Дискриминантный анализ

Стукач Олег Владимирович

Каф. КИСМ, пр. Ленина, д. 2, оф. 204

☏(3822)-701777*2754

tomsk@ieee.org

http://ieee.tpu.ru/pages/stukach.htm

Основная задача дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик объекта классифицировать его, то есть отнести к некоторому классу оптимальным образом

Халафян А.А. Statistica 6. Статистический анализ данных. - 3-е издание. - М: Бином-Пресс, 2007. - 512 с. - ISBN 978-5-9518-0215-6

Стр. 199, гл. 12

Содержание

Цель

Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы). В дискриминантном анализе рассматривается некоторая «зависимая» переменная, определяющая наше мнение относительно предстоящей группировки. Далее определяются линейные классификационные модели, которые позволяет «предсказать» поведение новых элементов в терминах зависимой переменной на основании измерения ряда независимых переменных (факторов, показателей), которыми они характеризуются

Модель более информативна, чем, например, модель в кластерном анализе, так как дает «силу влияния»

Виды дискриминантного анализа

Пошаговый анализ с включением. В пошаговом анализе дискриминантных функций модель дискриминации строится по шагам. Точнее, на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

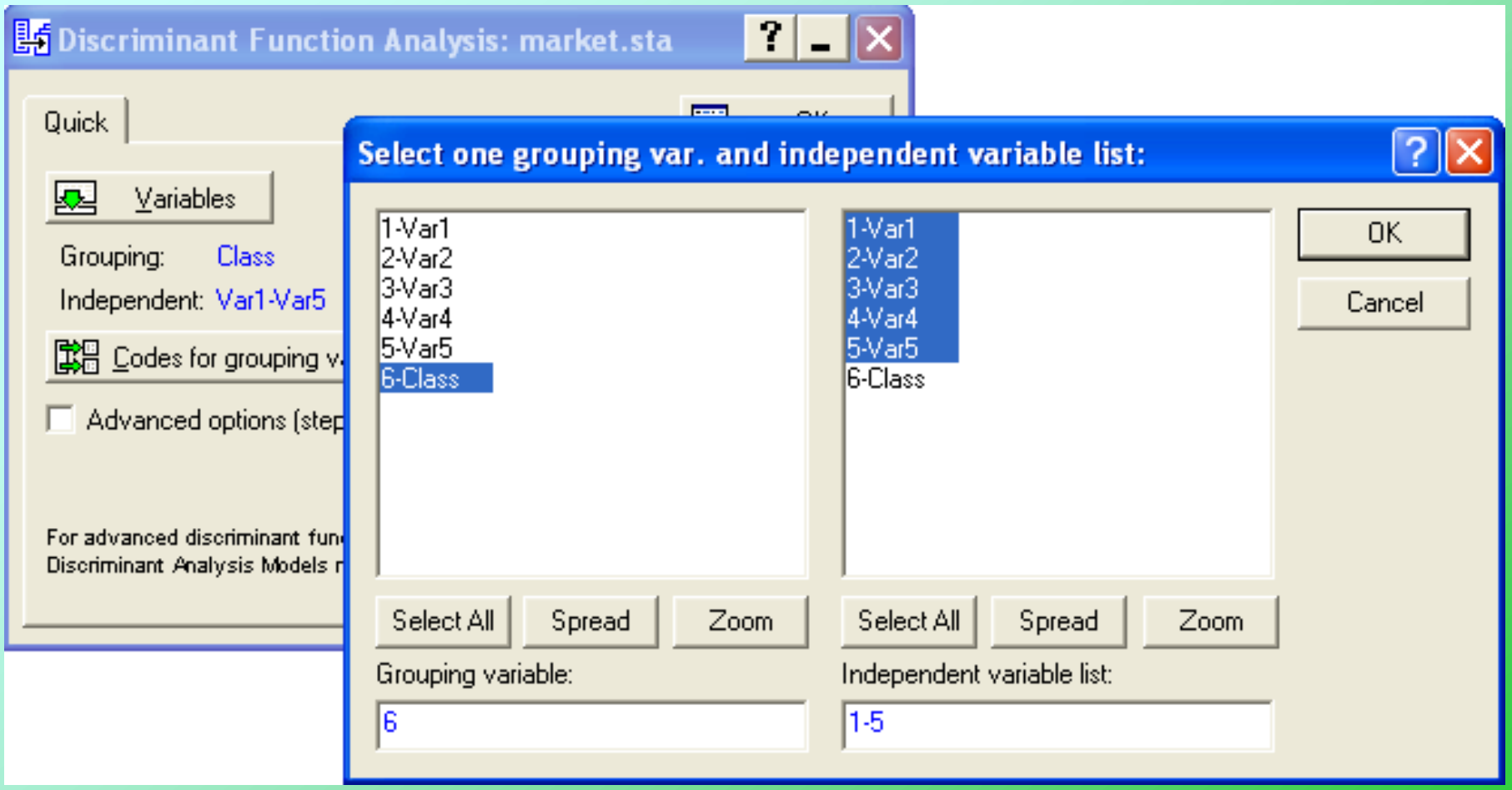
Пошаговый анализ с исключением. Можно также двигаться в обратном направлении. В этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания. Тогда в качестве результата успешного анализа можно сохранить только «важные» переменные в модели, то есть те переменные, чей вклад в дискриминацию больше остальных.

