Визуализация и кластеризация данных в задаче системного анализа результатов обследования пациентов на начальных стадиях когнитивных расстройств

Б.Р. Салем[1], В.И. Солодовников[1]

[1]Центр информационных технологий в проектировании РАН, ул. Маршала Бирюзова 7а, Одинцово, Россия, 143000

Аннотация. Существует несколько десятков болезней и состояний, в рамках которых развиваются когнитивные нарушения, представляющие собой ухудшение когнитивных функций по сравнению с индивидуальной нормой. В настоящее время причины появления и прогрессирование когнитивных нарушений слабо изучены. В данной работе представлен системный анализ данных, включающих в себя 35 пациентов и результаты их обследования. Для визуализации были использованы методы предварительной обработки данных, уменьшения размерности данных, кластеризации. Несмотря на небольшой набор данных удалось достичь точности диагностирования когнитивных нарушений на начальных стадиях в 63%

1. Введение

Когнитивные функции являются одной из наиболее важных интегративных функций головного мозга. В пожилом и старческом возрасте возникает их снижение, могут развиться и тяжелые расстройства, приводящие к деменции, распространенность которой в популяции людей старше 70 лет составляет около 13.9% [1]. В настоящее время эффективного лечения деменции или предшествующего ей так называемого умеренного когнитивного расстройства (УКР) [2] не существует. Начало патологического процесса может протекать без явных проявлений и пациенты не ощущают заболевание на ранней стадии. Несвоевременная диагностика приводит к необратимым последствиям в функционировании головного мозга. Поэтому наиболее действенным способом снизить связанные с ними значительные экономические и социальные потери является их профилактика, основанная на своевременном выявлении менее значительных когнитивных расстройств – субъективного когнитивного нарушения (СКН) и легкого когнитивного нарушения (ЛКН). Причиной последних могут быть естественные возрастные структурно-функциональные изменения головного мозга, а также различные патологические процессы, главными из которых являются сердечно-сосудистые, связанные с ними сосудисто-мозговые нарушения и нейродегенеративные заболевания, среди которых наиболее частой причиной деменции является болезнь Альцгеймера [3]. Для определения возможных причин когнитивных расстройств требуется проведение комплексного анализа, который базируется на анамнестических, клинических, а также инструментальных данных, подразделяющихся на лабораторные и нейровизуализационные методики. Учитывая сложность проведения всего комплекса подобных обследований, возникает необходимость выявить набор наиболее значимых признаков и их сочетаний для обнаружения начала патологического

процесса. В связи с этим, задачей является анализ возможных медицинских и биологических факторов, приводящих к развитию и прогрессированию субъективного и легкого когнитивного нарушений, что позволит значительно упростить их диагностику на ранних стадиях. Важной особенностью является изучение в комплексе и динамике взаимоотношений клинических, нейропсихологических, инструментальных, лабораторных показателей, отражающих состояние головного мозга, неврологической и сердечно-сосудистой систем. Предлагается провести системный анализ данных, полученных при прохождении пациентами ряда стандартных процедур и когнитивных тестов. На начальном этапе данные необходимо визуализировать и классифицировать, что позволит увидеть специфику явлений, их разнообразие, свойства, связи и зависимости параметров.

2. Исходные данные

Исходные данные представляют собой таблицу со сведениями о 35-ти пациентах. Каждый пациент имеет 76 параметров, которые были условно разделены на 2 категории:

1) физические данные – представляют собой характеристики, которые были получены при первом визите к врачу, это данные описывающие возраст, пол, количество глюкозы, общего холестерина и т.д.

2) Изменяющиеся характеристики, полученные в результате проведения некоторых процедур, таких как измерение давления, прохождение теста Струпа[4], прохождение Мока теста[5].

Изменяющиеся характеристики были учтены два раза, при первом и повторном визите пациента к врачу. Было решено разделить исходную выборку на две подвыборки: первая состоит из физических данных и изменяющихся характеристик, полученных при первом визите к врачу, вторая подборка состоит из физических данных и изменяющихся характеристик, полученных при последующем визите. Предполагается, что при измерении параметров пациента не было допущено ошибок со стороны медицинского персонала, что должно соответствовать достоверным результатам обследования.

3. Применяемые методы обработки данных

Системный анализ массива исходных данных включал решение следующих задач:

1. предварительная обработка исходных данных;
2. кластеризация обработанных данных;
3. уменьшение размерности данных;
4. визуализация данных меньшей размерности.

Предварительная обработка заключалась в разделении исходного массива данных на несколько подвыборок на основе таких характеристик как:

1) группа, к которой был отнесен пациент после медицинского обследования (ЛКН или СКН);

2) время визита (первый визит или последний).

