pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210580

Statistical Inference for Clustering Results Interpretation in Clinical Practice

Alexander KANONIROV[a,1], Ksenia BALABAEVA[a], Sergey KOVALCHUK[a,b]
[a] *ITMO University, Saint Petersburg, Russia*
[b] *Almazov National Medical Research Center, Saint Petersburg, Russia*

Abstract. The relevance of this study lies in improvement of machine learning models understanding. We present a method for interpreting clustering results and apply it to the case of clinical pathways modeling. This method is based on statistical inference and allows to get the description of the clusters, determining the influence of a particular feature on the difference between them. Based on the proposed approach, it is possible to determine the characteristic features for each cluster. Finally, we compare the method with the Bayesian inference explanation and with the interpretation of medical experts [1].

Keywords. Explainable artificial intelligence, interpretable machine learning, clustering interpretation, statistical inference, clinical pathways, k-means

1. Introduction

Machine learning is at the center of numerous domains in science and innovation. More and more human lives depend on their decisions. In placing so much responsibility on algorithms, we need to have complete confidence in how they work. "The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." [2] Determining trust in individual predictions is an important problem when the model is used for decision-making. For instance, when using machine learning for medical diagnosis, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

The awareness of this problem has led to a rapid increase in the number of scientific papers on the interpretation of machine learning algorithms. For example, Tim Miller [3] gives the following definition of interpretability: «Interpretability is the degree to which a human can understand the cause of a decision» or «Interpretability is the degree to which a human can consistently predict the model's result». The higher the interpretability of a model, the easier it is for someone to comprehend why certain decisions were made. A model has better interpretability than another model if its decisions are easier to comprehend for a human than decisions from the second model.

[1] Corresponding Author: Alexander Kanonirov, ITMO University, Birzhevaya Liniya, 4, Saint Petersburg, Russian Federation; E-mail: alexander.kanonirov@yandex.ru

There's a big number of works dedicated to interpretation in the tasks of classification and regression [5-9]. However, only several works are on the topic of clustering results interpretation [10,11]. Paper [10] uses dendrograms and works only with hierarchical clustering and [11] suggests three-step clustering frameworks which work only for single feature clustering. The fundamental drawback of present methods is that they are model-specific and can only be used to explain a single model.

2. Materials and Methods

This work develops the ideas of interpreting clustering results approach described in articles [1, 4]. This article uses Bayesian inference for the post-hoc interpretation of clustering provided by the K-Means algorithm. Differences and similarities between clusters are investigated by sampling and comparing the posterior distribution of features.

In the presented work, we propose an alternative way to compare and explain clusters based on statistical inference. As an input for interpretation, we get the matrix X (n × m) with n observations and m features and corresponding cluster labels – y, that we get from clustering modeling using K-means. In order to describe the clusters, we compare the distributions of m features between clusters by statistical hypotheses testing. See the pseudocode of the procedure below.

Various statistical tests exist for formally testing statistical hypotheses. To select the proper test, the algorithm divides the data into categorical and continuous. If the data is continuous, then the algorithm checks if the data obeys normal distribution law. After that, a user has to select whether the compared groups are independent and how many groups are compared. Based on all the above, the algorithm decides which test to use to interpret continuous data (see Figures 1, 2).

Let Continuous be a set of continuous features and Categorical be a set of categorical features, where m – the number of features.

```
FOR each feature $f_i, i \in [0, m]$:
    IF $f_i \in$ Continuous THEN
      1. Check whether $f_i$ obeys the normal distribution law
         IF $f_i$ is normally distributed THEN
         1. Asking the user if the groups are dependent (YES/NO)
            IF answer = YES THEN
               IF groups > 2 THEN
               1. Apply ANOVA for Repeated Observations
               ELSE
               1. Apply Student's t-test for paired samples
            ELSE
            1. Asking the user if a comparison with a given value (YES/NO)
               IF answer = YES THEN
               1. INPUT value
               2. Apply Z-test
               ELSE
                  IF groups > 2 THEN
                  1. Apply One-way ANOVA
                  ELSE
                  1. Apply Student's t-test for independent samples
```

Figure 1. Pseudocode for comparing procedure using normally-distributed continuous features.

Similarly, the Figures 1-2 show the algorithm if the features have an abnormal distribution or are categorical.

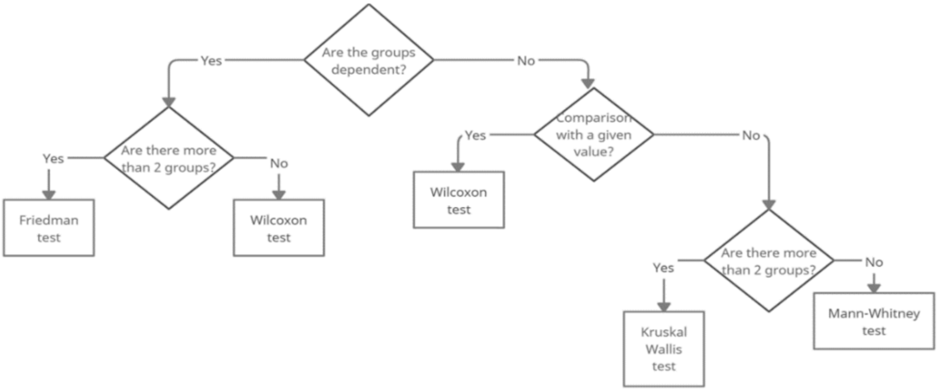
Are the groups dependent?
Yes
No
Are there more than 2 groups?
Yes
No
Friedman test
Wilcoxon test
Comparison with a given value?
Yes
No
Wilcoxon test
Are there more than 2 groups?
Yes
No
Kruskal Wallis test
Mann-Whitney test

