Автоматизация анализа данных экспериментальных исследований

С.А. Заруцкий[1,2], Е.А. Власенко[2]

[1]Ростовский государственный медицинский университет, Ростов-на-Дону
[2]Аналитическая компания Статзилла

Аннотация: Рассматриваются устройство и работа платформы для автоматизированной обработки данных, предоставляющей пользователю связный текстовый отчет с результатами анализа.
Ключевые слова: Анализ данных, статистика, обработка данных, автоматизация, статистические методы, статистические критерии, генерация отчета, интерпретация результатов, медико-биологические исследования, качество исследований, достоверность исследований.

Исходя из анализа современной методики прикладных исследований, можно прийти к выводу о том, что построение алгоритмов характерно для набора статистических методов и тестов, тогда как логика их использования, подготовки и обработки данных, а также написания текста с результатами анализа хотя и очень шаблонна, но до сих пор не автоматизирована. В то же время, автоматическая генерация текста позволит облегчить работу и экспертам по статистике, которые после получения оформленного по ГОСТ текста, таблиц и рисунков, могут при необходимости редактировать их вместо создания с нуля. В связи с этим, актуальность поставленной проблемы обусловлена высокой частотой ошибок в расчетной статистической части доказательных исследований в условиях сложной рутинной аналитики, которая занимает много времени при обработке данных представителями других отраслей науки, а также высокой стоимостью работ при обращении к специалистам, затратами на специализированное ПО.

Целью проводимой научно-исследовательской работы является создание платформы для автоматизированной обработки данных, предоставляющей пользователю связный текстовый отчет с результатами анализа. Объект исследования включает в себя практику применения методов статистической обработки данных в экспериментальных

исследованиях и алгоритмы генерации текста с интерпретацией результатов использования соответствующих методов. Основанием для выполнения работы является то, что согласно оценкам различных экспертов, в России около 80% статей в ведущих тематических журналах содержат ошибки в анализе данных, делающие их заведомо неверифицируемыми [1, 2]. От правильности анализа данных, собранных на выборке, напрямую зависит адекватность выводов исследования. Также примерно в 95% исследований для различных областей применения статистического анализа моделирование проблемной ситуации проводится экспертами по одному из небольшого числа шаблонов, которые, однако, до сих пор не алгоритмизированы. Таким образом, методика выполнения представленной работы, связанной с алгоритмизацией всех этих блоков, предполагает построение экспертной системы статистической обработки и интерпретации результатов анализа на основе международных практик (GCP, EBP), руководств (CONSORT, STROBE, STARD, STREGA, PRISMA, SQUIRE) и прочих правил, признанных научным сообществом [3-10].

Первый этап исследований является базисом всей научно-исследовательской работы (далее НИР), так как задает структурную схему работы платформы и определяет дальнейшее развитие работы. Цель НИР первого этапа состоит в создании концепции и основных разделов платформы для автоматизированной обработки данных. Достижение цели требует постановки следующих задач первого этапа исследований:

1) проанализировать частоту применения статистических методов в экспериментальных доказательных исследованиях;
2) проинтервьюировать аспирантов и исследователей в области медицины с целью выявления типичных задач, наиболее часто встречаемых в рутинной статистической обработке данных;
3) определить основные компоненты общей схемы работы сервиса;

4) реализовать основные компоненты в алгоритмах и программном коде.

Результатами первого этапа работы являются: разработка системы мета-свойств данных, разработка и реализация алгоритма получения мета-свойств от пользователя, написание кода для применения основных статистических методов в R, построение схемы алгоритма автоматического анализа на основе мета-свойств, его программная реализация.

Программа НИР второго этапа нацелена на развитие платформы для автоматической обработки данных до полной реализации с возможностью выбора одного из четырех вариантов аналитических задач: описательный анализ данных, сравнение групп и повторных измерений, анализ связей, кластерный анализ. Достижение цели требует постановки следующих задач второго этапа исследований:

1) Разработать методы генерации текста для отчета по анализу данных на основе дерева мета-свойств и результатов применения конкретных статистических методов.
2) Программно реализовать разработанные ранее алгоритмы анализа по задачам: "описание данных", "анализ связей".
3) Реализовать формирование отчета по всем аналитическим задачам, включающего описание выбранных методов, ссылки на литературу, таблицы, графики, текст с интерпретацией (на основе многовариантных шаблонов интерпретации), сгруппированные в иерархию заголовков (включая приложения) и оформленные по ГОСТ.
4) Разработать веб-интерфейс для получения мета-свойств по всем аналитическим задачам.
5) Запустить все разработанные алгоритмы (4 задачи, 6 модулей) на платформе (онлайн-сервисе online.statzilla.ru).

