Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 931–934
DOI: 10.15439/2022F210
ISSN 2300-5963 ACSIS, Vol. 30

# Impact of clustering of unlabeled data on classification: case study in bipolar disorder

Olga Kamińska, Katarzyna Kaczmarek-Majer, Olgierd Hryniewicz
Systems Research Institute Polish Academy of Sciences
ul. Newelska 6, 04-710 Warsaw, Poland
Email: {okaminska, k.kaczmarek, o.hryniewicz}@ibspan.waw.pl

*Abstract*—**Currently, it is possible to collect a large amount of data from sensors. At the same time, data are often only partially labeled. For example, in the context of smartphone-based monitoring of mental state, there are much more data collected from smartphones than those collected from psychiatrists about the mental state. The approach presented in this paper is designed to examine if unlabeled data can improve the accuracy of classification tasks in the considered case study of classifying a patient's state. First, unlabeled data are represented by clusters membership through Fuzzy C-means algorithm which corresponds to the uncertainty of the patient's condition in this disease. Secondly, the classification is performed using two well-known algorithms, Random Forest and SVM. The obtained results indicate a minimal improvement in the quality of classification thanks to the use of membership in clusters. These results are promising due to both, the accuracy and interpretability.**

## I. Introduction

MOTIVATION for this research comes from the practical problem of classifying partially labeled data. Within this work, we concentrate on a particular case study in monitoring the mental state of bipolar disorder (BD) patients which has a large dataset of sensor-based data with labels provided by doctors. Since we are limited by medical labels, the most frequent attempts to predict a patient's condition come down to using only a small part of the data from the entire set. Such selected data may not contain characteristics for each patient and the obtained results may not be accurate. To alleviate the aforementioned problem, we propose to incorporate the results of clustering into the classification task.

The collection of medical data in our possession indicates to use of a semi-supervised approach. For this purpose, it is worth enabling clustering to extract information from unlabeled data [1]. In related work, the accuracy of classification for patients with bipolar disorder using sensor data amounts to 76% [2]. Some works test the influence of clusters in the classification problems, e.g., [3] and indicate an improvement in the results. Other works include unlabeled data by means of the dynamic incremental fuzzy semi-supervised learning, see e.g., [4].

Experiments are performed on data about voice collected from smartphones of bipolar disorder patients. On the other hand, labels obtained from psychiatrists during visits are valid only for daytime, and of that day only so the amount of those labels is relatively small. Psychiatrists indicate that the symptoms of a given phase are visible several days before the visit, therefore the validity of this label can be extended. That procedure results in the preparation of a much larger range of data enabling the classification of the patient's condition. Therefore, we check whether the use of the remaining unlabeled data represented by the membership to clusters improves the quality of the classification labeled data.

## II. Methodology

The idea of the proposed experiment is to verify that unlabeled data assign as a clusters membership has an impact on the classification of patients states.

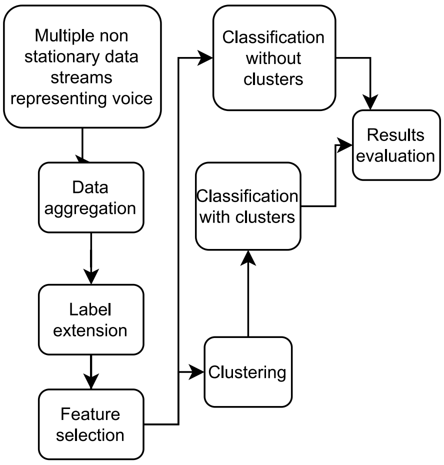
Multiple non stationary data streams representing voice
Data aggregation
Label extension
Feature selection
Classification without clusters
Clustering
Classification with clusters
Results evaluation

Fig. 1. Process flow

The process flow of the proposed experiment is presented in Figure 1 and in Algorithm 1. The experiment begins with retrieving all available multiple non-stationary data streams representing voice. All frames are then aggregated to the patient's phone call level with different aggregates methods. Psychiatric assessments obtained during visits, represented as labels are spread around the day of the patient's visit. Within the feature selection, the top-k most important voice parameters are selected for each patient and aggregated methods. Additionally, all available patient data are clustered to include unlabeled data as well. Simultaneously, the classification of the patient's condition is carried out on the data containing clusters membership and without this information. Finally, results are evaluated with multiple metrics.

