

УДК 007.001.362

Профессор В.К. Битюков, ассистент М.Л. Моторин,
ассистент Е.А. Саввина

(Воронеж. гос. ун-т инж. технол.) кафедра информационных и управляющих систем,
тел. (473) 255-38-75

E-mail: zhuravlyevaEA@yandex.ru

Professor V.K. Bitiukov, assistant M.L. Motorin,
assistant E.A. Savvina

(Voronezh state university of engineering technology) Department of information and control
systems, phone (473) 255-38-75

E-mail: zhuravlyevaEA@yandex.ru

Формирование классов объектов методом дискриминантного анализа

The formation classes of objects by the method of discriminant analysis

Реферат. В работе представлен метод дискриминантного анализа, как современный инструмент для классификации объектов, на примере мукомольного производства. Дискриминантный анализ является статистическим методом, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно и дает возможность классифицировать объекты по принципу максимального сходства. Содержанием дискриминантного анализа является разработка и исследование статистических методов для изучения различий между двумя и более группами объектов по нескольким переменным одновременно при помощи дискриминантной линии. В дискриминантном анализе, в отличие от кластерного, имеется обучающая выборка, в которой известно к каким классам относятся объекты. По обучающей выборке получают правила, которые в дальнейшем позволяют определить, к какому классу относятся новые объекты. Построены дискриминантные функции, графики распределения объектов по классам качества, графически представлена методика классификации. В ходе выполнения работы была сформирована база данных, состоящая из 595 анализов, характеризующих качество муки по 15 признакам. Каждый анализ описывался химическими (массовая доля белка, массовая доля золы, массовая доля жира, содержание клетчатки и водорастворимых углеводов) и органолептическими показателями качества муки (влажность, титруемая и активная кислотность, массовая доля и качество клейковины, вкус, запах, хруст и т.д.). Точность классификации методом дискриминантного анализа составила 576 (98,02 %).

Summary. The paper provides the method of discriminant analysis as a modern tool for the classification of objects by the example of flour production. Discriminant analysis is a statistical technique that allows us to study the differences between two or more groups of objects on multiple variables simultaneously and provides the ability to classify objects according to the principle of maximum similarity. Content of discriminant analysis is the development and study of statistical methods to examine the differences between two or more groups of objects on multiple variables simultaneously with the dominant line. In discriminant analysis, in contrast to the cluster, there is a training set, which is known what classes are objects. The training set is obtained rules, which further allow you to determine what class are new objects. Built discriminant functions, graphs of distribution of objects on quality classes, graphically presents classification methodology. During the performance was formed database consisting of 595 analyzes characterizing the quality of flour by 15 characters. Each assay described chemical parameters (mass fraction of protein mass fraction of ash, the mass fraction of fat, fiber content and water-soluble carbohydrates) and organoleptic quality of flour (moisture content, titratable acidity and active, and the mass fraction of gluten quality, taste, smell, and the crunch etc.). Classification accuracy of the method of discriminant analysis was 576 (98.02%).

Ключевые слова: каноническая дискриминантная функция, дискриминант, классы объектов, классификация

Keywords: canonical discriminant function, the discriminant, classes of objects, classification

Дискриминантный анализ является статистическим методом, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно и дает возможность классифицировать объекты по принципу максимального сходства.

Цель работы: рассмотреть классификационные различия между классами объектов и графически интерпретировать полученные результаты.

«Дискриминантный анализ» - это общий термин, относящийся к тесно связанным статистическим процедурам: интерпретации и классификации. Интерпретация предполагает поиск различий между классами при данном наборе характеристик (переменных). Метод, относящийся к классификации, связан с получением одной или нескольких функций, обеспечивающих возможность отнести каждый объект к одной из групп [1].

© Битюков В.К., Моторин М.Л., Саввина Е.А., 2014

В дискриминантном анализе используются методы максимального правдоподобия:

1) расчет априорных вероятностей принадлежности к классу, которые можно рассматривать как решающее правило, применяемое в том случае, когда нет никакой дополнительной информации об объектах;

2) расчет условных вероятностей принадлежности к классу, каждая из которых равна вероятности получить соответствующее значение дискриминантной функции при условии, что объект принадлежит классу.