6) Провести апробацию платформы у целевой аудитории.

Результатами второго этапа работы являются: тестирование и корректировка алгоритмов платформы, исследование методов генерации текста, разработка системы многовариантных шаблонов интерпретации результатов, а также разработка и реализация алгоритмов формирования связного текстового отчета с результатами анализа данных. Генерация отчета реализована с использованием библиотеки ReporterS на языке R в формате OOXML. Таким образом, основным результатом второго этапа данной НИР можно считать создание рабочей платформы для автоматизированного статистического анализа, реализующей ключевые задачи анализа данных, стоящие перед исследователями таких специальностей как медицина, психология, биология, социология, педагогика.

Описательная статистика является неотъемлемой частью анализа данных. Помимо описательной статистики данный программный модуль включает в себя определение вида распределения показателя, а также тесты на нормальность распределения. Алгоритм получения мета-свойств от пользователя для этого модуля был реализован в коде и протестирован в ходе этапа 1. Целью второго этапа являлись реализация статистических расчётов и внедрение в веб-сервис. В зависимости от полученных мета-свойств данных статистический отчёт включает частоты встречаемости и их диаграммы, средние, медианные и квартильные значения показателя, среднее квадратическое отклонение и значение ошибки среднего, а также гистограммы распределения показателя. Проверка данных на нормальность осуществляется с помощью метода Шапиро-Уилка.

Сравнение средних значений количественных признаков и частот качественных признаков в группах - одна из ключевых задач в биомедицинской статистике. Соответствующий модуль обработки исходных

данных реализует сравнение основных и контрольных групп, выявление статистически значимой динамики показателей, сравнение с нормой.

Модуль анализа связей и ассоциаций включает в себя расчёт корреляций, отношений шансов и рисков, их доверительных интервалов. Методы расчёта корреляций были подобраны на основе статей в Pubmed, а также в ведущих российских журналах. Для программной реализации модуля корреляций использовалось дерево мета-свойств, разработанное на этапе 1. Каждый уникальный набор значений всех свойств соответствует только одному методу, т.е. после установления свойств происходит однозначный выбор метода.

Кластерный анализ включает в себя выявление кластеров и анализ связи кластеров с группой, а также построение дендрограмм. Для кластерного анализа необходимо наличие уникального идентификатора наблюдения, набора кластеризуемых показателей и, возможно, группирующего показателя.

Характерной особенностью выбранного для программной реализации продукта языка R является векторизация вычислений. Векторизация представляет собой один из способов выполнения параллельных вычислений, при котором программа определенным образом модифицируется для выполнения нескольких однотипных операций одновременно. Очевидно, что такой подход потенциально может привести к значительному ускорению однотипных вычислений над большими массивами данных. Эта особенность языка была использована при реализации вычисления описательных статистик показателей. В базовой комплектации R имеется семейство функций *apply*, предназначенных для организации векторизованных вычислений над объектами.

В целом, после реализации 2 этапа НИР достигнуты все поставленные задачи. На основе рекомендаций оформления анализа данных (GCP и прочие)

разработаны шаблоны интерпретации результатов для каждого статистического метода. Проанализированы возможные сочетания свойств исходных данных и результатов и созданы варианты шаблонов для каждого сценария. Все разработанные алгоритмы (4 задачи, 6 модулей) запущены на платформе (онлайн-сервисе online.statzilla.ru). Всего сервис автоматизирует использование более 40 статистических методов в различных сочетаниях, автоматически строит 5 видов графиков. На разработку программного продукта получено свидетельство о государственной регистрации программы для ЭВМ, апробация платформы проведена на 6 кафедрах РостГМУ, а также производилось тестирование в ВолГМУ и РязГУ. Сервисом воспользовались более 200 пользователей, сгенерировавших почти 3 тыс. аналитических отчетов. В настоящее время продолжается процесс интеграции разработки с платформой федеральных медицинских клинических регистров "Росмед.инфо".

