УДК 519.254

Инструментарий верификации программных средств одномерного анализа данных об активных системах

Н.Б. Осипенко, М.Н. Васенда
Учреждение образования «Гомельский государственный университет им. Ф. Скорины»

В работе описывается метод одномерного анализа данных, реализованный в программном обеспечении «Strand», предназначенном для автоматизации исследования активных систем (АС). Разработанный комплекс алгоритмов может практически в полном объеме покрыть нужды эксперта при проведении одномерного и разведочного анализа данных и формировании на их основе представлений о характере исследуемой АС. Описывается также разработанный в рамках пакета Statistica модуль автоматизированного создания тестов, с использованием которого проведены апробация и верификация разработанных методов и функциональных возможностей программного обеспечения «Strand». Верификация результатов с использованием данного модуля тестирования практически полностью подтвердила их корректность. В работе определен 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 потребностей экспертов.

Ключевые слова: *автоматизация, активные системы, модуль тестирования, набор данных, одномерный анализ данных, расщепление смеси распределений, статистика.*

Verification tools for software of unidimensional data analysis of active systems

N.B. Osipenko, M.N. Vasenda
Gomel State University named after Francysk Skaryna

The algorithms of unidimensional analysis have been developed for the software «Strand», which are intended for automated research of active systems (AS). The algorithm complex, which has been elaborated, can practically completely cover the needs of an expert during unidimensional and reconnaissance data analysis as well as the picture of an investigated AS based on them. Also, the automated testing module was developed which uses Statistica packet for creation of tests and checks of results of performance that assists to reduce time of testing and to simplify its process. The result verification of unidimensional analysis algorithms with the use of this testing module has completely confirmed their correctness. The paper states a number of problems which are to be solved in future during other types of data analysis in Stand appendix because the algorithms described here do not meet the whole spectrum of experts' needs.

Key-words: *active systems, automation, data scope, splitting of distributions mix, testing module, statistics, unidimensional data analysis.*

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. Посредством «Strand» был проведен анализ данных [4], полученных при обследовании 129 практически здоровых добровольцев контрольной группы из популяции г. Гомеля и Гомельской области для определения фенотипа ацетилирования, результаты которого согласуются с данными исследований большинства регионов Европы. Целью настоящей работы является проведение полной верификации результатов работы данного инструментария.

Алгоритм одномерного анализа данных. Полное описание алгоритма одномерного статистического анализа данных представлено в работе [5]. Обобщенная блок-схема данного алгоритма, реализованного в «Strand», представлена на рис. 1.

В процессе всей работы над выборкой учитывается ее объем. Так, по размеру выборки подразделяются на 3 класса: малые – до 30, средние – от 30 до 100, большие – более 100 элементов. Данное деление учитывается при «ядерной» аппроксимации (в зависимости от класса выбирается размер «ядра»), при проверке гипотез (критическая область), в процессе квантильного анализа (количество квантилей).

Также имеется возможность для каждой выборки задать веса, которые будут учитываться при анализе (построении нормальной вероятностной бумаги на каждом этапе), что, например, позволяет корректно обрабатывать данные, полученные из различных источников. Часто встречающиеся на практике пропуски в данных могут в процессе анализа игнорироваться или заполняться согласно Zet-алгоритму [6, с. 414], как содержащие значения среднеарифметического или средневзвешенного по данному при-

знаку наиболее коррелируемых объектов.

Таким образом, осуществляется достаточно гибкий подход при одномерном анализе тех или иных выборок.

Обзор инструментария верификации. Для реализованных алгоритмов проведения одномерного анализа была предоставлена возможность проходить апробацию на серии создаваемых автоматически, специально разработанных тестов.

