Записки Ленинградского ордена Ленина и ордена Трудового Красного Знамени горного института им. Г. В. Плеханова, т. LXIV, вып. 2, 1974 г.

ОБ ОДНОЙ ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ ДИСКРИМИНАНТНОГО АНАЛИЗА В ГЕОЛОГИЧЕСКОЙ ПРАКТИКЕ

Э. Л. Нарымский

Задача дискриминантного анализа заключается в отнесении объекта, характеризуемого набором l признаков к одной из (в нашем случае двух) l-мерных совокупностей (A или B). Выборки из этих совокупностей назовем классом A или B соответственно. Обозначим через m_1 число эталонных образцов класса A; m_2 — число эталонных образцов класса B; n — число исследуемых образцов (объектов), относительно которых решается сформулированная выше задача дискриминантного анализа; l—число признаков, характеризующих все образцы — как эталонные, так и исследуемые; A_{ik}—i-й признак к-го эталонного образца класса $A\,(0<i\leqslant l,\ 0<\kappa\leqslant m_1)$; B_{ip} — i-й признак p-го эталонного образца класса B $(0<p\leqslant m_2)$; C_{is} — i-й признак s-го исследуемого образца $(0<s\leqslant n)$.

Коэффициенты линейной дискриминантной функции

$$F(C_s)=\lambda_1 C_{1s}+\lambda_2 C_{2s}+\ldots+\lambda_l C_{ls}=\sum_{i=1}^{l}\lambda_i C_{is} \qquad (1)$$

находим из системы линейных уравнений

$$\sum_{i=1}^{l}\lambda_i\delta_{ir}=\Delta_i(m_1+m_2-2),\ r=1,\ 2,\ \ldots,\ l, \qquad (2)$$

где δ_{ir} — сумма ковариаций i-го и r-го признаков в совокупностях A и B, умноженных, соответственно, на m_1 и m_2

$$\delta_{i,r}=\sum_{k=1}^{m_1}A_{ik}A_{rk}-\frac{\sum\limits_{k=1}^{m_1}A_{ik}\sum\limits_{k=i}^{m_1}A_{rk}}{m_1}+\sum_{p=1}^{m_2}B_{ip}B_{rp}-\frac{\sum\limits_{k=1}^{m_2}B_{ip}\sum\limits_{p=1}^{m_2}B_{rp}}{m_2}; \qquad (3)$$

Δ_i — разность средних значений i-го признака,

$$\Delta_i\ \frac{\sum\limits_{k=1}^{m_1}A_{ik}}{m_1}-\frac{\sum\limits_{p=1}^{m_2}B_{ip}}{m_2},$$

F_0 — порог,

$$F_0 = \sum_{i=1}^{l} \lambda_2 \frac{\sum_{k=1}^{m_1} A_{ik} + \sum_{p=1}^{m_2} B_{ip}}{m_1 + m_2}. \quad (4)$$

При $F(C_s) > F_0$ образец C_s следует отнести к классу A, при $F(C_s) < F_0$ — к классу B. При нормальном распределении матриц порог, рассчитанный по формуле (4), статистически наилучший [3]. В общем же случае порог устанавливается обычно эмпирически (подбором величины, при которой дискриминантная функция обеспечивает наилучшее, с точки зрения исследователя, разделение эталонных образцов).