В последние несколько лет появилось много систем анализа данных, ориентированных на пользователей, не являющихся специалистами в статистике и аналитике. Однако, почти все они созданы для решения бизнес задач: анализ маркетинговых данных - DataCracker, данных из сферы продаж - Tableau, данных социальных сетей [11] и медиа [12]. На данный момент в мире существует только несколько онлайн-платформ, которые ориентированы на задачи статистической обработки данных для доказательных исследований. Все, за исключением одной из них, сконцентрированы на анализе данных генома и протеома поэтому не генерируют текст с интерпретацией результатов. Только одна платформа производит генерацию текста, однако оставляет за пользователем возможность выбора необходимой методики статистического анализа. Кроме того, для одного исследования обычно необходимо использование нескольких методов, поэтому при проведении анализа данных с

использованием различных методов организуется выработка отдельного отчета с последующей интеграцией полученных результатов. Платформа online.statzilla.ru реализует подход, идущий от задачи, а не от метода. Предлагаемый сервис автоматически выбирает методы на основе мета-свойств и генерирует цельный отчет, включающий результаты всех использованных методов. Такой подход является следующим шагом в автоматизации, представляя конкурентоспособную техническую альтернативу на мировом рынке. Таким образом для большинства типичных задач необходимость обращения к статистику полностью исключается, что позволяет существенно ускорить последний этап исследования и снизить стоимость его проведения.

Литература

1. Гржибовский А. М. Использование статистики в российской биомедицинской литературе. Экология человека. – 2008. №12, С. 55-64.

2. Леонов В.П. Статистика в кардиологии. 15 лет спустя // Медицинские технологии. Оценка и выбор, 2014, №1, С. 17-28.

3. APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. American Psychologist, 61(4), pp. 271-285.

4. Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney S, SQUIRE Development Group. Publication guidelines for improvement studies in health care: evolution of the SQUIRE Project. Ann Intern Med. 2008; 149(9): pp. 670-676.

5. ICH harmonized tripartite guideline: Guideline for Good Clinical Practice. J Postgrad Med 2001;47: pp. 45-50.

6. Knottnerus JA, Tugwell P. The standards for reporting of diagnostic accuracy. J Clinical Epidemiology 2003, 56, Issue 11: pp.1118 - 1129.

7. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche P, et al. and the PRISMA Group (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Med. 2009 Jul; 6: e1000100. doi: 10.1371/journal.pmed.1000100.

8. Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, Von Elm E, et al. STrengthening the REporting of Genetic Association studies (STREGA)-an extension of the STROBE Statement. February 3, 2009 URL: doi.org/10.1371/journal.pmed.1000022

9. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1): pp. 28-55.

10. Vandenbroucke JP. The making of STROBE. Epidemiology 2007;18: pp. 797-799.

11. Носко В.И. Система автоматизированного построения графа социальной сети // Инженерный вестник Дона, 2015, №4-2 URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1428.

12. Носко В.И., Свечкарев В.П., Розин М.Д. Методика и фреймворк конструирования лингвистических моделей для сетевого мониторинга // Инженерный вестник Дона, 2015, №4 URL: ivdon.ru/ru/magazine/archive/n4y2015/3409.

References

1. Grzhibovskij A. M. Jekologija cheloveka. 2008. №12, pp. 55-64.

2. Leonov V.P. Medicinskie tehnologii. Ocenka i vybor, 2014, №1, pp. 17-28.

3. APA presidential task force on evidence-based practice: Evidence-based practice in psychology. American Psychologist, 61(4), pp. 271-285.

4. Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney S. Ann Intern Med. 2008;149(9): pp. 670-676.

5. ICH harmonized tripartite guideline: Guideline for Good Clinical Practice. J Postgrad Med 2001; 47: pp. 45-50.

6. Knottnerus JA, Tugwell P. J Clinical Epidemiology 2003, 56, Issue 11: pp. 1118 - 1129.

7. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche P, et al. and the PRISMA Group (2009) PLoS Med. 2009 Jul; 6: e1000100. doi: 10.1371/journal.pmed.1000100.

8. Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, Von Elm E, et al. February 3, 2009 URL: doi.org/10.1371/journal.pmed.1000022.

9. Schulz KF, Altman DG, Moher D, for the CONSORT Group. Int J Surg. 2012; 10(1): pp. 28-55.

10. Vandenbroucke JP. Epidemiology 2007; 18: pp. 797-799.

11. Nosko V.I. Inženernyj vestnik Dona (Rus), 2015, №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1428.

12. Nosko V.I., Svechkarev V.P., Rozin M.D. Inženernyj vestnik Dona (Rus), 2015, №4. URL: ivdon.ru/ru/magazine/archive/n4y2015/3409.