International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-018-0141-y

REGULAR PAPER

Identifying key factors of student academic performance by subgroup discovery

Sumyea Helal[1] · Jiuyong Li[1] · Lin Liu[1] · Esmaeil Ebrahimie[1,2,3] · Shane Dawson[4] · Duncan J. Murray[5]

Received: 11 May 2017 / Accepted: 8 June 2018

Abstract
Identifying the factors that influence student academic performance is essential to provide timely and effective support interventions. The data collected during enrolment and after commencement into a course provide an important source of information to assist with identifying potential risk indicators associated with poor academic performance and attrition. Both predictive and descriptive data mining techniques have been applied on educational data to discover the significant reasons behind student performance. These techniques have their own advantages and limitations. For example, predictive techniques tend to maximise accuracy for correctly classifying students, while the descriptive techniques simply search for interesting student features without considering their academic outcome. Subgroup discovery is a data mining method which takes the advantages of both predictive and descriptive approaches. This study uses subgroup discovery to extract significant factors of student performance for a certain outcome (Pass or Fail). In this work, we have utilised student demographic and academic data recorded at enrolment, as well as course assessment and participation data retrieved from the institution's learning management system (Moodle) to detect the factors affecting student performance. The results have demonstrated the effectiveness of the subgroup discovery method in general in identifying the factors, and the pros and cons of some popular subgroup discovery algorithms used in this research. From the experiments, it has been found that students, who have indigent socio-economic background or been admitted based on special entry requirement, are most likely to fail. The experiments on Moodle data have revealed that students having lower level of access to the course resources and forum have higher possibility of being unsuccessful. From the combined data, we have identified some interesting subgroups which are not detected using enrolment or Moodle data separately. It has been found that those students, who study off-campus or part-time and have a low level of contributions to the course learning activities, are more likely to be the low-performing students.

Keywords Subgroup discovery · Education data mining · Moodle · Enrolment data

1 Introduction

A huge volume of student enrolment data is accumulated by educational institutions each year. The data contain socio-demographic (e.g. age, ethnic origin, gender and disability status) and academic (e.g. admission basis and delivery mode) information of students. These data can be useful for institutions to detect potential at-risk students early in their course of study. Hence, this early identification affords greater, and more timely, opportunity to provide more pro-active support interventions [50].

The use of technology in learning has become an integral component of contemporary education. A long-standing enterprise system adopted to meet the pressures of delivering a flexible model of education has been the Learning Management System (e.g. Moodle [3], BlackBoard [1], and

✉ Sumyea Helal
sumyeahelal@gmail.com

Jiuyong Li
Jiuyong.Li@unisa.edu.au

1 School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia

2 School of Biological Sciences, The University of Adelaide, Adelaide, Australia

3 School of Biological Sciences, Flinders University, Adelaide, Australia

4 Teaching Innovation Unit, University of South Australia, Adelaide, Australia

5 Business Intelligence and Planning, University of South Australia, Adelaide, Australia

Published online: 21 June 2018

Desire2Learn [2]). These systems accumulate vast volumes of student information related to the course, study activities and outcomes. The information can be valuable for analysing student behaviour, predicting academic performance and assisting teachers in taking appropriate and timely measures to improve student learning experience and outcomes.

Identifying the factors affecting student academic performance is a challenging task as a huge number factors such as demographic, academic as well as students' participation in different course activities can influence their performance. A solution to this problem is to use data mining approach, a knowledge discovery process that allows automatic extraction of implicit and interesting patterns from large data collections [60].

A large number of data mining techniques have been developed. They mainly fall into two types according to their objectives: predictive (e.g. classification) and descriptive (e.g. association analysis) [28]. Descriptive techniques extract properties of the data, while predictive techniques induce the data for making predictions. This study focuses on extracting interesting features that influence student performance. Although the descriptive techniques discover interesting patterns of student characteristics, these approaches are not able to group students based on a certain outcome. On the contrary, the predictive methods group students into classes of similar behaviour; however, their main focus is on building models with higher predictive accuracy. We propose to extract the influencing student features by employing subgroup discovery [41] methods.

Subgroup discovery is halfway between descriptive and predictive techniques, and it can be used to identify interesting relations with respect to a special property of interest. Subgroup discovery has been successfully used in various domains. In medical domain, subgroup discovery has been widely applied to extract the subgroups of influenza virus [13] and to the pathogenesis of acute sore throat conditions in humans [15]. Subgroup discovery has also been successfully used in the area of technical behaviour analysis, for example, mining service processes [55], extract the features of a type of solar cell in the concentrating photovoltaic area [14], or fault analysis of production processes [34]. In marketing domain, subgroup discovery has been employed to do customer segmentation of an online customised fashion business [12]. Subgroup discovery has also been used to uncover structure–property relationship of materials [24] and community detection [7].

In this study, we use subgroup discovery to identify influencing factors of student academic performance as this method can discover subgroups of students while considering student behaviour (i.e. enrolment or Moodle data features) and the outcome (student performance) simultaneously. Carmona et al. [18] demonstrated that the implementation of subgroup discovery approach is promising in mining educational data. This work compares several classic and evolutionary subgroup discovery approaches to extract interesting features of student performance. However, the authors only considered a few of the vast number of Moodle attributes while ignoring other important (e.g. socio-demographic or academic) features.

In this study, we consider student enrolment and course participation attributes (gathered from Moodle) separately as well as jointly. Considering both types of features may reveal some new facts useful for detecting the vulnerable students more accurately. For example, when analysing only Moodle data, we may find out that a student having medium participation in assignment activities still fails the course. However, if we also analyse the enrolment data, we may discover that a student who has medium participation in assignment activities and studies part-time fail the course. Such discovery is more sensible because part-time students may have less time in general in their studies, including participating in Moodle activities, which may result in a low academic performance.

Existing subgroup discovery approaches can be broadly classified into three major categories according to the strategy for searching candidate subgroups: exhaustive search-based approaches, beam search- based approaches, and genetic algorithm (GA)-based approaches [31]. This study employs six different subgroup discovery methods: Statistically Non-redundant Subgroup discovery (SNS) [48], Diverse Subgroup Set Discovery (DSSD) [45], Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery (NMEEF-SD) [16], Bitset-based Subgroup Discovery (BSD), SD-Map, and APRIORI-SD each belonging to either one of the three categories. These methods are chosen as being very popular among different existing subgroup discovery methods. The diversity of the methods helps to evaluate the consistency of mined factors across them.

The aim of our study is to answer the following research questions, and hence, the contribution of the paper is the answers to these questions.

1. What are the advantages and disadvantages of each type of subgroup discovery methods in identifying interesting and useful factors of student performance? Firstly, different subgroup discovery methods are applied on the enrolment, Moodle and combined datasets to extract the factors influencing student academic achievement. In the next step, their pros and cons are discussed in terms of the knowledge derived from the discovered subgroups.
2. Is there a best-performing method in identifying those predictors? The performance of the subgroup discovery methods is evaluated and compared by using different measures.
3. What attributes are most effective predictors of student academic performance? The influencing factors found by

different methods are assessed in terms of their usefulness in decision-making for the course instructors and the educational institutions.

The rest of the paper is arranged as follows: Section 2 introduces the basic concept of subgroup discovery and its methodology and discusses some related works on educational researches in the area of data mining. The methodology followed in this work is described in Sect. 3. In Sect. 4, subgroup discovery methods are applied to the enrolment and Moodle datasets separately and jointly and the results are discussed. The performance of the methods and the usefulness of their findings are analysed in Sect. 5. Finally, Sect. 6 concludes the article and suggests future work.

2 Background

This section introduces the concept of subgroup discovery and its methodology, and briefly describes the use of different data mining approaches in the area of educational data.

2.1 What is subgroup discovery

Subgroup discovery searches for interesting characteristics of subgroups of a population with respect to a property of interest (also known as target variable) [41]. A subgroup is represented in the form of a rule S→t where the antecedent S is the conjunction of a set of features describing the subgroup (called subgroup descriptor hereafter) and the consequent t refers to the value of the target variable, T, i.e. the property of interest. Consider Table 1, where *Obesity* is the target variable. Some possible interesting subgroups are as follows.

$$SG_1: (W = 60\text{–}80\,\text{kg AND } L = \text{No}) \rightarrow \text{Obesity} = \text{yes}$$
$$SG_2: (A = 25\text{–}50 \text{ AND } W = 60\text{–}80\,\text{kg}) \rightarrow \text{Obesity} = \text{no}$$

2.2 Essential elements of subgroup discovery

When applying a subgroup discovery method, the following two elements are considered very important.

Table 1 An example dataset

Weight (W)	Lack of sleep (L)	Age (A)	Obesity
60–80 kg	No	< 25	Yes
< 60 kg	No	< 25	No
60–80 kg	No	25–50	No
> 80 kg	Yes	25–50	Yes
60–80 kg	Yes	> 50	No

Type of target variable

Target variables are of many types; among them, there are three major types as follows.

