Вычислительная химия, экономика и биофизика. Биоинформатика

УДК 519.68;633/635:577/2

Сравнительное исследование кластерного и нейросетевого подходов в задаче анализа белковых структур

Д. А. Баранов*, Г. А. Ососков*, А. А. Баранов†

* *Лаборатория информационных технологий*
Объединённый институт ядерных исследований
ул. Жолио-Кюри, д. 6, Дубна, Московская область, Россия, 141980

† *Московский государственный технический университет радиотехники, электроники и автоматики (технический университет)*
пр. Вернадского, д. 78, Москва, Россия, 119454

В данной статье описывается работа, которая является продолжением предыдущего исследования, направленного на поиски решения проблем, возникающих в задаче автоматизации процедуры распознавания генетических белковых структур по их электрофоретическим спектрам (ЭФ-спектрам).

Спектральная идентификация сортовой принадлежности зёрен пшеницы является одной из важных сельскохозяйственных задач, для решения которой было предложено использовать Искусственную Нейронную Сеть (ИНС), обученную на выборке из специально подготовленных экспертами сортов.

Рассматриваются особенности применения методов нейросетевой классификации и кластерного анализа на примере определения сортовой принадлежности ЭФ-спектров.

Правомерность использования предложенных алгоритмов подтверждается положительными результатами, полученными на основе специально подготовленных модельных данных в виде многомерных векторов, имитирующих особенности реальных ЭФ-спектров, прошедших предварительную обработку, которая включает оцифровку, устранение шумовых и фоновых составляющих, нормализацию.

По естественной причине генетического сходства, наблюдаемого у некоторых родственных сортов, ЭФ-спектры имеют трудно различимый характер, что оказывает неблагоприятное влияние на эффективность распознавания схожих экземпляров средствами ИНС. Это накладывает ограничение на количество одновременно распознаваемых сортов. Для преодоления данной особенности был предложен алгоритм кластерного разбиения всего множества сортов на отдельные сортовые группы с последующим применением нейросетевой обработки для каждой группы.

Ключевые слова: искусственные нейронные сети, классификация, кластеризация, генетический анализ, определение сортовой принадлежности, электрофоретический спектр.

1. Введение

Определение сортовой принадлежности зерновых культур является актуальной на сегодняшний день сельскохозяйственной задачей, направленной на контроль за поддержанием чистосортности зернового фонда, а также выведением новых экземпляров, обладающих хорошими посевными и сортовыми характеристиками. Сортовая идентификация проводится на основе электрофоретических спектров (ЭФ-спектров), получаемых методом электрофореза белкового материала (глиадина) зёрен пшеницы. Каждый отсканированный ЭФ-спектр, называемый денситограммой, является уникальным идентификатором исследуемого образца. Определённые комбинации тёмных полос (генетических маркеров) на спектре позволяют экспертам-генетикам делать вывод о сортовой принадлежности рассматриваемого образца.

Статья поступила в редакцию 30 декабря 2013 г.

