УДК 519.24:537.591

Анализ методик интерпретации данных в применении к эксперименту АНИ

А.А.Чилингарян (ЕрФИ)

Проводится анализ существующих методик классификации многомерных данных, методик определения истинной размерности данных, линейных и нелинейных преобразований признакового пространства с целью выделения наиболее эффективных признаков. Приводится схема процедуры принятия статистических решений для эксперимента АНИ .

ВВЕДЕНИЕ

Изучать сильные взаимодействия в области энергий $> 10^{15}$ эВ в настоящее время возможно только с помощью экспериментов с космическими лучами.

Экспериментальные установки регистрируют потоки вторичных частиц, образующих ядерно-электромагнитный каскад (ЯЭК), или, иначе говоря, широкий атмосферный ливень (ШАЛ). Замена измерения характеристик первичного взаимодействия измерением параметров ШАЛ ведет к значительной редукции информации о характере сильного взаимодействия. Прямо не измеряются три самых важных параметра - энергия и тип первичной частицы, высота первого взаимодействия. Неопределенность этих параметров приводит к очень большим флуктуациям экспериментально измеряемых величин. Поэтому моделирование ЯЭК является единственной возможностью сравнения экспериментальных данных о ШАЛ с различными моделями сильного взаимодействия. Вычислительные мощности современных ЭВМ позволяют создавать весьма детальные имитационные программы развития ЯЭК в пространстве и времени [1,2] . Однако принятие решения о выборе правильной модели из списка альтернатив затруднено из-за статистического характера и неизбежных упрощающих предположений моделирования, а также экспериментальных ошибок.

Обычно не удается установить взаимно-однозначного соответствия между моделью и полученной экспериментальной информацией.

Целью эксперимента АНИ [3] является получение данных в области энергий 10^{15}-10^{16} эВ, достаточных для выбора модели сильного взаимодействия.

Для планирования и интерпретации эксперимента АНИ предлагается использовать методы и процедуры, развитые в рамках подхода распознавания образов.

Разработка методики проведения статистического анализа позволит дать рекомендации об оптимальном составе измерительного комплекса, оценить необходимое время измерений.

Всякое распознающее устройство состоит из измерительного комплекса и системы принятия решений.

Задается список альтернатив - возможных состояний природы, отвечающих определенным модельным представлениям. Каждый тип (образ,класс) характеризуется некоторой областью в признаковом пространстве - совокупностью реализаций моделирующей программы. Задачей распознавания является выбор признакового пространства и решающего правила, позволяющего принять точные статистические решения о принадлежности результатов эксперимента одному из классов, не превышая допустимых затрат.

Признаковое пространство - N -мерное пространство измеряемых величин. Каждая координата соответствует экспериментально измеряемому параметру ШАЛ. Решающее правило - алгоритм принятия решения о сходстве или различии вектора измерения с имеющимися классами.

Стоимостные характеристики - время эксперимента и стоимость оборудования, задают ограничения на точность и количество измеряемых параметров и достижимую статистику.

Совокупность классов, заданных в признаковом пространстве, назовем обучающей выборкой, а ее элементы A^{κ}_{ij} - векторами, точками, представителями.

Вопросы атомной науки и техники.
Серия: Техника физического эксперимента. 1981, вып. 2(8), 1-84.

$$A_{ij}^{k}\qquad \begin{matrix} k = 1, \ldots, N & \text{– размерность пространства,}\\ i = 1, \ldots, M & \text{– число векторов в классе,}\\ j = 1, \ldots, L & \text{– число классов,}\end{matrix}$$

или $(\vec{A}_i, \omega_j)$ ω_i - i – й класс.
Если не делаем различия в классах, то

$$A_i^k \qquad i = 1, \ldots, ML,\ ML = M \times L.$$

Векторы экспериментальных величин, поступающие на распознавание, обозначим

$$\vec{X},\ x^k \qquad k = 1, \ldots, N.$$

Будем считать, что априорные вероятности появления представителей различных классов одинаковы и равны $1/L$. Для вычисления степени "близости" векторов в признаковом пространстве будем использовать функцию расстояния с обычными свойствами [4]. Из большого числа предложенных метрик в N-мерном пространстве [5] выберем евклидову

$$D^k(\vec{A}_i, \vec{X}) = A_i^k - X^k; \qquad D(\vec{A}_i, \vec{X}) = \Big[\sum_{k=1}^{N} D^{k\,2}\Big]^{1/2}.$$

Для примера используем простую модель классификации – случай $L = 2$. Общий случай можно привести к нему последовательным разбиением.
Через $P(\vec{x})$, $P(\vec{x})$; $P(\vec{x}/\omega_j)$, $P(\vec{x}/\omega)$ обозначим соответственно функцию распределения вероятности случайной величины и ее плотность и условную функцию распределения и ее плотность. Под ранжированием выборки A_i будем понимать такое её упорядочение, что $A_i \le A_{i+1}$ $i = 1, \ldots, M-1$ или $A_i \ge A_{i+1}$ $i = 1, \ldots, M-1$. В дальнейшем знак E –обозначает математическое ожидание, – T – транспонирование, $\wedge$ – оценку, $\mathcal{E}$ – дисперсию, tr – шпур.

