УДК 539.122.164.074.3

Войлов П. Ю., Шаповалов В. Л.

СНИЖЕНИЕ РАЗМЕРНОСТИ ПРОСТРАНСТВА ПРИЗНАКОВ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ РАДИОИЗОТОПНОГО СОСТАВА ИСТОЧНИКОВ ГАММА-ИЗЛУЧЕНИЯ

В статье показана возможность значительного снижения размерности пространства признаков спектральных данных в системе идентификации радио-изотопного состава источников ионизирующего излучение по их спектру. Экспериментальные исследования подтверждают эффективность подхода, что позволяет снизить размерность данных на два порядка.

Ключевые слова: *гамма спектрометрия, радионуклид, пространство признаков, идентификация.*

У статті показано можливість значного зниження розмірності простору ознак спектральних даних у системі ідентифікації радіоізотопного складу джерел іонізуючого випромінювання по їх спектру. Експериментальні дослідження підтверджують ефективність підходу, що дозволяє знизити розмірність даних на два порядки.

Ключові слова: *гамма спектрометрія, радіонуклід, простір ознак, ідентифікація.*

In the article there is shown a grounded possibility for a significant dimensionality decrease for a Radio-Isotope Identification (RIID) System. A sample RIID system built for ^{133}Ba ^{137}Cs and their mixtures' shows the significant dimensionality reduction with only a minor identification error probability increase.

Key words: *gamma-ray spectrometry, radionuclide, space of attributes, identification*

Постановка задачи. При разработке и эксплуатации систем автоматизированного мониторинга радиоактивных загрязнений, систем гамма – видения и т. п. регулярно возникает потребность в решении задачи идентификации радиоизотопного состава источника ионизирующего излучения по зарегистрированному спектру. Особый интерес представляют методы, позволяющие идентифицировать источники смешанного состава с двумя и более компонентами с неизвестным заранее соотношением удельных активностей. Так, для гамма – сканера кругового обзора с кодирующей маской [1] такая идентификация радионуклидного состава источника ионизирующего излучения по спектру, зарегистрированному с направления, является ключевым элементом процедуры дистанционной оценки активности этого источника.

Решение этой задачи представляет собой большое пространство для применения множества маремати-ческих методов и алгоритмов, однако и содержит в себе ряд трудностей. Так, в [2] описан успешный опыт такого решения при помощи нейронных сетей, однако оно получено для малого числа изотопов и сопряжено с известными сложностями при обучении сети. В [3] описан метод идентификации по фотопикам, однако оценка удельных активностей источников невозможна для неизвестной геометрии. В [4] описан подход, близкий к процедурам идентификации голоса; он предполагает разложение спектров на частотные составляющие и сравнение коэффициентов с библиотечными. В [5] приводится обзор большого числа существующих подходов и алгоритмов для идентификации радиоизотопов, однако отмечается, что вариабельность геометрий измерений значительно осложняет решение этой задачи.

Реализацию полевой системы идентификации для работы в условиях неопределенности относительно геометрии эксперимента представляется целесообразным основывать на методе прямого сравнения [5], сравнивая исследуемый спектр с некоторым набором спектров, полученных для широкого набора геометрий. Для организации поиска следует разделить процесс идентификации на несколько уровней, т.е. база спектров будет иметь иерархическую структуру, в простейшем случае двухуровневую. На первом уровне спектр сравнивается с набором центроид, полученных для групп спектров, снятых для одного и того же источника (или смеси) с разными геометриями. Задавшись некоторой метрикой и получив расстояние между исследуемым спектром и каждой из центроид, выбираются k ближайших кластеров. На втором уровне

вычисляется расстояние от до спектров, составляющих выбранные кластера, и проводится идентификация.

Очевидно, эффективность работы такой системы, в первую очередь, будет зависеть от репрезентативности выборки данных, т. е. от её размеров. Существенными проблемами при построении такой системы будут:

– хранение большого объема экспериментальных данных (сотни тысяч и миллионы 512- или 1024-канальных спектров);

– обоснованный выбор меры близости спектров для работы в таких многомерных пространствах;

– организация поиска в базах данных такого большого объема.

