ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 105

1st International Scientific and Practical Conference on Digital Economy (ISCDE 2019)

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

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

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

***Abstract* — The overall goal of our work is to find economic and robust supervised machine learning methods which adequate to both individual and collective Student Performance Forecast (SPF). The individual SPF are subject of well-known classification methods but collective SPF is subject of quantification learning algorithms dealing with the novel task to predict the frequency of classes in tested sample e.g. a number of students with unsatisfactory grade. The need for revise of classification methods shows review of 86 SPF in developing countries. The analysis depicts that most of SPF report the high overall accuracy of classifiers based on decision tree J48, Naïve Base NB, Multilayer Perception MLP, k-Nearest Neighbor k-NN, and Support Vector Machine SVM algorithms, but did not take into account the accuracy of the forecast of a minor presented class. So, given the imbalance in the sample, "useful forecast" with the F1 metric above 50% (75%) are given only in ½ (1/5) of cases of forecasts. The pivotal study of the efficacy factors of binary SPF (data type, algorithm, sample balancing, number of classes etc.). Another important finding is that classifiers with the probabilistic Naïve Bayesian kernel, have more stable behavior to classify different EDM datasets, overcoming MLP, J48, SVM and k-NN based classifiers which sometimes achieved good forecast but sometimes failed in prediction. After that, collected all the above experimental finds associated with relationship between algorithm and data information, we construct 3-15 member heterogeneous ensembles contained strong, moderate and weak classifiers for deterministic individual SPF by simple voting and heuristically proposed how individual probabilistic predictions could be generated and how to aggregate them for overall frequency forecasting, i.e. resolve the task of quantification. The proposed methods of ensemble forecasting and ensemble quantification can become the basis for new economic and robust solutions of various real-world problems in the field of machine learning.**

Keywords* — *Educational Data Mining, monitoring, individual and collective learning outcomes, Weka, classification, assessment, deterministic and probabilistic forecasts, ensemble classification and quantification

I. INTRODUCTION

This work is part of Educational Data Mining (EDM) branch of data mining, which deals with the digital methods of processing and analysis of information circulating in the education system [1]. In applied matters, including quality assessment, monitoring, and prediction of the results of various forms of traditional and constructivist learning, modeling of adaptive hypermedia systems and other tasks, EDM intersects with Learning Analytics (LA), which considers these issues from a broader, integrated perspective of learning optimization and management [2-5]. In our study, we explore the possibilities of EDM predicting the educational result of individual students, such as grades of student A on the future math exam and the LA task to evaluate the general successes of a large group of students, such as the number of students to get a scholarship in next year. For the brevity of these two types of the forecast, we will call as of individual and collective Student Performance Forecast (SPF). The overall goal of our work is to find economic and robust supervised machine learning methods which adequate for both types of SPF.

II. PROPOSED METHODOLOGY

- Base classification of SPF

The task of individual SPF is usually solved by supervised classification requires previously classified reference samples of student data in order to train the classifier and subsequently classify unknown data. At the moment, the many algorithms of classification are available in open software [6], so the focus of major part of research associated with SPF shifts towards the evaluation of SPF algorithms and the importance of student attributes [7,8]. It should be noted that the most popular way of SPF evaluation by calculating the overall accuracy of the classifier is not an adequate indicator in real-world cases. Thus, we revise of 86 SPF cases reported during 2010-2018 in South Asia. Firstly, the analysis demonstrates that there are five popular SPF algorithms: J48, NB, multi-MLP, k-NN and SVM. These algorithms are applied to four types of attributes: final grades and internal assessments (all cases), data about, forming a complete set with additional demographic or socio-economic attributes. In general, this is a well-known fact reflected in the reviews [7,8], which were also based on overall accuracy allegations about of high quality of the classifier. However, this is far from the case. For example, we found that two-thirds of the samples used were strongly class unbalanced and reflect the nature of educational data were that parity of classes is impossible: the number of excellent student and underachievers will always be markedly lower than students of middle achievement (sometimes the training sample can be artificially balanced, for example by using "oversampling", but the tested sample used to SPF must be naturally unbalanced). In this case, instead of accuracy, metrics reflecting the accuracy of a particular class (Precision)

and its completeness (Recall) or their average geometric value of F-measure are preferred [9]. When we found Precision, Recall, F1-measure or evaluated these metrics of the quality of the forecast of the individual class, the result of the analysis was even more depressing. For example, Fig. 1a presents the diapasons of overall accuracy (A) of classification and minimum values of precision, Recall, and F1-measure of the individual classes. Moreover, in 80% of cases, the overall A>50% deceptively pointed out the presence to predictive of the classifier, but metrics of individual classes demonstrated that the "useful forecast" e.g. with F>50% will be feasible only in half of the SPF cases. Moreover, if we take into account only the values of > 75% of both overall and individual class accuracy, we can say that the satisfactory SPF is achieved only in 20% of cases.

