УДК 63:681.2:001.89:004

Д.И. ЧАНЫШЕВ[1], научный сотрудник,
А.Ф. АЛЕЙНИКОВ[1,2], доктор технических наук, профессор, главный научный сотрудник,
И.Г. ГРЕБЕННИКОВА[1], кандидат сельскохозяйственных наук, заведующая лабораторией,
А.Ф. ЧЕШКОВА[1], кандидат физико-математических наук, заведующая лабораторией,
П.И. СТЁПОЧКИН[1,3], доктор сельскохозяйственных наук, ведущий научный сотрудник

[1]Сибирский физико-технический институт аграрных проблем СФНЦА РАН

630501, Россия, Новосибирская область, пос. Краснообск

e-mail: sibfti.n@ngs.ru

[2]Новосибирский государственный технический университет

630073, Россия, Новосибирск, пр. Карла Маркса, 20

e-mail: fti2009@yandex.ru

[3]Сибирский институт растениеводства и селекции — филиал института цитологии и генетики СО РАН

630501, Россия, Новосибирская область, пос. Краснообск

sibniirs@bk.ru

КЛАСТЕРИЗАЦИЯ КОЛЛЕКЦИОННЫХ ОБРАЗЦОВ ТРИТИКАЛЕ ДЛЯ ИСПОЛЬЗОВАНИЯ В СЕЛЕКЦИИ*

Представлен вариант прогнозирования селекционной ценности коллекционных образцов ярового тритикале на основе полученной информации по изучению количественных признаков. Приведен алгоритм прогноза – кластерный анализ, осуществленный тремя различными методами. Изучены особенности основных методов кластеризации на примере исследования коллекции образцов тритикале. Доказано, что метод Уорда, использующий методы дисперсионного анализа для изучения расстояний между кластерами, позволяет проводить оценку селекционного материала с более высокой эффективностью, чем остальные, повышает результативность процесса подбора родительских пар. Данный метод оптимизирует минимальную дисперсию внутри близко расположенных кластеров, направлен на их объединение и создание кластеров малого размера. Каждый шаг алгоритма объединяет пару кластеров, приводящих к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов отклонений. Полученные методом Уорда результаты дали наиболее объективную и информативную кластеризацию коллекции тритикале.

Ключевые слова: прогноз, ритикале, кластерный анализ, метод Уорда.

В процессе селекционной работы исследователю приходится обрабатывать огромный объем информации в конкретных условиях произрастания для подбора и вовлечения в гибридизацию подходящих родительских форм, прогнозируя передачу от них желаемых признаков потомству. При этом каждая форма оценивается по 20 признакам и более [1, 2]. Селекционные учреждения длительное время ведут работу по изучению исходного материала, выделяют из него множество селекционно-ценных форм. Интегральный анализ большого разнообразия привлеченного в гибридизацию исходного материала различных эколого-географических групп с учетом многолетнего характера наблюдений невозможен без использования средств автоматизации. Отсутствие или неполнота исходных данных существенно снижают качество вырабатываемых технологических реше-

*Статья опубликована при финансовой поддержке РФФИ (грант № 16-07-20001).

ний [3–5]. Исходя из сказанного, задача прогнозирования результатов селекционного процесса весьма актуальна и значима.

Цель исследования – выбор наиболее эффективного способа прогнозирования селекционной ценности коллекционных образцов ярового тритикале урожая 2011 г., полученного в лаборатории экспериментальных исследований (биополигон) Сибирского физико-технического института аграрных проблем (СибФТИ) СФНЦА РАН. Алгоритм прогноза – кластерный анализ.

Исследовали 186 образцов тритикале с 15 признаками с балльной оценкой из коллекции СибФТИ. В предшествующих работах авторы исследовали возможность применения нейронной сети PNN с обучающей выборкой и слоя Кохонена к данной коллекции урожая [6–11]. Нейронный слой Кохонена в зависимости от заданных циклов обучения разделяет коллекцию на кластеры, число которых неконтролируемо увеличивается с ростом циклов обучения, а кластеризация фактически начинается после 300 циклов [10, 11]. Так, при 300 циклах обучения кластеры исчисляются 25 единицами, при 575 – их 70. Следует отметить, что состав кластеров в зависимости от количества циклов обучения варьируется и четкой наследственности не прослеживается. Стабилизация наступает после 575 циклов обучения.