Разумеется, решение этих проблем должно быть направлено одновременно по нескольким путям. Существенной мерой, позволяющей обеспечить значительное снижение объема хранимых данных и уменьшение количества операций при выполнении поиске в базе, было бы снижение размерности хранимых данных, за счет выявления скрытых закономерностей внутри их структуры.

Целью данной работы является определение возможности снижения размерности исходного пространства признаков в базе данных спектров, регистрируемых сцинтилляционными детекторами. При этом решаются задачи собственно снижения размерности хранимых данных, определения минимально достижимого предела такого снижения и экспериментальной проверки эффективности различных мер близости в пространстве признаков малой размерности.

Снижение размерности пространства признаков.

Для построения макета системы идентификации из 440 моделей была выполнена серия из 80 базовых измерений для двух изотопов, ^{137}Cs и ^{133}Ba, размещаемых на малом удалении от детектора (0,25..2 м, шаг 0,25 м) за преградами различной толщины и состава (сталь толщиной 1 мм, 3 мм и 5 мм, дерево и пластик), излучение которых регистрировалось гамма-сканером (кристалл CsI(Tl) Ø50×100 мм). 360 смесевых спектров были получены аддитивно из базовых измерений. Центроиды кластеров (усредненные спектры по всему набору геометрий для данного изотопа или смеси) и два реальных спектра представлены ниже (рис. 1).

Снижение размерности проводилось методом главных компонентов [6], рассматривая в качестве класса допустимых преобразований **F** всевозможные линейные нормированные ортогональные комбинации энергетических каналов, т.е.:

$$z_j(X) = c_{j1}(x_1 - \bar{x}_1) + \ldots + c_{jn}(x_n - \bar{x}_n)$$

$$\sum_{v=1}^{n} c_{jv}^2 = 1,\ j = 1, 2, \ldots n \quad (1)$$

$$\sum_{v=1}^{n} c_{jv} c_{kv} = 0,\ j,k = 1, 2, \ldots n;\ j \neq k$$

где мера информативности

$$I_{n'}(Z(X)) = \frac{Dz_1 + \ldots + Dz_{n'}}{Dx_1 + \ldots + Dx_n} = \max_{Z \in F}\{I_{n'}(Z(X))\} \quad (2)$$

Где z_j – проекция спектра в новое пространство признаков, $X = \{x_1, \ldots x_n\}$ – исходный спектр, c_{ij} – некоторые коэффициенты.

В результате множество точек было спроецировано в новый ортогональный базис, причем в качестве оценки меры информативности (2) использовалось отношение накопительной суммы собственных значений матрицы ковариаций к полной сумме собственных значений (3)

$$J = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n'} \lambda_j},\ k = 1,..n' \quad (3)$$

где – λ_i – собственные значения.

По этому критерию для исследуемого набора спектров оказалось, что $J \geq 0.99,\ k \geq 10.$ Однако ниже будет показано, что для эффективной работы системы достаточно разложение до двух компонент.