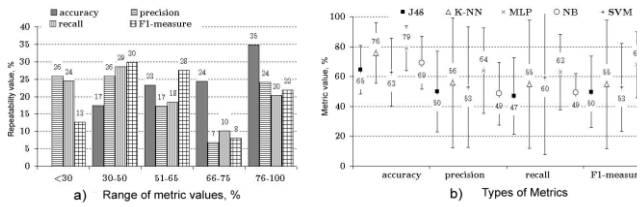

Fig.1. Assessment of SPF quality: a) values of different metrics; b) Average values of metrics of SPF classifiers

- Comparative effectiveness of classifiers

This task is often seen as a major challenge in many studies. There are also many misconceptions. It is easy to see from Fig. 1b that the overall accuracy of the MLP classifier forecast reaches 80% and the performance of the individual class forecast is up to 65%. Isn't that a good result? However, these results obtained with a significantly smaller number of SPF using MLP 8% compared to J48 22%, and NB 19%.

Moreover, the task of selecting the most effective classifier [10] is a more complex procedure than a simple comparison of the accuracy of the classifiers and it will be considered later in section III.

The collective SPF is based on a relatively new field of machine learning called quantification learning. Unlike the individual forecast produced by the classifier, a quantifier trained to predicts the frequency of certain class objects appearing in a tested sample [11]. More strictly, the quantifier q presents a mathematical model capable of predicting the prevalence of each class in the general population. The most developed task is to create a binary quantifier [11]. For example, you can do this with a binary classifier, followed by counting the frequencies of predictive classes by Classify &Count (C&C) method [12]. It is noted that the new methods of quantification are more effective than C&C [11]. However, comparing their effectiveness is noticeably more complex than classifiers, as quantifiers estimate unknown frequencies of objects and classifiers assume that the objects of the training and test sample are equally distributed.

- Individual and collective learning outcomes

In our task on the of individual and collective SPF, we initially assumed that in finding an effective algorithm, the optimal set of attributes (predictors) of the object and the detail of the forecast of the classification task and quantifications are the same, especially for the processing of big data [9-11] and the application of ensemble approaches [13]. In this work, we have developed an ensemble approach to classification with the training of various binary classifiers and allowing us to implement both acceptable deterministic and probabilistic individual forecasting, including suitable for use of unbalanced samples. Moreover, we have proposed a way to aggregate individual probabilities on the basis of which a new method of quantification has been obtained, which has a simple application as the C&C method but has much better efficiency and stability.

In the remaining part of the work, we investigate SPF quality on a set of predictors and predictantres in section III, and carried out a procedure to find the best classification algorithm. Then section IV substantiated the method of deterministic and probabilistic prediction of individual outcomes based on constructed heterogeneous ensemble classifiers with a variation of the data type. Next in section V, we have developed and tested a new algorithm of binary ensemble quantification, which allows improving the forecast of collective learning success. The conclusion concludes that further research in this area is needed.

- Description of experimental forecasting

Our Dataset that used of total sample of 160 students of 20 attributes, was divided into training and test samples in the ratio of 112 to 48 students (i.e. 70 to 30%) and the proportion of 80 to 80 (50 to 50), used free software for data analysis and machine learning Weka 3.6.9 (Waikato environment for knowledge analysis). For more experiments and details about a dataset are available online for interested readers at [17].

1) All dataset: The dataset of 20 attributes (Table I, lines 1-2), the entire classifiers trained and tested on sample 70/30 gave useful predictions of the results of training a computer and mathematics subjects. 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.

2) Socio-economic data (SED): when 9 Socio-economic attributes are removed from the set of predictors (Table I, 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 "?", 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.

3) 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 (Table I, lines 5,6). In this case, 4 classifiers coped

with the forecast in computer and mathematics ∑=4. The classifier with 1NN did not give a useful forecast.

4) Previous performance (PP): As expected, a cardinal deterioration in the forecast is observed when attributes characterizing the student's previous academic performance were removed from the all dataset (Table I, lines 7-8).

Comparing the effectiveness of the forecast in mathematics and computer with previous experiments where ∑≥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 didn't fall 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).

5) Sample size: The observations described above are confirmed by the results on sample 50/50 (Table I, 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 ∑=17, which is 1.5 times less than in the experiments with the 70/30 sample ∑=26.

