Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210204

Clustering Results Interpretation of Continuous Variables Using Bayesian Inference

Ksenia BALABAEVA[a,1] and Sergey KOVALCHUK[a]
[a] *ITMO University, Saint-Petersburg, Russia*

Abstract. The present study is devoted to interpretable artificial intelligence in medicine. In our previous work we proposed an approach to clustering results interpretation based on Bayesian Inference. As an application case we used clinical pathways clustering explanation. However, the approach was limited by working for only binary features. In this work, we expand the functionality of the method and adapt it for modelling posterior distributions of continuous features. To solve the task, we apply BEST algorithm to provide Bayesian t-testing and use NUTS algorithm for posterior sampling. The general results of both binary and continuous interpretation provided by the algorithm have been compared with the interpretation of two medical experts.

Keywords. explainable artificial intelligence, interpretable machine learning, clustering interpretation, Bayesian inference, clinical pathways, XAI, eXAI, NUTS, BEST, K-Means, posterior sampling

1. Introduction

With the raise of machine learning (ML) application to the real-world problems the issue of trust to the modelling results is growing rapidly. The more human depend on such solutions, the more urgent it become to understand the principles of modeling and a solid argumentation for predictions. Along with the accuracy and inference time of ML models, interpretation has become the third criteria of ML models evaluation.

The importance of interpretation is great, especially in the fields directly related with people's health and lives. The users of decision support systems based on the black box models require the explanation and logic that lies behind each prognosis. It is clear that each person interacting with a black-box model may benefit from interpretation. However, each potential user plays a specific role, has different goals, background, domain knowledge. Therefore, each potential user has different explanation requirements [1–3].

Therefore, the form of interpretation may differ. Sometimes it could be a list of features used for modelling, ranged according to their significance [4]. Other examples use graphical representation in the form of diagrams [5, 6] or pixel heatmaps on the processed image [7]. Text explanation could also be used, for instance in image

[1] Corresponding Author, ITMO University, Kronverkskyi prospect, 49, Saint Petersburg, Russia; E-mail: kyubalabaeva@gmail.com.

processing using deep learning. Regardless the form, interpretation, in broader sense, is something that provides useful insights into the intrinsic work of the model or its results.

The big amount of interpretation approaches can be classified as model-specific [8] and model agnostic methods [9]. Another classification divides them into intrinsic [8] and post-hoc [4, 10].

Our goal is to automatically explain clinical pathways clustering results, described in the work of Funkner et al [11]. The objective of the authors was to predict which pathway a patient would follow during hospitalization. The data sample includes acute coronary syndrome (ACS) patients. Each patient is represented by a sequence of hospital departments. This data was collected from hospital EHRs: the logs of clinical events. A part of the solution was a clustering method K-Means.

In our previous paper we proposed an approach to explain clustering results automatically [12]. In terms of aforementioned classification, the proposed approach is post-hoc, since we analyze only model's output, and model-agnostic, since it could be used for any clustering model.

However, it had a limitation, taking only binary features for processing. In the present paper we continue to develop the approach of interpretation. Here we are expanding the algorithms functionality to work with continuous features used for explanation and determine the way of posterior sampling for such data.

2. Methods

Let $x_{n_i c_k}$ be a number of successes a person belongs to cluster c_i with n_i being a corresponding total number of observations for cluster c_i and p_{n_j,c_i} is the probability a patient belongs to cluster c_i.

Our approach is based on Bayesian inference [13]. The algorithm consists of three stages: posterior sampling, comparison matrix calculation and identification of features typical for each cluster by comparing the sampled distributions [12].

2.1 Explanation algorithm

Step 1. Posterior Sampling

For each feature f_j , $j \in [0, m]$:

For each cluster c_i , $i \in [0, k]$:

1. Determine the priors for $\mathrm{x_{n_i C_l}}$ and $P(A)_{f_j c_l}$;
2. Calculate the posterior distribution $P(A/D)_{f_j c_l}$;
3. Sample W new observations from the posterior $P(A/D)_{f_j c_l}$;

Output: vector of sampled probabilities $p_{\mathrm{f_j c_l}}$.

Step 2. Comparison matrix

Let I be a 2D matrix with the number of rows and columns equal to the number of clusters. The value of each matrix element $i_{c_l c_{l+1}}$ is equal to the mean value of sampled probabilities comparison. The calculation depends on the hypothesis we want to check. Each hypothesis is described in chapter 3.2.

Step 3. Identification of features more typical for a cluster

Output*:* dictionary with keys equal to cluster numbers and values equal to array of specific features for this cluster [12].

2.2 Explanation algorithm modification

In order to adapt this procedure to continuous variables, we had to modify the first step of posterior sampling.

In case of binary distributions, we assigned a beta distribution as prior $Beta(\alpha, \beta)$ and $Binomial(n, p_{n_i,c_l})$ to observations. Since both distributions have conjugate relations, we could calculate it as follows and eliminate MCMC modeling. However, for continuous variables such as time, or blood test parameters we should select different priors and posterior modeling methods.

To solve this task, we applied BEST algorithm [14] to *Step 1*. This algorithm allows the comparison of two or more groups and replicates statistical T-test. BEST helps not only to tell if the groups differ, but also to estimate how different are two samples, which is more informative. Another advantage is that BEST algorithm also allows to estimate the uncertainty related with the model parameters due to the intrinsic stochasticity of the system.

