УДК: 004.942 DOI: 10.53816/23061456_2022_7–8_28

МЕТОДИЧЕСКИЙ ПОДХОД К РАСПОЗНАВАНИЮ СОСТОЯНИЯ СИСТЕМЫ УПРАВЛЕНИЯ НА ОСНОВЕ РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ПРОФИЛЕЙ ИНТЕНСИВНОСТИ ОБЪЕКТА

METHODOLOGICAL APPROACH TO RECOGNIZING THE STATE OF A CONTROL SYSTEM BASED ON SOLVING THE PROBLEM OF CLUSTERING OBJECT INTENSITY PROFILES

М.В. Розганов, Ф.Л. Шуваев

M.V. Rozganov, F.L. Shuvaev

ВКА им. А.Ф. Можайского

В данной статье приведен методический подход, применяемый при решении задач оценивания состояний сложных информационно-управляющих систем с учетом важности его признаков. В качестве признаков выбраны десять основных характеристик, описанных в теории временных рядов. Проведен сравнительный анализ методов определения важности признаков. В результате сравнительного анализа выбран наилучший, основанный на теории математической статистики, а именно — методе главных компонент. Апробация предложенных решений проведена на базе моделей временных рядов, полученные результаты свидетельствуют о необходимости анализа признакового пространства для повышения точности систем распознавания при решении задач оценивания состояния системы управления сложного объекта.
Ключевые слова: метод главных компонент, временной ряд, признак, вклад, важность.

This article presents a methodological approach used in solving problems of estimating the states of complex information and control systems, taking into account the importance of its features. Ten main characteristics described in the theory of time series are selected as features. A comparative analysis of methods for determining the importance of features has been carried out. As a result of a comparative analysis, the best one was chosen, based on the theory of mathematical statistics, namely, the method of principal components. Approbation of the proposed solutions was carried out on the basis of time series models, the results obtained indicate the need to analyze the feature space to improve the accuracy of recognition systems when solving problems of estimating the state of the control system of a complex object.
Keywords: principal component analysis, time series, feature, contribute, importance.

Введение

Определение режимов и способов функционирования известных объектов (явлений) наблюдения зачастую носит сложный характер ввиду отсутствия достаточного количества априорной информации об указанном объекте или неоднозначности их интерпретаций. Одним из возможных вариантов решения данной задачи может служить выявление профилей

интенсивности деятельности системы управления объекта исследования и соотнесение полученных результатов с вероятными режимами его функционирования. В рассматриваемом случае основным объектом наблюдения и анализа является временной ряд информационно-телекоммуникационной сети, которая лежит в основе исследуемой системы управления.

В настоящее время активно развивается научное направление анализа данных, базирующееся на положениях теории временных рядов и теории распознавания. Это обусловлено развитием вычислительных мощностей ЭВМ, появлением языков программирования высокого уровня и специального программного обеспечения, позволяющего эффективно исследовать сети. При этом под сетями подразумеваются сети различной природы: указанные информационно-коммуникационные сети (ИКС), сети взаимодействия белков в живых организмах, сети цитирования, социальные сети и др. [1, 2].

В теории распознавания одной из основных задач является отбор признаков, основанный на определении их важности, выявлении коллинеарных, что, в свою очередь, помогает лучшим образом понять структуру данных. В случае изобилия признаков нет понимания, какие же из них наиболее полно отражают топологические свойства сети. Следствием этого является высокая неопределенность распознавания исследуемых вариантов профилей. Таким образом, актуальность работы обусловлена необходимостью разработки такого методического подхода, который позволит обосновать состав признаков для решения задачи кластеризации профилей интенсивности информационно-телекоммуникационных сетей с целью оценивания состояния системы управления исследуемого объекта.

