METHOD OF FEATURES ANALYSIS ON TRANSITION DATA

Eduard Manziuk, Oleksander Barmak, 2022

1. Manziuk E., Barmak O., Krak I., Mazurets O., Pylypiak O. Method of features analysis on transition data. *The 2021 IEEE International Scientific and Technical Conference "Advanced Trends in Information Theory":* Proceedings (Kyiv (Ukraine), December 15–17, 2021). Pp. 272–277. URL: https://doi.org/10.1109/ATIT54053.2021.9678787.
2. Petrovych V., Kuznetsov V., Manziuk E., Krak I., Kasianiuk V., Barmak O., Kulias A. I. On development classification methods for hidden features separation in data. *CEUR-WS*. 2021. Vol. 3018. Pp. 25–31.
3. Barmak O., Krak Y., Manziuk E. Characteristics for choice of models in the ensembles classification *CEUR-WS*. 2018. Vol. 2139. Pp. 171–179.

Introduction

The classification of atypicality is complicated by the fact that a separate group of data is distinguished, which are considered atypical and should not belong to any class. For this purpose are being developed methods of group definition and classification with the use of ensembles.

The use of an ensemble of models allows compensating for inaccuracies and a more objective approach to obtaining information about data. The choice of features using model ensembles is a promising area.

At the same time, methods are being developed for using projections of mental models on the machine level to complement ensembles.

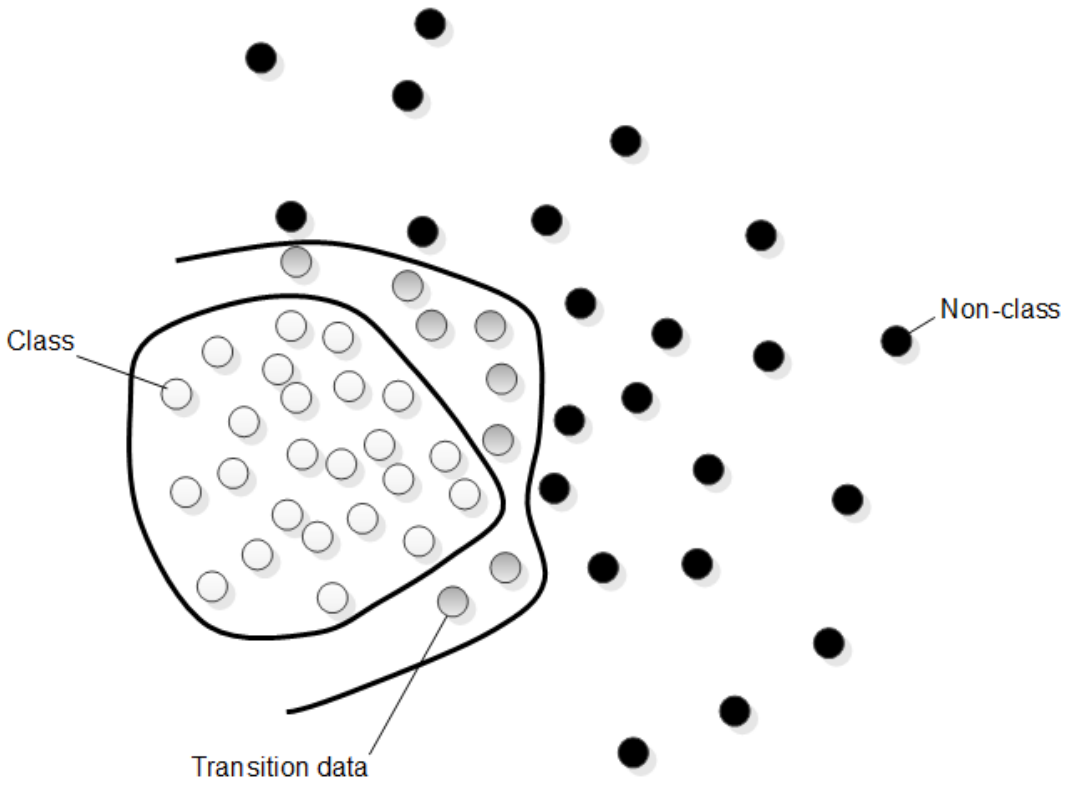
Purpose and objectives of the work

the purpose of the study is the development of a method for determining the features that most affect the local separability of transient data

