Using Data Quality to compare the Prediction Accuracy based on diverse annotated Tutor Scorings

Sylvio Rüdian
Humboldt-Universität zu Berlin,
Weizenbaum Institute
Berlin, Germany
ruediasy@informatik.hu-berlin.de

Niels Pinkwart
Humboldt-Universität zu Berlin,
Weizenbaum Institute
Berlin, Germany
niels.pinkwart@hu-berlin.de

ABSTRACT

Cross-validation is a wide-spread approach to understand how well a prediction model performs with unseen data. While this is the state of the art, machine learning is often used for educational purposes in educational data mining. Whether a system is applicable and generalizable in practical settings is based on the cross-validation accuracy. One major problem is that the quality of annotated data is often worse due to different raters that score equal tasks differently, even if they were trained before. In this paper, we did an experiment where 1.200 texts of three difficulty levels in an open writing task for language learning were scored by two tutors independently to get the inter-rater reliability score for measuring the similarity across their grades. We used the existing scorings of other tutors of the system to train a random forest regressor for predicting scorings based on the texts. We found out that the accuracy has a strong relationship to the inter-rater reliability score and propose a new measurement that combines both metrics for scenarios where data was annotated by tutors, that could principally be diverse.

Keywords

Tutoring systems, scorings, data labeling, inter-rater reliability, cross-validation

1. INTRODUCTION

As long as tutor scorings are used as a basis to train machine learning systems, there is a bias of subjectivity. Research has shown that the agreement among scores given by tutors often varies [1]. Depending on the task and scale, tutors reach different inter-rater reliability scores. Practical settings have shown that even when teachers were trained for grading, there is a gap. Thus, formal exams are often graded twice and in case that there is a huge gap, a third grader needs to be taken into account. For the field of machine learning, we need thousands of scored tasks, e.g. for automated essay grading. From the practical point, it is understandable that scorings cannot be done by the same tutor all the time. Tutors' time is a limited resource and thus there is the need to score tasks by different experts. If we consider machine learning approaches, there are many examples of prediction tasks, where researchers try to imitate teacher scorings, based on different features. As the reduction of a text or task to features removes information that could be important for a good evaluation, automatic scorings cannot be perfect. Using data gathered by tutors where even scorings for the same texts or tasks are not always equal we think, that it is not fair to compare the prediction accuracy in education in general if we use tutors' labeled datasets.

In machine learning, the proper way to decide whether a system generalizes well is to do cross-validation [2]. Therefore, the data is split into several pieces. The model will be trained on all the data, except from one piece. This piece is used to evaluate the model as we know features and the concrete label. Based on the features, the system creates a prediction using the trained model. The predicted label can be compared with the known one. With every piece, the leave-one-out method (or alternative ones) can be applied to get an averaged accuracy. The main advantage of this method is to create a prediction on previously unseen data. Thus the evaluation shows whether a model generalizes well. Observing this value in detail, we often notice that the accuracies are between 0.6 and 0.8, e.g. 0.6 for 8 classes and 0.78 for 4 classes in [3] or 0.7 in for 4 classes in [4]. From the perspective of machine learning, these are bad values as it means that 3-4 of 10 predictions are wrong.

To have a good and fair measurement for comparison it is necessary to take the inter-rater reliability of human raters into account as in general, the prediction cannot be better than the ratings among raters that have been used for training the machine. The inter-rater reliability is a score of consistency among raters. According to McGraw & Wong [5], the minimum value should be 0.6 as the cut-off for acceptability. Wang & Michelle [6] did a comparative study to compare human essay scoring and reached an inter-rater reliability score (IR score) of 0.62, using the Intraclass Correlation Coefficient. Williamson proposes that an IR score lower than 0.7 is not applicable [7].

The accuracy of predictions is often measured as the comparison of the prediction of the machine and the rater annotations. But the machine itself was trained based on the raters scores, which could differ among raters [8] [1]. It is not surprising that the predicted scorings by machines correlate with the human rater scorings as they are the training base [6]. In contrast, Williamson has shown statistically significant differences between human and machine rating scores [7]. The question remains: comparing all the systems, what is the best and most applicable one? Using the accuracy only fails as the major problem is the quality of the training data – and not the resulting accuracy in cross-validation.