I. КЛАССИФИКАТОРЫ

Классификацией назовем процесс установления соответствия между случайным вектором $\vec{X}$ и классами ω_j, $j = 1, \ldots, L$.

В N-мерном пространстве предъявляется вектор $\vec{X}$. Надо определить, какому из распределений, описанных в этом пространстве, он принадлежит.

Допустим, плотности вероятности (ПВ) этих распределений заданы явно, тогда решающее правило (РП) можно записать в виде [6]

$$\ell(\vec{x}) = p(\vec{x}/\omega_1)/p(\vec{x}/\omega_2) \gtrless 1 \rightarrow \vec{x} \in \begin{cases}\omega_1\\ \omega_2\end{cases}, \qquad (I.1)$$

$\ell(\vec{x})$ – отношение правдоподобия.
Вектор X относится к тому классу, чья ПВ в точке X наибольшая. РП (I.1) называют байесовским критерием, минимизирующим ошибку решения. Это РП оптимально в том смысле, что его использование обеспечивает наименьшую вероятность совершения ошибки классификации.

Можно сформировать РП, используя понятие штрафных функций. Пусть C_{ij} – штраф за решение $\vec{X} \in \omega_j$, если на самом деле $\vec{x} \in \omega_i$. Тогда байесовский критерий, минимизирующий средние потери, имеет вид

$$p(\vec{x}/\omega_1)/p(\vec{x}/\omega_2) \gtrless \frac{C_{21} - C_{22}}{C_{12} - C_{11}} \rightarrow \vec{x} \in \begin{cases}\omega_1\\ \omega_2\end{cases}. \qquad (I.2)$$

Если положить $C_{ii} = 0$; $C_{ij} = 1 - \delta$, $\delta = const$, то есть если в случае правильной классификации штраф равен нулю, а в случае любой ошибки одинаков, то (I.2) совпадает с (I.1).

Для конструирования оптимального классификатора необходимо вычисление ПВ для всех классов. Значение $p(\vec{x}/\omega_j)$ в явном виде редко бывает известно. Для их оценки используется обучающая выборка. Оценивание параметров ПВ известного вида и дальнейшую классификацию объектов на основании этих оценок называют параметрическим методом классификации. В случае отсутствия априорных знаний мы должны как-то оценить ПВ, не зная ее структуры. Этот случай называется непараметрическим оцениванием.

Широкое распространение в этом подходе получил метод потенциальных функций [7], в котором предполагается, что каждая точка обучающей выборки является источником "поля". На основании этой аналогии можно допустить существование эквипотенциальных поверхностей, которые описываются потенциальной функцией $\gamma(\vec{A}_i, \vec{x})$. Значение ПВ в точке $\vec{X}$ оценивается как суперпозиция

потенциальных функций

$$\hat{P}(\vec{x}/\omega_j) = \frac{1}{M}\sum_{i=1}^{M} \gamma(\vec{x}, \vec{A}_{ij}) \qquad (I.3)$$

Желательно выбирать потенциальные функции гладкими и монотонными [8] с выполнением условий:
1) $\gamma(\vec{A}_i, \vec{X})$ – максимальная при $\vec{X} = \vec{A}_i$;
2) $\gamma(\vec{A}_i, \vec{X}) \to 0$ при $D(\vec{A}_i, \vec{X}) \to \infty$.
Возможен и другой тип разложения ПВ [9]

$$\hat{P}(\vec{x}/\omega) = \vec{B}_j^T \vec{\phi}(\vec{x}) \qquad j = 1, \dots, L, \qquad (I.4)$$

где $\vec{\phi}^T = (\varphi_1, \varphi_2, \dots, \varphi_q)$,

$\vec{B}_j^T = (b_{j1}, b_{j2}, \dots, b_{jq})$,
Здесь $\varphi_i(\vec{x})$ – линейно независимый базис,
$\vec{B}_j$ – вектор неизвестных параметров, который должен быть определен так, чтобы ошибка аппроксимации $P(\vec{x}/\omega_j)$ была минимальна.
Обычно минимизируют критерий вида

$$J(B) = \sum_{i=1}^{ML} [\Pi(\vec{A}_i) - B^T\phi(\vec{A}_i)]^T [\Pi(\vec{A}_i) - B^T\phi(A_i)] , \qquad (I.5)$$

где $\Pi(\vec{A}_i) = [P(\vec{A}_i/\omega_1), P(\vec{A}_i/\omega_2), \dots, P(\vec{A}_i/\omega_2)]$,