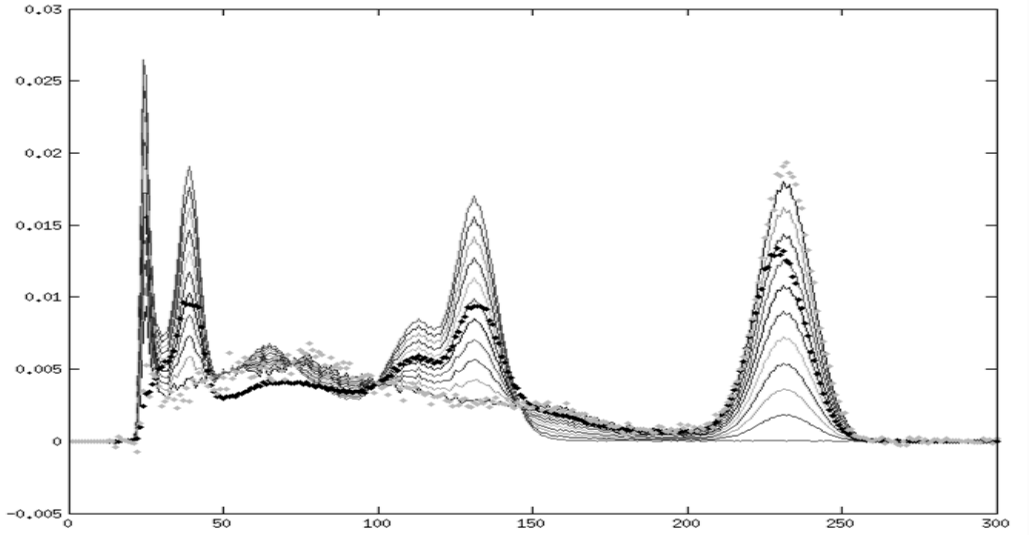

Рис. 1. Спектры – центроиды одиннадцати кластеров (сплошными линиями) и два реальных спектра (черными и серыми точками).

На рис. 2 показано представление центроид кластеров, разложенных на пять компонент. Для сравнения показаны те же два спектра, что были выделены на рис. 1, но спроецированные в новое пространство.

Сравнительная эффективность работы различных метрик в исходном и спроецированном пространствах оценивалась методом десятикратной стратифицированной кросс – проверки [7] для четырех метрик: метрики Минковского (4) для случаев манхэттенского ($p = 1$) и эвклидова ($p = 2$) расстояния, метрики Чебышева (5) и корреляционной метрики (6) [8]:

$$L_p\left[(x_1,\ldots,x_n),(y_1,\ldots,y_n)\right]=\sqrt[p]{\sum_{i=1}^{n}\left|x_i-y_i\right|^p} \quad (4)$$

$$L\left[(x_1,\ldots,x_n),(y_1,\ldots,y_n)\right]=\lim_{k\to\infty}\left(\sqrt[k]{\sum_{i=1}^{k}\left|x_i-y_i\right|^k}\right) \quad (5)$$

$$L\left[(x_1,\ldots,x_n),(y_1,\ldots,y_n)\right]=1-\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sigma_x\cdot\sigma_y}\right) \quad (6)$$

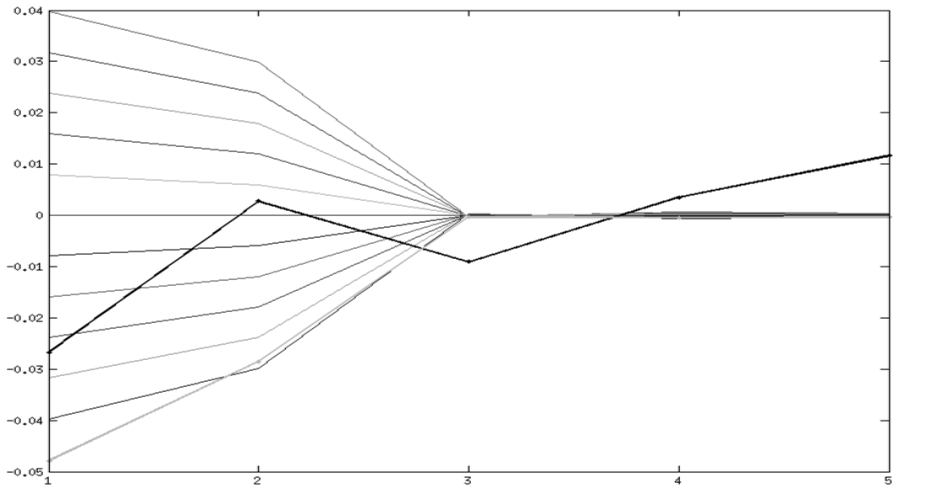

Рис. 2. Те же кластеры и спектры в пятимерном пространстве.

