УДК 519.254

МЕТОДИКА ОБРАБОТКИ МНОГОМЕРНЫХ ДАННЫХ МИКРОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ ЭНДОКРИННЫХ КАРЦИНОМ

[1]М.В. СПРИНДЖУК, [2]Л.М. ЛЫНЬКОВ

[1]*Объединенный институт проблем информатики НАН Беларуси, Республика Беларусь*

[2]*Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь*

Поступила в редакцию 3 июня 2018

Аннотация. Приводится описание практической методики обработки многомерных данных на примере данных текстуры изображений, получаемых с микроскопических изображений карцином яичников.

Ключевые слова: обработка многомерных данных, обработка изображений, биомедицинские большие данные, карцинома яичника.

Abstract. Practical technique for processing of multi-dimensional data using the example of image texture data obtained from microscopic ovarian cancer images is described.

Keywords: multidimensional data processing, image processing, biomedical big data, ovarian carcinoma.

Doklady BGUIR. 2018, Vol. 115, No. 5, pp. 72-76
A technique for the processing of multidimensional data of microscopic images of endocrine cancers
M.V. Sprindzuk, L.M. Lynkou

Введение

Проблема обработки многомерных данных актуальна в связи с увеличением объема данных и мощностей современных компьютерных систем. Основной математический аппарат данной предметной области происходит от опыта многочисленных эконометрических и маркетологических исследований. Биомедицинские данные отличаются своей неоднородностью, большим объемом и нередко трудностями, связанными с хранением данных, их валидацией и конвертацией. Имеется единичный опыт применения текстурных характеристик изображений ядрышек опухолевых клеток с окраской препарата по Feulgen-Schiff для прогноза рецидива карцином яичников [1], а также серия экспериментальных исследований по применению текстурных свойств изображений УЗИ, МРТ, КТ для диагностики опухолей различной локализации, в том числе щитовидной железы [2–5].

Цели и задачи исследования

Целью исследования являлась разработка практической методики обработки многомерных данных текстуры, получаемых с микроскопических изображений, для разработки моделей прогноза развития эндокринных карцином.

Материалы и методы

Для реализации разработанной методики обработки данных было использовано стандартное статистическое программное обеспечение JMP SAS 11, Statistica 8, Stats Direct 2004, Stata 8, Minitab 11, SPSS 20 в бесплатных испытательных версиях. Данные были получены с изображений карциномы яичников (материал проекта МНТЦ 2012) ранее разработанным программным скриптом, написанным на языке Python 2.7. Для получения текстурных характеристик изображений

применялся модуль-библиотека функций Mahotas [6].

Обсуждение результатов

Метод основных компонентов это наиболее распространенный практический метод уменьшения размерности данных [7–9]. Он позволяет из множества колонок данных получить одну или несколько компонент как квази-описательные данные, представляющие основные свойства всей выборки. Математиками критикуется метод основных компонент, так как основные компоненты лишены прямого физического смысла, представляют в своем роде эссенцию-усреднение выборки данных, а не реальный сигнал. Существуют и активно применяются различные алгоритмы пред- и постобработки многомерных данных, в том числе алгоритмы нормализации, аналогичные тем, что широко применяются для биоинформатических данных биочипов-микроматриц [10].

Кластеризация данных − другой стандартный математический подход к обработке многомерных биомедицинских, в том числе и спектроскопических данных [11–13]. Путем вычисления расстояния (Евклидова, Манхеттен) ряды данных группируются по степени подобия, формируют специфический образ-сигнатуру. Результаты кластеризации можно использовать для распознавания образов в потоках сигналов, для классификации данных.