- Binary: This type of variable possesses only two values—true or false. The algorithms that employ binary target generate subgroups for each of the different variable types.
- Nominal: For this type of analysis, a target variable can take different number of values. Subgroup discovery algorithms that implement nominal target extract subgroups for each of the values.
- Numeric: This type of target variables can take up to many values and hence very complex to analyse.

Description language

The representation of the generated subgroups should be comprehensible to obtain interesting and useful rules. The subgroup discovery algorithms refine subgroups using attribute-value pairs in conjunctive or disjunctive normal form. The values of the variables can be represented by using either the $=$, $\neq$, $<$, or $>$ operators.

2.3 Phases of subgroup discovery

A subgroup discovery algorithm possesses three phases for subgroup extraction—candidate subgroup generation, pruning and post-processing. In the following, we describe the phases in detail.

2.3.1 Candidate subgroup generation

Each subgroup discovery algorithm uses a specific strategy to search for the candidate subgroups. The search space is traversed by starting with simple descriptions and processing these in a more general to specific manner by adding up more attribute-value pairs. Different search strategies have been employed so far for subgroup discovery; among them, the most widely used strategies are beam search-, exhaustive search- and genetic algorithm (GA)-based search.

Exhaustive search

Exhaustive search is a very popular problem-solving technique which generates all possible candidates and verifies whether each candidate satisfies some specific constraints. The cost of this type of search is proportional to the number of candidates extracted; however, when the search space is too large, exhaustive search is not affordable.

Beam search

In beam search, only a predetermined number (known as beam width bw) of best partial solutions are considered as candidates. At individual level, the bw highest ranked candidates are generated according to the quality. Beam search restricts the memory usage by exploring part of the search space, but it does not guarantee solution at end.

Genetic algorithm

Genetic algorithm [23] is a search heuristic that follows the process of natural evolution; hence, the methods implementing this search heuristic are known as evolutionary methods. This type of heuristic is used to extract solution to different optimisation and search processes.

2.3.2 Pruning

In the second phase, a subgroup discovery algorithm needs to employ a pruning scheme for selecting only the significant candidates. A number of pruning strategies are used by different methods. The major types include minimum support pruning [9,22,39], optimistic estimate pruning [66] and constraint pruning [26,44].

2.3.3 Post-processing

The final phase of subgroup discovery algorithm implements a quality measure in the purpose of ranking subgroups. These measures are very vital for evaluating subgroups as the interest attained directly relies on them.

2.4 Discovery of knowledge from educational data

Discovering significant factors of student academic outcome requires an explanatory analysis of the data which can be achieved through Educational Data Mining (EDM), a knowledge discovery process that extracts valuable information from data coming from an educational setting [61]. The identification of the key indicators of student performance is one of the widely researched areas in EDM, with the ultimate goal to provide administrative and academic support to the vulnerable students.

Researches [47,51,64] found socio-demographic and academic features as significant indicators of student success. Some other works discovered that student participation in different course-related activities had direct influence on their academic achievements. For instance, Lokyer et al. [49] raised the issue of pedagogical context about how course design and activities influenced the way in which students interacted with a technical system. Similar findings can be seen in [65,69] which showed that students' level of interaction with the course had high impact on their academic outcome.

A number of predictive methods, e.g. classification [58], have been widely used for discovering potential factors of students academic performance. Classification is a supervised process of grouping objects of similar characteristics into classes. This approach has been widely used for modelling student academic performance. A decision tree-based approach [53,54] has been proposed which extracts some influencing factors separating successful students from unsuccessful ones. It was found that features regarding student socio-demographic information and extra-curriculum activities affect their academic achievement. Another interesting work [25] found that student socio-demographic features, e.g. age, gender, prior academic performance and some psychometric factors, influence their academic success.

A number of works have attempted to predict and improve student performance by considering the utilisation of course resources and participation in course activities. Researches found [40,52,62] that web-based courses supported with online forums enhanced student performance. A similar work [68] was conducted for capturing student activities by considering reading and posting messages, content creation contribution, quiz efforts and number of files viewed by them. Another work [67] used online Q and A discussions to predict student performance using regression analysis. In [61], student performance in a course was predicted by analysing both the quantitative information (e.g. number of forum messages read and written) and qualitative information (a score based on the usefulness of the content of forum messages manually set by the course teacher).

Other researches have categorised students into different groups according to their course Moodle usage to detect the factors behind a student's success/failure in a course. In [59], a comparison among different data mining methods was made for classifying students based on their usage data in a course Moodle. In a similar approach, Jovanovic et al. [36] clustered students based on identified cognitive styles. For improving e-learning system, student behaviour pattern was analysed in [10] for some e-courses where each student was represented with an activity value (e.g. submit, view or edit) and a module value (e.g. course or user). A reference model was developed in [42] to classify a student into the dropout or non-dropout group.

Descriptive methods, e.g. association analysis, have been widely used for discovering useful features of student performance, whose main purpose is to discover interesting knowledge from data. Association analysis aims to search for descriptive rules representing relations among different attributes. This approach has been employed on educational data to reveal those student characteristics tend to occur together frequently.

A work [63] considered student socio-demographic and basic academic features to discover the influencing factors behind student academic performance using association rule mining. It was found that students' GPA, secondary school type and their gender affect their academic outcome. The work in [56] profiled students according to their performance by generating two-way associations in search for interesting patterns among different demographic and academic attributes. Association rule mining has also been used in [11] which found that students' academic features, i.e. class attendance and their marks in assignments, are highly correlated with their academic achievement.

Association analysis has also been successfully used in the process of e-learning. An interesting work [37] attempts to identify different types of learners with the help of their distinguishing interaction behaviour relate to learning. Another work [20] has revealed that users who were provided with hints achieved higher average marks than those who were not and stayed engaged for longer with the course site. The authors in [33] proposed an association rule mining-based approach on discovering courses pattern for constructing suitable learning path. Another similar approach [32] has been used to exploit learners' model for e-learning to discover the learning paths in LMS beneficial to the course instructors. A group of data mining techniques including fuzzy association rule mining, statistical correlation analysis have been applied in [19] to support mobile formative assessment in order to help teachers understand the significant factors of learner performance.

Subgroup discovery methods have also been used for mining important factors of student performance. Subgroup discovery methods have been employed in [47] for identifying the influencing factors of student success. However, this study considers fewer socio-demographic and admission features. The authors in [17,60] have shown the effectiveness of subgroups discovery in mining educational data. It was found that higher participation in different Moodle activities led students to secure a good grade. A comparison among several subgroup discovery methods has been made in [18] for extracting factors influencing student academic performance. However, the experiments are confined to the analysis of Moodle data rather than considering other important features regarding their demographic, socio-economic, and academic status.

3 Materials and methods

This work proposes the use of different subgroup discovery methods to identify the factors of student academic performance. Firstly, students' enrolment and Moodle data are collected and pre-processed. In the next step, different subgroup discovery methods are applied to these datasets and finally the findings by these methods are analysed and the performance of the methods is compared using different evaluation measures.

Table 2 Summary of datasets

Dataset	#Instances	#Attributes
Enrolment	1311	20
Moodle	3567	14
Combined	1311	34

3.1 The datasets

The datasets used in this paper were collected at a division of an Australian university about their first-year domestic undergraduate students. Three types of datasets have been employed in this work—enrolment data, Moodle data and the combined data containing both the enrolment and Moodle features. The enrolment dataset contains 8 socio-demographic attributes (e.g. age, gender and economic status) and 12 academic attributes (e.g. attendance mode and attendance type). For each course, a student's performance (passing or failing the course) is also included in the datasets which is used as target variable for all the executions. A brief description of the datasets is given in Table 2.

3.1.1 Enrolment data

The enrolment data were collected during a student's entry into the institution. The dataset contains students' socio-demographic (i.e. age, gender and economic status) and academic features (attendance type and delivery mode). In our experiments, a student's overall performance was calculated based on the average marks of all the courses he/she has taken in a year.

3.1.2 Moodle data

The Moodle data records students' online participation in different activities (e.g. assignments, quiz, forum and others) and resources (e.g. book and file) gathered from individual course Moodle sites. Each record of the Moodle dataset contains a student's participation in different Moodle activities in a specific course. All the features in a course Moodle have been formed in three different ways:

– Module type features: These are the collection of action features for a particular module type such as assignment, quiz, page.

– Module/action code features: Module/action code features represent a student's participation in an individual activity of a specific module.
– Categorisation features: Categorisation features are combination of several features which fall into same category. For example, one of such features may be when a student only views an activity of all modules.

In our experiment with the Moodle data, we use the Module/action code features for identifying the factors affecting student performance in a course as these features represent a student's participation in a specific activity and hence are more helpful for predicting his/her performance in a course based on the specific actions.

3.1.3 Combined data

Each record of the combined dataset represents a unique student's information containing both the enrolment and Moodle features. As a record of the Moodle dataset corresponds to a course taken by a student, there can be multiple records for different courses taken by the same student in the dataset. Therefore in the combined dataset, for a student, the value of a Moodle feature is the average counts of the student's participation in this Moodle activity of all his/her courses. Similar to the enrolment dataset, for the combined dataset the performance of a student is represented as the average mark of all the courses he/she has taken in a year.