F для включения, F для исключения. Эта пошаговая процедура руководствуется соответствующим значением F для включения и соответствующим значением F для исключения. Значение F-статистики для переменной указывает на её статистическую значимость при дискриминации между совокупностями, то есть она является мерой вклада переменной в предсказание членства в совокупности.

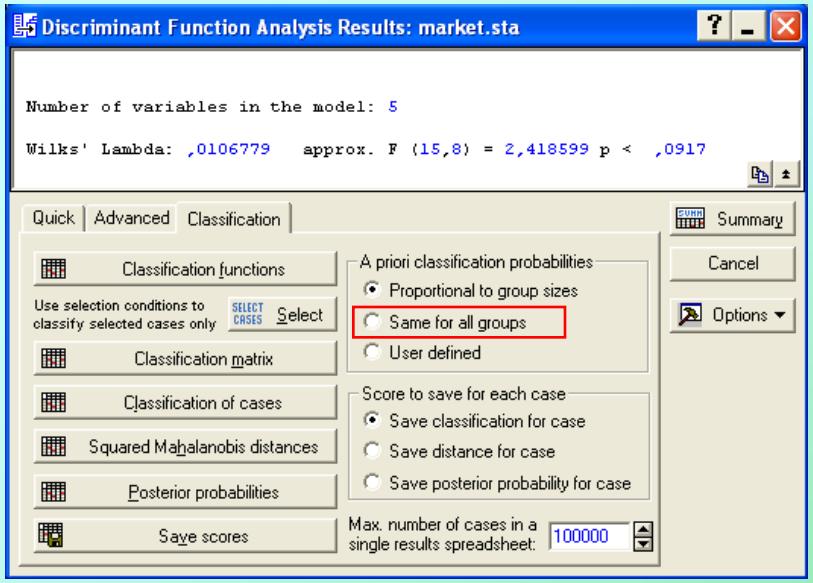
Данные

Объект	Var1	Var2	Var3	Var4	Var5	Class
1	5868	531	450	63	1608	1
2	6330	636	401	69	1768	1
3	4731	447	405	64	979	2
4	6793	620	487	104	1775	2
5	2902	161	182	22	631	3
6	3634	334	361	59	925	3
7	3499	204	129	27	398	3
8	6368	288	169	27	601	3
9	3058	169	86	23	307	4
10	5110	82	57	11	174	4
11	4166	207	183	32	487	4

Начало


Discriminant Function Analysis: market.sta
Quick
Variables
Grouping: Class
Independent: Var1-Var5
Codes for grouping v
Advanced options (step
For advanced discriminant fun
Discriminant Analysis Models r
Select one grouping var. and independent variable list:
1-Var1
2-Var2
3-Var3
4-Var4
5-Var5
6-Class
1-Var1
2-Var2
3-Var3
4-Var4
5-Var5
6-Class
OK
Cancel
Select All
Spread
Zoom
Select All
Spread
Zoom
Grouping variable:
6
Independent variable list:
1-5

Результаты анализа дискриминантных функций



Classification Matrix (market.sta)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
G_1:1	100,0000	2	0	0	0
G_2:2	100,0000	0	2	0	0
G_3:3	75,0000	0	0	3	1
G_4:4	100,0000	0	0	0	3
Total	90,9091	2	2	3	4

* Number of variables in the model (число переменных в модели) =5;
* Wilks lambda (значение лямбды Уилкса) = 0,0106779;
* Approx. F(15,8) (приближенное значение F-статистики, связанной с лямбда Уилкса) = 2,418599;
* p-уровень значимости F-критерия меньше 0,0917.

