ԵՐԵՎԱՆԻ ՖԻԶԻԿԱՅԻ ԻՆՍՏԻՏՈՒՏ
ЕРЕВАНСКИЙ ФИЗИЧЕСКИЙ ИНСТИТУТ



ЕФИ-544(3I)-82

A.A.CHILINGARIAN

# ON THE NEW METHODS OF ROENTGEN-EMULSION CHAMBER DATA TREATING

ԵՐԵՎԱՆ 1982 ЕРЕВАН

ЕФИ-544(31)-82

А.А.ЧИЛИНГАРЯН

# О НОВОЙ МЕТОДИКЕ ОБРАБОТКИ ДАННЫХ РЕНТГЕНОЭМУЛЬСИОННЫХ КАМЕР

Современные установки, изучающие космические лучи, регистрируют пространственные и энергетические характеристики разнотипных частиц. Задачи выбора модели сильного взаимодействия или определения химического состава первичного излучения приходится решать в условиях высокой априорной неопределенности, когда практически ничего не известно о виде функции распределения вероятностей в пространстве признаков. В работе описываются универсальные методы, позволяющие выбирать оптимальные комплексы признаков и проводить статистические выводы о принадлежности экспериментальных данных той или иной модели с наибольшей достоверностью. Методика опробована на задаче различения зарегистрированных рентгеноэмульсионными камерами $\gamma$ - семейств, генерированных первичными протонами или ядрами железа.

Ереванский физический институт
Ереван 1982

ЕФИ-544(3I)-82

A.A.CHILINGARIAN

# ON THE NEW METHODS OF ROENTGEN-EMULSION CHAMBER DATA TREATING

The modern installations for cosmic ray studies register the spatial and energy characteristics of various particles. One has to solve the problems of selecting the strong interaction models or determining the primary radiation chemical composition under conditions of high a priori uncertainty, when there is practically nothing known about the probability distribution function form in the indication space. In this paper the universal methods are reported that allow one with most confidence to select the optimal feature complexes and to make statistical decisions on the experimental data relation to one or the other model. The methods are tested on the problem of discriminating $\gamma$ -families produced by primary protons or iron nuclei and registered by roentgen-emulsion chambers.

Yerevan Physics Institute

Yerevan 1982

ЕФИ-544(3I)-82

YEREVAN PHYSICS INSTITUTE

A.A.CHILINGARIAN

# ON THE NEW METHODS OF ROENTGEN-EMULSION CHAMBER DATA TREATING

Yerevan 1982

Introduction.

Any information obtained on the physical value or the value system assumes either making a decision or estimation[1].

Most of the problems connected with the cosmic radiation study may be formulated in the following way[2]: " The general theoretical scheme and certain empirical data are given, the model that satisfies the both is to be found".

Usually the simplified "inverse" problem is solved: the limited list of models is given (in the physical phenomenon imitation model realization set form - a so-called "training sample") - a model satisfying the experimental data is to be determined. That is, the selection of the model "closest" to the experimental data, and not the parameter estimation is the purpose of the analysis - a pattern recognition problem. The classification algorithm is called a decision rule or a classifier.

The statistical decision optimization is connected with the choice of decision rules minimizing the classification possible error that occurs because of the statistical character of the measured physical processes and registration processes, as well as due to the limited number of imitation experiment realizations and experimental data scarcity.

At present a tendency to construct large measurement complexes investigating in detail both the core part of the wide atmospheric showers and the electromagnetic accompaniment is observed because of single cosmic radiation components low sensitivity to model parameters [3]. The body of primary experimental information from such installations is rather great and involves various particle characteristics.

Thus, one must not rely on the heuristic selection of the model-dependent features when analyzing composed multidimensional data and the empirical decision rules based on them. A universal data treating methods should be created that would allow one to extract the optimal feature complexes and to make statistical decisions with most possible confidence. In this paper the nonparametric decision rules based on the concept of nearest neighbourhood (NN) are applied to roentgen-emulsion chamber (REC) data analysis, and the probability measures of distinction between various models are used to select the model-dependent features.

