Праграмныя сродкі

УДК 004.9

Н.Б. Осипенко, М.Н. Васенда, Т.В. Сатырова

РАЗРАБОТКА ПРОГРАММНОГО ИНСТРУМЕНТАРИЯ ОДНОМЕРНОГО АНАЛИЗА ДАННЫХ ПРИ ИССЛЕДОВАНИИ МЕДИКО-ЭКОЛОГИЧЕСКИХ СИСТЕМ

В работе описываются алгоритмы и их реализация в программном обеспечении «Strand», предназначенном для автоматизации исследования активных систем (АС). В основу этой разработки положена концепция конструктивного использования методов анализа данных и проверки их работоспособности при автоматизации выделения экспертом причинно-следственных связей в модели АС. Описывается также разработанный модуль для автоматизации создания тестов при верификации функциональных возможностей программного обеспечения «Strand». В работе определён ряд проблемных мест, которые планируется решать в последующем при проведении других видов анализа данных в приложении «Strand», так как описываемые здесь алгоритмы не покрывают весь спектр возможных вариантов. Как подтверждение практической значимости и хороших эксплуатационных показателей кратко приводятся результаты исследования, проведённого посредством анализа данных в программном обеспечении «Strand», полученные при обследовании 129 добровольцев контрольной группы из популяции г. Гомеля и Гомельской области для определения фенотипа ацетилирования.

Ключевые слова: программный инструментарий, медико-экологические системы.

Введение. Разрабатываемое программное обеспечение «Strand», описанное в работе [1], предназначается для автоматизации исследования АС и является развитием комплекса работ [2–3]. Созданный программный продукт реализует концепции конструктивного использования методов анализа данных. В частности, последним расширением его функциональных возможностей применительно к методам разведочного анализа данных стало: реализация методов устранения пропусков выборки, применение алгоритмов по подбору преобразования выборки и разработка инструментария расщепления многомодальной смеси распределений. И хотя последнее, на первый взгляд, может показаться избыточным при проведении разведочного анализа данных, однако именно такой подход позволяет эксперту первоначально сделать правильное предположение о характере данных, что может предопределить весь ход и результативность дальнейшего исследования.

Многомодальность, как правило, возникает в случае, когда не была обеспечена однородность выборки и информация бралась для объектов из разных классов. При обнаружении многомодальности необходимо провести расщепление смеси с тем, чтобы получить теоретические параметры и корректно осуществить нормировку данных.

Алгоритм расщепления смеси распределений. Данный алгоритм является развитием описанного в работе [4] варианта расщепления смеси двух компонент и реализует возможность расщепления n-модальной ($n<6$) разнесённой смеси:

1. Нахождение эмпирических статистических характеристик, которые будут использоваться

Осипенко Наталья Борисовна, канд. физ.-мат. наук, доц., доц. каф. математических проблем управления ГГУ им. Ф. Скорины (Гомель).
Адрес для корреспонденции: e-mail: mpu@gsu.by.

Васенда Михаил Николаевич, магистрант каф. математических проблем управления ГГУ им. Ф. Скорины (Гомель); науч. рук.– канд. физ.-мат. наук, доц. Н.Б. Осипенко (Гомель).
Адрес для корреспонденции: e-mail: mpu@gsu.by.

Сатырова Татьяна Викторовна, аспирант каф. общей и клинической фармакологии с курсом анестезиологии и реаниматологии ГГМУ (Гомель); науч. рук. – канд. мед. наук, доц., зав. каф. общей и клинической фармакологии с курсом анестезиологии и реаниматологии ГГМУ Е.И. Михайлова (Гомель).
Адрес для корреспонденции: e-mail: mpu@gsu.by.

при дальнейшей работе с данными в том случае, если не удастся подобрать согласующееся теоретическое распределение и рассчитать по нему теоретические параметры.

2. Выполнение «ядерной» аппроксимация функции распределения с размером «ядра» зависящим от объема выборки.