**Algorithm 1** Pseudocode of the experiment

```
for patient in patients_list:
    for agg in aggregates:
        best_features = RFE(no_of_best_parameters = 10)
        cluster_mmbs = FuzzyCmeans(data[,best_features], unique_visits_no)
        rf_nocluster = RandomForest(data_without_clusters)
        rf_withcluster = RandomForest(data_with_clusters)
        SVM_nocluster = SVM(data_without_clusters)
        SVM_withcluster = SVM(data_with_clusters)
```

### A. Data collection and aggregation

Patients were enrolled and used a dedicated smartphone application in everyday life starting in September 2017 and ending in December 2018. All their recordings were divided into 10-20 ms frames. Next, for those frames, openSMILE [5] library was used to calculate acoustic features. The final dataset contains 86 data streams that describe the main acoustic features of voice such as e.g., loudness, voice energy, pitch, etc. All data from 2018 were selected for the experiment for each of the 6 patients who had several visits in the year of the study where different disease phases were observed.

Due to a large number of frames available for each connection, this data has been aggregated to the level of one phone call using mean, standard deviation, skewness and quartiles. Each of the available 85 voice parameters is aggregated in this way. The data were then normalized. Aggregating the data to the level of the conversation will allow you to slightly reduce the noise of the data and facilitate data processing in subsequent processes.

### B. Labels extension

The labels obtained by the patient during the visit are valid only on the day of the visit. The number of labels available for a given patient during the year did not exceed 7 for 1 patient, which is a negligible value throughout the year. The ability to extend the label to days around the visit increases the amount of labeled data. Other studies are considering extending the label to 7 days in advance as symptoms of the patient's condition may already be noticeable prior to the visit. In the present experiment, the label was extended 7 days before the visit and 2 days after the visit. This gives the label a validity period of 10 days. Results received from that method are shown in TABLE I. In an example of first patients with ID 1472 we can observe, that label extension increased the number of phone calls with labels from 42 to 391. There was much more unlabeled data, i.e. 1482, what is worth using it.

### C. Feature Selection

To obtain significant voice parameters we apply one of the automatic feature selection methods called Recursive Feature Elimination (RFE) [6]. The idea of the RFE technique is to build a model with all variables and after that, the algorithm successfully removes features until the desired number remains. The current implementation of that method used the Random Forest algorithm to train and create ranking features by importance, discarding the least important features, and re-fitting the model. To find the optimal number of features cross-validation is used with the RFE algorithm to obtain the best scoring collection of features. The final subset used the first 10 best parameters resulting from that method.

TABLE I
PATIENTS DETAILS

| ID patient | No. of visits | No of phone calls in day of visits | No of phone calls for extended label validity | No of all phone calls in 2018 |
|---|---|---|---|---|
| 1472 | 2 | 42 | 391 | 1482 |
| 2004 | 2 | 16 | 223 | 871 |
| 2582 | 3 | 31 | 169 | 645 |
| 5656 | 2 | 7 | 29 | 215 |
| 5736 | 2 | 20 | 71 | 1025 |
| 6139 | 3 | 22 | 90 | 254 |

Earlier studies [7] have shown that the introduction of the RFE method improves clustering results. In the present work, the RFE method is used for each patient and each data aggregation method separately considering only the labeled data. This allows the best voice parameters to be selected in a tailored way. This set of the 10 most important acoustic parameters is then used in the next stage of the experiment.

