# INFLUENCE OF THE TRAINING SET ON THE PREDICTION STABILITY IN ESTIMATION OF ACUTE PANCREATITIS SEVERITY

E. Mangalova, O. Chubarova, D. Melekh
*LLC RD Science, Siberian Federal University*
*Krasnoyarsk, RUSSIA*
e-mail: e.s.mangalova@hotmail.com

**Abstract**

The small sample size problem is often encountered in data analysis, especially for medical applications. It leads to unstable predictions when including or excluding several observations could change prediction significantly. Prediction stability visualization and measure were proposed and applied to estimation of acute pancreatitis severity. A simulation experiments were carried out to study the stability of ridge-regression, SVM, random forest trained with various subsets.
***Keywords:*** data science, small sample, prediction, acute pancreatitis severity

## 1 Introduction

Analysts in medicine face two contradictory problems due to the prohibition on disclosure and dissemination of personal data. Usually medical analysts deal with either large amounts of poorly matched data (health facilities have a different set of equipment with various accuracy, also medical institutions can depersonalize data in different ways) or small amounts of data (from one medical institution).

When training predictive models on a small data set is required, the analyst deals with the following challenges:

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

The work is devoted to the research of the influence of the training set on the prediction results. As example, acute pancreatitis severity classification task is considered.

## 2 Classification task

Acute pancreatitis severity is classified as mild, moderate or severe. Mild acute pancreatitis, the most common form, has no organ failure, local or systemic complications and usually resolves in the first week. Moderately severe acute pancreatitis is defined by the presence of transient organ failure, local complications or exacerbation of co-morbid disease. Severe acute pancreatitis is defined by persistent organ failure [1].

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 no. 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, $y$ is acute pancreatitis severity determined by medical expert based on patient clinical examination data defined by integer (1 - mild, 2 - moderate, 3 - severe).

The existing multi-class classification task can be transformed to binary classification task. One-vs.-rest strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. One-vs.-rest strategy requires the base classifiers to produce a real-valued class probability; class labels alone can lead to ambiguities, where multiple classes are predicted for a single observation.

Thus, there are two problems of binary classification:

- 1vsR: mild acute pancreatitis vs moderate and severe acute pancreatitis.
- 3vsR: severe acute pancreatitis vs mild and moderate acute pancreatitis.

The task 2vsR is excluded, since the moderate class is intermediate that requires the construction of a more complex separating surface, which is not desirable in small sample size conditions. For the study, three algorithms were chosen that allow the construction of simple separating surfaces: Ridge Regression, SVM and Random Forest.

# 3 Algorithm description

For $i$-th observation from the initial training set we estimate prediction stability of classification method using the following algorithm:

1. A set of $T$ training subsets is created such that each subset contains m different observations and do not contains $i$-th observation:

$$S_i = \{S_{i,1}, ..., S_{i,T}\}, S_{i,T} = ((\bar{x}_{j_{t,k}}, y_{j_{t,k}}), k = 1, ..., m : j_{t,k} \neq i), t = 1, ..., T. \quad (1)$$

   The $m/n$ ratio can range from 0.5 to $(n-2)/(n-1)$:

   - If $m/n$ is equal to $(n-2)/(n-1)$, we deal with some analogue of the leave-one-out cross-validation. Each pair of subsets is distinguished by one observation. This variant allows to show how one observation can change classifier prediction and identify specific observations that are similar to outliers.
   - If $m/n$ is smaller than $(n-2)/(n-1)$, the impact of sample size can be estimated. Changes in the classifier predictions trained on subsets with significant differences show how much information is contained in observations. The smaller the changes, the less information the observations contain. And the greater the changes, the greater the need to increase the training set.

It is also possible to vary the parameter $T$ (number of subsets):

- If $m/n$ is equal to $(n-2)/(n-1)$ training subsets contain (n-2) observations and only (n-2) different subsets can be formed. And accordingly, for small initial training set (number of observations $n$ allows to build $n^2$ classifiers in limited time) it is possible to form a complete set of subsets.
- If ratio $m/n$ is smaller then the number of possible variants becomes much larger (even for sufficiently small training set size). It means that the number $T$ should be limited to some reasonable value, and $T$ training samples for $i$-th observation should be chosen randomly. At the same time, we note that because the decision rule is tested for stability to a training set of observations, there is no need to ensure the preservation of the different classes objects proportion in the training subsets and the initial training set.

2. $T$ models $M_i, i = 1, 2, ..., T$ are built using the training subsets $S_i$ to obtain matrix of $T$ predictions $P_i = \{p^i_{t,z}, t = 1, ..., T, z = 1, ..., Z\}$, where $Z$ is the number of classes.

3. A convex hull of a set $P_i$ of points is constructed according to the predictions of the classifiers $M_i$.

If the problem of binary classification is solved, then there is a segment $H_i$ containing all predictions of classifiers for the $i$-th observation. The beginning of the segment is the minimum prediction, the end of the segment is the maximum prediction $H_i = [a_i, b_i] = [\min P_i, \max P_i]$.

If the problem of multiclass classification is solved, then there is such a convex hull $H_i$ containing all predictions of classifiers (rows of the matrix $P_i$). In mathematics, the convex hull of a set $P_i$ of points in the Euclidean space is the smallest convex set that contains $P_i$. Computing the convex hull means constructing an unambiguous, efficient representation of the required convex shape.

