**ЛЕКЦИЯ 8 ДИСКРИМИНАНТНЫЙ АНАЛИЗ**

*1 Понятие о дискриминантном анализе*

*2 Постановка задачи, методы решения, ограничения*

*3 Предположения и ограничения*

*4 Алгоритм дискриминантного анализа*

*5 Реализация дискриминантного анализа в системе STATISTICA*

**1 Понятие о дискриминантном анализе**

*Классификацией* называют разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы.

Различают:

- классификацию при наличии обучающих выборок (*дискриминантный анализ*);

- классификацию без обучения (автоматическая классификация – *кластерный анализ*).

Дискриминантный анализ является одним из методов многомерного статистического анализа. Цель дискриминантного анализа – на основе измерения различных характеристик (признаков, параметров) объекта классифицировать его, то есть отнести к одной из нескольких групп (классов) некоторым оптимальным способом. Под *оптимальным способом* понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. Этот вид анализа является многомерным, так как измеряется несколько параметров объекта, по крайней мере, больше одного, например, температура, влажность в технологическом процессе, давление, состав крови, температура больного и т.д.

Типичные области применения дискриминантного анализа – биология, медицина, управление производством, экономика, геология, контроль качества.

В медицине объектом исследования является пациент, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет, например, необходимо ли хирургическое вмешательство при лечении.

В управлении производством принимается решение по отнесению поступающего сырья или продукции к одному из нескольких типов.

В экономике важно решение по отнесению клиента к определенному классу при выдаче кредита.

Чрезвычайно интересно применение дискриминантного анализа в борьбе с терроризмом. Последствия действий против террористов могут привести к одному из двух результатов: успешное освобождение заложников и случаи, когда заложники пострадали. Здесь имеются следующие дискриминантные переменные: число террористов, степень поддержки, количество оружия и т.д.

Очевидно, лицам, ведущим переговоры, важно классифицировать возникшую ситуацию, с тем, чтобы определить свои действия.

Широкий круг задач, возникающих на практике и связанных с классификацией, можно решить методами дискриминантного анализа.

**2 Постановка задачи, методы решения, ограничения**

Предположим, имеется *n* объектов с *m* характеристиками. В результате измерений каждый объект характеризуется вектором x1 ... xm, *m*>1. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких групп (классов) G1,... Gk, k≥2. Иными словами, нужно построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Число групп заранее известно, также известно, что объект заведомо принадлежит к определенной группе.

Пусть X – пространство значений вектора измерений. Решающее правило называется *нерандомизированным*, если пространство X разбито на k непересекающихся областей; при попадании измерения параметров объекта в k–ю область объект относится к k–й группе.

Решающее правило называется *рандомизированным*, если для каждого вектора наблюдений х задана вероятность pi(x), с которой объект принадлежит *i*-й группе, pi(x)≥0; p1(x)+ ... +pk(x)=1; *i*=1,...k.

Очевидно, при использовании решающего правила возникают потери, вызванные тем, что объект неправильно классифицирован – отнесен к классу *i*, когда в действительности он принадлежит классу *j* (*i≠j*).

Если значение потерь трудно оценить численно, то при построении оптимального правила используют критерий минимальной вероятности ложной классификации.

В дискриминантном анализе можно задать априорные вероятности принадлежности объекта к определенному классу. На практике эти вероятности оцениваются из массива экспериментальных данных.

Так как массив экспериментальных данных накапливается, то эти оценки постепенно уточняются. При этом можно учесть различные факторы, влияющие на принадлежность объекта к определенному классу, например, если поступает мука в хлебное производство, то можно учесть сезонные факторы: вероятность того, что мука будет лучшего качества осенью выше той же вероятности весной.

В случае двух групп объектов дискриминантный анализ эквивалентен множественной регрессии (зависимой переменной является номер группы).

Независимые переменные с наибольшими стандартизированными коэффициентами регрессии дают наибольший вклад в предсказание принадлежности объекта к группе.

Для практических целей реализовано два общих метода дискриминантного анализа: *стандартный* и *пошаговый (включения и исключения)*. Данные методы дискриминантного анализа аналогичны методам множественной регрессии.

При **стандартом методе** в случае двух групп путём наименьших квадратов строится регрессионная прямая (зависимая переменная – номер группы, все остальные переменные – независимые). Если групп несколько, то можно представить себе, что вначале строится дискриминация между группами 1 и 2, затем между 2 и 3, и так далее.

