Acute pancreatitis severity classification: accuracy, robustness, visualization[1]

Ekaterina Mangalova[2], Olesya Chubarova[3], Daniil Melekh [3] and Anton Stroev[4]

[2] *LLC RD-Science, Krasnoyarsk, Russia*

[3] *Siberian Federal University, Krasnoyarsk, Russia*

[4] *Krasnoyarsk State Medical University named after Prof. V.F. Voino-Yasenetsky, Krasnoyarsk, Russia*

e-mail: e.s.mangalova@hotmail.com

Abstract

The work is devoted to the problem of acute pancreatitis severity classification. This problem is characterized by a small amount of data, which leads to unstable estimations for new patients and a strong influence of the training sample on the predictions. In this paper prediction stability visualization based on violin plot is proposed and applied. A simulation experiments are carried out to study the stability of linear regression, support vector machine, random forest trained with various subsets.

Keywords: classification, machine learning, visualization, violin plot, bootstrapping.

Introduction

Early recognition of disease severity is important to identify patients on admission or during the first 24 to 48 hours who will require aggressive resuscitation. These patients should be treated in an intensive care unit or transferred to a high-acuity care hospital.

Classification of acute pancreatitis defines 3 degrees of severity according to the morbidity: mild, moderately severe, and severe acute pancreatitis.

Mild acute pancreatitis lacks organ failure or local or systemic complications. Pancreatitis resolves rapidly, mortality is rare, pancreatic imaging is often not required.

Moderately severe acute pancreatitis has transient organ failure, local complications, and/or systemic complications but not persistent (>48 hour) organ failure. The morbidity is increased as is mortality ($< 8\%$) compared with that of mild acute pancreatitis.

Severe acute pancreatitis is defined by persistent organ failure and patients usually have 1 or more local and/or systemic complications. Patients with severe acute pancreatitis that develops within the early phase are at a markedly increased risk (36%-50%) of death [1].

[1] The reported study was funded by Krasnoyarsk Regional Fund of Science, to the research project: Development and implementation of decision support system for acute pancreatitis diagnosis and treatment in the Krasnoyarsk Territory

The study was based on a retrospective analysis of 130 cases of acute pancreatitis: 47 cases from Krasnoyarsk Regional Clinical Hospital and 83 cases from RSBHI Regional Interdistrict Clinical Hospital 20 named after I.S. Berzon in the period from 2015 to 2017.

The task is to estimate of acute pancreatitis severity by using patient clinical examination data $D = \{(\bar{x}_i, y_i), i = 1, ..., 130\}$, where $\bar{x} = \{x^1, ..., x^{27}\}$ is set of features (Clinical Blood Analysis, Biochemical Blood Analysis, Ultrasound of pancreas, the results of the examination of the patient) measured in 130 patients.

1 Data preparation

1.1 Feature Scaling

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the Support Vector Machine is based on the distances between points. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

All variables are preprocessed using the min-max scaling.

Min-max scaling is the simplest method and consists in rescaling the range of features to scale the range in [0, 1]. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

, where x is an original value, x' is the normalized value.

1.2 Filling missing values

Data scientists often check data for missing values and then perform various operations to fix the data or insert new values. The goal of such cleaning operations is to prevent problems caused by missing data that can arise when training a model.

Two types of operations for "cleaning" missing values are implemented:

- Replacing missing values with a linear regression. If two features are strongly correlated linear regression is used to fill missing values. For example, the size of the head, body or tail of the pancreas may be absent due to poor visualization of the pancreas on ultrasound examination of the abdominal cavity. However, the size of the head, body and tail of the pancreas is highly linearly correlated and can be filled.

- Replacing missing values with a within-class median. If features there are not correlated missing values are replaced using a within-class median. This technique allows to avoid reduction of the influence of feature with a large number of missing values as in the case of replacement with median for the whole sample.

2 Accuracy estimation

Since the three classes are strictly ranked, the multi-class classification problem can be solved as a regression problem. As a result, each new object (patient) instead of the class number (1 - mild acute pancreatitis; 2 - moderately severe acute pancreatitis; 3 - severe acute pancreatitis) will be assigned a value from 1 to 3, characterizing not only the class of disease severity, but also how likely this severity class. For example, if the first patient has prediction 1.1 and the second has prediction 1.3, then although they will both be assigned to patients with mild severity of acute pancreatitis, but the probability that the first patient has a mild severity is higher than the second has one.