Значение статистики Уилкса всегда находится в интервале [0,1]. Значения статистики Уилкса, лежащие около нуля, свидетельствуют о хорошей дискриминации, а значения, лежащие около единицы, свидетельствуют о плохой дискриминации.

По данным показателя Wilks lambda и по значению F-критерия можно сделать вывод, что данная классификация практически корректна.

Результаты анализа дискриминантных функций

Из классификационной матрицы можно сделать вывод, что объекты в классах 1, 2, 4 были правильно отнесены экспертным способом к выделенным группам. В классе 3 есть объекты, неправильно отнесённые к соответствующим группам. Их можно посмотреть, нажав в окне на рис. 2 кнопку Classification of cases (классификация наблюдений). В таблице классификации наблюдений некорректно отнесенные предприятия помечаются звёздочкой (*).

Новые наблюдения

С помощью дискриминантных функций можно будет в дальнейшем классифицировать новые наблюдения. Новые наблюдения будут относиться к тому классу, для которого классифицированное значение будет максимальным. Выбор метода окончательной классификации зависит от количества новых объектов, подлежащих классификации. Если количество новых наблюдений невелико, можно применить метод, основанный на статистических критериях. Если же количество новых наблюдений велико, то рациональнее по обучающим выборкам получить классификационные функции, получить формулы и провести окончательную классификацию.

Повторная классификация

На основе полученных обучающих выборок можно проводить повторную классификацию тех объектов, которые не попали в обучающие выборки, и любых других объектов, подлежащих группировке. Для решения данной задачи существуют два варианта: первый – провести классификацию на основе дискриминантных функций, второй – на основе классификационных функций

В первом случае необходимо, не закрывая диалогового окна Discriminant Function Analisis Results, добавить в таблицу исходных скорректированных данных новые наблюдения. Для того чтобы понять, к какому классу относится этот объект, нажмите кнопку Posterior probabilities (Апостериорные вероятности). Появится таблица с апостериорными вероятностями. К тем классам, которые будут иметь максимальные вероятности, можно отнести новые наблюдения.