Содержанием дискриминантного анализа является разработка и исследование статистических методов для изучения различий между двумя и более группами объектов по нескольким переменным одновременно при помощи дискриминантной линии.

В дискриминантном анализе, в отличие от кластерного, имеется обучающая выборка, в которой известно к каким классам относятся объекты. По обучающей выборке получают правила, которые в дальнейшем позволяют определить, к какому классу относятся новые объекты.

Функция с линейной комбинацией вида:

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \dots + u_p X_{pkm}, \quad (1)$$

где f_{km} – значение канонической дискриминантной функции для m -го объекта в группе k ; X_{1km} – значение дискриминантной переменной X_i для m -го объекта в группе k ; u_i – коэффициенты, обеспечивающие выполнение требуемых условий.

Коэффициенты u_i для первой функции выбираются таким образом, чтобы ее средние значения для различных классов как можно больше отличались друг от друга. Коэффициенты второй функции выбираются так же, то есть соответствующие средние значения должны максимально отличаться по классам, при этом налагается условие, чтобы значения второй функции были некоррелированы со значениями первой.

При классификации указанный объект «принадлежит к» или «очень похож на» данный класс. Такое решение принимается на основе информации, содержащейся в дискриминантных переменных.

Значимость дискриминантной функции при классификации определяется несколькими способами:

1. По расчету собственного значения функции:

$$\lambda = \frac{\sum b_{ij} v_i}{\sum w_{ij} v_i}, \quad (2)$$

где λ – собственное значение; v_i – последовательность коэффициентов; w_{ij} , b_{ij} – рассчитаны соотношениями (3) и (4) соответственно:

$$w_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - X_{ik})(X_{jkm} - X_{jk}), \quad (3)$$

$$b_{ij} = t_{ij} - w_{ij}. \quad (4)$$

2. По проценту объясненной дисперсии:

$$\lambda_{\text{общ}} = \lambda_1 + \lambda_2 + \dots + \lambda_n, \quad (5)$$

$$d = \frac{\lambda_1}{\lambda_{\text{общ}}}. \quad (6)$$

Процент объясненной дисперсии показывает, какова функция по сравнению с другими, то есть та, которая имеет большее процентное содержание, сильнее влияет на дискриминацию.

3. По расчету коэффициента канонической корреляции, который является мерой связи (степени зависимости между группами и дискриминантной функцией).

4. По расчету Λ – статистики Уилкса – меры различий между классами по нескольким дискриминантным переменным:

$$\Lambda = \prod_{i=k+1}^g \frac{1}{1 + \lambda_i}, \quad (7)$$

где k – число уже вычисленных функций; символ $\prod$ означает, что для получения окончательного результата необходимо перемножить все члены. Λ является «обратной» мерой, потому что величины Λ , близкие к нулю, говорят о высоком различии (то есть центроиды классов сильно отличаются друг от друга).

Каждый объект принадлежит одной из групп, поэтому вероятность принадлежности вычисляется для любой из групп. Вероятность того, что объект X является членом класса k , равна:

$$Pr(G_k|X) = \frac{Pr(X|G_k)}{\sum_{i=1}^g Pr(X|G_i)}, \quad (8)$$

где $Pr(X|G_k)$ – вероятность того, что объект, находящийся далеко от центроида, действительно принадлежит классу k .

Стандартизованной мерой эффективности для любого количества классов является t -статистика ошибок:

$$\tau = \frac{n_c - \sum_{i=1}^g p_i n_i}{n_{\text{общ}} - \sum_{i=1}^g p_i n_i}, \quad (9)$$

где τ – число правильно классифицированных объектов; p_i – априорная вероятность принадлежности к классу; а $\sum_{i=1}^g p_i n_i$ представляет

собой число объектов, которые будут правильно но предсказаны при случайной классификации.

Дискриминантные переменные – оси p -мерного евклидова пространства. Каждый объект описывается точкой этого пространства с координатами, представляющими собой наблюдаемые значения каждой переменной. Если классы отличаются друг от друга по наблюдаемым переменным, их можно представить как скопления точек в некоторых областях рассматриваемого пространства. Поскольку классы могут частично «перекрываться», соответствующие им «территории» не совпадают. Для определения положения класса вычисляется его «центроид» – точка, координаты которой есть средние значения переменных в данном классе.

Радиус ядра разброса наблюдений рассчитывается по формуле:

$$R_i = \bar{r}_i + \sqrt{2} * \sigma_i, \quad (10)$$

где $\bar{r}_i$ среднее расстояние между центром кластера и кластерными элементами, а σ_i среднеквадратическое отклонение величины.

В ходе выполнения работы была сформирована база данных, состоящая из 595 анализов, характеризующих качество муки по 15 признакам. Каждый анализ описывался химическими (массовая доля белка, массовая доля золы, массовая доля жира, содержание клетчатки и водорастворимых углеводов) и органолептическими показателями качества муки (влажность, титруемая и активная кислотность, массовая доля и качество клейковины, вкус, запах, хруст и т.д.). В соответствии с классификацией, предложенной в [3], наблюдения были разделены на 4 группы качества. Первая группа (класс 1 высшего качества) – 140 (23,3 %) наблюдений; вторая (класс 2 хорошего каче-

ства) – 195 (33,3 %); третья (класс 3 плохого качества) – 140 (23,3 %); четвертая (класс 4 очень плохого качества) – 120 (20,0 %).

Исходные качественные признаки были формализованы в категориальные. Качественный признак, например, «вкус», переведен к бинарному виду, где каждый признак имел 2 состояния (0 – признак отсутствует, 1 – присутствует): вкус свойственный (1 – присутствует, 0 – отсутствует); вкус заплесневелый (1 – присутствует, 0 – отсутствует); вкус затхлый (1 – присутствует, 0 – отсутствует); вкус кисловатый (1 – присутствует, 0 – отсутствует); горьковатый (1 – присутствует, 0 – отсутствует). Таким образом, количество признаков увеличилось до 25 [2].

Исходные данные представляют собой случайные величины. Полная информация о случайной величине дается ее функцией распределения. По критерию Колмогорова-Смирнова установлено, что для подавляющего большинства признаков распределение отличается от нормального на уровне значимости $P=0,05$. Нормальному закону распределения подчиняются признаки: газообразующая способность и содержание водорастворимых углеводов.

Для принятия решений об отнесении хлеба к определенному классу необходимо отобрать наиболее информативные признаки. Выявление наиболее информативных признаков и классификация осуществлялась в два этапа. На первом этапе использовался корреляционный анализ. На втором этапе формировалась классификационная система признаков методом дискриминантного анализа. Методика классификации методом дискриминантного анализа представлена на рисунке 1.

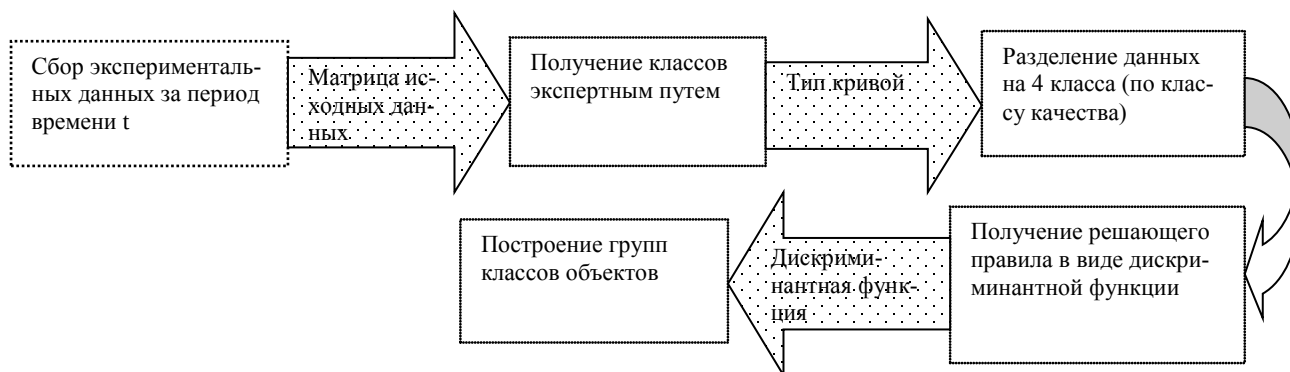


Рисунок 1. Методика классификации методом дискриминантного анализа

В качестве критерия информативности признаков принят коэффициент корреляции между признаком и классом качества муки, определенным экспертом.