3. С использованием нормальной вероятностной бумаги оцениваются параметры распределения, то есть математическое ожидание и среднеквадратическое отклонение.

4. Проверка гипотез о нормальности распределения. На практике часто оказывается, что выборка согласуется с нормальным или логнормальным законом распределения. Гипотеза о нормальности распределения проверяется по критерию согласия Колмогорова–Смирнова:

$$D_N = \max_{1 \le i \le N} \left| N(x_i, \tilde{\mu}, \tilde{\sigma}) - \tilde{F}_i \right|.$$

Обнаружение факта нормальности распределения упрощает дальнейшую работу с выборкой: происходит расчёт теоретических параметров и выход из рассмотрения текущей компоненты.

5. Проверка гипотез о логнормальности распределения. Для проверки на логнормальность элементы выборки логарифмируются и производится проверка на нормальность. Если предположение о логнормальности подтвердилось, то также происходит завершение рассмотрения текущей компоненты.

6. Если предположение о нормальности или логнормальности не подтвердилось, то рассматриваются точки $\tilde{X}$ на нормальной вероятностной бумаге.

7. Для точек выборки из интервала $\left[\tilde{X}_i; \tilde{X}_{i+4}\right]$ определяем a_1 – угловой коэффициент линии линейной регрессии на нормальной вероятностной бумаге.

8. Далее происходит поиск первого и последнего скачков a_{i_1} и a_{i_2} (то есть те значения угловых коэффициентов, которые в отношении к следующему превышают некоторый заданный порог, например, 2).

9. Если ни одного скачка найти не удалось, то происходит завершение алгоритма (выборка одномодальна) или же переход к пункту 15, в случае рассмотрения подкомпоненты, иначе – пункт 10;

10. Рассматриваются начальные значения параметров смеси:
если найдены два скачка, то

$$\tilde{a}_1 = \frac{\tilde{X}_{i_1} + \tilde{X}_{i_2}}{2};\ \tilde{\sigma}_1 = \frac{\left|\tilde{X}_{i_1} - \tilde{X}_{i_2}\right|}{6};\ \tilde{a}_2 = \frac{\tilde{X}_1 + \tilde{X}_N}{2};\ \tilde{\sigma}_1 = \frac{\left|\tilde{X}_1 - \tilde{X}_N\right|}{6};$$

если найден лишь один скачок, то

$$\tilde{a}_1 = \frac{\tilde{X}_1 + \tilde{X}_{i_1}}{2};\ \tilde{\sigma}_1 = \frac{\left|\tilde{X}_1 - \tilde{X}_{i_1}\right|}{6};\ \tilde{a}_2 = \frac{\tilde{X}_{i_1} + \tilde{X}_N}{2};\ \tilde{\sigma}_1 = \frac{\left|\tilde{X}_{i_1} - \tilde{X}_N\right|}{6}.$$

11. Для каждого q1=0.1,…,0.9, q2=1–q1, где q1 и q2 – веса первой и второй компоненты, в выборке осуществляется вариация параметров смеси в сторону их увеличения и уменьшения (во всех комбинациях). При этом рассчитывается значение критерия, в качестве которого выступает сумма модулей отклонений значений эмпирической интегральной функции распределения от функции распределения со сгенерированными параметрами.

12. Находится минимальное значение критерия среди всех рассчитанных альтернатив. Соответствующие значения параметров распределения запоминаются как текущие.

13. Выполнение пунктов 11 и 12 происходит до тех пор, пока отклонение текущего значения критерия от значения критерия предыдущего шага превышает 0.1% текущего значения.

