ISSN 2225-6016

ВЕСТНИК

Смоленской государственной медицинской академии

Том 18, №3

2019

УДК 519 253

СПОСОБ ОЦЕНКИ ИНФОРМАТИВНОСТИ ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ В МЕДИЦИНЕ И ФАРМАКОЛОГИИ

© Лямец Л.Л., Евсеев А.В.

Смоленский государственный медицинский университет, Россия, 214019, Смоленск, ул. Крупской, 28

Резюме

Цель. Оценка информативности диагностических признаков необходима для их объективного ранжирования по степени важности и определения порядка их рассмотрения в процессе постановки диагноза. Для признаков, которые являются непрерывными величинами с неизвестными функциями распределения вероятностей, оценка информативности по результатам выборочных исследований является сложной задачей. Цель теоретического исследования заключалась в разработке способа, который на основе индуктивных выводов, полученных в результате статистического исследования, позволяет произвести оценку их информативности на основе применения известных математических конструктов. Способ рекомендуется для оценки результатов как медико-биологических исследований, так и клинических данных при дифференциальной диагностике заболеваний, а также идентификации состояний, вызванных фармакологическим воздействием.

Методика. Проведен обзорный анализ публикаций по вычислительным диагностическим методам, основанным на статистических методах анализа экспериментальных медицинских данных. Описаны особенности признаков, являющихся непрерывными физическими величинами, для количественного представления которых используются действительные числа. Выявлены проблемные вопросы оценки информативности непрерывных признаков. Исследованы математические конструкты, которые на основании пороговых значений позволяют осуществить категоризацию непрерывных признаков по порогу и привести их к дихотомической порядковой шкале. Рассмотрено применение информационной меры Кульбака для оценки информативности категоризированных дихотомических признаков.

Результаты. Разработан способ оценки информативности непрерывных признаков. В основе способа лежит категоризация непрерывных признаков по пороговому значению, которое вычисляется на основании анализа эмпирических кумулятивных функций (ЭКФ). Оптимальное пороговое значение вычисляется для максимального модуля разности между исследуемыми ЭКФ. Пороговые значения диагностических признаков представляют самостоятельное научное и практическое значение для дифференциальной диагностики нозологических форм. Для выявления статистически значимого различия между ЭКФ показано применение критерия Колмогорова-Смирнова. Для выявления статистически значимого различия между категоризированными по пороговому значению признаками использовался критерий хи-квадрат (Пирсона) и точный критерий Фишера (для таблиц сопряженности 2×2). Показана возможность применения этих статистических критериев для обоснования оптимального порогового значения для категоризации непрерывных признаков. Для оценки информативности признаков в разработанном способе использовалась информационная мера Кульбака.

Заключение. В результате теоретического исследования разработан способ оценки информативности непрерывных диагностических признаков. В основе способа лежит категоризация непрерывных признаков по оптимальному пороговому значению. Категоризация позволяет преобразовать непрерывный признак в порядковый дихотомический признак с двумя упорядоченными градациями. Это преобразование позволяет применить для оценки информативности диагностических признаков информационную меру Кульбака. Разработанный способ является информационной технологией, позволяющей на основании первичных статистических данных производить отбор признаков для дифференциальной диагностики на основе оценки их информативности. Способ оценки информативности может представлять практический интерес для научных работников, осуществляющих исследования в области доказательной медицины на основе вычислительных диагностических методов и использующих в своей работе статистические методы анализа экспериментальных данных и принятия решений.

Ключевые слова: непрерывные случайные величины, эмпирическая кумулятивная функция, категоризация непрерывных величин, пороговые значения, информативность признаков, вычислительные диагностические методы, дифференциальная диагностика

METHOD OF EVALUATING THE INFORMATIVENESS OF DIAGNOSTIC FEATURES IN MEDICINE AND PHARMACOLOGY

Lyamec L.L., Evseev A.V.
Smolensk State Medical University, 28, Krupskoj St., 214019, Smolensk, Russia

Abstract

Objective. Assessment of the information content of diagnostic signs is necessary for their objective ranking by importance and determining the order of their consideration in the process of diagnosis. For the signs that are continuous quantities with unknown probability distribution functions, evaluating the information content based on the results of sample studies is a difficult task. The purpose of the theoretical study was to develop a method that, based on inductive conclusions obtained as a result of statistical research, allows us to evaluate their information content based on the use of well-known mathematical constructs. The method is intended for use in biomedical research and in practical medical activities related to the differential diagnosis of diseases, as well as the identification of conditions caused by pharmacological effects.