6) 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 (Table I, 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 all dataset of predictors in the 70/30 sample. The ternary classifiers in 4 experiments on the 50/50 sample (Table I, 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.

TABLE I. THE PREDICTION'S QUALITY OF CLASSIFIERS OF ALGORITHMS, PREDICTORS AND PREDICTANTS

№	Sample training/forecast, subject, output graduation	The number of predictors in the experience	J48		MLP		NB		SVM		1NN		Σ
			A	F	A	F	A	F	A	F	A	F	
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	Σ(20,11,13,16)	3		0		3		0		0		6
10	50/50, Mathematics, 33\47	Σ(20,11,13,16)	2		3		3		3		0		11
11	70/30, Computer,12\16\20	Σ(20,11,13,16)	1		1		0		0		1		3
12	70/30,Mathematics, 24\11\13	Σ(20,11,13,16)	1		1		0		0		0		2
13	50/50, Computer, 21\29\30	Σ(20,11,13,16)	0		0		0		0		0		0
14	50/50,Mathematics, 33\22\25	Σ(20,11,13,16)	0		0		0		0		0		0
15	Useful predictions, ∑	Lines 1-14	13		11		12		8		4		48

III. COMPARATIVE EVALUATION OF ALGORITHM EFFICIENCY

A comparative evaluation of the performance of algorithms, as mentioned, is a more complicated procedure than evaluating the effectiveness of an individual classifier Table II. Three problems arise here: the choice of an adequate comparison criterion, the observance of the conditions for its applicability, and the required number of tests to generate a statistically valid statement.

1) Overall accuracy

summarizing the quality of forecasts of individual algorithms as shown in (row1, Table II). shows that in 3×4=12 experiments on all dataset, the classifiers with NB showed the highest overall forecast accuracy (A) about 79%, and classifiers with a lower 67% with 1NN. For that, to reduce ambiguity, we cannot increase the number of experiments by using experiments with other data sets, since in these experiments the values of A are much lower. In particular, this demonstrates an estimate of the average value of A equal to 74±12% for all 12×5=60 cases of the application algorithms in experiments with a complete set of data. Strictly speaking, we cannot even increase the number of new series, since new data samples will be dependent on each other. All the ways that used for useful experiences will be discussed below.

TABLE II. CHARACTERISTICS OF THE PERFORMANCE OF ALGORITHMS FOR BINARY CLASSIFICATION

№	Characteristic	Number of experiments, N	J48	MLP	NB	SVM	1NN	Average
1	Overall accuracy A,% data set	12	75±12	74±11	79±13	76±10	67±15	74±12
2	Sum of useful forecasting, ∑	48	34	33	37	27	13	29±10
3	Average ranks <R_j>, χ^2_4=14,01, $CD_{0,1}$=1,37	16	2,59	2,56	2,25	3,78	3,81	3,0
4	Average ranks <R_j>, χ^2_4=19,59, $CD_{0,1}$=0,99	32	2,83	2,34	2,28	3,33	4,09	2,98
5	Average ranks <R_j>, χ^2_4=34,57, $CD_{0,1}$=0,79	48	2,90	2,69	2,04	3,18	4,11	2,98

2) Useful forecast

For the criterion of useful forecasts ∑ (row2, Table II), there is no prohibition on the use of data from all experiments. The best algorithms for binary classification are J48, MLP, and NB in three series that presented by ∑=33÷37 useful predictions. This represents 68-77% of the number of experiments, while SVM and NN gave useful forecasts in 56% and 27% of experiments respectively. In general, the criterion ∑ almost certainly points to the low efficiency of the k-NN algorithm.

3) Friedman's rank test

To clarify the previous conclusions and more reliable estimates of the comparative efficiency of the algorithms, we used the Friedman rank criterion used in [15-16]. The procedure of this criterion consists in ranking R_{jn} of the individual achievement of the *j*-th classifier in the N experiment and the subsequent calculation of the Friedman χ_F^2 statistics according to (5a).

