

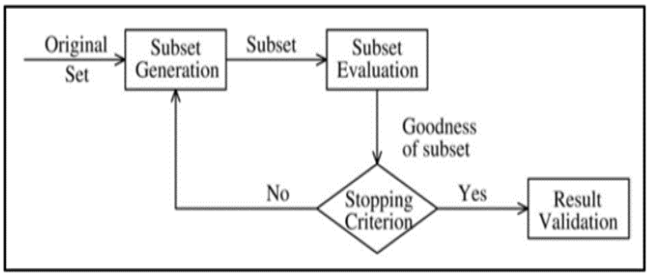

Figure 1 Basic Steps of Feature Selection Process [9].

Firstly, in subset generation, the new candidate of feature subsets is produced using search strategy. Then, the candidate subsets are evaluated and compared with previous subsets based on evaluation criterion. The previous feature subsets will be replaced if the latter-produced subset is better. Both processes are repeated until stopping criterion is satisfied. Lastly, the validation step on real data is needed to validate the best selected feature subsets. Fig. 1 demonstrates the flow of feature selection process.

In the second step, the subset evaluation for feature selection algorithm is divided into three approaches, which are filter model, wrapper model and embedded model [9]. The first model, namely filter model, is an independent measure without involving any learning algorithm for evaluating feature subset [3, 15]. The simple measurement done by filter approach may result in a simple and fast computation [11]. In the second model which is wrapper model, learning algorithm is applied for evaluating subset features. This model often produces good results through the interaction between feature subset search and classifier model [12, 13]. The last approach used in subset evaluation is embedded model. This approach is a hybrid of filter and wrapper model where it interacts with the learning algorithm in lower computational time [12, 13]. The next subsection will discuss the categorization of feature selection algorithms for educational data.

B. *Categorisation of Feature Selection Algorithms for Educational Data*

Based on the four basic steps in feature selection technique as mentioned in previous section, the feature selection algorithms can be categorized into three dimensions. They are search strategy, evaluation criteria and data mining tasks. Kumar and Liu & Yu, have proposed the three dimensional categorisation framework [11, 12]. This framework can be used as a guideline to select the suitable feature selection algorithm and also to carry out general process of feature selection technique. Table. 1. presents the previous studies that have used three dimensional categorisation framework for feature selection algorithms.

Table 1
Categorisation of Feature Selection Algorithms for Educational Data

Criteria	**Algorithm**	**Search Strategy**	**Data Mining Task**	**Author**
Filter	IGATE	Complete	Classification	[6][3][13]
	Gain Ratio			[6][13]
	Chi Square			[6][13][3]
	Symmetrical Uncertainty			[6]
	Correlation Based			[17] [5][1]
	Relief Attribute	Sequential		[6]
Wrapper	Predictive Accuracy	Complete	Classification	[14][3]
Hybrid	Filter + Wrapper	*NA*	*NA*	*NA*

Apart from the three dimensional categorisation, this study will focus on hybrid filter and wrapper approach to find the relevant features without affecting the accuracy result. This hybrid approach is supported by Uncu & Türkşen, where filter approach is a better alternative to perform the data in high dimensional space, while wrapper approach is more suitable for problems regarding accuracy [12].

As shown in Table. 1. filter method is commonly used due to its simplicity as well as being computationally less expensive. The preferable filter approach in this study will be explained in the next section under proposed framework.

III. Proposed Framework

Regarding the study on feature selection that has been discussed earlier, the proposed framework on applying feature selection technique to solve the high dimensional data in education is presented. The framework is starts with data pre-processing as shown in Fig. 2.

The objective of data pre-processing is to transform the raw data into valuable information. During this process, the feature selection techniques was applied to improve the quality of data. The two approaches used are filter and wrapper. Combining feature selection model will solve the problem of high dimensional data and provide good accuracy results [5]. The process of proposed framework is explained in the subsequent section.