In this paper, we propose an extension of the cross-validation to have a fair measurement for comparing educational predictions, where training data was gathered from tutors. We focus on language learning and examine two research questions:

RQ1: What is the correlation of the inter-rater reliability score in essay scoring for language learning, compared to the prediction accuracy?

RQ2: Combining the cross-validation with the inter-rater reliability score, what is a fair interpretable measurement taking both metrics into account?

2. METHODOLOGY

To address RQ1, two tutors had the task to score open text submissions of three tasks. All tasks had a different difficulty level, easy (1), medium (2), and difficult (3). For every task, we had 400 user submissions, in sum 1.200. Both tutors got access to the tasks and they got 10 typical scorings for a pre-training. Then, all submissions were scored by both instructors independently of each other, using scores of 1 (very good) to 4 (bad/not acceptable). The scoring procedure lasts 1 week for every tutor.

Then we prepared a random forest regressor [9] as a classifier to train a prediction model for essay scoring based on at least 1.200 scorings for each task, that are already existing in the learning system, independently of the scorings from the previous step. These scorings are created by different tutors, where each text was scored only once. So we did not use the data of the previous step for a comparison to avoid training with the new labeled dataset. From practical settings we know that intermediate grades are quite subjective, thus we concentrate on grades 2 and 3 only, which represent "good" and "satisfactory" that are used as labels for the classification problem. The accuracy for prediction in cross-validation (CV) was gathered for each task separately.

Within the next step, we compared the similarity among both tutors of the first step with the accuracy of the second step to examine a possible relation. Finally, we propose a combination of both metrics that allow a fair comparison of the prediction accuracy with the IR Scores to address RQ2.

3. RESULTS

Figure 1 shows all IR scores and the prediction accuracy in a 10-fold CV. We can see that there is a good correlation between these metrics (correlation 0.88). We used the same approach for all tasks, but the IR-scores vary from 0.45 to 0.74, and the accuracies in 10-fold CV range from 0.47 to 0.64. The results show that the accuracy, as well as the IR-score, vary depending on the task. But, there is a strong positive relationship between the maximal achieved prediction accuracy and the IR score.

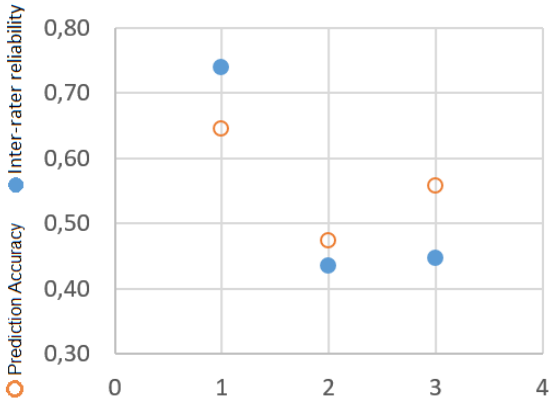

Figure 1. Inter-rater reliability score of two tutors for three tasks, separated by increasing difficulty level and prediction accuracies in cross-validation.

4. NEW MEASUREMENT

The main idea is to combine the classical cross-validation with the inter-rater reliability score. The CV addresses the accuracy of a trained prediction model. As there are multiple versions of the CV, e.g. leave-one-out or leave-p-out (where p is a range of the dataset), we use CV as a general concept and do not limit our approach to a specific version.

The similarity of tutor scorings can be measured by using a correlation coefficient. We chose the Pearson correlation coefficient (PCC) for applying to a sample [10]. It is not outlier resistant [11], but in the area of learning, large gaps can principally occur in ratings, e.g. the score from one rater is "very good / 1" and from another, it is "very bad / 4". This will impact the resulting correlation coefficient. For our new measurement, this is important as this gap influences the training data as well and thus, it influences the prediction accuracy negatively due to a large bias.

