УДК 59.002 + 004

© А. Л. Лобанов

ДИАГНОСТИЧЕСКАЯ ЦЕННОСТЬ КАЧЕСТВЕННЫХ И КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ В КОМПЬЮТЕРНЫХ ОПРЕДЕЛИТЕЛЯХ

[A. L. LOBANOV. DIAGNOSTIC VALUE OF QUALITATIVE AND QUANTITATIVE CHARACTERS IN COMPUTER IDENTIFICATION KEYS]

Биологическая диагностика — раздел таксономии, изучающий теорию и практику построения определителей биологических объектов (ключей) — имеет уже полувековую историю, которую можно начать с работ отечественных исследователей (Балковский, 1962; Кискин и др., 1965). Этот начальный этап совпал с появлением у биологов возможности использовать компьютеры. Именно компьютерные решения задач оптимизации определителей способствовали бурному развитию диагностики и привели к созданию целого ряда действующих систем (Goodall, 1968; Morse, 1968; Лобанов, 1974; Dallwitz, 1974; Pankhurst, 1975, 1978). В конце XX и в XXI в. значимые разработки в биологической диагностике уже неизменно связаны с компьютерами (Лобанов, Дианов, 1994; Smirnov, Lobanov, Dianov, 1996; Дианов, Лобанов, 1997; Лобанов, 1997; Dianov, Lobanov, 1999; Ryss, Lobanov, 1999; Лопатин, Довгайло, 2002; Dmitriev, 2006; Вахитов и др., Лобанов, 2007; Кирейчук и др., 2011).

Понятие «компьютерная биологическая идентификация» стало теперь достаточно широким и включает принципиально разные постановки задачи диагностики и методы ее решения (Лобанов, Рысс, 1999). Мы здесь рассматриваем только традиционные компьютерные таксономические диагностические системы, в которых признаки таксонов формирует составитель, а пользователь сам считывает признаки с определяемого экземпляра в процессе идентификации и сообщает их компьютерной программе, которая тем или иным способом помогает отнести экземпляр растения или животного к одному из заранее установленных составителем таксонов. К этому типу систем относится большинство разрабатываемых биологами программ. Компьютерные диагностические системы с использованием автоматической обработки изображений и диагностические системы на основе теории распознавания образов здесь не рассматриваются.

Классификация типов биологических определителей была разработана нами давно (Лобанов, 1972) и с тех пор практически не претерпела изменений. Изменилось только содержание термина «признак». Если раньше элемент распознавания на биологическом объекте назывался «рядом признаков», а его конкретные состояния — «признаками», то теперь говорят о «признаке» и его «состояниях». И, пожалуй, компьютерные ключи стоит сейчас подразделить на однопользовательские программы, работающие на локальном компьютере, и на многопользовательские, находящиеся на сер-

вере и доступные через Интернет. Большинство компьютерных определителей являются многовходовыми и политомическими. Важна для авторов компьютерных программ и еще одна характеристика — число возможных состояний одного признака у одного таксона, обозначаемая как мономодальность или полимодальность ключа.

Несмотря на внешние различия между компьютерными диагностическими системами, вызванные отличиями интерфейсов, степенью иллюстрированности и деталями организации диалога пользователя с компьютером, основной алгоритм множества программ разных авторов из разных стран стал практически однообразным, ибо уже давно достиг логического совершенства и целесообразности. В начале работы программы тем или иным способом выбирается и фиксируется исходный набор таксонов, к одному из которых должен принадлежать определяемый объект. В продвинутых диагностических системах используется информация об иерархических отношениях таксонов в этом наборе и часто предлагается выбрать таксономический уровень диагноза, нужный пользователю (например, при наличии в ключе видов из многих родов определяющему может быть достаточно узнать род). Затем обычно предусматривается выбор пользователем того подмножества признаков, с которым он предпочитает работать, и порядка их предъявления программой. Далее действия программы и пользователя состоят из повторяющихся одинаковых шагов диагноза. В начале каждого шага определяющему предоставляются пригодные для диагноза признаки, он выбирает один или несколько, а потом знакомится с выделенными в них состояниями и сопоставляет их со свойствами своего объекта. После ввода пользователем информации о выбранных состояниях программа отбирает для следующего шага те таксоны, которые ими обладают, и переоценивает оставшиеся признаки. Шаги повторяются до тех пор, пока не останется один возможный таксон. На этом диагноз заканчивается и выдается имеющаяся в системе информация об этом таксоне.