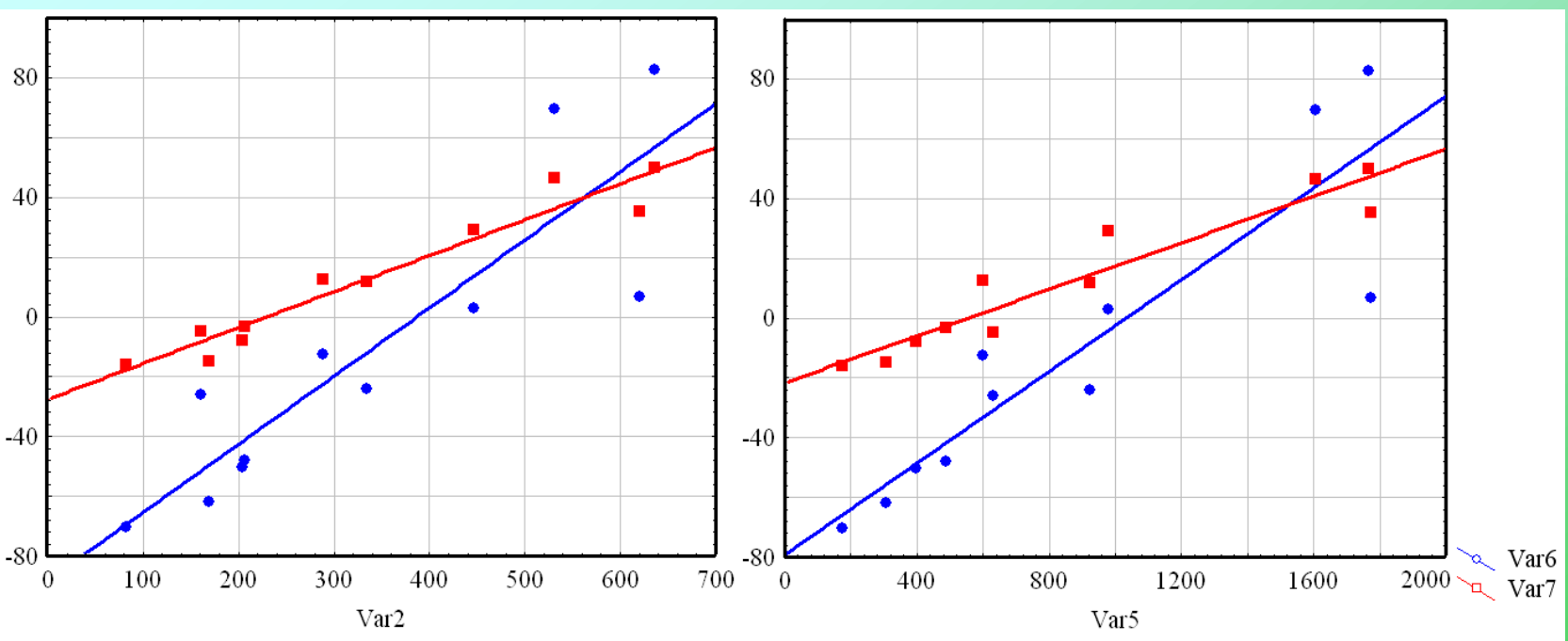
Во втором варианте необходимо в диалоговом окне Discriminant Function Analisis Results нажать кнопку Classification functions. Появится окно, из которого можно выписать дискриминантные функции для каждого класса

Variable	Classification Functions; grouping: Class (market.sta)			
	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
Var1	-0,0012	0,0021	0,0020	0,00317
Var2	0,2674	0,1179	0,0591	-0,00296
Var3	0,1772	0,1128	0,0764	0,03394
Var4	-2,4762	-0,6738	-0,5127	0,09802
Var5	0,0558	-0,0011	0,0030	-0,01184
Constant	-78,7186	-35,0491	-13,0743	-8,69528

Например, для первых двух классов функции имеют вид:

Ф1=-78,7186-0,0012*Var1+0,2674*Var2+0,1772*Var3-2,4762*Var4+0,0558*Var5
Ф2=-35,0491+0,0021*Var1+0,1179*Var2+0,1128*Var3-0,6738*Var4-0,0011*Var5