При использовании классических методов обработки данных, реализованных в пакете программ Snedecor и специально адаптированных О.Д. Сорокиным к биологическим исследованиям, не удается разделить коллекцию на кластеры [12]. Программа выстраивает коллекцию в виде пар элементов по ранжиру – по близости евклидова расстояния между элементами коллекции, оставляя за исследователем право самому формировать кластеры и наборы в них. Такая особенность кластеризации вынуждает использовать результаты программы Snedecor главным образом в качестве теста для проверки достоверности кластеризации по другим программам, т.е. пока евклидово расстояние достаточно мало, элементы каждой пары должны быть в одном кластере. При использовании программы Snedecor у последних 30 членов ранжира из 185 пар евклидово расстояние увеличилось на порядок по сравнению с первыми 10, так что этими 30 членами следует пренебречь.

Отклонение расчетов, выполненных по нейронному слою Кохонена, при увеличении циклов обучения от 300 до 575 изменяется от 23 до 13,4 %. Дальнейшее увеличение циклов не ведет к уменьшению отклонения.

Следует отметить, что задача выбора числа кластеров достаточно сложна. Для подтверждения удовлетворительного исхода часто требуется сделать значимые предположения о свойствах заранее заданного семейства распределений. Алгоритмы кластеризации обычно строятся как способ перебора числа кластеров и определения его оптимального значения в ходе этого процесса. Из известных методов кластерного анализа наиболее распространены иерархические агломеративные методы, суть которых состоит в последовательном объединении исходных элементов и соответствующем уменьшении числа кластеров [12, 13]. Из различных способов группировки выделяются методы одиночной связи (ближайших соседей), полной связи (наиболее удаленных соседей), средней связи и Уорда [13].

Метод Уорда имеет определенные преимущества. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, полученный в результате их объединения. Число кластеров таким образом контролируется, что создает определенный комфорт для исследователя; кластеры сохраняют четкую наследственную структуру, как ветви дерева, что позволяет провести как грубую кластеризацию, так и тонкую. В отличие от других методов здесь для оценки расстояний между кластерами применяются методы дисперсионного анализа. Метод направлен на объединение близко расположенных кластеров и оптимизирует минимальную дисперсию внутри них. Каждый шаг алгоритма объединяет пару кластеров, приводящих к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов отклонений:

$$s^2 = x_j^2 - \frac{1}{n}(\sum x_j)^2,$$

где x_j – значение признака j-го объекта, n – число объектов для классификации.

В рамках рассматриваемого метода объединяются те группы или объекты, для которых данная сумма получает минимальное приращение.

Использование кластерного анализа позволяет учесть всю совокупность изучаемых признаков пищевого сырья, дает возможность определить генетическую структуру имеющегося материала. Объединяются те группы или объекты, для которых данная сумма получает минимальное приращение. Полученные методом Уорда результаты дают наиболее объективную и информативную кластеризацию коллекции образцов тритикале. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

Ниже приведены итоги исследований, полученные в результате применения метода Уорда.

РАСПРЕДЕЛЕНИЕ НОМЕРОВ КОЛЛЕКЦИИ ТРИТИКАЛЕ ПО КЛАСТЕРАМ

число кластеров – 4:

[1] 23 29 30 33 40 41 42 43 45 47 48 49 50 51 56 59 62 64 92 127 173 187 192

[2] 24 32 35 37 38 58 60 61 63 66 70 72 74 75 76 77 79 81 82 86 87 108 114 115 123 124 128 132 137 138 140 143 144 146 153 154 155 156 157 164 165 166 167 169 170 174 176 182 205

[3] 25 53 54 57 68 71 83 90 93 95 97 100 102 110 142 148 149 152 160 161 163 191 194 195 203

[4] 26 27 28 31 34 36 39 44 46 52 55 65 67 69 73 78 80 84 85 88 89 91 94 96 98 99 101 103 104 105 106 107 109 111 112 113 116 117 118 119 120 121 122 125 126 129 130 131 133 134 135 136 139 141 145 147 150 151 158 159 162 168 171 172 175 177 178 179 180 181 183 184 185 186 188 189 190 193 196 197 198 199 200 201 202 204 206 207 208