$\vec{B}^T = [B_1, B_2, \dots, B_L]$.
Для минимизации функционала $J(B)$ вместо неизвестной функции $\Pi(\vec{A}_i)$ можно использовать векторы $[\vec{\alpha}_1, \vec{\alpha}_2, \dots, \vec{\alpha}_L)$:

$$\alpha_{ij} = \begin{cases} 1 & \vec{A}_i \in \omega_i \\ 0 & \vec{A}_i \quad \omega_i \end{cases} \qquad j = 1, 2, \dots, L .$$

Решая уравнение $\nabla J(B) = 0$, получаем решение в виде

$$B = \Delta^{-1} C ; \quad \Delta = \sum_{i=1}^{ML} \phi(\vec{A}_i)\phi^T(\vec{A}_i) ; \quad C = \sum_{i=1}^{ML} \phi(\vec{A}_i)\alpha_i^T . \qquad (I.6)$$

Следующий тип оценки определяет ПВ как математическое ожидание δ -функции [10]

$$P(x) = \int_{-\infty}^{\infty} \delta(x - A) P(A)\, dA = E\{\delta(x - A)\} . \qquad (I.7)$$

Вычислим математическое ожидание по обучающей выборке и заменим δ -функцию ее аппроксимацией:

$$K\left(\frac{x-A}{h}\right)/h \to \delta(x - A) \qquad \text{при} \qquad h \to 0,$$

где $K(x)$ – ядро оценки,

$$\hat{P}_M(x) = E_A\left\{\frac{1}{h} K\left(\frac{x-A}{h}\right)\right\} \approx \frac{1}{Mh}\sum_{i=1}^{M} K\left(\frac{x - A_i}{h}\right) , \qquad (I.8)$$

$\hat{P}_M(x) \to P(x)$, если $h \to 0$ при $M \to \infty$ и $Mh \to \infty$ при $M \to \infty$
В работе [II] (I.8) рассматривается как некорректная задача численного дифференцирования и $\hat{P}(x)$ определяется как решение интегрального уравнения Фредгольма I рода.

В работе [I2] для классификации предлагается использовать концепцию статистически эквивалентных блоков. Обучающая выборка $(\vec{A}_i, \omega_i)$ ранжируется и разбивается на блоки по K членов $K \sim \sqrt{M}$ и, если

$$\min_{i \neq j} \{C_{ij} \times N\omega_i\} < C_0 \times K, \qquad (I.9)$$

то блоку присваивается уровень ω_i ,
$N\omega_i$ – число членов, принадлежащих ω_i в блоке,
C_0 – штраф за отказ принять решение,
C_{ij} – штраф за ошибочное решение.
Далее, объединяя блоки с одинаковой классификацией и распределяя неопределенные блоки по соседним, получаем дробление признакового пространства на области с известной классификацией.

Наиболее распространенный классификатор основан на априорном представлении о "близости" объектов, принадлежащих одному классу [I3] . Вычисляются расстояния от точки $\vec{x}$ до всех представителей обучающей выборки и ранжируются в порядке возрастания.
Оценку ПВ в точке $\vec{x}$ получим в виде

$$P_{ML}(\vec{x}) = \frac{K}{ML}\,\frac{1}{V(K,N,\vec{x})}, \qquad (I.I0)$$

где $V(K, N, \vec{x})$ - объем гипершара радиуса R_K с центром в точке $\vec{x}$. P_K - расстояние от точки $\vec{x}$ до K - го ближайшего соседа (КБС). $\hat{P}(\vec{x})$ является несмещенной и состоятельной оценкой ПВ, если

$$\lim_{ML\to\infty} K = \infty \qquad \text{и} \qquad \lim_{ML\to\infty} K/ML = 0 \qquad (I.II)$$

Условие (I.II) не является необходимым. В работе [I4] показано, что даже для малых ($M \sim I0$) обучающих выборок КБС оценки близки к оптимальным.

Подставляя КБС оценку ПВ в (I.I), получаем

$$\text{где} \quad \begin{matrix} K1 \geq K2 \rightarrow \vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \\ K1 + K2 = K. \end{matrix}, \qquad (I.I2)$$

Таким образом, решение о принадлежности $\vec{x}$ можно принять, определив K ближайших соседей, и выбрав класс, представленный наибольшим числом векторов.
Развитие КБС метода идет в направлении увеличения его точности и уменьшения требований к памяти ЭВМ.

В работе [I5] предложено КБС правило с предварительным редактированием:
- для каждого члена обучающей выборки ($\vec{A}_i$, ω_j) находят K ближайших соседей и получают КБС оценку $\hat{\omega}_j$, и, если $\hat{\omega}_j \neq \omega_j$, $\vec{A}_i$ удаляют из обучающей выборки.