As accuracy criteria the following indicators were chosen:

- Mean Absolute Error (MAE);
- Mean Squared Error (MSE);
- Correlation Coefficient (Corrcoef);
- Number of Mistakes (NoM). If the prediction differs from the actual value by more than 0.5, it means that the classifier predict wrong class. Such forecasts will be called mistakes.
- Number of Mistakes x2 (NoM x2). If the prediction differs from the actual value by more than 1.5, it means that the classification error is more than one class (mild acute pancreatitis instead of severe acute pancreatitis or vice versa). Such forecasts will be called mistakes x2.

Table 1 contains accuracy of different algorithms calculated using leave-one-out cross-validation technique. Experiments show that SVM provides the greatest accuracy in all indicators.

Table 1: Accuracy of Linear Regression, SVM and Random Forest

	MAE	MSE	Corrcoef	NoM	NoM x2
Linear Regression	0.375	0.269	0.783	44	1
Support Vector Machine	0.354	0.243	0.808	35	0
Random Forest	0.413	0.293	0.765	43	1

3 Robustness

3.1 Small dataset problem

Acute pancreatitis severity classification task is characterized by small sample size for objective reasons. Analysts in medicine face with small dataset problem due to the prohibition on disclosure and dissemination of personal data. In such tasks, the analyst deals with the following challenges:

- Overfitting. With only a few data, the risk to overfit model is higher;
- Outliers. If analysts have millions of data, a couple of outliers will not be a problem. But with only a few, they will definitely skew prediction results.

The bootstrap procedure [2] can be used to evaluate the robustness of the predictions for the original sample and the effect of certain observations from the initial sample on the predictions.

3.2 Bootstrapping

The basic idea of bootstrapping is that inference about a population from sample data (training set) can be modelled by resampling the sample data and performing inference about a sample from resampled data. As the population is unknown, the true error in a sample statistic against its population value is unknown. In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data is measurable.

The bootstrap creates a large number of datasets that we might have seen and computes the statistic on each of these datasets. Thus we get a distribution of the statistic.

In our task, we are interested in the acute pancreatitis severity class of people worldwide. But we cannot measure all the people in the global population, so instead we sample only a tiny part of it, and measure that. Assume the sample (the training dataset) is of size N; that is, we measure the features (Clinical Blood Analysis, Biochemical Blood Analysis, Ultrasound of pancreas, the results of the examination of the patient) of N individuals. From that single sample, only one acute pancreatitis severity prediction can be obtained for each new patient. In order to reason about the population, we need some sense of the variability of the prediction that we have computed.

The most popular bootstrap method involves taking the original data set of N patients and randomly sampling from it to form a new sample (bootstrap sample) that is also of size N. The bootstrap sample is taken from the original by using sampling with replacement. On the first step, we randomly choose N_1 patients from the original data, On the second step, we randomly choose $N - N1$ patients from chosen on the first step. The key parameter for bootstrapping is the ratio between the number of unique observations in the bootstrap sample (N_1) and the initial sample size (N): $p = \frac{N_1}{N}$. This process is repeated a large number of times, and for each of these bootstrap samples we fit model (Linear Regression, Support Vector Machine and Random Forest) and make predictions for new patients.

After applying the bootstrap technique we can have a set of predictions for each new patient that can be analyzed and visualized to make the final decision.

4 Visualisation

4.1 Violin plot

Many different graphs and statistics interpret the characteristics of dataset.

While a box plot [3] only shows summary statistics such as median and interquartile ranges and gives information about location, scale, symmetry and tail thickness, the kernel density estimation shows the full distribution of the data. The difference between the box plot and kernel density estimation is particularly useful when the data distribution is multimodal. In this case a density trace shows the presence of different peaks, their position and relative amplitude.

Violin plots [4] combines the box plot and density trace smoothed by a kernel density estimator and can be used to show robustness of machine learning algorithms.