При обработке спектроскопических данных на уровне распознавания образов и бинарной (здоровый-больной) или тернарной (здоровый-предрак-рак) классификации широко применяются уже недавно ставшие традиционными алгоритмы машинного обучения, такие как метод опорных векторов [14–16], бинарная и пошаговая логистическая регрессия, методы, основанные на графах (случайные леса [17–20]), нейросетевые методы. На сегодняшний день все эти кибернетические методы (см. рис. 1) имеют множество модификаций и усовершенствований, наиболее робастные стабильные алгоритмы реализованы в программно-вычислительных комплексах SPSS, Stata, MiniTab, Mathematics, StatsDirect, JMP, SAS, Statistica, R. Python Pandas, SciKit .

Метод	Необходимая фильтрация	Предиктивная способность	Тенденция к переполнению	Интерпретабельность
Кластеринг	Средняя	Низкая	Низкая	Низкая
Логистическая и пошаговая регрессия	Значительная	Средняя	Средняя	Высокая
Дерево классификации	Средняя-значительная	Средняя	Средняя	Высокая
К-граф-ближайших соседей	Средняя-значительная	Средняя	Низкая	Средняя
Линейный ДА	Незначительная	Средняя	Низкая	Низкая
Метод ОВ	Незначительная	Высокая	Высокая	Низкая
Нейронные сети	Незначительная	Высокая	Высокая	Низкая

Рис. 1. Свойства различных статистических методов машинного обучения. По W. Hoffman (2003) (ОВ – опорные вектора)

Разработанная авторами практическая методика обработки данных заключается в рациональном пошаговом применении современного программного обеспечения и пакетов функций из библиотек языков программирования для оценки качества исходных данных, необходимой фильтрации, конвертации форматов, удаления шумов, артефактов, описательной статистики, сравнения спектров, объединения и вычленения нужных спектров для использования их в моделях машинного обучения, классификации, прогноза и т. п.

Алгоритм данной методики представляется следующими шагами-инструкциями:

1. Оценка качества и валидация исходных файлов в исходном формате (*.txt,*.xls, *.por, *.dat, .*spc). Визуализация содержимого исходных файлов или множества субфайлов и спектров.

2. Конвертация исходных файлов в удобный для чтения и манипуляции формат *.csv.

3. Оценка качества, валидация и визуализация полученных данных. Особое внимание нужно уделить типу конвертированных данных, если программа распознает численные данные как строковые, вычисления не смогут быть выполнены адекватно – нужно будет предварительно изменить тип данных.

4. Вычисление комплекса описательной статистики с генерацией бокс-плотов и гистограмм спектров для разведочного анализа данных.

5. В случаях использования реактивов разной концентрации и разбавления к обработке данных могут применяется алгоритмы нормализации и шумоудаления сигналов наподобие таковых, применяемых в сфере обработки данных биочипов-микроэррей.

6. Вычленение основных компонент спектров в нужном количестве в зависимости от дальнейших математических моделей. Можно суммировать как исходные, так и нормализированные данные спектров либо вычислять средние таковых. Таким образом, из множества столбцов данных получается один или несколько.

7. Выполнение комплекса описательной статистики, бокс-плотов, бабл-плотов, гистограмм и различных графиков распределения основных компонент для полноценной визуализации данных.

8. Формирования диагностической или прогностической модели, для которой помимо данных спектров нужны другие данные о пациенте и факторах, влияющих на него либо на диагностический процесс. Адекватно использовать двоичный исход болезни (умер-жив, рецидив-выздоровление и т. д.).

