УДК 675.03.031.81:577.15

Доцент Е.А. Балашова, профессор В.К. Битюков,
аспирант Е.А. Саввина
(Воронеж. гос. ун-т. инж. технол.) кафедра информационных и управляющих систем, тел. (473) 255-38-75

Сравнительный анализ методов классификации при прогнозировании качества хлеба

Проведен сравнительный анализ методов классификации двухэтапного кластерного, дискриминантного анализа и нейронных сетей. Предложена система информативных признаков, классифицирующая с минимумом ошибок.

The comparative analysis of classification methods of two-stage cluster and discriminant analysis and neural networks was performed. System of informative signs which classifies with a minimum of errors has been proposed.

Ключевые слова: двухэтапный кластерный анализ, дискриминантный анализ, искусственные нейронные сети.

В последние годы среди специалистов значительно выросла популярность систем интеллектуального анализа данных. Именно они играют ведущую роль в прогнозировании качества готовой продукции. Поэтому вопросы качества продукции наиболее важны в технологии хлебопечения. Это обусловлено большим объемом и сложным характером анализируемых данных, которые невозможно учесть при прогнозировании качества продукции. В таких системах используются методы кластерного, дискриминантного анализа и нейронные сети.

Целью данной работы был сравнительный анализ кластерных, дискриминантных и нейросетевых методов классификации, выявление наиболее информативных факторов, при которых количество ошибок сводится к минимуму.

В ходе выполнения работы была сформирована база данных, состоящая из 595 анализов, характеризующих качество хлеба по 20 признакам. Качество хлеба описывалось органолептическими (влажность, массовая доля и качество клейковины и т.д.), химическими показателями муки (массовая доля жира, клетчатки, содержание углеводов и т.д.), а также показателями хлеба (влажность мякиша, пористость и кислотность). В соответствии с классификацией, предложенной Пономаревой Е.И. [3], качество белого хлеба подразделяется на 4 основные группы: 1 группа (высшее качество) – 140 наблюдений (23,5 %), 2 группа (хорошее качество) – 195 (32,8 %), 3 группа (плохое качество) – 140 (23,5 %), 4 группа (очень плохое качество) – 120 (20,2 %).

Структуру базы данных составляют не только количественные (влажность муки, активная и титруемая кислотность, массовая доля клейковины, качество клейковины и т.д.), но и качественные признаки (наличие хруста, горькости вкуса, кислоты, зараженности вредителями и т.д.). Значения качественных признаков были кодированы цифрами и буквами. Исходные категориальные признаки были формализованы в бинарные, каждый из которых имел 2 состояния (0 – признак отсутствует, 1 – присутствует). В результате количество признаков в базе данных увеличилось до 27.

Обработка данных проводилась кластерными, дискриминантными и нейросетевыми методами. Метод двухэтапного кластерного анализа (Two Step Cluster) позволяет кластеризовать различные группы по отдельности, а после этого объединять полученные результаты в конечную структуру кластеров. Для измерения расстояния между объектами используется Евклидова метрика

$$d_{kl} = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{ij})^2}, \quad (1)$$

где d_{kl} - расстояние между объектом k и l, а x_{kj} -и x_{ij} - это j-е свойства объектов соответственно k и l.

Число кластеров в двухэтапном кластерном анализе может быть задано автоматически или рассчитано по критерию Акаике (AIC):

$$AIC_k = -2L_k + 2r_k, \quad (2)$$

где r_k - число параметров или информационный критерий Байеса

$$BIC_k = -2L_k + r_k \log n \quad (3)$$

Каноническая дискриминантная функция вычисляется по формуле:

$$F(x) = a_1 x_1 + a_2 x_2, \quad (4)$$

где a_1, a_2 – коэффициенты функции, x_1, x_2 - дискриминантные переменные.