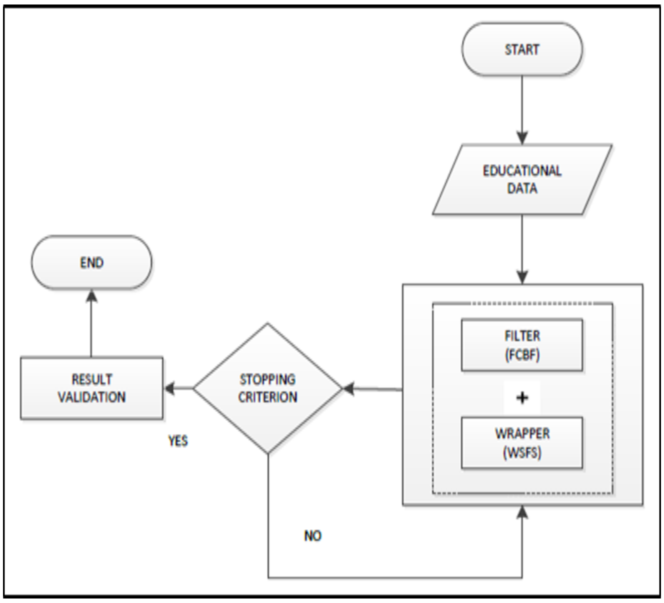

Figure 2 The Proposed Framework on Feature Selection Technique

A. *Description on Educational Dataset*

The sample dataset was collected from UCI Machine Learning Repository [15]. Around 395 of records can be maintained in data set for purpose of analyzing using feature selection techniques. The attributes that have influence of student performance are identified. The attributes include student demographic, academic information and lifestyles that were expected to affect student performance. The attributes are shown in Table 2.

The attributes selection is divided into three categories which are demographic data, academic information and student's lifestyle. Demographic data is looking as one of important features influencing students' performance. It is because in human being they have different learning style that would give different results on students' performance. While for academic information features, most of researchers reported that it has a tangible value as an indicator to identify the academic potential. Lastly, the most often attributes being used is under lifestyle category.

Previous researchers have used this attributes because when students have a lot of networking, leadership skills and active outdoor activities, it would give a positive impact on their study performance.

Table 2
Students' Data Description

Category	No	Attributes	Description	Domain
Demographic Data	1	Sex	Student's sex	Binary: 'F'-female or 'M'- male
	2	Age	Student's age	Numeric: from 15 to 22
	3	Address	Student's home address	Binary: 'U'-urban or 'R'-rural
	4	Famsize	Family size	Binary: 'LE3-less or equal or 'GT'-greater than 3
	5	Pstatus	Parent's cohabitation status	Binary: 'T'-living together or 'A'-apart
	6	Medu	Mother's education	Numeric: 0-none, 1-primary education (4th grade), 2-(5th-9th grade), 3-secondary school, 4-higher education
	7	Fedu	Father's education	Numeric: 0-none, 1-primary education (4th grade), 2-(5th-9th grade), 3-secondary school, 4-higher education
	8	Mjob	Mother's job	Nominal: 'teacher', 'health', 'service', 'at_home', 'other'
	9	Fjob	Father's job	Nominal: 'teacher', 'health', 'service', 'at_home', 'other'
	10	Guardian	Student's guardian	Nominal: 'mother', 'father', or 'other'
Academic Information	11	School	Student's school	Binary: 'GP'-Gabriel Pereira or 'MS'-Mousinho da Silveria
	12	Schoolsup	Extra educational support	Binary: yes or no
	13	Higher	Wants to take higher education	Binary: yes or no
	14	Absences	Number of school absences	Numeric: from 0 to 93
	15	G1	First period grade	Numeric: from 0 to 20
	16	G2	Second period grade	Numeric: from 0 to 20
	17	G2	Final period grade	Numeric: from 0 to 20
	18	Famsup	Family educational support	Binary: yes or no
Lifestyle	19	Studytime	Weekly study time	Numeric: 1-<15 min, 2-15 to 30 min, 3-30min to 1 hour, 4->1 hour
	20	Paid	Extra paid classes	Binary: yes or no
	21	Activities	Extra-curricular activities	Binary: yes or no
	22	Internet	Internet access at home	Binary: yes or no
	23	Romantic	With a romantic relationship	Binary: yes or no
	24	Famrel	Quality of family relationships	Numeric: 1-very bad to 5-excellent
	25	Freetime	Free time after school	Numeric: 1-very low to 5-very high
	26	Goout	Going out with friends	Numeric: 1-very low to 5-very high
	27	Health	Current health status	Numeric: 1-very bad to 5-very good

B. Fast Correlation Based Feature Selection (FCBF)

Among the three feature selection metrics, the correlation coefficient approach is more preferable for feature relevance measurement compared to information based and distances based. This is because correlation approach avoids the interdependencies between features and discretization of continuous features problem [16]. This statement is also supported by a study of Karegowda et. al., where FCBF was found capable of evaluating the worth feature subsets by weighing the individual predictive ability through the degree of redundancy between them [17]. In simple word, the FCBF identifies and eliminates the redundant features to increase the efficiency of learning process.