число кластеров – 10:

[1] 23 30 40 41 47 50 51 56 59 62 192

[2] 24 32 35 37 58 60 63 66 72 75 76 81 86 87 108 114 115 123 124 128 132 137 138 140 146 156 169 174 205

[3] 25 53 57 68 95 195
[4] 26 27 31 36 44 65 89 103 104 105 107 109 117 126 131 136 139 141 145 147 171 172 175 178 180 181 184 186 188 190 196 202
[5] 28 34 85 106 116 122 135 159 168 177
[6] 29 33 42 43 45 48 49 64 92 127 173 187
[7] 38 61 70 74 77 79 82 143 144 153 154 155 157 164 165 166 167 170 176 182
[8] 39 46 52 55 67 73 78 84 88 91 96 98 111 112 118 119 121 133 150 151 158 162 179 183 193 197 200 204 208
[9] 54 71 83 90 93 97 100 102 110 142 148 149 152 160 161 163 191 194 203
[10] 69 80 94 99 101 113 120 125 129 130 134 185 189 198 199 201 206 207

число кластеров – 25:

[1] 23 30 40 41 47 50 51 56 59 62 192
[2] 24 35 63 72 81 146
[3] 25 195
[4] 26 27 36 109 126 131 136 141 147 171 172 175 181 188 196 202
[5] 28 34 85 106 116 122 135 159 168 177
[6] 29 42 48
[7] 31 44 65 117 139 180 184
[8] 32 37 58 75 87 115 123 124 169
[9] 33 45 92 127 173 187
[10] 38 143
[11] 39 46 52 55 67 84 88 91 98 111 118 119 121 133 150 151 158 162 179 183 193 200 204 208
[12] 43 49 64
[13] 53 57 68 95
[14] 54 71 97 142 148 149 152 161 163 194 203
[15] 60 66 76 86 108 114 128 132 137 138 140 146 156 169 174 205
[16] 61 144 154 157 164 165 166 167 170
[17] 69 94 101 113 120 125 130 189 201 206 207
[18] 70 74 77 153 155 182
[19] 73 78 96 112 151 197
[20] 79 82 176
[21] 80 99 129 134 185 198 199
[22] 83 90 93 100 160 191
[23] 89 104 105 131 178 186 190
[24] 102 110
[25] 103 107

В качестве теста получена следующая динамика отклонений: для 4 кластеров (грубая кластеризация) – 5,4 %, 10 – 10,2 %, 25 – 12,9 % . Сравнение с расчетами по слою Кохонена для 25 кластеров выявило полное совпадение в 3 из них и частичное совпадение в 9. Таким образом, полученные по методу Уорда результаты дают наиболее объективную и информативную кластеризацию коллекции тритикале. Предложенная модель кластеризации в ходе экспериментов показала свою работоспособность.

Кластеризация пищевого сырья на примере образцов тритикале по методу Уорда предоставляет возможность проводить с высокой эффективно-