Постановка задачи

Система управления описывается в виде временного ряда или профиля интенсивности, который представляет собой собранный в отдельные моменты времени, с равными интервалами статистический материал о интенсивности работы наблюдаемого объекта. Множество таких профилей фиксированной длины является совокупностью наблюдений для задачи распознавания. В общем виде, при выборе в качестве объекта исследования какого-либо временного ряда, в состав характеризующих его признаков включают следующие характеристики, описанные в теории временных рядов [3–7]:

$$F = \begin{pmatrix} f_{\text{Стаб}}, f_{\text{КК}}, f_{\text{Ср}}, f_{\text{КХ}}, f_{\text{КП}}, \\ f_{\text{ЛМ}}, f_{\text{СКО}}, f_{\text{ПТ}}, f_{\text{СБ}}, f_{\text{Нел}} \end{pmatrix},$$

где F — множество характеристик временного ряда; $f_{\text{Стаб}}$ — стабильность временного ряда; $f_{\text{КК}}$ — коэффициент Квайтковски; $f_{\text{Ср}}$ — среднее временного ряда; $f_{\text{КХ}}$ — коэффициент Херста; $f_{\text{КП}}$ — коэффициент Филипса–Перрона; $f_{\text{ЛМ}}$ — локальные мотивы (англ. local motifs); $f_{\text{СКО}}$ — среднеквадратическое отклонение временного ряда; $f_{\text{ПТ}}$ — пересекающиеся точки; $f_{\text{СБ}}$ — коэффициент случайного блуждания; $f_{\text{Нел}}$ — коэффициент нелинейности.

Элементы множества F формируют образ рассматриваемого профиля интенсивности в пространстве параметров. В результате расчета F для N реализаций временных рядов получаем матрицу наблюдений $\mathbf{Y}_{[\mathbf{N,n}]} = \left\| \mathbf{f}_{\mathbf{ij}} \right\|_{\mathbf{N}}^{\mathbf{n}}$, где n — номер характеристики профиля.

Отбор наиболее важных признаков временного ряда предлагается осуществить на основе двух подходов: метода главных компонент (МГК) и расчета статистики Хопкинса [7] для комбинаций признаков, получаемых на основе прямого перебора. Данная статистика является одним из индикаторов тенденции данных к группированию. Значения статистики $H \geq 0,5$ говорят о том, что данные распределены случайно и равномерно. Значения статистики $H \leq 0,25$ на 90 % доверительном интервале указывают на имеющуюся тенденцию к группированию данных. Комбинация признаков, статистика Хопкинса для которой минимальна, будет считаться наилучшей для распознавания.

МГК представляет собой совокупность статистических приемов обработки данных, позволяющих сконцентрировать информацию, содержащуюся в исходном массиве данных, за счет перехода к меньшему числу наиболее информативных факторов — главных компонент (ГК) [7, 8].

В практической деятельности используется два ГК [9]. При этом существуют следующие

варианты использования результатов вычисления ГК:

– расчет вклада каждого наблюдаемого признака в ГК и их ранжирование по степени вклада;

– выявление мультиколлинеарных признаков, то есть имеющих тесную корреляционную взаимосвязь;

– разделение наблюдаемых признаков по группам (кластеризация).

Для расчета удельного вклада каждого признака (в данном исследовании — характеристики профиля) используется следующее выражение [9]:

$$\mathrm{Con}_i = \frac{f_1(y_i) f_2(y_i)}{\sum_{k=1}^{I} f_1(y_k) f_2(y_k)}, \ [i = 1(1)N],$$

где $f_1(y_i)$ и $f_2(y_i)$ — значение вклада i-го признака в ГК1 и ГК2.

Таким образом, значение удельного вклада Con (от англ. «contribution») является произведением вклада в ГК1 и ГК2 i-й характеристики профиля, нормированным по сумме вкладов остальных признаков профиля в ГК1 и ГК2.

Основные преобразования, осуществляемые в рамках методического подхода

На рис. 1 представлена структурная схема методического подхода по определению важности признаков различных профилей. Она состоит из 4 основных этапов, которые в свою очередь делятся на подэтапы. Рассмотрим их.

