JOURNAL COMPUTER SCIENCES INSTITUTE

JCSI 15 (2020) 214-218
Received: 1 May 2020
Accepted: 7 May 2020

Semi-supervised learning with the clustering and Decision Trees classifier for the task of cognitive workload study

Częściowo nadzorowane uczenie z zastosowaniem klasteryzacji oraz klasyfikatora Drzew Decyzyjnych w przypadku badania obciążenia poznawczego

Martyna Wawrzyk*

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

The paper is focused on application of the clustering algorithm and Decision Tress classifier (DTs) as a semi-supervised method for the task of cognitive workload level classification. The analyzed data were collected during examination of Digit Symbol Substitution Test (DSST) with use of eye-tracker device. 26 participants took part in examination as volunteers. There were conducted three parts of DSST test with different levels of difficulty. As a results three versions were obtained of data: low, middle and high level of cognitive workload. The case study covered clustering of collected data by using k-means algorithm to detect three clusters or more. The obtained clusters were evaluated by three internal indices to measure the quality of clustering. The David-Boudin index detected the best results in case of four clusters. Based on this information it is possible to formulate the hypothesis of the existence of four clusters. The obtained clusters were adopted as classes in supervised learning and have been subjected to classification. The DTs was applied in classification. There were obtained the 0.85 mean accuracy for three-class classification and 0.73 mean accuracy for four-class classification.

Keywords: clustering; semi-supervised learning; eye-tracker

Streszczenie

Celem artykułu było zastosowanie klasteryzacji wraz z klasyfikatorem Drzew Decyzyjnych jako częściowo nadzorowanej metody klasyfikacji poziomu obciążenia poznawczego. Dane przeznaczone do analizy zostały zebrane podczas badania DSST (z ang. Digit Symbol Substitution Test) z użyciem urządzenia eye-tracker. 26 wolontariuszów wzięło udział w badaniu. Zostały przeprowadzone trzy części testu DSST o różnych poziomach trudności. W wyniku tego, otrzymano trzy wersje danych: z niskim, średnim i wysokim poziomem obciążenia poznawczego. Do analizy danych został użyty algorytm klasteryzacji *k-means* do wyznaczenia trzech lub większej liczby klastrów. Uzyskane klastry zostały poddane ocenie przy użyciu trzech wewnętrznych indeksów w celu zmierzenia jakości klasteryzacji. Indeks David-Boudin'a wykazał najlepsze rezultaty w przypadku istnienia czterech klastrów. Na podstawie tej informacji można sformułować hipotezę, iż dane są podzielone na 4 klastry, co oznaczałoby istnienie dodatkowego poziomu poznawczego. Uzyskane klastry zostały zaadoptowane jako klasy w uczeniu pod nadzorem. Do klasyfikacji danych został użyty klasyfikator Drzew Decyzyjnych . Otrzymano średnią dokładność równą 0.85 w przypadku 3-klasowej klasyfikacji oraz 0.73 średnią dokładność dla 4-klasowej klasyfikacji.

Słowa kluczowe: klasteryzacja; uczenie częściowo nadzorowane; eye-tracker

*Corresponding author

Email address: martyna.wawrzyk@pollub.edu.pl (M.Wawrzyk)

1. Introduction

Clustering of unlabeled data is a commonly used method in unsupervised learning. It allows to detect the clusters of data with no labels. There were used many types of clustering methods in analysis of oculography data. In research [1], [2], [3] there were conducted the bidimensional clustering to detect the fixations. *K-means* algorithm can be used to identify the microsaccades [4] and areas of visual interest [5]. The Gaussian Mixture Models were applied for clustering gaze locations in dynamic scenes [6]. The hierarchical clustering was used in analysis of eye movements [7].

There is possibility to combine two methods from unsupervised learning and supervised learning [8]. In effect there is obtained the semi-supervised learning method. In research [9] there were performed the spectral clustering and semi-supervised Gaussian process regression to in order to analyze the tracking of gaze.