Method. A review of publications on computational diagnostic methods based on statistical methods for analyzing experimental medical data was carried out. Peculiarities of features that are continuous physical quantities are described, for the quantitative representation of which real numbers are used. The problematic issues of evaluating the information content of continuous signs are identified. Mathematical constructs are studied, which, based on threshold values, allow the categorization of continuous features by a threshold and bring them to a dichotomous ordinal scale. The application of the Kullback information measure for evaluating the information content of categorized dichotomous features is considered.

Results. A method for evaluating the informative value of continuous features was developed. The method is based on the categorization of continuous features by a threshold value, which is calculated based on the analysis of empirical cumulative functions (ECF). The optimal threshold value is calculated for the maximum modulus of the difference between the studied ECF. The threshold values of diagnostic features are of an independent scientific and practical value for the differential diagnosis of nosological forms. To identify a statistically significant difference between ECF, the Kolmogorov-Smirnov criterion is shown. To identify a statistically significant difference between the criteria categorized by a threshold value, the chi-square test (Pearson) and the Fisher exact test were used (for contingency tables 2×2). The possibility of applying these statistical criteria to substantiate the optimal threshold value for the categorization of continuous features is shown. To assess the information content of the signs in the developed method, the Kullback information measure was used.

Conclusion. As a result of theoretical research, a method was developed for evaluating the informative value of continuous diagnostic signs. The method is based on the categorization of continuous features by the optimal threshold value. Categorization allows to convert a continuous sign into an ordinal dichotomous sign with two ordered gradations. This transformation makes it possible to use the Kullback information measure to evaluate the information content of diagnostic signs. The developed method is an information technology that allows, on the basis of primary statistical data, to select features for differential diagnosis based on an assessment of their information content. The method of evaluating information content may be of practical interest for scientists carrying out research in the field of evidence-based medicine based on computational diagnostic methods and using statistical methods in their work to analyze experimental data and make decisions.

Keywords: continuous random variables, empirical cumulative function, categorization of continuous variables, threshold values, informative features, computational diagnostic methods, differential diagnostics

Введение

В наиболее простом случае диагностика представляет собой задачу распознавания двух состояний организма: в норме и при заболевании, которое однозначно определяется соответствующими патогномоничными признаками (симптомами). Особенность рассматриваемой в данном исследовании диагностической задачи состоит в том, что патогномоничные признаки заболевания в период обследования больного еще не проявились или по различным причинам не могут быть измерены и определены. Вместе с тем объективная реальность такова, что при обследовании больного имеется возможность измерять или оценивать диагностические признаки, которые в отличие от прямых (патогномоничных) признаков являются косвенными (непрямыми) и имеют

вероятностный характер, как в норме, так и при диагностируемом заболевании. Поэтому косвенные диагностические признаки (КДП) целесообразно рассматривать как случайные величины. Условия, при которых диагностическое решение принимается на основании КДП при отсутствии прямых (патогномоничных) признаков, в дальнейшем будем называть условиями неопределенности. При постановке диагноза в условиях неопределенности возможно возникновение диагностических ошибок первого рода (гипердиагностика) и ошибок второго рода (гиподиагностика). Для уменьшения неопределенности и снижения вероятностей диагностических ошибок можно использовать закономерности (знания) индуктивного характера. Они могут быть получены в результате специально спланированных выборочных статистических исследований и анализа статистических распределений КДП в норме и при диагностируемом заболевании. Эти знания в дальнейшем позволяют разработать дедуктивные диагностические критерии (правила) постановки диагноза в условиях неопределенности, практическое применение которых направлено на уменьшение вероятностей ошибок первого и второго рода. В качестве индуктивных закономерностей, полезных для диагностики, можно использовать значимые статистические различия между эмпирическими кумулятивными функциями, вычисленными для КДП в норме и при определенном заболевании. Также можно сказать, что полезные для диагностики знания могут содержаться в статистических показателях, характеризующих значимые различия между соответствующими эмпирическими кумулятивными функциями (ЭКФ). Косвенные диагностические признаки, у которых ЭКФ в норме и при диагностируемом заболевании значимо различаются, могут быть включены в диагностический критерий. Для этого необходимо оценить их информативность. Следовательно, для включения определенного КДП в диагностический критерий необходимо сначала выявить значимые различия между ЭКФ этого признака в норме и при заболевании, а затем оценить информативность этого признака. С формальной точки зрения диагностика заболевания в условиях неопределенности представляет собой задачу статистического распознавания нозологической формы на основе информативных КДП. Для этого могут быть использованы соответствующие математические методы.