Важное значение для дискриминантного анализа имеет величина

$$D^2 = \sum_{i=1}^{l} \lambda_i \Delta_i, \quad (5)$$

называемая обобщенным расстоянием Махаланобиса (при сделанных выше допущениях о нормальности распределения и равенстве ковариационных матриц). Легко заметить, что каждое слагаемое $\lambda_i\Delta_i$ в сумме (5) является как бы вкладом i-го признака в обобщенное расстояние. И. Девис и Р. Семпсон предложили нормировать эту величину, выражая ее в процентах, и назвать вкладом i-го признака

$$V_i = \frac{\lambda_i \Delta_i}{D^2} 100\%. \quad (6)$$

При отборе наиболее ценных признаков (в смысле их разделяющей способности) многие исследователи так или иначе руководствуются величиной вклада признака. Казалось бы, чем больше вклад признака, тем этот признак ценнее. Однако после исключения i-го признака величина D^2 изменяется на величину, не равную $\lambda_i\Delta_i$, так как переход от системы l уравнений (2) к аналогичной системе $l-1$ уравнений приводит к новым коэффициентам λ_j^1 ($j=1, 2, ..., i=1, i+1, ..., l$), вследствие чего обобщенное расстояние по новому набору признаков

$$D^2_{(i)} = \sum_{j=1}^{l-1} \lambda_j' \Delta_j.$$

Если бы признаки были бы независимыми, тогда действительно их ценность была бы тем выше, чем больше величина V_i. Но, с другой стороны, в такой ситуации нет необходимости прибегать к помощи дискриминатного анализа, так как тогда задача может быть решена в рамках общей кибернетической проблемы автоматического распознавания образов (применение методов детерминистской логики, вероятностной логики, метода фазового анализа и т. д.). Но от перечисленных методов дискриминантный анализ отличается более полным учетом коррелированности признаков, что сильно повышает разрешающую способность диагностической функции и тем уменьшает процент неверных отнесений объектов к соответствующим классам [2]. Несмотря на это, корреляция (особенно высокая) может существенно затруднить построение решающего правила.

При решении задач геологической практики почти во всех результатах счета корреляция создает отрицательные значения вкладов V_i.

Нельзя считать, что признаки, дающие отрицательные вклады, не несут никакой полезной информации (относительно интересующего нас отличия одних образцов от других). Дело осложняется еще и тем, что после исключения из набора l признаков наименее ценного из них относительная ценность признаков в оставшемся наборе l—1 признаков оказывается уже иной.

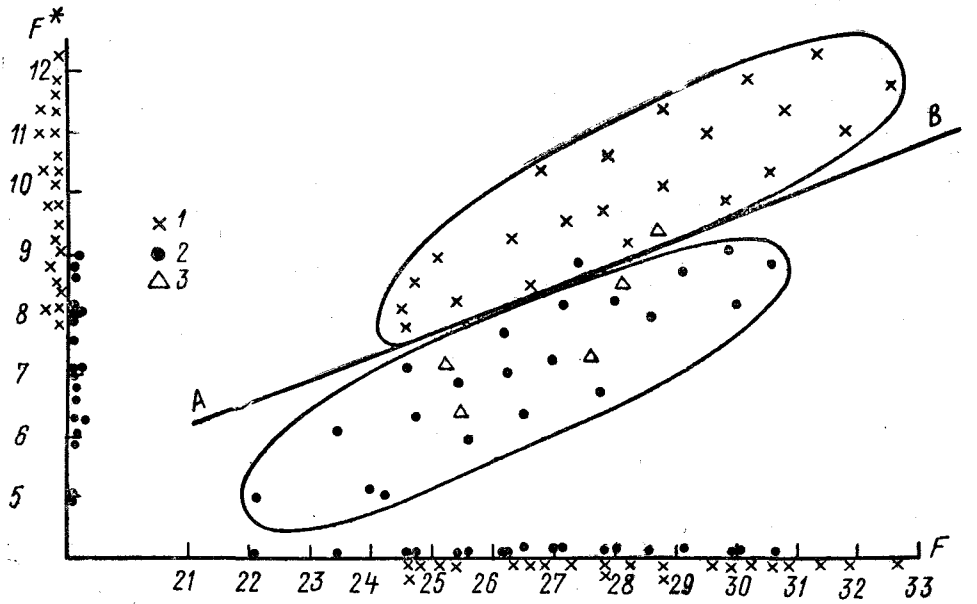
По описанному алгоритму в ВИРГе в 1968 г. была создана программа «Альтернатива» для ЭВМ «Минск-2», в которой последовательно исключались признаки, дающие отрицательный вклад, и строилась дискриминантная функция по набору оставшихся признаков. Обсчет материалов показал, что уменьшенный набор признаков часто дает лучшее разделение, чем первоначальный набор. Однако это не значит, что исключенные признаки не несут никакой полезной для разделения информации. Улучшение разрешающей способности нового набора признаков объясняется сложностью корреляционных связей между признаками, которую нельзя полностью учесть даже дискриминантным анализом. Поэтому информация, заключенная в исключенных признаках, может привести к повышению разрешающей способности набора оставшихся признаков. Возможность ее использования заключается в следующем.