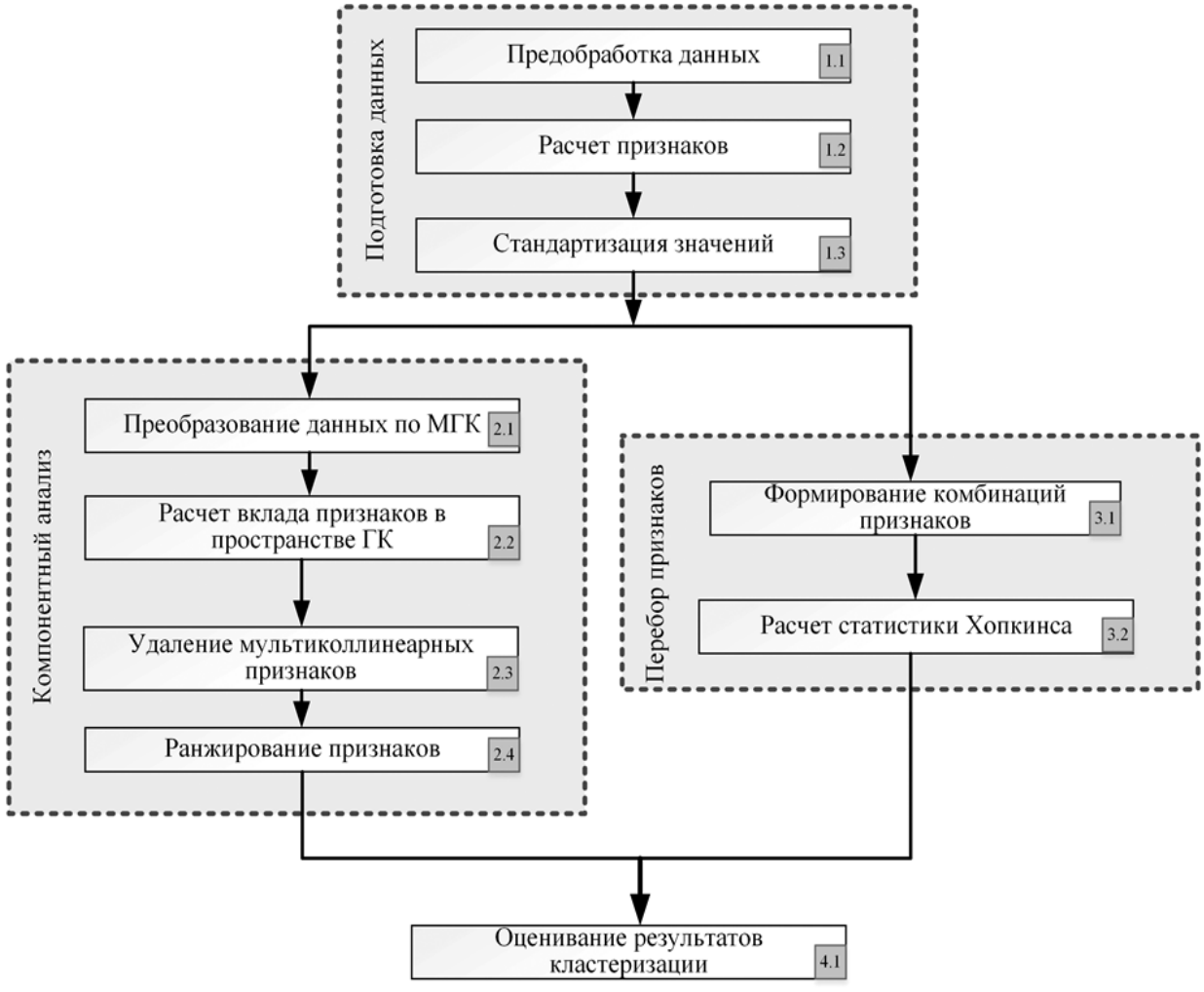


Рис. 1. Структурная схема методического подхода по определению важности признаков

Этап подготовки данных

1.1 Предварительная обработка данных. В настоящее время при решении задач машинного обучения и анализа данных около 30–40 % времени уходит на подготовку данных. Некорректно подготовленные данные могут существенно исказить результаты эксперимента.

1.2 Расчет признаков графов для каждой рассматриваемой модели (табл. 1). Полученные значения y_{ij} помещаются в матрицу наблюдений. При этом каждая строка матрицы соответствует одному исследуемому профилю.

1.3 Стандартизация значений. Результаты вычислений признаков имеют различную размерность, поэтому выполняется их нормировка по величине среднеквадратического отклонения для каждой меры.

Компонентный анализ

2.1 Осуществляются математические преобразования, согласно математическому аппарату МГК.

2.2 Производится расчет удельного вклада Con.

