Middle-East Journal of Scientific Research 23 (8): 1749-1755, 2015
ISSN 1990-9233
© IDOSI Publications, 2015
DOI: 10.5829/idosi.mejsr.2015.23.08.196

Proposing a New Framework Based on Hybrid Least Square Support Vector Machine and Fuzzy Logic for Student Handwrite Test Analysis

Tahmine Rasti

The General Directorate Education of Lordegan, Cheharmahale Bakhtiyari, Iran

Abstract: Understanding the factors that lead to success (or failure) of students at placement tests is an interesting and challenging problem. Since the centralized placement tests and future academic achievements are considered to be related concepts, analysis of the success factors behind placement tests may help understand and potentially improve academic achievement. Therefore in this paper we proposed a hybrid least square support vector machine (LS-SVM) and fuzzy logic for student handwrite test analysis. Here at first we create a vocabularies dataset and then make a quiz from students. Then extracting features and clustering using LS-SVM and scoring using fuzzy interface systems (FIS). Results show proposed method is capable to use in high school to correction of exams.

Key words: Least square support vector machine · Fuzzy interface systems · Data mining · Text mining · Correction of exams

INTRODUCTION

Data mining (DM) is a computer-based information system (CBIS) [1, 2] devoted to scan huge data repositories, generate information and discover knowledge. The meaning of the traditional mining term biases the DM grounds. But, instead of searching natural minerals, the target is knowledge. DM pursues to find out data patterns, organize information of hidden relationships, structure association rules, estimate unknown items' values to classify objects, compose clusters of homogenous objects and unveil many kinds of findings that are not easily produced by a classic CBIS. Thereby, DM outcomes represent a valuable support for decisions-making [3-6].

Concerning education, it is a novel DM application arena for knowledge discovery, decisions-making and recommendation [7]. Nowadays, the use of DM in the education arena is incipient and gives birth to the educational data mining (EDM) research field [8]. As we will see in Section 2, in a sense the first decade of the present century represents the kick-off of EDM.

EDM emerges as a paradigm oriented to design models, tasks, methods and algorithms for exploring data from educational settings. EDM pursues to find out patterns and make predictions that characterize learners' behaviors and achievements, domain knowledge content, assessments, educational functionalities and applications [9]. Source information is stored in repositories managed by conventional, open and distance educational modalities.

Some of the EDM trends are anticipated here. One of them corresponds to the standard integration of an EDM module to the typical architecture of the wide diversity of computer-based educational systems (CBES). Other tendency demands that EDM provides several functionalities during three stages of the teaching-learning cycle. The first stage corresponds to the provision of EDM proactive support for adapting the educational setting according to the student's profile prior to deliver a lecture. During the student-system interaction stage, it is desirable that EDM acquires log-data and interprets their meaning in order to suggest recommendations, which can be used by the CBES for personalizing services to users at real-time. In the next stage, EDM should carry out the evaluation of the provided education concerning: delivered services, achieved outcomes, degree of user's satisfaction and usefulness of the resources employed. What is more, several challenges (i.e., targets, environments, modalities,

Corresponding Author: Tahmine Rasti, The General Directorate Education of Lordegan, Cheharmahale Bakhtiyari, Iran.

functionalities, kinds of data,...) wait to be tackled or have been recently considered by EDM, such as: big data, cloud computing, social networks, web mining, text mining, virtual 3-D environments, spatial mining, semantic mining, collaborative learning, learning companions, ... [10].

Paper [11] analyses the online questions and chat messages automatically recorded by a live video streaming (LVS) system using data mining and text mining techniques. They apply data mining and text mining techniques to analyze two different datasets and then conducted an in-depth correlation analysis for two educational courses with the most online questions and chat messages respectively. The study found the discrepancies as well as similarities in the students' patterns and themes of participation between online questions (student–instructor interaction) and online chat messages (student–students interaction or peer interaction). The results also identify disciplinary differences in students' online participation. A correlation is found between the number of online questions students asked and students' final grades. The data suggests that a combination of using data mining and text mining techniques for a large amount of online learning data can yield considerable insights and reveal valuable patterns in students' learning behaviors [11].