In educational data, previous researchers have mostly used information based for relevant measurement. As shown in Table. 2, there are only three studies available showing the use correlation based metrics in filter approach [3, 8, 19]. The FCBF obtained with minimum features was discovered to give the highest measurement value compared to information based metrics including IGATE [3].

Even though the filter approach is used for filtering out the relevant features, the resulting features are still quite large. Therefore, wrapper approaches namely wrapper sequential forward selection can be utilized to help handling this problem.

C. Wrapper Sequential Forward Selection (WSFS)

Wrapper approach will be used to reduce the dimensionality of features subset. In wrapper approach, search strategy is used to determine the best possible features subset that maximises accuracy. This search strategy, there are divided into three categories, which are sequential, complete and random strategy.

The sequential search strategy was chosen for this study. The rationale for this is because this strategy is less computational and easy to be implemented [10, 12]. It also has the ability to quickly find the features and produce fast results where the search space is *O(N2)* [9]. In sequential search strategy, the most popular algorithms are sequential forward selection (SFS), sequential backward selection (SBS), sequential backward selection-SLASH (SBS-SLASH) and bi-directional search (BDS) [10]. However, this study will only describe the sequential forward selection (SFS) as it was used in the wrapper approach.

SFS refers to a search strategy that begins with an empty set of features. Then, the best features will be selected for every step in iteration and combined with the existing features chosen in an empty set. The iteration will be stopped until reaching a limit where no improvement is showed in the features [9].

IV. DISCUSSION

This section will discuss the existing feature selection techniques in educational data mining. Table 3 shows the summary of the existing feature selection techniques used in educational data mining since 2009 until 2015.

Table 3
Summary of Existing Feature Selection Techniques in Educational Data

Feature Selection Techniques	Algorithm	Results (%)	Author
Filter	IGATE	70.75%	[3]
	Gain Ratio	58.8%	[6]
	Chi Square	70%,59.2%	[4, 8]
	Symmetrical Uncertainty	59.2%	[6]
	Correlation Based	71.5%,58.4%, 87%	[4,8,19]
	Relief Attribute	58.2%	[6]
Wrapper	Predictive Accuracy	70.75%, 81.33%	[4, 14]

As shown in Table 3, most of the researchers have used filter based model to conduct feature selection process [6]. The main reason behind this is because of its fast computational time and simple measurement. This finding is also supported by Doshi et al., showing that Fast Correlation Base Filter has successfully removed the irrelevant and redundant features by (87%) [13]. As a result, the prediction accuracy increased while the computational time decreased. However, the filter based techniques also have their weaknesses. When applying in a wide feature set, it becomes unstable and produce the unsatisfactory results. Moreover, filter based technique avoids the interaction with the classifier which may affect the classification performance.

Wrapper based technique is also used by researchers in educational area for features selection process. The previous studies shows the result of performance accuracy is more than (70%) [4, 14]. Nonetheless, it is rarely applied compared to filter based technique. It is because the relevant features are directly measured using data mining algorithm. As a result, its computational time became slow. Another drawback of using wrapper model is more expensive as it requires more computation. However, this study is focused on performance accuracy and the improvement of data quality. By using wrapper method, it selects and evaluates only the important features using learning algorithm as there exists an interaction between features subset search and model selection. Therefore, the mining performance can be increased. So, wrapper based technique is a better choice to find the sub optimal features and increase the performance accuracy.

From the observation in educational data mining study, the trend is towards wrapper based method for selecting relevant features compared to filter based method. Consequently, by looking the advantages and disadvantages of both filter and wrapper based models, this research attempts to enhance the feature selection technique by combining the wrapper and filter based methods. The combination of filter and wrapper based methods is known as hybrid method. There have been a few studies in medical domain that apply embedded method for feature selection technique. But, in educational domain, the number of researches conducted within this area is still inadequate. Hence, it is a great research opportunity to apply the hybrid method for feature selection technique in educational area. It has been approved by the study commenced by Das, where the qualities of the filter and wrapper approaches have been successfully utilized in high dimension environment [18]. Inspired by the successful previous research, hybrid approaches (wrapper and filter) were further chosen for feature selection in this study.