2.3 На этом подэтапе осуществляется идентификация и удаление коллинеарных значений, что позволяет оптимизировать признаковое пространство.

2.4 На подэтапе ранжирования признаки упорядочиваются по их удельному вкладу Con в формирование главных компонент.

Перебор признаков

3.1 Рассчитывается статистика Хопкинса для 1–5 комбинаций признаков.

Оценивание результатов кластеризации

Заключительный этап, на котором рассчитывается коэффициент Фоулкса–Мэллова (*T*) [9] для оптимальной комбинации признаков, выбранной на основе расчета Con и статистики Хопкинса.

Апробация методического подхода

Апробация осуществлена на базе моделей искусственно сформированных временных рядов. Исходные параметры рядов представлены в табл. 1, а их графическое представление по одному экземпляру моделируемого профиля — на рис. 2.

Целью апробации методического подхода является сравнительный анализ МГК и метода перебора информативных признаков, чтобы показать влияние отбора признаков на оценивание состояний системы.

На этап 1.1 поступает 300 моделей временных рядов (табл. 1), после расчетных этапов 1.2–1.3 формируется матрица наблюдений $\mathbf{Y}_{[150,10]}$. Далее над этой матрицей осуществляются преобразования 2.1–2.4, в ходе которых получены матрицы удельного вклада Con в ГК для признаков профиля ($n = 10$) и для самих наблюдений ($N = 150$).

Таблица 1

Исходные данные для моделирования

Класс модели/ характеристика	Длина ряда	Вид распределения	Параметр распределения	Компоненты временного ряда
РЯД_1	100	Нормальное	СКО = 5	Без тренда и сезонности
РЯД_2	100	Пуассона	λ = 0,9	Без тренда и сезонности
РЯД_3	100	Логнормальное	СКО = 0,9	Без тренда и сезонности
РЯД_4	100	Нормальное	СКО = 9	Линейный тренд с гетероскедастичностью без сезонности
РЯД_5	100	Нормальное	СКО = 5	Линейный тренд без сезонности
РЯД_6	100	Пуассона	λ = 3	Линейный тренд без сезонности

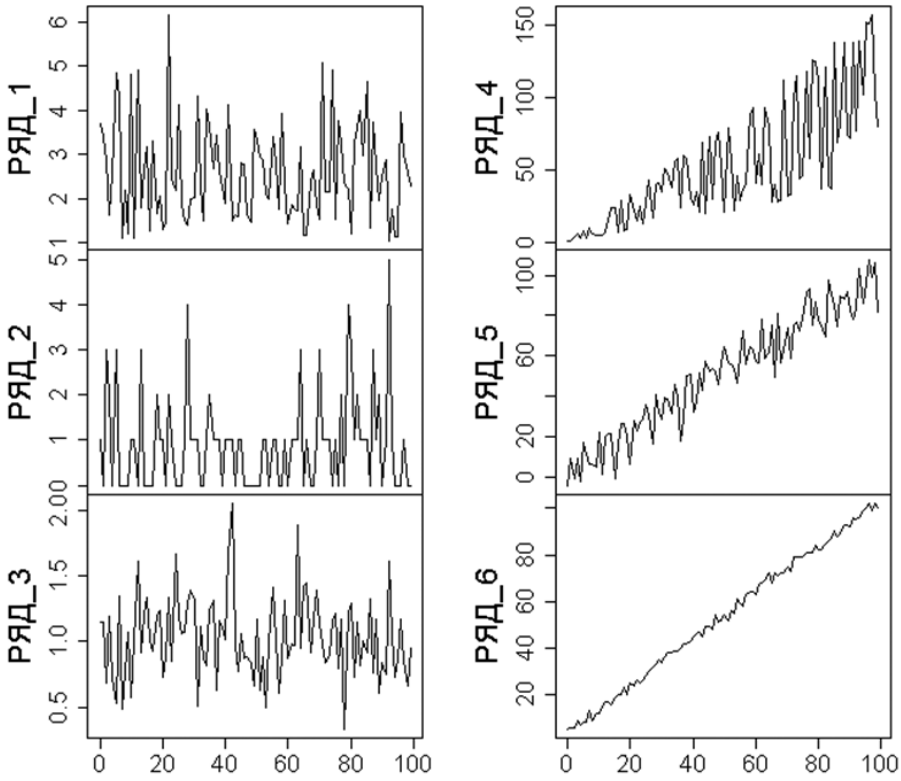


Рис. 2. Моделируемые временные ряды

На рис. 3 представлена диаграмма, показывающая ранжированный вклад признаков в ГК 1 и ГК 2.