But the one of the main problems of clustering is an evaluation of a model. There are two types of indices, which helps to assess the quality of clustering: external indices and internal indices [10]. Internal indices are used to measure the quality of a clustering without external information [11]. External indices evaluates the clustering with external information [12].

In case study of this paper there were used the digitalized version of the Digit Symbol Substitution Test (DSST) [16]. The DSST is a commonly used test in clinical neuropsychology in order to measure cogni-

tive dysfunction. The test allows to measure the speed of processing data, memory and others cognitive functions of a patient. For this reason is widely used in neuropsychology [17].

The aim of this paper is to apply a semi-supervised learning with K-means algorithm clustering with Decision Trees classifier. The assumption is the existence of three clusters, taking into account the fact of using a three-level cognitive workload study. The application of this algorithm is designed to check out if there are more than 3 clusters. The obtained clusters were evaluated by using three indices. Finally the clusters were adopted as a classes in supervised learning in order to classify the features of clusters.

2. The research procedure

In the study there were used the computerized version of DSST test. The examined person has to match symbols to the numbers according to a template located on the bottom of the screen. In order to match the number to the symbol the user needs to click the proper symbol on the template. The currently active letter is marked by graphical frame (Fig.1). After clicking the frame is moving to the next letter. The time to match the symbols is specified and letters are generated randomly.

The interface of application is presented in Fig.1. The application was developed in Java and is operated using a computer mouse. The case study was divided into three stages. There were performed DSST test three times with the following settings:

— 4 different symbols to assign; the test lasted 90 s.
— 9 different symbols to assign; the test lasted 180 s.

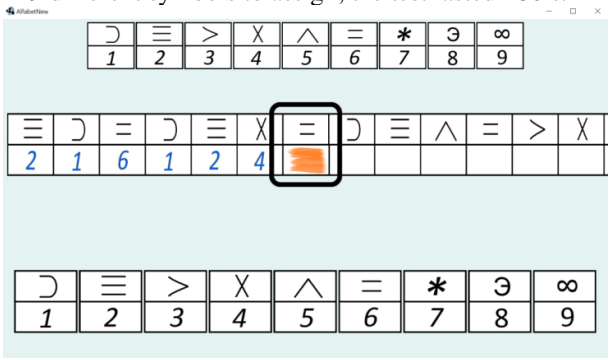


Figure 1: Procedure of data processing

2.1. Set up and equipment

An eye-tracker device and computer were used in the study. The experiment was performed in dedicated laboratory illuminated with standard fluorescent light. The data was recorded by Tobii Pro TX300 screen-based eye-tracker. The technology of Tobii Pro TX300 is based on video-oculography. It collects data using the dark pupil and corneal reflection method with the frequency of 300 Hz.

The Tobii Studio 3.2 was used to design the experiment. This software is dedicated for eye-tracker experiment and is compatible with eye-tracker device. The monitor with the following parameters: 23'' TFT monitor at 60 Hz was applied to present the visual stimuli.

The experiment was conducted in sitting position with distance between participant and monitor in range from 50 to 80 cm. The same procedures was issued for each participant.

2.2. Experiment

The study was performed with 26 participants aged 20-24. The duration of the examination of single participant lasted 15 minutes. The study was divided into three parts. There were conducted the process of calibration in each stage by using the eye-tracker device. The calibration process was consisted of 9-point built-in procedure. After the calibration, the instruction was displayed on the monitor to inform about the procedure of assignment the symbols to the numbers. The study consisted of three parts of DSST test. Each part had a different number of symbols to assign and lasted for different period of time. Each part had a short initial trial to familiarize the participants with the task. Each participant had to finish three parts of DSST test.

