УДК 004.421.2+005

ЗНАКОВЫЙ АЛГОРИТМ КЛАССИФИКАЦИИ СОСТОЯНИЙ БИОЦЕНОЗА

Ломакина Л.С., Ломакин Д.В., Блажнов И.Д., Пожидаева А.С.

Нижегородский государственный технический университет им. Р.Е.Алексеева (603950, ГСП-41, Н. Новгород, ул. Минина, д. 24), e-mail: llomakina@list.ru

Разработан знаковый алгоритм классификации состояний биоценоза, под которым в работе понимается совокупность микроорганизмов разного вида, населяющих желудочно-кишечный тракт (ЖКТ) человека. Состояние биоценоза описывается количеством микроорганизмов каждого из видов, которые в совокупности образуют многомерное пространство состояний биоценоза. Задача заключается в определении, какому из подмножеств в пространстве состояний биоценоза принадлежит состояние диагностируемого пациента. Подмножества формируются априорно и представляют собой выборки соответственно из здоровых и больных людей. Решение о состоянии пациента принимается на основании частных решений, каждое из которых принимается посредством сравнения двух признаков, один из которых принадлежит пациенту, а другой является признаком объектов, принадлежащих выбранному множеству. Сравнение осуществляется посредством вычисления разности между значениями выбранных признаков, при этом учитывается только знак разности. Окончательное решение принимается в пользу того множества, в пользу которого принято максимальное количество частных решений. Результаты экспериментальных исследований подтвердили более высокую эффективность разработанного алгоритма классификации по сравнению с традиционными методами и устойчивость к изменениям априорно неизвестных характеристик пространства состояний. Эффект достигается за счет учета скрытых структурных свойств пространства состояний.

Ключевые слова: знаковый алгоритм, классификация, биоценозы, скрытые параметры.

SIGN CLASSIFICATION ALGORITHM OF BIOCENOSIS STATE

Lomakina L.S., Lomakin D.V., Blazhnov I.D., Pozhidaeva A.S.

Nizhny Novgorod State Technical University n.a. R.E.Alekseev, Russia (603950, Nizhny Novgorod, street Minina, 24), e-mail: llomakina@list.ru

The sign classification algorithm of biocenosis state has been developed. It means in this article as aggregation of microorganisms of different species which inhabit a human large intestine tract. The condition of a biocenosis is described by number of all microorganisms of each species which together form a multidimensional state space of biocenosis. It should be determined which subset in the multidimensional space of biocenosis states contains a state of diagnosed patient. The subsets are formed a priori and there are sample of healthy and sick people. The patient's condition is determined on the basis of particular decisions. Every decision is made as a matching of two parameters, one of which belongs to some patient and the other is a object's parameter, which belongs to chosen set. The matching of parameters is made as calculation of subtraction between values of chosen parameters. Only the subtraction sign is considered. The final decision is made in favor of that set in favor of which the maximum quantity of particular decisions is accepted. Results of the experimental researches confirmed higher performance of the developed algorithm of classification in comparison with traditional methods and resistance to changes of a priori unknown characteristics of the state space. The effect is achieved by taking into account the hidden structural properties of the state space.

Keywords: sign algorithm, classification, biocenosis, latent parameters.

В настоящее время активно внедряются информационные технологии во все сферы научных и прикладных исследований, что позволяет получить новые научные результаты и избежать рутинной работы при обработке больших объемов различного рода экспериментальных данных. В связи с этим возникла необходимость разработки эффективных методов анализа и обработки многомерных экспериментальных данных, которые доставляют информацию об окружающей нас среде. В частности, информационные технологии нашли широкое применение в биологических и медицинских исследованиях.