Исходя из рис. 3 видно, что наибольший вклад вносит стабильность временного ряда, следующая по вкладу — коэффициент Квайтковски. Таким образом получены ранжированные по важности значения признаков, оценивающих свойства профилей интенсивности.

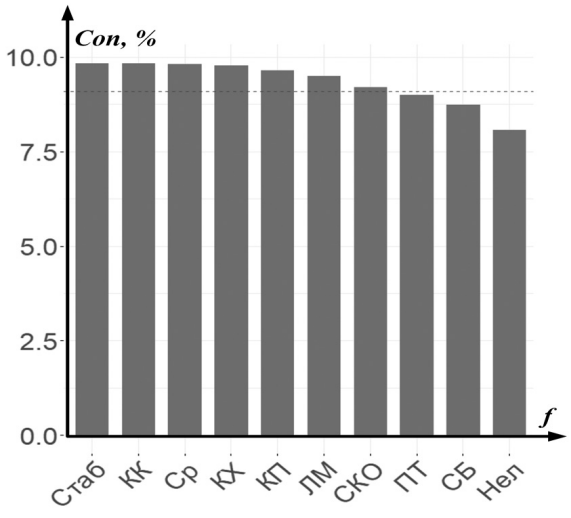


Рис. 3. Диаграмма вклада признаков в главные компоненты

Оценивание влияния важности признаков в задачах распознавания на примере решения задачи кластеризации

При апробации методического подхода была сформирована матрица наблюдений, состоящая из 6 различных по своей природе профилей интенсивности. Эти профили представлены различными видами распределения случайных величин. Таким образом, при решении задачи кластеризации теоретически возможно разделить их на классы. Но неясно — какие же признаки позволят наилучшим образом разделить смесь 6 классов моделей. Воспользуемся данными анализа на основе МГК и на основе расчета статистики Хопкинса [10].

Итак, проанализируем 300 наблюдений 6 моделей профилей интенсивности, представленных после расчета вклада Con и статистики

Хопкинса. При этом рассмотрим четыре варианта использования признаков:

– 1 вариант. Наблюдения представлены двумя признаками, с наибольшим вкладом Con, то есть Стаб и КК;

– 2 вариант. Наблюдения представлены двумя признаками с наименьшим значением статистики Хопкинса, то есть КХ и Ср;

