УДК 004.414

А. С. Бобылев*
Т. И. Ведерникова**

ВЫБОР СРЕДСТВ РАЗРАБОТКИ ПРОГРАММНОГО КОМПОНЕНТА ДЛЯ РЕАЛИЗАЦИИ МОС-МЕТОДА

В статье рассматривается проблема обработки многомерных статистических данных. В качестве решения предлагается использование метода моделирования статистик, различающих случайные величины (МОС-метод). Проводится обзор существующих средств разработки программ статистической обработки данных, а также осуществляется выбор одного из них для реализации МОС-метода, по заданным критериям.

Ключевые слова: МОС-метод, средства обработки многомерных данных, выбор средств разработки, статистическая обработка информации.

Существующие сегодня инструменты и средства позволяют детально описывать наблюдения за объектом (или явлением) исследования, что порождает огромные объемы информации. Часто такие данные представляют собой многомерные временные ряды. В большинстве случаев с практической точки зрения исследователей интересует не вся последовательность наблюдений, а аномальные явления или те явления, которые предшествуют или происходят сразу после него. Идентификация аномального явления по совокупности измерений является актуальной задачей распознавания образов.

Появление большого количества технических устройств, работа которых напрямую зависит от распознавания текущего состояния объектов, процессов, явлений и состояний, с которыми эти устройства работают, является одной из основных причин активного расширения области практического применения систем распознавания [1].

Эффективность решения задачи распознавания образов связанна с выбором признаков, характерных для данного объекта или явления. Набор таких признаков называют рабочим словарем признаков (РСП). Одним из методов построения рабочего словаря является метод моделирования статистик, различающий случайные величины. Этот метод базируется на переходе от многомерных случайных величин к одномерным, аккумулирующим отличительные особенности многомерных величин. Это позволяет существенно снизить трудоемкость решения задачи идентификации, т.к. анализ свойств одномерных случайных величин выполнить значительно проще, чем многомерных. Основополагающим для МОС-метода является предположение о том, что объекты разных областей являются реализациями различных случайных величин. Следова-

* Бобылев Андрей Сергеевич – магистрант, кафедра информатики и кибернетики, Байкальский государственный университет, г. Иркутск.

** Ведерникова Татьяна Ивановна – доцент, кафедра информатики и кибернетики, Байкальский государственный университет, г. Иркутск.

тельно, для каждой области существуют свои характерные преобразования, отражающие внутреннюю структуру объекта идентификации [2–3].

В настоящее время имеется большое количество пакетов программ, выполняющих статистическую обработку информации. Однако универсальных средств, позволяющих исследователю выполнить предварительно функциональные преобразования, специфические для изучаемого объекта или явления, с целью дальнейшей его идентификации, пока не обнаружено. Необходимо создать программу, способную: первое – выполнять различные «свертки» наблюдений за многомерными объектами, и второе – взаимодействовать с популярными средствами обработки и анализа статистических данных на программном уровне.

Статистические пакеты анализа данных можно разделить на две большие группы [4]:

– ориентированные на анализ и программирование: R, MATLAB, OCTAVE, NumPy (библиотека Python);

– ориентированные только на анализ данных: SAS, SPSS, Stata, MS Excel.

Пакеты первой группы позволяют произвести более гибкий подход к разработке, а также дают возможность в будущем расширять и дополнять программный компонент новыми функциями.

Одним из главных критериев по выбору инструмента реализации является его стоимость, поэтому рассматривались только бесплатные и с открытым программным кодом наиболее популярные программные пакеты:

– Microsoft R;
– Python (NumPy, SciPy, Pandas);
– GNU Octave.

Эти инструменты гибкие и простые в использовании для векторизации и матричных операций. Кроме того, они являются не только программами для анализа данных, но и языками программирования для создания собственных функций и пакетов, что делает их подходящими инструментами для реализации программного компонента МОС-метода.

Microsoft R – это мощная свободно распространяемая статистическая среда, которая включает в себя: программирование, интерактивную оболочку и широкие возможности по отображению графической информации. Более того, R имеет огромный набор математических и статистических функций, а также дополнительные возможности, которые предоставляются в подключаемых пакетах. R применяется везде, где нужна работа с данными. R может использоваться там, где принято использовать специализированные программы математического и статистического анализа. Географически R распространен очень широко. Трудно найти американский или западноевропейский университет, где бы не работали с R. Очень многие серьезные компании (например, Boing) устанавливают R для работы [5].

Python (NumPy, SciPy, Pandas) – высокоуровневый язык программирования с динамической типизацией, поддерживающий объектно-ориентированный, функциональный и императивный стили программирования. Это язык общего назначения, на котором можно одинаково успешно разрабатывать системные

приложения с графическим интерфейсом, утилиты командной строки, научные приложения, игры, веб-приложения и много другое. Поскольку Python является языком общего назначения, то работа в различных предметных областях осуществляется с помощью специальных библиотек [6].

