Educational data mining for decision-making: a framework based on student development theory

Xiao-Feng Lei[1,2]
[1]School of Ideological and Political Theory, Beihang University
Beijing, 100191, China
[2]Graduate School of Education, Peking University
Beijing, 100871, China
E-mail: leixf@buaa.edu.cn

Ming Yang[1]
1School of Ideological and Political Theory, Beihang University
Beijing, 100191, China
E-mail:514138646@qq.com

Yi Cai[†,3]
3College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China
[†]E-mail: caiyi@mail.buct.edu.cn

Applying Educational Data Mining (EDM) for decision making is an emerging interdisciplinary research field. From the view of student development, this paper presents a framework for educational decision making, which can explore some laws and characteristics in student development to improve educational decision and quality. The paper also presents a case study to verify the effectiveness of the framework.

Keywords: educational data mining; decision making; student development theory.

1. Instruction

At present, using data to make decisions is not new [1]. It often happens in the business domains as well as the educational fields. For educational institutions, one of the biggest challenges is the exponential growth of educational data and the use of this data to improve the quality of managerial decisions [2]. As an increasingly emerging interdisciplinary area, Educational Data Mining (EDM) is concerned with developing methods to explore the unique types of data that come from educational environments [3]. One critical and prominent objective of data mining in education is to enable data-driven decision-making for

improving current educational practice and learning materials, and then for serving students' learning and boosting educational quality.

In educational environment, higher education including universities or colleges will find larger and wider applications for data mining than its counterpart in other educational sectors because higher educational institutions carry at least three duties that are data mining intensive: 1) teaching that concerns with the transmission of knowledge, 2) scientific research that relates to the creation of knowledge, and 3) institutional research that pertains to the use of knowledge for decision making [4]. It means that higher educational institutes seek more efficient technology to better manage and support decision making procedures or assist them to set new strategies and plan for a better management of the current processes [5].

In previous studies, data mining technology has been applied in higher education fields, such as improving students' academic performance [6][7], selecting courses, measuring their retention rate [8], and managing the grant fund of an institution. However, most of these studies mainly focus on the data of the educational processes or didactical contents, and few of them regard the data organized according to educational objects as the research view. These objects usually comprise student, academic faculty, administrative staff, and alumni. For example, if student is regarded as the research object, the term of student's development from enrollment to graduation in a university can be observed. Furthermore, the data of this term can be stored and organized based on student development theory to data mining for decision-making. It will benefit to explore some new laws and characteristics in student development to improve educational decision and quality.

In order to understand the progress that how data mining works for decision making from student's view, this paper presents a research framework that descripts an idea based on Student Development Theory Model. The paper is structured as follows. Section 2 presents the review of EDM including its concept, process, application methods and tools. Section 3 presents how EDM applies for decision-making. Section 4 presents the improved model of Student Development Theory. Section 5 presents a framework for decision-making, and Section 6 gives a case study of managerial decision making on the development of graduate students. Section 7 presents the conclusion and future work.

2. Educational data mining (EDM)

Since data mining technique has been applied in many domains such as business and biology, it seems not necessary to regard education as a specific research sector. However, educational data has several typical characteristics, such as multiple hierarchies, complexity in context, interaction among objects, and

transformation along with time. These data can be generated by any type of information system supporting education and administration. For instance, for a university, these data can include the background data of students, didactical data throughout a course or an academic quarter, daily data of recreation or sport activities, and so on. These characteristics lead to further consideration and research on the application of data mining in educational field.

As an interdisciplinary area, EDM is associated with at least three disciplines: computer science, education, and statistics. Conceptually, it is also closely related with knowledge discovery in databases (KDD), computer-based education, machine learning, and learning analytics (LA). According to the definition of the website of International EDM Society, EDM is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings in which they learn [9]. In short, it is the application of data mining techniques in educational environment as a specific knowledge discovery.