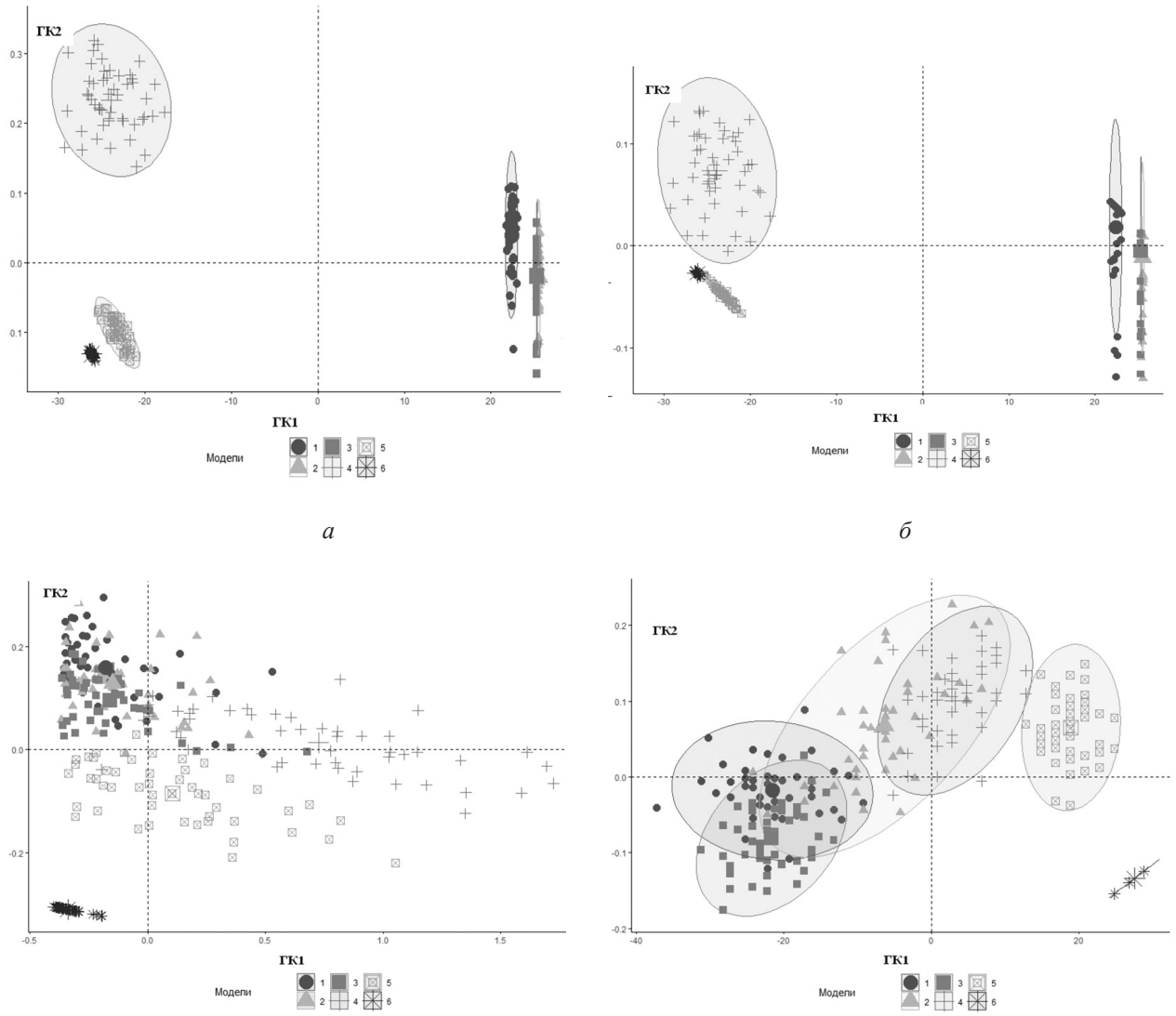
– 3 вариант. Наблюдения представлены двумя признаками с наименьшим вкладом Con, то есть СБ и Нел.

– 4 вариант. Наблюдения представлены двумя признаками с наибольшим значением статистики Хопкинса, то есть ПТ и СБ.

Результаты наблюдений представлены в пространстве первой и второй главной компонент по вариантам на рис. 4 (*а–в*). Характерно, что при переборе признаков наилучшими и наихудшими значениями статистики Хопкинса обладают комбинации из двух признаков.

Исходя из визуального анализа рис. 4 видно, какое влияние оказывает на разрешающую способность использование различных комбинаций признаков. Наилучшая способность к разделению данных достигается путем использования наиболее важных признаков, худшая — наименее важных.

Осуществим иерархическую кластеризацию данных для шести классов и измерим ка-



а *б*

в *г*

Рис. 4. Представление наблюдений в пространстве ГК1 и ГК2:
а — 1 вариант; б — 2 вариант; в — 3 вариант; г — 4 вариант

Таблица 2

Оценка результатов кластеризации и тенденции данных к группированию

Варианты использования признаков	*T*
1 вариант	0,89
2 вариант	0,83
3 вариант	0,11
4 вариант	0,09

чество кластеризации на основе коэффициента Фоулкса-Мэллова (*T*) [9, 10]. Коэффициент Фоулкса–Мэллова — это мера сходства между двумя результатами кластеризации (в этом исследовании: между истинными значениями классов, полученных на этапе моделирования (реальных данных, и значениями классов, полученных в результате кластеризации). Результаты измерений представлены в табл. 2.