The Friedman test checks how significant the deviations of the average rank $<R_j>$ of individual classifiers of the average rank, calculated from the equivalence condition of all classifiers (equal to 3 if 5 classifiers). The null hypothesis of the test says that all classifiers are equivalent; therefore their average ranks should be equal. The hypothesis is tested by comparing the calculated values χ_F^2 with tabulated values of chi-square when k-1 degrees of freedom and a given level of rejection of the hypothesis.

$$\chi_F^2 = \chi_{k-1}^2 = \frac{12}{N\,k(k+1)}\left[\sum_{j=1}^{k}\left(\sum_{n=1}^{N} R_{jn}\right)^2\right] - 3N(k+1) \quad (5a);$$

$$CD = q\,(\alpha,k)\sqrt{\frac{k(k+1)}{6N}} \quad (5b)$$

We applied the Friedman criterion (5a) to the ranking results of classification algorithms in 48 experiments with different data sets and samples (lines 1-10, Table I) . In our case, the rank of the binary classifier algorithm could be from 1 to 5, where the number 1 meant the highest, and the number 5 the lowest rank.

The rank was set depending on the comparative value of the minimum F-measure, when the higher among others, the rank of the algorithm was higher. we explained this with an example of the results of the first experiment contained in (line 1, Table I).

The NB algorithm obtained the highest F-measure value equal to 80% and it obtained the first rank R_{31} = 1, the J48

(SVM) algorithm got the second (third) rank R_{11}=2 (R_{41}=3). If the values of the F-measure for two binary classifiers were the same, then the rank was set taking into account the value of the general accuracy A. If both indicators were equal, the classifiers get the intermediate rank. For example, in the same experiment, MLP and 1NN had F = 52% and A = 77%, both got a rank of 4,5.

A preliminary study revealed the following, that when the ranking of algorithms is based on the results of data forecasts with a significant contribution of useful forecasts, the null hypothesis is discarded at once. The average ranks of the five classification algorithms are presented in Table II. So the calculation according to (5a) for N=16 experiments Table II gives χ_4^2=14.01, which means the rejection of the null hypothesis with a critical level of 99%.

Obviously, with an increase in the number of experiments to N=48, the level of rejection of the hypothesis of equivalence of algorithms reaches to 99.9%.

4) Retrospective (Nemenyi test)

The conclusion of the Friedman test about the lack of equivalence of the five classifiers does not yet indicate the statistical significance of this difference. To solve this problem, after the rejection of the null hypothesis can be applied retrospectively Nemenyi criterion [13, 15, 16]. The test is based on the calculation of the critical difference (CD) according to (5b) for various significance levels q (α, k) (for α=0.1 and k=5 q(α, k) = 2,498 [15]), and this allows to detect a group of classifiers causing the rejection of the null hypothesis. Estimates (Table II) show that, the critical difference CD ranges of the best NB algorithm and the worst k-NN algorithm overlap after the first 16 experiments at the level of 90% significance (α=0.1). However, with an increase in the number of experiments, there is a confident "dispersal" of the average ranks of NB and k-NN among themselves. When N=48 the difference in average rank got 4,11-2,04 = 2.07> 2 × $CD_{0,1}$=1,56 and it is clear that the effectiveness of classifiers based on NB and k-NN varies with the level of confidence of 90%. For that, J48, MLP and SVM Algorithms remain equivalent to the average rank ≅3.

IV. ENSEMBLE FORECAST

In practice, various technologies for building ensemble forecasting are used [9,13]. Usually, independent "base" classifiers are created first, and then uses their a combination. Moreover, the predictions of the ensemble of classifiers will be more accurate than the forecast of the best of the single classifiers, if only the members of the ensemble give useful forecasts [13]. Thus, to get a good ensemble, it is necessary to choose the most accurate useful basic classifiers with various models for the connection of features and classes of objects h_t [13].We have already done the bulk of this work. Firstly, we were convinced that all binary classifiers h_t give useful forecasts with the most informative set of predictors (lines 1-2, Table I). Secondly, the ranking-based criteria revealed that the NB classifier is the best classification algorithm, followed by in descending order of effectiveness: MLP, J48, SVM, and k-NN.

Note that at N=48, NB and k-NN differ even with the level of 95%, since q(0.05.5)=2.728 and $CD_{0.05}$=0.88. Therefore, further forecasting by an ensemble can be consciously performed based on combinations of these classifiers. For example, choosing the three best classifiers for an ensemble, or diversifying it by taking all five algorithms.