4.2 Comparison of Machine Learning algorithms

Figure 1 illustrates the influence of the training set on the prediction stability for typical observations from different classes (classes were determined by a medical expert): a - mild acute pancreatitis; b - moderately severe acute pancreatitis; c - severe acute pancreatitis. The ratio p between the number of unique observations in the bootstrap sample (N_1) and the initial sample size (N) is equal to 0.9. The density trace is plotted symmetrically to the upper and the lower of the horizontal box plot. Symmetric plot makes it easier to see the magnitude of the density. The black vertical line shows the median of the predictions, while the gray rectangle depicts interquartile range.

The graph demonstrates ambiguity of severity predictions produced different machine learning algorithms. Note that the Random Forest makes different predictions even with the same training set because of the elements of randomness in the model. When different bootstrap samples are used to fit model, the range of possible forecasts becomes even higher for almost all patients. On the contrary, SVM predicts based on several basic observations. In the case when both bootstrap subsets contain the same basic observations (support vectors), the models trained on them give very close the acute pancreatitis severity estimations. The diversity of SVM forecasts is achieved by subsets that do not contain one or more support vectors.

The Figure 2 shows a comparison of predictions made by different algorithms for patients of the same class (severe acute pancreatitis):

- The predictions of algorithms can be inconsistent, as in the case of Figure 2.a. While Random Forest tends to determine the moderately severe acute pancreatitis, Linear Regression and Support Vector Machine predict a severe acute pancreatitis;

- The predictions of algorithms can be consistent, as in the case of Figure 2.b. This is observed for typical class members for whom the initial training set contains many similar patients.

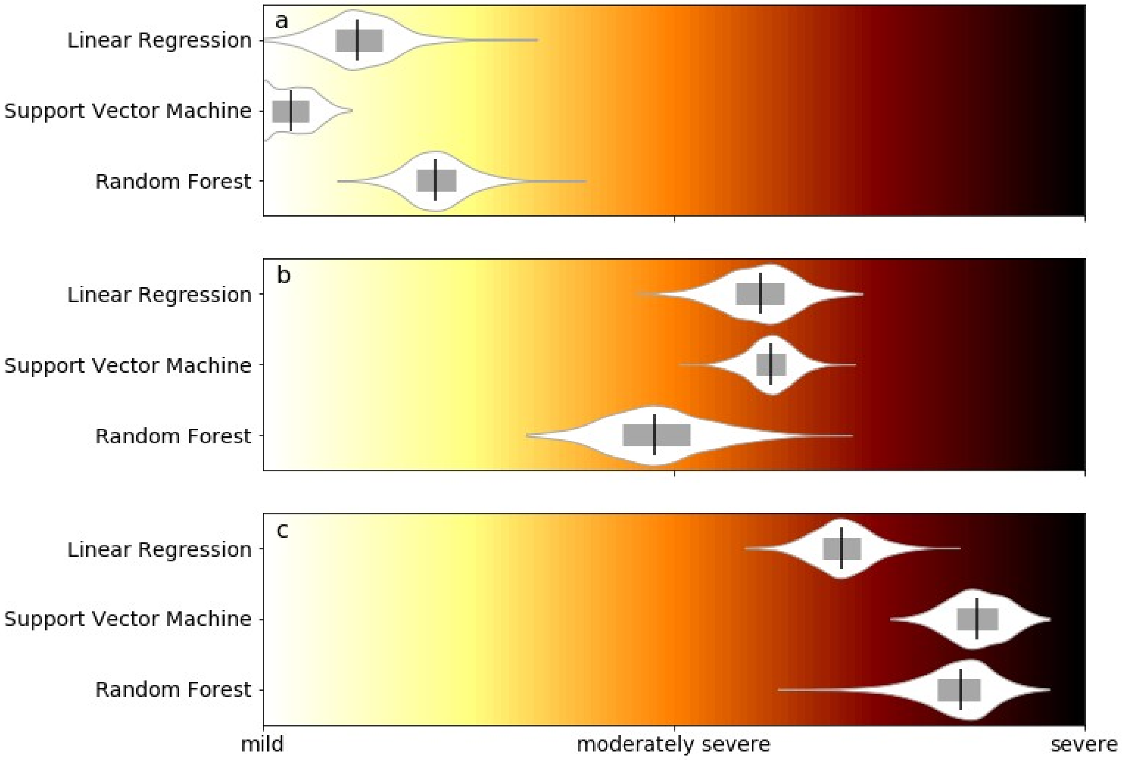

Figure 1: Violin plots based on various model predictions for typical observations from different classes: a - mild acute pancreatitis; b - moderately severe acute pancreatitis; c - severe acute pancreatitis

- The predictions of the algorithms can be incorrect, as in the case of Figure 2.c. Note the large scatter of the random forest predictions to the side of severe acute pancreatitis class that can be interpreted as classifier hesitation.