Study objectives:

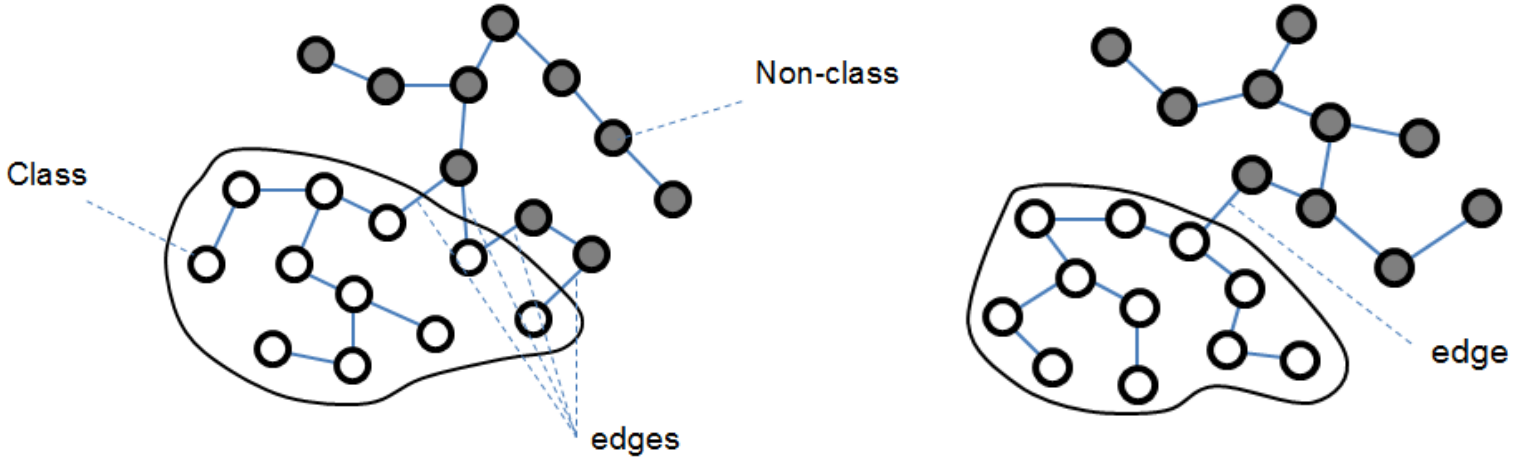
1. Creation of a model for selecting the attributes that most affect the separability of transient data.
2. Study the degree of influence of the features on the separability of data on the basis of the proposed method.
3. Development of a method of local compaction in the transient data area by minimizing the spatial displacement of all the data.
4. Development of a method for detecting outliers of features as the most influential on the local separability of data.

One-class classification in (transition data)



Example of a cluster in one-class classification and the presence of a transition data area.

Method of features analysis on transition data



a) b)

Transition zone data D_{Trans} with MST and marked edges that connect class data D_{Class} with data outside the class $D_{NonClass}$: a) data in the initial state with marked edges of the combination of objects D_{Class} and $D_{NonClass}$; b) transition zone data after the regularization of features and the increase of compactness of class data and the formation of ideal conditions for the spatial arrangement of class data, which correspond to the presence of one edge of the combination with data outside the class

Method of features analysis on transition data

The subgraph is formed from a minimal spanning tree and is built on the vertex of subsets $V = V_{Class} \cup V_{NonClass}$, where the total number of vertices is identical to the data of the transition zone. Since the data is a subset of vertices in general, we can say that a subset is formed from individual parts of MST, i.e. the forest of MST subgraphs.

$$FMST_{Class} = \bigcup subMST_{Class} \quad FMST_{NonClass} = \bigcup subMST_{NonClass}$$

$$FMST_{Class-NonClass} = \bigcup subMST_{Class-NonClass}$$

The ultimate goal and ideal case of regularization of the transition zone data is to increase the compactness of the class data, which ultimately corresponds to the merging of the forest into a single graph.