Для первого класса были выявлены 2 специфических признака: цвет муки белый X_{10} и содержание водорастворимых углеводов X_{28} , коэффициент корреляции которых варьируется от 0,755 до 0,819, то есть теснота связи сильная. Для 6 признаков (титруемая кислотность X_2 , активная кислотность X_3 , массовая доля клейковины X_4 , число падения X_7 , цвет муки X_8 , белизна X_{26} , зольность X_{27}) значение коэффициента корреляции превышает 0,5 по модулю, то есть теснота связи средняя.

Для второго класса выявлен 1 специфический признак: цвет муки с желтоватым оттенком X_{12} , коэффициент корреляции которого 0,826, то есть теснота связи сильная. Для 3 признаков (вкус свойственный X_{14} , отсутствие затхлого вкуса X_{18} , запах свойственный X_{19} , отсутствие хруста X_{23}) коэффициент корреляции находится в диапазоне от 0,508 до 0,655 по модулю, то есть теснота связи средняя.

Установлено, что третий класс имеет 1 специфический признак: цвет муки серый X_{11} , значение r равно 0,748. Для 4 признаков: вкус несвойственный X_{14} , вкус кислый X_{15} , вкус затхлый X_{18} , запах несвойственный X_{19} , значение коэффициента корреляции варьируется от 0,550 до 0,691, то есть теснота связи средняя.

Для четвертого класса было получено 7 специфических признаков: титруемая кислотность X_2 , массовая доля клейковины X_4 , число падения X_7 , крупность помола X_9 , вкус плесневелый X_{17} , зольность X_{27} , коэффициент корреляции которых лежит в диапазоне от 0,717

до 0,952. Для 9 признаков (активная кислотность X_2 , цвет серый с частицами X_{13} , вкус несвойственный X_{14} , вкус плесневелый X_{17} , активная кислотность X_3 , запах с посторонними примесями X_{22} , наличие хруста X_{23} , наличие металлопримесей X_{25} , белизна X_{26} , содержание водорастворимых углеводов X_{32} , коэффициент корреляции находится в диапазоне от 0,563 до 0,656 по модулю и имеется средняя теснота связи с классом качества.

Для повышения точности классификации был использован метод дискриминации. Преимуществом данного метода является автоматическое построение графиков распределения классов и функций дискриминации.

Была сформирована обучающая выборка из 50 (8,4 %) наблюдений, содержащая объекты всех классов муки. Также как и для кластерного анализа были приняты признаки, отобранные на первом этапе методом корреляционного анализа, имеющие значимую корреляцию с классом качества. Классификационная система признаков включает в себя: цвет муки белый X_{10} , цвет муки серый X_{11} , цвет муки с желтоватым оттенком X_{12} , вкус плесневелый X_{17} , титруемая кислотность X_2 , массовая доля клейковины X_4 , число падения X_7 , цвет муки X_8 , крупность помола X_9 , зольность X_{27} , содержание водорастворимых углеводов X_{28} .

Полезность дискриминантной функции оценивалась при помощи расчета коэффициента канонической корреляции (3); различие между классами - с помощью λ -статистики Уилкса (4); значимость дискриминантных функций - с помощью распределения χ^2 (хи-квадрат). Результат дискриминантного анализа представлен в таблице 1.