3.2 Data pre-processing

It is necessary to pre-process the data before applying the subgroup discovery methods to them. In this study, pre-processing is performed in the following two steps:

1. Discretisation: The software tools used in this experiment only work on categorical attributes. Hence, discretisation is performed on the enrolment attributes age, AUST-SES[1] and ATAR[2] as well as on all Moodle attributes. All Moodle attributes have been categorised into four quartiles.
2. Data Transformation: The subgroup discovery methods studied in this work require the data to be in the C4.5 [58], ARFF [27] and KEEL [6] formats. The original enrolment and Moodle dataset were in excel format. They are transformed to the above-mentioned formats applicable to the subgroup discovery methods, respectively.

[1] $AUST\text{-}SES$: *Australian Social Economic Status.*

[2] $ATAR$: *Australian Tertiary Admission Rank*, the primary criterion of entry into the most undergraduate programs in any university of Australia and represents a student's ranking relative to his/her peers upon completion of their secondary education.

3.3 Discovering significant factors of student performance

In this work, subgroup discovery methods are applied to the enrolment, Moodle and combined datasets to detect the risk factors of student academic performance. This section subgroup discovery methods used in this study and the measures used to evaluate the methods.

3.3.1 The methods used

In this paper, we employ some of the most recent and popular subgroup discovery methods, namely SNS, DSSD, NMEEF-SD, BSD, SD-Map and APRIORI-SD. A brief description of the methods is given as follows.

SNS

SNS [48] is an optimal subgroup discovery algorithm which extracts all statistically significant subgroups. It forms subgroups as conjunction of attribute-value pairs by employing the '=' operator. It works on a single binary target attribute. SNS uses odds ratio (OR) [21] to measure the quality of a subgroup.

In the contingency table below for subgroup $S \rightarrow t$, a_{11} represents the number of examples satisfying both S and t, while a_{12} represents the examples which satisfy S and do not satisfy t; a_{21} denotes the number of examples which contain t but do not satisfy S; a_{22} represents the number of examples which do not satisfy both S and t.

Based on Table 3, the odds ratio of the subgroup is given by the following equation:

$$\text{OR} = \frac{a_{11} * a_{22}}{a_{12} * a_{21}} \tag{1}$$

The higher the odds ratio, the stronger the association between the subgroup features and the target. Consider the above subgroup SG_1. From Table 1, we can see that $a_{11} = 1$, $a_{12} = 1$, $a_{21} = 1$ and $a_{22} = 2$. Therefore the odds ratio of SG_1 is 2.

DSSD

DSSD [45] is a beam search-based subgroup discovery algorithm which generates a fixed number of candidate subgroups during each iteration. It uses both '=' and '!=' operators for

Table 3 Contingency table of subgroup S

	t	$\neg t$
S	a_{11}	a_{12}
$\neg S$	a_{21}	a_{22}

refining subgroups. DSSD can use a number of measures to evaluate the interestingness of a subgroup. Among them, Unusualness (UN) [43] is the most popular one, and it is defined as follows:

$$\text{UN} = \frac{|coverset(S)|}{N}\left(\frac{|supportset(S,t)|}{|coverset(S)|} - \frac{|\{t\}|}{N}\right) \quad (2)$$

Here $coverset(S)$ is the set of examples in a dataset D satisfying S, $supportset(S,t)$ is the set of examples in D satisfying both S and target value t, $\{t\}$ represents the set of examples containing t, and N is the total number of examples.

The above definition of unusualness reflects the trade-off between rule generality (or the relative size of a subgroup, $(|coverset(S)|/N)$) and relative accuracy (the difference between the fraction of the examples covered within the subgroup and the fraction of all examples containing target value t in the whole dataset).

From Table 1, it is seen that the antecedent part of subgroup SG_2 covers only one example, so $|coverset(S)| = 1$. There is only example in the dataset for the subgroup, so $|supportset(S,t)| = 1$. The total number of examples $N = 5$ and the total number of examples containing the target value "No" is 3, so according to Eq. (2), the unusualness of SG_2 is 0.08.

NMEEF-SD

NMEEF-SD [16] is a genetic algorithm-based approach which follows the process of natural evolution such as inheritance, mutation, selection and crossover. According to genetic algorithm, each solution is composed of several variables and equipped with a fitness score. The solutions with higher fitness values are given the opportunity to evolve. This method forms subgroup by using conjunction of attribute-value pairs and employ only '=' operator.

For the extraction and evaluation of subgroups, this method can also use *unusualness* as defined in Eq. (2).

BSD

BSD [46] is an exhaustive subgroup discovery algorithm which utilises a vertical bitset-based data structure and uses a depth-first search approach. In this algorithm, the refinement of the patterns is applied using logical AND operations on the respective bitsets. It employs only '=' operator for subgroup refinement. It can use a number of quality measures, e.g. Piatetsky-Shapiro [41], unusualness and others for ranking subgroups.

SD-Map

SD-Map [9] is an exhaustive subgroup discovery method which is an extension of the popular Frequent Pattern (FP) growth [29] based association rule mining method. It implements a depth-first search for candidate generation and also is able to handle missing values for different domains. This method uses only '=' operator for forming subgroups. It uses several quality functions such as Piatetsky-Shapiro and unusualness are the most popular ones.

APRIORI-SD

APRIORI-SD [38] is an extension of the classification rule learning algorithm, APRIORI-C [35]. Discovered subgroups are post-processed by using unusualness as the quality measure. All the positive examples covered by a rule are not removed from the training dataset, rather each time an example is covered, and its weight decreases. An example is removed only when its weight falls below a given threshold or when an example has been covered more than k times. This method uses '=', '<' and '>' operator for refining subgroups.

Note that given a dataset, all the above methods firstly filter out subgroups with low support. The support of a subgroup is given by the following equation where as introduced previously $|supportset(S,t)|$ is the number of examples in the subgroup, i.e. the number of examples in the dataset that satisfy both S and t, and N is the total number of examples in the dataset [5].

$$\text{Sup} = \frac{|supportset(S,t)|}{N} \quad (3)$$

With the remaining subgroups, each of the methods applies its respective measure (as introduced above) to rank and select the top subgroups as the output.

3.3.2 Assessing discovered subgroups

The performance of a subgroup discovery method is evaluated by examining the subgroups generated by the method, in various aspects as introduced in the following.

Simplicity

This measure is related to the simplicity of knowledge obtained from the subgroups extracted by a method. In studies [31], the simplicity of a method is measured in terms of the number and length of subgroups which can be stated as below.

Number of subgroups It measures the number of discovered subgroups. In a beam search-based method, the number

of discovered subgroups is restricted by beam width. In a top-k method, the number of generated subgroups is limited by the value of k.

Length of subgroups It represents the number of variables contained in a subgroup. Let $l(S)$ be the length of a subgroup S and n_s be the total number of subgroups induced by a method; the average variable length of a subgroup set is given as follows.

$$\text{AvLength} = \frac{1}{n_s}\sum_{i=1}^{n_s} l(S_i) \tag{4}$$

Unusualness

This measure is defined as the Weighted Relative Accuracy (WRAcc) of a rule. It can be described as the trade-off between rule generality and relative accuracy (the difference between rule accuracy and default accuracy). It is represented as Eq. (2).

Redundancy [30]

This measure evaluates the proportion in which a subgroup set contain extraneous information. It can be defined as follows.

It measures the fraction of redundant subgroups to the discovered subgroups. A subgroup $S_k \rightarrow t$ is redundant if there is another subgroup $S_j \rightarrow t$ such that

1. $S_j \subseteq S_k$, and
2. $\varphi(S_j \rightarrow t) \geq \varphi(S_k \rightarrow t)$.

where φ is the quality measure used by the subgroup discovery method, e.g. odds ratio for SNS and unusualness for DSSD.

Comprehensibility

It is a subjective measure which represents the understandability of the discovered knowledge. It considers the semantics (e.g. use of different operators) and explanation of the patterns. In this paper, we classify the (relative) comprehensibility of the findings of a subgroup discovery method as high, medium or low based on the level of interpretability of the discovered subgroups.

Reliability

This criterion evaluates the trustworthiness of the knowledge represented by the discovered subgroups. This study has used the following metrics for evaluating reliability.

Precision [57] The precision (also known as *confidence*) measures the proportion of correctly classified examples that have actual matches. It is given by the following equation.

$$\text{Precision} = \frac{\sum_{i=1}^{n_s} |supportset(S_i, t)|}{\sum_{i=1}^{n_s} |coverset(S_i)|} \tag{5}$$

Recall [57] Recall (also known as *sensitivity*) measures the proportion of the actual matches that have been classified correctly. It is represented by the following equation.

$$\text{Recall} = \frac{\sum_{i=1}^{n_s} |supportset(S_i, t)|}{|t|} \tag{6}$$

4 Experiments

This section discusses the influencing features of the unsuccessful students as identified by different subgroup discovery methods. The experiment is conducted in three different phases by mining—(i) enrolment data, (ii) Moodle data and (iii) combined data. Each of the methods (SNS, DSSD, NMEEF-SD, BSD, SD-Map and APRIORI-SD) is applied to these three datasets, respectively. SNS employs odds ratio for ranking subgroups, while the rest use unusualness for ranking them. The minimum rule support is 5% .