A) Mathematical sheet

For the convenience of forecasting by an ensemble of binary classifiers, we will make the following useful transformation of the output information. Let the symbol

classes of binary classification "not passed" and "passed" correspond to integers - ± 1. Then the forecast of the classifier of the binary classifier h_t for some vector x_i will take two values $h_t(x_i)=y_t \in\{-1,1\}$. In this case, when by "a simple majority vote", the forecast result is S(x) of the ensemble $H_T(x)$ consisting of an odd number T=2n+1 of binary classifiers is equal to the sign of the algebraic sum of forecasts $h_t(x_i)$:

$$S(\mathbf{x}_i) = sign\left[H_T(\mathbf{x}_i)\right] = sign\left(\sum_{t=1}^{T} h_t(\mathbf{x}_i)\right),\ where\ T=2n+1, \left[H_T(\mathbf{x}_i)\right] \in \{\pm 1, \cdots, \pm T\} \quad (6)$$

Under conditions (6), an ensemble gives an erroneous forecast only when mistaken n+1 classifiers, i.e. more than half of the members of the ensemble. The main advantage of the ensemble forecast is that by the deterministic forecast (6) it is relatively easy to turn to a probabilistic forecast.

Suppose, that each object has probabilities P_{-1} assigning it to class "-1" and probability P_{+1} assigning it to class "+1". We will measure the probabilities in percent. Then every single classifier can generate two probability values P_{-1} 0% and 100%, and two corresponding values of P_{+1} 100% and 0%. It is easy to observe that the sum of the predictions of the classifiers from the ensemble H_T (x_i) referred to the number of classifiers T represents the probability difference ΔP (x_i)= P_{+1}–P_{-1} (7).

$$\Delta P(\mathbf{x}_i) = P_{+1} - P_{-1} = \frac{H_T(\mathbf{x}_i)}{T} \times 100\%\ ,\ where \quad P_{\pm 1} = 50\% \pm \frac{\Delta P(\mathbf{x}_i)}{2} \quad (7)$$

Where the probabilities P_{-1} and P_{+1} of assigning an object to one or another class are easily calculated.

B) Deterministic forecast

In Fig. 2a, illustrates the forecast of performance of 48 students by two ensembles of classifiers. The first of the H9 ensembles included classifiers using J48, MLP, and NB trained on three sets of predictors:

All data set and two sets of predictors without (SED) and (DD) (Table I), only 3×3=9 classifiers. In the second ensemble H15 used 3×5=15 classifiers: forecasts of classifiers using SVM and k-NN were added to the ensembles. As shown the ensemble H9,H15 made 4,6 erroneous forecasts. This is a quite good performance since the best classifier based on NB for this sample of data students gave 5 erroneous predictions on a complete set. For that, to select the best ensemble method, conditionally called by us "heterogeneous ensembles with a variety of predictors", it is possible to use the above procedure for finding the best ensemble. For this, we need a sufficiently large number of new series of experiments and data of large dimension.

Therefore, in this study, we limited to comparing the performance of ensembles for 8 different student samples, for which the sum of erroneous forecasts of various ensembles and classifiers was calculated. It can definitely be considered that the aggregate forecast quality will be higher with fewer errors.

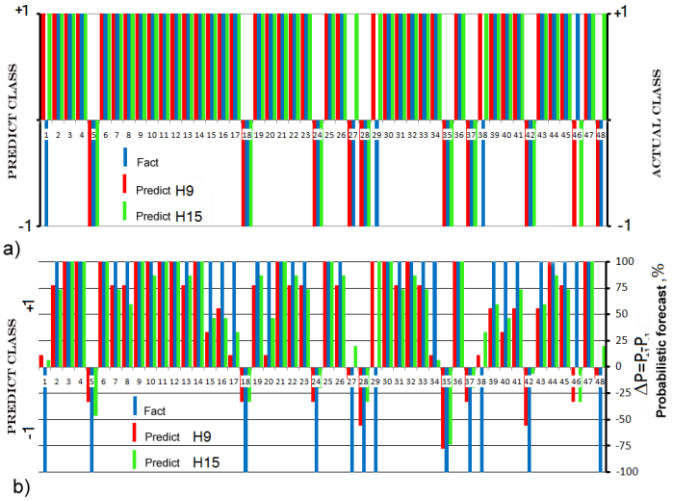

Fig. 2 a) Deterministic and b) probabilistic forecast by ensembles of 9 (H9) and 15 (H15) binary classifiers with variation of predictors. The horizontal numbers of 48 students.

Then, if the quality of the best classifier in each experiment is taken as 100%, it turns out that the forecast quality of the NB classifier was 84%, the best of the ensembles was 92%, the quality of the ensembles H3 and H5, consisting of three (NB, MLP, J48) and five algorithms (NB, MLP, J48, SVM, k-NN), and trained on a set of data, 91 and 89% respectively, and ensembles with a variation of the predictors H9 =91% and H15=86%.

Obviously, in each specific case of the forecast, we cannot predict in advance which of the algorithms or ensembles will be the best. Therefore, the noticeably better results of ensembles H3, H5 without variation of predictors, and ensembles H9 and H15 with the variation of predictors in comparison with single classifiers, indicate the prospect of increasing the quality of forecasting using the ensembles we created.