Они успешно внедряются в сферу здравоохранения с целью повышения эффективности обработки результатов анализов и автоматизации процесса принятия решения о состоянии здоровья человека. В данной работе разработан алгоритм классификации состояний диагностируемого пациента на основе априорно заданных множеств, которые представляют собой репрезентативные выборки соответственно из здоровых и больных людей. Множества строятся на основании экспертных оценок и методов кластеризации. Классификация заключается в определении, какому из множеств принадлежит состояние диагностируемого пациента. Индикатором состояния здоровья человека является качественный и количественный состав микрофлоры желудочно-кишечного тракта (ЖКТ), которая является частным случаем биоценоза, который представляет собой некоторую системно-организованную совокупность растений, животных или микроорганизмов, обитающих в определённой среде. Состояние микрофлоры описывается совокупностью значений информативных признаков, в качестве которых используется количество микроорганизмов данного вида. Упорядоченную последовательность значений выбранных признаков будем рассматривать как вектор в некотором метрическом пространстве. В настоящее время разработано большое количество методов классификации, которые в основном различаются выбранной метрикой пространства. К сожалению, использование традиционных метрик не позволяет полностью раскрыть структурные свойства объекта [2,3]. Поэтому возникает необходимость в поиске таких скрытых признаков (параметров), которые обладают максимальной информацией, необходимой для решения поставленной задачи. С целью выявления более глубоких структурных свойств и их использования при синтезе правил принятия решения в работе предлагается использовать знаковые алгоритмы, которые обладают сравнительно высокой эффективностью и устойчивостью по отношению к изменениям априорно неизвестных характеристик многомерных данных. В силу указанных причин выбранное направление исследований является актуальным.

Модель и описание метода классификации

Базовая математическая модель многомерных экспериментальных данных представляет собой n-мерное векторное пространство, в котором отдельный вектор $\boldsymbol{\xi}=(\xi_1, \xi_2, \ldots, \xi_n)$ описывает состояние одного из объектов, образующих репрезентативную выборку. Координатами вектора являются значения n признаков, характеризующих состояние объекта, в нашем случае, состояние желудочно-кишечного тракта, а, следовательно, и состояние самого человека. Значения признаков представляют собой результаты медицинских анализов ЖКТ человека.

Предметом исследования являются структурные свойства базовой модели, на основании которых выбирается структурная модель многомерных данных, которая должна

не только адекватно отображать структуру базовой модели, но и соответствовать цели решаемой задачи. Цель решаемой задачи заключается в нахождении эффективных алгоритмов классификации, то есть в определении принадлежности диагностируемого пациента к одному из множеств. В зависимости от выбранной структурной модели будут изменяться эффективность и достоверность классификации состояний биоценоза ЖКТ. В данной работе предлагается использовать знаковый алгоритм классификации многомерных данных, который раскрывает более глубокие структурные свойства многомерных данных по сравнению с известными, сохраняя при этом устойчивость к изменениям априорных характеристик многомерных данных.

Основное содержание метода классификации заключается в следующем. Окончательное решение принимается на основании частных решений. Решение о принадлежности пациента одному из множеств принимается в пользу того множества, в пользу которого принято максимальное количество частных решений. Частные решения принимаются посредством сравнения расстояний между значением j-го признака пациента и значениями i-го признака индивидов, принадлежащих соответственно первому и второму множествам. Все частные решения можно записать в виде матрицы, размером ($n \times n$), где n – количество признаков, характеризующих состояние индивида. Поскольку количество значений i-го признака в множестве равно количеству в нем индивидов, то в качестве расстояния выбрана оценка вероятности превышения значениями i-го признака индивидов порога, равного значению j-го признака пациента. Поскольку оценка вероятности посредством вычисления относительной частоты появления события является непараметрической, то и алгоритм в целом является непараметрическим, устойчивым к изменениям характеристик многомерных данных.

Описание алгоритма классификации

Исходные данные представляют собой результаты бактериологических исследований состояний биоценозов здоровых и больных людей. Состояние отдельного индивида из множества описывается вектором $\boldsymbol{\xi}=(\xi_1, \xi_2, \ldots, \xi_n)$ в n-мерном пространстве, а вектором $\boldsymbol{\eta}=(\eta_1, \eta_2, \ldots, \eta_n)$ – состояние диагностируемого пациента. Координатами векторов являются значения количества микроорганизмов данного вида. Расстояние от значения j-го признака пациента до значений i-х признаков индивидов множества вычисляется как отношение количества индивидов, у которых значение i-го признака превысило порог, равный значению j-го признака пациента, к количеству индивидов в множестве.

Блок схема разработанного алгоритма представлена на рис. 2, при этом приняты следующие обозначения. Множества здоровых и больных людей обозначены соответственно через **X** и **Y**; x – номер выбранного индивида в множестве **X**, y – номер

выбранного индивида в множестве **Y**. Количество индивидов в множестве **X** обозначим как m_x, количество индивидов во множестве **Y** обозначим m_y. Введем величину I, которая равна количеству признаков у каждого индивида множеств **X** и **Y**. Количество признаков у пациента обозначим J. Номер признака индивида из множества обозначим i, номер признака диагностируемого пациента j.

Результаты экспериментальных исследований

С целью проверки эффективности разработанного знакового алгоритма классификации состояний биоценозов была сделана тестовая выборка из данных бактериологических исследований качественного и количественного состава микрофлоры желудочно-кишечного тракта 50 индивидов, подлежащих диагностированию. Данная выборка индивидов на основании экспертных оценок была разделена на 2 группы по 25 человек, соответственно больных и здоровых. В качестве априорных данных использовались результаты бактериологических исследований 100 человек, которые аналогичным образом были разделены на здоровых и больных индивидов. Эффективность разработанного алгоритма сравнивалась с эффективностью классического метода, в котором использовалась евклидова метрика для измерения расстояния от пациента до множества. Это расстояние вычисляется как среднее расстояние от пациента до индивидов множества. Эффективность разработанного метода сравнивалась с эффективностью алгоритма классификации состояний объекта с использованием весовых коэффициентов, при вычислении среднего расстояния [4]. Разработанный знаковый алгоритм классификации был отдельно протестирован для трех случаев: $i=j$, $i\neq j$, и когда i и j могут свободно изменяться от 0 до n, где n – количество признаков, характеризующих состояние индивида.

Метод с использованием весовых коэффициентов основывается на введении степени принадлежности какого-либо состояния объекта данному множеству. В данном алгоритме степень принадлежности предлагается оценивать посредством сравнения состояния объекта с остальными состояниями этого множества. В методе с использованием весовых коэффициентов предлагается вычислять функцию принадлежности по формуле:

$$\mu(\mathbf{x_i}) = \frac{1-\delta_i}{\sum_{i=1}^{n-1}(1-\delta_i)} = \frac{1-\delta_i}{n-1},\ \delta_i = \frac{\rho_i}{\sum_{i=1}^{n-1}\rho_i}$$

где ρ_i – среднее евклидово расстояние между вектором $\mathbf{x}_i$ и всеми остальными $\mathbf{x}_j$ векторами $(j \neq i)$ множества $\mathbf{X}$, а δ_i его нормированное значение.

Полученные значения характеристических функций используются в качестве весовых коэффициентов при вычислении среднего расстояния от состояния диагностируемого

объекта соответственно до множеств **X** и **Y**. Результаты проведенных экспериментов представлены в таблице 1, графически результаты представлены на рис. 1.

Таблица 1

Результаты экспериментальных исследований

Название метода	Количество правильных результатов	Количество неправильных результатов	Количество правильных результатов в %
Классический алгоритм классификации	39	11	78 %
Алгоритм с использованием весовых коэффициентов	40	10	80 %
Знаковый алгоритм, $i=j$	41	9	82 %
Знаковый алгоритм, $i\neq j$	41	9	82 %
Знаковый алгоритм, i и j могут свободно изменяться от 0 до n	42	8	84 %