Полученные в результате расчета на основе полного набора признаков $F(A_к)$ и $F(B_р)$ ($к=1,2, ..., m_1$, $р=1, 2, ..., m_2$) будем считать координатами точек, расположенных на оси абсцисс. При этом часть точек, соответствующих образцам класса A, может попать в область значений $F(B_р)$ и наоборот (на практике именно так чаще всего и бывает). После исключения t признаков будут получены новые значения $F^*(A_к)$ и $F^*(B_р)$, которые также будем считать координатами точек, но расположенных на оси ординат (понятно, что области значений $F^*(A_к)$ и $F^*(B_р)$ тоже могут пересекаться). Таким образом, на плоскости вводится декартова система координат. Ось абсцисс является осью значений дискриминантной функции, построенной на основе исходного набора признаков со сложной корреляционной зависимостью между ними. Ось же ординат является осью значений дискриминантной функции, построенной на основе сокращенного набора признаков с более простой структурой корреляционной зависимости. Понятно, что уже на плоскости область точек с координатами $[F(A_к), F^*(A_к)]$ и область точек с координатами $[F(B_р), F^*(B_р)]$ могут не пересекаться. Образец же C_s будет отнесен к совокупности A или B в зависимости от того, в какую область попадает точка с координатами $[F(C_s), F^*(C_s)]$.

Предлагаемый прием может в ряде случаев сильно повысить разрешающую способность первоначального набора признаков. Очевидно, функции F и F^* осуществляют линейное преобразование исходных признаков. Если последние имели не нормальное распределение, то полученное решающее правило по полному набору не является оптимальным (во всяком случае оптимальность его не доказана). Линейное же преобразование, согласно центральной предельной теореме, может привести к нормально распределенным признакам F и F^*. Таким образом, улучшение распознавания можно связать с улучшением модели, для которой применяется используемый алгоритм. Когда такое преобразование не приводит к нормальному распределению, не происходит и улучшения распознавания.

Пример. В процессе опробования программы «Альтернатива» решалась задача разделения перидотитов и пироксенитов одних массивов от таких же пород других массивов Кольского полуострова. При

этом исследователей больше всего интересовал вопрос о корректности в постановке задачи (ибо, как уже говорилось, основной целью было опробование возможностей программы). Разделение производилось на основе 11 геохимических признаков (содержании в образцах SiO_2, TiO_2, Al_2O_3, FeO, Na_2O, K_2O, NiO, MnO, MgO, CaO и S). В результате решения задачи дискриминантного анализа были получены значения $F(A_к)$ ($к=1, 2, ..., 23$) от 24,5 до 32,6 и $F(B_р)$ ($р=1, 2, ..., 20$) от 22,1 до 30,5, причем область пересечения значений $F(A_к)$ и $F(B_р)$ была очень велика: 84% всех значений F. После исключения пяти признаков, кото-



Разделение образцов $A_к$ и $B_р$ и распознавание образцов C_s на плоскости (F, F^*): *1* — образцы совокупности A; *2* — образцы совокупности B; *3* — контрольные образцы из совокупности B, не вошедшие в число $B_р$

рые давали отрицательные вклады (TiO_2, MnO, MgO, CaO и NiO), решение задачи на основе шести оставшихся признаков дало следующие результаты: значения $F^*(A_к)$ изменялись от 8,0 до 12,6, значения $F^*(B_р)$—от 5,1 до 9,4. При этом в область пересечения значений $F^*(A_к)$ и $F^*(B_р)$ попало меньшее число значений, но все же оно достигло 35%. На координатной же плоскости область точек с координатами $[F(A_к), F^*(A_к)]$ и $F(B_р)$, $F^*(B_р)$ почти не пересекается.