14. Проверяется гипотеза о соответствии исходных данных найденному закону распределения. Рассчитываются математические ожидания и дисперсии двух компонент смеси, их доли и точки расщепления по формуле, представленной ниже:

$$x_{1,2} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sqrt{\left(\mu_1\sigma_2^2 - \mu_2\sigma_1^2\right)^2 - \left(\sigma_1^2 - \sigma_2^2\right)\left(\mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2 + 2\sigma_1^2\sigma_2^2 Ln\frac{\sigma_2}{\sigma_1}\right)}}{\sigma_1^2 - \sigma_2^2}.$$

15. Предварительное изъятие элементов одной из компонент смеси. В первом случае, для левой компоненты удаляются все точки, которые располагаются за пределами m-2s правой компоненты, иначе для правой за пределами m+2s левой, с учётом её доли. То есть удаляемые точки должны соответствовать следующему условию:

$$X_{del} \in \begin{cases} (-\infty; \mu_2 - 2\sigma_2] - & \text{для левой компоненты,} \\ [\mu_1 + 2\sigma_1; +\infty) - & \text{для правой компоненты.} \end{cases}$$

Такое удаление преследует две практические цели: ускорение работы алгоритма и устранение проблемы «хвостов», возникающих при удалении элементов из подвыборки, для которой неправильно определён закон распределения.

16. Основное изъятие в цикле одной из компонент смеси выборки происходит следующим способом: зная интервалы, на которых располагается левая и правая компоненты, через математическое ожидание и дисперсию, генерируется на этом интервале число, согласно закону распределения, и изымается из выборки ближайшее к нему, так повторяется до исчерпания доли одной компоненты.

17. Для другой компоненты переходим к первому пункту. Если в пунктах 4, 5 или 9 для неё будет завершение, то в пунктах 15–17 левая и правая компоненты «меняются местами» и изъятию подлежит уже другая компонента.

18. Перерасчёт эмпирических и теоретических параметров для найденной одномодальной компоненты относительно её собственного закона распределения.

19. Сохранение каждой полученной одномодальной компоненты вместе со всеми проанализированными признаками.

20. При условии выбора двойной точности проведения алгоритма происходит повторное выполнение в цикле пунктов 15–16, то есть удаление всех компонент смеси, кроме одной.

21. Для каждой отобранной компоненты происходит повторный поиск теоретических и эмпирических параметров согласно пунктам 1–5, что позволяет уменьшить погрешность произведённого анализа.

Автоматизация проверки работы реализованных алгоритмов. Для разработанных алгоритмов была реализована возможность проходить апробацию на серии создаваемых автоматически, специально разработанных тестов. Например, на тестах для расщепления смеси, каждый из которых является результатом смеси 2, 3, 4 или 5 нормально распределённых генеральных совокупностей, с различными соотношениями в разнесённости математических ожиданий у компонентов смеси и разных дисперсий.

Для автоматизации при верификации программных средств статистического анализа данных был разработан модуль на языке программирования SVB (Statistica Visual Basic), предназначенный для создания тестовых данных, модуль работает в пакете Statistica 7.

Подход, реализованный в модуле, при создании тестов предполагает последовательность из трёх этапов, описанных ниже.

1. Создание шаблона с заполнением его первоначальными параметрами в соответствии с предполагаемым видом создаваемых тестовых наборов. В рамках этого этапа пользователю модуля автоматизированного тестирования предоставляется возможность выбрать типы подготавливаемых тестов и объёмы начальных данных. В результате пользователь получает на выходе шаблон для заполнения начальными данными, используемыми для построения набора тестов, который в дальнейшем будет использоваться на последующих этапах автоматизации тестирования.

2. Создание набора одномодальных выборок. В ходе этого этапа используется заранее подготовленный пользователем шаблон с введёнными начальными данными. Например, реализация автоматизации создания тестов для расщепления смеси распределений с параметрами, указанными в шаблоне, осуществляется следующим образом: получение выборки равномерно-распределённых чисел на отрезке [0;1] (первая выборка) с заданным пользователем в диалоге объёмом; получение значений обратной функции по значениям заданной функции распределения с параметрами N(0;1) по первой выборке (вторая выборка); получение различных выборок заданных распределений с предустановленными параметрами, например: для текущей потребности тестирования расщепления смесей нормальных распределений – N(a; σ) путём умножения значений второй выборки на σ и

Таблица 1 – Значения параметров нормального распределения (a, σ)

a	σ
0.2	1; 3; 6
1	0.2; 2; 4
2	1; 3; 6
3	0.2; 2; 4
4	1; 3; 6
6	0.2; 2; 4

прибавлением *a*, где *a* – оценка математического ожидания, σ – среднеквадратическое отклонение (таблица 1). В преддверии этого этапа пользователю предоставляется возможность выбрать файл шаблона из текущих открытых файлов, диапазон используемых начальных значений для генерации из таблицы шаблона и тип теста или закон распределения для генерируемых компонент.