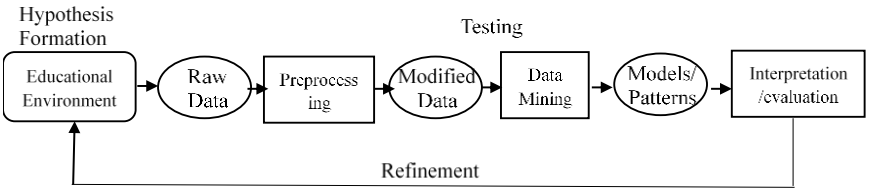

Fig. 1. Application Process of Educational Data Ming

Different from that of the general data mining, the application process of EDM can be seen as an iterative cycle of hypothesis formation, testing, and refinement (see Fig. 1) [3]. In this process, hypothesis formation and refinement present the educational views in particular. In fact, due to the presence of hypothesis formation and refinement, the iterative cycle of EDM looks like growing up to a robust educational system, which is not only to realize knowledge discovery, but also to feedback useful information of decision-making for better educational environment.

Hypothesis formation defines the educational object of data mining differing from other domains, and provides a prerequisite hypothesis model based on educational theories. Therefore, the research object of EDM is not a disordered dataset but a meaningful dataset of the educational entity. In this paper, Student Development Model is presented as the hypothesis formation.

Refinement is feedback to propose educational decisions to improve or optimize the educational environment. It means that refinement involves some matters of decision-making methods to make EDM more efficient and accurate.

At present, there are a number of popular methods within EDM. Some of them are widely acknowledged to be universal across types of data mining, such as prediction, clustering, relationship mining, outlier detecting, social network association (SNA), process mining, and text mining. Based on these popular methods, a lot of general free and commercial DM tools and frameworks emerged to be used for mining datasets from any domain or research area [10], such as Weka, SAS Enterprise Miner, SPSS Clementine, IBM Intelligent Miner, and DB Miner.

In Romero's study [3], some other methods, particularly for EDM including the distillation of data for human judgment, discovery with models, knowledge tracing (KT) and nonnegative matrix factorization, were concluded. Correspondingly, there are some tools to specify these methods to solve different educational problems. In the case study of this paper, data mining is run by Weka because the data is formed with Student Development Model but not complex.

3. Educational Data Mining for Decision Making

As a thought behavior of human being, the process of decision making generally includes the following procedures: problem finding, goal settlement, matter materials analysis, decision formation and implement. Each step content of this process may differ from various domains that demonstrate essential differences in natural laws or rules. Such as in educational fields, the process presents the characteristics of decision according to the educational phenomena and laws.

In educational areas, there are large amounts of educational materials to be collected for decision making, such as administrative documents, course syllabus, academic scores, and record data of students' activities. Among these materials, educational data has been applied to decision making, which is also called data-driven decision making (DDDM) in education [11].

Data may be used for several purposes according to the different demand levels that influence the nature of the decision making process. Teachers can use assessment data to reflect on their own teaching practice, to create and review intervention strategies for individuals, and to improve students' academic performance. Students can use self-evaluation data to manage study habit, to boost academic or sport exercises, and to select enrollment in or drop from education. Data can also be used to help administrative policymakers or leaders to response the demands from teachers and students, monitor educational environment, react timely from class deviation, and evaluate the quality of education.

Nowadays, data in higher educational fields has been placed squarely on the contemporary agenda [12]. A number of papers had focused on the higher educational data to improve the efficiency and effectiveness of the higher

educational process. In this paper, relevant data of student development in a college or university is present at administrative level in order to guide a range of decisions to help improve the success and development of students.

Educational data cannot directly be used to decision making because data provides no judgment or interpretation, and no sustainable basis of action [13]. Usually analysis and interpretation of the data take place through a variety of methods to be adaptive with decision making. Therefore, data mining with the features of information technology happens to help improve decision making in education.

From the view of educational research, there are two popular accesses including quantitative orientation and qualitative orientation. These two accesses explore the laws in education. Obviously, the process of EDM pertains to quantitative access that presents mining outcomes being used to some qualitative analysis and interpretation. Furthermore, several or a number of interpretations have come to a decision profile that are ready for the policymakers to make the final selection. In fact, a same process of EDM may propose various, even contrary outcome interpretations to policymakers, who can make the appropriate, even not optimal decision with the help of educational laws, experts, or previous experiences.

4. An Improved Model of Student Development Theory

There are two models: 1) Inputs-Environment-Outcomes model, 2) Student Involvement Theory, to explain the development of students in higher educational environment [14-15]. Those two models contribute to student development theory in the interaction of student and educational environment, and also promotes the rapid progress of quantitative research within student development theory.