На основе этих данных можно верифицировать смещение показателей между первым и последним визитом, корреляцию параметров исходных данных, а также сравнить показатели пациентов с СКН и ЛКН. В рамках данной задачи, для проверки линейной сепарабельности среди пациентов с ЛКН и СКН было решено провести классификацию без учителя. Для кластеризации был выбран метод k-средних[6], который также предусматривает выбор ожидаемого количества кластеров. Для последующей визуализации требуется уменьшить размерность данных, что позволяет представить данные в двумерном или трехмерном пространстве. Многие параметры в исходном массиве данных коррелируют между собой, так как получены при проведении стандартных медицинских процедур, таких как измерение артериального давления, общий холестерин, глюкоза и т.д. С учетом этой особенности для уменьшения размерности был выбран метод главных компонент[7], что позволяет выделить ключевые параметры и уменьшить размерность, потеряв при этом минимум информации.

4. Результаты

Для визуализации использовались наборы данных, полученные на этапе предварительной обработки исходной выборки. Для наглядности было необходимо представить результаты в интуитивно-понятной для человека форме, что в свою очередь было достигнуто с помощью описания результатов кластеризации в двухмерном или трехмерном пространствах. Кластеризация была проведена с помощью программной библиотеки scikit-learn[8], для визуализации использовалась программная библиотека matplotlib[9].

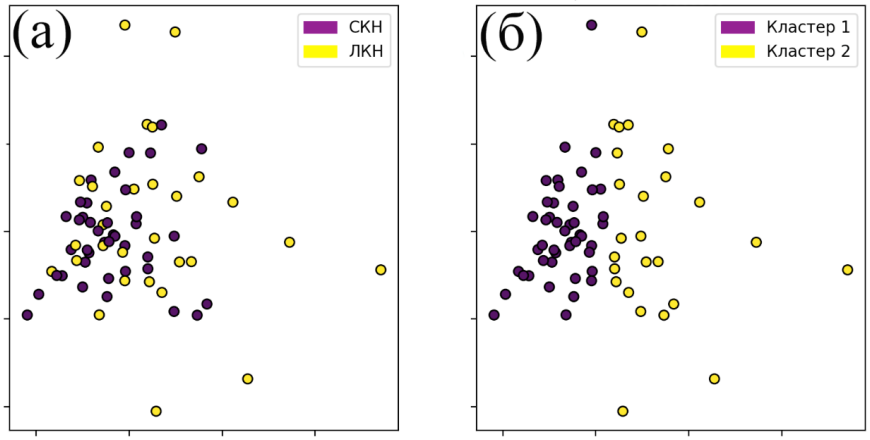

Рисунок 1. Сравнение исходных данных с результатами кластеризации в двухмерном пространстве. (а) — исходные данные, (б) — результат кластеризации.

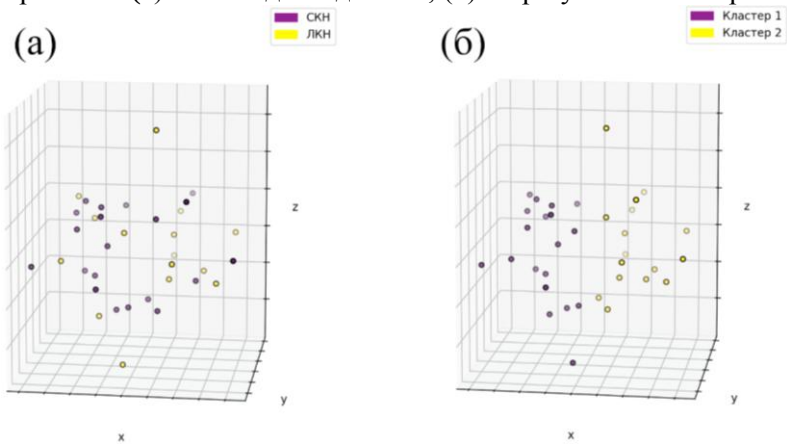

Рисунок 2. Сравнение исходных данных с результатами кластеризации в трехмерном пространстве. (а) — исходные данные, (б) — результат кластеризации.

Результаты кластеризации были проанализированы с помощью двух показателей:

$$\text{Точность} = \frac{tp}{tp+fp} \quad (1)$$

$$\text{Полнота} = \frac{tp}{tp+fn} \quad (2)$$

где *tp* – количество совпадений результатов кластеризации и исходных данных, *fp* - количество ложных срабатываний (ошибка первого рода[10]), *fn* - количество пропусков события (ошибка второго рода[10]).

Таблица 1. Точность и полнота для СКН.

Учитываемые характеристики	Точность	Полнота
Только 1-го визита	0.631	0.850
Только 2-го визита	0.628	0.601
Обоих визитов	0.652	0.653

Таблица 2. Точность и полнота для ЛКН.