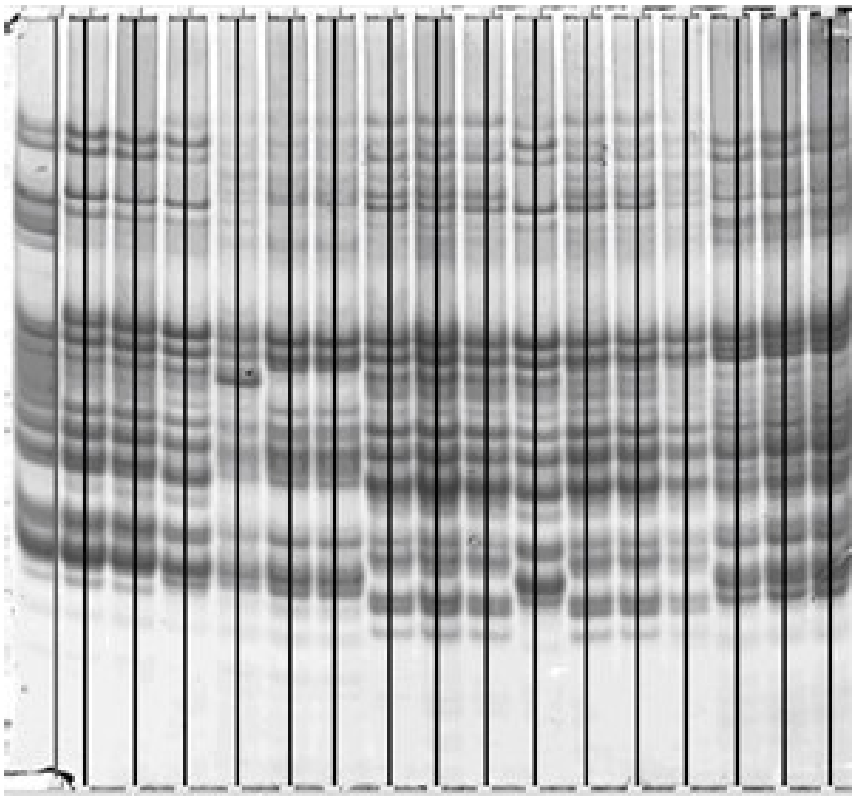

Рис. 1. Электрофоретическая дорожка и её денситограмма

Обрабатывая большое количество образцов, эксперты тратят значительное количество времени, «вручную» просматривая каждый спектр. Применение методов нейросетевой классификации позволит экспертам, проводящих анализ, существенно ускорить и в значительной степени автоматизировать процедуру обработки большого объёма данных.

2. Особенности применения нейросетевого подхода

Использование нейросетевого подхода [1] в задаче классификации подразумевает, во-первых, большую предварительную работу экспертов-генетиков, направленную на подготовку специальной выборки, которая включает для каждого из нескольких десятков сортов более сотни ЭФ-спектров. Большая часть этой выборки используется для обучение нейросети, остальные 20–25% — для тестирования эффективности нейросетевой классификации. Во-вторых, исходные данные должны быть преобразованы к форме, допустимой для их ввода в нейросеть. Физические особенности получения денситограмм методом электрофореза приводят к различного рода искажениям, которые вносят отрицательный вклад в информативность данных. В ходе проведённого нами исследования был разработан *алгоритм предварительной обработки*, который включает сглаживание сигнала, вычисление и вычитание фоновой составляющей, выравнивание амплитуд и положений спектров. Данные преобразования позволяют минимизировать влияние негативных факторов, что повышает информативность данных.

Избыточная размерность спектров (около 4000 измерений) не позволяет непосредственно вводить их в Искусственную Нейронную Сеть (ИНС) для обработки — требуется сокращение размерности при сохранении максимальной информативности. Проблема редукции данных тесно связана с проблемой выделения характерных признаков, являющихся ключевыми особенностями того или иного сорта.

Такими признаками было принято считать пики гауссовой формы на спектре. Данные пики, с биологической точки зрения, являются генетическими маркерами, определённые комбинации которых дают информацию о сортовой принадлежности исследуемого спектра.

В работах [2, 3] предлагался ряд алгоритмов сокращения размерности ЭФ-спектров с минимальной потерей информативности, таких, например, как *ранжирование*, суть которого состояла в расстановке значимых пиков в порядке убывания их рангов, а также алгоритмы, основанные на применении *метода главных компонент*, *Фурье- и вейвлет-анализа*. Однако все эти алгоритмы, хотя и осуществляют сокращение размерности входных данных более, чем на порядок, дают недостаточно высокую эффективность по результатам нейросетевой классификации.