Method of features analysis on transition data

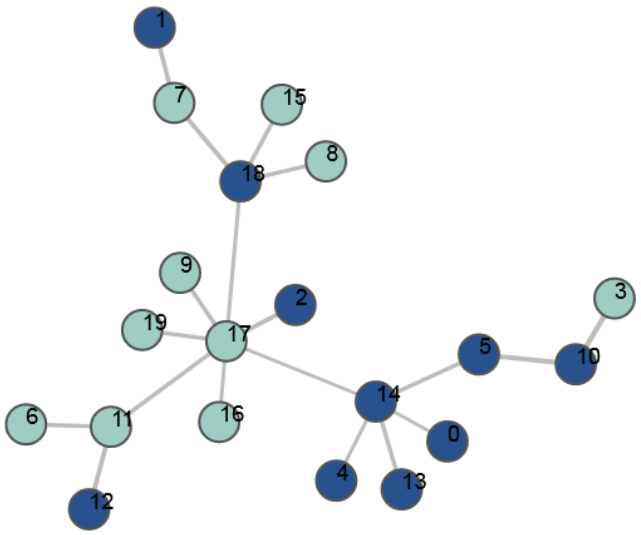
Consider the boundary conditions for data separation with their spatial bias and the use of MST.
The MST is built on all transient data. There are three types of subgraphs from MST. Subgraph with edges connecting vertices that are data belonging to the class $subMST_{Class}(V,E)=\left\{v\middle|\forall v\in D_{Class},\right\}$ V - set vertex, E - set edges. Subgraph with edges that combine data outside the class $subMST_{NonClass}(V,E)=\left\{v\middle|\forall v\in D_{NonClass}\right\}$. Subgraph with edges connecting vertices belonging to the data of these two previous sets

$$subMST_{Class-NonClass}(V,E)=\left\{e\middle|e_{x_i,x_j}\in E,x\in D_{Trans},i\in\left[1..\left|V_{Class}\right|\right],j\in\left[1..\left|V_{NonClass}\right|\right]\right\}$$

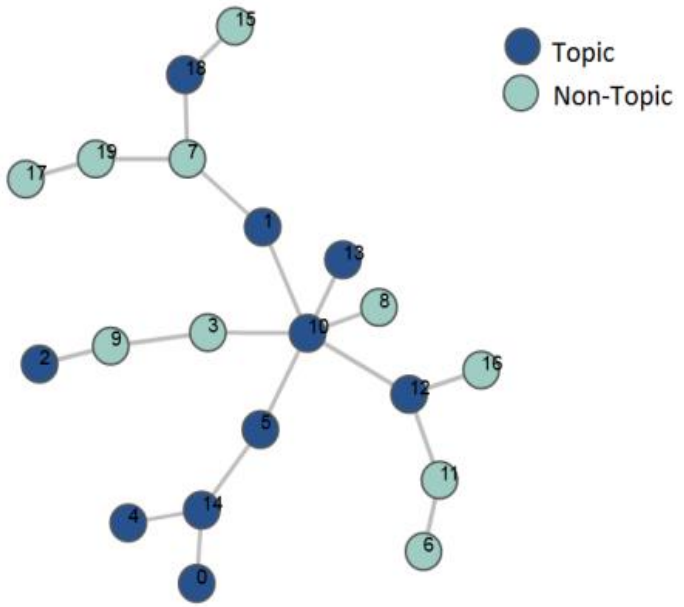
$$\lim_{\lambda\to a}\left(FMST_{Class-NonClass}\right)\to\exists!subMST_{Class-NonClass}$$

This λ is the regularization parameter, a - boundary value of the regularization parameter

Experimental studies



MST for the initial state of the transient data n = 20, constructed taking into account only the common features of the number 187. The number of edges of the bipartite graph 9, the number of all features 778

