Unsupervised Analysis for Early-Stage Diagnosis of Cognitive Disorders

Olga Georgieva*, Preslav Hadzhitsanev*, Dessislava Petrova-Antonova **
* Sofia University "St Kliment Ohridski", Sofia, Bulgaria
** Institute GATE, Sofia University, Sofia, Bulgaria
o.georgieva@fmi.uni-sofia.bg,

***Abstract*—Cognitive disorder is a condition severe enough to compromise social and/or occupational functioning. It has a huge social significance to all the parties involved in their diagnosis and treatment. The timely diagnosing and early treatment have a great social and economic importance. The neuropsychological and neuropsychiatric assessments provide a solid base for the diagnosis and prediction of cognitive disorders. The present research work aims to find dependencies between factors and conditions for disease's recognition. It is based on data collected for three very commonly applied examinations for functional, mental and neuropsychological assessment. Clustering analysis is applied as appropriate datamining technique. The results show that the functional assessment data present more clear structuring than the separation ability of the other examinations. The other part of the research conclude that the diagnosis's prediction based on groups found by the clustering analysis is improved by analysis of the merged datasets instead the individual ones.**

I. INTRODUCTION

Cognitive disorders are impairments of brain function that interfere the people to cope with everyday tasks and work. More often they are caused by neurodegenerative diseases, vascular diseases, brain damage and the various combinations between them. Cognitive disorders and especially dementia are severe prolonged and often irreversible conditions that compromise social and/or professional functioning [5]. Given the ageing and growing world population, the tasks of its timely diagnosing and thus treating, including delaying irreversible impairment are of great social and economic importance. However, the diagnosis of the cognitive disorders meets difficulties as follows:

- A reliable diagnosis could be done by invasive methods or/and expensive and time-consuming examinations. An exact diagnosis could be done after a lot and different types of medical investigations.
- Usually, the diagnosis takes place at a later stage of the disease development, when the opportunities for adequate patient care are very few.
- Different cognitive states resulting from other diseases may have similar appearances and in this sense to blurry the right diagnosis and then consequent treatment.

In this context, it is very important to know for existing dependencies among different factors and appearances of the cognitive disorders. Recently, a large amount of data has been collected for clinical research using different medical, psychological methods. Some strong statistical data analysis about possible dependences and relations in the cognitive states exist [2]. However, a deep understanding of the existing interdependence between diseases symptoms and appeared cognitive states is still lacking.

Our special attention is focused on the exploration of the factors and symptoms of dementia, which can be easily and quickly used to diagnose the disease. Such knowledge could help for an early-stage diagnosis by saving human efforts and resources for patient caring. In addition, it could reveal new understandings about the disease mechanisms and causes.

II. RESEARCH QUESTIONS

The neuropsychological and neuropsychiatric assessments provide a solid base for the diagnosis and prediction of cognitive impairments [1,3]. Vascular Cognitive Impairment Harmonization Standards (VCIHS) are used for the evaluation of Vascular Cognitive Impairment (VCI) [4] by covering four cognitive domains - memory, visuospatial, executive/activation and language.

We based our research on commonly applied neuropsychological examinations. Each of them evaluates the subject's status by specific questions and observations. The examinations are not invasive and do not need special medical equipment. The commonly accepted approach is patient evaluation according to statistical assessment by the total score of each examination.

In contrast, our consideration is that the collected data of each neuropsychological and neuropsychiatric assessments serve as a data space for analysis. By investigating the characteristics of these data space, more insights into particular dependences for the disease's recognition could be obtained. An additional concern is to find more informative data sets - for each examination as well as combining the important features of different data sets. Explorations of these concerns would give valuable knowledge for early-stage cognitive disorder diagnosis and by that lower rates of health complications, fewer emergency hospital visits and by that to achieve a better overall quality of life.