### D. Clustering

The algorithm used for clustering is Fuzzy C-mean [8] with squared Euclidean distances as a parameter of dissimilarities between observations. We assume a patient may have symptoms of several conditions at the same time during unlabeled days. This happens mainly in a mixed state where the symptoms of mania and depression occur simultaneously. Furthermore, on unlabeled days the patient may not have obvious symptoms characteristic of only one BD state. Such uncertainty resulted in the choice of the Fuzzy C-mean algorithm which introduces cluster membership. In that case, the number of clusters corresponds to the known number of different phases diagnosed in a given patient.

In fuzzy clustering, each observation is “spread out” over the various clusters. The output of that clustering is the membership to each of the clusters. The memberships are nonnegative, and for a fixed observation it sums to 1. So each phone call could be partly assigned to one class and partly to another class. We don't assign a specific cluster to BD state. We just looking for a variety between classes.

Clustering was performed on all available data for each patient (also unlabeled data) in order to capture the variability over all available observations. These data were aggregated to the level of the patient's conversation and then the 10 most important acoustic parameters were selected by the RFE method.

### E. Classification

The classification was made in 2 ways using 2 well-known classifiers, the Random Forest [9] and Support Vector Machine [10]. The first method (lines 5 and 7 in Algorithm 1) assumes that aggregated data with selected voice parameters by the RFE method are classified. The second method (lines 6 and 8 in Algorithm 1) additionally joined the membership of each cluster to that dataset. Clusters membership were used for labeled data only. Then the data is divided into a training set and a test set in the proportion of 75:25 in such a way that each set contains data from each BD patient's state. Predicted classes depend on how many different phases the patient received during the whole study. The classification is performed on each patient for each of the 6 available aggregation methods. The "best aggregate" is then selected according to the Accuracy comparison of each aggregate for that patient. The classification results for the selected aggregate are summed up from all patients and a common confusion matrix is created for the selected algorithm and data without clusters and with clusters. The values in the confusion matrix are presented for the test set not used during training.

### F. Evaluation metrics

In total 4 confusion matrices have been created. The first two matrices concern the comparison of values from all patients and their best aggregation methods for the Random Forest algorithm in the first case without the use of clusters and in the second additionally including cluster membership.

The next 2 matrices also compare the values without clusters and with clusters, this time using the SVM algorithm.

Additionally, for each of the above-mentioned matrices, classification coefficients were calculated, such as Accuracy - correctly forecasted patient states concerning all forecasts, Precision - i.e. the fraction of relevant instances among the retrieved instances, and Recall - i.e. the fraction of relevant instances that were retrieved.

## III. Experimental Results

### A. Selection of aggregation operators and acoustic features

Selecting an appropriate aggregation operator for the acoustic data is not obvious, therefore, several such methods were tested in this study

The best aggregates were selected separately for each patient. The results are presented in TABLE II. The best aggregating methods turned out to be the mean and skewness. They have been selected 7 times. Then Q1 was selected 5 times, Q3 - 3 times, and Standard deviation - 2 times. Interestingly, the Q2 aggregate known as the median was not selected even once. Moreover, the parameters that best characterize the normal distribution, i.e. mean, standard deviation, and skewness, were selected twice as often (16 hits) as the parameters characterizing the quantile distribution (8 hits). The differences are also noticeable concerning the classifiers used. However, it does not affect their further interpretability.

The relevant acoustic parameters received from RFE methods differ for each model as well. The following 10 parameters were most often selected by the model are: *spectralRollOff90*, *LOGenergy*, *mfcc 11*, *fband1000-4000*, *f3frequency*, *f2frequency*, *fband0-650*, *hammarbergindex*, *audSpec*, *spectralharmonicity*.