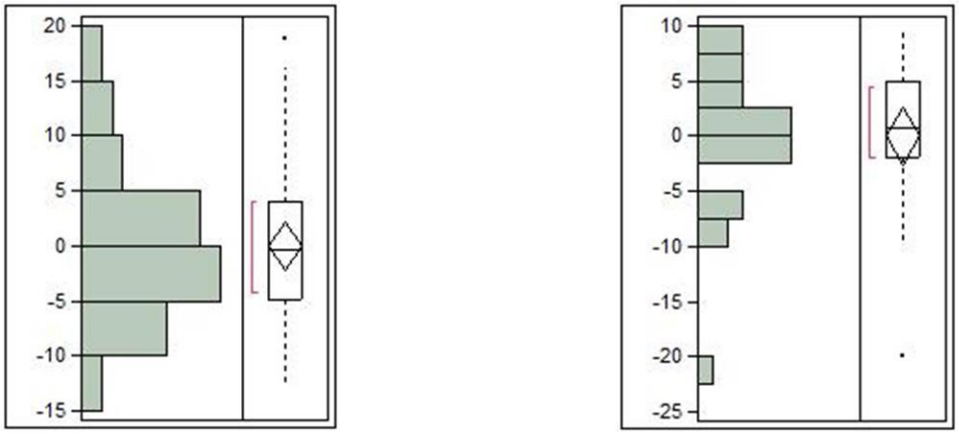
9. Получение предикторного веса данных или параметров точности классификации по типу здоров-болен, опухоль-воспаление и т.д.

10. Формирование заключения и рекомендаций для практического применения с учетом ограничений, преимуществ и недостатков дизайна исследования.

Пример практического применения разработанной методики.

1. Данные текстуры с изображений, обработанных маркерами CD31, Ki67 и D2-40, были разделены по наличию рецидива.

2. С каждого полученного класса данных была вычислена первая основная компонента (применялось стандартное программное обеспечение JMP SAS в испытательной бесплатной версии, см. рис. 2). Основные компоненты были подвергнуты линейному дискриминантному анализу с целью выяснить разделимость классов по схеме здоровый-больной.



а *б*

Рис. 2. Распределение величин первых основных компонент, полученных со спектров функций Р. Харалика: *а* – нет рецидива; *б* – есть рецидив, маркер Ki-67

3. Значения общей классификации прогноза дискриминантного анализа составили 50, 56 и 52 % для групп данных маркеров CD31, Ki67 и D2-40 соответственно ($P < 0{,}005$). Таким образом, были получены доказательства о значимых различиях между классами текстурных параметров у пациентов с рецидивом болезни и без такового. Дискриминирующую силу можно увеличить до 80 % и более отбором исходных изображений и нормализацией данных (такие алгоритмы реализованы, например, в программном комплексе StatsDirect). В данном случае, в отличие от существующих подходов, изучению были подвергнуты специфические микроскопические изображения лимфангиогенеза и ангиогенеза. Текстура изображения вычленялась как спектр численных параметров непосредственно с изображения, введенного в компьютер как матрица, техническими средствами программных библиотек. Текстурные характеристики использовались без сегментации сосудов на изображении, чтобы получить характеристики изображений, не зависимые от настройки аппарата распознавания микрососудов. Таким образом, были получены доказательства о значимых различиях между классами текстурных параметров у пациентов с рецидивом болезни и без такового.

Выводы и заключение

Предлагаемая практическая методика обработки многомерных числовых данных позволяет формировать модели классификации данных, которые можно использовать для разработки способов прогноза развития и исхода заболеваний. Разработанная методика приемлема для обработки не только данных текстуры изображений, но и для биоинформатических (микроэррей-биочипы) и данных лазерной спектроскопии Рамана. Данной методикой также можно формировать ансамбли предикторов одного или различного происхождения. При отсутствии стандартных программных комплексов для статистических вычислений данную методику можно реализовать средствами бесплатных языков программирования R и Python.