В работе [I6] показано, что использование отредактированной обучающей выборки, наряду с экономией памяти ЭВМ, приводит к значительному улучшению асимптотических свойств оценок.

Если в обучающей выборке каждому классу соответствует неодинаковое количество представителей, то вводится взвешенное расстояние [I7]

$$D(\vec{A}_{ij}, \vec{x}) = \left[\frac{M_j}{\lambda_j \sum_{j=1}^{L} M_j}\right]^{1/N} D(\vec{A}_{ij}, \vec{x}), \qquad (I.I3)$$

где M_j - число представителей в классе ω_j ;
λ_j - априорная вероятность j -го класса.

Для того, чтобы избежать неправильной классификации отдельных, далеко отстоящих точек, предложен [I8] (K, K') БС алгоритм, в котором требуется, чтобы K ближайших соседей вектора $\vec{x}$ находились на расстоянии, не большем R_{zp} .
R_{zp} - пороговое значение, при превышении которого $\vec{x}$ не может считаться близким к классу ω_j . При выполнении этого условия $\vec{x}$ относится к классу ω_j , если не менее K' из K соседей $\vec{x}$ принадлежат ω_j $(K' < K)$.

КБС правило зависит от выбранной метрики и, следовательно, от выбранных единиц измерения признаков. Если измеряются качественно равные признаки (энергия, координаты и т.д.) трудно выбрать согласованный масштаб переменных. В работе [I9] предложено эмпирическое расстояние, инвариантное относительно монотонных преобразований координатных осей. Сначала множества

$$A_i^k \;,\; X^k \qquad \begin{matrix} i = 1, \ldots, ML \\ k = 1, \ldots, N \end{matrix}$$

ранжируются в порядке возрастания. Обозначим ранги A_i^k и X^k через r_i^k и r^k . Расстояние между $\vec{X}$ и $\vec{A}$ определяется так:

$$D(\vec{A}_i, \vec{x}) = \max_{1<k<N} \left| r_i^k - r_x^k \right|.$$

При сборе информации с крупной экспериментальной установки трудно добиться постоянной I00% эффективности всех каналов регистрации. Поэтому векторы измерений будут содержать пропуски, и расстояния между ними и векторами обучающей выборки не будут определены. В работе [20] исследованы различные алгоритмы "заполнения" пропусков:
I. Алгоритм нормализации расстояния

$$D^k\left(A_i^k, x^k\right) = \begin{cases} 0, & \text{если } x^k \text{ отсутствует} \\ A_i^k - x^k \end{cases} \qquad (I.I^4)$$

$$D^2(\vec{A}_i, \vec{x}) = \frac{N}{N-NB} \sum_{k=1}^{N} D_k^2 ,$$

NB – число пропусков.

2. Усреднение

$$D^k\left(A_i^k, x^k\right) = \begin{cases} A_i^k - x^{*k}, & \text{если } x^k \text{ отсутствует} \\ A_i^k - x^k \end{cases} \qquad (I.I5)$$

$$x^{*k} = \frac{2}{M_э(M_э-1)} \sum_{ij} \left| x_i^k - x_j^k \right| ,$$

$M_э$ – число экспериментальных векторов.

3. Зануление

$$D^k\left(A_i^k, x^k\right) = \begin{cases} 0, & \text{если } x^k \text{ отсутствует} \\ A_i^k - x^k. \end{cases} \qquad (I.I6)$$

Эти алгоритмы проверялись для широкого диапазона данных и разного процента пропусков. Лучшие результаты получены при использовании формулы (I.I5).

2. ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА И ИСТИННАЯ РАЗМЕРНОСТЬ ДАННЫХ

Выбор множества признаков, обеспечивающих необходимое качество классификации, представляет собой одну из наиболее трудных задач построения распознающих систем.

Очевидно, что наиболее важными являются признаки, различающиеся для разных классов, а одинаковые для всех классов признаки – **малоценны.** Выбор признаков можно рассматривать как процесс преобразования исходных измерений в более эффективные. Если это преобразование является линейным, то задача сводится к нахождению коэффициентов линейной функции, максимизирующей или минимизирующей некоторый критерий, экстремум которого можно найти с помощью методов линейной алгебры, или, в случае сложного критерия, с помощью методов оптимизации.

Следует учесть, что разделимость классов зависит не только от распределения векторов, но и от используемого классификатора. Лучшим критерием эффективности признаков является малая вероятность ошибки классификации. Из-за того, что вероятность ошибки часто не имеет явного математического выражения, ее подсчитывают экспериментально, выбрав набор признаков и построив классификатор.

В работе [2I] предложено несколько критериев разделимости классов. Они строятся исходя из следующих требований:

I) монотонная связь с вероятностью ошибки или с верхней и нижней границей вероятности;

2) инвариантность относительно взаимно-однозначных отображений;

3) аддитивность по отношению к независимым признакам.