TABLE II
Aggregates Methods Selection

| | **RF**<br>**non-cluster** | **RF**<br>**with cluster** | **SVM**<br>**non-cluster** | **SVM**<br>**with cluster** |
|---|---|---|---|---|
| *aggregate* | *cardinality* | | | |
| **mean** | 1 | 1 | 2 | 3 |
| **standard deviation** | 1 | 0 | 1 | 0 |
| **skewness** | 2 | 2 | 2 | 1 |
| **Q1** | 1 | 2 | 1 | 1 |
| **Q2** | 0 | 0 | 0 | 0 |
| **Q3** | 1 | 1 | 0 | 1 |

### B. Classification

*1) Random Forest:* Table III contains the confusion matrices for the test sets from all patients where the data included in the Random Forest classifier did not include cluster membership (left) and where the data contained cluster membership (right). Values on the diagonal of the matrix indicate the correct classification of each of the states. The remaining values indicate what were the forecasts for the remaining cases where the observed label does not agree with the predicted value. The results are promising for both models. The total number of correctly classified for the non-clustering model is 209 and for the clustering model is 211, so we see a slight improvement in the results. The only place where clusters join in shows a slight weakening of the results is in the prediction of the state of depression. The other values are slightly better or equal. However, these differences are subtle, so it should be tested on more examples.

Table IV contains the results calculated based on the above-mentioned confusion matrices. As observed, the precision and recall values for each state are similar or slightly better for the method containing cluster memberships. Accuracy, i.e. the ratio of correctly predicted values to all values, also increases in the method containing clusters from 83.27% to 84.06%.

*2) SVM :* Table III contains the confusion matrix summed for the test set from all patients where the data included in the Support Vector Machine classifier did not include cluster membership (left) and where the data contained cluster membership (right). Results obtained for the SVM are similar to the previously considered RF. The total number of correctly classified for the non-clustering model is 203 and for the clustering model is 210, so there is again a slight improvement. In that case, there is no place where the method using clusters received a worse number of incorrect predicted values in any class.

TABLE III
CONFUSION MATRICES FOR RF AND SVM CLASSIFIERS WITHOUT (LEFT) AND WITH INFORMATION ABOUT MEMBERSHIP TO CLUSTERS.

| | | *actual* | | | |
|---|---|---|---|---|---|
| RF-nc | | **E** | **X** | **D** | **M** |
| *predicted* | **E** | 79 | 15 | 13 | 1 |
| | **X** | 4 | 76 | 0 | 0 |
| | **D** | 9 | 0 | 39 | 0 |
| | **M** | 0 | 0 | 0 | 15 |

| | | *actual* | | | |
|---|---|---|---|---|---|
| RF-c | | **E** | **X** | **D** | **M** |
| *predicted* | **E** | 83 | 11 | 13 | 1 |
| | **X** | 3 | 77 | 0 | 0 |
| | **D** | 12 | 0 | 36 | 0 |
| | **M** | 0 | 0 | 0 | 15 |

| | | *actual* | | | |
|---|---|---|---|---|---|
| SVM-nc | | **E** | **X** | **D** | **M** |
| *predicted* | **E** | 79 | 17 | 11 | 1 |
| | **X** | 2 | 78 | 0 | 0 |
| | **D** | 17 | 0 | 31 | 0 |
| | **M** | 0 | 0 | 0 | 15 |

| | | *actual* | | | |
|---|---|---|---|---|---|
| SVM-c | | **E** | **X** | **D** | **M** |
| *predicted* | **E** | 82 | 17 | 13 | 1 |
| | **X** | 2 | 78 | 0 | 0 |
| | **D** | 15 | 0 | 35 | 0 |
| | **M** | 0 | 0 | 0 | 15 |