Представленные в табл. 2 результаты свидетельствуют о влиянии выбора признака на результаты кластеризации, а равно и иной задачи распознавания, например классификации и регрессии. Полученные результаты для варианта 3, когда выбраны два наиболее важных признака, показывают, что данные четко сгруппированы, и кластеризация при таком их расположении происходит без ошибок. Использование наименее важных признаков приводит к низкой группировке данных и высоким ошибкам кластеризации. Кроме того, проведенный эксперимент показывает преимущество МГК перед методом перебора признаков (T = 0,89 против 0,83).

Выводы

Данное исследование показало, что при решении широкого спектра задач кластеризации профилей интенсивности, важным аспектом является выбор признакового пространства. Признаковое пространство выбирается на основе анализа важности признаков, в результате которого удаляются коллинеарные признаки, а остальные ранжируются на основе их удельного вклада. Проведенный эксперимент по кластеризации профилей подтвердил теоретические результаты анализа признаков, что подтверждает работоспособность методического подхода.

Разработанный методический подход может быть использован в ходе решения задач распознавания структур информационно-телекоммуникационных сетей с целью оценивания состояния системы управления объекта исследования, а также в других отраслях народного хозяйства.

Литература

1. Шуваев Ф.Л., Татарка М.В. Анализ динамики мер центральности математических моделей случайных графов // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 2. С. 249–256.

2. Шуваев Ф.Л., Татарка М.В. Анализ математических моделей случайных графов, применяемых в имитационном моделировании информационно-коммуникационных сетей // Вестник Санкт-Петербургского университета ГПС МЧС России. 2020. №2. С. 67–77.

3. https://ranalytics.github.io/tsawithr

4. Hyndmanand R., Athanasopoulos G. Forecasting: Principlesand Practice. OTexts, 3rd edition. 2019. 41 p.

5. Liao T.W. Clustering of time series dataasurvey. Pattern Recognition, 38. 1857–1874. 2005.

6. Scott S.L., Varian H.R. Predicting the present with bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation. 5 (1/2). 4–23. 2014.

7. Юсупов Р.М., Петухов Г.Б., Сидоров В.Н., Городецкий В.И., Марков В.М. Статистические методы обработки результатов наблюдений: Монография // Под ред. Р.М. Юсупова. — М.: МО СССР, 1984. 786 с.

9. Кобзарь А.И. Прикладная математическая статистика. — М.: Физматлит, 2012. 813 с.

9. Еремеев И.Ю., Татарка М.В., Шуваев Ф.Л., Цыганов А.С. Анализ мер центральности узлов сетей на основе метода главных компонент // Информатика и автоматизация. 2020. № 19 (6). С. 1307–1331.

10. Kassambara A. Practical guide to cluster analysis in R // STDHA. 2017. 187 p.

References

1. Shuvaev F.L., Tatarka M.V. Analysis of the dynamics of measures of centrality of mathematical models of random graphs // Scientific and technical bulletin of information technologies, mechanics and optics. 2020. V. 20. № 2. Pp. 249–256. (In Russ.).

2. Shuvaev F.L., Tatarka M.V. Analysis of mathematical models of random graphs used in the simulation of information and communication networks // Bulletin of St. Petersburg University State Fire Service of the Ministry of Emergencies of Russia. 2020. №2. pp. 67–77. (In Russ.).

3. https://ranalytics.github.io/tsawithr

4. Hyndmanand R., Athanasopoulos G. Forecasting: Principlesand Practice. OTexts, 3rd edition. 2019. 41 p.

5. Liao T.W. Clustering of time series dataasurvey. Pattern Recognition. 38. 1857–1874. 2005.

6. Scott S.L., Varian H.R. Predicting the present with bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation. 5 (1/2). 4–23. 2014.

7. Jusupov R., Petuhov G., Sidorov V., Gorodeckij V., Markov V. Statistical Methods for Processing Observation Results: Monograph // Edited by R.M. Yusupov. — M.: Ministry of Defense of the USSR, 1984. 786 p.

8. Kobzar' A. Applied Mathematical Statistics. — M.: Fizmatlit, 2012. 813 p.

9. Eremeev I., Tatarka M., Shuvaev F., Cyganov A. Comparative analysis of centrality measures of network nodes based on principal component analysis, Informatics and automation. 2020. Vol. 19 (6). P. 1307–1331.

10. Kassambara A. Practical guide to cluster analysis in R // STDHA. 2017. 187 p.