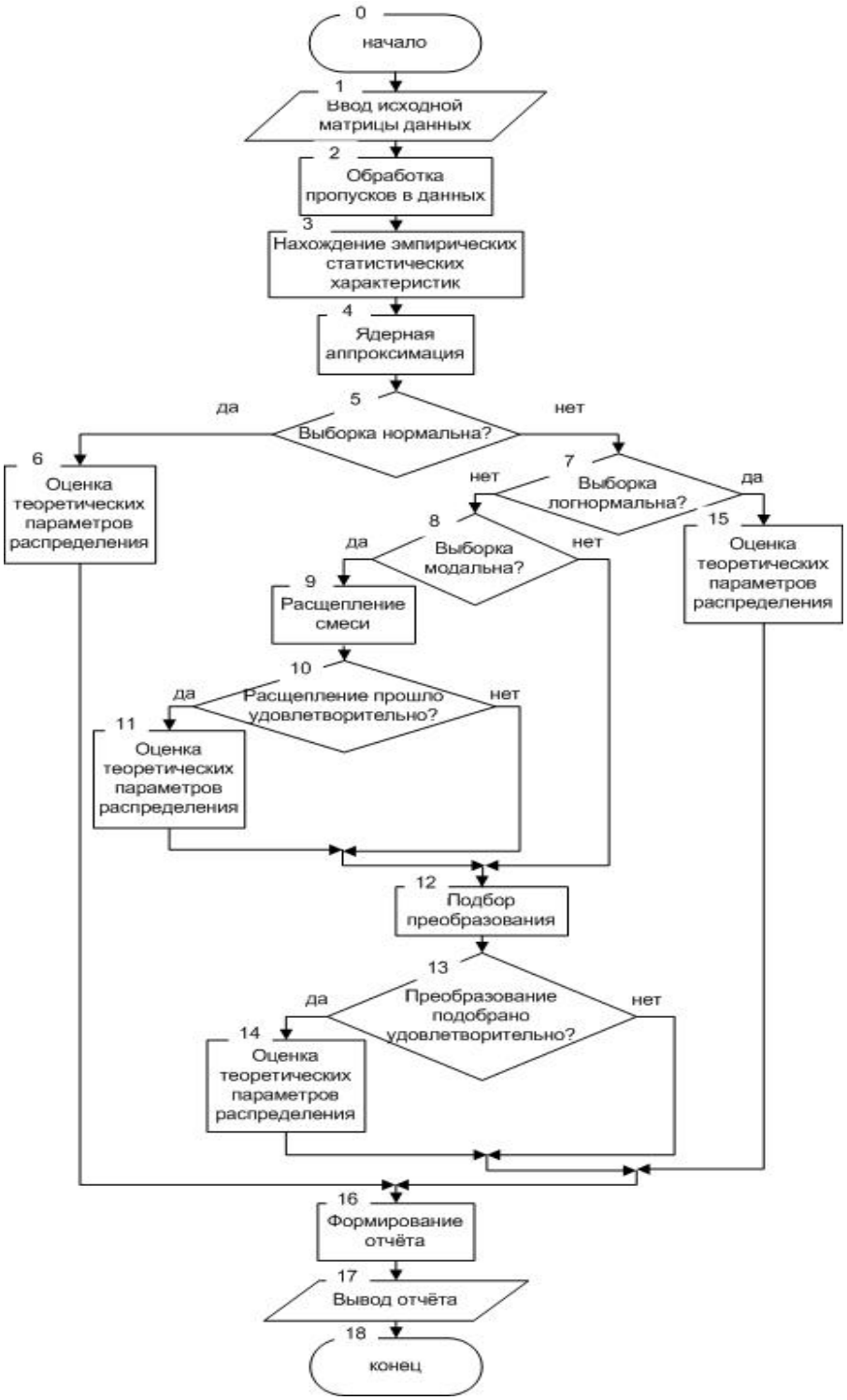

Рис. 1. **Схема программного инструментария одномерного анализа данных разработанного приложения «Strand».**

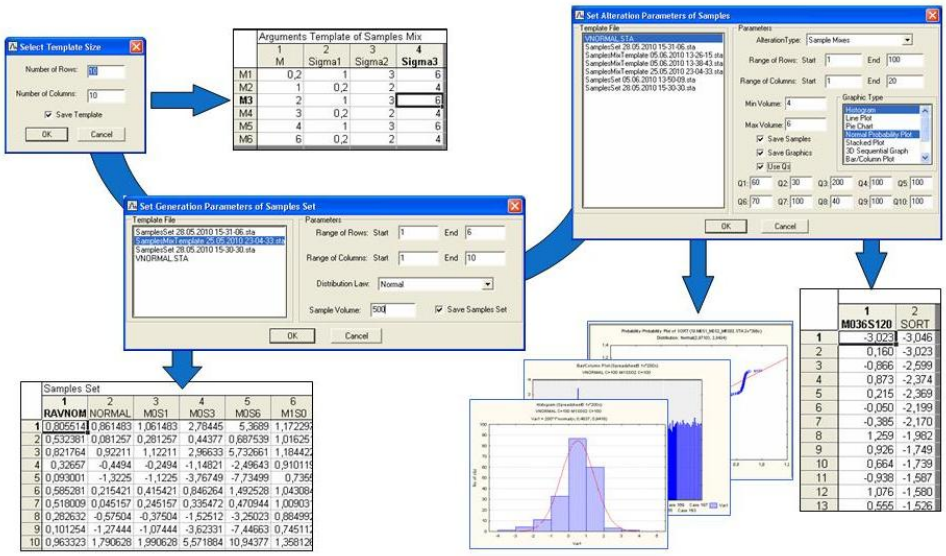

Рис. 2. **Схема работы модуля автоматизированного создания тестов в пакете Statistica.**

Для автоматизации при верификации разработанных программных средств статистического анализа данных и описанных выше элементов алгоритма одномерного анализа был разработан модуль на языке программирования SVB (Statistica Visual Basic), предназначенный для создания тестовых данных. Модуль работает в пакете Statistica.

Подход, реализованный в модуле, при создании тестов предполагает последовательность из трех этапов, работа которых представлена на рис. 2 и описана ниже.

1. Создание шаблона с заполнением его первоначальными параметрами в соответствии с предполагаемым видом создаваемых тестов. На этом этапе пользователю предоставляется возможность выбрать объемы используемых начальных данных и параметр сохранения. В результате пользователь получает на выходе шаблон для заполнения начальными данными, используемыми для построения набора тестов.