III. METHODOLOGY

The research work is based on data collected for three very commonly applied neuropsychological examinations: Functional Activities Questionnaire (FAQ) as a powerful tool for screening dementia, Mini-Mental State

Examination (MMSE) that evaluates the global cognitive function and assessment the possible presence of 12 symptoms in dementia cases by Neuropsychiatric Inventory Questionnaire (NPI-Q) providing an index to score the corresponding severity of each symptom. The data is provided by Alzheimer's Disease Neuroimaging Initiative (ADNI)[i]. The main goal is to find data spaces that distinguish the three subjects' groups: Controls (CN), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD).

Data mining techniques can be applied, in order to uncover hidden meaningful patterns in the considered data sets. Due to the lack of a reference model, studies of the data spaces can be carried out using cluster analysis. This machine learning technique does not need preliminary information about the data structure. An effective solution could be done by different clustering algorithms as those based on objective function minimization (K-means, Fuzzy-C-Means (FCM)) or based on cluster density assessment (DBscan). The objective function clustering finds groups represented by a cluster center. Fuzzy clustering by FCM is a good opportunity to deal with the existing data uncertainty and lack of a clear boundary between groups. DBscan separates the valuable clusters and outliers is one pass. In this research work a commonly accepted K-means algorithm is applied.

The research procedure follows the steps:

1. Data clustering;
2. Evaluation of clustering;
3. Analysis of the obtained clusters to evaluate to what extent the clusters cover the patients' status;
4. Exploring combined data sets to improve the predictability of the diagnosis result;
5. Interpretation of the results to describe the dependencies identified in the data.

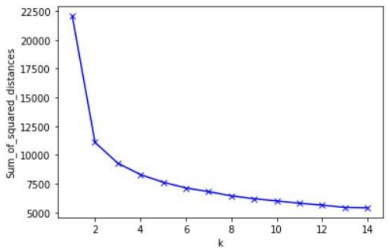
IV. Results and Discussion

Three data sets of psychological examination and questionnaires provided by the ADNI database are investigated. In this research work, only baseline data i.e. first visit data of ADNI1 data set are examined. It comprises 819 person examinations.

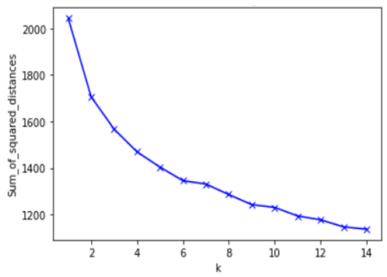
A. Data clustering and evaluation

The three data sets were filtered for features that are not able for data structure identification as values are missing or are nominal. Thus, 11 features of FAQ and 31 of MMSE remain for evaluation. As many of NPI-Q columns are empty, only 12 features of them remain for analysis.

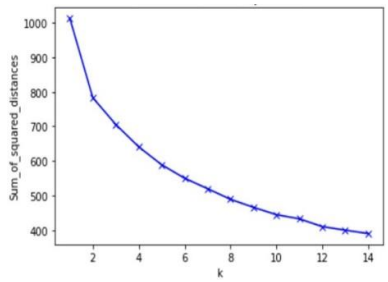
Further K-Means clustering is applied to each filtered dataset. In order to define the proper cluster number *k*, an Elbow method was applied (Figure 1) to each dataset.



a) FAQ data



b) MMSE data



c) NPI-Q data

Figure 1. Results from application of Elbow method, *k* - number of clusters

Whereas the FAQ data tends to group in 3 clusters, *k*=3 (Fig.1.a), the rest two sets show relatively equivalently quality of grouping as for *k*=3 as well in more clusters, *k*=4 or 5 (Fig.1.b, Fig.1.c). In conclusion FAQ data assessment presents a more reliable separation ability than the other examined data. Additional investigation of MMSE and NPI-Q data are needed to define clearly the *k* value. However, for the followed comparative analysis provided here we explore clustering in three clusters to all investigated sets.