Методика

Для получения новых знаний о вероятностных особенностях проявления КДП в норме и при диагностируемом заболевании необходимо организовать и провести выборочное статистическое исследование. Это исследование предполагает формирование двух выборочных совокупностей, одна из которых представляет норму, а другая – диагностируемое заболевание. В основе формирования выборочных совокупностей лежит принцип случайного отбора. Включение случайно отобранных единиц наблюдения в выборочные совокупности происходит на основе определенных в ходе обследования прямых (патогномоничных) признаков. Желательно, чтобы сформированные выборочные совокупности не являлись малыми, т.е. объем каждой выборки должен быть более 30 единиц наблюдения. Исходя из целей и задач исследования, а также фактического наличия диагностических приборов и устройств формируется список признаков, подлежащих исследованию. В выборочных статистических совокупностях у каждой единицы наблюдения регистрируются результаты измерения КДП. В результате для каждого исследуемого признака получают две выборочные статистические совокупности, одна из них представляет множество измерений признака для единиц наблюдения, отнесенных к норме, а другая – для единиц наблюдения, у которых по зафиксированным патогномоничным признакам диагностировано заболевание. Априорно предполагается, что эти выборки могут содержать информацию, которая может быть использована для построения диагностического правила в условиях неопределенности при отсутствии патогномоничных признаков.

Для понимания сущности методики следует коротко пояснить смысловое значение признака при диагностике заболеваний. В живом организме одновременно протекает множество сложных и взаимосвязанных процессов. Для количественного описания свойств организма в норме и в определенных нозологических формах используются признаки (параметры). Признаком будем называть величину, количественно характеризующую свойство процесса, явления или организма в целом. Организм характеризуется множеством параметров, которые необходимо измерить для описания его состояния. Медицинские параметры могут определяться на организменном, органном и тканевом уровнях. В общем случае признаки измеряются в определенных шкалах. Наиболее часто в медицинской практике используются следующие шкалы: номинальная шкала, порядковая шкала, шкала интервалов и шкала отношений [1, 3]. Результатом измерения признака является число, свойства которого и допустимые математические операции определяются используемой измерительной шкалой. Для каждой измерительной шкалы разработаны соответствующие методы анализа экспериментальных данных.

Численные значения признаков организма содержат в себе важную медицинскую информацию. Для измерения и регистрации медицинских признаков широко применяются специальные технические устройства (приборы и системы). Измерение медицинских признаков (параметров) при помощи технических устройств обычно производится в шкале интервалов или шкале отношений. По своей сути эти измерения являются физическими, а их результаты описываются действительными числами. Следовательно, признаки, измеряемые в шкале интервалов и шкале отношений, являются непрерывными величинами. Поскольку результат измерения является заранее неизвестным, то результат физического измерения представляет собой непрерывную случайную величину. При проведении выборочных статистических исследований одной из задач, предшествующей оценке информативности признаков, является обоснование закона распределения для исследуемых непрерывных случайных величин. Обычно при помощи критериев согласия проверяется гипотеза о соответствии выборочных данных нормальному закону распределения. По причине малого объема выборочных данных или в силу особенностей исследуемых признаков обосновать нормальность их распределения или соответствие иному теоретическому закону распределения вероятностей не всегда представляется возможным. В этом случае для статистического описания исследуемых признаков на основании выборочных данных можно вычислять эмпирическую кумулятивную функцию (ЭКФ) [1, 3]. Исследуемый признак является информативным для диагностики определенной нозологической формы, если соответствующие кумулятивные функции в норме и при данной нозологической форме значимо различаются. Чем больше различие между ЭКФ, тем выше диагностическая информативность признаков.