4.3 The effect of the bootstrap parameter p to the prediction diversity

The ratio between the number of unique observations in the bootstrap sample and the initial sample size p has an impact on predictions. The smaller the value of the parameter p, the smaller the subsets intersect and the greater the differences in the forecasts.

Figure 3 shows the effect of the parameter p on the prediction diversity by the example of one patient. If the parameter p is 0.95, the subsets differ by a maximum of 7 observations and the predictions of the class are compact on the numerical axis. If the parameter p is 0.9, the subsets differ by a maximum of 14 observations, medians change slightly, but the prediction diversity increases significantly for all models. And further, with a decrease in the parameter, this trend continues. When the parameter

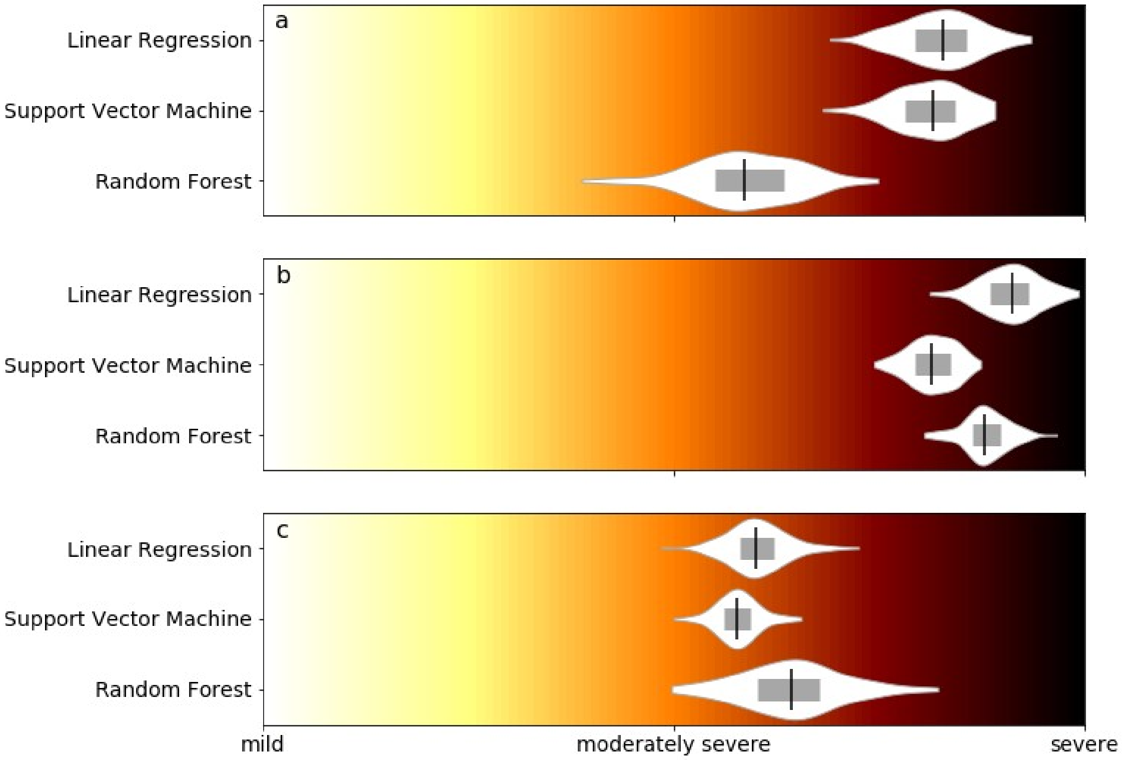

Figure 2: Violin plots based on various model predictions for patients with severe acute pancreatitis

p reaches 0.7, the linear regression and random forest predictions cover almost half of the numeric axis in the range $[1, 3]$.