C) Probability forecast

The features of performance in a probabilistic form are illustrated in the graph in Fig. 2b, where, for the convenience of comparison with a deterministic forecast, represented the difference classes of probability $\Delta P(x)=P_{+1}–P_{-1}$ (7).

Obviously, when $\Delta P(x)> 0$ (<0), then the predicted class of student success "1", ("-1"). As we see, the deterministic forecast was carried out in some cases with a very small difference in the probabilities of the classes $\Delta P(x)$. For example, for student number 1, an erroneous deterministic forecast was obtained with a probability difference $\Delta P(x)=10\%$, when the probability of an erroneous class "+1" was 55%, and the probability of the correct class was 45%.

The ability to detect students with close probabilities of belonging to different classes in the prediction, or the detection of large erroneous probabilities in test mode, allows us to detect hidden data anomalies. In a semantic formulation, such anomalies can be expressed in the first case, as "the results of his / her training are not predictable", and in the second as "he/she should not (or should) have failed the test".

Obviously, such an interpretation of the ensemble probabilistic forecast and assessment, allows us to automatically identify groups of students with different levels of learning problems and substantiate actions that fix the problem.

V. NEW METHOD OF QUANTIFICATION

The obvious estimate of the unknown frequencies ($\hat{p}^+$, $\hat{p}^-$) of the binary distribution of student success can be as a sum of N^+ forecasts relating to the class of "+1" binary classification [11-12]. This is the previously mentioned in Classify & Count approach [12]. Indeed at the output of a binary classifier, we have:

$$\hat{p}^+ = N^+/N; \quad \hat{p}^- = N^-/N = 1 - N^+/N = 1 - \hat{p}^+ \qquad (8)$$

It is also obvious, that (8) remains valid for a deterministic forecast produced by an ensemble classifiers. In Fig.6, shows the actual and forecast distributions of learning outcomes in mathematics (Fig.6a) and computer (Fig.6b), obtained by a single Bayesian classifier NB and ensembles of 3, 5, 9 and 15 classifiers, trained and tested on samples with an equal number of objects N=80.

Predictions of ensembles with probabilistic correction of classes are indicated by star, Fig. 6a shows the unrounded forecasts of $N^{(+)}$, in Fig. 6b mixed forecasts are added H3&=(H3 + H3*)/2 and H9&=(H9 + H9*) / 2.

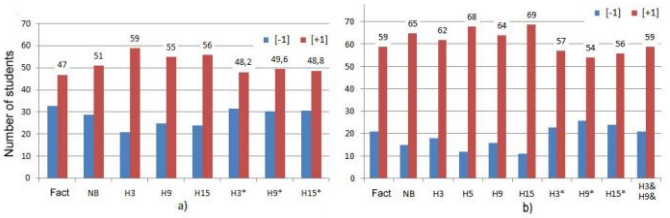

Fig.3. Binary quantification using ensembles classifiers. Predictions of collective success in (a) mathematics and (b) computer in a sample of 80 students..

Quantification quality assessment [11] can be made by the difference between the actual and forecasted sum of the N^+ class "+1". So in the forecast in mathematics, NB predicted better than ensembles and gave an estimate of the number of successful students N^+=51 by 4 units greater than the actual number equal to 47. On the contrary, in the forecast for computer, the best ensemble was of three H3 classifiers, which increased the number of collective successes by 3 units. When comparing the results of quantification of single and determinate ensemble quantifiers, obviously, the same qualitative conclusions are obtained as in the case of an individual deterministic forecast: The ensembles H3 and H9 make slightly more mistakes than the best quantifier in this experiment. Nevertheless, the uncertainty of choosing the best single quantifier in each of the experiments makes us consider the determined ensemble quantification as a convincing alternative to the prediction of a single quantifier.

A) Quantification probabilistic of ensemble

The probabilistic form of the ensemble forecast of the collective achievements of students has a more rigorous and rational justification than the individual probabilistic forecast.

Indeed, in our previous discussions based on the student's probability of success (7), we tried to predict the evolutionary paths of each individual student, and some imaginary opponent could always challenge our probabilistic forecast, citing an example from a real-life case. If we consider the obvious practical benefits of probabilistic forecasting as a way of identifying a cluster of students with a small probability difference $\Delta P \cong 0$. however, here, when we further study this group problems, we would need some way to clarify the class of each object, in other words, a new classification stage (such an idea is embedded in the Adaboost technology) [13].