Критерии разделимости классов формулируются с использованием матрицы рассеяния внутри классов

$$G_\omega = \sum_{j=1}^{L} E\left\{(\vec{A} - \vec{M}_j)(\vec{A} - \vec{M}_j)^T\right\} = \sum_{j=1}^{L} \Delta_j , \qquad (2.I)$$

Δ_j – выборочная ковариационная матрица и матрица рассеяния между классами

$$G_б = \sum_{j=1}^{L} (\vec{M}_j - \vec{M}_o)(\vec{M}_1 - \vec{M}_o)^T , \qquad (2.2)$$

$\vec{M}_j$ – математическое ожидание векторов класса,

$\vec{M}_o$ – математическое ожидание смеси распределений.

Матрицы G_ω и $G_б$ инвариантны относительно трансляций системы координат. Критерий разделимости должен увеличиваться при увеличении рассеяния между классами и при уменьшении рассеяния внутри классов. Поэтому можно использовать критерий вида

$$J(\vec{A}_i) = t_r\left(G_\omega^{-1} G_б\right). \qquad (2.3)$$

Преобразование исходного пространства , максимизирующее J , получаем следующим образом. Нахо-

дим собственные функции и собственные значения матрицы $\{G_\omega^{-1} \times G_\ell\}$. Выбираем K "существенных" собственных функций, соответствующих наибольшим собственным значениям, $K < N$, и строим из них матрицу C , которую используем для преобразования N -мерных векторов в новое K -мерное пространство:

$$\vec{A}_i' = C^T \vec{A}_i \qquad C^T = (\vec{\psi}_1, \vec{\psi}_2, \ldots, \vec{\psi}_K) .$$

Отметим, что ранг матрицы G_ℓ равен L , поэтому матрица $G_\omega^{-1} \times G_\ell$ имеет только L ненулевых собственных значений. Остальные $N - L$ признаков не вносят вклад в J .

Алгоритмы линейного преобразования просты и не требуют больших вычислительных мощностей, однако в информации о ШАЛ имеются важные признаки, являющиеся существенно нелинейными функциями исходных измерений.

В настоящее время не существует сколько-нибудь общей теории нелинейных преобразований, поэтому процедуры выделения признаков в сильной степени зависят от конкретной задачи и обычно носят итеративный характер.

Размерность вектора исходной информации с эксперимента АНИ может достигать $\sim 10^3$. Поэтому важно выбрать преобразование, проецирующее эти векторы в пространство наименьшей возможной размерности.

Минимальное число параметров, необходимое для генерации многомерных данных называют истинной или топологической размерностью. Кривая в N -мерном евклидовом пространстве обладает истинной размерностью $N_t = 1$, поверхность любой формы $N_t = 2$ и т.д... Если N -мерные данные лежат на гиперповерхности размерности $K < N$, то их истинная размерность равна K . K задает нижнюю границу возможного понижения размерности и определяет необходимое и достаточное число эффективных признаков.

Идея, используемая для определения истинной размерности многомерных данных, основана на изучении расстояний соседства в локальных областях N -мерного пространства. Рассмотрим векторы, равномерно распределенные в гипершаре радиуса r . Нормализуем евклидово расстояние между двумя векторами из этого шара,

$$R_N = |\vec{x}_1 - \vec{x}_2| / 2r .$$

Исследование распределения расстояний R_N показало [22] , что дисперсия величины R_N обратно пропорциональна размерности пространства N ,

$$N \delta (R_N) \simeq const , \qquad (2.6)$$

то есть с уменьшением размерности пространства происходит "расширение" локальных сгустков. Следовательно, ключевая процедура уменьшения размерности – локальное увеличение дисперсии межточечных расстояний. Этого можно добиться, увеличивая расстояния, большие среднего, и уменьшая расстояния, меньшие среднего. Такая деформация, однако, может сильно исказить исходное распределение. Поэтому надо определить самые существенные свойства исходного распределения и потребовать их инвариантность относительно преобразований, понижающих размерность. Предложенные к настоящему времени инварианты строятся как функции расстояния между точками исходного пространства.

В первых работах по нелинейному проецированию и многомерному шкалированию [23-25] за инвариант был принят ранговый порядок межточечных расстояний в исходном пространстве.