В данной работе даётся краткое описание и результаты тестирования нового метода сокращения размерности, основанного на *векторном представлении* ЭФ-спектра относительно характеристик значимых пиков. Идея метода возникла в результате рассмотрения способов генетического кодирования ЭФ-спектров, принятых в работах генетиков (например, [4]), где электрофореграммы глиадина после оцифровки образуют сцепленные в силу генетических связей блоки дискретных компонент, что позволяет их записывать в виде простых генетических формул. Суть предлагаемого нами алгоритма заключается в разбиении ЭФ-спектра на определённое количество зон (как правило, до 40 зон). Значение зоны приравнивается значению амплитуды пика, попавшего в зону. В случае, если в зону не попало ни одного пика, она считается «пустой» и её значение равно нулю. В результате получается многомерный вектор, компонентами которого являются последовательность зон. Размерность вектора соответствует количеству зон. При таком подходе одновременно учитываются два параметра пика — положение, как относительный номер зоны, и амплитуда, как значение зоны.

Адекватность алгоритма разбиения ЭФ-спектров на зоны, а также его эффективность при последующем применении классификационных и кластерных методов были протестированы на *модельных данных*, характеристики которых соответствуют особенностям реальных сортов. Были подготовлены несколько выборок с различными уровнями амплитудного зашумления, полученных на основе 10 типовых незашумлённых образцов. Максимальный уровень зашумления, равный 50%, был выбран с запасом — дисперсии значений по амплитуде и смещению реальных образцов не превышают 30% (по результатам статистического анализа).

Последующие обучение и тестирование искусственной нейронной сети на предложенных зашумленных выборках показали превосходные результаты — эффективность нейросетевого распознавания не падала ниже 95% даже на сильно зашумленных выборках, что позволяет утверждать об эффективности применения нейросетевого подхода в задаче классификации зашумленных многомерных данных.

3. Особенности применения кластерного подхода

В работах [2, 3] затрагивалась проблема резкого падения эффективности нейросетевой классификации при увеличении количества распознаваемых сортов. Для решения данной проблемы был предложен метод *кластерного разбиения* множества сортов на отдельные сортовые группы. Использование кластерного анализа [5] в качестве вспомогательного инструмента классификации даёт возможность обучения нейросети на отдельных группах, позволяя охватить все сортовое множество, предварительно разбитое на части.

Для оценивания потенциальной возможности применения кластерного подхода в задаче классификации многомерных сортовых данных было проведено исследование, заключающееся в разбиении выборок модельных данных при различных уровнях амплитудного зашумления, используя различные методы кластерного анализа.

В качестве критерия эффективности кластеризации применялся следующий алгоритм: после выполнения кластеризации по тому или иному методу на большой обучающей выборке, сгенерированной из K идеальных сортовых векторов с заданным коэффициентом зашумления, получался набор из K кластеров с их центрами. Для проверки правильности кластеризации на тестовой выборке находилось расстояние от каждого элемента тестовой выборки до центров всех кластеров и выбрался наиближайший из них. Если признак этого тестируемого элемента совпадал с номером этого кластера, то кластеризация считалась выполненной правильно. Если тестируемый элемент попадёт не в «свой» кластер (несовпадение номеров), то счётчик эффективности не увеличивался. В случае кластеризации по методу одиночной связи (ближайшего соседа) критерием удачной кластеризации бралась близость тестируемого элемента не к центрам кластеров, найденных на этапе обучения, а к любому из элементов этих кластеров.

Так как заранее было известно, к какому сортовому типу принадлежит тот или иной вектор, то мы могли дать оценку правильности отнесения любого вектора к какому-либо кластеру. По данному критерию, стопроцентный результат эффективности достигается при условии попадания векторов одного типа в отдельный кластер.

В приведённой ниже таблице представлены результаты кластерного разбиения множества модельных данных, имеющих в своём составе 10 типовых образцов, на 10 кластеров соответственно. Эффективность кластеризации определялась как процент правильно попавших в свой кластер векторов.

По полученным результатам можно судить, что кластерные методы эффективны при небольшой степени зашумления исходных выборок (до 30%). При более высоком уровне зашумления кластеры начинают пересекаться между собой. Однако, принимая во внимание тот факт, что реальные данные имеют до 30% амплитудного зашумления, а уровень зашумления модельных данных выбирался с некоторым запасом (до 50%), то можно утверждать о правомерности применения кластерного подхода в задаче идентификации сортовой принадлежности ЭФ-спектров.