All the experiments were conducted on a computer with 4 core Intel i7-3370 CPU @ 3.40 GHz, 16 GB RAM and 64-bit windows operating system. For SNS, we used the implementation by SNS software tool [48]; for DSSD, the implementation by DSSD software tool [45], for NMEEF-SD, implementation by KEEL data mining tool [6], for BSD and SD-Map, implementation by Vikamine [8], and for APRIORI-SD, implementation by Orange data mining tool [4].

In the following, we present the results of the experiments for the target *fail* for the above-mentioned datasets, respectively, in the following subsections.

4.1 Mining enrolment data

Referring to Table 4, using the enrolment dataset, SNS has identified that student socio-demographic features, i.e. age, parent education and social economic status, are strongly associated with their academic performance. It is also found that their academic features such as ATAR, admission basis and being admitted into multiple programs in current year contribute highly to their academic achievement. It follows from the very first subgroup of Table 4 that students who have medium ATAR and come from a lower-income family are most likely to fail. It is also seen from the second subgroup that students, who get admission based on professional qualification and have parents who completed only year 12

Table 4 Top-10 subgroups discovered by mining enrolment dataset

Method	No.	Subgroup description	Subgroup size	Quality (OR/UN)
SNS	1	Social economic status = Low AND ATAR = Medium	88	5.11
	2	Admission basis = Professional qualification AND Parent Education = Year 12	64	4.96
	3	Age = Mature	164	4.69
	4	HECS exempt type = Deferral AND High School State = SA	89	4.29
	5	ATAR = Low AND In multi program this year = Yes	42	3.92
	6	Admission basis = Mature Age Special Entry	75	3.91
	7	In multi program any year = Yes AND ATAR = Medium	81	3.79
	8	Attendance mode = Internal AND Social economic status = Low	69	3.43
	9	High School State = SA AND In multi program this year = Yes	42	3.09
	10	ATAR = Medium	181	3.00
DSSD	1	ATAR != High AND Admission Basis != Higher Education Course	203	0.06
	2	HECS exempt type != DiscountAND Birth Country != Ireland	184	0.05
	3	ATAR != High AND High School State != Tasmania	220	0.05
	4	Attendance mode != Internal AND Parent Education != Postgraduate Degree	227	0.05
	5	ATAR != High AND Admission Basis != Secondary Education	217	0.05
	6	ATAR != HighAND Birth Country != Poland	163	0.05
	7	HECS exempt type != Discount AND Admission Basis != Secondary Education	147	0.05
	8	Age = Normal	217	0.05
	9	HECS exempt type != Discount AND Parent Education = Year 12	205	0.05
	10	Birth Country != Thailand AND Social economic status != High	216	0.05
NMEEF-SD	1	Program type = Bachelor's Pass	290	0.05
	2	Application was first preference = Yes	265	0.05
	3	In multi program any year = No	235	0.04
	4	Attendance type = Full time	243	0.04
	5	Application was first preference = Yes AND In multi program any year = No	180	0.03
	6	HECS exempt type = Deferral AND Application was first preference = Yes	197	0.03
	7	Program type = Bachelor's Pass AND In multi program any year = No	154	0.02
	8	HECS exempt type = Deferral	166	0.02
	9	Program type = Bachelor's Pass AND In multi program this year = No	189	0.01
	10	In multi program any year = No AND In multi program this year = No	135	0.01
BSD/SD-Map	1	ATAR = Good AND Age = Normal	249	0.03
	2	ATAR = Good AND Admission Basis = Secondary Education	211	0.03
	3	ATAR = Good AND High School State = SA	213	0.03
	4	ATAR = Good AND Application was first preference = Yes	307	0.03
	5	ATAR = Good AND HECS exempt type = Deferral	293	0.03
	6	ATAR = Good AND Attendance Mode = Internal	265	0.02
	7	ATAR = Good	325	0.02
	8	ATAR = Good AND Program type = Bachelor's Pass	325	0.02
	9	Attendance type = Part-time AND HECS exempt type = Deferral	159	0.02
	10	Attendance type = Fulltime AND ATAR = Good	282	0.02
APRIORI-SD	1	Gender = M AND Admission basis = Mature Age Entry AND Age = Mature	112	0.02
	2	Gender = M AND Age = Mature AND Attendance mode = external	107	0.02
	3	Attendance type = Part-time	137	0.02

Table 4 continued

Method	No.	Subgroup description	Subgroup size	Quality (OR/UN)
	4	Admission basis = VET AND Attendance mode = external	126	0.02
	5	Gender = M AND Age = Mature AND In multi program any year = No	103	0.02
	6	Admission basis = Mature Age Entry AND Age = Mature	168	0.02
	7	HECS exempt type = Deferral AND Application was first preference = No	135	0.01
	8	Program type = Bachelor's Pass AND In multi program this year = No	189	0.01
	9	Attendance type = Part-time AND HECS exempt type = Discount	128	0.01
	10	Age = Mature AND Attendance mode = external AND Admission basis = VET	178	0.01

schooling, have a very high possibility to fail. The results have also shown that mature-aged students are mostly low performers which is stated by the third subgroup.

DSSD has found that student socio-demographic and academic background affect their academic outcome. DSSD has also discovered that student academic features have significant influence on their academic performance. Most of these factors have also been identified by SNS. The first subgroup (in Table 4) discovered by DSSD states that students, whose admission basis is other than higher education course and who do not have a high ATAR, possess higher possibility to fail. It is also found that students, who do not study on-campus and do not have postgraduate parent, have higher chance to fail (subgroup 4).

NMEEF-SD has revealed that different academic features such as program type, attendance type, application preference and being admitted into multiple program in current year or any year are highly related to their academic performance. Although SNS has found that being admitted into multiple programs in current year affects a student's performance, it was unable to extract the remaining factors affecting their academic outcome. These influencing factors were not found by DSSD as well.

BSD has identified student age as an influencing factor of their academic outcome. It has also discovered a number of admission features that affect student academic outcome, e.g. ATAR, admission basis, HECS exempt, attendance type, mode and being admitted into multiple program. Most of these factors have been found by SNS, DSSD and NMEEF-SD; however, student attendance type has not been picked up by either SNS or DSSD. SD-Map discovered similar factors as BSD.

APRIORI-SD has discovered that student demographic features, such as age and gender, affect their academic achievement. It has also extracted that student admission features, e.g. admission basis, attendance type, attendance mode, HECS exempt and being admitted into multiple program, have significant influence on their academic outcome. All of these features have been found by other methods as well.

The subgroups discovered from the enrolment data by different methods have revealed some influencing factors of student performance. Such discovery is very helpful for the educational institutions to detect the at-risk students. They can take pro-active measures accordingly to reduce the failure rate. As an example, it is seen that students, who have a poor academic and social background, are most likely to fail. By learning the risk factors, an educational institution can take necessary steps to support the students possessing these features, such as monitoring the progress by conducting a routine assessment of their study throughout the term. Moreover, the institution can also provide additional academic support, e.g. forming smaller groups of such students to allow them to take several extra classes along with a small weekly seminar on a specific topic regarding their interest.

4.2 Mining Moodle data

From the experiments on the Moodle dataset as illustrated from Table 5, SNS has discovered that students, who have less participation in viewing some specific resources and activities, have higher possibility to fail in that course. Although student participation in different Moodle activities is not assessable, they have significant influence on student academic outcome. In Table 5, the first subgroup discovered by SNS states that when a student has fewer number of visits to course home page, he/she has higher possibility to fail. Followed from the second subgroup, it is also seen that students, who have lower participation in viewing course forum and its discussions, are most likely to fail.