4.1. *Inputs-Environment-Outcomes Model (I-E-O Model)*

From the view of higher educational assessment, an educational project or program can be regarded as a model including data with three variables: 1) student inputs, 2) student outcomes, and 3) the educational environment to which the student is exposed (See Fig. 2).

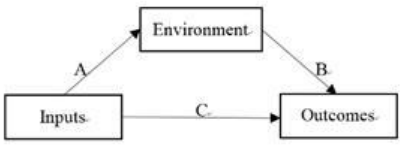

Fig. 2. Inputs-Environment-Outcomes model

Inputs refers to those personal qualities the student brings initially to the educational project, such as the student's initial level of developed talent at the time of entry, individual characteristics including gender, race, family background, educational experiences, and also personal expectation to the future project.

Outcomes refers to the talents we are trying to develop through the current project. It can be descripted through three dimensions: type of outcomes, type of data, and time. And then, all the educational outcomes including incremental knowledge, school retention, academic certificate, career orientation, can be classified into several different categories.

Environment refers to the student's actual experiences during the educational project. In fact, environment is most complex among three variables. It may be depicted through a number of ways. For example, if the educational project is a university, environment can be divided into two aspects: between-university environment and within-university environment, by which the university environment can be presented. Another example, if an educational system associated with students only, it is regarded as the project. Environment can focus on data to record the development activities of students.

The three arrows depict the relationships among the three classes of variables. Actually, the outcome variables are influenced not only by educational environment variables, but by the input variables.

4.2. *Student Involvement Theory*

It is a primary fact for student development theory that the student should not be treated as a kind of 'Black Box', and some mediating mechanism can explain how these educational programs and policies are translated into student achievement and development. The theory of student involvement reveals some truth hiding within the so-called 'Black Box'.

Student development is closely related with subject-matter, resource, and individual talent and investment, among which the theory of student involvement can provide a link from the view of student. Student involvement refers to the quantity and quality of the physical and psychological energy that students invest in the college experience. Of course, Involvement has both quantitative and qualitative features. And also, there are some evidences to show that more effective educational programs for student can promote student involvement, and to improve student achievement and development in the future.

In practice, some specific forms of involvement happen to influence the development of student, such as place of residence, honors programs, academic involvement, student-faculty interaction, athletic involvement, involvement in student government, and research on cognitive development. Policymakers and

leadership in a university can manage these factors of involvement to increase retention and other outcomes.

Although this theory emphasizes the importance of student involvement, and gradually turns 'Black Box' into 'White Box' in student development, necessarily, it must provide more evidences and applications to demonstrate the truth, and also pay more attention to inputs, outcomes and the interactions among them.

4.3. *An Improved Model*

The new improved model of student development is likely a collaboration between 'I-E-O' Model and student involvement theory (See in Fig. 3). But it draws on some advantages of both models and appeals to me for some reasons.

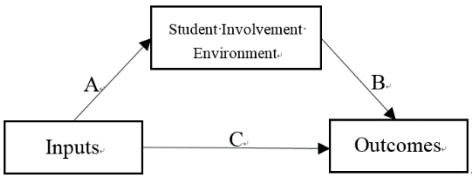

Fig. 3. An Improved Model

Firstly, the improved model better respects the subjectivity of student in educational involvement environment. And it provides clear data and information associated with student involvement to present a sophisticated environment. Secondly, the improved model is more dynamic. It means that the new model focuses on not only the static data and document in education, such as course contents and exam papers, but also the student activities and interactions among student peer and faculty. In addition, since student involvement theory places time of student as a kind of educational resource, the new model is also interested in the time characteristic of student development.

Of course, the improved model narrows the application range of 'I-E-O' Model, which may lead to lack of some evidences about student development. Moreover, from the view of educational research, the new model is inclined to quantitative study.

5. A Decision Framework Based on Student Development Theory

The educational object can be depicted based on Student Development Theory with data in higher education, and also the interpretation of data mining outcome can be used to decision making. Also, the educational object is defined by the 'hypothesis formation' step of educational data mining process. Therefore, this paper delineates a research framework including three parts: student development system, educational data mining, and decision process (See Fig. 4).

As the goal, this framework is to mine student development data, and then to help decision making in order to promote the development of student.