In paper [12] using a large and feature rich dataset from Secondary Education Transition System in Turkey, developed models to predict secondary education placement test results and using sensitivity analysis on those prediction models identified the most important predictors. The results showed that C5 decision tree algorithm is the best predictor with 95% accuracy on hold-out sample, followed by support vector machines (with an accuracy of 91%) and artificial neural networks (with an accuracy of 89%). Logistic regression models came out to be the least accurate of the four with and overall accuracy of 82%. The sensitivity analysis revealed that previous test experience, whether a student has a scholarship, student's number of siblings, previous years' grade point average are among the most important predictors of the placement test scores [12].

This paper produces in 4 sections, in section 2, we describe proposed method in details and in section 3 introduced results and discussions and in last part we have conclusion.

Proposed Method

Dataset Creation: We create database in Lordegan's high schools, Cheharmahale Bakhtiyari province. For producing dataset, create a geography test from 300 students. Fig. 1 show an example of this exam. As shown in Fig. 1 we need two dataset, one of them is answer key and the other on is handwritten dictionary. Handwritten dictionary is useful for text mining. Then we create a word dictionary and ask students to handwrite each word. We create an answer key with score.

Framework of Algorithm: The framework of this algorithm was shown in Fig. 2. As was shown in Fig. 2, proposed algorithm has 4 steps.

Step 1: Creating answer key with score and handwrite dictionary,

Step 2: Exam from students and scanned them,

Step 3: Feature extracting and reduced feature space,

Step 4: Using LS-SVM and fuzzy systems for scoring; now we introduce these steps separately.

In step 1 and step 2, as mention later, for creating data set we need student, exam from students, answer key score and handwrite dictionary. This step explains in dataset creation completely.

Feature Extraction: After students handwrite dictionary and participating in exam, we need to extract feature. For feature extracting used Dinesh Dileep toolbox in Matlab, here we explain it.

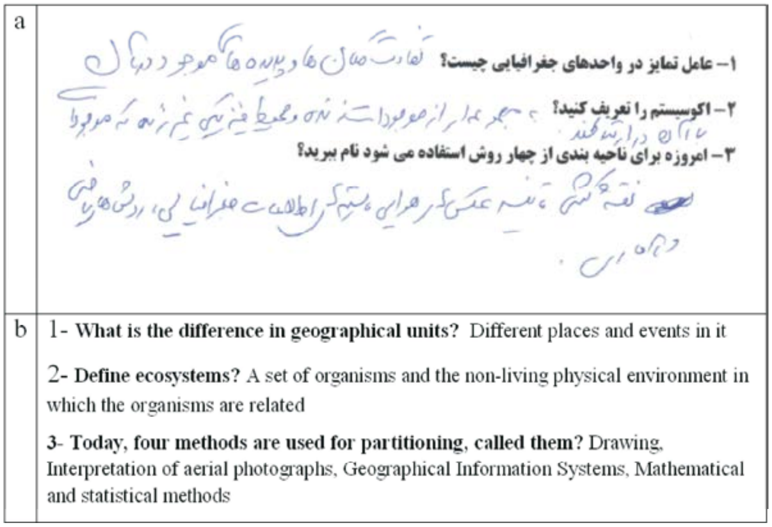
Image Preprocessing Involves the Following Steps:

- Character Extraction from Scanned Document.
- Binarization.
- Background Noise removal.
- Skeletonization.

The paper assumes that the input Image is available after undergoing all this processes. Some excellent papers on these steps are given in the references.

Universe of discourse is defined as the shortest matrix that fits the entire character skeleton. The Universe of discourse is selected because the features extracted from the character image include the positions of different line segments in the character image (Fig. 3).