Одна из важных характеристик каждого компьютерного определителя — способ предъявления пользователю признаков для их использования на очередном шаге диагноза. Возможен их показ в фиксированном порядке (по алфавиту или по заранее определенным составителем предпочтениям). В продвинутых программах порядок признаков изменяется на каждом шаге в соответствии с их ценностью для различения оставшихся возможными таксонов. Эта изменяющаяся с шагами диагноза оценка полезности каждого признака именуется диагностической ценностью. Для вычисления диагностической ценности было предложено несколько вариантов формул (Pankhurst, 1970; Лобанов, 1974; Гамбарян, 1975). В рамках диссертационной работы (Лобанов, 1983) нами было проведено сравнение 6 различных формул на 16 реальных многовходовых определителях растений и животных разных авторов (с числом таксонов от 8 до 73). Формулы были, на первый взгляд, разными, но все основывались на подсчетах числа отдельных состояний очередного признака в матрице таксоны/признаки (полной исходной или сокращенной для оставшихся на втором и последующих шагах таксонов). Алгоритм оценки формул был очень прост. Матрица таксоны/признаки конкретного ключа вводилась в компьютер, и специальная версия диагностической программы автоматически имитировала диагноз каждого включенного в определитель таксона. На каждом шаге диагноза признаки ранжировались по исследуемой формуле и использовался признак с наивысшей диагностической ценностью. Числа пройденных до окончания диагноза шагов для всех таксонов суммировались и делились на число таксонов в ключе. Так вычислялась средняя длина пути определения для каждой формулы. Предложенная нами формула или не уступала другим, или оказывалась лучшей. Незначительно проиграла она (средняя длина 2.79 против 2.74

по другой формуле) только один раз на ключе для 19 таксонов, поэтому в дальнейших наших разработках использовалась только эта формула:

$$d = \frac{N^2}{\sum_{i=1}^{s} n_i^2},$$

где N — общее число таксонов на данном шаге; *s* — число различных состояний оцениваемого признака; n — число таксонов, имеющих состояние *i*; d — значение диагностической ценности.

Проведенное почти 40 лет назад сравнение формул основывалось преимущественно на мономодальных ключах (т. е. таких, в которых у каждого таксона для каждого признака было приведено только одно состояние). При работе с полимодальными ключами, где число состояний у одного таксона по одному признаку не ограничено и может доходить до s, применение используемой формулы иногда приводило к ошибке, когда на последнем шаге оставалось всего два таксона, и в программы приходилось вносить поправки. Но принципиальный недостаток этой формулы даже не в этом, а в том, что она годится только для работы с качественными признаками, где число состояний, выделенных в одном признаке, ограничено. Обычно в соответствии с рекомендациями психологов (Miller, 1956), не выделяют в одном признаке более 7—9 состояний (показано, что объем кратковременной памяти человека ограничен 7 ± 2 блоками сопоставляемой информации). До сих пор нам приходилось превращать мерные признаки (длина тела, соотношение длины и ширины какой-то части тела, число члеников в многочлениковом органе и т. п.) в качественные, разбивая возможный диапазон значений признака на фиксированные интервалы, которые и являлись «состояниями». Это создает неудобство и для автора определителя, и для пользователя, который вынужден выбирать два интервала, когда его измерение попадает на границу между ними.

Решение очевидно — в определитель нужно включать и количественные признаки, которые в матрице таксоны/признаки будут представлены двумя значениями (минимум и максимум) для каждого таксона. Пока признаки предъявляются в фиксированном порядке, проблема не возникает. Но при ранжировании признаков по их диагностической ценности встает задача сопоставимых оценок качественных и количественных признаков. Ниже нами предлагается новый способ расчета ценности признаков, который позволяет совместно использовать на равных правах качественные и количественные признаки.

Алгоритм оценки признака на текущем (для данного шага диагноза) множестве возможных таксонов легко описывается словами. Производится перебор всех доступных на данном шаге признаков (если шаг не первый, то уже использованные признаки отбрасываются). Для каждого признака делается анализ матрицы таксоны/признаки. Каждый оставшийся возможным таксон сравнивается с каждым — перебираются все возможные пары. Если множество состояний качественного признака первого таксона хоть одним состоянием совпадает с множеством второго таксона (или диапазоны значений количественного признака перекрываются или соприкасаются границами), то засчитывается совпадение, в противном случае — несовпадение. Сумма несовпадений для данного признака делится на сумму совпадений (для исключения деления на нуль ко второй сумме прибавляется единица). Частное и принимается за оценку диагностической ценности признака.

Явным достоинством этого метода является полная сопоставимость оценок для качественных и количественных признаков. Это упрощает компью-

терные программы ключей и позволяет группировать признаки по биологически важным аспектам. Для доказательства практического преимущества было проведено сравнение нового способа со старой формулой на реальном определителе жуков (Лобанов, Дианов, 1996) — 130 таксонов и 24 признака. Алгоритм сравнения был описан выше. Средняя длина пути определения по старой формуле = 5.36, а новым способом = 5.15. Налицо небольшой, но вполне ощутимый выигрыш.