Otherwise, the situation is in the case of forecasting collective success. Here, the probabilistic forecast can be considered as an assessment of the possibility that among the set of N^+ persons assigned to successful students, there will be a certain number of persons Δ^- which according to the results of training, will be in another set (class "-1").

Similarly, among the group of students that we rank among the set N^-, there is a group of some unknown students, the numerical strength Δ^+ which contrary to expectations will overcome the test. Suppose that in some way the abundances of these groups Δ^- and Δ^+ are estimated, then using such estimates we can try to correct the forecasts issued earlier on the abundance of the groups N^- and N^+. We propose the following estimation method, which uses the previously obtained results of an individual probabilistic forecast of the ensemble (9).

$$\Delta^+ = \sum_{j=1}^{N^-} \left(1 - \frac{H_T(\mathbf{x}_j)}{T} sign\left(H_T(\mathbf{x}_j)\right) \right); \qquad (9a)$$

$$\Delta^- = \sum_{j=1}^{N^+} \left(1 - \frac{H_T(\mathbf{x}_j)}{T} sign\left(H_T(\mathbf{x}_j)\right) \right) \qquad (9b)$$

Formula 9, it is assumed that the group's Δ^+ and Δ^- can be considered as the sum of the probabilities of students moving from one predictive class to another. Let us explain this with the following example, Let's assume there is a test sample of 10 students, the ensemble forecast showed in (7), that the probability of success of each is 0.6 (P_{+1}=60%).

In the case of a deterministic forecast, we assigned all these students to a sample of successful N^+=10 and we got N^-=0. However, if we consider that each of N^+ students has a probability of assignment (him/her) to another class (P_{-1}=40%) and summarize all these probabilities (9b), we formally get a group of Δ^-=10×(1- 0.6)=4 students.

Obviously, by adding Δ^- to N^-, and subtracting the same number from N^+, we obtain a new forecast sample with $N^{(+)}$= 6 and with $N^{(-)}$= 4.

In the general case, Δ^+ and Δ^- can take non-integer values, in addition to direct "hard" exchange with subtracting Δ^+ and Δ^- from the initial values of N^- and could lead to a number of undesirable effects.

Therefore, when forecasting by the ensemble, we performed the following "soft" method of correcting N^+ with

subsequent rounding of the correction result to the whole $N^{(+)}$:

$$N^{(+)} = round\left(\frac{\left(N^{+}+\Delta^{+}\right)\cdot N}{N+\Delta^{-}+\Delta^{+}}\right);\quad N^{(-)} = N - N^{(+)} \qquad (10)$$

Thus, applying (10) to the initial frequencies N^+ and the corrector Δ^- given in the previous example, instead of a sample with 6 units in the class "+1" by 4 units in the class "-1", we get a corrected of 7 to 3 sample. The results of calculations of samples using (9) and (10) for the correction of the initial predicted frequencies of the ensembles N^+ and N^- are shown in Fig 3. As see, the ensemble quantifiers passed the probabilistic correction (marked with *), estimates of the distribution of classes improved significantly, comparison with the initial quantifiers. This is practically noticeable in the results of the $H3^*$ ensemble, which gave both the best estimate of $N^{(+)}$ from above with one error in the computer (Fig. 3a) and the best estimate of $N^{(+)}$ from below with two errors (Fig. 6b).

As see, in the first case, the relative error in the number of the smallest class decreased from 6% in the best quantifier to 3% in H3* for probabilistic correction, and from 14% to 9% in the second case. We also note that some linear combinations of deterministic and probabilistic forecasts give absolutely accurate predictions (Fig. 3b).

B) Preliminary conclusions

obviously, the determinate forecasts of ensemble classifiers suffer from an underestimation of the number N^- minority class. The natural imbalance of the sample leads to the fact that single classifiers, and after them, their ensembles overestimate the number of objects of the majority class N^+.

Improving the accuracy of individual forecasts when voting strong classifiers, ensemble members not leads to an improvement in the proportions between N^- and N^+.

In contrast, probabilistic correction can significantly improve the results of the Classify & Count method. This is due to the fact that with a larger number of the majority class, large Δ^- groups are simultaneously generated in it, allow to increasing the predicted number of the N^- minority class.

VI. Conclusion

In this work we evaluate the machine learning algorithms in the context of their applicability to student's performance forecast. Firstly, the analysis of 86 cases of SPF in south Asia countries demonstrate that there are five most popular SPF algorithms: J48, NB, MLP, k-NN SVM. These algorithms are applied to four types of data: final grades, internal assessments, DD and SED, forming a complete set of data. Generally, the relative high overall accuracy of forecast is reported. Unfortunately, the detailed analysis of these reports depicts that useful forecast, when single class F1-score achieved 50%, only in half of cases.

