УДК 675.03.031.81:577.15

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

Влияние типа данных на результаты классификации объектов

В работе показано влияние типа данных на классификацию объектов, выявлены наиболее информативные признаки для разных классов качества, результаты классификации подтверждены дискриминантным анализом.

This work shows the influence of the type of data on the classification of objects and revealed the most informative signs for different classes of quality, the results of the classification are confirmed by discriminant analysis.

Ключевые слова: двухэтапный кластерный анализ, дискриминаный анализ, коэффициент корреляции Пирсона.

Качество белого хлеба из пшеничной муки зависит от качества рецептурных компонентов и точности соблюдения норм технологического процесса. При одних и тех же параметрах протекания технологического процесса возможно получение хлеба различного качества, в зависимости от качества ингредиентов, основным из которых является мука. Поэтому задача прогнозирования качества готовой продукции по информации о рецептурных компонентах весьма актуальна.

Исходная цель работы: определить взаимосвязь между показателями муки и качеством хлеба, выявить наиболее информативные признаки, построить алгоритм классификации данных.

В ходе выполнения работы была сформирована база данных, состоящая из 80 анализов, характеризующих качество белого хлеба по семи количественным признакам. Каждый анализ описывался органолептическими показателями качества муки (влажность, массовая доля и качество клейковины и т.д.) и показателями качествами хлеба (влажность, кислотность и пористость). В соответствии с классификацией, предложенной Пономаревой Е.И. данные были разделены на 4 группы. Первая группа (класс 1 высшего качества) – 20 наблюдений (25,0 %); вторая (класс 2 хорошего качества) – 14 (17,5 %); третья (класс 3 плохого качества) – 26 (32,5 %); четвертая (класс 4 очень плохого качества) – 20 (25,0 %).

Для принятия решений об отнесении хлеба к определенному классу необходимо отобрать наиболее информативные признаки.

Выявление наиболее информативных признаков осуществлялось в три этапа. На первом этапе использовался корреляционный анализ. На втором этапе формировалась классификационная система признаков методом двухэтапного кластерного анализа. На третьем этапе строилась дискриминантная функция.

Одним из методов определения типов сходства является коэффициент корреляции Пирсона, который рассчитывается:

$$r_{xy} = \frac{\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2}}, \qquad (1)$$

где x_i - значения, принимаемые переменной X, y_i - значения, принимаемые переменной Y, $\bar{x}$ - средняя по X, $\bar{y}$ - средняя по Y.

Кластерный анализ позволяет группировать «однородные» или «близкие» объекты в классы по какому-либо признаку. Наиболее распространены иерархическая кластеризация и k-средними. Недостатком иерархических методов кластеризации является то, что модель предлагает несколько вариантов разбиения или объединения данных в кластеры, выбор результирующей модели остается за человеком. Кластеризация k-средними, или «метод ближайшего соседа» основан на том, что число кластеров задается изначально. Затем элементы перераспределяются по кластерам, улучшая качество модели. Недостатком данного метода является то, что необходимо применять процедуры несколько раз для различного числа кластеров, не всегда разбиение оптимально для заданной задачи.

Модель двухэтапного подхода (TwoStep Cluster) позволяет кластеризовать различные группы по отдельности, а после этого объединять полученные результаты в конечную структуру кластеров. Двухэтапный кластерный анализ используется как основной инструмент для сокращения размерности данных при создании кластеров или подгрупп данных, более удобных для анализа. Последующий многомерный анализ выполняют над кластерами, а не над отдельными наблюдениями. Для измерения расстояния между объектами используется Евклидова метрика:

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

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

На первом этапе двухэтапного кластерного анализа рассчитывается межкластерная дисперсия, логарифмическая функция правдоподобия и первоначальное количество кластеров через критерии Акаике и Байеса.

Дисперсия ξ_i в кластерах *v=(i,s):*

$$\xi_i = -n_i\left(\sum_{j=1}^{p}\frac{1}{2}\log(\delta_{ij}^2 + \delta_j^2) - \sum_{j=1}^{q}\sum_{i=1}^{m_j}\pi_{ij}\log(\pi_{ij})\right) \qquad (3)$$

состоит из двух частей:

$-n_i\left(\sum_{j=1}^{p}\frac{1}{2}\log(\delta_{ij}^2 + \delta_j^2)\right)$ - мера дисперсии непрерывных переменных x_i в пределах кластера и $\sum_{j=1}^{q}\sum_{i=1}^{m_j}\pi_{ij}\log(\pi_{ij})$ мера дисперсии категориальных переменных. Кластеры с минимальным расстоянием *d*(*i*, *s*) будут объединены на каждом шаге итерации. Логарифмическая функция правдоподобия для шага с *k*-кластерами вычисляется по формуле:

$$1.\quad L_i = \sum_{i=1}^{k}\xi_i \qquad (4)$$

Число кластеров в двухэтапном кластером анализе может быть задано автоматически. Информационный критерий Акаике (AIC):

$$AIC_k = -2L_k + 2r_k\,, \qquad (5)$$

где r_k - число параметров или Информационный критерий Байеса

$$BIC_k = -2L_k + r_k \log n\,. \qquad (6)$$

Информационные критерии (5) и (6) определяют максимальное число кластеров.