В **пошаговом методе** модель строится последовательно по шагам. Для *метода включения* на каждом шаге оценивает вклад в функцию дискриминации не включенных в модель переменных. Переменная, дающая наибольший вклад, включается в модель, далее система переходит к следующему шагу. Если применяется так называемый *пошаговый метод исключения*, то вначале в модель включаются все переменные, затем производится их последовательное исключение.

Близкими к методам дискриминантного анализа являются методы дисперсионного анализа, кластерного и факторного анализов, а также, как уже говорилось, методы множественной регрессии. Отличие кластерного анализа от дискриминантного – в нем заранее не фиксировано число групп (кластеров).

**3 Предположения и ограничения**

Дискриминантный анализ «работает» при выполнении ряда предположений.

1) **Предположение о том, что наблюдаемые величины** (измеряемые характеристики объекта) **имеют нормальное распределение**. Это предположение следует проверять. Следует заметить, что умеренные отклонения от этого предположения не являются фатальными.

2) **Предположение об однородности дисперсий** наблюдаемых переменных в разных классах (отличие между классами имеется только в средних). Умеренные отклонения от этого предположения также допустимы.

Методы, реализуемые в практических расчетах, являются линейными. Функции классификации и дискриминантные функции являются линейными комбинациями наблюдаемых величин.

Необходимо сделать важное замечание о проверке предположений анализа. **Дискриминантный анализ может быть проведен и когда основные предположения не выполняются** (предположение о нормальности и однородности дисперсий). Задача состоит в интерпретации результатов. В конечном счете, наиболее важным критерием правильности построенного классификатора является практика. И если окажется, что в результате построен классификатор, «работающий» на практике, то это будет достижением.

**4 Алгоритм дискриминантного анализа**

Решение задач дискриминации (дискриминантный анализ) состоит в разбиении всего выборочного пространства (множества реализации всех рассматриваемых многомерных случайных величин) на некоторое число областей.

Пусть имеются две генеральные совокупности X и Y, имеющие многомерный (трехмерный) нормальный закон распределения с неизвестными, но равными ковариационными матрицами.

Из этих совокупностей взяты обучающие выборки объемами n1 и n2 соответственно **(!)**:

; 

Целью дискриминантного анализа в этом случае является отнесение нового наблюдения (строки) из матрицы:



либо к X, либо к Y.

Для решения задачи по обучающим выборкам проводятся оценки векторов средних и ковариационных матриц **(!)**

; 

Затем определяется граница дискриминации – константа *С*.

Оценку дискриминантной функции Ui для *i*–й строки матрицы Z, которая характеризует *i*-e наблюдение, подлежащее дискриминации, получается из уравнения **(!)**:



Если Ui ≥ C, то *i*–e наблюдение следует отнести к совокупности X, если же Ui < C, то *i*–e наблюдение относится к совокупности Y.

Дискриминантный анализ допускает наличие более двух обучающих выборок, однако в этом случае задача существенно усложняется и не всегда приводит к однозначной дискриминации, т.е. не все объекты удается отнести к какому-либо классу.

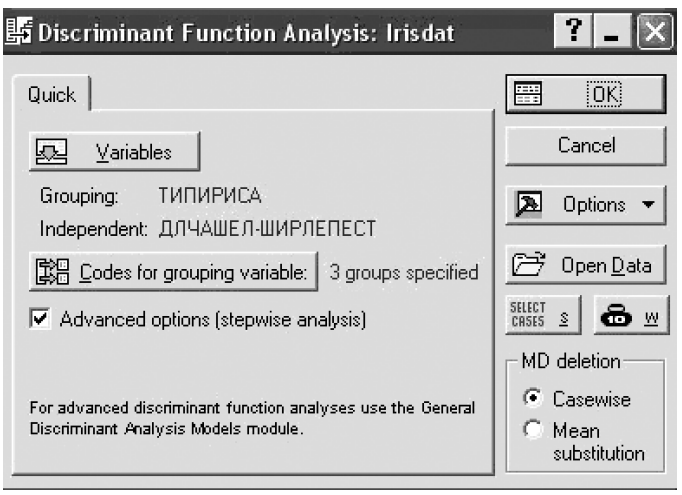
**5 Реализация дискриминантного анализа в системе STATISTICA (!)**

Знакомство с возможностями проведения дискриминантного анализа в системе STATISTICA лучше всего начать с разбора апробированного примера. Таким примером является классический **пример Фишера** – анализ цветков ириса.