In order to perform the comparison, we have to build a probabilistic model. In our experiments we tested the difference between age of patients and the period of time they spent in the hospital before operation. For age we chose the Student-t distribution (1) as prior, since it is more robust to the outliers than Normal distribution.

$$f(x, \mu, \lambda, \nu) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)}\left(\frac{\lambda}{\pi\upsilon}\right)^{\frac{1}{2}}\left[1 + \frac{\lambda(x-\mu)^2}{\upsilon}\right]^{-\frac{\upsilon+1}{2}} \quad (1)$$

There are three parameters in Student-t distribution: μ - mean, λ – inverse variance, ν – the degrees of freedom parameter, responsible for the measure of "normality" in data. In probabilistic model we used normal distribution for μ as prior and uniform distribution for λ , and exponential for ν.

For the time feature we used gamma distribution as prior with α as a shape parameter and β as an inversed scale parameter.

$$f(x, \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad (2)$$

As a sampling method (*Step 1.3)* for both tasks, we applied NUTS algorithm, which is an MCMC algorithm, that allows to converge more quickly than Gibs or Metropolis sampling [15]

3. Results

Considering the posterior modelling of age and time before operation we have to conclude that the probabilistic models converged. However, the significant difference of mean and standard deviations based on these features wasn't found.

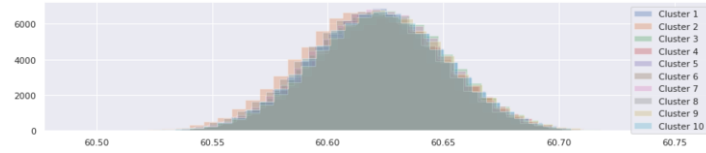

Figure 1. Samples from Posterior Distribution of μ for Age Feature

We would also like to introduce the updated results where we compared the clustering results interpretation of two medical experts (D1 and D2) with algorithms' interpretation (A). Each of the medical experts provided a set of features which he or she connects with patients in a particular cluster during analysis. Based on that we have calculated the DICE coefficient and measure the extent to which opinions coincide. The higher the metric – the higher is the concordance of opinions.

Table 1. Comparison of algorithms' and human's interpretations

Cluster	D1	D2	A	D1vsA	D2vsA	D1vsD2
1	rehabilitation, rehospitalization, additional surgeries	surgery, coronarography, cardio_department, other_department, comorbidity, icu	'coronarography', 'icu', 'rehabilitation', 'stenting'	0.22	0.44	0
3	outcome death, comorbidity, coronarography, no surgery, revascularization,	no_surgery, outcome_death, cardio_department, icu	'death', 'icu'	0.3	0.66	0.22
8	serious_condition, transfer_from_stationar, emergency_operation, stroke, clinical_death, cardiac_shock	surgery, outcome_death, cardio_department, icu	'death', 'stenting'	0.43	0.33	0
9	outcome death, coronarography, no_operation	no_surgery, coronarography, outcome_death, icu	'death', 'coronarography', 'icu'	0.35	0.86	0.57
MEAN DICE on 10 Clusters	-	-	-	**0.185**	**0.438**	**0.155**

All in all, the proposed approach could be applied to make clustering more transparent, informative and efficient. Based on such interpretation, we can automatically build a typical patient portrait in a certain cluster to use the results more proficiently (Figure 2). We believe that such interpretation based on data exploration and statistical inference can help in providing better patients treatment.

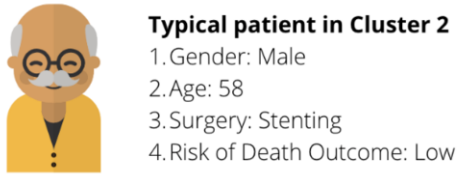

Figure 2. Portrait of a typical patient in cluster 2

4. Conclusions

Taking everything into consideration, we would like to say that the proposed objective is reached, and the functionality of continuous variables processing is developed. The proposed approach may be applied to explain the output of any clustering algorithm trained on the variables with categorial or continuous distributions. Moreover, it may be built into the decision support systems based on clustering algorithms to explain the results of the model to a doctor and make the interface of such systems more user-friendly. Such improvements contribute to the higher trust between models and users and reveal new opportunities for validation [16].

Concerning the future works, we are planning to test this methodology on different clustering tasks, using various data and clustering methods.

Acknowledgement

This work was supported by the Ministry of Science and Higher Education of Russian Federation, goszadanie no. 2019-1339

References

[1] Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. arXiv. 2018. p. 1811
[2] Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552. 2018.
[3] Amarasinghe K, Rodolfa K, Lamba H, Ghani R. Explainable machine learning for public policy: Use cases, gaps, and research directions. arXiv preprint arXiv: 2010.14374. 2020.
[4] Sundararajan M, Najmi A. The many Shapley values for model explanation. arXiv preprint arXiv:1908.08474. 2019.
[5] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann. Statist. 2001;29(5):1189-1232.
[6] Apley DW. Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468. 2016.
[7] Olah C, Mordvintsev A, Schubert L. Feature Visualization. Distill. 2. 10.23915/distill.00007. 2017.
[8] Patel H, Prajapati P. Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 2018;6:74-78.
[9] Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2020; URL: https://christophm.github.io/interpretable-ml-book/index.html (Last update: 02.02.2021).
[10] Balabaeva K, Kovalchuk S. Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients. Procedia Computer Science. 2019;156:87-96.
[11] Funkner A, Yakovlev A, Kovalchuk S. Data-driven modeling of clinical pathways using electronic health records. Procedia Computer Science. 2017;121:835-842.
[12] Balabaeva K, Kovalchuk S. Post-hoc Interpretation of Clinical Pathways Clustering using Bayesian Inference. Procedia Computer Science. 2020;178:264-273.
[13] Davidson-Pilon C. Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference. Addison-Wesley. 2019.
[14] Kruschke JK. Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General. 2013;142(2):573-603.
[15] Hoffman MD. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 2014;15:1593-1623.
[16] Kovalchuk SV, Kopanitsa GD, Derevitskii IV, Savitskaya DA. Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. arXiv preprint 2020; arxiv2007.12870.