1. Decision Rules and Classification Error.

The decision rule selection depends on the a priori information available If the training sample vector distribution function density is known, then the Bayesian decision rule can be used for the unknown vector classification [4] *)

$$P_1(\vec{x}/\omega_1) / P_2(\vec{x}/\omega_2) \gtrless 1 \rightarrow \vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \qquad (1)$$

*) Here and below the vector appearance a priori probabilities from $\omega_1$ and $\omega_2$ $\omega_2$ classes are assumed to be identical and equal to 0.5.

That is, vector $\vec{X}$ relates to the class $\omega_1$ or $\omega_2$ depending on which density function is greater at the point $\vec{X}$ .

The Bayesian decision rule corresponds to the maximum a priori information and ensures the smallest error of the classification.

If the density function type is known only, its parameters can be estimated by the training sample , and the parametrical decision rules can be used:

$$\hat{P}_1(\vec{X}/\omega_1) / \hat{P}_2(\vec{X}/\omega_2) \lesssim 1 \rightarrow \vec{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \qquad (2)$$

And if it is impossible to describe the training sample by a function with a finite number of parameters, then the nonparametrical decision rules are used that relate the vector introduced for recognition to the class to which it is "closer". These decision rules are based on the local density function estimation

$$P(\vec{X}/\omega_i) \approx K_i / m_i \phi_N \quad ; \quad \sum_{i=1}^{2} K_i = K \qquad (3)$$

where $K_i$ is the number of $\omega_i$ class representatives among $K$ , i.e. the $\vec{X}$ vector nearest neighbours (KNN), $m_i$ is the total number of $\omega_i$ class vectors, $\phi_N$ is the volume of the $N$ dimensional hyperball with the centre at the point $\vec{X}$ including $K$ nearest neighbours.

We shall obtain the nonparametrical decision rule using the estimation (3)

$$K_1 \gtrless K_2 \rightarrow \vec{X} \in \begin{matrix} \omega_1 \\ \omega_2 \end{matrix} \qquad (4)$$

The distances between $\vec{X} = (X_1 \ldots X_n)$ and $\vec{A}_i = (A_{i1}, \ldots A_{iN})$ vectors in the N-dimensional indication space are determined according to the formula

$$D_e^2(\vec{X}, \vec{A}_i) = \sum_{j=1}^{N} (x_j - A_{ij})^2 \qquad (5)$$

or

$$D_M^2(\vec{X}, \vec{A}_i) = (\vec{X} - \vec{A}_i)^T \Sigma^{-1} (\vec{X} - \vec{A}_i) , \qquad (6)$$

here $\Sigma$ is the covariance matrix of the class the vector $\vec{A}_i$ belongs to.

The distances (6) are invariant relative to the linear transformation of the coordinates, being independent of the used measurement units. The Bayesian classifier error depends on the density function and on the space dimensionality. The (2) and (4) classifiers error mathematical expectation depends, besides that, on the training sample size ( $m_i$ ) and on the decision rule parameters (e.g., on $K$ in the KNN rule) *). The most informative parameter characterizing both the classification decision rule and the "identity" of various models is the misclassification probability.

Therefore the measures assuming a relation to the misclassification probability are of the greatest interest among a number of suggested measures of discrimination (see reviews 6, 7 ) between N-dimensional vector classes. The Bhattacharya distance [8] **)

$$R_B = -\ln \int_{-\infty}^{\infty} \sqrt{P_1(\vec{X}/\omega_1) \times P_2(\vec{X}/\omega_2)}\, d\vec{X} \qquad (7)$$

helps one to express the classifier error upper and lower limits

$$\varepsilon_B \leq 1/2 \exp(-R_B) \qquad (8)$$

$$\varepsilon_H \geq 1/2 - 1/2 (1 - 4\varepsilon_B^2)^{1/2}$$

The values $R_B$ are positive, equal to zero if $\omega_1$ and $\omega_2$ classes overlap, and $\infty$ - if they don't.