Дискриминантные функции для каждого класса



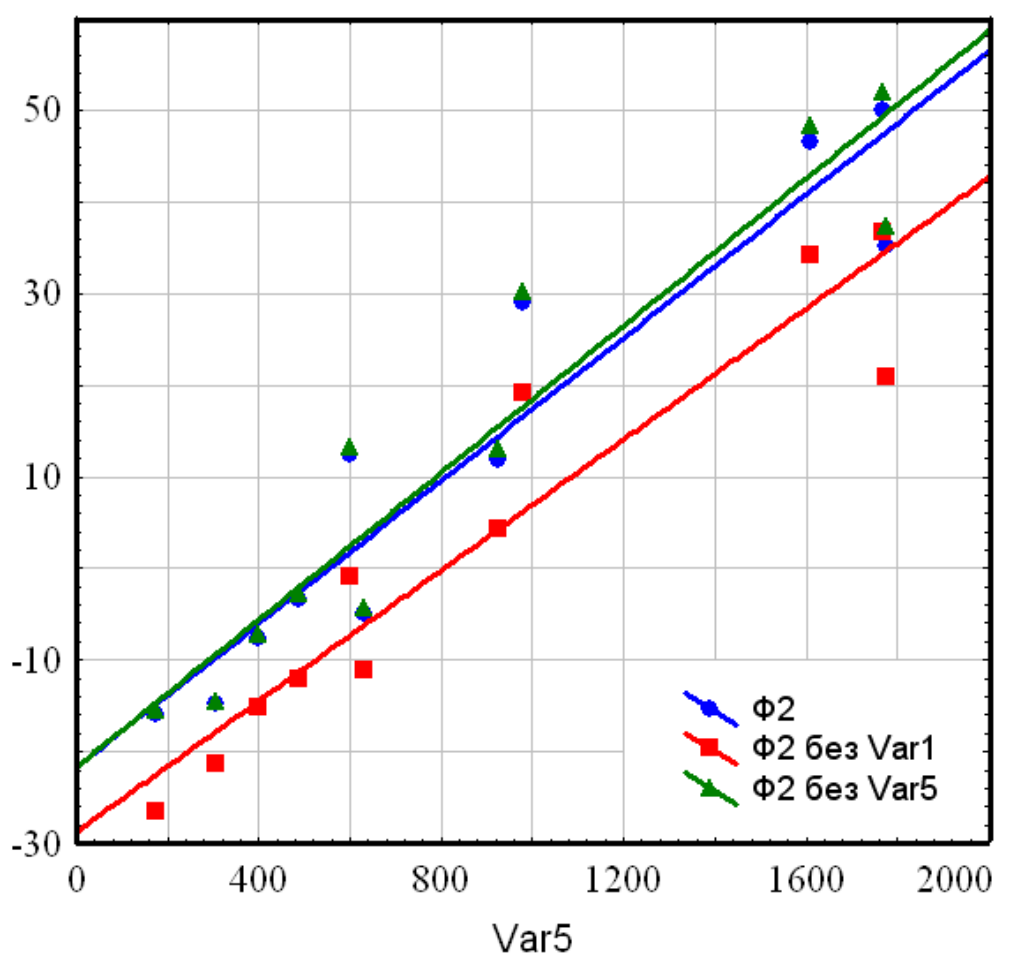
Ф1=-78,7186-0,0012*Var1+0,2674*Var2+0,1772*Var3-2,4762*Var4+0,0558*Var5
Ф2=-35,0491+0,0021*Var1+0,1179*Var2+0,1128*Var3-0,6738*Var4-0,0011*Var5

Для построения графиков значения функций Ф1 и Ф2 рассчитаны по формулам и записаны в переменные Var 6 и Var 7 соответственно. Затем в модуле Statistics/ Scatterplots построены графики рассеяния Var 6 и Var 7 от переменных Var 2 и Var 5 (можно выбрать и остальные переменные). Как видим, классы 1 и 2 хорошо различаются.

Значимость дискриминантной функции

Ф1=-78,7186-0,0012*Var1+0,2674*Var2+0,1772*Var3-2,4762*Var4+0,0558*Var5

Ф2=-35,0491+0,0021*Var1+0,1179*Var2+0,1128*Var3-0,6738*Var4-0,0011*Var5



Результаты дискриминантного анализа

Результаты классификации наблюдений можно вывести в терминах расстояний Махаланобиса, апостериорных вероятностей и собственно результатов классификации, а значения дискриминантной функции для отдельных наблюдений можно просмотреть на обзорных пиктографиках и других многомерных диаграммах, доступных непосредственно из таблиц результатов. Все эти данные можно автоматически добавить в текущий файл данных для дальнейшего анализа. Можно вывести также итоговую матрицу классификации, где указано число и процент правильно классифицированных наблюдений. Имеются различные варианты задания априорных вероятностей принадлежности классам, а также условий отбора, позволяющих включать или исключать определенные наблюдения из процедуры классификации (например, чтобы затем проверить её качество на новой выборке).

Результаты

Основная идея дискриминантного анализа заключается в том, чтобы определить, отличаются ли совокупности по среднему значению какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе

Результаты

Задачи дискриминантного анализа часто встречаются в производственной практике. Допустим, что мы располагаем информацией о некотором числе бракованных деталей, дефект каждой из которых может быть следствием ряда разладок производственного процесса. На основе этой информации нужно найти функцию, позволяющую определить, какая разладка (несоблюдение температурного режима, качество сырья) вызвала причину конкретного дефекта

Результаты

Задачи второго типа связаны с предсказанием будущих событий на основании имеющихся данных. Примером может служить определение вероятности, с которой, если соответствующие предписания производственного были соблюдены, деталь окажется стандартной (с какой вероятностью покупатель купит продукт, если ... и т.д.).

Результаты

В целом, дискриминантный анализ – это очень полезный инструмент для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп, и для классификации наблюдений в различные группы.