Список литературы / References

1. The prognostic value of adaptive nuclear texture features from patient gray level entropy matrices in early stage ovarian cancer / B. Nielsen [et al.] // Anal Cell Pathol (Amst). 2012. № 4. P. 305–314.
2. Cord A., Bach F., Jeulin D. Texture classification by statistical learning from morphological image processing: application to metallic surfaces // J. Microsc. 2010. № 2. P. 159–166.
2. The value of nuclear DNA and texture analysis by digital image processing in the diagnosis of lipomatous and leiomyomatous tumours / M. Remmelink [et al.] // Anal Cell Pathol. 1996. № 1. P. 45–58.
3. Classification of melanocytic lesions with color and texture analysis using digital image processing / T. Schindewolf [et al.] // Anal Quant Cytol Histol. 1993. № 1. P. 1–11.
4. An interactive processing system for ultrasonic compound imaging, real-time image processing and texture analysis / E. Schuster [et al.] // Ultrason Imaging. 1986. № 2. P. 131–150.
5. Coelho L.P. Mahotas: Open source software for scriptable computer vision // J. of Open Research Software. 2013. № 1. P. 1–6.
6. Ferraty F., Romain Y. The Oxford handbook of functional data analysis. Oxford, New York, 2011. 494 p.
7. Gray V. Principal component analysis: methods, applications, and technology. Nova Science Publishers, 2017. 130 p.
8. Härdle W., Mori Y., Vieu P. Statistical methods for biostatistics and related fields. Berlin, New York, 2007. 370 p.
9. Huang H.C., Qin L.X. Empirical evaluation of data normalization methods for molecular classification // Peer J. 2018. № 1. P. 4584.
10. Nahid A.A., Mehrabi M.A., Kong Y. Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering // Biomed Res Int. 2018. № 2. Article ID 2362108.
11. Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types / J.K. Rhee [et al.] // Cancer Immunol Res. 2018. № 1. P. 87–97.
12. Zhang E., Ma X. Regularized Multi-View Subspace Clustering for Common Modules Across Cancer Stages // Molecules. 2018. № 5. P 22–29.
13. Support Vector Machines (SVM) classification of prostate cancer Gleason score in central gland using multiparametric magnetic resonance images: A cross-validated study / J. Li [et al.] // Eur J. Radiol. 2018. № 3. P. 61–67.
14. Moteghaed N.Y., Maghooli K., Garshasbi M. Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine // J. Med Signals Sens. 2018. № 1. P. 1–11.
15. Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study / I. Vidic [et al.] // J. Magn Reson Imaging. 2018. № 5. P. 1205–1216.
16. Genetic Variants in Metabolic Signaling Pathways and Their Interaction with Lifestyle Factors on Breast Cancer Risk: A Random Survival Forest Analysis / S.Y. Jung [et al.] // Cancer Prev Res (Phila). 2018. № 1. P. 44–51.
17. Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes / Y.O. Nunez Lopez [et al.] // Rep Pract Oncol Radiother. 2018. № 1. P. 6–20.
18. Prognostic value of cancer antigen -125 for lung adenocarcinoma patients with brain metastasis: A random survival forest prognostic model / H. Wang [et al.] // Sci Rep. 2018. № 1. P. 5670.
19. Hofmann W.-K. Gene expression profiling by microarrays: clinical implications. Cambridge, New York, 2006. 246 p.

Сведения об авторах

Спринджук М.В., научный сотрудник Объединенного института проблем информатики НАН Беларуси.

Лыньков Л.М. д.т.н., профессор, профессор кафедры защиты информации Белорусского государственного университета информатики и радиоэлектроники.

Адрес для корреспонденции

220012, Республика Беларусь
г. Минск, ул. Сурганова, 6,
Объединенный институт
проблем информатики НАН Беларуси
тел. +375-33-682-57-55;
e-mail: bioinformatics_bel@yahoo.com;
Спринджук Матвей Владимирович

Information about the authors

Sprindzuk M.V., researcher of United institute of informatics problems of National academy of sciences of Belarus.

Lynkov L.M., D.Sci, professor, professor of information security department of Belarusian state university of informatics and radioelectronics.

Address for correspondence

220012, Republic of Belarus
Minsk, Surganova st., 6,
United institute of informatics problems
of National academy of sciences of Belarus
tel. +375-33-682-57-55;
e-mail: bioinformatics_bel@yahoo.com
Sprindzuk Matvey Vladimirovich