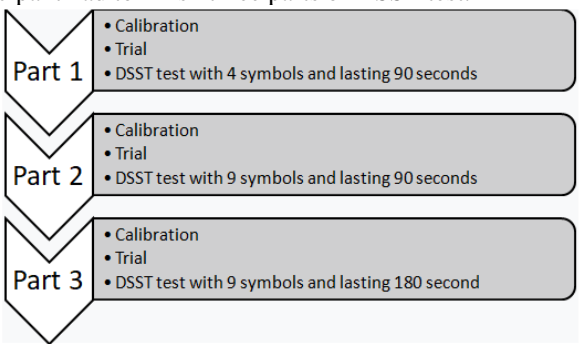


Figure 2: Procedure of the examination

2.3. Data set

There were obtained 156 files: 78 files generated from eye-tracker device and 78 files generated from the application. From each participant there were received 6 files: 3 from eye-tracker device and 3 from the application (the data from each part of test were saved in separate files).

3. Methods applied

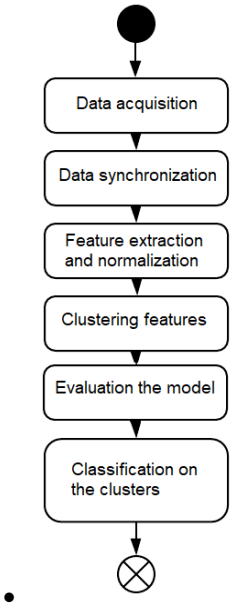
3.1. Data processing

The data processing was divided into six stages (Fig. 3). There were obtained two separate files from each part of experiment. Consequently after data acquisition there were performed data synchronization.

In sequence the features were extracted. The following metrics were selected as features for further data processing data:

- number of blinks, mean and duration of blinks;
- mean, standard deviation, maximal and minimal duration of saccades;
- mean, standard deviation, maximal and minimal duration of fixations;

- mean and standard deviation of left/right pupil diameter;
- number of responses;
- number of mistakes in responses;
- mean time of response;



- Figure 3: Procedure of data processing

3.2. Unsupervised learning

Clustering is the method performed in unsupervised learning, which is used to for extracted clusters of unlabeled data. The k-means algorithm were chosen for the clustering of extracted features. K- means clustering is based on defining number of centroids and assigning of each data point to the nearest cluster.

Also specific internal indices to evaluate the model were used. It is necessary to assess to quality of clustering. In this paper there were considered three type of internal indices. The Calinski-Harabasz index is defined as a ratio between the within-cluster dispersion and the between-cluster dispersion [13]. In the case of this index there is a need to make a line-plot (dependence of index value on the number of clusters). If the peak on the line-plot were observed, this point (number of clusters) should be chosen as the best clustering. The next index is Silhouette index, which is a coefficient between mean-intra cluster distance to mean nearest-cluster distance [14]. The value 1 gives the best results, the value -1 means incorrect clustering, the value close to 0 means clustering overlaps. The last index is a David-Boudin index, which is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances [15]. The lowest values of index means the better clustering.

3.3. Supervised learning

The received clusters were adopted as classes in supervised learning. The Decision Trees (DTs) were conducted in classification of features. It was performed for k-class classification. The model of DTs and clustering were implemented in Google Colab environment.

4. Results

4.1. Clustering

There were generated plots to visualize the distribution of clusters. The PCA algorithm were used to reduce data before the generated of plots. The Fig. 4 illustrates the arrangement of features assuming existence of three clusters, in turn Fig. 5 presents the arrangement in case of four clusters. The red points represent the centroids of the clusters. The Fig.5 presents that the points of cluster 4 are visible separated from other clusters**.**