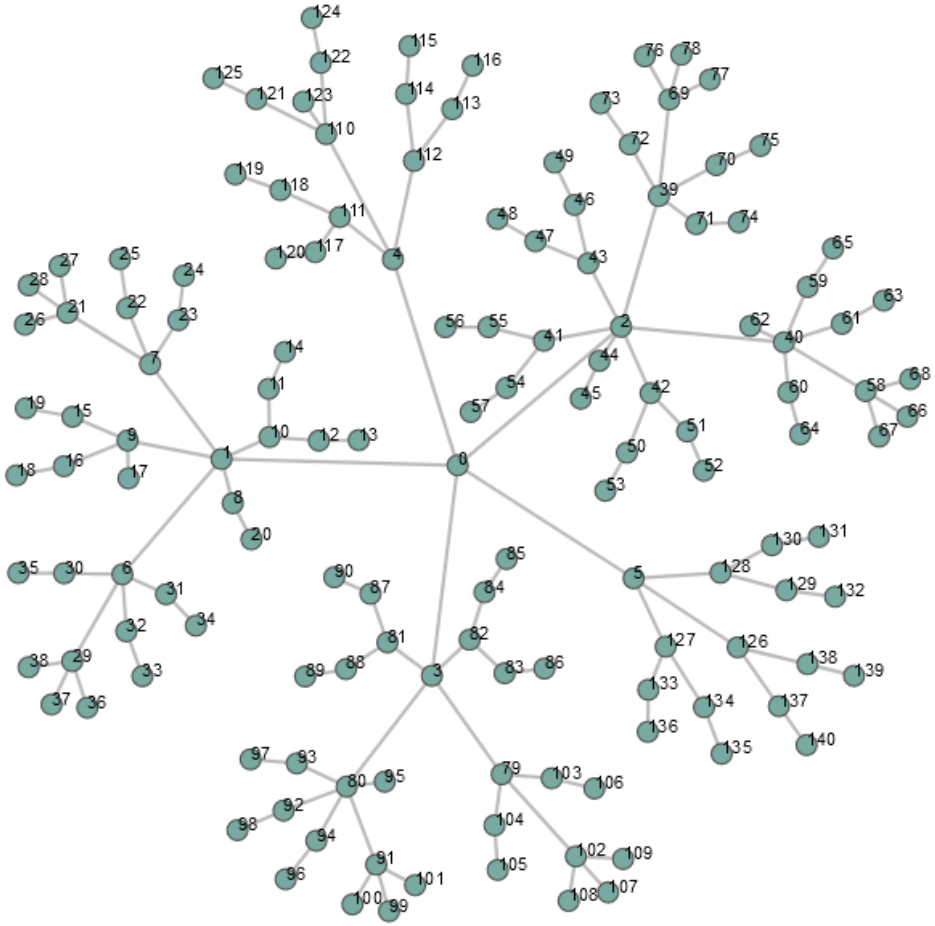


MST for the initial state of the transient data n = 20, the number of edges of the bipartite graph 8, the number of all features 778

Under this condition we obtain three sets
{372, 306, 522, 455}, {616, 306, 522, 455}, {455, 306, 522, 445}.

Counting the features for their presence in these sets.
522 -> 3 ,306-> 3 ,455 -> 3, 372 -> 1, 616 -> 1, 445 -> 1
Features number 522, 306, 455 are available in each trajectory which leads to the necessary minimization. This suggests that these features can be considered outlier.

Experimental studies



Graph of trajectories on a set of common features. A vertex marked "0" indicates the beginning of the forest graphs and is not a feature

CONCLUSIONS

The data of a class at one-class classification in a transition zone form more compact areas. Local delimitation by minimizing the number of edges of a bipartite graph determines the outliers of features as they affect the distinction between class data and data outside the class.

The proposed approach allows determining outlier features among the search sets. A local compaction method was developed on the transient domain with minimizing the spatial bias of all data.

Using the proposed approach allows detecting features set that negatively affect the differentiation of data and detect measure of influence and obtain their dimensional characteristics.