К недостаткам нового способа можно отнести более высокую трудоемкость вычислений. Для оценки по старой формуле нужно произвести N извлечений из матрицы таксоны/признаки, а для оценки новым способом — $N \cdot (N - 1)/2$, т. е. примерно в $N/2$ раз больше, однако постоянно нарастающая скорость работы современных процессоров делает этот недостаток несущественным.

Выразить одной формулой новый алгоритм затруднительно. Мы попытались это сделать с использованием обозначений теории множеств:

$$d = \frac{\sum_{i=1,\, j=i+1}^{N-1,\, N} (T_i \cap T_j \neq \varnothing)}{1 + \sum_{i=1,\, j=i+1}^{N-1,\, N} (T_i \cap T_j \neq \varnothing)}.$$

К использованным выше обозначениям здесь добавлено T — множество состояний (или возможных числовых значений для количественных признаков) у данного таксона (для оцениваемого признака). В числителе формулы — сумма случаев, когда пересечение содержимого ячеек матрицы таксоны/признаки для двух сравниваемых таксонов является пустым, а в знаменателе — не пустым. Очевидно, что более ценным в диагностическом отношении будет тот признак, у которого сумма пустых пересечений больше.

СПИСОК ЛИТЕРАТУРЫ

Балковский Б. Е. О повышении диагностической значимости признаков, используемых для определения растений // Бот. журн. 1962. Т. 47, № 9. С. 1309—1314.

Вахитов А. Т., Граничин О. Н., Кирейчук А. Г., Лобанов А. Л. Алгоритмы построения вопросника минимальной длины для биологического определителя в Интернете и успехи их реализации. Научный сервис в сети ИНТЕРНЕТ: многоядерный компьютерный мир. 15 лет РФФИ: Тр. Всерос. науч. конф. (24—29 сентября 2007 г., г. Новороссийск). М.: Изд-во МГУ, 2007. С. 293—295.

Гамбарян П. П. Числовой определитель водных цветковых Армении // Биол. журн. Армении. 1975. Т. 28, № 9. С. 108—111.

Дианов М. Б., Лобанов А. Л. PICKEY — Программа для определения организмов с интерактивным использованием изображений // Базы данных и компьютерная графика в зоологических исследованиях. СПБ., 1997. С. 35—39. (Тр. Зоол. ин-та РАН, т. 269).

Кирейчук А. Г., Лобанов А. Л., Смирнов И. С., Иночкин А. А., Степаньянц С. Д. Интернет-определители биологических объектов. 5 лет спустя // Научный сервис в сети Интернет: экзафлопсное будущее: Тр. Междунар. суперкомпьютерной конф. (19—24 сентября 2011 г., г. Новороссийск). М.: Изд-во МГУ, 2011. С. 449—453.

Кискин П. Х., Печерская И. Н., Печерский Ю. Н. Автоматизация диагностического поиска сортов винограда на ЭВМ «Минск-1» // Виноделие и виноградарство СССР. 1965. № 1. С. 21—22.