Chan's algorithm [2] is an optimal output-sensitive algorithm to compute the convex hull of a set $P_i$ of $T$ points in two- and three-dimensional space. The algorithm takes $O(T \log h)$ time, where $h$ is the number of vertices of the output (the convex hull). In the planar case, Chan's algorithm combines Graham scan algorithm with time complexity $O(T \log T)$ with Jarvis march algorithm with time complexity $O(Th)$, in order to obtain an optimal $O(T \log h)$ time.

The convex hull allows to display on a two-dimensional graph (for three classes) all possible classifier predictions based on different training subsets. Figure 1 illustrates the stability of various classifiers predictions (Ridge Regression, Support Vector Machine, Random Forest) for new observation. For this observation a set of $T = 500$ training subsets ($n = 130$, $m = 117$) was generated to fit classifiers. All three machine learning algorithms do not classify the patient as a severe acute pancreatitis, but there is ambiguity regarding classification as mild acute pancreatitis. Random forest estimates probability of mild class in the range $[0.4, 0.7]$, it is significantly less than the predictions of the two other algorithms.

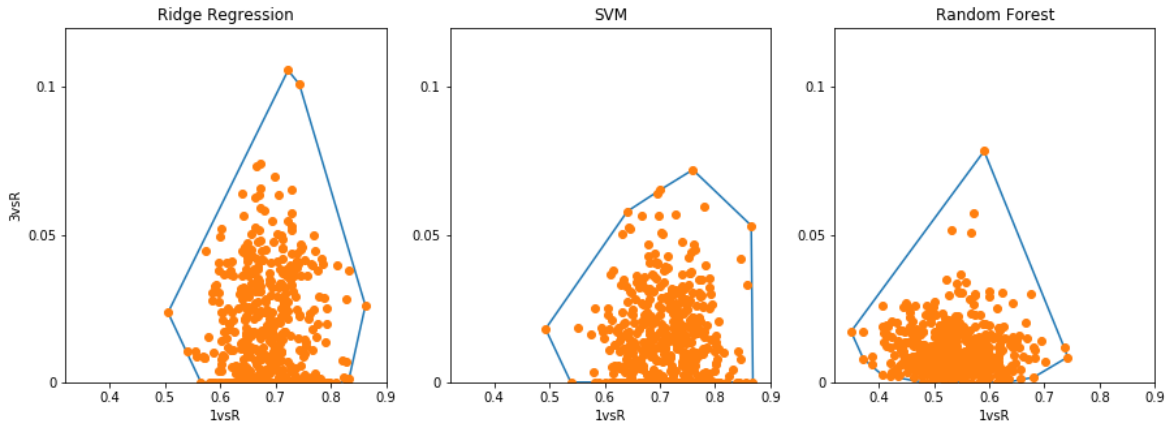


Figure 1: The stability of predictions for new observation: points are predictions of classifiers trained on different subsets of the training data, polygon is the convex hull constructed by these points

# 4 Experimental results

The proposed visualization algorithm allows to view the spread of predictions for multiple observations on a single graph. Figure 2 shows the stability of classifiers predictions for set of observations. In general random forest turns out to be a less stable algorithm, in other words, the convex hull area is larger for most observations. At the same time there is more compact area of observations with severe acute pancreatitis than in case of Ridge Regression and SVM. The wide scatter of the some predictions for the Ridge Regression and Random Forest indicates the presence of outliers. Note that the predictions scatter for patients with severe acute pancreatitis is higher than in others cases.

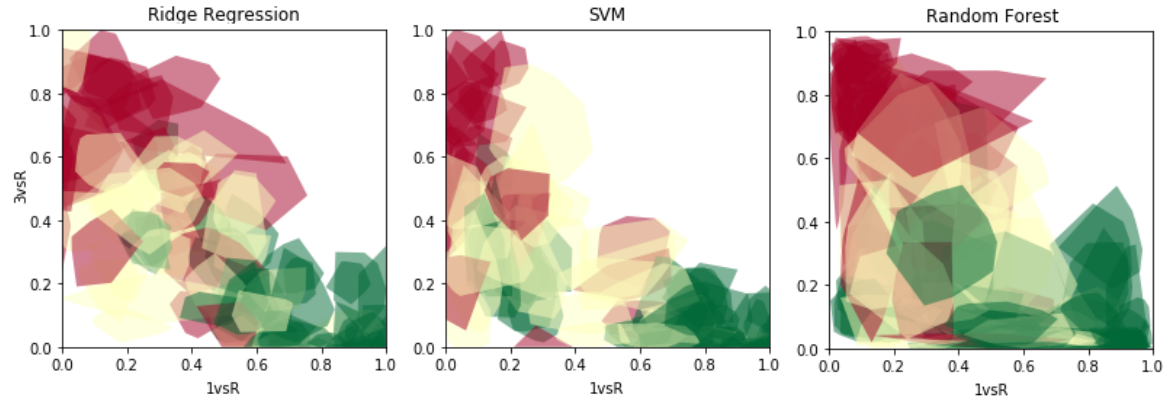


Figure 2: The stability of predictions for set of observations. The severity estimated by doctors are marked by the following colors: mild acute pancreatitis (green), moderate acute pancreatitis (yellow), severe acute pancreatitis (red)

# 5 Conclusion

Prediction stability visualization and measure were proposed and applied to estimation of acute pancreatitis severity. Visualization allows to evaluate the spread of predictions for multiple observations on a single graph and compare various machine learning algorithms. This study can be useful to estimate the current dataset quality and to justify the need dataset increasing.

Also the study shows the need for a combination of several algorithms for the final forecast because different methods have their advantages and disadvantages for different observations from various classes.

# 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] Chan T.M. (1996). Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete & Computational Geometry*. Vol. **16**, pp. 361–368.