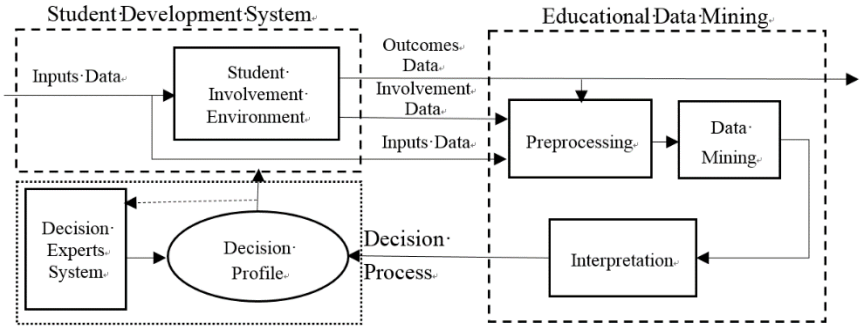

Fig. 4. A Decision Framework Based on Student Development Theory

5.1. *Student Development System*

As the higher educational hypothesis formation, this part is well organized by the improved model of student development theory. Among three variables depicting the system, both inputs and involvement are types of independent variables that are only influenced by educational decisions but not directly interfered by outcomes. Data related to inputs and involvement can be used for the following data mining. Outcomes is a type of dependent variable being affected by inputs and involvement. Especially, it is observed to demonstrate the effect of decision, and also to evaluate the quality of student development system. Outcomes data, also being used to data mining, is the quantitative presentation of the goal which is responding to this framework.

5.2. *Educational Data Mining*

In educational contexts, due to complexity and multiple sources, data from student development system must be preprocessed before it will be used to data mining. It is necessary to convert the data to an appropriate form for solving a specific student development problem. Normally, a number of variables/attributes with information about each student can be integrated into a summary table for better analysis. And then, according to the decision goal, data mining techniques including but not limited to classification, clustering, and association analysis techniques, can be applied into mining the data. In the end of this part, it is very important to show educational suggestions through interpreting the data mining results. In educational viewpoints, this step of interpretation is eventually a transform process between quantitative results to qualitative suggestions.

5.3. *Decision Process*

From the view of educational decision, outcome of educational data mining may provide some suggestions belonging to different aspects, such as administrative reform, faculty improvement, student inspiration, and alumni involvement. We collect all these suggestions to a decision profile, which is alike a number of educational decision papers presented on a desk in front of policymakers or leadership. In fact, not every policymaker or leadership is the educational master that can deal with these suggestions in a good way. Therefore, decision expert system including previous experience and knowledge, senior educational experts or committee, and computer-aid intelligent decision, can help policymakers or leadership to make the final decision. Furthermore, on the one hand, the final decision may be used to improve student development system. On the other hand, it may be recorded into decision expert system, and play a role to next round of decision making. As a result, the decision expert system can become more intelligent and robust as the decision framework is working in an iterative process.

6. A Case Study

This section presents a case study with a dataset that includes 2,266 instances about master graduate students in a Chinese university. There are 11 attributes with information about each graduate student (See Table 1). Obviously, the data of this case study is formed with three parts according to the improved model of student development theory. In order to be adaptive to the following data mining, the data is stored with the .csv file format.

Table 1. Eleven attributes with each instance.

Attribute Number	Part of Student Development Theory	Attribute Name	Type of Attribute	Data of Attribute
1	Inputs	Gender	Nominal	M=Male,F=Female
		Code of Gender		
2		Academic Type	Nominal	A1=Academic,A2=Professional
		Code of Academic Type		
3		School	Nominal	24 schools and their codes {S1,…,S24}
		Code of School		
4		Origin District	Nominal	District1, District2, District3 ,District4 and their codes {D1,…, D4}
		Code of Origin District		
5		Politics	Nominal	B1=Party Member,B2= Probationary Party Member;B3=Youth League;B4=Masses
		Code of Politics		
6		Ethnic	Nominal	56 ethnic groups in China and their code {M1,…,M56}
		Code of Ethnic		
7	Involvement	Average Score	Numeric	Centesimal
8		National Prize	Nominal	Y:Achieved,N:Not achieved
9		General Prize	Ordinal	1:First Prize,2:Second Prize,3:Third Prize,4:Fourth Prize,N:No
10	Outcomes	Job Place	Nominal	District1, District2, District3 ,District4 and their codes {D1,…, D4}
		Code of Job Place		
11		Job Type	Nominal	Job1, Job2,…,Job9 and their codes {J1,…,J9}
		Code of Job Type		

Additional Information:

1. Ditrict1 refers to the area including Beijing, Shanghai, and Tianjin; District2 refers to the area including several provinces located in eastern China; District3 refers to the area including several provinces located in middle China; District4 refers to the area including several provinces located in western China.