So that we evaluate relative SPF efficiency of classifier by using Friedman rank test and post-hoc Nemenyi test, and show that the strong and weak binary classifiers are based on NB and k-NN, but MLP, J48 and SVM classifier perform intermediate efficiency.

The deterministic and probabilistic ensembles for resolution of SPF and quantification problems are proposed. It is shown that the predictive ensemble of heterogeneous classifiers using datasets with different completeness significantly improves SPF. Also it is shown that the aggregation of probabilities generated by ensemble may be useful to construction of new quantification method that will outperform the Classify & Count algorithm and will be applied as the tool to the collective achievements of the students. However, the results of the ensemble forecast and ensemble quantification suggest that multi-class methods of classification and quantification will be found.

References

[1] Romero C., Ventura S. Educational data mining: A review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 2010. vol. 40, no. 6. pp. 601–618.DOI:10.1109/TSMCC.2010.2053532

[2] U.S. Department of Education, Office of Educational Technology. Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. Washington, D.C., 2012. Available at: https://tech.ed.gov/learning-analytics/edm-la-brief.pdf (accessed: 03.07.2018).

[3] Baker R.S., Inventado P.S. Educational Data Mining and Learning Analytics. In: Larusson J., White B. (eds) Learning Analytics. Springer. New York, NY. 2014. pp. 61-75. DOI:10.1007/978-1-4614-3305-7_4

[4] Calvet Liñán L., Juan Pérez Á.A. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC. Universities and Knowledge Society Journal. 2015. v. 12, n. 3. pp. 98-112. DOI: 10.7238/rusc.v12i3.2515

[5] Berland M., Baker R.S.. Blikstein P. Educational Data Mining and Learning Analytics: Applications to Constructionist Research. Tech Know Learn. 2014. vol. 19. pp. 205–220. DOI: 10.1007/s10758-014-9223-7

[6] Slater S., Joksimovic S., Kovanovic V., et al. Tools for Educational Data Mining: A review. Journal of Educational and Behavioral Statistics. 2017. vol. 42 no 1, pp. 85-106. DOI: 10.3102/1076998616666808

[7] Shahiri A.M., Husain W., Rashid N.A. A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Comput. Sci. 2015. vol. 72. pp. 414–422.DOI: 10.1016/j.procs.2015.12. 157.

[8] Mukesh Kumar, Y.K. Salal, Systematic Review of Predicting Student's Performance in Academics, International Journal of Engineering and Advanced Technology (IJEAT), 2019, vol.8 no.3, pp.54-61.

[9] Jiawei H., Kamber M., Han J., Pei J. Data Mining: Concepts and Techniques. Elsevier Inc. 2012. 740 p. DOI: 10.1016/B978-0-12-381479-1.00001-0.

[10] Halkidi M., Vazirgiannis M. Quality Assessment Approaches in Data Mining. In Maimon O., Rokach L. (eds.). Data Mining and Knowledge Discovery Handbook. 2nd ed. Springer Science + Business Med. 2010. Ch. 30. pp. 661-696. DOI: 10.1007/978-0-387-09823-4_1

[11] González P, Castaño A., Chawla N.V, Del Coz J.J. A Review on Quantification Learning. ACM Comput. Surv. September 2017. 2017. v. 50, no. 5, Article 74, 40 p. DOI: 10.1145/3117807

[12] Forman G. Quantifying counts and costs via classification. Data Mining and Knowledge Discov.Oct. 2008. vol. 17, no.2. pp 164–206. DOI: 10.1007/s10618-008-0097-y

[13] Zhou Z.-H. Ensemble Methods: Foundations and Algorithms. Series: Chapman & Hall/CRC Machine Learning & Pattern Recognition, CRC press. 2012.236 p.

[14] Hämäläinen W., Vinni M. Classifiers for Educational Data Mining. In: Romero C. et al. (eds) Handbook of Educational Data Mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor and Francis Group, LLC. 2011. pp. 57-71.

[15] Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. 2006. no. 7, pp. 1–30.

[16] Sheskin, D. Handbook of parametric and nonparametric statistical procedures. David J. Sheskin.-2nd ed. Chapman & Hall/CRC. 2000. 973 p.

[17] Y. K. Salal, S. M. Abdullaev. Reposotory of Github,2019. URL: https://github.com/YassKhudheir1983/Educational-Data-mining