Задача состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветков ириса отнести ирис к одному из трех типов: SETOSA, VERSICOL, VIRGINIC.

Данные для этого примера имеются в файле Irisdat.sta. В файле содержатся результаты измерений 150 цветков ириса, по 50 каждого типа.

**Шаг 1.** Из **Переключателя модулей** STATISTICA откройте стартовую панель модуля **Discriminate function analysis** **(дискриминантный функциональный анализ)** (рис. 1).



*Рис. 1 –Стартовая панель модуля* ***Дискриминантный анализ***

**Шаг 2**. Нажмите кнопку **Open Data (открыть данные)** и откройте файл данных *Irisdat.sta* из каталога **Examples (примеры)**. Следующий файл данных появится на экране (рис. 2).

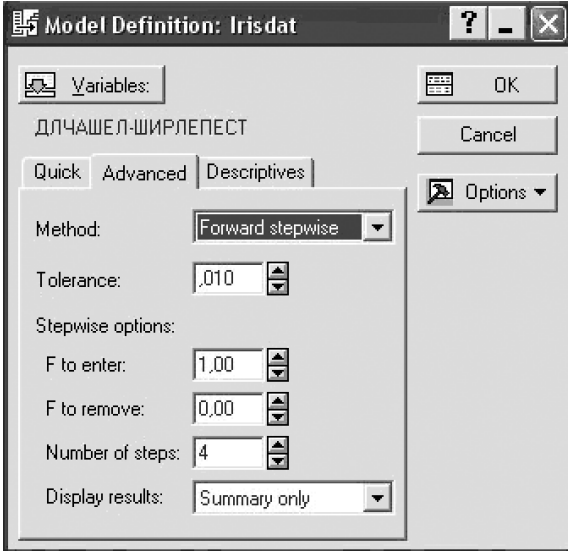


*Рис. 2 – Файл данных Iris.sta*

**Шаг 3.** Нажмите кнопку **Variables (переменные)** и выберите переменные для анализа. В качестве **Grouping variable (группирующая переменная)** выберите переменную **IRISTYPE (тип ириса)**. В качестве **Independent variables (независимые переменные)** выберите переменные **ДЛЧАШЕЛ**, **ШИРЧАШЕЛ**, **ДЛЛЕПЕСТ**, **ШИРЛЕПЕСТ**.

Сделайте установки, как показано на рис. 1.

**Шаг 4**. Нажмите кнопку **ОK** и откройте диалоговое окно **Model Definition (определение модели)** – рис. 3.



*Рис. 3 – Окно определения модели дискриминантного анализа*

Сделайте установки, как показано на рис. 3. Нажмите кнопку **OK** и запустите вычислительную процедуру, реализующую пошаговый метод включения.

**Шаг 5**. Всесторонне просмотрите итоги в диалоговом окне **Discriminant Function Analysis Results (результаты дискриминантного анализа)** (рис. 4).

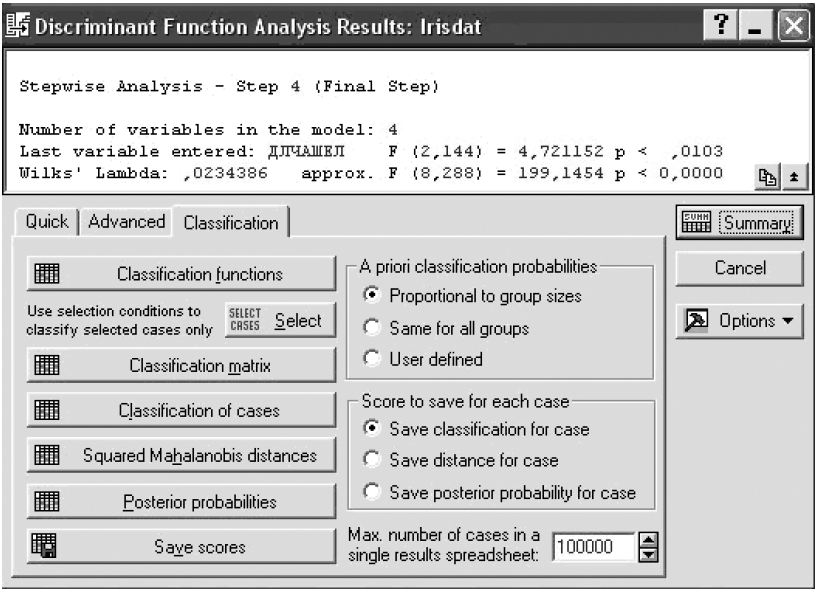


Рис. 4 – Окно результатов дискриминантного анализа данных из файла Iris.sta

Информационная часть окна сообщает, что использован:

- **Stepwise analysis (пошаговый анализ), Step 4 Final step (шаг 4 заключительный шаг)**;

- **Number of variables in the model (число переменных в модели): 4**;

- **Last variable entered (последняя включенная переменная): ДЛЧАШЕЛ**, соответствующее значение статистики **F-критерия F(2, 144) = 4,72**, уровень значимости **р < 0,0103**;

- **Wilks lambda (значение лямбды Уилкса): 0,0234**;

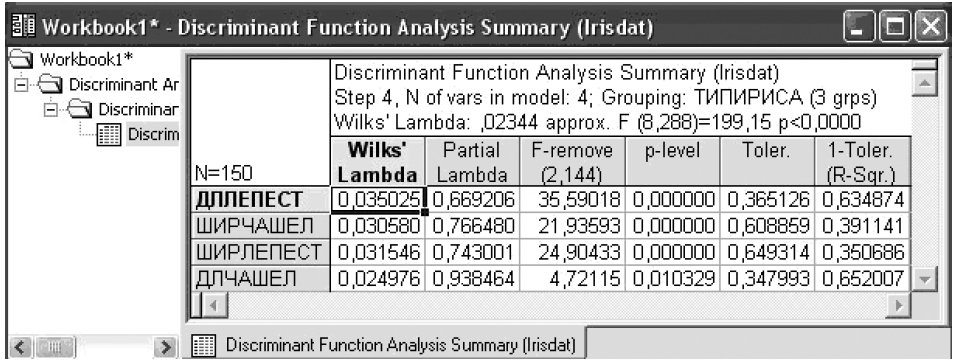
- **approx. F (4,292) = 199,1454 (приближенное значение F- статистики)**, связанной с **лямбдой Уилкса**;

- **р** – уровень значимости **F-критерия** для значения **199,1454**;

- значения статистики **лямбда Уилкса** лежат в интервале [0, 1].

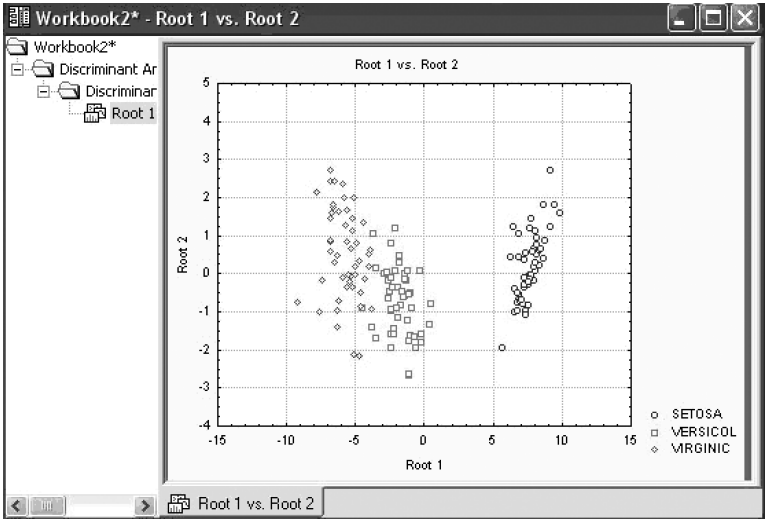
Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации. Иными словами, это можно выразить следующим образом: если значения лямбды Уилкса близки к 0, то мощность дискриминации (мощность = 1– вероятность ошибки) близка к 1, если лямбда Уилкса близка к 1, то мощность близка к 0.

**Шаг 6**. Нажмите кнопку **Variables in the model (переменные, включенные в модель).** На экране появится итоговая таблица анализа (рис. 5).