2. Job1 refers to the job provided by state-owned enterprises; Job2 is for scientific institutions; Job3 is for multiple-capital enterprises; Job4 is for government organs; Job5 is for educational institutions; Job6 is for medical and healthy institutions; Job7 is for some public services; Job8 is for army; Job9 is for other type of occupation.

6.1. *Educational goal for case study*

In this case study, it aims to set a classifier that can predict the job type of graduate students, and to discover the classifying rules, which may provide educational decision suggestions to improve the distribution of graduate students' job type.

6.2. *Data mining algorithm and tool*

Decision tree is widely used to generate a classifier that can predict a target attribute or single aspect of the data from some combination of other aspects of the data. In details, it is a flowchart-like structure in which each internal node represents a 'test' on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. At present, some algorithms, such as ID3 and C4.5, contribute to the application of decision tree for data mining. In this paper, we select pruned C4.5 algorithm to analysis the data of graduate students.

Although there are a variety of data mining tools being used in various fields, Weka is a good choice for researchers and educators. It is an open source software developed by The University of Waikato, New Zealand. With a collection of data mining algorithms, it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [16]. In Weka, we can easily find C4.5 algorithm with another label 'J48' in 'Classify'. J48 is for generating a pruned or unpruned C4.5 decision tree through editing properties in convenient ways. In this case study, after series of trials, we finally set the properties as follows: Cross-validation with 10 folds as Test Option, JobType as the Class, value 0.1 as Confidence Factor, number 3 as numFolds, and accepting all other default settings.

6.3. *Data mining results*

After data mining starts, it is really rapid that data mining results are appearing with text information in the Classifier Out (See examples in Fig. 5). Besides of the descriptive information about the case study, classifier model and result summary are most useful for interpretation. Classifier model reveals several rules that may predict the job type of graduate students. For example, in Figure 6, if 'School=S7', 'JobPlace=D3', and 'Gender=M', outcome is J2. In the educational view of the case study, it means that male graduate students in the seventh school have an aptitude to select a job serving for scientific institutions in middle China's provinces. Result summary presents the information of accuracy with the test. Although 45.2% of accuracy in this case is not enough good, to an extent, it can be beneficial to achieve some decision suggestions for policymakers or leadership.

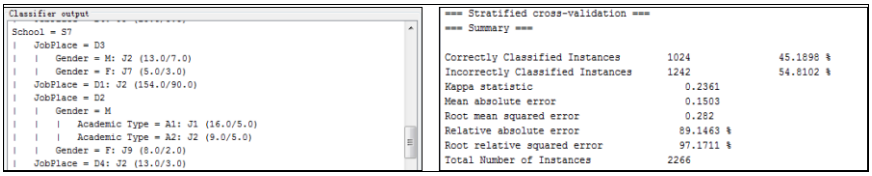

Fig. 5. Text result examples in the Classifier Out

In addition, a visualizer can be used to display the tree-like structure, in which ‘School’ attribute is the root node, and ‘JobPlace’ and ‘Politics’ attributes appear in the second level of internal nodes. Furthermore, some attributes including ‘Place’, ‘General Prize’, ‘National Prize’ and ‘Gender’ are in the third level, and ‘Academic Score’ is in the fourth level. However, we cannot find the attributes of ‘Ethnic’ and ‘Academic Type’ in this decision tree structure.

6.4. *Interpretation of result*

According to the classifying rules obtained from educational data mining, we can get to the following conclusions about the job type of master graduate students in this university:

1) Job type orientations of graduate students obviously differs among schools, where they had enrolled in. For example, most of students in the fifth school (S5) are favorite with jobs in scientific institutions (J2), and the majority of students in the ninth school (S9) focuses on positions in J9 (other types of occupations).