DSSD has found the students as low performers who have lower participation in viewing and contributing to different resources and activities. Although SNS has discovered that having lower participation in viewing different activities and resources leads to a lower grade, it was unable to reveal that student contribution to different activities also affects their performance. As shown in Table 5, the first subgroup

Table 5 Top-10 subgroups discovered by mining Moodle dataset

Method	No.	Subgroups of failure	Subgroup size	Quality (OR/UN)
SNS	1	Visiting Course home page = Q1	204	9.33
	2	Viewing Forum activity = Q1 AND Viewing Discussion in Forum = Q1	120	6.73
	3	Viewing File Resource = Q1 AND Reviewing Quiz activity = Q1	104	6.30
	4	Viewing Forum activity = Q1 AND Viewing Quiz activity = Q1	130	6.04
	5	Viewing Quiz Activity = Q1 AND Viewing File Resource = Q1	124	4.80
	6	Viewing Discussion in Forum = Q1	163	4.56
	7	Viewing Forum activity = Q1	161	4.32
	8	Viewing File Resource = Q1	180	4.01
	9	Commencing Quiz Attempt = Q1	269	3.75
	10	Viewing Quiz activity = Q1	245	3.23
DSSD	1	Visiting Course home page = Q1 AND Adding Discussion in Forum = Q1	196	0.06
	2	Viewing Choice activity != Q4 AND Adding Post in Forum != Q3	287	0.06
	3	Editing course Wiki != Q3 AND Viewing Lesson activity != Q4	274	0.06
	4	Editing course Wiki != Q4 AND Adding Discussion in Forum = Q1	282	0.06
	5	Adding Post in Forum != Q4 AND Editing course Wiki != Q4	264	0.06
	6	Viewing Lesson activity != Q4 AND Viewing Choice activity != Q4	253	0.06
	7	Viewing File resource != Q3 AND Adding Post in Forum != Q2	248	0.05
	8	Viewing Choice activity != Q3 AND Adding Post in Forum != Q4	269	0.05
	9	Adding Post in Forum != Q3 AND Adding Discussion in Forum = Q1	244	0.05
	10	Editing course Wiki != Q4 AND Adding Post in Forum != Q3	227	0.05
NMEEF-SD	1	Visiting Course home page = Q1	204	0.08
	2	Editing course Wiki = Q1	282	0.08
	3	Adding Post in Forum = Q1	290	0.08
	4	Viewing Discussion in Forum = Q1	163	0.07
	5	Visiting Course home page = Q1 AND Adding Post in Forum = Q1	235	0.07
	6	Visiting Course home page = Q1 AND Viewing Discussion in Forum = Q1	247	0.06
	7	Viewing Discussion in Forum = Q1 AND Adding Post in Forum = Q1	211	0.06
	8	Visiting Course home page = Q1 AND Editing course Wiki = Q1	247	0.06
	9	Viewing Choice activity = Q1 AND Editing course Wiki = Q1	220	0.05
	10	Viewing Discussion in Forum = Q1 AND Editing course Wiki = Q1	216	0.04
BSD/SD-Map	1	Adding Discussion in Forum = Q1	324	0.04
	2	Visiting Course home page = Q1 AND Adding Discussion in Forum = Q1	332	0.03
	3	Visiting Course home page = Q1 AND Adding Post in Forum = Q1	277	0.03
	4	Visiting Course home page = Q1 AND Viewing Choice activity = Q1	257	0.03
	5	Visiting Course home page = Q1 AND Editing course Wiki = Q1	271	0.03
	6	Visiting Course home page = Q1 AND Viewing Wiki Activity = Q1	269	0.03
	7	Visiting Course home page = Q1 AND Viewing Lesson activity = Q1	266	0.03
	8	Visiting Course home page = Q1 AND Commencing Quiz Attempt = Q1	234	0.03
	9	Visiting Course home page = Q1 AND Reviewing Quiz activity = Q1	251	0.03
	10	Visiting Course home page = Q1 AND Viewing Book Resource = Q1	288	0.03
APRIORI-SD	1	Visiting Course home page = Q1 AND Commencing Quiz Attempt = Q1	234	0.03
	2	Viewing Discussion in Forum = Q1 AND Viewing Wiki activity = Q1	218	0.03
	3	Adding Discussion in Forum = Q1 AND Adding Post in Forum = Q1	207	0.03
	4	Viewing Choice activity = Q1	278	0.03
	5	Commencing Quiz Attempt = Q1	269	0.03

Table 5 continued

Method	No.	Subgroups of failure	Subgroup size	Quality (OR/UN)
	6	Viewing Choice activity = Q1 AND Adding Post in Forum = Q1	203	0.02
	7	Commencing Quiz Attempt = Q1 AND Viewing Wiki activity = Q1	212	0.02
	8	Viewing Choice activity = Q1 AND Viewing Wiki activity = Q1	178	0.02
	9	Viewing File Resource = Q1	180	0.01
	10	Reviewing Quiz activity = Q1	191	0.01

extracted by DSSD represents that having lower participation in visiting course home page and in adding discussion in forum leads students to secure a poor grade. Another interesting finding can be seen from the fifth subgroup which states that students, who do not have a very high participation in adding post in forum and in editing course wiki, are most likely to fail.

Similar to DSSD, NMEEF-SD has extracted that students, having lower involvement in viewing and contributing to different resource and activities, are most likely to fail. The first subgroup extracted by this method (Table 5) states that students, who have lower frequency in visiting course home page, got a very high chance to fail. It is also found that having lower participation in editing course wiki or adding post in forum lead students to become unsuccessful followed from subgroups 2 and 3, respectively.

BSD and SD-Map have discovered that those students, who have lower participation in viewing and contributing to course activities and resources, are the low performers. As followed from the first subgroup, it is seen that students, who has lower contribution to participate in forum discussion, obtain a poor grade. This method also found that students having lower view in course home page and choice activity fail in that course as depicted from subgroup 4. All of these factors have been identified by DSSD, while most of them found by SNS and NMEEF-SD.

Alike to DSSD, NMEEF-SD and BSD, APRIORI-SD has revealed that both student participation frequency in viewing and contributing to course resources and activities affect their academic performance. The first subgroup found by this method states that students having lower visits in course home page and lower attempts in commencing quiz activity are most likely to be the low achievers.

The findings of this experiment have revealed that lower engagement in certain activities and resources lead to a poor grade. For example, the experiments on Moodle dataset show that a student who has lower participation in discussion forum got a very high chance to fail in that course. Using such information, a course teacher should direct his/her attention to the group of students who have a very high chance to fail and also encourage them to participate in such activities because of their strong association with student academic performance in a course.

4.3 Mining combined data

This part of experiment has identified subgroups that were not extracted when mining the enrolment and Moodle data separately. SNS revealed that students, possessing specific academic and demographic features and having lower participation in different Moodle activities, got higher chance to fail. The first subgroup in Table 6 states that students, who are external and have lower participation in adding discussion in forum, are most likely to fail. Another interesting finding can be observed from the third subgroup which states that students who are part-time and have lower participation in editing course wiki are most likely low performers. The experiment shows that those external students, who have lower contribution to activities of different modules, have a higher possibility to fail. The last subgroup extracted by this method reveals that male students, who have lower frequency in adding post in forum, have a very high chance to fail.

DSSD demonstrated that students, not from some specific country, or state and have less engagement in viewing and contributing to different activities, have higher possibility to fail. SNS has discovered that students, who possess specific demographic features and participate less in different Moodle activities, are most likely to fail; however, it was unable to find that a student's ethnic origin has influence on his/her academic performance. The second subgroup found by DSSD states that students, who are not from Ireland and have lower participation in viewing wiki activity, are most likely low achievers.

NMEEF-SD showed that a student, who possesses certain academic features and is less involved in different Moodle activities, has higher possibility to fail. Most of these factors were not discovered by the other methods. The first subgroup represents that students, who did not get admitted in multiple programs in any year of their study and had lower score in adding discussion in forum, have a very high chance to fail. Although a number of new factors have been found by this method, none of these are able to present any useful indicators of student performance.

Similar to NMEEF-SD, BSD and SD-Map found that students with specific admission features and having lower engagement in viewing certain Moodle activities are most

Table 6 Top-10 subgroups discovered by mining combined dataset

Method	No.	Subgroups of failure	Subgroup size	Quality (OR/UN)
SNS	1	Attendance mode = External AND Adding Discussion in Forum = Q1	87	12.80
	2	Admission Basis = VET (Non-Secondary) AND Viewing Book Resource = Q1	56	11.54
	3	Attendance type = Part-time AND Editing course Wiki = Q1	76	10.75
	4	Attendance mode = External and Adding Post in Forum = Q1	91	10.20
	5	Attendance mode = External AND Editing course Wiki = Q1	90	9.79
	6	Attendance type = Part-time AND Adding Discussion in Forum = Q1	64	8.26
	7	High School State = SA AND Viewing Book Resource = Q1	78	7.01
	8	Admission Basis = Mature Age Entry AND Viewing Book Resource = Q1	41	6.88
	9	Attendance type = Part-time AND Viewing File Resource = Q2	46	5.85
	10	Gender = Male AND Adding Post in Forum = Q1	58	4.75
DSSD	1	Application was first preference != No AND Visiting Course home page = Q1	240	0.08
	2	Birth Country != Ireland AND Viewing Wiki activity = Q1	242	0.08
	3	High School State != ACT AND Adding Post in Forum = Q1	236	0.08
	4	Attendance type!= Fulltime AND Visiting Course home page = Q1	226	0.07
	5	Application was first preference != No AND Viewing Book resource != Q3	218	0.07
	6	Attendance mode != Internal AND Viewing Quiz activity = Q1	206	0.07
	7	ATAR != Low AND Viewing Forum activity = Q1	213	0.07
	8	Birth Country != France AND Viewing Forum activity != Q2	286	0.07
	9	Social economic status != High AND Visiting Course home page != Q4	289	0.07
	10	High School State != NSW AND Visiting Course home page != Q3	298	0.06
NMEEF-SD	1	In multi program any year = No AND Adding Discussion in Forum = Q1	189	0.06
	2	Program type = Bachelor's Pass AND Viewing Wiki activity = Q1	180	0.06
	3	In multi program this year = No AND Viewing Wiki activity = Q1	210	0.06
	4	Program type = Bachelor's Pass AND Viewing Quiz activity = Q1	232	0.06
	5	In multi program any year = No AND Adding Post in Forum = Q1	168	0.05
	6	Program type = Bachelor's Pass AND Adding Discussion in Forum = Q1	175	0.05
	7	In multi program any year = No AND Editing course Wiki = Q1	154	0.05
	8	In multi program this year = No AND Viewing Forum activity = Q1	190	0.05
	9	Application was first preference = Yes AND Adding Post in Forum = Q1	202	0.05
	10	Program type = Bachelor's Pass AND Editing course Wiki = Q1	230	0.04
BSD/SD-Map	1	Program type = Bachelor's Pass AND Visiting Course home page = Q1	264	0.04
	2	HECS exempt type = Deferral AND Visiting Course home page = Q1	251	0.04
	3	In multi program this year = No AND Visiting Course home page = Q1	284	0.04
	4	Application was first preference = Yes AND Visiting Course home page = Q1	276	0.04
	5	In multi program any year = No AND Visiting Course home page = Q1	275	0.04
	6	Birth Country = Australia AND Visiting Course home page = Q1	296	0.04
	7	Program type = Bachelor's Pass AND Viewing Discussion in Forum = Q1	296	0.04
	8	Application was first preference = Yes AND Viewing Discussion in Forum = Q1	296	0.03
	9	HECS exempt type = Deferral AND Viewing Discussion in Forum = Q1	296	0.03
	10	In multi program any year = No AND Viewing Discussion in Forum = Q1	296	0.03
APRIORI-SD	1	HECS exempt type = Deferral AND Viewing Forum activity = Q1	210	0.04
	2	Program type = Bachelor's Pass AND Viewing Wiki activity = Q1	178	0.04
	3	Application was first preference = Yes AND Viewing Choice activity = Q1	211	0.04
	4	In multi program this year = No AND Viewing Wiki activity = Q1	232	0.04
	5	Application was first preference = Yes AND Viewing Forum activity = Q1	169	0.03