3. Создание окончательного набора тестовых выборок для проведения верификации алгоритмов. Модуль предоставляет пользователю определиться с файлом, содержащим набор исходных выборок, подлежащих дальнейшему изменению, типом преобразования (например, для расщепления – типом смешивания), с максимальным количеством компонент, используемых в создаваемых наборах, и определением необходимых визуализаций для построенных наборов тестов. Например, объединение полученных выборок в смеси с количеством компонент от 1 до 6 и сохранение их в отдельные файлы, построение для всех выборок гистограмм и полигонов частот.

При увеличении предполагаемых компонент полученный набор тестов может оказаться достаточно большим, но всё же его создание происходит быстрее ручного создания части из набора тестов для проверки правильности программы.

С помощью этого модуля по представленным выше данным для смесей, содержащих до шести компонент, может быть построено 43776 тестовых выборок и графических их представлений при проведении верификации модуля и алгоритма расщепления. Но, конечно же, целесообразней воспользоваться возможностью выборки части данных из шаблона, содержащего начальные параметры. Как дальнейшее развитие данной тематики предполагается расширять данный модуль для использования при отладке работы остальных компонентов разрабатываемого программного обеспечения и реализовать механизм, позволяющий автоматизировать непосредственно сам процесс тестирования текущей версии продукта.

Апробация алгоритма и разработанного программного обеспечения на тестовых и экспериментальных данных.

В ходе проведения тестирования с использованием сгенерированного набора тестов была выявлена неспособность программы определять нахождение смеси распределений у выборок, компоненты которых слабо разнесены (незначительно разнятся математические ожидания у больших выборок). К такому типу выборок относится, например, выборка, представленная на рисунке 1. У данной выборки соответственно математические ожидания и дисперсии имеют следующие значения: $a_1 = 1$ и $\sigma_1 = 2$ – для первой компоненты смеси, $a_2 = 2$ и $\sigma_2 = 1$ – для второй компоненты смеси, $a_3 = 3$ и $\sigma_3 = 2$ – для третьей компоненты смеси. Объём объектов, входящих в каждую равен 200, соответственно веса у каждой компоненты одинаковы и равны 1/3.

В то же время, программа и, соответственно, реализованный в ней алгоритм хорошо выполняют расщепление смесей распределений при достаточно разнесённых, как представлено на рисунке 2, компонентах исследуемой выборки данных. У неё соответственно математические ожидания и дисперсии имеют следующие значения: $a_1 = 0.2$ и $\sigma_1 = 1$ – для первой компоненты смеси, $a_2 = 3$ и $\sigma_2 = 2$ – для второй компоненты смеси, $a_3 = 6$ и $\sigma_3 = 0.2$ – для третьей компоненты смеси. Объём объектов, входящих в каждую так же, как и в предыдущем случае, равен 200, соответственно и веса у каждой из компонент одинаковы и равны 1/3.

Выявленный недостаток не противоречит теории в таких случаях и преодолим при дальнейшем

Рисунок 1 – Выборка с близкорасположенными компонентами

Рисунок 2 – Выборка с разнесенными компонентами

исследовании в рамках программы «Strand» с использованием других функциональных модулей приложения и проведении с их помощью, например, кластерного анализа.

Посредством «Strand» был проведен анализ данных [5], полученных при обследовании 129 практически здоровых добровольцев контрольной группы из европеоидной популяции г. Гомеля и Гомельской области для определения фенотипа ацетилирования. Конечная выборка данных представлена на рисунке 3.