Лобанов А. Л. Логический анализ и классификация существующих форм диагностических ключей // Энтомол. обозр. 1972. Т. 51, вып. 3. С. 668—681.
Лобанов А. Л. Оценка диагностической ценности рядов признаков в многовходовых определителях, рассчитанных на использование ЭВМ // Тез. докл. VI Коми республиканской молодежной научной конференции. Сыктывкар, 1974. С. 125—126.
Лобанов А. Л. Принципы построения определителей насекомых с использованием электронных вычислительных машин. Автореф. дис. ... канд. биол. наук. Л.: ЗИН АН СССР, 1983. 19 с.
Лобанов А. Л. Компьютерные определители в биологии: результаты 30-летней эволюции // Компьютерные базы данных в ботанических исследованиях. Сб. науч. тр. СПб., 1997. С. 51—55.
Лобанов А. Л., Дианов М. Б. Диалоговая компьютерная диагностическая система BIKEY и возможности ее использования в энтомологии // Энтомол. обозр. 1994. Т. 73, вып. 2. С. 465—478.
Лобанов А. Л., Дианов М. Б. Мир жуков (Wir bestimmen Käfer). CD-ROM и краткое руководство. Berlin: dialobis edition, 1996.
Лобанов А. Л., Рысс А. Ю. Компьютерные идентификационные системы в зоологии и ботанике: современное состояние и перспективы / Рысс А. Ю., Смирнов И. С. (ред.) // Информационно-поисковые системы в зоологии и ботанике. Тез. междунар. симпоз., май 1999. СПб., 1999. С. 17—19. (Тр. Зоол. ин-та РАН, т. 278).
Лопатин И. К., Довгайло К. Е. Жуки рода Cryptocephalus (Chrysomelidae) Палеарктики. CD-ROM определитель 398 видов и база данных на основе системы «Lysandra». Минск, 2002.
Смирнов И. С., Лобанов А. Л. Компьютерный определитель по офиурам как база данных для хранения таксономической информации // Бюл. Моск. общ-ва испыт. прир. (МОИП). Отд. геологии. 1999. Т. 72, вып. 1. С. 87—88.
Dallwitz M. J. A flexible program for generating identification keys // Syst. Zool. 1974. Vol. 23, N 1. P. 50—57.
Dianov M. B., Lobanov A. L. BIKEY8 — the biological diagnostic software for Windows // Information Retrieval Systems in Biodiversity Research (Abstracts of the International Symposium. May 1999). St. Petersburg, 1999. P. 74. (Proc. Zool. Inst. RAS, vol. 278).
Dmitriev D. A. 3I, a new program for creating Internet-accessible interactive keys and taxonomic databases and its application for taxonomy of Cicadina (Homoptera) // Rus. Entomol. J. 2006. Vol. 15, pt 3. P. 263—268.
Goodall D. W. Identification by computer // BioScience. 1968. Vol. 18, N 6. P. 485—488.
Miller G. A. The magical number seven, plus or minus two // Psychol. Rev. 1956. Vol. 63. P. 81—92.
Morse L. E. Construction of identification keys by computer // Amer. J. Botan. 1968. Vol. 55, N 6. P. 737.
Pankhurst R. J. A computer program for generating diagnostic keys // Computer J. 1970. Vol. 13, N 2. P. 145—151.
Pankhurst R. J. (ed.). Biological Identification with Computers. London: Academic Press, 1975. 333 p.
Pankhurst R. J. Biological Identification. The Principles and Practice of Identification Methods in Biology. London: Edward Arnold, 1978. 104 p.
Ryss A., Lobanov A. Principles of taxonomic identification illustrated on nematode computer key // Problems of Nematology. St. Petersburg, 1999. P. 22—23. (Proc. Zool. Inst. RAS, vol. 280).
Smirnov I., Lobanov A. L., Dianov M. B. Creation of computer picture identification key for the arctic ophiuroids // 9th International Echinoderm Conference. August 5 to 9, 1996. San Francisco, California. Program and Abstracts. San Francisco, 1996. P. 132.

Зоологический институт РАН,
Санкт-Петербург.
E-mail: all@zin.ru

Поступила 12 V 2014.

SUMMARY

A new method of estimations of diagnostic value of characters for computer keys to biological objects is proposed. The algorithm described is applicable to correct comparison of qualitative and quantitative characters. A brief review of the history of biological diagnostics is given.

УДК 59.002 + 004

ДИАГНОСТИЧЕСКАЯ ЦЕННОСТЬ КАЧЕСТВЕННЫХ И КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ В КОМПЬЮТЕРНЫХ ОПРЕДЕЛИТЕЛЯХ. Лобанов А. Л. Энтомол. обозр., 2015, том 94, вып. 1.

Предлагается новый метод вычисления диагностической ценности признаков для компьютерных определителей биологических объектов. Описанный алгоритм годится для сравнения качественных и количественных признаков при их предъявлении пользователю. Сделан краткий обзор истории развития биологической диагностики.
Ключевые слова: компьютерные определители, диагностические ключи, таксономические объекты, качественные и количественные признаки, диагностическая ценность признаков.

УДК 51.76

МОДЕЛЬ РАЗВИТИЯ СПОНТАННОЙ ВСПЫШКИ ЧИСЛЕННОСТИ НАСЕКОМОГО С АПЕРИОДИЧЕСКОЙ ДИНАМИКОЙ. Переварюха А. Ю. Энтомол. обозр., 2015, том 94, вып. 1.

В статье рассматривается разработка динамической системы, позволяющей моделировать описанный в литературе сценарий вспышки численности листоблошки *Cardiaspina albitextura* Taylor, 1962. Предлагается непрерывно-событийная вычислительная структура модели, учитывающая изменение смертности на разных стадиях развития насекомого и фактор резкого исчерпания пищевых ресурсов на пике развития вспышки. Ил. 6.
Ключевые слова: модели динамики насекомых, вспышки численности, инвазии опасных вредителей, устойчивые и переходные состояния популяции, бифуркации в биологических моделях.