Как видно из приведенного рисунка, один контрольный образец оказался распознан неверно (попал в область точек, соответствующих образцам совокупности A), т. е. ошибка на стадии экзамена составила 17%.

На этом примере можно показать, что предлагаемый прием фактически сводится к усилению значимости признаков, вошедших в уменьшенный набор. Действительно, прямая AB, очевидно, наилучшим образом разделяет точки, соответствующие образцам совокупности A, от точек, соответствующих образцам совокупности B. Уравнение прямой

$$\alpha_1 F + \alpha_2 F^* = a \qquad (7)$$

можно получить тем же методом дискриминантного анализа, использовав в качестве набора признаков два признака — F и F^*. Если в уравнение (7) подставить выражение F из уравнения (1) и аналогичное выражение для F^*, то относить образец C_s к совокупности A или B можно по значению выражения

$$\alpha_1 \sum_{i=1}^{l} \lambda_i C_{is} + \alpha_2 \sum_{i=1}^{l-t} \lambda_i^* C_{is}, \qquad (8)$$

сравнивая его с пороговым (здесь через λ^* обозначены коэффициенты дискриминантной функции, вычисленные на основе уменьшенного набора признаков). Но выражение (8) фактически является не чем иным, как усилением значимости признаков уменьшенного набора с использованием информации, содержащейся в t исключенных признаках.

Пусть, например, $l=5$, $t=2$, причем исключенными оказались второй и четвертый признаки. Тогда

$$a = \alpha_1 \sum_{i=1}^{5} \lambda_i C_{is} + \alpha_2 \sum_{\substack{i=1 \\ i\neq 2,\ i\neq 4}}^{5} \lambda_i^* C_{is} = (\alpha_1 \lambda_1 + \alpha_2 \lambda_1^*) \times C_{1s} + \lambda_1 \lambda_2 C_{2s} + (\alpha_1 \lambda_3 + \alpha_2 \lambda_3^*) C_{3s} + \alpha_1 \lambda_4 C_{4s} + (\alpha_1 \lambda_5 + \alpha_2 \lambda_5^*) C_{5s} = F(C_s).$$

Таким образом, значения первого, третьего и пятого признаков образца C_s уже оказываются с коэффициентами $\alpha_1\lambda_1+\alpha_2\lambda_i^*$ $(i=1, 3, 5)$, тогда как значения второго и четвертого признаков только с коэффициентами $\alpha_1\lambda_t$ $(t=2, 4)$. Таким образом, сокращение набора l признаков до двух признаков F и F^* (корреляция между которыми, как видно из рисунка, довольно велика) также сводится к усилению значимости признаков, оставшихся в уменьшенном наборе при использовании вместе с тем информации, содержащейся в исключенных признаках.

ЛИТЕРАТУРА

1. Нарымский Э. Л. Программа построения линейной дискриминантной функции («Альтернатива») для ЭВМ «Минск-2». Изд. ВИРГ, 1968.
2. Урбах В. Ю. Дискриминантный анализ: основные идеи и приложения (обзор и библиография). В сб.: Статистические методы классификации. Вып. 1, МГУ, 1969.
3. Рао С. Р. Линейные статистические методы и их применение. Наука, 1968.
4. Davis I. C., Sampson R. J. State Geol. Surv. Univ. Kansas, 1966.