JACOBI–4 – ПРОГРАММНО-АЛГОРИТМИЧЕСКИЙ КОМПЛЕКС ДЛЯ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ

В.М. Ефимов, д-р. биол. наук, профессор **(НГАСУ (Сибстрин), Новосибирск), И.А. Штайгер, Д.А. Полунин,** магистранты **(НГУ, Новосибирск)**

Несмотря на огромное количество специализированных программных средств, предназначенных для анализа многомерных данных, на практике ими не слишком удобно пользоваться, не говоря уже о том, что часть из них является коммерческими. Основная проблема – закрытая реализация, не позволяющая пользователю в полной мере осуществить необходимую ему обработку своих экспериментальных результатов. Поэтому разработка простого и удобного программного комплекса для анализа многомерных данных, доступного для любого пользователя без специального образования, представляется весьма актуальной.

Пользователю нужна программная система, обслуживающая реальные практические запросы и сочетающая свойства статистических пакетов и систем управления базами данных. Очевидно, что эти задачи противоречивы и что компромисс может быть достигнут только за счет потери эффективности на задачах, типичных для каждой из названных областей. Однако, по нашему мнению, удобство использования одной системы вместо нескольких и неуклонно растущая производительность вычислительных средств делают такой компромисс рентабельным.

Разрабатываемая нами система состоит из головной программы-диспетчера и комплекса независимо реализованных исполняемых модулей. Будут реализованы различные алгоритмы многомерного анализа, включая главные компоненты, множественную регрессию, дискриминантный анализ, кластерный анализ, многомерное шкалирование (метрическое и неметрическое), многомерный PLS-анализ и нейронные сети.

Целью проекта является реализация методологии объединения разнотипных описаний одного и того же набора объектов

в виде программного комплекса с удобным и дружественным для пользователей интерфейсом. Назначение программного комплекса – решение научных проблем, связанных с обработкой разнородных описаний, в первую очередь, в биологических и экологических исследованиях, хотя ими, конечно, не ограничивается. Предлагаемая методология заключается в геометрическом представлении любого набора объектов множеством точек в евклидовых пространствах через вычисление матриц сходства-различия между их описаниями и сведение их к матрицам евклидовых расстояний. Все звенья предлагаемой методологии по отдельности давно и хорошо проработаны в современной науке, однако объединение их в единый программно-алгоритмический комплекс предлагается впервые. Реализация программного комплекса позволит решать широкий класс содержательных научных задач в различных областях биологии и экологии.

Дан набор объектов, для которых имеется несколько разнородных описаний. Например: количественные, ранговые и качественные признаки; текстовые последовательности; матрицы коэффициентов сходства-различия; иерархические деревья (дендрограммы); данные геометрической морфометрии о форме биологических объектов и т.д.

Требуется: а) оценить конгруэнтность этих описаний; б) выявить в обоих описаниях общие направления изменчивости; в) построить объединенное описание, пригодное для многомерного анализа.

Исследуемая гипотеза состоит в том, что, поскольку все описания с разных сторон отражают изменчивость одного и того же набора объектов, то их общая часть, если она есть, является наиболее перспективной для дальнейших исследований, так как отражает наиболее глубинные свойства объектов, проявляющиеся во всех описаниях. В рамках данного проекта предлагается методология выявления и сопоставления общей части всех описаний, а также построения объединенного описания. Методология базируется на возможности геометрического представления объектов в виде точек евклидова пространства невысокой размерности для любых типов описаний и заключается в сле-

дующем. Для каждого типа данных строится своя матрица расстояний между объектами. Если она не является евклидовой, то предварительно переводится в евклидову с помощью многомерного шкалирования. Оценивается конгруэнтность матриц расстояний. Матрицы нормируются, и по ним вычисляется объединенная евклидова матрица расстояний между объектами. По всем матрицам расстояний между объектами вычисляются представляющие их евклидовы пространства. Системы координат в этих пространствах вращаются до получения максимального соответствия между конфигурациями объектов. Далее конфигурации объектов исследуются средствами многомерного анализа.

Все звенья предлагаемой методологии, как частные случаи, по отдельности давно и хорошо проработаны в современной науке, однако в разных предметных областях. В связном виде она предлагается впервые. Нам неизвестна ни одна работа, кроме наших [1–3], в которых эта логика была бы последовательно проведена от начала до конца. Причина, по-видимому, заключается в том, что исследователи редко заглядывают в «чужие» области.

Кроме того, самой существенной частью проекта является реализация новой методологии в виде программного комплекса с удобным и дружественным для пользователей интерфейсом, поскольку только в таком виде возможно ее массовое испытание, применение и распространение. Судя по литературе, легко доступных и специально предназначенных для биологов программных средств, нацеленных на решение именно этого типа задач, в мире нет. Ближайшими аналогами являются R-пакет (по универсальности) и PAST (по удобству для биологов). Но их возможностей явно недостаточно для массового применения предлагаемой нами методологии, так как обработка с их помощью всего спектра возможных вариантов входных данных весьма затруднительна, а для непрофессионалов в области обработки данных практически нереальна.

Следует особо оговорить, что вариант с использованием стандартных для пользователей с техническим образованием

инструментальных средств (Matlab, R, Pyton) нами, конечно, рассматривался как самый очевидный, но был отвергнут, прежде всего потому, что разрабатываемый нами инструмент предназначен для пользователей, практически не знакомых с программированием. У биологов другое образование, и это надо учитывать в первую очередь. Кроме того, на сегодня и в универсальных языках типа C + + имеются достаточно богатые библиотеки анализа данных, позволяющие быстро включить их в состав любого программного комплекса. Основная проблема вовсе не в этом, а в том, чтобы обеспечить максимальные удобства по увязке этих возможностей в единый, легко и просто собираемый, и при этом заранее неизвестный и даже, возможно, корректируемый по ходу обработки сценарий выполнения задания. Особенно важно это еще и потому, что для биологов имеет огромное значение расчет достоверности получаемых результатов. На сегодня наилучшим универсальным средством для этого является бутстреп-анализ, который требует многократного (от тысяч до миллионов раз) просчета тех же данных с намеренно вносимым случайным искажением. Поэтому всякие, даже минимальные технические затруднения или ограничения, а тем более ручное вмешательство пользователя (например, для экспорта файлов из одного пакета или формата в другой), должны быть просто исключены. Точно так же должны быть максимально устранены все проблемы, связанные с объемом обрабатываемых данных (у биологов на сегодня это могут быть сотни тысяч и даже миллионов признаков для нескольких сот или тысяч объектов). По нашему замыслу, пользователь должен только ясно представлять конечную задачу и уметь скомпоновать (или даже только подправить) не слишком сложный сценарий обработки его данных, составленный на понятном и доступном ему языке, а остальное – дело разработчиков.

Разработка программного комплекса поддержана грантом РФФИ № 13–07–00315 «Интеллектуальный анализ и комбинирование гетерогенных данных».

Список литературы

1. Программно-алгоритмический комплекс для многомерного анализа микрочиповых данных / В. М. Ефимов [и др.] // Материалы II Междунар. науч.-практ. конф. «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика». – Новосибирск, 2011. – С. 120.
2. Efimov V. M. Heterogenic data mining and combining / V. M. Efimov, V. Y. Kovaleva // 8-th Int. Conf. on Bioinformatics of Genome Regulation and Structure\Systems Biology. – Novosibirsk, 2012.
3. Анализ соответствия и комбинирование молекулярно-генетических и морфологических данных в зоологической систематике / Ковалева В. Ю. [и др.] // Известия РАН. – 2012. – № 4. – С. 404–414.