Коэффициенты дискриминантной функции a_i определяются таким образом, чтобы средние значения функций $\overline{f_1}(x)$ и $\overline{f_2}(x)$, как можно больше различались между собой, т.е. чтобы для двух множеств (классов) было максимальным выражение

$$\overline{f_1}(x) - \overline{f_2}(x) = \sum_{i=1}^{n_1} a_1 x_{1i} - \sum_{i=1}^{n} a_1 x_{2i}, \quad (5)$$

Вектор коэффициентов дискриминантной функции определяется по формуле:

$$A = S_*^{-1}(\overline{X_1} - \overline{X_2}), \quad (6)$$

где S_*^{-1}- объединенная ковариационная матрица признаков

$$S_* = \frac{1}{n_1 + n_2 - 2}(X_1^{/} X_1 + X_2^{/} X_2), \quad (7)$$

где X – матрицы отклонений наблюдаемых значений исходных переменных от их средних величин в группах.

Методы нейронных сетей моделируют функции биологического нейрона, то есть формируют выходной сигнал в зависимости от сигналов, поступающих на его входы. Состояние нейрона характеризуется величиной синаптической связи (весом w_i) и определяется по формуле:

$$NET = \sum_{i=1}^{n} x_i w_i \quad (8)$$

где *NET* – суммирующий блок, который складывает взвешенные входы алгебраически, создавая выход, x_i – множество входных сигналов поступающих на искусственный нейрон, w_i – множество весов сигнала.

Для классификации с высокой точностью необходимо выявление наиболее информативных признаков. Информативность признаков определяется коэффициентом корреляции Пирсона, то есть чем больше корреляция, тем больше сходство между объектами.

С помощью корреляционного анализа в общей выборке было выявлено, что признаки коррелируют с классом качества на уровне 0,01. Для класса 1 был выявлен один специфический признак (содержание водорастворимых углеводов X_{23}), коэффициент корреляции равен 0,819, теснота связи сильная. Класс 2 не имеет специфичных признаков, лишь для 3 признаков коэффициент корреляции превышает 0,5, теснота связи средняя. В классе 3 информативных признаков обнаружено не было, только один признак (зараженность вредителями X_{17}) имеет коэффициент корреляции более 0,5. Для класса 4 было выявлено 9 специфичных признаков, коэффициент корреляции которых превышает 0,7 и лежит в диапазоне от 0,717 до 0,801, теснота связи сильная. 6 признаков имеют среднюю тесноту связи и коэффициент корреляции более 0,5. Данный набор признаков был использован для всех методов классификации.

Выявление информативных признаков позволяет сделать вывод о возможности выделения 4 класса качества. Классы 1, 2 и 3 выявить невозможно из-за небольшого количества специфических признаков, в этой связи была построена иерархическая схема классификации, представленная на рис. 1.

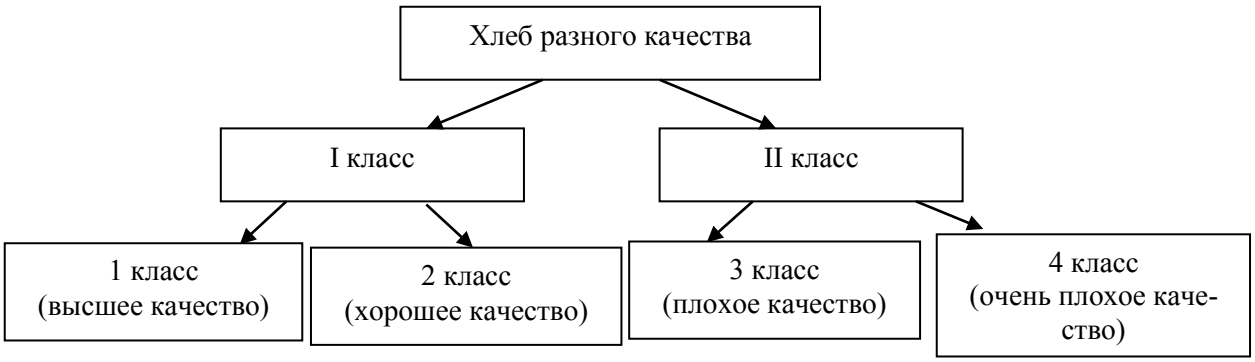

Рис. 1. Иерархическая схема классификации

Для классификации методом двухэтапного кластерного анализа были выбраны признаки, имеющие значимую корреляцию с классом качества. Была установлена двухкластерная структура данных. К первому классу (I) относится хлеб высшего, хорошего и плохого качества, ко второму (II) – очень плохого качества (табл. 1). Процент правильно классифицированных наблюдений составил 93,27 %.