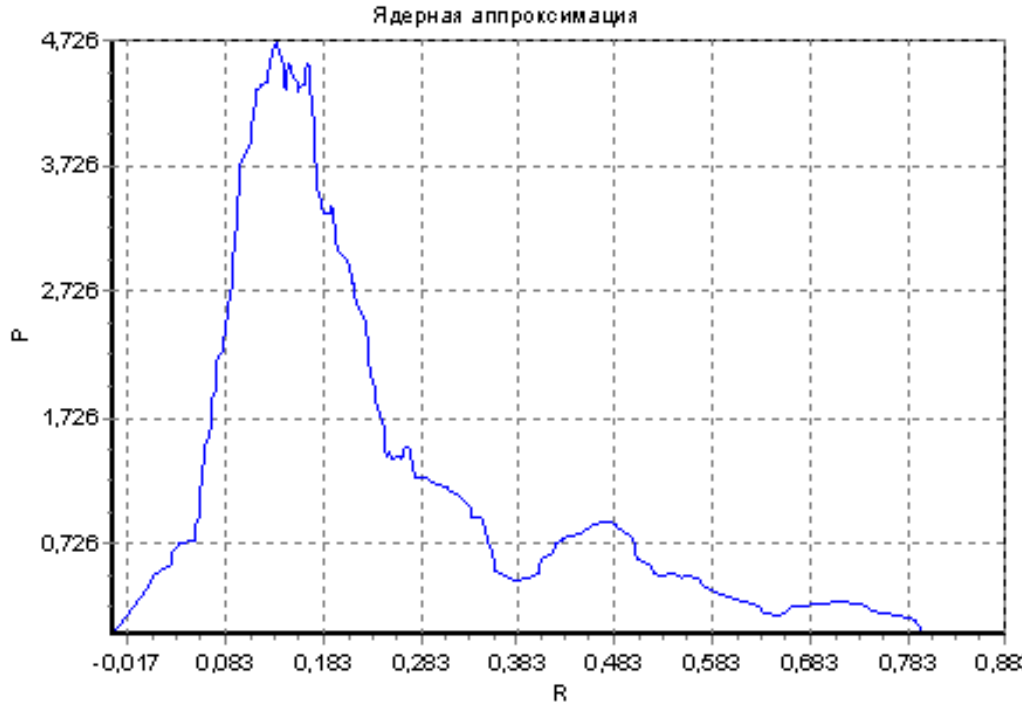

Рисунок 3 – Выборка определения фенотипа ацетилирования

Фенотип N-ацетилирования определялся как скорость ацетилирования изониазида и рассчитывался как отношение ацетизониазида к изониазиду. В группу здоровых добровольцев входили лица в возрасте от 25 до 58 лет: 45 (35%) мужчин и 84 (65%) женщины. Было выделено четыре одномодальных компонента смеси с точкой расщепления медленных и быстрых ацетиляторов 0,28. Соотношение медленных и быстрых ацетиляторов составило 66% и 34% соответственно, что согласуется с данными большинства регионов Европы.

С целью практического использования полученной оценки было выделено три интервала отношения R: (0 – 0.28]; (0.28 – 0.37]; [0.37 – 1). При этом левый интервал относится к достоверно медленным, правый интервал – к достоверно быстрым, а средний – к промежуточным или предбыстрым ацетиляторам. В ходе проводимого исследования не установлено взаимосвязи ацетиляторного фенотипа с полом добровольцев, их возрастом, массой тела и табакокурением.

Глубокое и детальное изучение ферментативных систем биотрансформации ксенобиотиков в конкретной популяции позволяет приблизиться к индивидуализированной терапии, подразумевающей эффективное и безопасное применение лекарственных средств. Результаты апробации разработанного программного обеспечения одномерного анализа данных по расщеплению смеси распределений подтверждены удостоверением на рационализаторское предложение за № 1067 от 23.12.2009, принятое к использованию Учреждением образования «Гомельский государственный медицинский университет», что демонстрирует его применимость при проведении исследований и практическую значимость.