Table 6 continued

Method	No.	Subgroups of failure	Subgroup size	Quality (OR/UN)
	6	Birth Country = Australia AND Viewing Discussion in Forum = Q1	196	0.03
	7	In multi program any year = No AND Viewing File resources = Q1	202	0.03
	8	Gender = M AND Viewing Forum activity = Q1	160	0.02
	9	Age = Mature AND Adding Post in Forum = Q1	175	0.02
	10	Application was first preference = Yes AND Viewing Lesson activity = Q1	168	0.02

likely to obtain a lower grade. This method has identified that if a student is not admitted into multiple program and has lower visits in course home page, he/she is most likely to fail as followed from subgroup 3.

APRIORI-SD discovered that students possessing specific demographic and academic features and participate less in different Moodle activities are prone to fail. As illustrated from subgroup 8, this method found that students who are male and have lower contribution in viewing forum are the low performers. Although most of the factors have been found by all other methods, student gender has been only identified by SNS.

The results obtained from the previous experiments revealed that both student enrolment and Moodle features affect their academic performance. However, the experiments on combined dataset have discovered that students possessing specific demographic features and participating less in different Moodle activities are highly prone to fail. Such finding is helpful for the educational institutions to detect the at-risk students more accurately. For example, consider the first subgroup discovered by SNS which states that external students, who have lower participation in contributing to the discussion forum, are most likely to be low performers. Similar findings can be observed from the sixth subgroup which represents that part-time students, who have a lower participation in adding discussion in forum, have a higher chance to fail. The institution may suggest these groups of students to attend a course internally or full-time if they study onshore and may provide them with financial aid to support their study. However, if they study offshore, the institution can direct the course teachers to monitor their progress in participating in different course activities and provide valuable suggestions when needed.

5 Discussion

The methods employed in the present study identified various subgroups associated with student academic performance. According to the discovered subgroups found by the methods, SNS found higher number of factors as compared to the other methods as indicated in Fig. 1. A number of these factors are common across different methods as illustrated in Fig. 2.

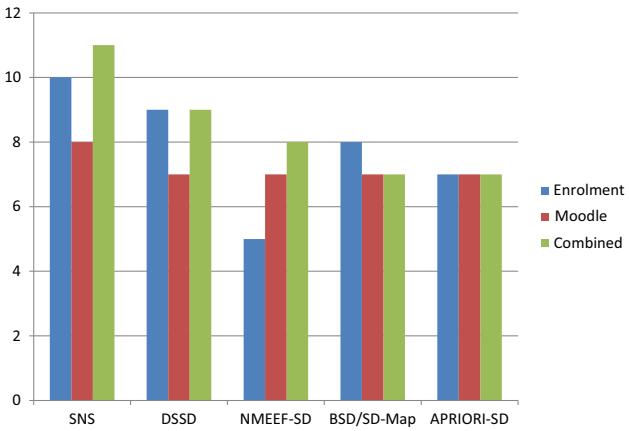

Fig. 1 Number of factors found by different methods

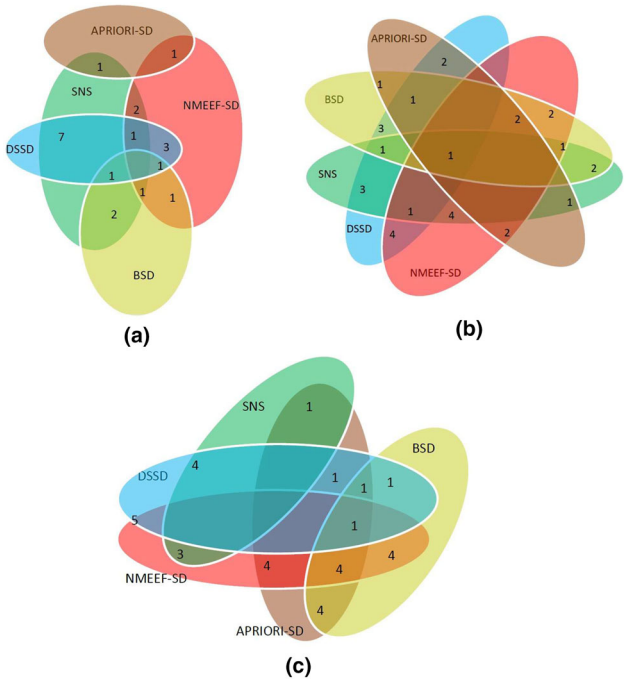

Fig. 2 Common factors found across different methods. **a** Enrolment, **b** Moodle, **c** combined

5.1 Performance of subgroup discovery methods in identifying useful and interpretable subgroups

Although some interesting subgroups have been picked up by all the methods, some of them produced conflicting results. Consider subgroup 8 (Table 4) discovered by DSSD which states that normal-aged students are most likely to fail. This finding contradicts with the result obtained by SNS, stating that mature-aged students have got higher probability to fail (subgroup 3). Similar surprising discovery has been made by NMEEF-SD by subgroup 4 which identifies that full-time students have higher chance to fail. The reason behind such discovery is that both DSSD and NMEEF-SD evaluate a subgroup by measuring its distributional unusualness. According to this property, the occurrence frequency of a subgroup contributes positively to be of high quality. Most of the bachelor students are full-time and normal-aged; hence, these subgroups are found by these methods because of their frequent occurrence in the dataset.

For all the three datasets, DSSD generates a number of non-interpretable subgroups. For example, consider the first subgroup of Table 4 which identifies that the students, whose ATAR is not high and admission basis is other than higher education course, are most likely to fail. A student's ATAR may be low, medium or good apart from being high. Similar discovery was found from the experiment of Moodle data. Subgroup 7 followed from Table 5 states that students, not having medium participation in viewing file resources and adding post in forum, prone to fail. The experiments on the combined dataset also discovered some incomprehensible subgroups. Followed from Table 6, subgroup 9 states that when a student's social economic status is not high and his/her participation in visiting course home page is not high, he/she is most likely to fail. Although the use $!=$ operator of this method helps to get rid of a number of tuples; it makes the presentation of result obscure.

The quality of the subgroups discovered by different methods are evaluated in terms of the criteria discussed in Sect. 3.3.2. This study employed different types of criteria to compare the results in terms of the generated subgroups. Table 7 presents the performance of different methods for the top-10 subgroups. From the table, it is seen that DSSD generated subgroups with higher average length, while the subgroups discovered by APRIORI-SD are shorter in length. Moreover, APRIORI-SD generates fewer rules than the other methods. Hence, this method achieves higher simplicity in terms of the discovered knowledge. The results also show that SNS does not generate any redundant subgroup, while NMEEF-SD generates quite a high number of redundant subgroups. From Tables 4, 5 and 6, we can observe that except DSSD, the knowledge represented by the subgroups generated by all the methods is understandable.