B. Feature selection

Feature selection was accomplished by step-wisely assessment of Silhouette Coefficient [6] for finding the best features of each dataset clustered in three clusters, *k*=3 (TABLE I). Silhouette Coefficient *SC* is calculated using the mean intra-cluster distance *a* and the mean nearest-cluster distance *b* calculated of each sample as follows:

$$SC = \frac{b-a}{\max(a,b)}. \quad (1)$$

TABLE I.
THE MOST INFORMATIVE FEATURES OF EACH DATASET

Dataset	Features
FAQ	,'FAQBEVG', 'FAQTV', 'FAQEVENT'
MMSE	'MMDATE', 'MMFLAG', 'MMD'
NPI-Q	'NPIE', 'NPIA', 'NPIB'

The meaning of the selected features is the following given in the order of their importance:

- For FAQ

'FAQBEVG' - Heating water, making a cup of coffee, turing off the stove.

'FAQTV' - Paying attention to and understanding a TV program, book, or magazine.

'FAQEVENT' - Keeping track of current events.

- For MMSE

'MMDATE' - What is today's date?

'MMFLAG' - Verbatim response: “Flag”

'MMD' - Verbatim response: “Letter D”

- For NPI-Q

'NPIE' - Does Patient become upset when separated from you? Does he/she have any other signs of nervousness, such as shortness of breath, sighing, being unable to relax, or feeling excessively tense?

'NPIA' - Does Patient believe that others are stealing from him/her, or planning to harm him/her in some way?

'NPIB' - Does Patient act as if he/she hears voices? Does he/she talk to people who are not there?

The attempts to cover the three states – CN, MCI and AD by the defined clusters in each individual data set is not successful as the three diagnosis are presented in each cluster almost in parity.

C. Clustering of the merged datasets

In seeking an improvement of the clustering informativity we explore clustering of the merged sets FAQ, MMSE, NPI-Q. This clustering improves the predictability in comparison with the individual data set in sense of covering the subject diagnosis. However, the best results are obtained with merged data sets comprising all data features instead only the most informative ones. Clustering for k=3 gives best disease prediction results. In this data structuring Cluster 1 recognizes mostly MCI, Cluster 2 – CN with parity MCI and Cluster 3 – AD (TABLE II, in bold). The results clearly show that improvement of diagnosis prediction is achieved by accounting for all data.

TABLE II.
THE PERCENTAGE OF SUBJECTS SEPARATED IN EACH CLUSTER ACCORDING TO THE DIAGNOSIS BY MERGED DATASET

Clustering of merged dataset		CN (%)	MCI (%)	AD (%)
K-Means, k=2	Cluster 1	41	54	5
	Cluster 2	0	36	64
K-Means, k=3	Cluster 1	2	**70**	28
	Cluster 2	**50**	48	2
	Cluster 3	0	25	**75**

The experiment for clustering merged data presented by only the best three selected features of TABLE 1 does not give any improvement. The results are given in TABLE III, where in order to balance the three subject groups only 193 subject samples - randomly selected, are taken to evaluate the percentage values. That is a way to balance the diagnosis groups.

TABLE III.
THE PERCENTAGE OF SUBJECTS SEPARATED IN EACH CLUSTER ACCORDING TO THE DIAGNOSIS

Clustering of merged dataset with first 3 best features		CN (% of 193)	MCI (% of 193)	AD (% of 193)
K-Means, k=2	Cluster 1	48	37	15
	Cluster 2	0	25	75
K-Means, k=3	Cluster 1	0	13	**87**
	Cluster 2	**48**	37	15
	Cluster 3	1	30	**69**

The third experiment also uses the merged data set formed by the total score data of each cognitive dataset. The total score provides information about the summary estimation of each test for every person. A new clustering is applied in a three-dimensional space formed by the total score values of datasets FAQ, MMSE and NPI-Q. According to the Elbow method the best structuring is obtained for division in three clusters (Fig. 2).