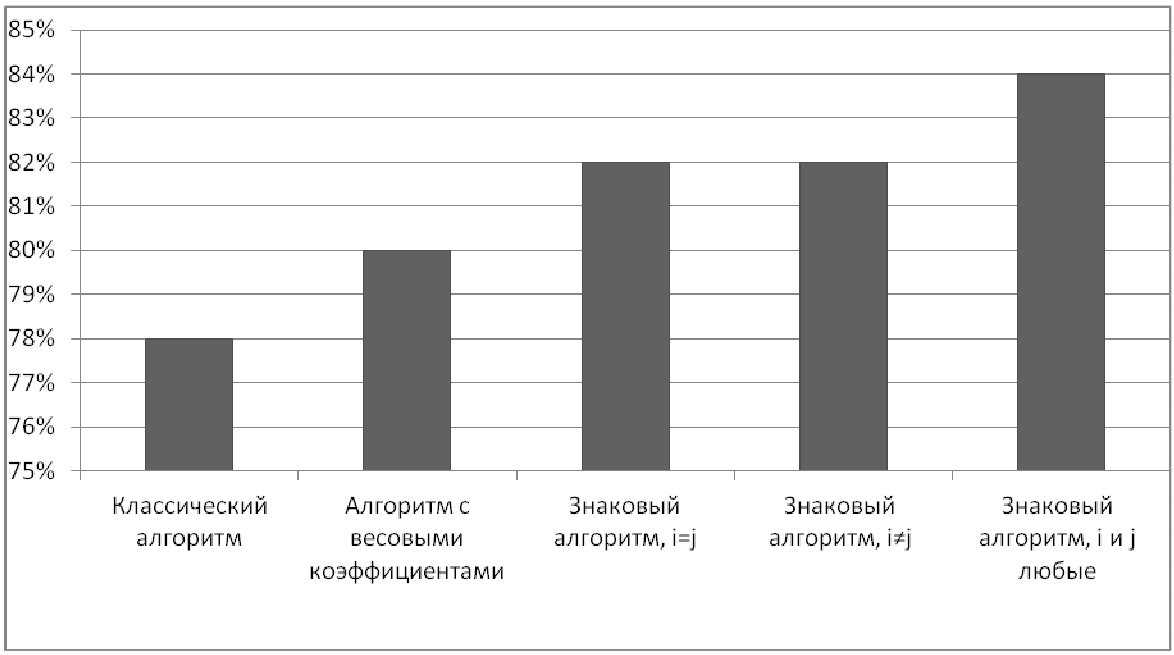

Рис. 1. Диаграмма экспериментальных исследований

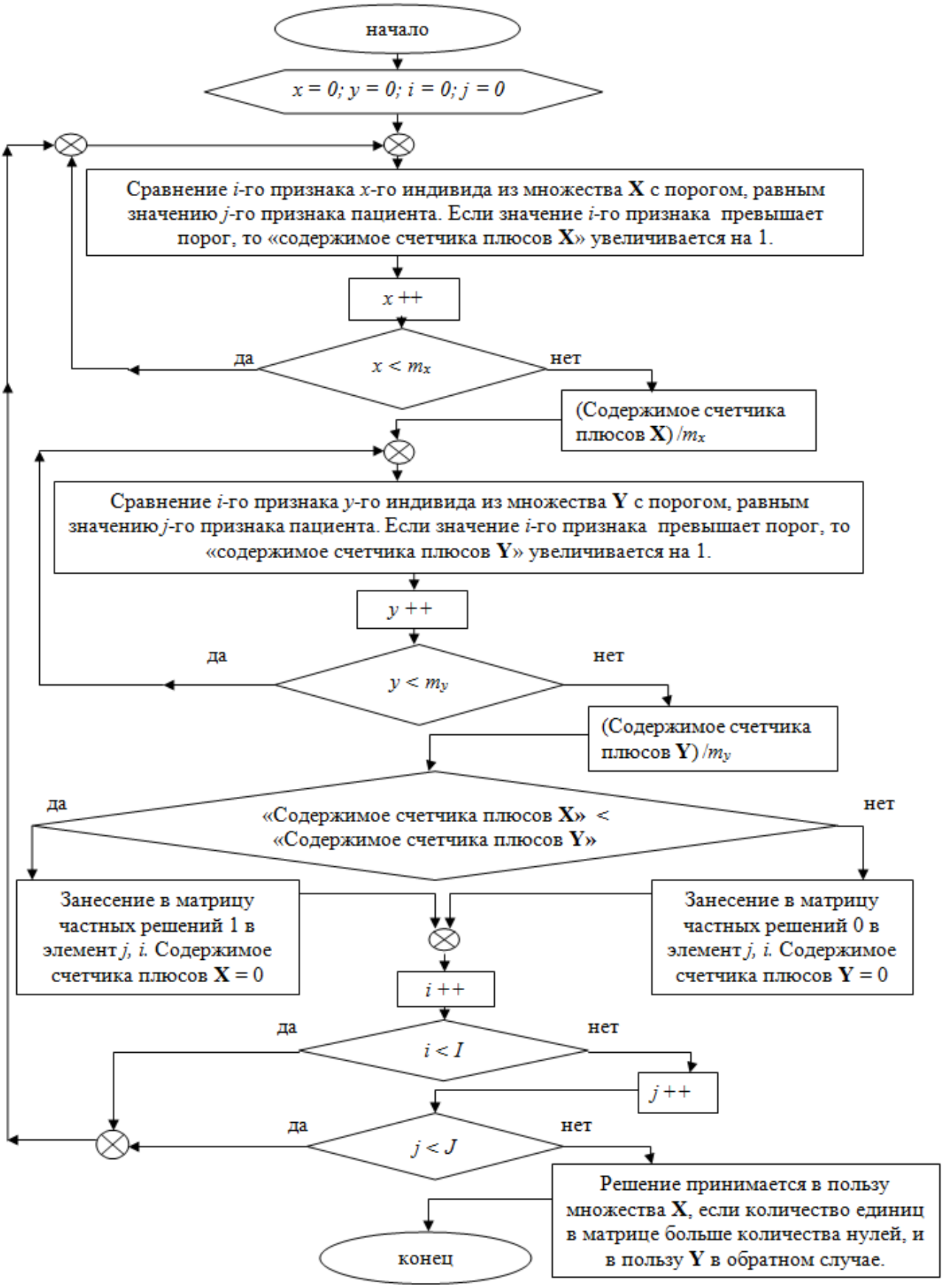

Рис. 2. Блок-схема знакового алгоритма классификации состояний биоценозов

При решении задачи классификации с применением классического алгоритма было получено 78% верных результатов. При использовании разработанного знакового алгоритма,

в котором i и j могут изменяться от 0 до n, на тех же данных было получено 84 % верных результатов, что доказывает увеличение эффективности классификации. При этом эффективность знакового алгоритма при условиях $i=j$ и $i\neq j$ составила 82 %, что является более успешным результатом, как в сравнении с классическим алгоритмом, так и с методом, в котором используются весовые коэффициенты.

Выводы

Разработанный знаковый алгоритм, программная реализация которого подтверждена свидетельством о государственной регистрации программы для ЭВМ №2015611346 от 28.01.2015 г., апробирован на конкретных примерах и при этом обнаружил более высокую эффективность как по сравнению с классическим методом, так и по сравнению с методом, основанным на введении весовых коэффициентов [5]. Эффект достигнут благодаря раскрытию скрытых свойств многомерных данных. Подтверждена устойчивость знаковых алгоритмов по отношению к изменениям априорно неизвестных характеристик многомерных данных.

Список литературы

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.
2. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов: учебное пособия для вузов. – М.: Горячая линия - Телеком, 2007. – 520 с.
3. Горелик А.Л., Скрипкин В.А. Методы распознавания: учебное пособие для вузов. – М.: «Высшая школа», 1977. – 222 с.
4. Ломакин Д.В., Блажнов И.Д. Алгоритм классификации состояний объекта на основе априорных многомерных данных. // Материалы XX Международной научно-технической конференции «Информационные системы и технологии» ИСТ-2014, 322.
5. Ломакина, Л.С. Программа классификации состояний биоценозов с использованием знаковых алгоритмов / Л.С. Ломакина, Д.В. Ломакин, И.Д. Блажнов, А.С. Пожидаева // Свидетельство об официальной регистрации программы для ЭВМ № 2015611346. Зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности РФ (Роспатент) от 28 января 2015 г.
6. Эсбенсен К. Анализ многомерных данных. Избранные главы / пер. с англ. С.В. Кучерявского; под ред. О.Е. Родионовой. – Черноголовка: ИПХФ РАН, 2005. – 160 с.

Рецензенты:

Баландин Д.В., д.ф.-м.н., профессор, заведующий кафедрой численного и функционального анализа. Нижегородский государственный университет им. Н.И. Лобачевского, Национальный исследовательский университет, г. Нижний Новгород;
Федосенко Ю.С., д.т.н., профессор, заведующий кафедрой «Информатика, системы управления и телекоммуникации» Волжской государственной академии водного транспорта, г. Нижний Новгород.