We propose a combination of both metrics, namely CV_{PCC}, defined by the following formula:

$$CV_{PCC} = 1 - |CV - PCC^{+}|$$

under the constraint $0 \leq CV, PCC^{+} \leq 1$. The CV accuracy is defined as a number between 0 and 1 [2] and $PCC^{+} = |PCC|$ as we only consider the similarities, not whether the PCC is positive or negative. The CV_{PCC} is a new value where $CV_{PCC} \in [0,1]$, similar to the CV. In the following paragraph, we show that CV_{PCC} cannot be smaller than 0 and never more than 1. Let CV and PCC^{+} be defined as above. Then we examine whether $\exists CV, PCC^{+}$: $1 - |CV - PCC^{+}| < 0$ or $1 - |CV - PCC^{+}| > 1$.

$$1 - |CV - PCC^{+}| < 0 \iff 1 < |CV - PCC^{+}|$$

With $PCC^{+} \geq 0$ we set $PCC^{+} = 0$ to maximize the value for $|CV - PCC^{+}|$. As $0 \leq CV \leq 1$, the maximum value for CV is 1. This follows: $1 < |1 - 0| \iff 1 < 1$, which is a contradiction.

$$1 - |CV - PCC^{+}| > 1$$
$$\iff 1 > 1 + |CV - PCC^{+}|$$
$$\iff 0 > |CV - PCC^{+}|$$

The absolute value x is defined as $\forall x \in \mathbb{R}: |x| \geq 0$ [12]. Thus, with $x = CV - PCC^{+}$: $0 > |x| \iff |x| < 0$. According to the definition of the absolute value, this is not existing in $\mathbb{R}$. Finally, we showed the second contradiction and can conclude that $CV_{PCC} \in [0,1]$. □

In Figure 1 we can see that the CV accuracy, as well as the IR scores, range from 0.44 to 0.74. If we just compare the CV accuracy, we can conclude that there is a high fluctuation. Using the new CV_{PCC}, the scores range from 0.89 to 0.96. Here, the fluctuation is much lower and we now can compare this value with other tools and different datasets.

5. DISCUSSION

In general, we know that having a low inter-rater reliability score is an indicator of a bad quality of training data. Although we used the same amount of data to train the classifier for each task we can observe that it is not fair to compare the achieved accuracy in prediction only. As there is a strong positive relationship between the accuracy and the inter-rater reliability score we propose to combine both metrics when comparing the result with other datasets. Otherwise, that is what our results show, the accuracy differs across tasks, and results are based on the task selection.

In an optimal setting, where all scorings are the same for equal texts across different raters ($PCC^{+} = 1$), follows $CV_{PCC} = \text{CV}$. Observing the other "extreme" side, where the PCC^{+} and the CV values are very low, we can still achieve a high CV_{PCC}, as the

accuracy will be low if labels are diverse for equal feature values. The higher the range between PCC^+ and CV is, the lower CV_{PCC} will be, which means that the relation between both metrics is low. With that information we address RQ2. Thus, this is an indicator of whether the model needs improvement or whether the accuracy cannot become better as the training base has a low quality due to diverse labeling based on different quality expectations of tutors. This interpretation of the value can be helpful to optimize the model. As we use cross-validation as a general metric, our approach is not limited to specific classification methods. We used the random forest regressor, but we can use other classification-based methods like neural networks, support vector machines, or others as long as we get access to the CV score.

We need to emphasize that our method requires a further labeling step to get the inter-rater reliability score across at least two tutors, where each text needs to be labeled twice. This increases the labeling costs. To reduce the amount of work, we could principally use a subset of already labeled texts that has to be labeled by a new tutor to understand the data quality. If a low value will be detected, we know that the resulting accuracy will differ from experiments with other datasets due to the low agreements. We can argue that knowing the problem of diverse scorings is a good fundament to optimize further scorings by a better pre-training of raters. But in praxis, often thousands of labels are existing based on the data that was collected over the last years. Thus, only for future data collection, there can be optimization. If we want to use existing datasets, we propose to use the CV_{PCC} for a fair comparison in relation to other datasets.