Таблица 1

Результат классификации кластерным анализом

Кластер	Распределение по кластерам		% ошибок
	N	% объединенных	
Класс I Хлеб хорошего качества	435	73,1 %	
Класс II Хлеб очень плохого качества	160	26,9 %	6,73 %
Итого	595	100 %	6,73 %

Проведем разбиение класса I на подклассы (табл. 3). Класс I объединяет 435 наблюдений. Для данной подвыборки проведем корреляционный анализ (табл. 2).

Таблица 2

Таблица коэффициентов корреляции

Признак	Класс 1	Класс 2	Класс 3
Влажность муки (X_1)	0,742**	-0,307**	-0,462**
Титруемая кислотность (X_2)	0,722**	-0,238**	-0,520**
Вкус свойственный (X_{10})	-0,307**	-0,382**	0,756**
Вкус кислый (X_{11})	-0,248**	0,362**	0,704**
Массовая доля золы (X_{20})	0,143**	0,741**	-0,145**
Зараженность вредителями	-0,337**	-0,441**	0,895**
Содержание водорастворимых углеводов (X_{23})	0,834**	-0,327**	-0,540**

* - корреляция значима на уровне 0,05 ** - корреляция значима на уровне 0,01

Таблица 3

Результат классификации кластерным анализом

Кластер	Распределение по кластерам		% ошибок
	N	% объединенных	
Класс 1 Хлеб хорошего качества	175	40,2 %	20 наблюдений 4,6 %
Класс 2 Хлеб очень хорошего качества	196	45,1 %	
Класс 3 Хлеб плохого качества	64	14,7 %	36 наблюдений 8,3 %
Итого	435	100,0 %	

Недостатком такого метода является классификация несколькими этапами: на первом этапе выделяются 2 класса (I класс – хорошее и очень хорошее качество, II класс – плохое и очень плохое качество). На втором этапе данные классы разделяются на подклассы 1,2,3,4.

Пошаговым дискриминантным анализом с критерием отбора статистики Уилкса (λ Уилкса) были построены уравнения дискриминантных функций (их значения представлены в таблице 4) D_1, D_2, D_3 разделяющие выборку на классы:

$$D_1 = -2{,}384+0{,}246X_2-0{,}317X_4-0{,}928X_{16}+1{,}604X_{20}+0{,}370X_{21}+0{,}774X_{23}+0{,}189X_{24}$$
$$D_2 = -6{,}506+1{,}760X_2+0{,}425X_4-0{,}880X_{16}-6{,}014X_{20}-2{,}362X_{21}-3{,}739X_{23}+0{,}363X_{24} \quad , \qquad (8)$$
$$D_3 = -25{,}940+2{,}751X_2+0{,}407X_4-0{,}543X_{16}-0{,}663X_{20}-3{,}077X_{21}-1{,}138X_{23}-0{,}329X_{24}$$

Таблица 4

Значения дискриминантных функций

Функция	Собственное значение	% объясненной дисперсии	Каноническая корреляция	λ – Уилкса	χ – квадрат
D_1	9,843	69,2	0,953	0,012	2605,37
D_2	3,727	26,2	0,888	0,128	1206,22
D_3	0,651	4,6	0,628	0,606	294,44

Установлено, что значимость по коэффициенту Уилкса для дискриминантных функций не превышает 0,0001, следовательно, использование данных функций для дискриминации целесообразно. Наибольший вклад в дискриминацию вносит первая дискриминантная функция, так как внутригрупповые корреляции между дискриминантной функцией и каноническими переменными имеют среднюю тесноту связи, коэффициент корреляции превышает 0,5.

Результаты расчетов показали, что число случаев ложной тревоги составило 18 (3,37 %), причем 5 (0,8 %) из них это отнесение хорошего качества к плохому, и 16 (2,69 %) – распознавание плохого качества как очень плохое. Один случай (0,7 %) – отнесение плохого качества хлеба к хорошему качеству. По результатам классификации было выявлено, что высокая точность 100% достигается в 4 классе (очень плохое качество). В первом классе точность классификации составила – 99,3 %, во втором – 97,4 %, в третьем – 87,9 %. Методом дискриминантного анализа 96,1 % наблюдений были классифицированы верно.