Table 7 Performance of SNS, DSSD and NMEEF-SD in terms of the generated subgroups

Dataset	Method	No. of subgroups	Subgroup length (average)	Unusualness (average)	Redundancy (%)	Comprehensibility
Enrolment	SNS	23	1.7	0.01	0	High
	DSSD*	–	1.9	0.05	10	Medium
	NMEEF-SD	24	1.5	0.03	50	High
	BSD**	–	1.9	0.03	2	High
	SD-Map**	–	1.9	0.03	2	High
	APRIORI-SD	18	1.4	0.02	1	High
Moodle	SNS	29	1.4	0.01	0	High
	DSSD	–	2.0	0.06	8	Medium
	NMEEF-SD	30	1.6	0.05	60	High
	BSD	–	1.9	0.03	1	High
	SD-Map	–	1.9	0.03	1	High
	APRIORI-SD	20	1.0	0.02	3	High
Combined	SNS	23	2.0	0.01	0	High
	DSSD	–	2.0	0.07	6	Medium
	NMEEF-SD	26	2.0	0.05	10	High
	BSD	–	2.0	0.06	0	High
	SD-Map	–	2.0	0.04	0	High
	APRIORI-SD	21	1.3	0.03	2	High

*This is a beam search-based method. So the number of generated subgroups is determined by beam width.
**These are the top-k methods; hence, the number of rules generated depends on k value

5.2 Predicting ability of different methods in classifying students according to their academic outcome

The factors represented by the discovered subgroups are helpful for identifying the vulnerable students and improving their performance by taking further actions. Therefore, it is very crucial to evaluate their usefulness in decision-making. In this regard, the discovered subgroups are tested in terms of their predictive ability in classifying a new student according to their academic outcome by using two measures—precision and recall (Eqs. 5, 6). It is depicted from Table 8 that BSD and SD-Map attain higher precision for the enrolment dataset, while DSSD achieves a higher precision for the Moodle and combined datasets. It was found from the experiment of Moodle datasets that 60% of students, who participate less in different Moodle activities, are predicted as unsuccessful. The use of ! = operator helps to get rid of a number of tuples containing those attribute-values more likely to be associated with the alternate class (pass).

NMEEF-SD attains a very high recall for all the datasets. This is due to the fact that this method generates shorter subgroups and hence covers a large proportion of the target examples. SNS obtains higher recall for the top-20 subgroups, which was observed from the result, stating that above 75% of the students, possessing the features containing in the discovered subgroups, are predicted low achievers. It is also observed from the result that the top-10 and top-20 subgroups discovered by DSSD and NMEEF-SD show same precision and recall rate. It depicts the fact that the low ranked subgroups produced by these methods do not provide any new knowledge which can also be followed from Table 7, stating that these methods generate a number of redundant subgroups. It has been found that APRIORI-SD attains lower precision but higher recall for top-20 subgroups as compared to the top-10 subgroups. On the contrary, SNS achieves higher precision but lower recall for top-10 subgroups than the top-20 subgroups.

Table 8 Precision and recall of top-10 and top-20 subgroups achieved by different methods

K	Dataset	Method	Precision	Recall
10	Enrolment	SNS	0.33	0.37
		DSSD	0.25	1
		NMEEF-SD	0.21	0.8
		SD-Map	0.42	0.29
		BSD	0.42	0.29
		APRIORI-SD	0.21	0.43
	Moodle	SNS	0.35	0.8
		DSSD	0.6	0.64
		NMEEF-SD	0.24	1
		SD-Map	0.21	0.6
		BSD	0.21	0.6
		APRIORI-SD	0.22	0.47
	Combined	SNS	0.38	0.613
		DSSD	0.76	0.44
		NMEEF-SD	0.25	1
		SD-Map	0.51	0.78
		BSD	0.51	0.78
		APRIORI-SD	0.26	0.52
20	Enrolment	SNS	0.23	0.9
		DSSD	0.25	1
		NMEEF-SD	0.21	0.8
		SD-Map	0.43	0.26
		BSD	0.43	0.26
		APRIORI-SD	0.19	0.46
	Moodle	SNS	0.24	1
		DSSD	0.6	0.64
		NMEEF-SD	0.24	1
		SD-Map	0.2	0.55
		BSD	0.2	0.55
		APRIORI-SD	0.21	0.5
	Combined	SNS	0.26	0.76
		DSSD	0.76	0.44
		NMEEF-SD	0.25	1
		SD-Map	0.5	0.76
		BSD	0.5	0.76
		APRIORI-SD	0.24	0.56

5.3 Major findings based on different datasets

It is followed from the experiment that there is no single method that achieves the best results in all aspects. In this regard, the common knowledge found across different methods should be used for decision-making. From the experiment of enrolment dataset, different methods found that students' socio-demographic features, i.e. family economic status and parent education level, have significant influence on their educational achievement. It is also noticed that different academic features such as ATAR and admission basis also affect their study outcome. For example, it was found that students, who got admission based on special entry requirements, e.g. mature age entry, or professional qualification, are most likely to be unsuccessful. This may be due to the fact that they are not regular students and hence less involved in their studies.

The mining of Moodle data has revealed that the degree of participation in different Moodle activities affects student performance in a course; hence, they work as predictors of their academic outcome. The result shows that lower participation in viewing different resources (e.g. course home page and file resources) and in viewing or contributing to differ-

ent course activities such as discussion forums leads students to achieve a poor grade in the courses. In fact, a discussion forum contains student discussions regarding the course contents. A student may get benefited by viewing or participating in the discussions as they may be helpful for sharing different problems and resolving doubts by receiving feedback. Hence, less participation in these activities may make them less knowledgeable about the course.

From the experiments of enrolment or Moodle datasets, we are able to find some interesting subgroups; however, the results are in view of either student enrolment or Moodle features. The experimental results on the combined dataset have shown that those students, who possess specific demographic and academic features and participate less in certain Moodle activities, are the low achievers. For example, it is identified that those external or part-time students, who have lower participation in adding post or discussion in forum or editing course wiki, are most likely to fail. This is because of the fact that external or part-time students may be involved in some other works, i.e. job, and hence less focused in their study.

6 Conclusion

This work has employed different subgroup discovery methods to analyse students' enrolment and Moodle data separately as well as jointly to find potential factors influencing student academic performance. The result shows that a number of socio-demographic as well as academic and course assessment features affect student academic performance. The resulting enrolment factors will be helpful for the educational institutions to undertake approaches for early identification of vulnerable students and provide them with additional academic support. After learning the significant Moodle factors, course teachers may promote those activities as key course assessment criteria. The analysis on the combined dataset has revealed some interesting and useful information which was not observed when considering demographic and Moodle features separately. The outcome may not be able to detect vulnerable student at an early stage, but the course teachers can direct their attention to those students who hold certain socio-demographic and academic characteristics as possessed by the unsuccessful students.

From the experiments, it was found that there is not an individual method that shows the best performance in terms of all the evaluation criteria. It has been found that APRIORI-SD generated subgroups are simple in terms of presentation of knowledge. The subgroups discovered by BSD/SD-Map covers a large proportion of examples. SNS performs best in terms of generating non-redundant subgroups, DSSD achieves higher precision and NMEEF-SD performs best in terms of attaining high recall. Therefore in practice, it is useful to employ a number of different methods in order to achieve comprehensive results.

Further research will develop a student performance prediction model by considering different influencing features of student outcome and also evaluate the usefulness of knowledge discovered by them.

References

1. BlackBoard. http://www.blackboard.com/. Accessed 05 Mar 2018
2. Desire2Learn. http://www.brightspace.com/. Accessed 05 Mar 2018
3. Moodle. https://moodle.org/. Accessed 05 Mar 2018
4. Orange. https://orange.biolab.si/. Accessed 05 Mar 2018
5. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
6. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Mult. Valued Logic Soft Comput. **17**(2–3), 255–287 (2011)
7. Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-oriented community detection using exhaustive subgroup discovery. Inf. Sci. **329**, 965–984 (2016)
8. Atzmueller, M., Lemmerich, F.: VIKAMINE—Open-source subgroup discovery, pattern mining, and analytics. In: Proceedings of ECML/PKDD 2012: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Heidelberg, Germany (2012)
9. Atzmueller, M., Puppe, F.: SD-Map–A Fast Algorithm for Exhaustive Subgroup Discovery, pp. 6–17. Springer, Berlin (2006)
10. Blagojević, M.: Živadin Micić: a web-based intelligent report e-learning system using data mining techniques. Comput. Electr. Eng. **39**(2), 465–474 (2013)
11. Borkar, S., Rajeswari, K.: Predicting students academic performance using education data mining. Comput. Sci. Mobile Comput. **2**, 273–279 (2013)
12. Brito, P.Q., Soares, C., Almeida, S., Monte, A., Byvoet, M.: Customer segmentation in a large database of an online customized fashion business. Robot. Comput. Integr. Manuf. **36**, 93–100 (2015)
13. Carmona, C., Chrysostomou, C., Seker, H., del Jesus, M.: Fuzzy rules for describing subgroups from influenza a virus using a multi-objective evolutionary algorithm. Appl. Soft Comput. **13**(8), 3439–3448 (2013)
14. Carmona, C., González, P., García, B., del Jesus, M., Aguilera, J.: Mefes: an evolutionary proposal for the detection of exceptions in subgroup discovery. An application to concentrating photovoltaic technology. Knowl. Based Syst. **54**, 73–85 (2013)
15. Carmona, C., Ruiz-Rodado, V., del Jesus, M., Weber, A., Grootveld, M., González, P., Elizondo, D.: A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans. Inf. Sci. **298**, 180–197 (2015)
16. Carmona, C.J., González, P., del Jesus, M.J., Herrera, F.: NMEEF-SD: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans. Fuzzy Syst. **18**, 958–970 (2010)
17. Carmona, C.J., González, P., del Jesus, M.J., Romero, C., Ventura, S.: Evolutionary algorithms for subgroup discovery applied to e-learning data. In: IEEE EDUCON 2010 Conference, pp. 983–990 (2010)