The predictability ability of this division is estimated according to the percentage of coverage of each diagnosis group in each cluster. For this a balanced data of 193 subject is chosen and represented (TABLE IV). Again, as in case of space formed by best selected features (TABLE III) the group of Alzheimer is covered by two clusters.

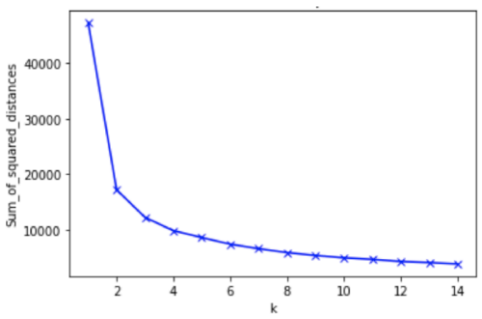


Figure 2. Results of Elbow method for clustering in three-dimensional data space formed by the total score values of each dataset.

TABLE IV.
THE PERCENTAGE OF SUBJECTS SEPARATED IN EACH CLUSTER IN THREE-DIMENSIONAL DATASET

Clustering of three-dimensional dataset		CN (% of 193)	MCI (% of 193)	AD (% of 193)
K-Means, k=2	Cluster 1	0	15	85
	Cluster 2	47	41	12
K-Means, k=3	Cluster 1	0	6	**94**
	Cluster 2	**53**	40	7
	Cluster 3	1	28	**72**

This result could be explained with the well-done individual questionnaires that form the distinct data sets and thus each feature makes an equal contribution to the overall assessment. Indirectly this result is supported by the feature selection experiments, which show that best selected features do not differ much in value from the others features according to the selection assessments.

V. CONCLUSIONS

The research is focused on the investigation of factors and symptoms of cognitive disorder and in particular dementia, which can be easily and quickly used to diagnose the disease. The work is based on data collected for three very commonly applied examinations as functional, mental and neuropsychological. Based on clustering analysis technique conclusions about the ability to separate the individual dataset as well as the merged dataset are provided.

It is found that the Functional Activities Questionnaire data are tend to separate in three clusters, whereas for the mental and neuropsychiatric assessment data the number of the clusters is not clearly found. The other part of the research conclude that the diagnosis's prediction based on groups found by the clustering analysis is improved by analysis of the merged datasets instead the individual ones

ACKNOWLEDGMENT

This research work has been supported by GATE project, funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement No.857155 and Operational Programme Science, by Operational Programme Science and Education for Smart Growth under Grant Agreement no. BG05M2OP001-1.003-0002-C01 and by the Bulgarian National Science fund under project no. KP-06-N32/5.

REFERENCES

[1] K. Bokenberger, P. Ström, A. Aslan, T. Åkerstedt, N. Pedersen, Shift work and cognitive aging: A longitudinal study, Scandinavian Journal of Work, Environment & Health, 43(5), pp. 485-493, 2016.

[2] E. Lee, H. Zhu, D. Kong, Y. Wang, K. Giovanello, J. Ibrahim, BFLCRM: A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease," The Annals of Applied Statistics, 9(4), pp. 2153-2178, 2015.

[3] W. Li, L. Sun, G. Li, S. Xiao, Prevalence, Influence Factors and Cognitive Characteristics of Mild Cognitive Impairment in Type 2 Diabetes Mellitus, Frontiers in aging neuroscience, 11, 180, 2019.

[4] S. H. Park, M. K. Sohn, S. Jee, S. S. Yang, The Characteristics of Cognitive Impairment and Their Effects on Functional Outcome After Inpatient Rehabilitation in Subacute Stroke Patients, Annals of rehabilitation medicine, 41(5), pp. 734–742, 2017.

[5] Hugo, J., & Ganguli, M. (2014). Dementia and cognitive impairment: epidemiology, diagnosis, and treatment. Clinics in geriatric medicine, 30(3), 421–442.

[6] Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20, 53-65.

---

[i] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.
A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf