

Supporting Information for paper

Economic deterministic ensemble classifiers with probabilistic output using for robust quantification: study of unbalanced educational datasets

S. M. Abdullaev

Department of System Programming
South Ural State University
Chelyabinsk, Russian Federation
abdullaevsm@susu.ru

Y.K. Salal

Department of System Programming
South Ural State University
Chelyabinsk, Russian Federation
Yasskhudheirsalal@gmail.com

Analysis of the results of training

Let us now compare the students' examination scores with the results of an independent test containing 30 questions (15 in computer science and 15 in mathematics) based on the material covered in the semester. (In Tab. II) shows three gradations of grades of girls, boys and the entire population of schoolchildren. The analysis identified the following three statistically significant conclusions.

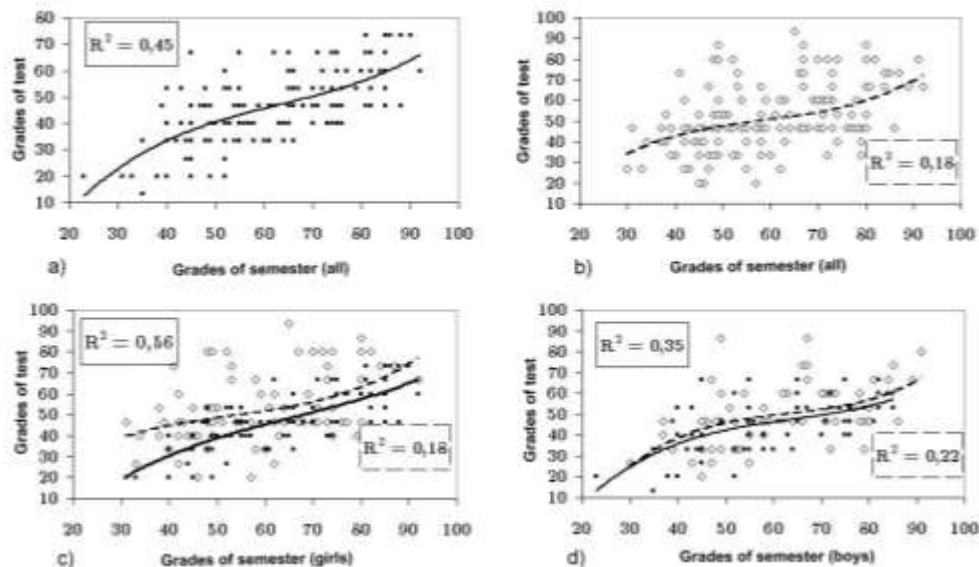


Fig. 1. The relationship between the results of the test and the semester exam: in mathematics (a) and in computer science (b) for all students; separately for girls (c) and boys (d). Continuous and discontinuous lines are polynomials of the third degree, approximating the connections between estimates in mathematics and computer science.

- **The first observation:** obvious independent testing lowers grades. So, the average score in the math (computer science) exam was 62.5 ± 15 (60.2 ± 14), and in the test, it was 46.6 ± 13 (51.1 ± 15). Using a non-parametric chi-square criterion (two degrees of freedom) to compare two distributions with three gradations, we find that the minimum value = 17.9 is observed in the case of computer science in a female gymnasium. In other cases, it is much larger. This means that with a confidence level of 99.9%, it can be assumed that the distribution of grades in testing differs from the distribution of grades of previous exams. The most critical is the increase in the number of students failing the test in both subjects by 1.8-2.5 times.
- **The second observation:** Testing level both subject differences and the difference in grades between schools, showing almost the same percentage of $\cong 60\%$ of the number of students who fail the test.
- **Third observation:** the distribution of scores for a Girl's school in an exam in computer and mathematics differs from each other with a probability of 97%, but the corresponding distribution of marks for boys is practically the same. However, there are no correlations between the student's success in computer science and his success in mathematics in either case. This was most strikingly reflected in the fact that 18% of students believed that computer sciences were not related to mathematics, and only 30% of respondents were confident in the firm connection of these sciences (question 15, Table I).
- **Fourth observation**
supplementing the previous one, it concerns the individual relationships between the results of the student's knowledge assessment produced at school and the manifestation of this knowledge on an "independent test" (Fig. 1). As you can see from the graph, a connection of slightly more average tightness is found between current and past knowledge of mathematics (Fig. 1a). Here, the R coefficient of linear Pearson correlation, $R = 0.67 \pm 0.07$. Moreover, the degree of connection in the results of girls in mathematics ($R = 0.74 \pm 0.10$, Fig. 1c) is significantly higher than the average observed in boys ($R = 0.59 \pm 0.12$, Fig. 1d). The connection between the marks in the school exam and the computer science test reveals a tightness of the links below the average $R = 0.42 \pm 0.07$ (Fig. 1b). A slightly higher tightness of connections in informatics among boys is due to the fact that students of the Zeytun school had a computer class and showed a degree of connection $R = 0.52 \pm 0.16$, which is noticeably higher than that of boys from Jawahiriya school, where as in the girl's school there is no computer class.

Experiment evaluated the effectiveness

The experiment evaluated the effectiveness of classifiers using five algorithms: a decision tree algorithm (J48), a two-layer perceptron with feedback (MLP), Naive Bayes classifier NB, the k-nearest-neighbor method (k-NN, for $k = 1$), and the support vector method (SVM). For the training of classifiers, 13 to 20 predictors of 16 attributes of the questionnaire were used (Table II) and 4 grades: 2 grades for examinations in mathematics and computer, and 2 intermediate grades for theory and practice in computer in the first semester.

Two or three classes of students' grades on tests in computer and mathematics were predictants. Another task, including preliminary theoretical analysis, was experiments on the selection of the best classification algorithm, ensemble prediction, and quantification.

Describe the training and test samples.

The total sample of 160 students was divided into training and test samples in the ratio of 112 to 48 students (i.e. 70 to 30%) and in the proportion of 80 to 80 (50 to 50%). In the first case, the forecast was simulated when training classifiers on the data for the previous 2 years, in the second more typical case, when there is data for the past year, it is necessary to classify the same number of newly enrolled students.

Balancing of sample

Obviously, with a real students' success threshold of at least 8 correct answers, 59% and 62% of students should be assigned to the class [-1], meaning "low knowledge" in computer and mathematics subjects. In this case, the classes of binary classification would be moderately balanced, but then there would be a strong imbalance in the gradations of the ternary classification (Table II).

On the other hand, in the routine process of school education, it is impossible to imagine a situation where the number of students completed the course with negative grades will be greater than the number of their fellow students with positive grades. Therefore, classes of binary classification were formed by a threshold value of 7 questions. (There were 15 questions in the test), the 7 correct answers or more "passed" (class [+1]), and "not passed" (class [-1]) were 6 or less correct answers. With such a breakdown on a test in a computer (mathematics), 25% (38%) of high school students would have received a "not passed" rating, which is closer to the real situation. . Further, when selecting, as a condition of average gradation of knowledge required 7 correct answers was achieved relative balance in the provision of classes "low", "medium" and "high" level of knowledge in computer (mathematics) in the proportion 25/34/41 (38/25/37).

Experiment Objectives

After preliminary studies of Weka's capabilities and the available database, the following 6 experimental tasks were set: to investigate the dependence of the forecast quality on the choice: 1) the type of predictors and predictants; 2) the volume of the training and test sample; 3) the number of output classes, 4) the classifier algorithm, as well as explore 5) the prospects of ensemble prediction and 6) quantification

All tasks were repeated at least three times when in each of the 3 series a new learning and test sample was randomly assigned. In tab. 4, we demonstrate the summary results of one series consisting of 32 experiments, in which under the same conditions the forecast accuracy A and F-measure of five individual classifiers were compared, dependence of the quality of prediction on the choice of predictors and predictants. The dependence of the performance of binary classifiers, trained (tested) on a sample of 112 (48) students on the number of predictors and the type of predictants are shown in lines 1-8, Tab.4.