Учитываемые характеристики	Точность	Полнота
Только 1-го визита	0.750	0.410
Только 2-го визита	0.562	0.601
Обоих визитов	0.650	0.603

Динамика изменения результатов тестов была выражена в виде сравнения характеристик, полученных при первом и втором визитах, в результате было замечено, что большая часть данных не отклоняется более чем на 20%, кроме показателей, отвечающих за прохождение тестов 10 слов [11]. Также проведена визуализация изменения состояния пациента между первым и последним визитом, на графике разница состояний между визитами представлена в виде стрелки на рисунке 3.

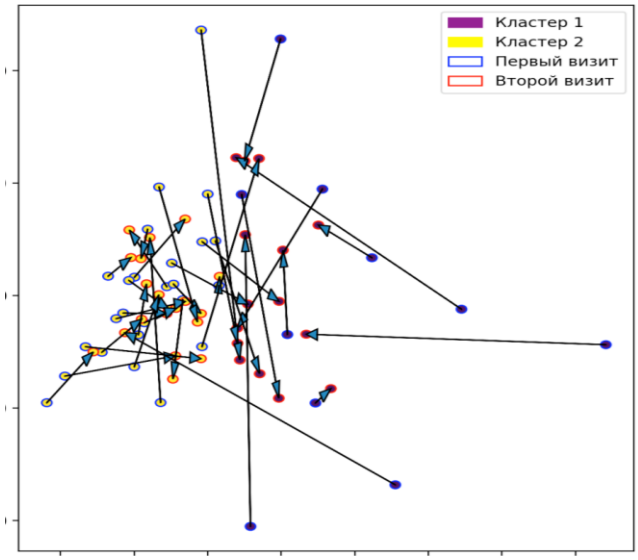

Рисунок 3. Динамика изменения характеристик пациентов между первым и вторым визитом.

В результате проверки корреляции параметров была получена корреляционная тепловая карта [12], изображенная на рисунке 4.

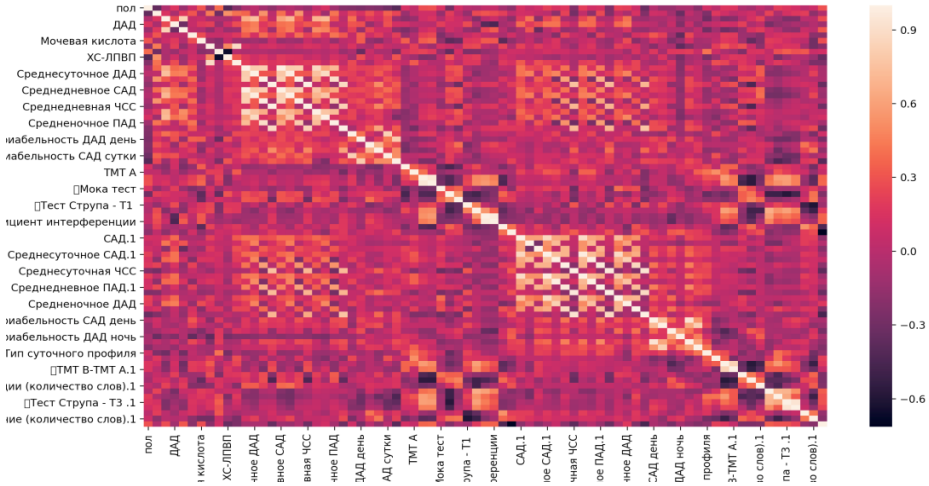

Рисунок 4. Тепловая карта корреляции характеристик пациентов.

Наиболее важной является визуализация результатов обследования пациентов при первом и втором визите, что позволит в дальнейшем выделить значимые для кластеризации параметры, на основе которых будет проводиться диагностика пациентов.

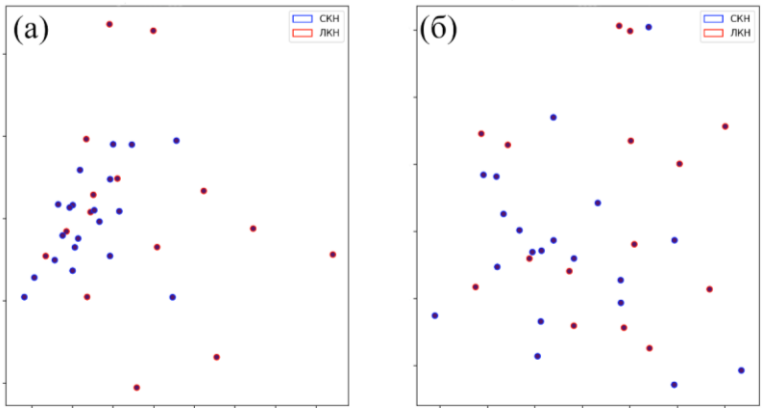

Рисунок 5. Характеристики пациентов в двухмерном пространстве. (а) — первый визит, (б) — второй визит.