В области анализа данных и интерактивных научно-исследовательских расчетов с визуализацией результатов Python неизбежно приходится сравнивать со многими предметно-ориентированными языками программирования и инструментами такими, как R, MATLAB, SAS, Stata и другими. Появление улучшенных библиотек для Python (прежде всего, pandas) сделало его серьезным конкурентом в решении задач манипулирования данными. В сочетании с достоинством Python как универсального языка программирования это делает его отличным выбором для создания приложений обработки данных [7].

GNU Octave является языком высокого уровня и, в первую очередь, предназначен для численных расчетов. Он предоставляет удобный интерфейс командной строки для решения линейных и нелинейных задач, а также для выполнения других численных экспериментов с использованием языка, который в основном совместим с Matlab. Он также может быть применен в качестве пакетно-ориентированного языка. Octave имеет обширные средства для решения общих численных задач линейной алгебры, нахождения корней нелинейных уравнений, интегрирования функций, манипулирования полиномами и интегрирования обыкновенных дифференциальных и дифференциально-алгебраических уравнений. Он легко настраивается и расширяется с помощью определяемых пользователем функций, написанных на родном языке Octave, или с использованием динамически загружаемых модулей, написанных на C ++, C, Fortran, или других языках [8].

Выбор средства реализации программного компонента МОС-метода определялся следующими критериями:

1) квалификация программиста;
2) работа с большими объемами многомерных данных;
3) визуализация результатов обработки;
4) возможность функционального расширения;
5) взаимодействие с внешними программами;
6) сложность изучения.

Больших затрат требуют услуги высококвалифицированных программистов, которых на рынке труда не так много. Пользователи, на которых ориентирован программный компонент МОС-метода, это в большей степени ученые и аналитики, для которых программирование не является основным видом деятельности. Но иногда в работе требуется изменить программу под свои нужды, не прибегая к помощи программиста.

Наблюдение за объектом исследования порождает большое количество данных, поэтому необходим инструмент, с помощью которого можно создать программный компонент, способный обрабатывать большое количество данных. Представление обработанных статистических данных в графическом виде является неотъемлемой частью аналитических пакетов, это делает информацию более удобной для восприятия и дальнейшего анализа. Поэтому визуализация является одним из важнейших критериев при выборе средства разработки.

Возможность дополнения компонента новыми функциональными преобразованиями является решающей при выборе средства реализации. На данный момент существует большое количество продуктов, которые проводят статистическую обработку, требуется, чтобы компонент МОС-метода взаимодействовал со сторонними программами без сложной надстройки и доработки. Сложность изучения языка программирования напрямую влияет на скорость разработки и расширения компонента.

В соответствии с определенными критериями выполнено сравнение рассматриваемых средств разработки (2 – «отлично», 1 – «хорошо», 0 – «плохо»). Выбор был сделан в пользу системы Microsoft R. Результаты представлены в таблице.

Сравнение средств разработки данных

Средство разработки	Критерий						Итого
	1-й	2-й	3-й	4-й	5-й	6-й	
R	2	1	2	2	2	1	10
Python	2	2	1	2	1	1	9
Octave	2	1	1	2	1	1	8

Список использованной литературы

1. Волченко Е. В. Метод w-mief построения рабочего словаря признаков на основе взвешенных обучающих выборок [Электронный ресурс] / В. С. Степанов, Е. В. Волченко // Вестник НТУ «ХПИ»: науч.-метод. журн. – 2012. – № 18. – Режим доступа: http://repository.kpi.kharkov.ua/bitstream/KhPI_Press/ 9948/1/vestnik_HPI_2012_18_Volchenko_Metod.pdf.

2. Ведерникова Т. И. Один подход к построению рабочего словаря признаков для идентификации реализаций векторных случайных величин / Т. И. Ведерникова, Н. Ю. Дунаева // Применение математических методов и информационных технологий в экономике. – Иркутск: Изд-во ИГЭА. – 2001. – С. 33–35.

3. Ведерникова Т. И. Способы построения рабочего словаря признаков для решения задач идентификации / Т. И. Ведерникова [Электронный ресурс] // Известия ИГЭА (БГУЭП). – 2012. – № 4. – Режим доступа: http://eizvestia.isea.ru

4. Чучуева И. Сравнение программных продуктов для анализа данных [Электронный ресурс] / И. Чучуева. – Режим доступа: http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-r-matlab-scipy-ms-excel-sas-spss-stata.

5. Шипунов А. В. Наглядная статистика. Используем R! [Электронный ресурс] / А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников, С. А. Назарова, С. В. Петров В. Г. Суфиянов. – Режим доступа: http://herba.msu.ru/shipunov/school/books/rbook.pdf.

6. Справочник по языку программирования Python [Электронный ресурс] – Режим доступа: http://pythonz.net/promo.

7. Маккинли У. Python и анализ данных / пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 482.

8. Официальный сайт GNU Octave [Электронный ресурс] – Режим доступа: https://www.gnu.org/software/octave/