Evaluation of the quality of classification

Among the many assessments of the quality of the classifiers produced by Weka, to demonstrate the various effects, we selected the two simplest characteristics: accuracy (A) characterizing the overall accuracy, and the minimum value of the F1 measure (F) characterizing the predictability of the individual class (Table. III).

The forecast was considered useful if the F value was greater than 50%. The sum of useful predictions for various classifiers is denoted as Σ .

Table I. The results of a survey of students (January, 2018).

	Demographic data		Socio-economic data
1	Gils's School El-Diwaneyah (88) 830; Male schools: Jawahiriya (35) 450, Zeytun (37) 660 *	8	Father's profession: worker (56), teacher (21), law enforcement agencies (21), technician (14), physician (10), other (38)
2	Scientific (128) 1377, humanitarian (32) 563 *	9	Father's income: not working (21), low (41), medium (24), high (74)
		10	Mother income: housewife (83), teacher (27), medic (26), technician (14), other (10)
3	age: 14-15 (37), 16-17 (95), 18 and older (28)	11	Father's education: none (1), primary (57), secondary (84), higher (18)
4	The family consists of 2-4 (17), 5-8 (80), > 9 people (63)	12	Mother's education: none (22), primary (56), secondary (75), higher (7)
5	I am in the family: 1-3rd (84), 4-6th (61) and later child (15)	13	Brother / sister education.: none (15), primary (56), secondary (81), higher (8)
6	Parents married (133), divorced (17), widowed (10)	14	Home Internet: yes (134), no (26); PC: yes (106), no (54)
7	I live with my parents (148); with relatives and separately (12)	15	Q/Computer science related to math? Yes (48), Rather, Yes (83), Rather No (26), No (3)

* Second number the total number of students

Table II. Exam results and testing students

Contingent	Test	Computer science, points			Mathematics, points		
		<50	50 – 65	>65	<50	50 – 65	>65
Gir's School	Exam, %	38	28	34	22	25	53
	Test, %	58	18	24	63	23	15
Boy's school	Exam, %	25	31	44	28	26	46
	Test, %	61	24	15	61	29	10
All Schools	Exam, %	32	29	39	24	26	50
	Test, %	59	21	20	62	25	13

Table. III The prediction's quality of classifiers of algorithms, predictors and predictants

№	Sample training/forecast, subject,	The number of predictors in the	J48		MLP		NB		SVM		1NN		Σ
			A	F	A	F	A	F	A	F	A	F	

	output graduation	experience												
1	70/30, Computer, 12\36	20, All dataset	81	67	77	52	90	80	83	60	77	52	5	
2	70/30, Mathematics, 24\24	20, All dataset	71	67	75	73	67	64	69	65	56	51	5	
3	70/30, Computer, 12\36	11, without SED	73	0	73	44	77	56	75	?	71	46	1	
4	70/30, Mathematics 24\24	11, without SED	69	65	60	51	69	67	65	58	60	51	5	
5	70/30, Computer, 12\36	13, without DD	79	58	81	69	88	83	85	67	69	40	4	
6	70/30, Mathematics, 24\24	13, without PP	65	56	73	70	67	64	65	58	50	33	4	
7	70/30, Computer, 12\36	16 without PP	79	64	73	43	75	33	73	23	75	50	1	
8	70/30, Mathematics, 24\24	16, without PP	50	43	65	56	46	23	50	8	56	43	1	
9	50/50, Computer, 21\59	$\Sigma(20,11,13,16)$	3		0		3		0		0		6	
10	50/50, Mathematics, 33\47	$\Sigma(20,11,13,16)$	2		3		3		3		0		11	
11	70/30, Computer, 12\16\20	$\Sigma(20,11,13,16)$	1		1		0		0		1		3	
12	70/30, Mathematics, 24\11\13	$\Sigma(20,11,13,16)$	1		1		0		0		0		2	
13	50/50, Computer, 21\29\30	$\Sigma(20,11,13,16)$	0		0		0		0		0		0	
14	50/50, Mathematics, 33\22\25	$\Sigma(20,11,13,16)$	0		0		0		0		0		0	
15	Useful predictions, Σ	Lines 1-14	13		11		12		8		4		48	