---

*) For more details on decision rules see [5].

**) The distance (7) is called also a Hellinger distinction coefficient.

If a priori ideas on distribution density in the indication space are adequate, the $\varepsilon_B$ and $\varepsilon_H$ boundaries will outline the expected classification error calculated by means of the training sample, that allows one to compare the distinctive values of feature complexes eliminating the classification procedure. Unfortunately the $R_B$ analytical calculation is possible only for the case of Gaussian distributions $N_1(\vec{\mu}_1, \Sigma_1)$ and $N_2(\vec{\mu}_2, \Sigma_2)$

Then $R_B = 1/4\,\Delta + 1/8\,D^2$

$$\Delta = 2\ln\left(|\Sigma| / |\Sigma_1| \times |\Sigma_2|\right), \quad \Sigma = \frac{\Sigma_1 + \Sigma_2}{2} \qquad (9)$$

where

$$D^2 = (\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_2)$$

where $|\Sigma_i|$ are the covariance matrix determinants of the normally distributed vectors with the averages $\vec{\mu}_1$ and $\vec{\mu}_2$ .

The general discrepancy between the two classes consists of the part caused by the difference of averages - the Mahalonobis distance $D^2$ , and the covariance matrix differences $\Delta$ .

2. $\gamma$ -Family Recognition

Two coordinates and the energy of $\gamma$ -quanta and hadrons of the so-called "families" are the values observed in "Pamir" experiment. The designed "ANI" experiment will allow one to measure the primary particle energy. At present a number of simulation calculations exist orientated on "Pamir" experiment [9]. The artificial $\gamma$ -family selecting procedure is similar to the experimental one.

Discrimination of the families produced by protons and iron nuclei was the purpose of the analysis. The training sample consisted of the artificial $\gamma$ -families generated by means of the MA model [9]. The artificial events

produced by iron nuclei were presented for recognition.

The distances between vectors are not determined because of the various number of particles in $\gamma$ -families and therefore the training sample vectors different dimensions. Two ways in which the information was reduced for the training sample vectors to have identical dimensionality are: 1.- selection of the two most energetic $\gamma$ -quanta in the family; 2. - calculation of the first two moments of coordinates and energy distribution within $\gamma$ - family. The primary energy was added to the six features obtained. The distance between classes (7), the upper and lower limits of the expected classification error (8) were calculated for the both ways of representing the information. The calculations were performed "including" the features successively to have a possibility to measure their discriminative value. The distance between classes increase and the misclassification probability decreases and achieves the minimum value when using all the seven features. However, as the figure shows, if both the averages and the covariance matrices are different in the second way (curves $D^2$ and $\Delta$ ), then the difference betwee covariance matrices makes the main contribution to distances in the first way of reduction.

The total distance $R_B$ appeared to be somewhat greater in the first way indicating that in this case the ordered statistics is more informative than the distribution moments. The primary energy (the 7-th coordinate) is the most characteristic model-dependent feature for both ways.

The "pseudoexperimental" event recognition was performed by KNN classifier (4) using the Euclidean (5) and invariant metrics (6)(labeled by $\square$ and $\bigcirc$ in the figure ). The primary energy account reduces the classification error from 15 ÷ 20% to 3 ÷ 5%. The classification error being within upper and lower limits means that the (6) metrics application takes into account the

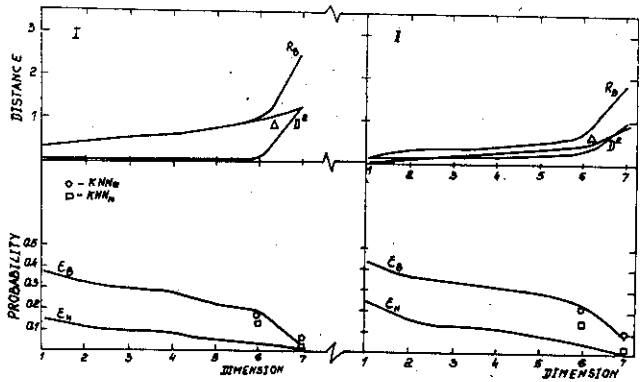
discriminative information totally, the (5) metric application leads to error overvaluation indicating the possible unequal choice of feature measuring units.