Для выявления значимых различий между ЭКФ исследуемого признака в норме и при определенном заболевании можно использовать критерий Колмогорова-Смирнова [3]. Реализация этого критерия позволяет вычислить пороговое значение признака, при котором имеет место максимум модуля разности между ЭКФ в норме и при заболевании. В случае если модуль разности между ЭКФ превосходит критическое значение для заданного уровня значимости (вероятности ошибки первого рода), то можно различия между ЭКФ обоснованно считать значимыми, а исследуемый признак отнести к классу информативных. Относительно порогового значения можно осуществить категоризацию информативного признака, который является результатом физического измерения и с формальной точки зрения представляет собой непрерывную величину. Категоризация по порогу позволяет преобразовать непрерывную величину в порядковую дихотомическую величину, имеющую две категории: допороговую категорию, включающую в себя значения меньшие или равные порогу; надпороговую категорию, которая включает в себя значения большие порога. Для категоризированных признаков в норме и при заболевании можно составить таблицы сопряженности признаков формата 2×2. Для составленных таблиц сопряженности признаков проверяется статистическая гипотеза об отсутствии статистической зависимости (сопряженности) между номинальным признаком, характеризующим состояние организма (норма и заболевание), и порядковым дихотомическим признаком, имеющим две градации – допороговую и надпороговую. Проверить статистическую гипотезу о независимости указанных признаков можно при помощи критерия хи-квадрат или точного критерия Фишера, если анализируются выборки малого объема. Отклонение статистической гипотезы о независимости признаков будет являться подтверждением состоятельности порогового значения. Пороговые значения, вычисленные для анализируемых КДП, имеют важное самостоятельное практическое и научное значение. Категоризация КДП позволяет вычислить их информативность. Для этого можно использовать информационную меру Кульбака. Вычисление информативности КДП по Кульбаку позволяет осуществить их ранжирование и определить порядок их включения в диагностический критерий. Описанная методика, по сути, является информационной технологией, которая определяет порядок действий и математических вычислений, необходимых для вычисления информативности КДП, являющихся непрерывными физическими величинами.

Цель исследования заключалась в разработке способа, который на основе индуктивных выводов, полученных в результате статистического исследования, позволяет произвести оценку их информативности на основе применения известных математических конструктов. Способ рекомендуется для оценки результатов как медико-биологических исследований, так и клинических данных при дифференциальной диагностике заболеваний, а также идентификации состояний, вызванных фармакологическим воздействием.

Результаты исследования и их обсуждение

В результате проведенного теоретического исследования был разработан способ оценки информативности КДП, являющихся непрерывными физическими величинами. Способ

представляет собой последовательность действий с первичными данными, полученными в результате выборочного статистического исследования. В основе этих действий лежит последовательное преобразование первичных данных и их математическая обработка в соответствии с изложенной выше методикой. Используемое ниже математическое описание приводится без строгого обоснования и адаптировано для понимания исследователями, которые не имеют специальной математической подготовки.

Для представления и обсуждения результатов исследования сформулируем задачу, для решения которой будет применен и подробно описан разработанный способ. Пусть имеются две выборочные совокупности V_1 и V_2, одна из которых (V_1) представляет норму, а другая (V_2) – заболевание. Объемы выборочных совокупностей соответственно равны N_1 и N_2. В этих выборочных совокупностях исследуется некоторый КДП, в отношении которого предполагается, что он содержит полезную диагностическую информацию и, следовательно, может быть использован в диагностическом критерии. Исследуемый КДП является непрерывной случайной величиной. В результате сбора первичной информации проведено измерение исследуемого КДП у каждой единицы наблюдения в выборочных совокупностях V_1 и V_2. Для обоснованного включения исследуемого КДП в диагностический критерий необходимо оценить его информативность. Разработанный способ оценки информативности признаков, являющихся непрерывными величинами, состоит из следующих пяти этапов.