стью оценку исходного материала на ранних стадиях селекции зерновых культур.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Гребенникова И.Г., Алейников А.Ф., Стёпочкин П.И.** Диаллельный анализ числа колосков в колосе яровой тритикале // Сиб. вестн. с.-х. науки. – 2011. – № 7–8. – С. 77–85.
2. **Гребенникова И.Г., Алейников А.Ф., Стёпочкин П.И.** Диаллельный анализ длины колоса у яровой тритикале // Сиб. вестн. с.-х. науки. – 2010. – № 12. – С. 103–109.
3. **Гребенникова И. Г., Алейников А.Ф., Стёпочкин П.И.** Информационное обеспечение селекционного процесса тритикале // Вестн. НГАУ. – 2011. – № 4 (20). – С. 10–15.
4. **Гребенникова И. Г., Алейников А.Ф., Стёпочкин П.И., Чанышев Д.И.** Структура комплекса информационного обеспечения селекционного процесса тритикале // Новейшие направления развития аграрной науки в работах молодых учёных. Ч. 1: сб. тр. Междунар. науч. конф. молодых ученых. – Новосибирск, 2010. – С. 247–249.
5. **Grebennikova I. G., Aleynikov A. F., Stepochkin P. I.** Diallel Analysis of the spike lets per spise in spring triticale // Bulgarian Journal of Agricultural Science. – 2011. – № 6. – P. 755–759.
6. **Алейников А.Ф., Стёпочкин П.И., Чанышев Д.И., Голышев Д.Н.** Программно-алгоритмические средства и искусственные нейронные сети в селекции растений: метод. реком. – Новосибирск: Изд-во ИПФ «АГРОС», 2008. – 16 с.
7. **Чанышев Д.И., Алейников А.Ф.** Алгоритм прогнозирования показателей качества пищевого сырья растительного происхождения // Пища. Экология. Качество: труды VII междунар. научн.-практ. конф. – Новосибирск, 2010. – С. 258–260.
8. **Чанышев Д.И., Гребенникова И. Г., Алейников А.Ф., Стёпочкин П.И.** Алгоритм прогнозирования селекционной ценности образцов тритикале на основе искусственных нейронных сетей // Информационные технологии, системы и приборы в АПК: материалы V Междунар. науч.-практ. конф. «АГРОИНФО-2012». – Новосибирск, 2012. – С. 107–113.
9. **Чанышев Д.И., Алейников А.Ф., Гребенникова И.Г., Чешкова А.Ф., Стёпочкин П.И.** Кластерный анализ показателей качества пищевого сырья при селекции // Информационные технологии, системы и приборы в АПК: материалы VI Междунар. науч.-практ. конф. «АГРОИНФО-2015». – Новосибирск, 2015. – С. 280–286.
10. **Алейников А.Ф., Чанышев Д.И., Чаплина М.А**. Автоматизированный синтез патентоспособных технических решений преобразователей сигналов // Сиб. вестн. с.-х. науки. – 2009. – № 2. – С. 86–92.
11. **Notebook** «Нейронные сети» 2008 [Электронный ресурс]: http://pandia.org/text/78/102/560-3php
12. **Сорокин О.Д.** Прикладная статистика на компьютере. – Краснообск: РПО СО РАСХН, 2009. – 222 с.
13. **Жамбю М.** Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 345 с.

Поступила в редакцию 24.11.2016

D.I. CHANYSHEV[1], Researcher,
A.F. ALEYNIKOV[1,2], Doctor of Science in Engineering, Professor, Head Researcher,
I.G. GREBENNIKOVA[1], Candidate of Science in Agriculture, Laboratory Head,
A.F. CHESHKOVA[1], Candidate of Science in Physics & Mathematics, Laboratory Head,
P.I. STEPOCHKIN[1,3], Doctor of Science in Agriculture, Lead Researcher

[1]*Siberian Physical-Technical Institute of Agrarian Problems, SFSCA RAS*

Krasnoobsk, Novosibirsk Region, 630501, Russia

e-mail: sibfti.n@ngs.ru

[2]*Novosibirsk State Technical University*

20, Karl Marx Ave, Novosibirsk, 630073, Russia

e-mail: fti2009@yandex.ru

[3]*Siberian Research Institute of Plant Production and Breeding – Branch of the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences*

Krasnoobsk, Novosibirsk Region, 630501, Russia

e-mail: sibniirs@bk.ru

CLUSTERING OF TRITICALE COLLECTION SAMPLES TO BE USED IN BREEDING

There is given an approach to forecasting breeding values of spring triticale collection samples based on information obtained from studies on their quantitative traits. The algorithm of forecasts is cluster analysis carried out by three different ways. The features of clustering methods were studied by way of example of a research into triticale collection samples. The Ward's method, using dispersion analysis to study distances between clusters, has proven to allow more fully evaluating breeding material, and to increase effectiveness of selecting parental pairs. This method optimizes the minimum dispersion within the closest clusters, and aims at their integration and creation of small-sized clusters. Each step of the algorithm unites a pair of clusters, resulting in the minimum increase in the objective function that is the intra-group sum of squared deviations. The results obtained by the Ward's method have produced most objective and informative clustering of triticale collection.

Keywords: forecast, triticale, cluster analysis, Ward's method.