After the universe of discourse is selected, the image is divided into windows of equal size and the feature is done on individual windows. For the system implemented, two types of zoning were used. The image was zoned into 9 equal sized windows. Feature extraction was applied to



a

۱– عامل تمایز در واحدهای جغرافیایی چیست؟ تفاوت مکان ها و پدیده های موجود در آن

۲– اکوسیستم را تعریف کنید؟ مجموعه ای از موجودات زنده و محیط فیزیکی غیر زنده که موجودات در آن ارتباط دارند

۳– امروزه برای ناحیه بندی از چهار روش استفاده می شود نام ببرید؟ نقشه کشی، تفسیر عکس های هوایی، سیستم اطلاعات جغرافیایی، روش های ریاضی و آماری

b

1- **What is the difference in geographical units?** Different places and events in it

2- **Define ecosystems?** A set of organisms and the non-living physical environment in which the organisms are related

3- **Today, four methods are used for partitioning, called them?** Drawing, Interpretation of aerial photographs, Geographical Information Systems, Mathematical and statistical methods

Fig. 1: Geography test, a- originale exam and b- translated

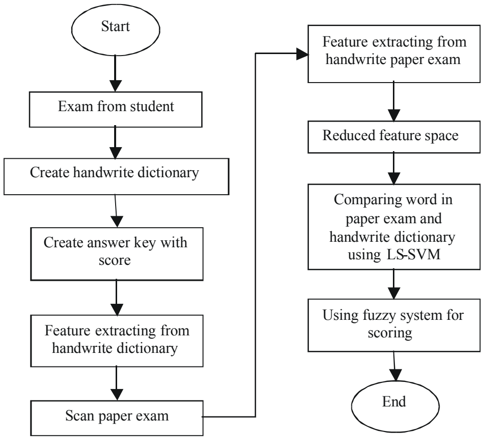

Start
Exam from student
Create handwrite dictionary
Create answer key with score
Feature extracting from handwrite dictionary
Scan paper exam
Feature extracting from handwrite paper exam
Reduced feature space
Comparing word in paper exam and handwrite dictionary using LS-SVM
Using fuzzy system for scoring
End

Fig. 2: Framework of proposed algorithm

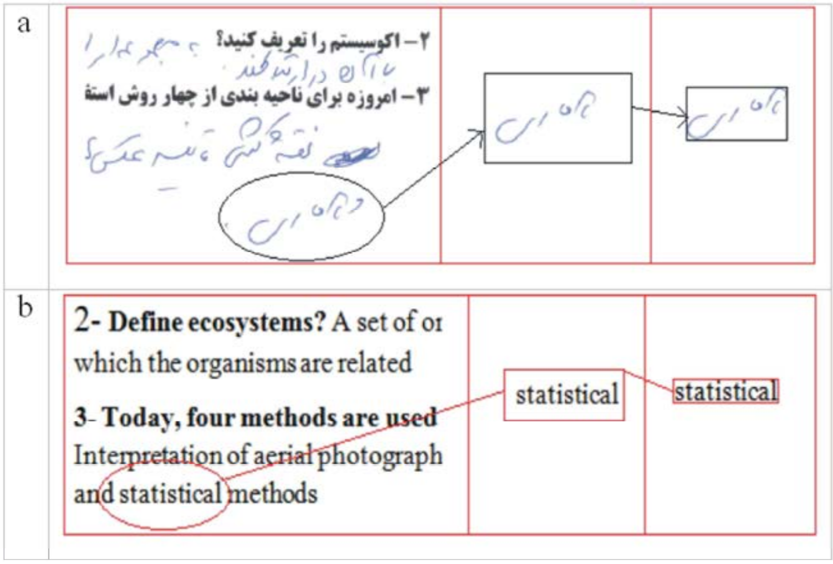

a
۲– اکوسیستم را تعریف کنید؟ مجموعه ای
ارتباط دارند
۳– امروزه برای ناحیه بندی از چهار روش استفا
نقشه کشی، تفسیر عکس
و آماری
و آماری
و آماری
b
2- Define ecosystems? A set of or
which the organisms are related
3- Today, four methods are used
Interpretation of aerial photograph
and statistical methods
statistical
statistical

Fig. 3: Word selecting in proposed method, a- a- originale exam and b- translated