## Conclusion

The results obtained show that the nonparametrical classifier application allows one to discriminate $\gamma$ -families produced by protons and iron nuclei. The calculation of the distance probability measures between the training sample classes permits one to outline the model-dependent features and evaluate the minimum-possible error of classification.

However for making statistical decisions with maximum confidence it is necessary to reason out the information reduction procedure (how many moments or ordered statistics are enough to take into account), to define the sufficient number of training sample vectors (only 100 vectors were used because of material scarcity), to work out the distance definition methods without the compulsory condition of Gaussian assumption. Besides, it is necessary to take into account the intermediate nuclei and examine the sensitivity of the classification results to the strong interaction models.

The author is very grateful to E.A.Mamidjanian and E.I.Tukish for their support of the work and also to A.M.Dunaevskii for numerous discussions and valuable remarks.




The distances between the training sample classes ( $\Delta + D^2 = R_B$ ), the upper ( $\mathcal{E}_B$ ) and lower ( $\mathcal{E}_H$ ) limits of the expected classification error of the KNN classifiers ( $\square$ , $\bigcirc$ ) as functions of the used features number.

## References


1. Helstrom C. Quantum Detection and Estimation Theory, Academic PRESS New-York, San Francisco, London, 1976.
2. Bunge M. Philosophy of Physics, Reidel D. Publ. Comp. Dordrecht, 1973.
3. Никольский С.И., Тукиш Е.Н., Фейнберг Е.Л. и др. Исследование взаимодействий адронов и ядер космического излучения при энергиях $10^3 - 10^5$ ТэВ. Препринт ЕФИ, 353(16)-74, Ереван, 1974.
4. Fukunaga K. Introduction to Statistical Pattern Recognition, Academic PRESS, New York and London, 1972.
5. Чилингарян А.А. Анализ методик интерпретации данных в применении к эксперименту АНИ. Вопросы атомной науки и техники, сер. Техника физического эксперимента, вып.2/8/, 1981г. Харьков, 1981.
6. Буреев В.А. Клоков Ю.К., Кудрявцев Т.В. и др. Методы сокращения вычислительных затрат в задачах распознавания изображений. Зарубежная радиоэлектроника, 1980, № 4.
7. Кутин Г.И. Методы ранжировки комплексов признаков. Зарубежная радиоэлектроника, 1981, № 9.
8. Rao S.R. Cluster-Analysis as Applied to Study of Race Mixing in Human Populations. Classification and Clustering. Academic PRESS, Inc New York, San Francisco, London 1977.
9. Dunaevskii A.M., Emelyanov Yu.A., Shorin B.F., Urysson A.U. The Calculation of Nuclear-Electromagnetic Cascades, P.N.Lebedev Physical Institute, preprint N.149, 1980.


The manuscript was received 3 March 1982.

А.А.ЧИЛИНГАРЯН

О НОВОЙ МЕТОДИКЕ ОБРАБОТКИ ДАННЫХ
РЕНТГЕНОЭМУЛЬСИОННЫХ КАМЕР

(на английском языке, перевод З.Н.Абрамян,А.Н.Арутюнян)

Ереванский физический институт

Тех.редактор А.С Абрамян

Заказ 170 ВФ- 05186 Тираж 299

Препринт ЕФИ Формат издания 60x84/16

Подписано к печати 16/IУ-82г. 0,8 уч.изд.л. Ц. 5 к.

Издано Отделом научно-технической информации
Ереванского физического института,Ереван-36,пер.Маркаряна 2

индекс 3624

16