Таблица 1

Результат дискриминантного анализа

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция	Лямбда Уилкса	Хи-квадрат
D1	5901,95	98,4	98,4	1,000	0,000	9118,57
D2	83,95	1,4	99,8	0,994	0,001	4051,92
D3	11,208	0,2	100,0	0,958	0,082	1459,97

Функция D1 обладает большими дискриминирующими возможностями, так как ее собственное значение равно 5901,95. Процент объясненной дисперсии 98,4 %, значение канонической корреляции 1,0 и распределение χ^2 9118,57 подтверждает дискриминационные

возможности функции D1. Так как значения функций D2 и D3 по данным критериям несколько ниже, данные функции не сильно влияют на процесс дискриминации.

Канонические дискриминантные функции D1, D2, D3, разделяющие выборку на 4 класса:

$$D1 = -163,01 - 0,2X_2 + 0,01X_4 - 0,01X_7 + 0,01X_9 - 0,56X_{12} + 0,27X_{15} - 0,3X_{16} - 0,07X_{17} - 0,4X_{20} - 0,41X_{22} + 223,14X_{25}; \quad (11)$$

$$D2 = -6,17 + 0,3X_2 - 0,01X_4 + 0,03X_7 - 0,01X_9 + 0,17X_{12} - 4,1X_{15} + 1,05X_{16} - 0,84X_{17} + 0,59X_{20} - 0,2X_{22} + 1X_{25}; \quad (12)$$

$$D3 = 12,61 - 2,01X_2 - 0,561X_4 - 0,027X_7 - 0,026X_9 + 2,21X_{12} + 2,48X_{15} + 5,33X_{16} + 2,33X_{17} + 1,18X_{20} + 1,46X_{22} + 3,275X_{25}. \quad (13)$$

Уравнения линейных простых классифицирующих функций $D1_k$, $D2_k$, $D3_k$, $D4_k$,

позволяющие относить объекты к одному из 4 классов качества:

$$D1_k = -3346,35 + 85,73X_2 + 48,19X_4 + 2,29X_7 + 1,57X_9 - 31,29X_{12} + 115,86X_{15} + 28,91X_{16} + 29,21X_{17} - 60,80X_{20} + 70,56X_{22} + 13936,45X_{25}; \quad (14)$$

$$D2_k = -5034,72 + 77,66X_2 + 44,98X_4 + 1,99X_7 + 1,44X_9 - 31,74X_{12} + 137,58X_{15} + 59,14X_{16} + 41,51X_{17} - 63,9X_{20} + 69,42X_{22} + 47887,02X_{25}; \quad (15)$$

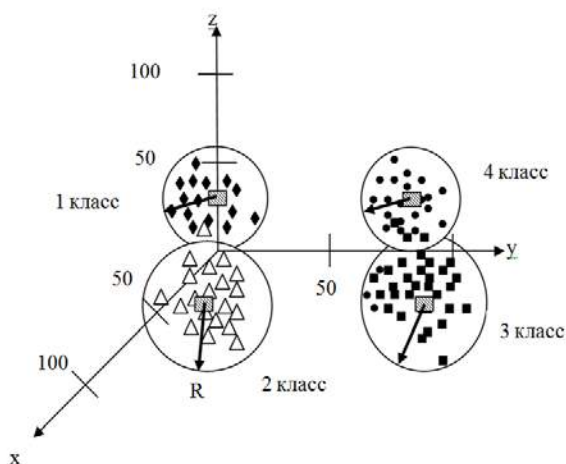
$$D3_k = -39368,68 + 115,34X_2 + 47,81X_4 + 1,074X_7 + 1,71X_9 - 144,32X_{12} + 164,55X_{15} - 15,2X_{16} + 18,39X_{17} - 143,9X_{20} - 13,02X_{22} + 155295,13X_{25}; \quad (16)$$

$$D4_k = -39421,92 + 113,14X_2 + 47,94X_4 + 0,79X_7 + 1,72X_9 - 146,33X_{12} + 201,51X_{15} - 25,4X_{16} + 25,79X_{17} - 149,62X_{20} - 11,99X_{22} + 155276,21X_{25}. \quad (17)$$

Первые две наиболее дискриминирующие канонические дискриминантные функции и центры в классах качества представлены на рисунке 2. Центроиды четырех классов имеют следующие координаты: (-68,74; 14,46; -0,11), (-66,85; -10,51; 0,05), (86,59; 0,33; 4,62), (87,81; -0,18; -5,43).