Figure 2. An interpretation algorithm based on statistical inference for not normally distributed continuous variables

If the data is categorical, then the algorithm only needs to check whether the compared groups are independent and how many groups are compared. Based on this, it is decided which test to use to interpret categorical data (see Figure 3). After the test is selected and applied to the data, the algorithm displays in a form understandable for a medical expert, which features have a significant effect on the difference between the compared groups, and also indicates the features characteristic for each group. We will talk in more detail about what the data that the algorithm outputs and their interpretations mean in the next chapter.

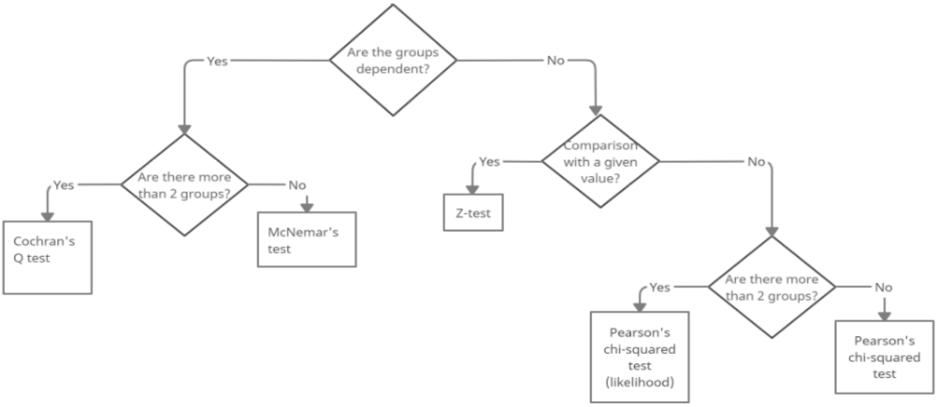
Are the groups dependent?
Yes
No
Are there more than 2 groups?
Yes
No
Cochran's Q test
McNemar's test
Comparison with a given value?
Yes
No
Z-test
Are there more than 2 groups?
Yes
No
Pearson's chi-squared test (likelihood)
Pearson's chi-squared test

Figure 3. An interpretation algorithm based on statistical inference for categorical variables

3. Results and Discussion

The dataset consists of 3312 observations. By observation, we understand a clinical episode, a single hospitalization of a patient with a Diagnosis of Acute Coronary Syndrome. The initial feature set included the clinical pathways - a sequence of departments a patient passes during hospitalization. In order to improve the interpretation, the feature set was extended with additional information from the electronic health records. All features are boolean. For instance, surgery, death outcome, stroke, stenting, coronarography, rehabilitation, clinical death, cardiogenic shock, rehospitalization.

The data presented were interpreted using Bayesian inference and a medical expert in [1]. Next, we compare these results with the results obtained by the algorithm described in the previous chapter. As a comparison metric, we use the percent of a coincidence for each cluster, calculated as a proportion of a number of features in the intersection to the number of features provided by the doctor or Bayesian inference. The main and most interesting results of the experiment are presented in Table 1.

Table 1 The results of the statistical inference interpretation algorithm and its percentage of agreement with the results of the Bayesian interpretation algorithm and the results of the interpretation of the medical expert.

Cluster	Num Observations	Bayesian inference (BI)	Statistical inference	Doctor’s Interpretation (DI)	% of explanation match with BI	% of explanation match with DI
1	116	'icu', 'rehospitalization', 'stenting', 'nevrology_dep', 'delayed_surgery', 'additional_surgeries'	'age', 'minutes_before_first_operation', 'stenting', 'serious_condition', 'icu', 'revascularization', 'rehospitalization', 'nevrology_dep'	'rehabilitation', 'rehospitalization', 'additional surgeries'	66.7	66.7
2	821	'outcome better', 'stenting', 'stroke', 'no surgery'	'num_operations', 'age', 'minutes_before_first_operation', 'no_surgery', 'outcome better', 'rehospitalization', 'revascularization'	'optimal path', 'transfer_from_stationar', 'rehospitalization', 'revascularization'	25	50
3	460	'outcome death', 'transfer_from_inhospital, 'no surgery’	'outcome_death', 'revascularization', 'no_surgery'	'outcome death', 'comorbidity', 'coronarography', 'no surgery, 'revascularization'	66.7	60
4	193	'cardiogenic shock', 'coronarography', 'rehospitalization', 'wheelchair_transporting', 'nevrology_dep_rehosp', 'vessel_surgery_dep', 'more_surgeries	'num_operations', 'minutes_before_first_operation', 'no_surgery'	'no surgery'	0	100
9	287	'outcome death', 'cardiogenicshock', 'seriouscondition', 'vessel_surgery_dep', 'transfer_from_stationar', 'no_surgery'	'age', 'operation', 'coronarography', 'clinical_death', 'vessel_surgery_dep', 'no_surgery'	'outcome death', 'coronarography', 'no_surgery	50	100
10	203	'outcome better', 'coronarography', 'urgent_operation', 'rehabilitation', 'cardio_dep', 'vessel_surgery_small', 'optimal_path'	'outcome better', 'coronarography', 'rehabilitation', 'cardio_dep', 'vessel_surgery_small'	'rehabilitation'	71.4	100

Cluster 1. Both methods found that patients in the cluster were more likely to be rehospitalized and undergo additional surgery during treatment. However, the algorithm,

based on statistical inference, determined that these patients needed additional rehabilitation. All of the above was also confirmed by a medical expert.

Cluster 2. Both algorithms determine a positive outcome for patients without surgery. But the doctor claims that patients from this cluster follow the optimal clinical path, which is not supported by algorithms in any way.

Cluster 3. Both algorithms establish death without surgery as the characteristic outcome for this group of patients. There were also no significant differences found in other characteristics suggested by the medical expert.

Cluster 4. An algorithm based on statistical inference revealed the features characteristic of the cluster - the presence or absence of surgical intervention. An algorithm based on Bayesian inference has shown that patients can have more than one operation. According to the doctor, patients here tend to get conservative treatment, without surgery.

Cluster 5. Both algorithms identified patients with complications characteristic of this cluster, as well as a favorable outcome for them, which emphasizes the algorithm based on Bayesian inference, indicating the sign of the optimal clinical pathway. In this group, the surgeon admits patients with complications and the necessity of additional post-surgery recovery treatment in other departments.

Cluster 6. Algorithms have revealed that stenting surgery is typical for patients, but for some reason, it is postponed. The medical expert mentions a complex diagnostics process for patients in this group.

Cluster 7. Both approaches in interpretation revealed multiple complications and the death of patients. However, unlike the medical expert and Bayesian inference, the algorithm based on statistical inference did not reveal that postponed operations are a characteristic feature of this group.

Cluster 8. The main diagnosis for both interpretations is stroke, but the algorithm based on statistical inference also indicates some complications for patients in the form of cardiac shock. All of the above was also confirmed by a medical expert.

Cluster 9. Algorithms indicate problems with blood vessels in patients, as well as a possible death.

Cluster 10. Both approaches determined that patients in this group are undergoing rehabilitation, therefore, they are likely to have a favorable outcome.

4. Conclusion

In conclusion, we would like to say that the goal of the study has been achieved: an approach has been developed to interpret the clustering results using statistical inference. This method allows you to get an idea of the clusters by determining the influence of a particular feature on the difference between them, on the basis of which it is possible to determine the characteristic features for each cluster.

When comparing the two approaches, no significant contradictions were found in the interpretation. In some cases, both algorithms are also capable of complementing each other's work, which is confirmed on the basis of the conclusions of a medical expert.

These are the primary results of our experiments, but we can already talk about a fairly good accuracy and availability of interpretation, so we are confident that this work will serve well in the application of machine learning models in clinical practice and other fields.

Acknowledgment

This research is financially supported by The Russian Scientific Foundation, Agreement #17-15-01177.

References

[1] Balabaeva K, Kovalchuk S. Post-hoc Interpretation of clinical pathways clustering using Bayesian inference. *Procedia Computer Science*. **178** (2020), 264-273.

[2] Doshi-Velez F., and Been K. Towards A Rigorous Science of Interpretable Machine Learning. *no. Ml:* **1–13** (2017), http://arxiv.org/abs/1702.08608.

[3] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv Preprint* (2017), arXiv:1706.07269.

[4] Balabaeva K, Kovalchuk S. Clustering Results Interpretation of Continuous Variables Using Bayesian Inference. *Studies in Health Technology and Informatics*. **281** (2021), 477-481

[5] Molnar C, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. URL: https://christophm.github.io/interpretable-ml-book/index.html. (Last update: 21.05.2020)

[6] Ribeiro M, Singh S., Guestrin C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA:1135–1144. (2016).

[7] Lundberg S, Su-In L. A Unified Approach to Interpreting Model Predictions. *Neural Information Processing Systems Conference 2017*. (2017).

[8] Sundararajan M, Najmi A. The many Shapley values for model explanation. *arXiv Preprint* (2019), arXiv:1908.08474

[9] Wachter S, Mittelstadt B., Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. **31(2)** (2018)

[10] Randriamihamison N., Vialaneix N., Neuvial Pierre. Applicability and Interpretability of Hierarchical Agglomerative Clustering With or Without Contiguity Constraints. (2019), arXiv:1909.10923

[11] Yang C., Shi X et al. I Know You'll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application. (2018), 914-922.