2. Создание набора одномодальных выборок. В ходе этого этапа используется заранее подготовленный пользователем шаблон с введенными начальными данными. Например, реализация автоматизации создания тестов для расщепления смеси распределений с параметрами, указанными в шаблоне, осуществляется следующим образом: получение выборки равномерно распределенных чисел на отрезке [0;1] (первая выборка) с заданным пользователем в диалоге объемом; 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. 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 определить файл, содержащий набор исходных выборок, подлежащих дальнейшему изменению, тип преобразования (например, для расщепления – тип смешивания), максимальное количество компонент, используемых в создаваемых наборах, их априорные вероятности, параметры сохранения выборок и графиков, и определить необходимые визуализации для построенных наборов тестов.

При увеличении предполагаемых компонент полученный набор тестов может оказаться достаточно большим, но все же его создание происходит быстрее ручного. При создании больших наборов выборок целесообразно пользоваться возможностью использования только части данных из шаблона, содержащего начальные параметры.

Результаты апробации алгоритма и разработанного программного обеспечения. В ходе проведения верификации, с использованием сгенерированного набора тестов, была проверена работа всех реализованных в приложении алгоритмов. Ниже приведены некоторые результаты апробации.

Оценка качества восстановления пропусков проводилась при помощи тестового набора, в котором была удалена часть данных. Мерой качества восстановления пропущенных элементов служила мера отклонения (сумма квадратов отклонений) истинных значений от значений, полученных в результате восстановления пропусков. Реализованный в приложении Zet-алгоритм показал хорошие результаты работы.

Проверка преобразований по формулам Бокса–Кокса проводилась на тестовых наборах, также построенных в модуле тестирования. Так, наиболее классическими примерами проверки являются обратное преобразование от логнормального к нормальному и преобразование выборок с коэффициентом асимметрии, отличном от 0, к выборкам с коэффициентом, близким к 0. Рассматривая итоги, можно отметить, что преобразование по формуле с двумя параметрами [6, с. 337] приводит к более сглаженному графику и, как следствие, более симметричной функции распределения.

В ходе проведения тестирования алгоритма расщепления смеси, с использованием сгенерированного набора тестов, была выявлена неспособность программы определять наличие смеси распределений у выборок, компоненты которых слабо разнесены (незначительно разнятся математические ожидания у больших выборок). В то же время программа и, соответственно, реализованный в ней алгоритм, хорошо выполняют расщепление смесей распределений при достаточно разнесенных компонентах. Выявленный недостаток преодолим при дальнейшем исследовании в рамках программы «Strand», с использованием других функциональных модулей приложения и при проведении с их помощью, например, кластерного анализа.

Оценка результатов работы данного алгоритма расщепления проводилась посредством вычисления средних ошибок математического ожидания, дисперсии и объемов компонент. Рассчитанные оценки показали, что в тех случаях, когда возможно проведение расщепления смеси, реализация алгоритмов в программе допускает малые средние ошибки, которые напрямую связаны с соотношением весов, составляющих смесь, компонент, а также с их разнесением, которое зависит от разницы математических ожиданий и объемов компонент.

Заключение. Практическая значимость метода выражена в его ориентированности на повышение интерпретируемости получаемых результатов экспертом даже на начальных этапах, что может предопределить ход и результативность всего дальнейшего исследования. Разработанный модуль позволяет создавать тесты в автоматическом режиме и проводить на их основе верификацию, что позволяет уменьшить время тестирования и упростить его процесс.

Ограниченность данного метода связана с необходимостью привлечения к моделированию высококвалифицированных экспертов в статистическом анализе данных. В модуле тестирования недостаточно проработана защита от некорректных действий пользователя и отсутствует его сопровождение.

Универсальность разработанных алгоритмов и программных средств состоит в том, что они обеспечивают почти в полном объеме проведение одномерного статистического анализа, а также, что верификация результатов работы приложения проходит в полуавтоматическом режиме.

Как дальнейшее развитие этой тематики предполагается расширять данный модуль автоматизированного тестирования для использования при отладке работы других, уже существующих и новых компонентов разрабатываемого программного обеспечения и реализовать механизм, позволяющий автоматизировать непосредственно сам процесс тестирования текущей версии продукта.

ЛИТЕРАТУРА

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 Франциска 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 Франциска Скорины. – 2006. – № 4(37). – С. 113–115.
4. 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.
5. Осипенко, Н.Б. Математическое обеспечение корректности при оценке распределения экологического параметра по разнотипным статистическим описаниям / Н.Б. Осипенко, А.Н. Осипенко, М.Н. Васенда // Юбилейная научно-практическая конференция, Гомель, 11 июня 2009 года: посвящена 40-летию Гомельского государственного университета имени Франциска Скорины: материалы: в 4 ч. / редкол.: О.М. Демиденко (отв. ред.) [и др.]. – Гомель: ГГУ им. Ф. Скорины, 2009. – Ч. 4. – С. 159–162.
6. Айвазян, С.А. Прикладная статистика: основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 472 с.

Поступила в редакцию 23.08.2010

Адрес для корреспонденции: г. Гомель, ул. Жукова, д. 52, кв. 197, +375-44-969-59-15 – Осипенко Н.Б.