Supporting Information for paper

An economic deterministic ensemble classifiers with probabilistic output using to robust quantification: study of unbalanced educational datasets

Y.K. Salal
Department of System Programming
South Ural State University
Chelyabinsk, Russian Federation
Yasskhudheirsalal@susu.ru

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 ≅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 ± 0.07. Moreover, the degree of connection in the results of girls in mathematics (R = 0.74 ± 0.10, Fig. 1c) is significantly higher than the average observed in boys (R = 0.59 ± 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 ± 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 ± 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.

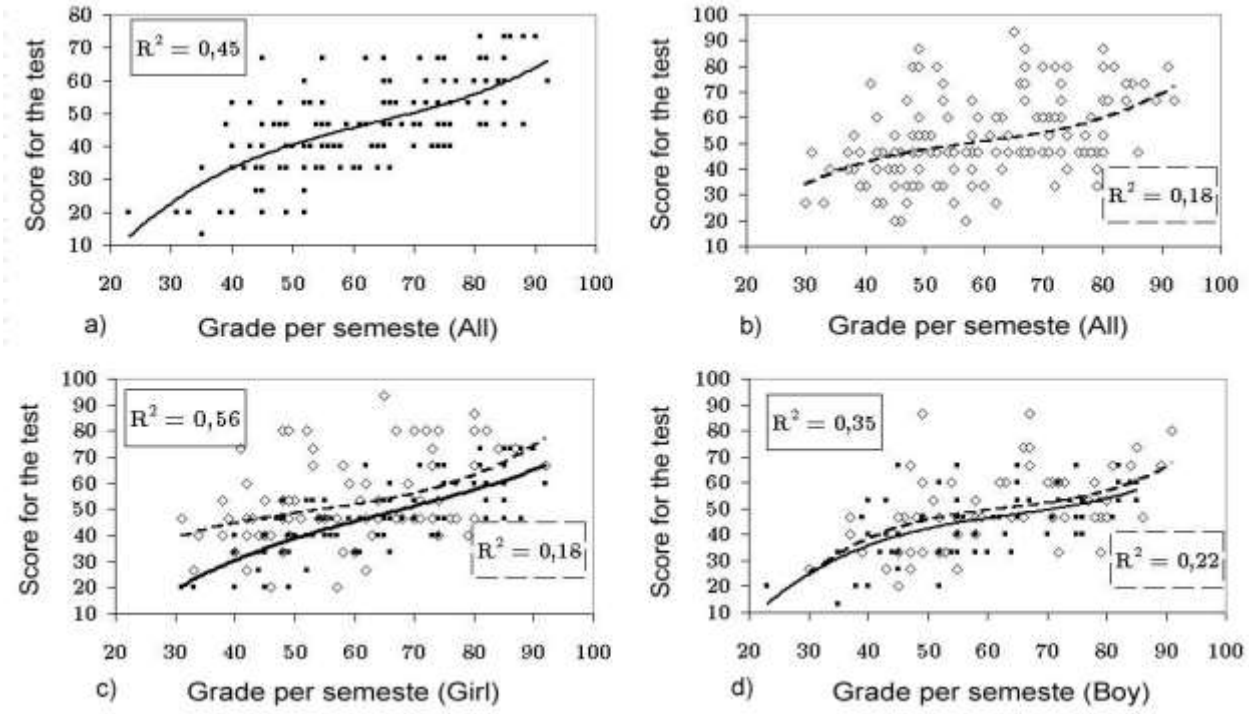

Fig. 1. The relationship between the results of the test and the semester exam: in mathematics (a) and in computer science (b) for all students; separately for girls (c) and boys (d). Continuous and discontinuous lines are polynomials of the third degree, approximating the connections between estimates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

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 science and mathematics. 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 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.

All dataset

The dataset of 20 attributes (lines 1 and 2), all the classifiers trained and tested on sample 70/30 gave useful predictions of the results of training a computer and mathematics. The overall accuracy of forecasts "A" depended on the choice of the predictor: when forecasting the estimates of the test in computer, it was in the range of 77-90%, which is about 20% higher than in mathematics.

Socio-economic data (SED): when 9 socio-economic attributes are removed from the set of predictors (lines 3.4), a slight decrease of up to 10% in the overall accuracy A and a sharp deterioration in the predictability of a particular class in a computer are noticeable, $\sum = 1$.

Moreover, in the case indicated by "?", The SVM algorithm did not predict low student performance at all. The decrease in the predictability of the total in computer test was combined with the fact that the number of useful predictions $\sum$ in mathematics remained at the same level, $\sum = 5$.

On the other hand, the gradations of the binary classification of results in the computer were not balanced: the ratio of classes 1/3. In a series of experiments, it was found that the lack of

predictability of an individual class with an imbalance in the selection of data is a frequent occurrence. Recall that at this conclusion.

Demographic Data (DD)

The assertion that the nature of predictors affects the quality of predicting a smaller class to a greater extent, becomes obvious when removed 7 predictors characterizing demographic data from the full dataset of attributes (DD, lines 5,6).

In this case, 4 classifiers coped with the forecast in computer science and mathematics ($\sum$ = 4). The classifier with 1NN did not give a useful forecast. A slight decrease in the quality of predictions in mathematics in the absence of DD compared to previous experiments is explained by the remarks made in Section 1. It noted the importance of a number of DDs (the age and sex of the student, the direction of study, and others) for clarifying the relationship between the grades of the math test and the student's previous successes.

Previous performance (PP)

As expected, a dramatic deterioration in the forecast is observed when attributes characterizing the student's previous academic performance were removed from the general set (PP, lines 7.8).

Comparing the effectiveness of the forecast in mathematics and computer with previous experiments where $\sum \geq 4$, it is easy to notice one more known effect of unbalanced samples, the general accuracy of the forecast A and other weighted average quality estimates within certain limits do not feel a drop in the predictability of the less well-off class of objects. Indeed, A practically did not fall below 75% (50%) when predicting the results of training computer (mathematics).

Sample size

The observations described above are confirmed by the results on sample 50/50 (lines 9-10). So the number of useful predictions of binary classifiers on a balanced sample in mathematics is twice as many as on an unbalanced sample in computer science. However, the correlation of test results and grades in the math exam obviously helps. In general, we see that in the experiments on the 50 to 50 sample, the sum of the useful forecasts is $\sum$ = 17, which is 1.5 times less than in the experiments with the 70/30 sample ($\sum$ = 26).

Ternary classification

An increase in the number of classes (predictors) is accompanied by a sharp drop in predictability. So the sum of useful predictions of ternary classifiers (lines 11-14) is three to five times less than binary classifiers. Moreover, all useful predictions were obtained with just two classifiers J48 and MLP in experiments with a full set of predictors in the 70/30 sample. The ternary classifiers in 4 experiments on the 50/50 sample (lines 13-14), despite the close sizes of the output classes, did not give a useful forecast.

Our subsequent estimates will be based only on experiments in which at least one binary classifier gave useful predictions.

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.

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. 3 shows three gradations of scores of girls, boys and the entire population of schoolchildren. The analysis identified the following three statistically significant conclusions.