УДК 004.422.837

АНАЛИЗ И ПРИМЕНЕНИЕ СПЕЦИАЛЬНЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕК ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ

К. т. н. Н. О. Комлевая, М. А. Крупенко, Д. Д. Бондаренко

Одесский национальный политехнический университет
Украина, г. Одесса
nokoml@yandex.ua

Выполнен краткий анализ математических библиотек, имеющих в своем составе функции для проведения статистической обработки информации. Рассмотрены средства библиотеки Armadillo для C++, работающей с матричными функциями. Реализован с их помощью алгоритм дискриминантного анализа для проведения статистического анализа медико-биологических данных.

Ключевые слова: математическая библиотека, статистический анализ, Armadillo C++.

На сегодняшний день существует большое количество библиотек математических функций, позволяющих облегчить реализацию сложных вычислительных задач и алгоритмов для обработки данных. При решении ряда медико-биологических задач часто возникает необходимость выполнения математической обработки и анализа многомерных данных. Для автоматизации этих процессов удобно использовать специальные программные средства, которые, в частности, содержат в себе функции статистики, генераторы случайных чисел, функции линейной алгебры, комплексных чисел, преобразований Фурье (и других преобразований), матричные функции (перемножение матриц, обратные матрицы, определители всех порядков), функции для работы с числами повышенной точности и т. д.

Целью работы является выполнение статистической обработки медико-биологических данных с помощью специальных функций математических библиотек. При этом необходимо получить комбинированную оценку наличия либо отсутствия существенного различия в наблюдаемых объектах по величине нескольких признаков.

Для проведения статистического анализа данных предложено использовать дискриминантный анализ – совокупность методов, позволяющих решать задачи идентификации и классификации объектов по заданному набору характерных признаков [1]. Изучаемые объекты редко различаются по величине только одного признака, обычно их несколько. Если производится оценка существенности – значимости различия только по величине отдельных признаков, характеризующих изучаемые случаи, можно получить неверные односторонние сведения [2]. В данной работе рассматривается учебная задача, предусматривающая статистическую обработку больших массивов медико-биологических данных. Решение задачи должно учитывать 32 независимых признака для построения разграничительной функции.

Для решения статистических задач существуют специальные многофункциональные программные пакеты. Однако их использование в учебных проектах зачастую оборачивается рядом неприятных аспектов: необходимость приобретения лицензионного ПО, отсутствие встроенного типизированного языка программирования, низкое быстродействие вследствие интерпретации кода во время выполнения, сложности с автоматизацией импорта/экспорта данных.

Рассмотрим особенности некоторых математических библиотек, позволяющих, в-первых, автоматизировать математическую и статистическую обработку данных, и, во-вторых, пройти последовательно этапы обработки данных в учебных целях. Среди таких библиотек для языка программирования Java хотелось бы отметить следующие: COLT – библиотека, содержащая матричные функции, различные статистические распределения, генераторы случайных чисел; Commons–Math – библиотека, позволяющая работать со статистическими функциями, генераторами случайных чисел, линейной алгеброй, комплексными числами, однако не содержащая матричных функций. Аналогичные библиотеки для языка программирования C#.NET: Numerical Methods on C# – мощная библиотека, со-

держащая в себе аппроксимации, интерполяции, интегрирование функций, линейные системы уравнений, матрицы, элементы статистики (генераторы чисел с нормальным распределением, гамма-распределением и т. д.); Npack – библиотека, созданная для операций над матрицами и векторами, однако реализация некоторых функций не доведена до конца. Библиотеки для Python: SymPy – активно развивающаяся библиотека для символьных вычислений, включающая в себя модуль для работы с матрицами (модуль линейной алгебры), модуль геометрии, модуль статистики, с помощью которого можно получать случайные величины с заданной функцией распределения плотности вероятности, модуль для отображения трехмерных поверхностей, заданных в виде уравнений с символьными переменными; NumPy – математический пакет, включающий работу с матрицами и векторами, быстрое преобразование Фурье, компиляцию модулей на фортране, работу с полиномами, функции для линейной алгебры. Библиотеки для C++: библиотека матричных функций Armadillo, поддерживающая операции нахождения обратных матриц, определителей, перемножения матриц и многое другое.

Для реализации учебной задачи выбран язык программирования C++, что обусловило использование средств библиотеки Armadillo. Все матрицы в этой библиотеке представлены как объект класса mat. Для того, чтобы задать матрице определенные значения, используется метод .set_size (rows, cols), принимающий как параметры количество строк и столбцов будущей матрицы. Методы .save (name, format) и .load (name, format) позволяют загружать и сохранять матрицы из различных файлов. Благодаря возможности переопределения операторов в C++, такие операции как суммирование двух матриц, их перемножение и т. д. выполняются с помощью обычных операторов +, –, ×. Метод inv(X) позволяет создать матрицу, обратную данной, det(A) – метод, позволяющий найти определитель данной матрицы, rank(A) – ее ранг, доступ к каждому элементу в матрице осуществляется с помощью конструкции A(*r*, *c*), где *r*, *c* – строка и, соответственно, столбец матрицы. С помощью данных средств пройдены следующие этапы обработки: 1) проверка соблюдения условий для модели дискриминации [3]; 2) инициализация переменных; 3) вычисление значений переменных, используемых в предсказании принадлежности к классу объектов, в указанном наблюдении в определенной группе; 4) расчет внутригрупповой матрицы рассеяния наблюдаемых переменных от средних; 5) вычисление матрицы рассеяния; 6) выбор переменных для анализа [4]; 7) формирование массива средних значений; 8) формирование массивов коэффициентов и констант, классифицирующих функции; 9) формирование функции классификации [5]. Таким образом, выполнена статистическая обработка медико-биологических данных с помощью средств библиотеки Armadillo.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Ким Дж.-О. Факторный, дискриминантный и кластерный анализ.— Москва: Финансы и статистика, 1989.

2. Комлевая Н. О. Построение системы диагностических признаков с использованием метода дискриминантного анализа в офтальмологических исследованиях // Радіоелектронні і комп'ютерні системи.— 2010.— Вип. 6 (47).— С. 250 – 253.

3. Комлевая Н. О., Бондаренко Д. Д., Крупенко М. А., Боренко А. С. Методика нормализации несвязанных данных в условиях неоднородности выборки // Труды XV МНПК «Современные информационные и электронные технологии».— Украина, г. Одесса..— 2014.— Т. 1.— С. 16–17.

4. Комлевая Н. О., Махиненко А. Ю., Луговской В. А., Провоторов В. В. Использование критерия Стьюдента для выявления статистически значимых различий в задачах классификации // Труды XV МНПК «Современные информационные и электронные технологии».— Украина, г. Одесса.— 2014.— Т. 1.— С. 18–19.

5. Реброва О. Ю. Статистический анализ медицинских данных.— Москва: Медиа Сфера, 2006.

N. O. Komlevaya, M. A. Krupenko, D. D. Bondarenko

Analysis and application of special mathematical libraries for statistical processing of medical and biological data.

A brief analysis of mathematical libraries in which there are functions for statistical data processing is made. Armadillo library facilities for C++, which works with matrix functions, are considered and used for realization of discriminant analysis algorithm for statistical analysis of medical and biological data.

Keywords: *mathematical library, statistical analysis, Armadillo C++*.