На втором этапе кластерного анализа рассчитывается расстояние для *k*- кластеров:

$$R(k) = \frac{d_{k-1}}{d_k}\,, \qquad (7)$$

где d_{k-1} – расстояние, в котором кластер k слит с кластером (k−1). Минимальное расстояние между кластерами:

$$d_k = L_{k-1} - L_k \qquad (8)$$

В следующем этапе анализа был использован дискриминантный анализ. Данный метод заключается в разработке методов решения задач различения (дискриминации) объектов наблюдения по определенным признакам. Процедуры дискриминантного анализа позволяют не только интерпретировать различия между существующими классами, но и проводить классификацию новых объектов в тех случаях, когда заранее неизвестно, к какому из существующих классов они принадлежат. Методы пошагового дискриминантного анализа предполагают проверку (в начале каждого шага) всех дискриминантных переменных на соответствие двум условиям: необходимой точности вычисления (толерантности) и превышению заданного уровня различения (использование статистик *F*-ввода и *F*-исключения). Статистика *F*-ввода оценивает улучшение различения благодаря использованию данной переменной по сравнению с различением, достигнутым с помощью отобранных переменных. Статистика *F*-исключения определяет значимость ухудшения различения после удаления переменной из списка уже отобранных. Переменная с наибольшим значением *F*-исключения дает наибольший вклад в различение.

Каноническая дискриминантная функция вычисляется по формуле:

$$F(x) = a_1x_1 + a_2x_2\,, \qquad (9)$$

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:

$$\overline{f_1}(x) - \overline{f_2}(x) = \sum_{i=1}^{n_1} a_1 x_{1i} - \sum_{i=1}^{n} a_1 x_{2i}, \quad (10)$$

Вектор коэффициентов дискриминантной функции определяется по формуле:

$$A = S_*^{-1}(\overline{X_1} - \overline{X_2}), \quad (11)$$

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

$$S_* = \frac{1}{n_1 + n_2 - 2}(X_1^{/} X_1 + X_2^{/} X_2), \quad (12)$$

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

Константа детерминации для классификации рассчитывается по формуле:

$$C = \frac{1}{2}(\overline{f_1} + \overline{f_2}), \quad (13)$$

С помощью корреляционного анализа в общей выборке было установлено, что признаки коррелируют на уровне значимости 0,05 с классом качества хлеба. Информативных признаков, коэффициент корреляции для которых превышает 0,7, выявлено не было. На основании проведенного корреляционного анализа можно утверждать, что выделение классов 1, 2, 3, 4 в общей выборке невозможно, из-за отсутствия специфических признаков в классах.

Кластерный анализ, проведенный на основе 7 признаков показал следующие результаты. Для получения приемлемых результатов классификации необходимо построение иерархической схемы, показанной в работе [3].

Для повышения точности классификации исходный набор количественных признаков был преобразован в категориальные, так как для описания класса важнее не само значение признака, а попадание этого значения в категорию (диапазон значения от и до), определяющую принадлежность к классу качества.

Полученные категориальные признаки были преобразованы в бинарные, где каждый признак имел 2 состояния (0 – признак не принадлежит диапазону, 1 – принадлежит). В результате в базе данных количество признаков увеличилось с 7 до 37.

С помощью корреляционного анализа в общей выборке было установлено, что признаки коррелируют на уровне значимости 0,01 с классом качества хлеба. В качестве наиболее информативных были отобраны признаки с коэффициентом корреляции превышающем 0,7.