На первом этапе в каждой выборочной совокупности V_1 и V_2 на основании первичных данных для исследуемого КДП строится вариационный ряд и ЭКФ [1]. Для построения вариационного ряда для исследуемой выборки определяются варианты x и соответствующие им абсолютные частоты $f(x)$ встречаемости в выборке. Варианты – это конкретные значения признака, которые реализовались в исследуемой выборке. В качестве значений аргумента ЭКФ используются выборочные значения вариант построенного вариационного ряда. В данном случае ЭКФ является функцией $F(x)$ действительного аргумента x, определяющая относительную частоту (эмпирическую вероятность) события $X \le x$, где X – выборочные значения исследуемого признака; x – выборочные значения вариант в соответствующей выборке. Множество вариант в выборке V_1 обозначим через W_1, а множество вариант в выборке V_2 – через W_2 Эмпирическую вероятность события $X \le x$, запишем как $P(X \le x)$. В общем случае формальное определение ЭКФ имеет вид: $F(x) = P(X \le x) = \frac{n_x}{N}$, где n_x – число выборочных значений признака, которые не превосходят значение варианты x, т.е. выполняется условие $X \le x$. Соответственно для выборочной совокупности V_1 строится ЭКФ $F_1(x) = \frac{n_{1x}}{N_1}$, а для выборочной совокупности V_2 – ЭКФ $F_2(x) = \frac{n_{2x}}{N_2}$. Аргументом x для функции $F_1(x)$ являются значения вариант из множества W_1, а аргументом x для функции $F_2(x)$ – выборочные значения вариант из множества W_2. Значения n_{1x} вычисляются на основании анализа первичных данных в выборке V_1 при выполнении условия $X \le x$, где значения x принадлежат множеству W_1. Аналогично значения n_{2x} вычисляются при выполнении условия $X \le x$ в выборке V_2, где значения x принадлежат множеству W_2.

Для автоматизации преобразования первичных данных и математических вычислений можно использовать статистические функции табличного процессора Microsoft Excel. Целесообразно представить эмпирические кумулятивные функции $F_1(x)$ и $F_2(x)$ в табличном и графическом виде. Графическое представление ЭКФ для значений исследуемых выборок позволяет визуально оценить различия между ними. На этом первый этап можно считать завершенным.

На втором этапе вычисляется модуль максимальной разности между функциями $F_1(x)$ и $F_2(x)$, а также значение аргумента, при котором достигается этот максимум. Для этого необходимо построить ЭКФ для выборок V_1 и V_2 на одном множестве значений аргумента. Такое множество W формируется в результате объединения множества вариант W_1 выборки V_1 и множества вариант W_2 выборки V_2, т.е. $W = W_1 \cup W_2$. Множество вариант W, полученное в результате объединения необходимо проранжировать. Значения полученного ранжированного множества обозначим через y, т.е. $y \in W$. Эти значения используются в качестве аргумента для построения

ЭКФ $F_1(y)=\frac{n_{1y}}{N_1}$ и $F_2(y)=\frac{n_{2y}}{N_2}$ в исследуемых выборках V_1 и V_2. Для построенных функций $F_1(y)$ и $F_2(y)$ вычисляется максимальное значение модуля разности между ними, т.е. $d_{\max}=\max\left|F_1(y)-F_2(y)\right|$. Одновременно определяется значение аргумента y_p, при котором выполняется это условие. Значение y_p будем называть пороговым. В дальнейшем пороговое значение будет использоваться для категоризации первичных данных в выборках V_1 и V_2. На этом второй этап можно считать выполненным.

На третьем этапе проверяется статистическая гипотеза H_0 об отсутствии значимых различий между ЭКФ $F_1(y)$ и $F_2(y)$. По сути, под проверкой этой статистической гипотезы понимаются действия, направленные на ее опровержение. Для проверки гипотезы H_0 необходимо зафиксировать ошибку первого рода (уровень значимости) α и выбрать соответствующий статистический критерий. Уровень значимости, например, можно зафиксировать на уровне $0{,}05$. Для проверки гипотезы H_0 следует использовать критерий Колмогорова-Смирнова [3]. В том случае, когда гипотеза H_0 отклоняется по причине ее малой вероятности, то принимается альтернативная гипотеза H_1, состоящая в том, что ЭКФ $F_1(y)$ и $F_2(y)$ значимо различаются. Мерой различия $F_1(y)$ и $F_2(y)$ может являться статистика Колмогорова-Смирнова λ, которая вычисляется по формуле: $\lambda=d_{\max}\sqrt{\frac{N_1\cdot N_2}{N_1+N_2}}$ Для принятия решения в отношении проверяемой гипотезы H_0 необходимо сравнить эмпирическое значение статистики Колмогорова-Смирнова λ с критическим значением $\lambda_{кр}$, которое зависит от выбранного уровня значимости. Для уровня значимости $\alpha=0{,}05$ критическое значение статистики Колмогорова-Смирнова равно $\lambda_{кр}(0{,}05)=1{,}36$. Решение в отношении гипотезы H_0 принимается на основании следующего правила: если $\lambda<\lambda_{кр}$, то нет основания отклонить гипотезу H_0, различия между ЭКФ $F_1(y)$ и $F_2(y)$ статистически незначимы; если $\lambda\geq\lambda_{кр}$, то есть основание отклонить гипотезу H_0 и принять гипотезу H_1, для выбранного уровня значимости α различия между ЭКФ $F_1(y)$ и $F_2(y)$ статистически значимы. На этом третий этап вычислений можно считать выполненным.