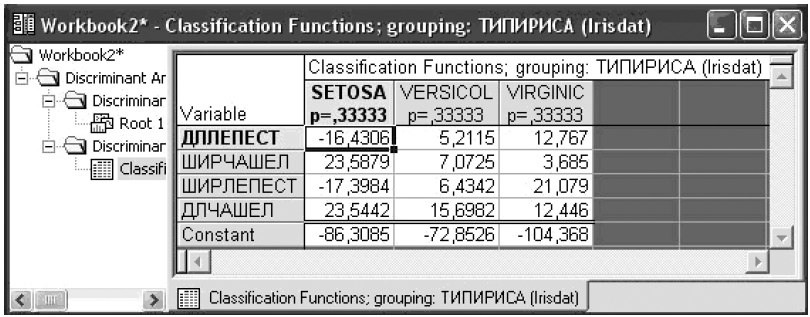
*Рис. 5 – Итоговая таблица анализа данных из файла iris.sta*

**Шаг 7**. Просмотрите разделение групп на графике. Для этого инициируйте кнопку **Canonical analysis & graphs (канонический анализ и графики)**. В появившемся диалоговом окне **Canonical Analysis (канонический анализ)** нажмите кнопку **Scatterplot of canonical scores (диаграмма рассеяния канонических значений)**. На экране появится следующий график (рис. 6).



*Рис. 6 – Разделение трех типов ириса*

**Шаг 8**. Просмотрите функции классификации. В диалоговом окне **Результаты дискриминантного анализа** нажмите кнопку **Classification functions (функции классификации)** (рис. 7).



*Рис. 7 – Функции классификации, построенные пошаговым*

*методом вперед (Forward stepwise)*

С помощью этих функций можно вычислить классификационные значения (метки) для вновь наблюдаемых цветков по формулам:

**SETOSA = – 16,43\*ДЛ + 23,69\*ШЧ – 17,4\*ШЛ + 23,54\*ДЧ – 86,31;**

**VERSICOL = 5,21\*ДЛ + 7,07\*ШЧ – 6,43\*ШЛ + 15,70\*ДЧ – 72,85;**

**VIRGINIC = 12,76\*ДЛ + 3,69\*ШЧ – 21,08\*ШЛ + 12,5\*ДЧ – 104,37,**

где:

**- ДЛ – ДЛЛЕПЕСТ;**

**- ШЧ – ШИРЧАШЕЛ;**

**- ШЛ – ШИРЛЕПЕСТ;**

**- ДЧ – ДЛЧАШЕЛ.**

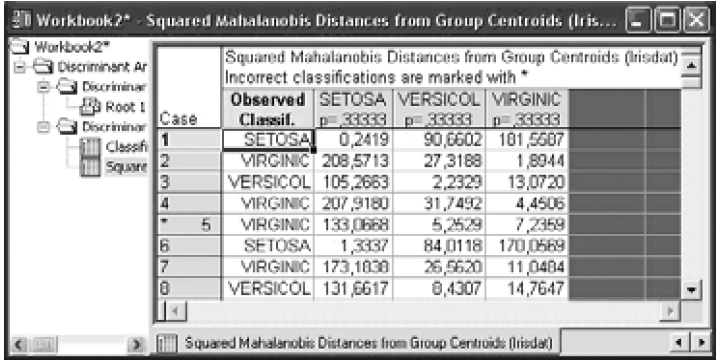
Пусть вы имеете новый цветок со значениями: ДЛЛЕПЕСТ, ШИРЧАШЕЛ, ШИРЛЕПЕС, ДЛЧАШЕЛИ.

К какому типу ириса его отнести? Формально следует подставить эти значения в приведенные выше формулы и вычислить классификационные значения SETOSA, VERSICOL, VIRGINIC.

Новый цветок относится к тому классу, для которого классификационное значение максимально. Конечно, построенные классификационные функции могут быть определены в электронных таблицах как формулы, и для каждого добавленного случая по ним могут быть вычислены классификационные метки. Таким образом, каждый новый объект автоматически относится к определенному классу.

**Шаг 9.** **Расстояния Махаланобиса**

Нажмите кнопку **Squared Mahalanobis distance (квадрат расстояния Махаланобиса)** и вы увидите таблицу с квадратами расстояния Махаланобиса от точек (случаев) до центров групп (рис. 8).

