IntSol 2021: II International Scientific Symposium
"Intelligent Solutions", September 28–30, 2021, Kyiv, Ukraine

On development classification methods for hidden features separation in data

Valentina Petrovych[1], **Vladislav Kuznietsov**[1], Eduard Manziuk[3], Iurii Krak[1,2], Veda Kasianyuk[2], Alex Barmak[3], Anatolii Kulias[1]

Glushkov Institute of Cybernetics [1],
Taras Schevchenko National University of Kyiv [2],
Khmelnitsky National University [3]

IntSol 2021: II International Scientific Symposium
"Intelligent Solutions", September 28–30, 2021, Kyiv, Ukraine

Related work

- On Classification Hidden Concepts Language in Specialized Texts Based on Methods of the Intellectual Data Processing [Y.Krak[1,2], O.Barmak[3], A.Kulias [1], V.Petrovych[1], E.Manziuk[3], V.Kuznetsov[1],] 2021
- Visual Analytics to Build a Machine Learning Model [Y. Krak [1,2], O.Barmak [3], E.Manziuk [3]] 2021
- Classification Based Hierarchical Clustering Prediction Variability in the Ensembles of Models Using a Statistical Approach [Y. Krak [1,2], O.Barmak [3], E.Manziuk [3]] 2020
- Recognition of emotional expressions using the grouping crowdings of characteristic mimic states [Y. Krak [1,2], O. Barmak [3], E.Manziuk[3], O. Kalyta[3], V.Kuznetsov[1], A. Kulias [1]] 2020
- Multivariate Scaling of the Characteristic Features Based on Pseudo-Inverse Operations for Recognition Problems Solving [Y.Krak [1,2], O.Barmak[3], V.Kasianiuk[2], H.Kudin [1,2], E.Manziuk[3]] 2020

Afiiliation: Glushkov Institute of Cybernetics [1], Taras Schevchenko National University of Kyiv [2], Khmelnitsky national university [3]

Problems Statement

- to create a datasets on different topics;
- to propose a feature representation;
- to study feature vector representation;
- to analyze density in feature space of data;
- to discover hidden features' in data;
- to conduct tests on different algorithms;
- to estimate the algorithms' stability.

Dataset

- Requirements:
  - The data is available to read and parse;
  - Different datasets are available;
  - The classes are well balanced within dataset.

- Our datasets are based on different topics :
  - Miograms classification (muscular activity);
  - Natural Language Processing (scientific texts);
  - Sound processing (audible signals).

Dimensionality reduction and grouping of features

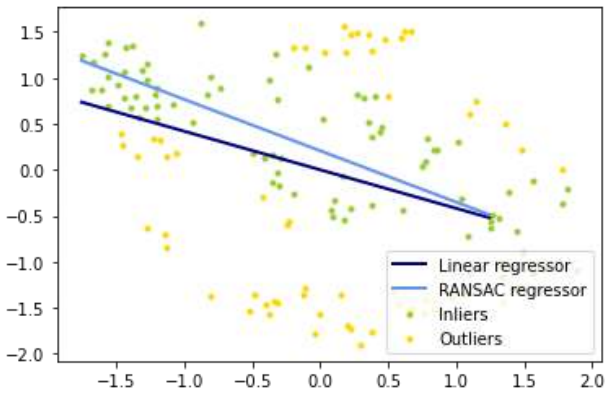
- Type of feature and vector representation:
  - Eigenvectors and eigenvalues;
  - T-stochastic grouping of features.
- Properties of the representation:
  - Grouping of data points in clusters by maximal like hood of data points and features;
  - Easy to visualize (2d diagram).

IntSol 2021: II International Scientific Symposium
"Intelligent Solutions", September 28–30, 2021, Kyiv, Ukraine

Anomalies detection

- Data points are located along 1 or more axes;
- Centroids show the areas of informational density;
- Method used: RANSAC and linear regression.

- Example of usage:



- Outlier points have a great distance from centroids of data clusters;
- Distance from centre of cluster for outliers show increase in variance and entropy;
- Big variance in data may shift decision boundary for specific class or dataset

IntSol 2021: II International Scientific Symposium
"Intelligent Solutions", September 28–30, 2021, Kyiv, Ukraine

Dataset balancing using augmentation

- Type of classifier: Bayesian classifier;
- Original dataset size: ~1500 data entries.

Table 1. Algorithm performance on original dataset

precision	recall	f1-score	support	
Class 1	0.59	1.00	0.74	366
Class 2	0.98	0.95	0.96	454
Class 3	0.90	0.64	0.75	332
Class 4	1.00	0.59	0.74	333
accuracy			0.81	1485
macro avg	0.87	0.79	0.80	1485
weighted avg	0.87	0.81	0.81	1485

Table 2. Algorithm performance on augmented dataset

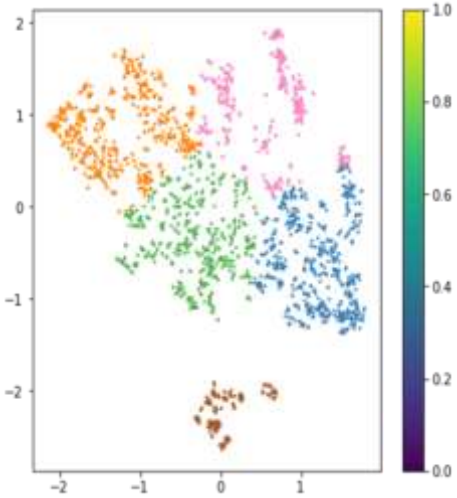
precision	recall	f1-score	support	
Class 1	0.81	0.97	0.88	1313
Class 2	0.92	0.96	0.94	1239
Class 3	0.95	0.76	0.85	1274
Class 4	0.98	0.93	0.96	1254
accuracy			0.91	5080
macro avg	0.92	0.91	0.91	5080
weighted avg	0.91	0.91	0.91	5080

IntSol 2021: II International Scientific Symposium
“Intelligent Solutions”, September 28–30, 2021, Kyiv, Ukraine

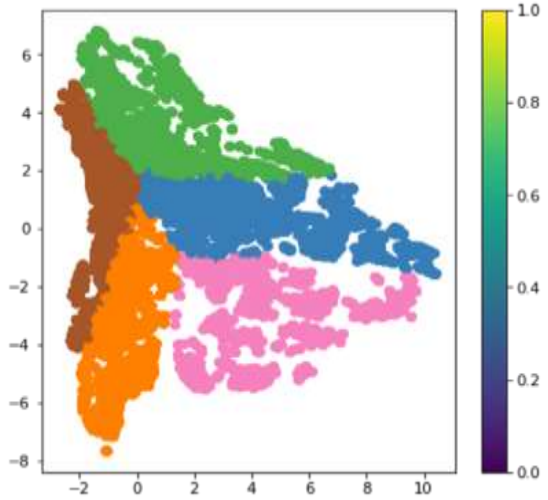
Testing algorithm stability

- Used method: variational autoencoder (VAE).
- Feature matrix representation before and after perturbation of data using VAE:

Before



After



Conclusions

- Classification of data using hidden features of the data and mean clustering is suitable for visual analysis and make decisions.
- Data scaling using hidden properties of the data, data augmentation and engineering gives better overall performance.
- The decision makings and hypotheses are stable, whether the data has the original dimension or some transformations applied.
- In further works, we plan to focus on the data preparation, feature extraction, visual analysis for each dataset presented in the paper as well its properties and structure.
- We will conduct a series of experiments using deep learning methods on this data, for instance, convolution neural networks and long-short term memory.

Thank you!