На четвертом этапе выполняется категоризация исследуемого непрерывного КДП, для которого обосновано значимое различие между ЭКФ $F_1(y)$ и $F_2(y)$. На этом же этапе производится проверка гипотезы об отсутствии значимых различий между категоризированными распределениями КДП, построенными для выборок V_1 и V_2. Категоризация статистических распределений в выборочных совокупностях V_1 и V_2 осуществляется на основе вычисленного значения y_p, которое выбирается в качестве порогового значения. В результате категоризации образуются две категории. Первая категория – это категория допороговых значений, для которой выполняется условие $X\leq y_p$. Вторая категория – это категория надпороговых значений, для которой выполняется условие $X>y_p$. Указанные категории будем обозначать соответственно K_1 и K_2. Категоризация непрерывного КДП, по сути, является преобразованием непрерывного признака в порядковый дихотомический признак, имеющий две упорядоченные категории. Упорядоченность категорий выражается в том, что значения КДП, отнесенные к категории K_2, больше значений признака, отнесенных к категории K_1. В образованных категориях производится подсчет абсолютного числа значений и соответствующих им относительных частот.

Таблица 2. Таблица сопряженности категоризированного косвенного диагностического признака (КДП)

	Категории признака	
Выборочные совокупности	K_1	K_2
V_1	$f_{11};\ \ p_{11}=\frac{f_{11}}{N_1}$	$f_{12};\ \ p_{12}=\frac{f_{12}}{N_1}$
V_2	$f_{21};\ \ p_{21}=\frac{f_{21}}{N_2}$	$f_{22};\ \ p_{22}=\frac{f_{22}}{N_2}$

В таблице 2 использованы следующие обозначения: N_1 – число единиц наблюдения в выборке V_1; N_2 – число единиц наблюдения в выборке V_2; f_{11} – число единиц наблюдения (абсолютная частота) в выборке V_1, у которых зафиксирована категория K_1; f_{12} – число единиц наблюдения в выборке V_1 у которых зафиксирована категория K_2; f_{21} – число единиц наблюдения в выборке V_2, у которых зафиксирована категория K_1; f_{22} – число единиц наблюдения в выборке V_2, у которых зафиксирована категория K_2. Величины p_{11}, p_{12}, p_{21}, p_{22} являются относительными частотами (долями), которые соответствуют абсолютным частотам f_{11}, f_{12}, f_{21}, f_{22}. Относительные частоты p_{11} и p_{12} в первой строке таблицы 2 представляют выборочное дискретное статистическое распределение $D_1(K)$ категоризированного признака K в выборке V_1. Относительные частоты p_{21} и p_{22} во второй строке таблицы 2 представляют выборочное дискретное статистическое распределение $D_2(K)$ категоризированного признака K в выборке V_2. Категоризированный признак K, являющийся аргументом дискретных статистических распределений $D_1(K)$ и $D_2(K)$, может принимать два значения: K_1 и K_2. Если существует значимое различие между ЭКФ $F_1(y)$ и $F_2(y)$, то после категоризации по порогу и преобразований оно должно сохраниться и между дискретными статистическими распределениями $D_1(K)$ и $D_2(K)$.

Для проверки гипотезы H_0 об отсутствии значимых различий между дискретными распределениями $D_1(K)$ и $D_2(K)$, представленными таблицей 2, можно использовать критерий хи-квадрат (Пирсона) [1, 3] или точный критерий Фишера [2], если анализируются выборки малого объема. Сущность и алгоритмы реализации этих критериев описаны, например, в [2]. Указанные критерии, в случае отклонения гипотезы H_0 и принятия гипотезы H_1 о существовании значимых различий, можно также использовать для обоснования оптимального порогового значения для категоризации непрерывного КДП. При оптимальном по критерию Колмогорова–Смирнова пороговом значении y_p величина статистики хи-квадрат, вычисленная по таблице 2, будет иметь наибольшее значение. При этом вероятность, вычисляемая в точном критерии Фишера, будет минимальной. После завершения описанных выше преобразований и вычислений четвертый этап можно считать законченным.