Заключение. Практическая значимость методов выражена в их ориентированности на повышение интерпретируемости результатов экспертом даже на начальных этапах проведения исследований, что может предопределить ход и результативность всего дальнейшего исследования. Практическая значимость предложенного в данной работе программно-технологического обеспечения состоит в том, что на его основе можно реализовать гибкие технологии анализа целевого функционирования, позволяющие в рамках заданных ресурсов на исследование достигать оптимально возможной глубины охвата данных и знаний и находить наиболее точные прогнозные оценки, адекватные модели и классификации. Это подкрепляется тем обстоятельством, что на практическом примере исследования ацетиляторного фенотипа программа неплохо справилась с поставленной перед нею задачей.

Ограниченность данных методов и программного инструментария, в первую очередь, связана с необходимостью привлечения к моделированию и анализу высококвалифицированных экспертов, так как в программном обеспечении всё ещё присутствует ряд недостатков, основными из которых являются: недостаточная проработанность уровня защиты от некорректных действий пользователя, отсутствие на данный момент контекстной помощи, а также сопровождения пользователя в процессе проведения им анализа данных, концептуального моделирования, формирования целевого свойства. И также, как уже отмечалось выше, выявленная невозможность применения описанного алгоритма к смеси неразнесённых компонент хоть и преодолима при дальнейшем исследовании в рамках программы «Strand» при использовании других функциональных модулей приложения и проведении с их помощью иных форм анализа данных, но разрешить такую ситуацию также не под силу неспециалисту.

СПИСОК ЛИТЕРАТУРЫ

1. *Васенда, М.Н.* Программное обеспечение статистического описания и регрессионного анализа экспериментальных данных / М.Н. Васенда // Творчество молодых 2009: сб. науч. работ студентов и аспирантов УО «ГГУ им. Ф.Скорины»: в 2 ч. – Гомель: ГГУ им. Ф. Скорины, 2009. – Ч. 1. – С. 101–104.

2. *Осипенко, Н. Б.* Методические и программно-технологические средства оценки и анализа сезонной динамики доз внутреннего облучения жителей населенных пунктов / Н. Б. Осипенко [и др.] // Известия Гомельского государственного университета имени Ф. Скорины. – 2004. – № 6 (27). – С. 171–176.

3. *Осипенко, Н.Б.* Построение модели факторов здоровья сельского населения по данным скринингового обследования / Н. Б. Осипенко [и др.] // Известия Гомельского государственного университета имени Ф. Скорины. – 2006. – № 4 (37). – С. 113–115.

4. *Стрибук, П.Н.* Метод и средства одномерного анализа данных при исследовании социальных и природных объектов / П.Н. Стрибук // Современные проблемы математики и вычислительной техники: материалы II региональной конференции молодых ученых и студентов, Брест, 28-30 ноября 2001 г. – Брест: БГТУ, 2001. – С. 155–158.

5. *Сатырова, Т.В.* Вариабельность фенотипа N-ацетилтрансферазы у пациентов с язвенным колитом / Т.В. Сатырова [и др.] // Вестник Витебского государственного медицинского университета. – 2010. – Т. 9, №1. – С. 42–47.

Поступила в редакцию 26.05.10.

The algorithms and their realization in the software program «Strand» are described in the article. It allows to carry out the data analysis and implement flexible technologies of the analysis of the target functioning. It also allows to achieve the optimum possible depth of the data scope, to gain knowledge and to find the most precise predictable estimates, adequate models and classifications. The automated testing module was developed. It uses Statistica 7 for creation tests and checks performance results. It serves to reduce time of testing and to simplify its process. The research of acetylator phenotype on inhabitants of the Gomel region with reference to the task of individualization pharmacological therapy was carried out using the data analysis in «Strand» program.

Key words: software program, medico-ecological system.