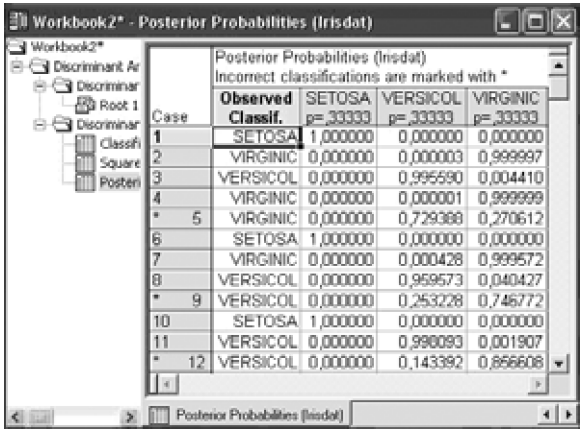


*Рис. 8 – Расстояния Махаланобиса для данных из файла Iris.sta*

Случай относится к группе, до которой расстояние Махаланобиса минимально.

**Шаг 10**. **Апостериорные вероятности**.

Рассмотрите группу опций справа внизу диалогового окна **Результаты дискриминантного анализа: A priori classifications probabilities (априорные вероятности классификации)**. До анализа вы задаете для каждого случая (в данном примере цветка) вероятность, с какой он принадлежит к определенному классу. После того как анализ выполнен, можно пересчитать эти вероятности и получить апостериорные вероятности классификации. Нажав кнопку **Posterior probabilities (апостериорные вероятности)**, вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу (рис. 9).



*Рисунок 11.9 – Таблица апостериорных вероятностей*

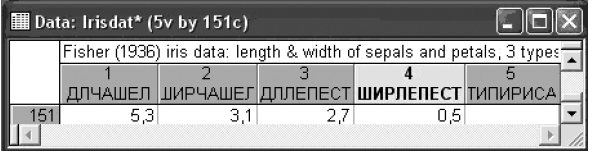
Интерпретация данной таблицы очень проста. В первом столбце указан тип ириса для каждого случая. Во втором, третьем, четвертом столбцах даны апостериорные вероятности отнесения каждого цветка к определенному типу.

Цветок относится к группе с максимальной апостериорной вероятностью.

Знаком \* отмечаются неправильно классифицированные при использовании данного правила случаи (5, 9, 12).

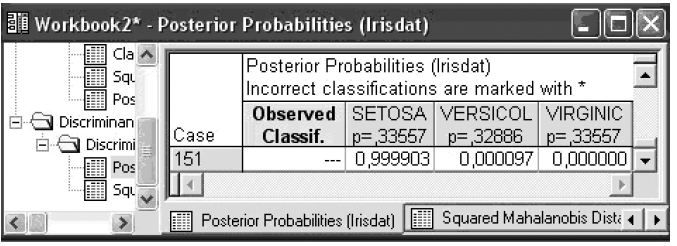
**Шаг 11.** **Классификация новых случаев**.

Не закрывая диалога Результаты дискриминантного анализа, добавьте в таблицу исходных данных новый случай (рис. 10).



*Рисунок 11.10 – Новое наблюдение в данных Iris.sta*

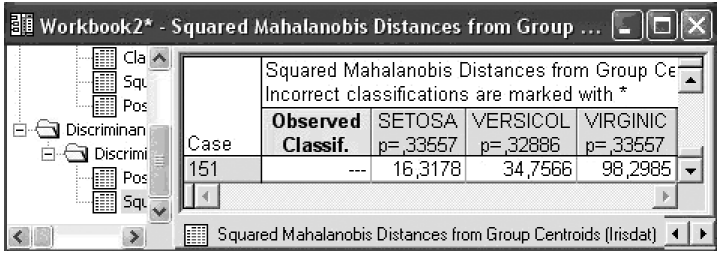
Для того чтобы понять, к какому классу относится этот объект, нажмите кнопку **Posterior probabilities (апостериорные вероятности)**, вы увидите ту же таблицу с постериорными вероятностями, к которой будет добавлена строка (рис. 11).



*Рис. 11 – Классификация нового наблюдения*

Итак, новое наблюдение с вероятностью 0,999 можно отнести к типу SETOSA.

Нажмите кнопку **Squared Mahalanobis distance (квадрат расстояния Махаланобиса)**, и вы увидите таблицу (рис. 12) с квадратами расстояния Махаланобиса. В последней строке этой таблицы показаны расстояния нового случая до групповых центров.



*Рис. 12 – Расстояние Махаланобиса от нового наблюдения до центров групп*

Опять расстояние от нового наблюдения до центра групп минимально для группы SETOSA. Следовательно, с высокой степенью вероятности новый цветок – это ирис типа SETOSA.