ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ И КЛАСТЕРНЫХ МЕТОДОВ В ОДНОЙ ЗАДАЧЕ ГЕНЕТИЧЕСКОГО АНАЛИЗА ПРОТЕИНОВ

Ососков Г.А.

ОИЯИ, Россия, 141980 Дубна, ул.Жолио-Кюри 6, +79168374079, ososkov@jinr.ru

Рассматривается генно-биологическая задача определения сортовой принадлежности сельскохозяйственной культуры пшеницы на основе анализа электрофоретических спектров белкового материала зёрен пшеницы. Определение сортовой принадлежности зерновых культур является актуальной на сегодняшний день сельскохозяйственной задачей, направленной на контроль за поддержанием чистосортности зернового фонда, а также выведение новых экземпляров, обладающие хорошими посевными и сортовыми характеристиками. Исходные данные представляют собой денситограммы электрофоретических спектров (ДЭФС), получаемые путём проведения электрофореза и последующего сканирования белкового материала (глиадина) зерен пшеницы. ДЭФС содержит упорядоченный набор тёмных и светлых полос - генетических маркеров, позволяющих экспертам-генетикам делать вывод о сортовой принадлежности рассматриваемого образца. Для автоматизации процедуры классификации было предложено использовать нейросетевой подход. Как было показано в работах [1,2], помимо большой предварительной работы экспертов-генетиков по созданию представительной выборки денситограмм для обучения и тестирования нейросети, этот подход требует сложных преобразований входных данных к форме, допустимой для их ввода в Искусственную Нейронную Сеть (ИНС) и кардинального сокращения их размерности. Предложенные ранее алгоритмы предобработки денситограмм, хотя и осуществляли сокращение объема входных данных более, чем на порядок, но давали недостаточно высокую эффективность по результатам нейросетевой классификации. В данной работе дается краткое описание и результаты тестирования нового метода, основанного на векторном представлении денситограммы относительно положений и амплитуд найденных на ней пиков. В предыдущей работе [2] затрагивалась проблема резкого падения эффективности нейросетевой классификации при увеличении количества распознаваемых сортов выше 15-20 в то время как на практике экспертам приходится иметь дело с многими десятками сортов. Для решения данной проблемы был предложен метод кластерного разбиения множества сортов на отдельные сортовые группы по 10 типов сортов, Такое использование кластерного анализа в качестве вспомогательного инструмента нейросетевой классификации дало возможность обучения ИНС на отдельных группах и позволило охватить все сортовое множество, предварительно разбитое на части.

Адекватность и эффективность алгоритма, использующего векторное представление ДЭФС, тестировались на специально подготовленных модельных данных, характеристики которых соответствовали характеристикам денситограмм реальных сортов. Результаты исследования показывают, что кластерные методы эффективны при степени амплитудного зашумления исходных выборок до 30%.

Литература

1. А.М. Кудрявцев и др., Использование искусственных нейронных сетей при определении сортовых качеств семян твердой пшеницы, Сельскохозяйственная биология, 2002, №1, с. 121-124
2. Баранов Д. А., Ососков Г. А., Баранов А. А. Сравнительное исследование кластерного и нейросетевого подходов в задаче анализа белковых структур, Вестник РУДН, Серия Математика. Информатика. Физика, №2, 2014, 234-238