Рисунок 2 показывает, что группы 1 и 2 вполне различимы, то есть центроиды хорошо отделимы друг от друга. На рисунке четко видно попадание одного ошибочно отнесенного наблюдения второго класса к первому. Однако выявлено перекрытие областей между 3 и 4 классами качества. Это обусловлено тем, что значения дискриминантных функций в центроидах групп для 3 и 4 классов близки друг к другу: 86,59 и 87,81 соответственно.

Рисунок 2. Канонические дискриминантные функции для четырех классов качества и центры групп



Результаты классификации методом дискриминантного анализа представлены в таблице 2.

Таблица 2

Результат классификации

Категория класса качества		Предсказанная принадлежность к группе				Итого
		1	2	3	4	
Частота	1 класс	140	0	0	0	140
	2 класс	1	194	0	0	195
	3 класс	0	0	137	3	140
	4 класс	0	0	3	117	120
%	1 класс	100	0	0	0	100,0
	2 класс	0,5	99,5	0	0,0	100,0
	3 класс	0,0	0	97,2	2,8	100,0
	4 класс	0,0	0,0	1,7	98,3	100,0

С учетом изложенного, по результатам расчета методом дискриминантного анализа выявлено, что число случаев ложной тревоги составило 3 (2,8 %) - это отнесение плохого качества к худшему. Выявлено 4 (2,3 %) случая пропуска сигнала, из них в 1 (0,5 %) наблюдении хорошее качество было неверно классифицировано как очень хорошее качество, в 3 (1,7 %) других случаях очень плохое качество расценено как плохое качество. По результатам дискриминации было выявлено, что в первом классе было верно классифицировано 140 (100,0 %) наблюдений, во втором классе – 194 (99,5 %), в третьем – 137 (97,2 %), в четвертом – 117 (98,3 %). Таким образом, точность классификации методом дискриминантного анализа составила 576 (98,02 %). При этом ошибки отнесения плохого качества к очень плохому не стоит принимать во внимание, так как мука и плохого и очень плохого качества не должна использоваться в хлебопечении.

ЛИТЕРАУРА

1 Клекка У.Р., Ким Дж.-О., Мьюллер Ч.У. Факторный, кластерный и дискриминантный анализ. М.: Финансы и статистика, 1989. 215 с.

2 Саввина Е.А., Балашова Е.А., Битюков В.К. Использование методов дискриминантного анализа для классификации качества муки // Финансы. Экономика. Стратегия. 2013. №3. С. 20-23.

3 Балашова Е.А., Битюков В.К., Саввина Е.А., Пономарева Е.И. Формирование системы информативных признаков для прогнозирования качества муки/ Е.А. Балашова, В.К. Битюков, Е.А. Саввина, Е.И. Пономарева // Сборник трудов 3-ей Международной научно-практической конференции «Ключевые вопросы в современной науке», 2013. С. 74-77.

REFERENCES

1 Klekka U.R., Kim Dzh.-O., M'iuller Ch.U. Faktornyi, klasternyi i diskriminantnyi analiz [Factor, cluster and discriminant analysis]. Moscow, Finansy i statistica, 1989. 215 p. (In Russ.).

2 Savvina E.A., Balashova E.A., Bitiukov V.K. Using the discriminant analysis methods for the classification of the quality of the flour. *Finansy. Ekonomika. Strategiya*. [Finance. Economy. Strategy], 2013, no. 3, pp. 20-23. (In Russ.).

3 Balashova E.A., Bitiukov V.K., Savvina E.A., Ponomareva E.I. Formation of the system of informative signs for prognose quality flour. *Sbornik trudov tret'ei Mezhdunarodnoi nauchno-practicheskoi konferentsii "Kliuchevye voprosy v sovremennoi nauke"* [Proceedings of the 3rd international scientific and practical conference «the Key issues in modern science»], 2013. pp. 74-77. (In Russ.).