Taking the final decision on the severity of acute pancreatitis, it is important to consider not only the average value of the forecasts, but also the variance of the forecasts.

Conclusions

Prediction stability visualization procedure was proposed and applied to estimation of acute pancreatitis severity. Visualization method allows to evaluate the prediction diversity of different machine learning algorithms for observation on a single graph. The study compared the stability of forecasts of Linear Regression, Support Vector Machine, Random Forest. This research can be useful to estimate the current dataset quality and to justify the need initial dataset increasing.

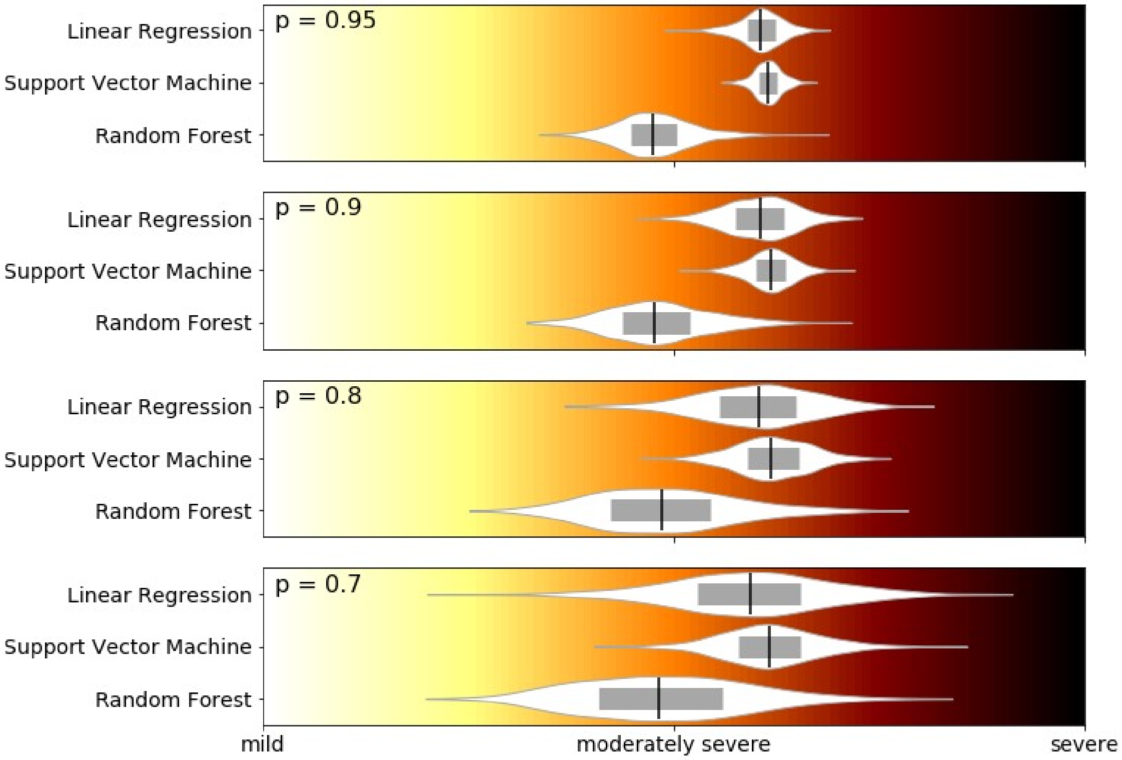

Figure 3: Violin plots based on various model predictions and influence of the ratio p between the number of unique observations in the bootstrap sample and the initial sample size on the stability of predictions

References

[1] Banks P.A., Bollen T.L., Dervenis C., et al. (2013). Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus. *Gut*. Vol. **62**, pp. 102-111.

[2] Efron B., Tibshirani R. J. (1994). An introduction to the bootstrap. – CRC press.

[3] Williamson D.F., Parker R.A., Kendrick J.S. (1989). The box plot: a simple visual method to interpret data. *Annals of internal medicine*. Vol. **110**, pp. 916-921.

[4] Hintze J. L., Nelson R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*. Vol. **52**, pp. 181-184.