Методы кластерного анализа дают хорошие результаты при обработке многомерных векторов, имеющих такую степень зашумления, при которой кластеры, образованные определённые типами, не пересекаются между собой. Также необходимо заметить, что кластеры рассматриваемых модельных образцов имеют компактную гиперсферическую форму в многомерном пространстве с разбросом, подчиняющемуся нормальному закону распределения. Оценка адекватности работы кластерных методов на данных, типовые группы которых имеют другую форму кластеров в многомерном пространстве, не проводилась.

4. Заключение

В данной работе был приведён сравнительный анализ эффективности применения нейросетевого и кластерного подходов в задаче классификации многомерных данных на основе реальных ЭФ-спектров белкового материала зёрен пшеницы. Как правило, многомерные данные носят многокритериальный характер, а процедура их типовой идентификации является трудно формализуемой с точки зрения компьютерной обработки. Кроме того, влияние различных шумовых и искажающих факторов значительно усложняют процесс классификации. Устойчивость к подобному типу проблем и прекрасные классификационные способности искусственных нейронных сетей объясняют актуальность и растущую популярность их использования в подобного рода задачах.

Литература

1. *Haykin S.* Nueral Networks. A Comprehensive Foundation. — New Jersey: Prentice Hall, 2006.

2. *Ososkov G. A., Baranov D. A.* Feature Extraction for Data Input to Neuro-Classifiers // Mathematical Modeling and Computational Physics (MMCP 2009): Book of Abstract of the International Conference (Dubna, July 7–11). — 2009. — Pp. 110–111.
3. *Ososkov G. A., Baranov D. A.* Extraction of Data Features for Neuro-Classifier Input // Bulletin of PFUR. Series "Mathematics. Information Sciences. Physics". — 2010. — No 3. — Pp. 142–148.
4. *Ruanet V. V., Kudryavtsev A. M., Dadashev S. Y.* The Use of Artificial Neural Networks for Automatic Analysis and Genetic Identification of Gliadin Electrophoretic Spectra in Durum Wheat // Russian Journal of Genetics. — 2001. — Vol. 37, No 10. — Pp. 1207–1209.
5. *Duran B. S., Odell P. L.* Cluster Analysis. — New York: Springer Verlag, 1947.

UDC 519.68;633/635:577/2

Comparative Study of Cluster and Neural Network Methods in the Problem of Protein Structure Analysis

D. A. Baranov*, G. A. Ososkov*, A. A. Baranov†

* *Laboratory of Information Technologies*
Joint Institute for Nuclear Research
6, Joliot-Curie str., Dubna, Moscow region, Russia, 141980
† *Moscow State Institute of Radio Engineering, Electronics and Automation*
78, Vernadsky av., Moscow, Russia, 119454

This work continues the previous study where the important problem of automatization of differentiation methods of the genetic protein structures according to their electrophoretic spectrums (EPS) was considered. The multicriterion problem of the agriculture cultivar identification by their spectra caused the idea of its solution by an artificial neural network (ANN) trained on an expert data base.

In the given paper peculiarities of the neural net use as well as the purposefulness of cluster analysis applications for the EPS classifying are studied.

A special model of multidimensional vectors adequately imitating the most essential characteristics of real data obtained after EPS digitalization, denoising and normalization is developed. A numerical experiment is fulfilled on such simulated data stream to study the influence of contamination and distortion factors on the ANN efficiency in order to suppress those factors and improve ANN functioning.

Various methods of cluster analysis are also applied to simulated multidimensional data as either an ANN alternative or more soundly as a prior stage of a coarse data classification in some set of detached cultivar groups to be classified next by ANN.

Key words and phrases: artificial neural networks, classification, clusterization, genetic analysis, electrophoretic spectra.