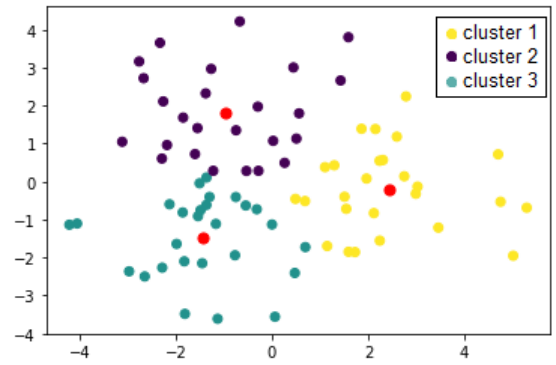


Figure 4: The data distribution into 3 clusters

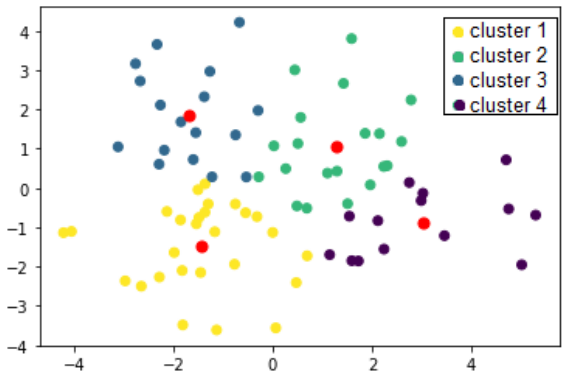


Figure 5: The data distribution into 4 clusters

The Table 1 contains the results of three chosen indices to evaluate the quality of clustering. The comparison was conducted between five clusters. Each metrics must be interpreted individually. In case of Caliniski-Harabsz index it is worth to visualizing by plot and find the characteristic point. The Fig.6 presents the dependence of Caliniski-Harabasz index value on the number of clusters. Base on the chart, it cannot be determined which number of clusters represents a real distribution of data. There were not observed the peak which means the best clustering. The silhouette index also does not specify which number of clusters gives the best results. The Davies-Boundin score with the lowest values indicating better clustering. The number of four clusters presents the lowest values in compared to others. Based on this information it is possible to formulate the hypothesis of the existence of four clusters.

Table 1. The values of internal indexes for k-clusters

Clustering metrics/ k-clusters	2	3	4	5
Calinski-Harabasz index	16.04	13.48	12.28	11.12
Davies-Boudin index	2.04	1.97	1.80	1.83
Silhouette index	0.16	0.14	0.14	0.13

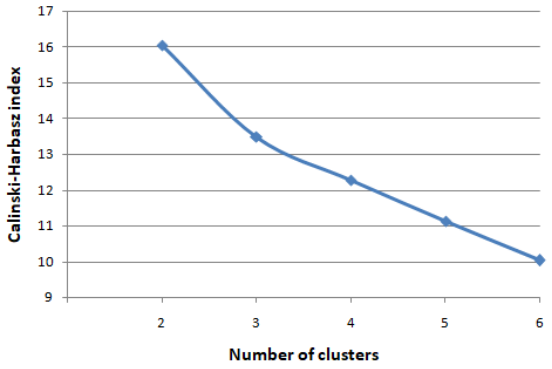


Figure 6: The line-plot of Calinski-Harabsz index

4.2. Classification results

The Table 2 presents the mean accuracy values of k-class classification. The best results were obtained for two-class and three- class classification. In case of four-class and five-class classification there were obtained good results.

Table 2. The values accuracyfor k-class classification

K-class	2	3	4	5
Mean accuracy of classification	88.5%	85.75%	73.85%	75.75%

Table 3 and Table 4 present the confusion matrix of three-class and four-class classification. There were observed the correct classification the class 2 in case of three-class classification. The confusion matrix of four-class classification presents the correct classification of class 3 and class 4.

Table 3. Mean confusion matrix for three-class classifaction

	class 1	class 2	class 3
class 1	3	2	0
class 2	0	5	0
class 3	3	0	3

Table 4. Mean confusion matrix for four-class classifaction

	class 1	class 2	class 3	class 4
class 1	4	0	2	0
class 2	0	2	1	0
class 3	0	0	4	0
class 4	0	0	3	0

5. Discussion and conlusion

The aim of this paper was to applied a semi-supervised learning in case study of cognitive workload based on eye-tracker data. There were conducted study with the computerized version of Digit Symbol Substitution Test (DSST). The collected data contained eye-tracking features related to blinks, fixation, saccades and pupil diameter.

