где i, j – точки; $d_{(i,j)}$ – расстояние между точками i, j; $h_{(i,j)}$ – разница уровней высоты точек i, j; $t_{(i,j)}$ – значение проходимости между точками i, j.

В итоге получаем взвешенный граф, который в дальнейшем можно использовать во многих операциях. Данный метод является надежным для определения непреодолимых объектов и правильного образования связей, что поспособствует корректной работе с данным графом. С помощью весов ребер можно находить оптимальные пути. Каждая точка, а, значит, и последовательность хранит информацию об объекте, и задает границы этих объектов, отсюда можно работать именно с объектами выбранной карты.

Библиографический список

1. Карпов Д.В. Теория графов [Электронный ресурс]: учеб. пособие. СПб государственный университет. – СПб.: 2017. – 525 с.

2. Асельдеров З.М., Донец Г.А. Представление и восстановление графов. – Киев: Наукова думка, 1991. –192 с

3. Алексеев В.Е., Таланов В.А. Графы. Модели вычислений. Структуры данных: учебник. – Нижний Новгород: Изд-во ННГУ, 2005. – 307 с.

4. Игнатьев Ю.Г., Агафонов А.А. Аналитическая геометрия евклидового пространства. Учебное пособие. – Казань: Казанский университет, 2014, – 204 с.

5. Marathe M.V., Breu H., Hunt III H.B., Ravi S.S., Rosenkrantz D.J. Simple heuristics for unit disk graphs // Networks. – V.25 (2). – P. 59–68.

УДК 519.23

Ансамбль алгоритмов фильтрации для отбора значимых признаков биомедицинских данных

И.Ю. Бойко

АлтГУ, г. Барнаул

В статье рассмотрены подходы к отбору признаков биомедицинских данных, реализован метод ансамблирования алгоритмов фильтрации, набирающий популярность в последние годы, с применением разработанного ранее ledge-критерия. Использование рассмотренного подхода потенциально позволяет улучшить качество классификации и получать более стабильные результаты.

Ключевые слова: *отбор признаков, бинарная классификация, ledge-коэффициент, ДНК-микрочипы.*

Биомедицинские данные с микрочипов имеют две отличительные особенности: большое количество признаков при малом числе объектов и высокая зашумленность. В связи с этим, для эффективной классификации таких данных широко применяются методы отбора признаков, поскольку это позволяет сократить шумовую составляющую и уменьшить используемые вычислительные ресурсы [1].

Методы отбора признаков традиционно делят на три группы: алгоритмы фильтрации, алгоритмы обертки, встроенные алгоритмы [2]. При использовании методов первой группы происходит оценка и упорядочивание признаков согласно заданной мере. После чего с применением некоторого отсекающего правила выбирается искомое подмножество признаков. В алгоритмах фильтрации не используются методы машинного обучения, поэтому они обладают наибольшей скоростью работы, что делает их самыми распространенными в применении на практике к наборам данных с большим количеством признаков [2].

Однако, алгоритмы фильтрации зачастую не предназначены для выявления сложных взаимосвязей между признаками, поэтому не всегда в полном объеме могут выявить полезную информацию [1]. Алгоритмы обертки, наоборот, слабо применимы к данным с микрочипов, потому что для их реализации происходит построение определенной модели машинного обучения на различных подмножествах признаков, после чего выбирает подмножество, на котором модель достигает максимального качества [2, 3]. В алгоритмах третьей группы отбор признаков выполняется в ходе построения модели машинного обучения. Такие методы работают быстрее, чем алгоритмы обертки, но медленнее, чем фильтры. Существующие подходы к отбору признаков имеют как преимущества, так и недостатки, что мотивирует дальнейшие исследования в этом направлении.

В последние годы все чаще появляются новые методы отбора признаков, заключающиеся в создании ансамбля алгоритмов фильтрации, что позволяет производить отбор признаков на основе многообразия критериев, а также получать более стабильные результаты классификации различных данных, в силу того, что такой метод не полагается на какой-либо один определенный алгоритм фильтрации [4].