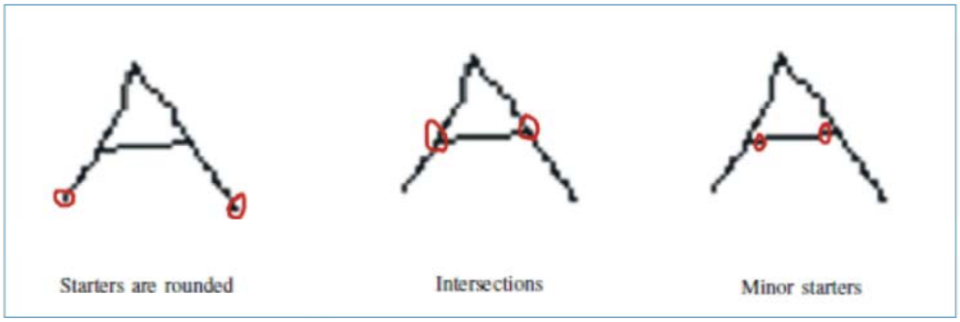


Fig. 4: Extract feature from a word or letter

individual zones rather than the whole image. This gives more information about fine details of character skeleton. Also positions of different line segments in a character skeleton become a feature if zoning is used. This is because, a particular line segment of a character occurs in a particular zone in almost cases. For instance, the horizontal line segment in character 'A' almost occurs in the central zone of the entire character zone.

To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton were defined as starters, intersections and minor starters.

Starters (Fig. 4): Starters are those pixels with one neighbor in the character skeleton. Before character traversal starts, all the starters in the particular zone is found and is populated in a list.

Intersections (Fig. 4): The definition for intersections is somewhat more complicated. The necessary but insufficient criterion for a pixel to be an intersection is that it should have more than one neighbor. A new property called true neighbors is defined for each pixel. Based on the number of true neighbors for a particular pixel, it is classified as an intersection or not. For this, neighboring pixels are classified into two categories, direct pixels and diagonal pixels. Direct pixels are all those pixels in the neighborhood of the pixel under consideration in the horizontal and vertical directions. Diagonal pixels are the remaining pixels in the neighborhood which are in a diagonal direction to the pixel under consideration. Now for finding number of true neighbors for the pixel under consideration, it has to be classified further based on the number of neighbors it have in the character skeleton. Pixels under consideration are classified as those with.

3 Neighbors: If any one of the direct pixels is adjacent to anyone of the diagonal pixels, then the pixel under consideration cannot be an intersection, else if none of the neighboring pixels are adjacent to each other than its an intersection.

4 Neighbors: If each and every direct pixel have an adjacent diagonal pixel or vice-versa, then the pixel under consideration cannot be considered as an intersection.

5 or Neighbors: If the pixel under consideration have five or more neighbors, then it is always considered as an intersection Once all the intersections are identified in the image, then they are populated in a list.

Minor Starters (Fig. 4): Minor starters are found along the course of traversal along the character skeleton. They are created when pixel under consideration have more than two neighbors. There are two conditions that can occur

Intersections: When the current pixel is an intersection. The current line segment will end there and all the unvisited neighbors are populated in the minor starters list.

Non-Intersections: Situations can occur where the pixel under consideration has more than two neighbors but still it's not an intersection. In such cases, the current direction of traversal is found by using the position of the previous pixel. If any of the unvisited pixels in the neighborhood is in this direction, then it is considered as the next pixel and all other pixels are populated in the minor starters list. If none of the pixels is not in the current direction of traversal, then the current segment is ended there and all the pixels in the neighborhood are populated in the minor starters list.

When the algorithm proposed is applied to character 'A', in most cases, the minor starters found are given in Fig. 4.

Individual Properties for Dimension Reduction: One of techniques are available for the selection of suitable properties, using the mean and variance of data. If the data is normally distributed, this technique is clearly accountable. When using these techniques. Data are labeled class. Comparison of the mean values of properties that have been normalized by the variance, the statistical techniques are used.

The following function can be used to select appropriate data when the two classes A and B are used. Threshold value, TherVal, is determined by the user.

$$\frac{|Mean(\mathrm{A}) - \mathrm{Mean(B)}|}{\sqrt{\frac{Var(\mathrm{A})}{N1} + \frac{Var(\mathrm{B})}{N2}}} > TherVal \tag{1}$$

where Var and Mean functions are, respectively, the mean and variance are calculated for the respective classes. The number of samples in each class N1 and N2 are shown. This value is calculated for each of the feature. The user can then set a threshold to identify suitable properties.