На пятом, заключительном, этапе производится оценка информативности исследуемого КДП. Для оценки информативности КДП используется информационная мера Кульбака, в основе которой лежит расхождение (дивергенция) между дискретными статистическими распределениями $D_1(K)$ и $D_2(K)$. Информативность КДП по Кульбаку вычисляется по формуле:

$$J[D_1(K), D_2(K)] = \left[10\lg\left(\frac{p_{11}}{p_{21}}\right)\cdot 0{,}5(p_{11}-p_{21})\right] + \left[10\lg\left(\frac{p_{12}}{p_{22}}\right)\cdot 0{,}5(p_{12}-p_{22})\right].$$

В приведенной формуле используются эмпирические вероятности p_{11}, p_{12}, p_{21}, p_{22}, вычисленные ранее в таблице 2. Величины $10\lg\left(\frac{p_{11}}{p_{21}}\right) = S(K_1)$ и $10\lg\left(\frac{p_{12}}{p_{22}}\right) = S(K_2)$ являются диагностическими коэффициентами для категорий K_1 и K_2 соответственно. Коэффициент 0,5 в формуле используется для усреднения и приближения к среднему диагностическому порогу при последующем использовании КДП в диагностическом критерии. С учетом введенных обозначений формула для вычисления информативности КДП имеет вид:

$$J[D_1(K), D_2(K)] = [S(K_1)\cdot 0{,}5(p_{11}-p_{21})] + [S(K_2)\cdot 0{,}5(p_{12}-p_{22})].$$

Вычисление величины $J[D_1(K), D_2(K)]$ для исследуемого КДП завершает пятый этап и в целом весь разработанный способ оценки информативности признака.

Заключение

Способ оценки информативности непрерывных КДП содержит пять основных этапов обработки первичных статистических данных. Для автоматизации вычислений можно, например, использовать математические и статистические функции табличного процессора Microsoft Excel. Оценка информативности непрерывных КДП с использованием разработанного способа позволяет выделить из множества исследуемых косвенных признаков те, которые несут полезную

информацию. Выделенные признаки могут быть проранжированы по их информативности и в дальнейшем использованы для построения диагностического алгоритма. В заключении следует отметить, что оценка информативности и выделение непрерывных признаков, имеющих значение для диагностики, являются первоосновой для построения диагностического алгоритма. Разработка диагностического алгоритма под конкретное заболевание является отдельной самостоятельной исследовательской задачей.

Литература (references)

1. Медик В.А., Токмачев М.С., Фишман Б.Б. Статистика в медицине и биологии: Руководство в 2-х томах. Т.1. Теоретическая статистика / Под редакцией Ю.М. Комарова. – М.: Медицина, 2000. – 412 с. [Medik V.A., Tokmachev M.S., Fishman B.B. *Statistika v medicine i biologii: Rukovodstvo v 2-h tomah. T.1. Teoreticheskaja statistika.* Statistics in medicine and biology: Guide in 2 volumes. V.1. Theoretical statistics. – Moscow: Medicine, 2000. – 412 p. (in Russian)]
2. Гланц С. Медико-биологическая статистика. Пер. с англ. – М.: Практика, 1998. – 459 с. [Glanc S. *Mediko-biologicheskaja statistika.* Medical and biological statistics. Per. with English. – Moscow: Practika, 1998. – 459 p. (in Russian)]
3. Сидоренко Е.В. Методы математической обработки в психологии. – СПб: ООО «Речь», 2003. – 350 с. [Sidorenko E.V. *Metody matematicheskoj obrabotki v psihologii.* Methods of mathematical processing in psychology. – Saint-Petersburg: OOO "Rech", 2003. – 350 p. (in Russian)]

Информация об авторах

Лямец Леонид Леонидович – кандидат технических наук, доцент, заведующий кафедрой физики, математики и медицинской информатики ФГБОУ ВО «Смоленский государственный медицинский университет» Минздрава России. E-mail: LLL190965@yandex.ru

Евсеев Андрей Викторович – доктор медицинских наук, профессор, заведующий кафедрой нормальной физиологии ФГБОУ ВО «Смоленский государственный медицинский университет» Минздрава России. E-mail: hypoxia@yandex.ru