Далее представим решение задачи классификации с применением рассмотренного подхода. Нами использован набор данных с ДНК-микрочипов, содержащий сведения о 168 пациентах, у которых был

диагностирован рак молочной железы. Данные о каждом объекте представлены значениями 2905 числовых признаков. Бинарная метка класса соответствует появлению метастазов в течение 5 лет наблюдений [5]. Двадцать процентов (34 объекта) данных были выделены в тестовую выборку.

Для отбора признаков использовались три алгоритма фильтрации, основанных на коэффициентах корреляции (Пирсона, Фехнера, Ледж), а также смешанный ансамбль этих алгоритмов. Отсекающее правило для фильтров – 20% лучших значений меры. После ансамблирования были выбраны 20 признаков. Для классификации использовался алгоритм Random Forest, обученный при 5-сегментной перекрестной проверке, который затем применялся к тестовой выборке. Расчеты выполнены на языке программирования Python 3.9 с использованием модулей NumPy, Pandas, scikit-learn, ITMO_FS. Ниже представлены основные результаты бинарной классификации, выполненной по вышеописанной методике.

Использование ансамбля фильтров, в свою очередь, позволило выбрать признаки, на которых точность классификации тестовой выборки составила 0.765. Таким образом, разработка гибридных моделей отбора признаков содержит потенциал для повышения эффективности классификации данных с микрочипов в сравнении с использованием традиционных подходов.

Таблица 1 – Количество отобранных признаков и результаты классификации тестовой выборки при использовании различных алгоритмов фильтрации

№	Мера алгоритма фильтрации	Количество признаков	Точность
1	-	2905	0.647
2	Pearson correlation	521	0.618
3	Fechner correlation	667	0.618
4	F-ratio	3	0.559
5	Ledge correlation	27	0.618

Библиографический список

1. Saeys Y., Inza I., Larranaga P. A review of feature selection techniques in bioinformatics // Bioinformatics. – 2007. – V.23. № 19. – P. 2507-2517.

2. Hira Z., Gillies D. A review of feature selection and feature extraction methods applied on microarray data // Advances in Bioinformatics. – 2015. – V.2015. – P. 1-13.

3. Aboudi N., Benhlima L. Review on wrapper feature selection approaches // 2016 International Conference on Engineering & MIS (ICEMIS). – 2016. – P. 1-5.

4. Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification // Pattern Recognition. – 2012. – V.45. № 1. – P. 531-539.

5. Gravier E. A prognostic DNA signature for T1T2 node-negative breast cancer patients // Genes, Chromosomes and Cancer. – 2010. – V.49(12). – P. 1125–1134.

УДК 528.854.2

Разработка веб-сервиса для управления модулями анализа фракционного состава зерновой смеси по фотографиям

Воронков А.Е., Жилин С.И., Жирнов Д.С., Козлов Д.Ю.
АлтГУ, г. Барнаул

Компания «СиСорт» [1] занимается разработкой и производством высокотехнологичного оборудования для сортировки сыпучих продуктов. Одно из направлений деятельности ООО «СиСорт» связано автоматизацией анализа фракционного состава зерновой смеси, поскольку при каждой перепродаже на пути от производителя к конечному потребителю требуется оценка качества продукции путем установления доли сорной примеси. Один из типовых способов автоматизации бизнес-процессов – это разработка и внедрение веб-приложения, которое для рассматриваемой предметной области может сочетать в себе набор инструментов автоматического анализа зерновых смесей с доступностью из любой точки планеты.

На рисунке 1 представлена общая схема работы с мобильным анализатором смеси сыпучих материалов, в котором рассматриваемое в данной статье веб-приложение выделено в рамку.

Предполагается, что аналитик (пользователь мобильного анализатора) перед запуском процесса анализа сможет подключить в веб-приложении необходимые модули анализа, которые выполнятся на вычислительном сервере. У компании «СиСорт» уже имеется мобильное приложение, которое позволяет с помощью камеры смартфона сделать и разместить в хранилище изображения зерновой смеси, которые и будут, собственно, анализироваться. Т.е. роль