Т а б л и ц а 1

Таблица информативных признаков

Показатели	1 класс	2 класс	3 класс	4 класс
Массовая доля клейковины 32-33 (X_7)	0,814**	-0,075	-0,323**	-0,367**
Качество клейковины 66-75 (X_{11})	0,788**	-0,302**	-0,163	-0,315**
Качество клейковины 35-50 (X_{12})	-0,279*	-0,230*	0,721**	-0,289**
Газообразующая способность 1400-1500 (X_{17})	0,742**	-0,249*	-0,425**	-0,031
Кислотность мякиша 3 (X_{29})	0,764**	-0,441*	0,238*	0,389*
Пористость мякиша 67-68 (X_{34})	-0,218	0,806**	0,230*	-0,300**
Пористость мякиша 69-70 (X_{35})	0,965**	-0,248*	-0,374**	-0,311**
Пористость мякиша ниже 63 (X_{37})	-0,333**	-0,275*	-0,111	0,705**

Для класса 1 было выявлено 5 информативных признаков (массовая доля клейковины 32-33, качество клейковины 66-75, газообразующая способность 1400-1500, кислотность мякиша 3, пористость мякиша 69-70), для которых коэффициент корреляции превышает 0,7. Для 2 признаков r находится в диапазоне 0,624 до 0,684 по модулю, и имеет среднюю тесноту связи с классом качества. Во второй группе специфических признаков не обнаружено, значение r находится в диапазоне от 0,485 до 0,689. В группе 3 обнаружено 2 специфических признака с теснотой связи от 0,721 до 0,806. Для 6 признаков коэффициент корреляции находится в диапазоне от 0,586 до 0,664 с средней теснотой связи. Класс 4 имеет один специфический признак (пористость мякиша) со значением коэффициента корреляции более 0,7; теснота связи сильная. Для 4 признаков в данной группе коэффициент корреляции r находится в диапазоне от 0,525 до 0,656, теснота связи средняя (больше 0,5).

На основании проведенного корреляционного анализа можно утверждать, что возможно выделение 4-х классов.

С помощью двухэтапного кластерного анализа была получена четырехкластерная структура данных, представленная на рис. 1.

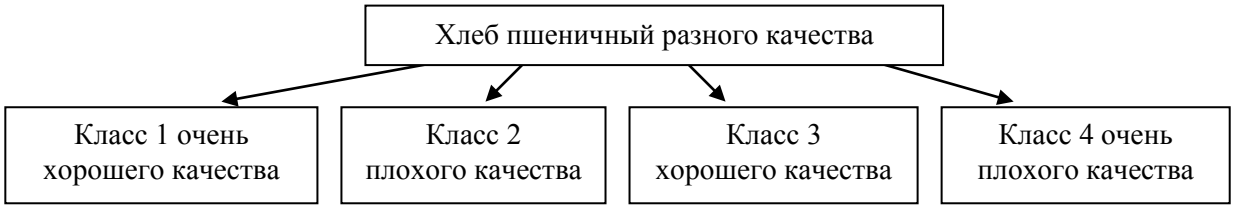

Рис. 1. Четырехкластерная структура данных.

К классу 1 (22,5 %) относится хлеб очень хорошего качества, к классу 2 (30,0 %) – хлеб плохого качества, 3 класс (21,3 %) – хлеб хорошего качества, класс 4 (26,3 %) – хлеб очень плохого качества. Было допущено 11 ошибок (13,75 %). Из них: 4 ошибки первого рода (5 %), класс плохого качества был ошибочно отнесен к классу хорошего качества; 2 ошибки второго рода (2,5 %); 5 ошибок попадания наблюдений плохого качества в очень плохое (6,25 %) не являются существенными для классификации, так как классы (3 и 4) не должны использоваться в хлебопечении. Следовательно, в классификации задан порог чувствительности выше необходимого. Результат классификации 86,25 %.

Результаты двухэтапного кластерного анализа представлены в табл. 2.

Таблица 2

Результат двухэтапного кластерного анализа

	Распределение по кластерам		% ошибок
	N	% объединенных	
1 класс очень хорошего качества	18	22,5 %	
2 класс плохого качества	24	30,0 %	2,5
3 класс хорошего качества	17	21,3 %	5
4 класс очень плохого качества	21	26,3 %	6,25
Объединенный	80	100,0 %	13,75

Были построены дискриминантные функции и оценена их значимость по коэффициенту Уилкса (λ):

$$D_1(X) = -3{,}994+2{,}206X_7+3{,}390X_{11}+0{,}413X_{12}+3{,}486X_{17}+1{,}496X_{29}+1{,}811X_{30}+2{,}739X_{31}; \quad (13)$$

$$D_2(X)=-0{,}649-2{,}619X_7+2{,}130X_{11}+1{,}987X_{12}+3{,}034X_{17}-2{,}341X_{29}-2{,}475X_{30}+4{,}353X_{31}; \quad (14)$$

$$D_3(X)=-1{,}148-2{,}754X_7+1{,}419X_{11}+0{,}670X_{12}+1{,}098X_{17}+1{,}111X_{29}+3{,}169X_{30}+0{,}967X_{31}; \quad (15)$$