LS-SVM: For solving quadratic programming model support vector machine is used. When we calculate the dimensions of the problem, you may encounter a computationally expensive. In this regard, the method of least squares support vector machine was designed for the optimization problem as follows:

$$Min\ 1/2\ \|w\|^2 + \frac{C}{2}\sum_i \varepsilon_i^2 \tag{2}$$

$$S.t \quad y_i(<\mathrm{w.x_i}> + \mathrm{b}) \geq 1 - \varepsilon_\mathrm{i}, \varepsilon_\mathrm{i} \geq 0 \quad \forall i \tag{3}$$

According to this method, the quadratic programming problems can be easily solved by a set of linear equations. Based on the Lagrange equation is as follows:

$$L(\mathrm{w,b},\varepsilon_i,\alpha_i) = \frac{1}{2}w^T w + \frac{C}{2}\sum \varepsilon_i^2 - \sum \alpha_i[\mathrm{y_i}(\mathrm{w^T}\,\varphi(\mathrm{x_i}) + \mathrm{b}) - 1 + \varepsilon_i] \tag{4}$$

where α_i is ith Lagrange multipliers.

To simplify the equation above, we consider the following assumptions:

$$Y = (\mathrm{y_1, y_2, \ldots, y}_N)^T, \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T, \quad 1 = (1,1,\ldots,1)^T,$$
$$\Omega_{ij} = y_i y_j^T \varphi(\mathrm{x_j}) + ((1)/\mathrm{C})\mathrm{I}$$
$$= y_i y_j^T K(\mathrm{x_i, x_j}) + (1/\mathrm{C})\mathrm{I}, \mathrm{i} = 1,2,\ldots,\mathrm{N} \tag{5}$$

In the above equation, $K(\mathrm{x_i, x_j})$ kernel function and I is the identity matrix.

Ccording to the above equation, we get:

$$\begin{bmatrix} \Omega & Y \\ Y^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{6}$$

Parameter Ω is positive with respect to the previous equation, Lagrange multipliers can be obtained using the above equation, the formula is as follows:

$$\alpha = \Omega^{-1}(1 - \mathrm{bY}) \tag{7}$$

Placement of above equation with the previous matrix the following equation is obtained:

$$b = \frac{Y^T \Omega^{-1} 1}{Y^T \Omega^{-1} Y} \tag{8}$$

Due to the fact that Ω is positive definite, Ω^{-1} is also positive given. In addition, since Y is non-zero vector, ie $Y^T\ \Omega^{-1} Y$ will be greater than zero. So b is obtained. b and w using the least squares support vector machine to classify the input space as follows:

$$F(x) = sign[f(x)] = sign[\sum \alpha_i y_i K(\mathrm{x, x_i}) + \mathrm{b}] \tag{9}$$

Fuzzy Interface System: For scoring, here, we proposed a fuzzy interface system. A FIS has three parts: part one is inputs, part two is processing unit and last part is output. Here each question equipped with a FIS, therefore, output is a score for a question start from 0 to high score.

RESULTS AND DISCUSSIONS

Simulation and extracting results was happen using MATLAB 2014 and for analyzing results using SPSS 22. For detect the accuracy of proposed method, compare our method with traditional methods in T-Test (99% significantly level).

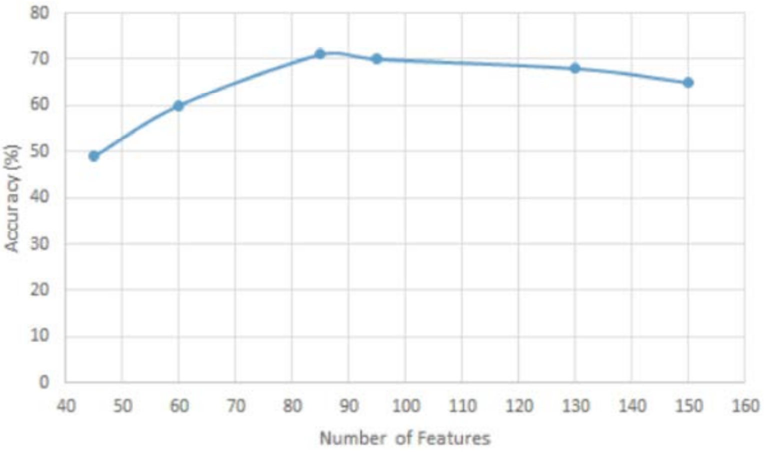


Fig. 5: Effect of feature space on accuracy

Table 1: Means of corrections and proposed method

Correction	Mean
Correction 1	14.66
Correction 2	14.67
Correction 3	15.31
Correction 4	15.21
Correction 5	15.03
Correction 6	13.91
Correction 7	15.70
Correction 8	13.76
Correction 9	14.55
Correction 10	15.75
Proposed method	13.21

