Н.П. Фот, О.С. Бравичева

ОБОСНОВАНИЕ ИСПОЛЬЗОВАНИЯ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ВИДОВОЙ ИДЕНТИФИКАЦИИ СТАФИЛОКОККОВ

В статье рассматриваются проблемы идентификации сложных биологических объектов (на примере рода стафилококков) с использованием многомерных статистических методов обработки данных. Для решения поставленной проблемы обосновано применение кластерного анализа.

Произошедшие в последнее десятилетие значительные изменения в таксономии и номенклатуре микроорганизмов, приведшие к резкому расширению идентифицируемых видов, определили необходимость разработки методов их эффективного распознавания и дифференциации. В полной мере данное замечание относится к микроорганизмам рода Staphylococcus, количество зарегистрированных видов которых к началу XXI века перевалило за третий десяток.

Несмотря на выраженное биологическое своеобразие отдельных видов стафилококков, безошибочное выявление их представителей остается одной из сложных задач в повседневной микробиологической практике, решение которой возможно только через тестирование целого спектра биологических и биохимических характеристик. Перечень же подобных тестов, приведенный в международном определителе Bergey's [7], оказывается достаточно широк, что существенно осложняет процедуру идентификации и делает ее чрезвычайно трудоемкой. Дополнительным моментом, затрудняющим решение задачи видовой идентификации стафилококков, является тот факт, что определяемыми для этих целей признаками одновременно могут обладать (или не обладать) представители нескольких видов, причем даже внутри одного вида могут встречаться представители с противоположными градациями признака, что полностью исключает возможность использования дихотомического подхода.

Соответственно столь сложный характер распределения признаков в условиях необходимости проведения многоальтернативной дифференциации определяет необходимость использования для этих целей вероятностно-статистических подходов, позволяющих на базе всей совокупности использованных тестов найти выраженное в числах наиболее вероятное решение. Среди подобных подходов в настоящее время наиболее хорошо разработаны метод диагностических порогов [4], надежность и эффективность которого проверяется методом определения ошибок [6], и некоторые другие. При этом их использование уже результировалось в создании ряда компьютерных программ, предусматривающих визуальное или автоматическое (с использованием фотометра – ридера) считывание результатов биохимических тестов [8].

Однако данные методы не лишены ряда существенных недостатков, основным из которых является жесткая привязка к определенному и обычно достаточно ограниченному перечню бактериальных видов, а также к той или иной используемой при их изучении коммерческой тест-системе, что требует кардинального пересмотра всей процедуры идентификации при изменении размерности матрицы за счет введения в нее ранее неизвестных видов микроорганизмов или увеличении спектра идентификационных тестов.

Один из возможных подходов к решению данной проблемы заключается в использовании современных многомерных статистических методов, позволяющих существенно оптимизировать и унифицировать процедуру идентификации, что и явилось целью настоящего исследования.

При проведении работ в данном направлении вся совокупность накопленных к настоящему времени исходных данных о 36 известных видах стафилококков, каждый из которых охарактеризован по 36 таксономически значимым биологическим характеристикам, была представлена матрицей объясняющих переменных размерностью 36x36. При этом все нечисловые компоненты, приведенные в определителе бактерий Bergey's [7] и определяющие биологические характеристики микроорганизмов, были переведены в числовые, где знаку «+» соответствовало значение вероятности наличия данного признака 0,9; знакам «+, -» и «ND» по 0,5 ; знаку «-» – 0,1. Соответственно, значения параметров каждого микробного вида были выражены вектором – столбцом Y с размерностью 36x1.

Обсуждая выбор вероятностно-статистического метода обработки полученных данных и их последующего использования для решения задачи ви-

довой идентификации стафилококков, следует указать на невозможность использования в имеющихся конкретных условиях дискриминантного анализа, ранее созданного непосредственно для решения подобного типа задач. Наиболее существенным ограничением названного метода в данном случае является выполнение условия о нормальности распределения исходных векторов, которое в большинстве случаев представленных данных не соблюдается. В этой связи наше внимание было обращено на методы, свободные от данного условия, в качестве наиболее перспективного из которых выбран кластерный анализ.

Традиционно кластерный анализ рассматривается как совокупность методов, позволяющих классифицировать многомерные наблюдения через формирование групп схожих между собой объектов, которые принято называть кластерами. При этом, в отличие от комбинационных группировок, кластерный анализ приводит к разбиению на группы с учетом всех признаков одновременно. Названные особенности определили сферу применения кластерного анализа, обычно используемого для построения научнообоснованных классификаций, а также выявления внутренних связей между единицами наблюдаемой выборки. В частности, ранее кластерный анализ был использован нами для установления уровней гомологии основных видов стафилококков по совокупности их биохимических и некоторых патогенетически значимых характеристик [2], что позволило оценить связи между их фенотипическими проявлениями и экологическими особенностями.

Дополнительной особенностью кластерного анализа является возможность ассоциации каждого из впервые наблюдаемых объектов с ранее уже существующим наблюдением или группой наблюдений. Последний момент как раз и положен нами в основу использования кластерного анализа для решения задачи видовой идентификации стафилококков на основе вычисления наибольшей близости каждого из вновь анализируемых объектов (штаммов) с одним из видов стафилококков с известной таксономической характеристикой как неких «точек» в созданном 36-мерном пространстве. Естественно предположить, что геометрическая близость двух или нескольких точек означает идентичность таксономических характеристик соответствующих объектов.