На рис. 2 и 3 приведены объединенные графики распределения всех классов с центроидами.

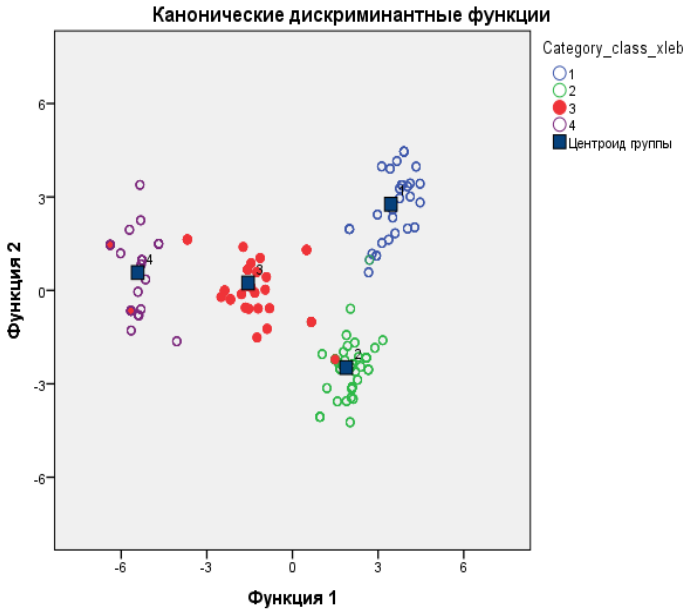

Рис.2 Диаграмма рассеяния для всех групп

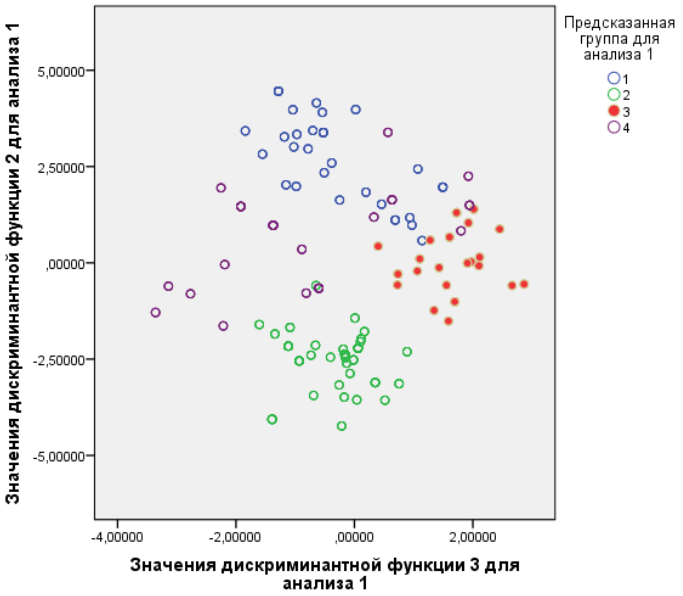

Рис.3 Расположение наблюдений для двух дискриминантных функций 3 и 2

При классификации методом нейронных сетей из общей выборки случайным образом были отобраны 348 наблюдений – для обучающей выборки, 121 – для контрольной, 126 – для проверочной.

Оценка качества функционирования диагностической системы проводилась на проверочной.

Была построена архитектура нейронной сети состоящей из 8 факторов (X_2, X_4, X_7, X_9, X_{20}, X_{21}, X_{22}, X_{26}) и 2 стандартизированных ковариатов (X_8, X_{16}). Нейронная сеть содержит 1 скрытый слой и 4 нейрона на скрытом слое. Архитектура нейронной сети представлена на рис. 4.