5. Заключение

В результате проведенного исследования удалось провести системный анализ данных 35-ти пациентов. Применение методов визуализации и кластеризации позволило выявить следующие характерные особенности, присутствующие в исходных данных:

1. Пациенты с СКН в целом продемонстрировали схожие результаты. Это можно наблюдать на двумерном и трехмерном графиках, на которых большинство точек, отвечающих за СКН, располагаются в одной области.
2. При втором визите, пациенты с ЛКН проявили схожие результаты с пациентами с СКН, что может свидетельствовать о готовности пациентов проходить когнитивные тесты, из-за чего результаты могут оказаться недостоверными
3. Исходя из второго пункта, а также учитывая показатели точности и полноты для пациентов, можно заметить, что наиболее важным в исходных данных является первый визит. Данные второго визита являются малоинформативными.
4. Как и ожидалось, многие параметры исходной выборки коррелируют между собой, что было продемонстрировано на тепловой карте корреляции.

В дальнейшем планируется проверить полученные признаки на выборке большего размера, разработать алгоритм диагностики и прогнозирования течения СКН и ЛКН с целью оценки прогноза дальнейшего развития когнитивного нарушения, разработать структуру интеллектуальной системы прогнозирования рисков развития СКН и ЛКН, разработать рекомендации по оценке состояния когнитивных функций пациентов с когнитивными жалобами с целью оценки прогноза дальнейшего развития когнитивного нарушения, адаптировать алгоритмы ведения пациентов в зависимости от характера и причин когнитивного нарушения.

6. Благодарности

Исследование выполняется в рамках темы № 0071-2019-0001.

7. Литература

[1] Plassman, B.L. Prevalence of dementia in the United States: the aging, demographics, and memory study // Neuroepidemiology. – 2007. – Vol. 29(1-2). – P. 125-32. DOI: 10.1159/000109998.

[2] Petersen, R.C. Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology // Neurology. – 2018. – Vol. 90(3). – P. 126-135. DOI: 10.1212/WNL.0000000000004826.

[3] Коберская, Н.Н. Болезнь Альцгеймера: новые критерии диагностики и терапевтические аспекты в зависимости от стадии болезни // МС. – 2017. – № 10.

[4] Scarpina, F. The Stroop Color and Word Test / F. Scarpina, S. Tagini // Front Psychol. – 2017. – Vol. 8. – P. 557. DOI: 10.3389/fpsyg.2017.00557.
[5] Nasreddine, Z.S. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment / Z.S. Nasreddine, N.A. Phillips, V.R. Bã©dirian // Journal of the American Geriatrics Society. – 2005. – Vol. 53(4). – P. 695-699.
[6] Lloyd, S.P. Least squares quantization in PCM // Information Theory IEEE Transactions on. – 1982. – Vol. 28(2). – P. 129-137.
[7] Abdi, H. Principal component analysis / H. Abdi, L.J. Williams // WIREs Comput. Stat. – 2010. – Vol. 2(4). – P. 433-459. DOI: 10.1002/wics.101.
[8] Pedregosa, P. Scikit-learn: Machine Learning in Python // JMLR. – 2011. – Vol. 12. – P. 2825-2830.
[9] Hunter, J.D. Matplotlib: A 2D Graphics Environment // Computing in Science & Engineering. – 2007. – Vol. 9(3). – P. 90-95.
[10] Banerjee, A. Hypothesis testing, type I and type II errors / A. Banerjee, U.B. Chitnis, S.L. Jadhav, J.S. Bhawalkar, S. Chaudhury // Ind Psychiatry J. – 2009. – Vol. 18(2). – P. 127-131.
[11] Лурия, А.Р. Заучивание 10 слов // Альманах психологических тестов – М., 1995. – С. 92-94.
[12] Wilkinson, L. The History of the Cluster Heat Map / L. Wilkinson, M. Friendly // The American Statistician. – 2009. – Vol. 63. – P. 179-184. DOI: 10.1198/tas.2009.0033.

Data visualization and clustering in the task of system analysis of the patients examination results in the initial stages of cognitive impairment

B.R.Salem[1], V.I. Solodovnikov[1], V.N. Gridin[1]

[1]Center of Information Technologies in Engineering RAS, Marshal Biryuzov Str. 7a, Odintsovo, Russia, 143000

Abstract. There are several numbers of diseases and conditions within which cognitive impairment develops, which is a deterioration of cognitive functions compared to an individual norm. Currently, the causes and progression of cognitive impairment are poorly researched. This paper presents a system analysis of data including 35 patients and the results of their examination. For visualization, methods of data preprocessing, data dimensionality reduction, and clustering algorithms were used. Despite a small data set, 63% of the accuracy of diagnosis of cognitive impairment in the initial stages was achieved.