Сначала вычисляются ММ расстояний $MM = M \cdot (M-1)/2$, M – общее число векторов обучающей выборки, множество расстояний ранжируется и ранги всех расстояний запоминаются. Затем выбирается произвольная конфигурация точек в пространстве минимальной размерности, вычисляются ММ расстояний в этом пространстве и рассматривается критерий сходства двух конфигураций:

$$S_{KR} = \sum_{i<j} S_{ij} \, (D_{ij} - D_{ij}^*)/D_{ij} , \qquad (2.7)$$

или

$$S_{SP} = \sum_{i<j} f_{ij} \, (R(D_{ij}) - R(D_{ij}^*)) , \qquad (2.8)$$

где D_{ij}^* межточечные расстояния в новом пространстве, сохраняющие ранговый порядок исходной выборки; D_{ij} – текущее значение межточечных расстояний; $R(D_{ij})$ – ранг расстояния между точками выборки $\vec{A}_i$ и $\vec{A}_j$. f_{ij} – весовой множитель. В частности, f_{ij} можно использовать для определения локальной области соседства, положив $f_{ij} = 0$, если $D_{ij} > \bar{D}_{ij}$, или, если $\vec{A}_i$ и $\vec{A}_j$, принадлежат разным классам обучающей выборки. Критерий минимизируется изменением положения всех точек. Итеративный процесс продолжается до тех пор, пока S не перестает уменьшаться. Если значение S все еще достаточно велико, то размерность исходного пространства увеличива-

ется на I и процесс минимизации продолжается. Когда S достигнет величины 5 - 10%, можно считать, что размерность пространства совпадает с топологической.

Определить истинную размерность можно и другим путем [22] . В исходном пространстве увеличиваем дисперсии локальных расстояний, следя за тем, чтобы ранговые соотношения в локальной области не изменялись. Итеративный процесс происходит до тех пор, пока дальнейшее увеличение дисперсии становится невозможным. Затем вычисляется ковариационная матрица полученной конфигурации точек, и истинная размерность определяется количеством "существенных" собственных векторов этой матрицы.

Но как определить размеры локальных областей? Ответ на этот вопрос можно получить, определив структурный инвариант, автоматически определяющий локальную близость [26] .

Минимальным графом (МГ) М точек в метрическом пространстве назовем древовидную структуру, соединяющую все точки и имеющую минимальную длину.

Основные свойства МГ:

I) Ближайшие соседи (БС) соединены между собой в МГ;

2) МГ инвариантен относительно любых преобразований, сохраняющих ранговый порядок межточечных расстояний;

3) МГ легко вычисляется и его чертеж ясно отражает структуру выборки.

Локальная область определяется соединенными на МГ точками.

Понижение размерности осуществляется специальным преобразованием, сохраняющим структуру МГ. Концепция МГ была использована для построения алгоритма геометрической реконструкции сложных событий в дрейфовых и пропорциональных камерах [27] . Быстродействие алгоритма увеличилось на порядок в сравнении с традиционными методами.

Отрицательной стороной изложенных методик определения истинной размерности является итеративный характер алгоритмов и эвристические правила остановки процесса минимизации.

В работе [28] предложен неитеративный алгоритм, основанный на информации ближнего соседства. Используя оценку плотности вероятности расстояния до K - го ближайшего соседа τ_K , удается получить линейное уравнение, угловой коэффициент которого можно использовать как оценку истинной размерности:

$$\log \vec{\tau}_K = (1/\hat{N}) \log K + \text{const} . \qquad (2.9)$$

$\vec{\tau}_K$ - усредненное по выборке расстояние до K - го соседа,

$\hat{N}$ - оценка истинной размерности.

Вычислив $\vec{\tau}_K$ для нескольких значений K , можно получить значение $\hat{N}$ методом линейной регрессии. Практически, за истинную размерность берется ближайшее к $\hat{N}$ целое число.

Алгоритм проверялся с использованием данных выборок различного типа: многомерного нормального распределения, равномерного и т.д. Чем больше значений K и $\vec{\tau}_K$ используется, тем точнее будет линейная аппроксимация, но даже для прямых, построенных по трем значениям τ_K (K = = I,2,3), получена хорошая сходимость алгоритма. Алгоритм вычисления прост и включает вычисление БС информации, используемой при классификации, так что объем добавочных вычислений невелик. Для получения проекций M векторов исходного пространства в пространство "истинной" размерности можно использовать алгоритмы случайного поиска с возвратом при неудачном шаге [29] .

Итеративный процесс начинается с M произвольных векторов в пространстве "истинной" размерности. На каждом шаге минимизации критерия сходства двух выборок генерируется случайный вектор $\vec{u}$ и новые координаты выборки вычисляются по формуле

$$\vec{A}_i = \vec{A}_i + \vec{u} . \qquad (2.10)$$

Если критерий, вычисленный с новыми координатами, уменьшился, то новые значения принимаются, в противном случае $\vec{u}$ вычитается и генерируется новый случайный вектор. Процесс продолжается до тех пор, пока дальнейшего улучшения получить не удается. Минимизируется критерий вида

$$S_{CAL} = \sum_{i<j} f_{ij} \, D^{2\,*}_{ij} / D^{2}_{ij} , \qquad (2.11)$$

В алгоритме используются ограничения, запрещающие сближение точек, принадлежащих разным классам.