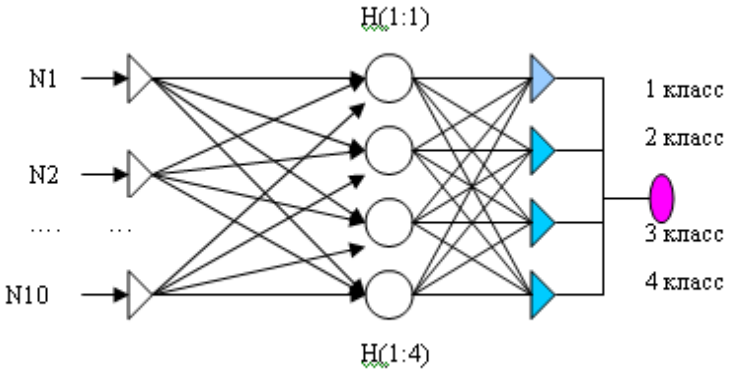

Рис.4. Архитектура нейронной сети

Анализ полученных результатов по проверочной выборке показал, что число ложных тревог и пропуска сигнала сократилось до 2,4 %. Наивысшая точность (100 %) была достигнута в следующих классах: 1 высшего качества и 4 очень плохого качества. В группе с хорошим качеством одно наблюдение (2,5 %) неправильно классифицировано как высшее качество. Точность классификации в данной группе составила 97,5 %. В 3 группе с плохим качеством было выявлено 2 ошибки (5,9 %) неправильной классификации как очень плохого качества. Данную ошибку не стоит принимать во внимание, так как данные два класса с плохим качеством не должны использоваться в хлебопечении. Таким образом, система выполняет небольшую гипердиагностику. Точность всей классификационной системы составила 97,6 %. Результаты сравнительного анализа представлены в табл. 5.

Таблица 5

Сводная таблица результатов по всем методам

Метод	Точность метода	Процент ошибок
Двухэтапный кластерный анализ	96,3 %	3,7 %
Дискриминантный анализ	96,1 %	3,9 %
Нейросетевой анализ	97,6 %	2,4 %

В заключение можно подвести некоторые итоги:

- предложен корреляционный анализ для отбора наиболее информативных признаков. Проведена классификация наблюдений двухэтапным кластерным, дискриминантным и нейросетевым методом. Показано, что коэффициент корреляции влияет на точность классификации объектов.

- выявлена система наиболее информативных признаков, позволяющая классифицировать качество пшеничного хлеба на классы. Методом двухэтапного кластерного анализа была получена двухкластерная структура данных, один кластер образует высшее и хорошее качество, другой – плохое и очень плохое качество. Дискриминантный и нейросетевой методы позволили выделить 4 класса качества за одну итерацию.

- проведенные исследования показали возможность применения методов кластерного, дискриминантного анализа и нейронных сетей для диагностики качества хлебобулочных изделий с точностью 96,3 %, 96,1 %, 97,6 % соответственно.

ЛИТЕРАТУРА

1 Битюков, В. К. Итерационный алгоритм диагностики систем, описываемых набором качественных признаков [Текст] / В. К. Битюков, Е. А. Балашова, К. О. Сунцов // Системы управления и информационные технологии. – 2008. – № 4.1 (34). – С. 134-138.

2 Ким, Дж.-О. Факторный, дискриминантный и кластерный анализ [Текст] / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.– М.: Финансы и статистика, 1989. – 215 с.

3 Санина, Т. В. Балльная оценка качества хлебобулочных изделий [Текст] / Т. В. Санина, Е. И. Пономарева. – Воронеж: ВГТА, 2008. – 144 с.

REFERENCES

1 Bityukov, V. K. Iterative algorithm diagnostic systems described by qualitative features of boron-[Text] / V. K. Bityukov, E. A. Balashov, S. C. Suntsov // Control systems and information technology. - 2008. - № 4.1 (34). - C. 134-138.

2 Kim, Dzh.-O. Factorial, discriminant and cluster analysis [text] /Dzh.-O. Kim, C. / Ch.U. Myuller, U.R. Klekka andothers - Moscow: Finance and Statistics, 1989. - 215 p.

3 Sanin, T. V. Scopingquality of bakery products [Text] / T. V. Sanin, E. I. Ponomarev. - Voronezh: VSTA, 2008. - 144 p.