Выбор меры близости является узловым моментом исследования, от которого решающим образом зависит окончательный вариант идентификации. При этом решение данного вопроса зависит в основном от главных целей исследования, биологической и статистической природы исходных данных, полноты априорных сведений о характере их вероятностного распределения. В этой связи следует также указать, что поскольку информация о характере распределения каждого из видов стафилококков обычно отсутствует (последнее является необходимым условием для проведения параметрического варианта кластерного анализа), при проведении дальнейшей работы мы сконцентрировали свое внимание на его непараметрическом варианте.

При этом понятие однородности объектов определяется заданием правила вычисления величины O_{ij}, характеризующей либо расстояние $d(O_i; O_j)$ между объектами O_i и O_j из исследуемой совокупности O (i,j = 1…n), либо степень близости (сходства) $r(O_i; O_j)$. Если задана функция $d(O_i; O_j)$, то близкие в смысле этой метрики объекты считаются однородными, принадлежащими одному классу, т. е. в случае решения задачи настоящего исследования – одному виду стафилококков.

В качестве же примеров расстояний и мер близости нами были рассмотрены обычное эвклидово и взвешенное эвклидово расстояние, расстояние Махалонобисса и Хеммингово расстояние, из которых в качестве оптимального выбрано обычное эвклидово расстояние как удовлетворяющее следующим требованиям: 1) компоненты вектора наблюдения однородны по своему биологическому смыслу, причем все они одинаково важны с точки зрения решения вопроса об отнесении объекта к тому или иному виду; 2) понятие близости объектов совпадает с понятием геометрической близости в анализируемом пространстве. При этом расчет расстояния осуществляется по формуле:

$$d(y,x_j)=\sqrt{\sum_{k=1}^{36}(y^k-x_j{}^k)^2}\,, \qquad (1)$$

где $d(y,x_j)$ – расстояние между неизвестным (идентифицируемым) микроорганизмом y и известным x_j видом стафилококков, $j=\overline{1,36}$;
k – номер признака, $k=\overline{1,36}$.

На основе приведенного выше обоснования в системе программирования Delphi 5 нами разработана программа, позволяющая автоматизировать процедуру видовой идентификации стафилококков, укрупненная блок-схема алгоритма работы которой представлена на рисунке 1.

После запуска программы на экране появляется основная форма, вид которой приведен на рисунке 2.

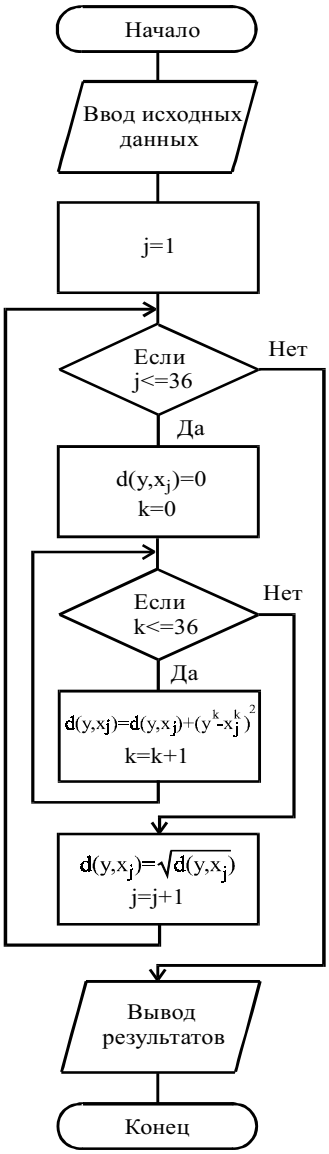

Рисунок 1. Описание алгоритма работы программы

При этом верхняя таблица содержит сведения о всех известных к настоящему времени 36 видах стафилококков: порядковый номер, название вида и его характеристики. Во вторую таблицу вводятся характеристики идентифицируемого штамма стафилококка. При нажатии на кнопку «Расчет» происходит вычисление расстояния неизвестного идентифицируемого штамма до эталонных представителей каждого вида, а результаты подобного расчета, упорядоченные по убыванию расстояния, приводятся в нижней таблице. Соответственно, первый столбец содержит порядковый номер соответствующего вида стафилококка из верхней таблицы, второй столбец – название этого вида, третий – расстояние до него. При нажатии на кнопку «Выход» происходит выход из программы.

Рисунок 2. Основная экранная форма программы

Список использованной литературы:
1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – Москва: ЮНИТИ, 1998.
2. Дерябин Д.Г. Стафилококки: экология и патогенность. – Екатеринбург, УрО РАН, 2000.
3. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – Москва: «Финансы и статистика»,ю 2000.
4. Генкин А.А. Отнесение наблюдений к одному из двух возможных классов (диагностика и прогнозирование); статья.
5. Нечмирев А.Б. Вероятностно–статистический подход к дифференциации родов энтеробактерий ; статья.
6. Нечмирев А.Б., Нечмирева Т.С. Надежность и эффективность дифференциации микроорганизмов; статья.
7. Определитель бактерий Берджи // Пер. с англ. //Под редакцией Дж. Хоума, Н. Крига, П. Снита и др. – М., Мир, 1997.
8. Смирнова А.М., Трояшкин А.А., Падерина Е.М.Микробиология и профилактика стафилококковых инфекций.– Ленинград – Медицина, 1977.