Таким образом, можно получить обучающую выборку компактных , хорошо разделенных классов. Но как использовать эту выборку для классификации? Ведь у нас нет явного математического выражения преобразования признакового пространства, а применять итеративное преобразование к векторам, поступающим на распознавание, бессмысленно.

Имея результаты действия неизвестного нелинейного преобразования выборки $\vec{A}_i' - \vec{A}_i$, можно поставить задачу его аппроксимации, предположив

$$\vec{A}' = \sum_{j=1}^{q} b_j \varphi_j(\vec{A}_i),$$

где φ_j - известные функции.
Коэффициенты b_j находятся из условия минимизации среднеквадратичной ошибки аппроксимации по всей обучающей выборке:

$$J(\vec{b}, \vec{\varphi}) = \sum_{i=1}^{NL} |\hat{\vec{A}}_i' - \vec{A}_i'|^2. \qquad (2.12)$$

Если удастся получить хорошую точность для не очень больших значений q, то (2.11) можно использовать для преобразования экспериментальных векторов и их последующей классификации.

Рассмотрим теперь специальный случай проецирования на признаковое пространство размерности $N=2$. Это преобразование дает возможность отображать данные на экране индикатора и анализировать данные в режиме диалога с ЭВМ.

В работе [30] для целей двухмерного проектирования предложено минимизировать критерий

$$S_{SAM} = \sum_{i<j} \delta_{ij} (D_{ij} - D_{ij}^*), \qquad (2.13)$$

то есть ищется конфигурация в двухмерном пространстве с теми же межточечными расстояниями, что и в исходном пространстве. Итеративный процесс начинается с произвольной конфигурации, для достижения минимума S используется метод наискорейшего спуска - минимизирующий шаг делается в антиградиентном направлении.

В работе [31] предложен алгоритм последовательного проецирования N -мерных точек на плоскость. Первые три точки можно спроектировать так, чтобы все 3 межточечных расстояния сохранились. Последующие точки можно спроектировать с сохранением лишь двух расстояний. Всего сохраняется $3+2\times(M-3)=2M-3$ расстояний из $\frac{M(M-1)}{2}$. Как уже указывалось, минимальный граф МГ содержит $M-1$ расстояние. Поэтому при проецировании можно сохранить структуру МГ. Оставшиеся $M-2$ расстояния можно использовать для сохранения расстояний до второго БС, или можно потребовать сохранения всех расстояний до специально выделенной точки. Такой подход позволит получить проекции, каждая из которых отвечает определенной точке зрения на структуру данных.

ЗАКЛЮЧЕНИЕ

Назначение статистических методов анализа - интерпретация данных с ощутимой случайной изменчивостью. Традиционный способ статистического анализа - выбор семейства вероятностных моделей с последующим оцениванием параметров этих моделей. Если, из-за сложности явления, трудно описать физическую ситуацию адекватной вероятностной моделью, используются непараметрические методы, которые позволяют формировать статистики с распределениями, одинаковыми для более общего семейства плотностей вероятности, чем параметрические семейства.

Для задач классификации достаточно оценить, какое из распределений обеспечивает наибольшую плотность вероятности в исследуемой точке признакового пространства.

Результаты моделирования ядерно-электромагнитного каскада в атмосфере трудно представить вероятностной моделью с небольшим числом параметров. Кроме того, моделирование ЯЭК с энергией $>10^{15}$ эВ требует очень больших машинных мощностей и, следовательно, количество реализаций каскадов с такими энергиями будет невелико.

Поэтому для интерпретации данных с эксперимента АНИ предлагается использовать методы, развитые в рамках концепций распознавания образов - ведущего подхода в попытках применения математических методов обработки информации, относящейся к сложным и плохо формализованным задачам [32].

Схема процедуры принятия статистических решений может быть представлена следующим образом:

1. Получение обучающей выборки - реализация имитационной программы для разных моделей сильного взаимодействия.
2. Определение истинной размерности векторов обучающей выборки.
3. Нелинейное преобразование векторов обучающей выборки и выделения эффективных признаков.
4. Редактирование обучающей выборки.
5. Заполнение "пропусков" в векторах экспериментальной информации.
6. Исследование качества классификации и выбор оптимального K в правиле K ближайших соседей.

7. Аппроксимация нелинейного преобразования и преобразование экспериментальных векторов.
8. Проведение классификации и выбор модели сильного взаимодействия.

Автор благодарен Т.Л.Асатиани и Э.А. Мамиджаняну за интерес к работе и поддержку, А.А.Аглинцеву и А.М.Дунаевскому за полезные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

I. Grieder P.K.F., The relation between extensive air shower data and high energy particle production models, Revista del N. Cim., 1977, vol. 7, p. 1.