Table 2: One-Sample T-Test at 99% significantly levels

	Test Value = 14.210					
					99% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Scores	3.009	9	.015	.64500	-.0517	1.3417

At first used LS-SVM for detecting the word, then calculating recognition rate. Figure 5 was shown, recognition rate in different feature space. At first we have 150 features, with decreasing to 84 features recognition rate was improved. Because some noise effect on features and then affected on recognition rate. Feature space with 84 members is the best accuracy and its use for fuzzy systems and scouring.

As a mention later, we use feature space with 84 members in LS-SVM and FIS for Scouring and compare results with 10 corrections (teachers) with T-Test. Table 1 shows means of all students according to corrections and in proposed method. We use one sample T-Test and compare mean of ten corrections with mean of proposed method.

Table 2 proved mean of score in proposed method has not significant differences with means of corrections. Therefore we can use this system with acceptable accuracy and high speed to correct the exams and record the results.

CONCLUSION

Correction of exams in high school and record on computer is a challenge in high schools. Researchers try to produces methods till improve speed accuracy of records of result and correction of exams. Therefore in this paper proposed a hybrid least square support vector machine (LS-SVM) and fuzzy logic for student handwrite test analysis. At first we create a vocabularies dataset and make a quiz from students. Then extracting features and clustering using LS-SVM and scoring using fuzzy interface systems (FIS). Results show proposed method has no significant different with traditional methods, but improves corrections treatment.

REFERENCES

1. Vlahos, G.E., T.W. Ferratt and G. Knoepfle, 2004. The use of computer-based information systems by German managers to support decision making. Journal of Information and Management, 41(6): 763-779.
2. Akcapinar, G., E. Cosgun and A. Altun, 2011. Prediction of perceived disorientation in online learning environment with random forest regression. In Proceedings of the 4th International Conference on Educational Data Mining, pp: 259-263.
3. Barracosa, J. and C. Antunes, 2011. Mining teaching behaviors from pedagogical surveys. In Proceedings of the 4th International Conference on Educational Data Mining, pp: 329-330.
4. Yuseni Ab Wahab, Abd Samad Hasan Basari and Burairah Hussin, 2014. Replacement Model for Hostel Building Case Study: ICYM, Middle-East Journal of Scientific Research, 21(11): 1977-1981.
5. Naveed Hasan, Shahab Alam Malik and Muhammad Majid Khan, 2013. Measuring Relationship Between Student’s Satisfaction and Motivation in Secondary Schools of Pakistan, Middle-East Journal of Scientific Research, 18(7): 907-915.
6. Taisir Mohammed Hameed, Zainuddin Bin H.J. Hassan and Rosnafisah Sulaiman, 2015. Is Social Network an Effective E-learning Tool: A Survey, Middle-East Journal of Scientific Research, 23(1): 119-126.
7. Vialardi-Sacin, C., J. Bravo-Agapito, L. Shafti and A. Ortigosa, 2009. Recommendation in higher education using data mining techniques. In Proceedings of the 2nd International Conference on Educational Data Mining, 190-199.
8. Anjewierden, A., B. Kolloffel and C. Hulshof, 2007. Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes. In Proceedings of the international workshop on applying data mining in e-Learning, pp: 23-32.
9. Luan, J., 2002. Data mining and its applications in higher education. Journal of New Directions for Institutional Research, 113: 17-36.
10. Alejandro Pena-Ayala, 2013. Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems with Applications, ESWA 8846, pp: 1-31.
11. Wu He, 2013. Examining students’ online interaction in a live video streaming environment using data mining and text mining, Computers in Human Behavior, 29: 90-102.
12. Baha Sen, Emine Ucar and Dursun Delen, 2012. Predicting and analyzing secondary education placement-test scores: A data mining approach, Expert Systems with Applications, 39: 9468-9476.