2) As one of two outcomes for the case student development system, the job type is closely related to another outcome, the job place. For example, students in School 4 seem either to select jobs for state-owned enterprises (J1) in Beijing, Tianjin and Shanghai, or to select jobs for scientific institutions (J2) in other districts (D2, D3, D4) of China.

3) In this case study, although the attributes about student involvement, such as ‘General Prize’, ‘National Prize’ and ‘Academic Score’ play an important role to the outcomes, school associated with majors greatly contributes the job type distribution of master graduate students.

6.5. *Decision making*

Considering the educational goal of this case study that helps to make decisions to improve the distribution of graduate students’ job type, the interpretations of mining result may provide several decisions to policymakers or leadership in the university.

1) For the inputs of student development system, the university can promote energy and initiative among schools to encourage students in major study and job seeking.

2) For the outcomes, faculty or staff can inspire student to focus on not only job type but also job place.

3) For the involvement environment of students, besides of academic achievements, such as academic score and prize, educators can pay more attention on their other involvement factors, for instance, the interaction faculty and student.

6.6. *Conclusion for the case study*

It is a simple case study to demonstrate the effectiveness of the framework in this paper. On the one hand, since it is not abundant with student development data, especially that lacks involvement data, the case study can be more helpful to decision making by obtaining more data. On the other hand, the case study also engages the attention of policymakers or leadership in the university to improve the information system according to student development theory.

7. Conclusion and future work

Different from other studies about EDM, this paper focuses on 'hypothesis formation' and 'refinement' in the iterative cycle of EDM, and presents a framework of decision making based on an improved model of student development theory. Initially, a case study of master graduate student testifies the effectiveness of the framework.

In the future, the framework may be deeply used to more complex data with more attributes for discovering more rules of decision making. And also, it may be widely used to other student objects in a number of higher educational institutions.

Acknowledgement

This work is supported by China Scholar Council (CSC) Fund, and File No. is 201406025111.

References

1. M. Bienkowski, M. Feng, and B. Means, Enhancing teaching and learning through educational data mining and learning analytics: An issue brief(US Department of Education, Office of Educational Technology, 2012).
2. M. Bala and D. Ojha, Study of applications of data mining techniques in education, International Journal of Research in Science and Technology, V**1**(2012).
3. C. Romero and S. Ventura, Data mining in education, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, V**3**, 12(2013).

4. J. Luan, Data mining and knowledge management in higher education-potential applications. In the Annual Forum for the Association for Institutional Research, (Toronto, Canada, 2002).
5. M. R. Beikzadeh, S. Phon-Amnuaisuk, N. Delavari, et al., Data mining application in higher learning institutions, Informatics in Education-An International Journal, V**7**, 31(2008).
6. M. M. A. Tair and A. M. El-Halees, Mining educational data to improve students' performance: a case study, International Journal of Information, V**2**, 2(2012).
7. C. Lei and K. F. Li, Academic performance predictors, in Advanced Information Networking and Applications Workshops (WAINA2015), (Gwangju, Korea, 2015).
8. M. Goyal and R. Vohra, Applications of data mining in higher education, International journal of computer science,V**9**, 113(2012).
9. EDM, Educational data mining, (international educational data mining society,2015) http://www.educationaldatamining.org/.
10. R. Mikut and M. Reischl, Data mining tools, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, V**1**, 431(2011).
11. J. A. Marsh, J. F. Pane, and L. S. Hamilton, Making sense of data driven decision making in education, (RAND Corporation 2006).
12. D. Naeimeh, S. Mohammad, and B. Mohammad, A new model for using data mining technology in higher educational systems, in Proceedings of the IEEE Conference on Information Technology Based Higher Education and Training(ITHET2004), (Istanbul, Turkey, 2004).
13. K. Schildkamp and W. Kuiper, Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors, Teaching and teacher education, V**26**, 482(2010).
14. A. W. Astin et al., Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.(Rowman& Littlefield Publishers, 2012).
15. A. W. Astin, Student involvement: A developmental theory for higher education. Journal of College Student Development, V**40,** 18(1999).
16. Weka 3: Data mining software in java(2016), http://www.cs.waikato.ac.nz/