2. Dunaevskii A.M., Emelyanov Yu.A., Shorin B.F., Urysson A.V., The calculation of nuclear-electromagnetic cascades, Preprint FIAN, N 149, 1980.

3. Никольский С.И.,Тукиш Е.И.,Фейнберг Е.Л. и др. Исследование взаимодействия адронов и ядер космического излучения при энергиях 10^3 - 10^5 ТэВ (проект эксперимента АНИ). -НСЕФИ, 358(I6),I974,Ереван,I974.

4. Елисеева И.И.,Рукавишников В.О. Группировка, корреляция, распознавание образов ,М.:Статистика, I977,с.35.

5. Дюран Б.,Оделл П. Кластерный анализ ,М.:Статистика, I977,с.I7.

6. Фукунага К. Введение в статистическую теорию распознавания образов,М.:Наука, I979,с.57,I76, I88,267.

7. Айзерман М.А.,Браверман Э.М.,Розоноэр Л.И. Теоретические основы метода потенциальных функций в задачах об обучении автоматов разделению входных ситуаций на классы. -Автоматика и телемеханика,I964, т.25,с.9I7.

8. Meisel W.S., Potential function in mathematical pattern recognition, IEEE Trans. comput. 1969, vol. C 18, p. 911.

9. Kittler J., Classification of incomplete pattern vectors using modified discriminant functions.-IEEE Trans. comput. 1978, vol. C 27, p. 367.

I0. Ванняк В.И.,Стефанюк А.Р. Непараметрические методы восстановления плотности вероятности. -Автоматика и телемеханика, I978, т.39,с.38.

II.Фукунага К. Введение в статистическую теорию распознавания образов, М.: Наука, I979,с.57, I76,I88,267.

I2. Henrichson E.G., Fu K.S., A nonparametric Partitioning procedure for pattern classification, IEEE Trans. comput. 1969, vol. C 18, p. 614.

I3. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука, I979,с.57, I76,I88,267.

14. Levine A., Lustick L., Saltzberg B., The nearest-neighbor rule for small samples drawn from uniform distributions.-IEEE Trans. inform theory, 1973, vol. IT-19, p. 697.

15. Wilson D.C., Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. syst., man. and cybernet., 1972, vol. SMC 2, p. 408.

16. Wagner T.J., Convergence of the edited nearest neighbor.-IEEE Trans. inform theory, 1973, vol. IT 19, p. 696.

17. Brown T.A., Koplowitz T., The weighted nearest neighbor rule for class dependent sample size.-IEEE Trans. comput., 1978, vol. C 27, p. 453.

18. Ben-Bassat M., Newhouse M., Bailis E.I., Object recognition and learning of known and unknown types,-IEEE Trans. comput., 1978, vol. C 27, p. 66.

19. Devroye L.P., A universal k-nearest neighbor procedure in discrimination.-IEEE Trans. comput, 1978, vol. C 27, p. 142.

20. Dixon J.K., Pattern recognition partly missing data, -IEEE Trans. syst., man. and cybernet., 1979, vol. SMC9, p. 617.

21. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука I979,с.57, I76,I88,267.

22. Bennet R.S., The intrinsic dimensionality of signal collections, IEEE Trans. inform. theory, 1969, vol. IT 15, p. 517.

23. Shepard R.N., The analysis of proximities: multidimensional scaling with an unknown distance function.-Psychometrika, 1962, vol. 27, p. 12, 5, 219.

24. Kruskal J.B., Multidimensional scaling by optimizing goodness of fit to a nonparametric hypothesis.-Psychometrika, 1964, vol. 29, p. 115.

25. Kruskal J.B., Nonlinear multidimensional scaling: a nuclear method, Psychometrika, 1964, vol. 29, p. 1.
26. Schwartzmann D.H., Vidal J.J., An algorithm for determining the topological dimensionality of point clusters.–IEEE Trans. comput., 1975, vol. C 24, p. 1175.
27. Kowalski H., Pattern recognition in track chamber spectrometers, DESY - 80/72, 1980.
28. Pettis K.W., Baiby T.A., Tain A.K., An intrinsic dimensionality estimator from near-neighbor information.–IEEE Trans. Pattern Analysis Machine Intell., 1979, vol. PAMI 1,p.25
29. Calvert W., Young T.Y., Randomly generated nonlinear transformations for pattern recognitio –IEEE Trans. syst. science cybern., 1969, vol. SSC5, p. 266.
30. Sammon T.W., A nonlinear mapping for data structure analysis.–IEEE Trans. comput., 1969, vol. C 18, p. 401.
31. Lee R.C.T., Slagle J.R., Blum H., A trangulation method for the sequential mapping of points from N-space to two-space.–IEEE Trans. comput., 1977, vol. C 26, p. 288.
32. Сборник "Современные проблемы кибернетики", М.: Знание, № II, 1979.

Статья поступила в редакцию 15 декабря 1980 года.