This study used Wrapper Sequential Forward Selection (SFS) with Fast Correlation Based Feature Selection (FCBF) as a feature selection technique to produce the quality dataset and improve the performance accuracy.

The result shows when hybrid of SFS and FCBF techniques, the performance accuracy is around (88%). Followed by filter based correlation technique which is (87%) of accuracy. The lowest accuracy is predictive accuracy by (70%). It approves that hybrid of feature selection technique can improve the performance accuracy during feature selection technique process compared to previous technique. Fig. 3. shows the result comparison between filter, wrapper and hybrid feature selection techniques by using same dataset.

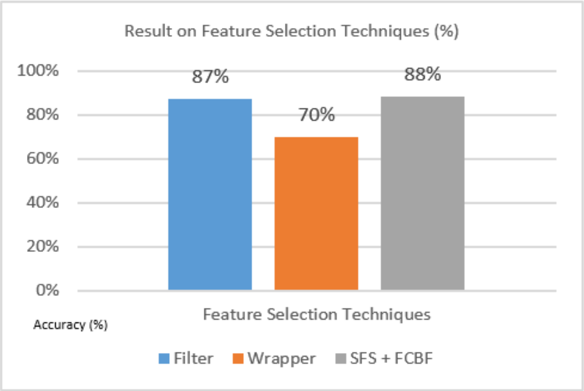

Figure 3. Result Comparison on Feature Selection Techniques

V. Conclusion

From the literature review, feature selection technique is appropriate technique to reduce high dimensional data in educational dataset. The finding from this study is expected to give insight to other researchers in educational domain especially on ways to produce the quality dataset. For future work, the experiment will be conducted to prove the proposed framework can improve quality of students' dataset and increased the prediction accuracy.

Acknowledgment

We would like to show our gratitude to Universiti Sains Malaysia for supporting this research.

References

[1] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, Mar. 2014.

[2] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] A. Acharya and D. Sinha, "Application of Feature Selection Methods in Educational Data Mining," *Int. J. Comput. Appl.*, vol. 103, no. 2, pp. 34–38, 2014.

[4] A. Bidgoli and M. N. Parsa, "A Hybrid Feature Selection by Resampling , Chi squared and Consistency Evaluation Techniques," *Eng. Technol.*, vol. 6, no. 8, pp. 276–285, 2012.

[5] B. Singh, N. Kushwaha, and O. P. Vyas, "A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty," *J. Data Anal. Inf. Process.*, no. November, pp. 95–105, 2014.

[6] M. Ramaswami and R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining," *J. Comput.*, vol. 1, pp. 7–11, 2009.

[7] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015.

[8] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," *Data Classif. Algorithms Appl.*, pp. 37–64, 2014.

[9] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *Knowl. Data Eng. IEEE Trans.*, vol. 17, no. 4, pp. 491–502, 2005.

[10] V. Kumar, "Feature Selection: A literature Review," *Smart Comput. Rev.*, vol. 4, no. 3, 2014.

[11] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," *Int. J. Eng. Res. Technol. data Min.*, vol. 1, no. 6, pp. 1–6, 2012.

[12] Ö. Uncu and I. B. Türkşen, "A novel feature selection approach: Combining feature wrappers and filters," *Inf. Sci. (Ny).*, vol. 177, no. 2, pp. 449–466, 2007.

[13] M. Doshi and S. K. Chaturvedi, "Correlation Based Feature Selection (Cfs) Technique To Predict Student Perfromance," *Int. J. Comput. Networks Commun.*, vol. 6, no. 3, pp. 197–206, 2014.

[14] H. M. Harb and M. A. Moustafa, "Selecting Optimal Subset of Features for Student Performance Model," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 253–262, 2012.

[15] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance," *5th Annu. Futur. Bus. Technol. Conf.*, vol. 2003, no. 2000, pp. 5–12, 2008.

[16] K. Michalak and H. Kwasnicka, "Correlation-based Feature Selection Strategy in Neural Classification," *Sixth Int. Conf. Intell. Syst. Des. Appl.*, vol. 1, pp. 741–746, 2006.

[17] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[18] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," *Engineering*, pp. 74–81, 2001.