TABLE IV
RESULTS OF RF AND SVM CLASSIFIERS WITHOUT (LEFT) AND WITH CLUSTERS(RIGHT)

| **RF-nc** | **E** | **X** | **D** | **M** |
|---|---|---|---|---|
| **precision (PPV)** | 0.73 | 0.95 | 0.81 | 1 |
| **recall (TPR)** | 0.86 | 0.84 | 0.74 | 0.94 |
| **accuracy** | 83.27% | | | |

| **RF-c** | **E** | **X** | **D** | **M** |
|---|---|---|---|---|
| **precision (PPV)** | 0.78 | 0.96 | 0.75 | 1 |
| **recall (TPR)** | 0.85 | 0.88 | 0.74 | 0.94 |
| **accuracy** | 84.06% | | | |

| **SVM-nc** | **E** | **X** | **D** | **M** |
|---|---|---|---|---|
| **precision (PPV)** | 0.73 | 0.98 | 0.65 | 1 |
| **recall (TPR)** | 0.80 | 0.82 | 0.74 | 0.94 |
| **accuracy** | 80.88% | | | |

| **SVM-c** | **E** | **X** | **D** | **M** |
|---|---|---|---|---|
| **precision (PPV)** | 0.73 | 0.98 | 0.70 | 1 |
| **recall (TPR)** | 0.83 | 0.82 | 0.73 | 0.94 |
| **accuracy** | 81.40% | | | |

Table IV contains the results calculated based on the abovementioned Confusion Matrix. Values received by that classifier are similar to the previous one. Both precision and recall values for each state are similar or slightly better for the method containing cluster memberships. Accuracy, i.e. the ratio of correctly predicted values to all values, also increases in the method containing clusters from 80.88% to 81.40%.

Results obtained by both classifiers are promising. Both classifiers achieved high precision and recall that indicate the correctly predicted class. Moreover, joining a cluster membership has a slightly better effect on the quality of the classification.

## IV. CONCLUSION

Unlabeled data appear to have a positive effect on the accuracy of classifying patients with bipolar disorder. In the study, a model was prepared that fits each patient individually. The presented solution individually sets the list of important acoustic parameters, the appropriate method of aggregating these data, and the number of clusters in which the patient may be. Such a model was compared with a model that did not include membership of clusters (so used amount of data from each patient was used for the model). The presented results indicate that adding information about clusters slightly improves the classification performance. The current results seem to be promising, and the study will be repeated for the remaining patients.

## ACKNOWLEDGMENT

Olga Kamińska and Katarzyna Kaczmarek-Majer are supported by the Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project: "Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)".

## REFERENCES

[1] A. Bouchachia and W. Pedrycz, “Data clustering with partial supervision,” *Data Min. Knowl. Discov.*, vol. 12, no. 1, p. 47–78, jan 2006. [Online]. Available: https://doi.org/10.1007/s10618-005-0019-1

[2] A. Grünerbl, A. Muaremi, and V. Osmani, “Smartphone-based recognition of states and state changes in bipolar disorder patients,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19(1), 2015.

[3] T. Chakraborty, “Ec3: Combining clustering and classification for ensemble learning,” in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 781–786.

[4] G. Casalino, G. Castellano, F. Galetta, and K. Kaczmarek-Majer, “Dynamic incremental semi-supervised fuzzy clustering for bipolar disorder episode prediction,” in *Discovery Science. DS 2020*, A. Appice and et al., Eds., 2020.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proc. of the 21st ACM Int. Conf. on Multimedia*, 2013, pp. 835–838.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[7] O. Kamińska, K. Kaczmarek-Majer, and O. Hryniewicz, “Acoustic feature selection with fuzzy clustering, self organizing maps and psychiatric assessments,” *Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2020, Lisbon*, 2020.

[8] J. Bezdeck, R. Ehrlich, and W. Full, “Fcm: Fuzzy c-means algorithm,” 1984.

[9] M. Pal, “Random forest classifier for remote sensing classification,” *Int. J. Remote Sens.*, pp. 217–222, 2005.

[10] K. Srinivasan, N. Mahendran, D. R. Vincent, C.-Y. Chang, and S. Syed-Abdul, “Realizing an integrated multistage support vector machine model for augmented recognition of unipolar depression,” *Electronics*, vol. 9, no. 4, p. 647, 2020.