Таблица 3

Результаты дискриминантного анализа

Функция	Собственное значение	% объясненной дисперсии	Каноническая корреляция	λ -Уилкса	Хи-квадрат
$D_1(X)$	13,745	75,4	0,965	0,007	366,024
$D_2(X)$	3,027	16,6	0,867	0,101	168,245
$D_3(X)$	1,450	8,0	0,769	0,408	65,864

По результатам дискриминантного анализа (таблица 3) было выявлено, что наибольший вклад в дискриминацию вносит функция $D_1(X)$. На основании внутригрупповой корреляции между дискриминантными переменными и дискриминантными функциями было выявлено, что наибольший вклад в дискриминацию вносят переменные качество клейковины 66 - 75 (X_7) 0,402*, газообразующая способность 1400 - 1500 (X_{11}) 0,385* и кислотность мякиша 3 0,362*.

Результаты классификации методом дискриминантного анализа показали, что высокая точность достигнута в первой, третей и четвертой группах (100 %). Менее точные результаты получены во второй группе (8,75 %), где 5 наблюдений были ошибочно отнесены к плохому качеству (6,25 %), 2 наблюдения (2,5 %), классифицированные в базе данных как хорошее качество, были неправильно распознаны как очень плохое качество.

Результаты классификации свидетельствуют о том, что для 91,25 % наблюдений классификация проведена корректно.

Подводя итоги работы, можно сделать выводы:

- был предложен трехэтапный анализ для отбора наиболее информативных признаков, где на первом этапе проводится корреляционный анализ, на втором – двухэтапный кластерный анализ, на третьем – дискриминантный анализ. Показано, что коэффициент корреляции между признаками и классом определяет точность классификации.

- была предложена категориальная структура. Показано, что структура базы данных влияет на классификацию.

- выполнена классификация качества хлеба. При использовании метода двухэтапного кластерного анализа было допущено 11 ошибок (13,75 %): 4 ошибки первого рода (5 %), класс плохого качества был ошибочно отнесен к классу хорошего качества; 2 ошибки второго рода (2,5 %); 5 ошибок попадания наблюдений плохого качества в очень плохое (6,25 %), не являются существенными для классификации, так как классы (3 и 4) не должны использоваться в хлебопечении. Метод дискриминантного анализа классифицирует с точностью 91,3 %. Было допущено 7 ошибок: 5 наблюдений были ошибочно отнесены к плохому качеству (6,25 %), 2 наблюдения (2,5 %), классифицированные в базе данных как хорошее качество были неправильно распознаны как очень плохое качество.

ЛИТЕРАТУРА

1 Бююль, А. SPSS: искусство обработки информации, анализ статистических данных и восстановление скрытых закономерностей [Текст] / А. Бююль, П. Цёфель. – СПб.: ООО «ДиаСофтЮП», 2002. – 608с.

2 Бессокирная, Г. П. Дискриминантный анализ для отбора информативных переменных [Текст] / Г. П. Бессокирная // Статистические методы и анализ данных. – 2003. – №16. – С. 25-26.

3 Балашова, Е. А. Классификация качества хлеба методом двухэтапного кластерного анализа [Текст] / Е. А. Балашова, В. К. Битюков, Е. А. Журавлева. – Сборник трудов конференции ММТТ-25. – Волгоград: ВолгГТУ, 2012. – С. 67-70.

4 Сидоренко, Е.А. Информационное описание и диагностика состояния иерархически организованных систем [Текст] / Е. А. Сидоренко. – Воронеж,2001. – с. 192.

5 Bacher, J. SPSS TwoStep Cluster – A First Evaluation [Text] / J. Bacher, K. Wenzig. – Nurnberg: Universitet Erlanger, 2004.

REFERENCES

1 Byuyul, A. SPSS: art of treatment of information, analysis of statistical data and renewal of the hidden conformities to law [Text] / And. Byuyul, P. of Cefel. – SPb.: LTD. «DiaSoftYUp», 2002. – 608 p.

2 Bessokirnaya, G.P. Discriminant analysis for the selection of informativevariables [Text]/ G.P. Bessokirnaya // the Statistical methods and analysis of data. – 2003. – №16. – P. 25-26.

3 Balashova, E.A. Classification of quality of bread by the method of a twostage cluster analysis [Text] / E.A. Balashova, V.K. Bityukov, E.A. Zhuravleva. it is Collection of labours of conference of MMTT-25. - Volgograd: VolgGTU, 2012. – P. 67-70.

4 Sidorenko, E.A. Informative description and diagnostics of the state of the hierarchically organized systems [Text] / E.A. Sidorenko. – Voronezh,2001. – p. 192.

5 Bacher, J. SPSS TwoStep Cluster – A First Evaluation [Text] / J. Bacher, K. Wenzig. – Nurnberg: Universitet Erlanger, 2004.