Firstly there were used a k-mean algorithm to detect clusters of analyzed data. The obtained data were evaluated by three clustering metric to assess the quality of clustering. The David-Boudin metric indicates better clustering in case of four clusters compared to others. Based on this it may be concluded that the data is divided into 4 clusters. Fig. 5 illustrates the distribution of data with four clusters. The four cluster is visible separately from others clusters. The cluster no 4 may represent the different level of cognitive workload.

The results of classification present very good mean accuracy in case of three-class classification and also four-class classification. Also there were extracted the results in case of two-class and five-class to make a comparison.

In summary, a hypothesis for the collected data being able to be divided into four has been suggested. But there cannot be made an assumption that the data has a distribution into four clusters. The clustering metrics cannot indicate which number of clusters gives the best results. These indices may inform which number of cluster is better than other.

References

[1] T. Urruty, S. Lew, N. Ihadaddene and D. A. Simovici, Detecting eye fixations by projection clustering. ACM Transaction on Multimedia Computing, Communications and Application, 3 (4), 5:1–5:20, 2007

[2] N. Flad, T. Fomina, H. H. Buelthoff and L. L. Chuang, Unsupervised Clustering of EOG as a Viable Substitute for Optical Eye Tracking. Eye Tracking and Visualization, Cham, 2017, 151–167

[3] R. S. Hessels, D. C. Niehorster, C. Kemner and I. T. C. Hooge Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). Behaviour Research Methods, 49 (5), 1802–1823, 2017

[4] J. Otero-Millan, J. L. A. Castro, S. L. Macknik and S. Martinez-Conde Unsupervised clustering method to detect microsaccades. Journal of Vision, 14 (2), 18–18, 2014

[5] A. Santella and D. DeCarlo Robust clustering of eye movement recordings for quantification of visual interest. Proceedings of the 2004 symposium on Eye tracking research & applications, San Antonio, Texas, 2004, 27–34

[6] P. K. Mital, T. J. Smith, R. L. Hill and J. M. Henderson, Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. Cognitive Computation, 3 (1), 5–24, 2011

[7] Z. Kang and S. J. Landry An Eye Movement Analysis Algorithm for a Multielement Target Tracking Task:

Maximum Transition-Based Agglomerative Hierarchical Clustering. IEEE Transactions on Human-Machine Systems, 45 (1), 13–24, 2015

[8] M. Aamir and S. M. A. Zaidi Clustering based semi-supervised machine learning for DDoS attack classification. Journal of King Saud University - Computer Information Sciences, 2019

[9] K. Liang, Y. Chahir, M. Molina, C. Tijus and F. Jouen Appearance-based gaze tracking with spectral clustering and semi-supervised Gaussian process regression. Proceedings of the 2013 Conference on Eye Tracking South Africa, Cape Town, South Africa, 2013, 17–23

[10] K. Wang, B. Wang and L. Peng CVAP: Validation for Cluster Analyses. Data Science Journal, 8 (0), 88–93, 2009

[11] A. Thalamuthu, I. Mukhopadhyay, X. Zheng and G. C. Tseng Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics (Oxford, England), 22 (19), 2405–2412, 2006

[12] S. Dudoit and J. Fridlyand A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology, 3 (7), 2002

[13] T. Caliński and J. Harabasz A dendrite method for cluster analysis. Communications in Statistic, 3 (1), 1–27, 1974

[14] P. J. Rousseeuw Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65, 1987

[15] D. L. Davies and D. W. Bouldin A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2), 224–227, 1979

[16] C. Boake From the Binet-Simon to the Wechsler-Bellevue: tracing the history of intelligence testing. Journal of Clinical and Experimental Neuropsychology, 24 (3), 383–405, 2002

[17] V. Sicard, R. D. Moore, i D. Ellemberg Sensitivity of the Cogstate Test Battery for Detecting Prolonged Cognitive Alterations Stemming From Sport-Related Concussions. Clinical Journal of Sport Medicine: Official Journal Canadian Academy Sport Medicine,29 (1), 62–68, 2019