18. Carmona, C.J., González, P., del Jesus, M.J., Ventura, S.: Subgroup discovery in an e-learning usage study based on moodle. In: 7th International Conference on Next Generation Web Services Practices, pp. 446–451 (2011)
19. Chen, C.M., Chen, M.C.: Mobile formative assessment tool based on data mining techniques for supporting web-based learning. Comput. Educ. **52**(1), 256–273 (2009)
20. Dominguez, A.K., Yacef, K., Curran, J.R.: Data mining for individualised hints in elearning. In: Proceedings of the International Conference on Educational Data Mining, pp. 91–100 (2010)
21. Fleiss, J.: Statistical Methods for Rates and Proportions Rates and Proportions. Wiley, New York (1973)
22. Gamberger, D., Lavrač, N.: Expert-guided subgroup discovery: methodology and application. J. Artif. Intell. Res. **17**, 501–527 (2002)
23. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, 1989. Addison Wesley, Reading (1989)
24. Goldsmith, B.R., Boley, M., Vreeken, J., Scheffler, M., Ghiringhelli, L.M.: Uncovering structure-property relationships of materials by subgroup discovery. New J. Phys. **19**(1), 13–31 (2017)
25. Gray, G., McGuinness, C., Owende, P.: An application of classification models to predict learner progression in tertiary education. In: Advance Computing Conference (IACC), 2014 IEEE International, pp. 549–554 (2014)
26. Grosskreutz, H., Stefan, R.: On subgroup discovery in numerical domains. Data Min. Knowl. Discov. **19**(2), 210–226 (2009)
27. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
28. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc, Burlington (2005)
29. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. SIGMOD Rec. **29**(2), 1–12 (2000)
30. Helal, S.: Subgroup discovery algorithms: a survey and empirical evaluation. J. Comput. Sci. Technol. **31**(3), 561–576 (2016)
31. Herrera, F., Carmona, C.J., González, P., del Jesus, M.J.: An overview on subgroup discovery: foundations and applications. In: Knowledge Information system, pp. 495–525 (2011)
32. Holzhüter, M., Frosch-Wilke, D., Klein, U.: Exploiting Learner Models Using Data Mining for E-Learning: A Rule Based Approach, pp. 77–105. Springer, Berlin (2013)
33. Hsieh, T.C., Wang, T.I.: A mining-based approach on discovering courses pattern for constructing suitable learning path. Exp. Syst. Appl. **37**(6), 4156–4167 (2010)
34. Jin, N., Flach, P., Wilcox, T., Sellman, R., Thumim, J., Knobbe, A.: Subgroup discovery in smart electricity meter data. IEEE Trans. Ind. Inform. **10**(2), 1327–1336 (2014)
35. Jovanoski, V., Lavrač, N.: Classification rule learning with APRIORI-C. In: Proceedings of the 10th Portuguese Conference on Artificial Intelligence, pp. 44–51 (2001)
36. Jovanovic, M., Vukicevic, M., Milovanovic, M., Minovic, M.: Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. Int. J. Comput. Intell. Syst. **5**(3), 597–610 (2012)
37. Kardan, S., Conati, C.: A framework for capturing distinguishing user interaction behaviours in novel interfaces. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 126–138 (2012)
38. Kavšek, B., Lavrač, N., Jovanoski, V.: APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery, pp. 230–241. Springer, Berlin (2003)
39. Kavšek, B., Lavrač, N.: Apriori-SD: adapting association rule learning to subgroup discovery. In: Advances in Intelligent Data Analysis V, pp. 543–583 (2006)
40. Khan, T.M., Clear, F., Sajadi, S.S.: The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 226–229 (2012)
41. Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)
42. Lara, J.A., Lizcano, D., Martínez, M.A., Pazos, J., Riera, T.: A system for knowledge discovery in e-learning environments within the european higher education area: application to student data from open university of madrid, udima. Comput. Educ. **72**, 23–36 (2014)
43. Lavrac, N., Flach, P.A., Zupan, B.: Rule evaluation measures: a unifying view. In: Proceedings of the 9th International Workshop on Inductive Logic Programming, pp. 174–185 (1999)
44. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. J. Mach. Learn. Res. **5**, 153–188 (2004)
45. Leeuwen, M.V., Knobbe, A.: Diverse subgroup set discovery. Data Min. Knowl. Discov. **25**(2), 208–242 (2012)
46. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast exhaustive subgroup discovery with numerical target concepts. Data Min. Knowl. Discov. **30**(3), 711–762 (2016)
47. Lemmerich, F., Ifl, M., Puppe, F.: Identifying influence factors on students success by subgroup discovery. In: Educational Data Mining 2011, pp. 345–346 (2011)
48. Li, J., Liu, J., Toivonen, H., Satou, K., Sun, Y., Sun, B.: Discovering statistically non-redundant subgroups. Knowl. Based Syst. **67**, 315–327 (2014)
49. Lockyer, L., Heathcote, E., Dawson, S.: Informing pedagogical action: aligning learning analytics with learning design. Am. Behav. Sci. **57**(10), 1439–1459 (2013)
50. Macfadyen, L.P., Dawson, S.: Mining lms data to develop an "early warning system" for educators: a proof of concept. Comput. Educ. **54**(2), 588–599 (2010)
51. Marschark, M., Shaver, D.M., Nagle, K.M., Newman, L.A.: Predicting the academic achievement of deaf and hard-of-hearing students from individual, household, communication, and educational factors. Except. Child. **81**(3), 350–369 (2015)
52. Mwalumbwe, I., Mtebe, J.: Using learning analytics to predict students' performance in moodle learning management system: a case of mbeya university of science and technology. IEEE Trans. Learn. Technol. **79**, 1–13 (2017)
53. Elakia, G., Aarthi, N.J.: Application of data mining in educational database for predicting behavioural patterns of the students. Int. J. Comput. Sci. Inf. Technol. **5**(3), 4469–4472 (2014)
54. Natek, S., Zwilling, M.: Student data mining solution knowledge management system related to higher education institutions. Exp. Syst. Appl. **41**(14), 6400–6407 (2014)
55. Natu, M., Palshikar, G.K.: Interesting Subset Discovery and Its Application on Service Processes, pp. 245–269. Springer, Berlin (2014)
56. Ogor, E.N.: Student academic performance monitoring and evaluation using data mining techniques. In: Proceedings of the Electronics, Robotics and Automotive Mechanics Conference, pp. 354–359 (2007)
57. Perry, J.W., Kent, A., Berry, M.M.: Machine literature searching x. machine language; factors underlying its design and development. Am. Doc. **6**(4), 242–254 (1955)
58. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc, San Francisco (1993)
59. Romero, C., Espejo, P.G., Zafra, A., Romero, J.R., Ventura, S.: Web usage mining for predicting final marks of students that use moodle courses. Comput. Appl. Eng. Educ. **21**(1), 135–146 (2013)
60. Romero, C., González, P., Ventura, S., del Jesus, M., Herrera, F.: Evolutionary algorithms for subgroup discovery in e-learning: a practical application using moodle data. Exp. Syst. Appl. **36**(2, Part 1), 1632–1644 (2009)

61. Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Comput. Educ. **68**, 458–472 (2013)
62. Shaw, R.S.: A study of the relationships among learning styles, participation types, and performance in programming language learning supported by online forums. Comput. Educ. **58**(1), 111–120 (2012)
63. Tair, M.M.A., El-halees, A.M.: Mining educational data to improve students' performance: a case study. Inf. Commun. Technol. Res. **2**, 140–146 (2012)
64. Thiele, T., Singleton, A., Pope, D., Stanistreet, D.: Predicting students' academic performance based on school and socio-demographic characteristics. Stud. High. Educ. **41**(8), 1424–1446 (2016)
65. Wei, H.C., Peng, H., Chou, C.: Can more interactivity improve learning achievement in an online course? Effects of college students' perception and actual use of a course-management system on their learning achievement. Comput. Educ. **83**, 10–21 (2015)
66. Wrobel, S.: An Algorithm for multi-relational discovery of subgroups. In: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, pp. 78–87 (1997)
67. Yoo, J., Kin, J.: Predicting learner's project performance with dialogue features in online Q and A discussions. In: Intelligent Tutoring Systems ITS, pp. 570–575 (2012)
68. Zacharis, N.Z.: A multivariate approach to predicting student outcomes in web-enabled blended learning courses. Internet High. Educ. **27**, 44–53 (2015)
69. Zheng, B., Warschauer, M.: Participation, interaction, and academic achievement in an online discussion environment. Comput. Educ. **84**, 78–89 (2015)