In our experiments, we used two separate datasets, one that contains the scorings of the two tutors and one much larger set, where more texts were scored by other tutors. The first was used to get the PCC^+ score and the other to train the classifier based on the maximum achievable CV score. To benefit from the extra labeling, we could enhance the training dataset by the data where the two tutors had equal scorings for the same texts.

Our proposed metric is limited to datasets that were annotated manually. If we have labels that are automatically processed (e.g. the achieved scores in interactive tasks in an online course or whether a student drops out), normally we do not have a diverse annotated dataset. Thus we recommend using the CV_{PCC} in all scenarios where tutors are involved and where diverse annotations (e.g. in scorings) play a role. This is early-stage research, limited to three difficulty levels of specific open-writing tasks. To generalize our findings, the next step is to compare more tasks and the resulting CV_{PCC}. Besides, further studies in other learning domains are required to verify the found relations of the metrics. Our first findings are promising.

6. CONCLUSION

In this study, we examined the relation of the inter-rater reliability of tutor scorings and the accuracy that can be achieved to predict two concrete ratings. In our setting of language learning, we focused on three open writing tasks of different difficulty levels, those accuracies in prediction differ. Based on our results, we observe that there is a strong relationship between both scores, even though both metrics were derived using datasets from multiple raters. Thus we can see that datasets, labeled by tutors, can differ. This infers the data quality and the maximum achievable accuracy in prediction. To use possibly diverse annotated data by tutors and for comparing the prediction results, we propose a new method of combining both metrics to allow fair comparison across different datasets. This new metric can help scientists in educational data mining to compare results of different tutor-based labeled datasets and it helps to understand whether a model or the dataset needs improvement.

7. ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII127 (Weizenbaum-Institute). The responsibility for the content of this publication remains with the authors.

8. REFERENCES

[1] S. Elliot, "A study of expert scoring, standard human scoring and IntelliMetric scoring accuracy for statewide eighth grade writing responses (RB-726)" Newtown, PA, Vantage Learning, 2002.

[2] J. Sha, "Linear Model Selection by Cross-Validation" in Journal of the American Statistical Association, Taylor & Francis, Ltd, 1993, pp. 486-494.

[3] S. R. Bowman, G. Angeli, C. Potts and C. D. Manning, "A large annotated corpus for learning natural language inference" in arXiv 1508.05326, 2015.

[4] S. Rüdian, J. Quandt, K. Hahn and N. Pinkwart, "Automatic Feedback for Open Writing Tasks: Is this text appropriate for this lecture?" in DELFI 2020 - Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V., 2020, pp. 265-276.

[5] K. McGraw and S. Wong, "Forming inferences about some intraclass correlation coefficients" in Psychological Methods, 1(1), 1996, p. 30–46.

[6] J. Wang and M. S. Brown, "Automated Essay Scoring versus Human Scoring: A Comparative Study" in The Journal of Technology, Learning, and Assessment, 2007.

[7] D. Williamson, "A framework for implementing automated scoring" in Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, SanDiego, 2009.

[8] P. D. Nichols, "Evidence for the interpretation and use of scores from an Automated Essay Scorer" in American Educational Research Association (AERA), San Diego, CA, 2004.

[9] L. Breimann, "Random Forests" in Machine Learning 45, 2001, p. 5–32.

[10] R. Hunt, "Percent Agreement, Pearson's Correlation, and Kappa as Measures of Inter-examiner Reliability" in Journal of Dental Research, 1986.

[11] Y. Kim, T.-H. Kim and T. Ergün, "The instability of the Pearson correlation coefficient in the presence of coincidental outliers" in Finance Research Letters, Volume 13, Elsevier, 2015, pp. 243-257.

[12] E. H. Moore and H. L. Smith, "A General Theory of Limits" in American Journal of Mathematics Vol. 44, No. 2, The Johns Hopkins University Press, 1922, pp. 102-121.