На рис. 3 приведена зависимость вероятности выбора правильного кластера в числе найденных одного, двух и трех ближайших кластеров от количества новых признаков для этих мер. Как видно из рис. 3, максимальная эффективность достигается при $n' = 2$ для чебышевской метрики, однако, учитывая её дальнейшее снижение на $n' = 3$, эвклидова мера представляется более стабильной. Можно считать возможным в данной базе использование эвклидовой меры для $n' = 2$.

Из графиков на рис. 3 также следует, что для данной базы знаний и исследуемых метрик использование $n' > 2$ вообще нецелесообразно, поскольку не приводит к улучшению качества распознавания. В табл. 1 приведены значения вероятностей нахождения истинного кластера как 1-го, 2-го и 3-го ближайшего среди исходных (в диапазоне от 0-го до 300-го каналов) и спроецированных в двухмерное пространство спектров, а также вероятность ошибки (отсутствия истинного кластера среди трех ближайших).

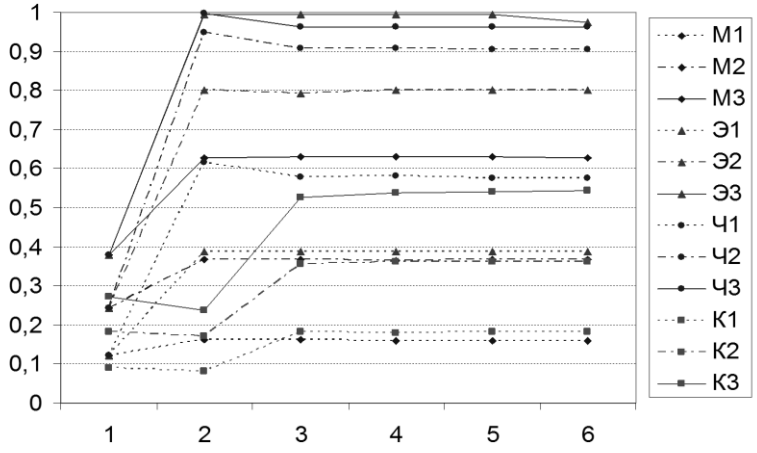

Рис. 3. Зависимость эффективности различных метрик в спроецированном пространстве от его размерности (М – манхэттенская мера, Э – эвклидова, Ч – чебышевская, К – корреляционная, число означает нахождение истинного кластера среди одного, двух или трех ближайших, соответственно).

Таблица 1

Сравнительная эффективность метрик в исходном, и в сниженном до двух измерений пространствах

Мера	**Исходные спектры**				**Спроецированные спектры**			
	I	**II**	**III**	**-**	**I**	**II**	**III**	**–**
Манхэттенская	0,72	0,25	0,03	0	0,16	0,20	0,26	0,38
Эвклидова	0,68	0,30	0,02	0	0,39	0,41	0,19	0,01
Чебышевская	0,53	0,26	0,14	0,08	0,62	0,33	0,05	0
Корреляционная	0,54	0,30	0,13	0,03	0,08	0,09	0,07	0,76

Очевидно, манхэттенская и корреляционная меры, весьма эффективные при идентификации в исходном пространстве, дают значительную ошибку в пространстве меньшей размерности. Следует отметить также рост эффективности чебышевской меры. Из приведенных результатов видно, что наиболее устойчивой к преобразованиям пространства признаков, и обеспечивающей высокую эффективность поиска, является эвклидова мера.

Выводы. Описан способ снижения размерности представления спектральных данных на первом уровне двухуровневой системы идентификации радиоизотопного состава источников. Показана эффективность подхода на реальных спектрах, при увеличении вероятности ошибки идентификации не более, чем на 1 %. Показано, что снижение размерности может приводить к необходимости замены используемой в системе метрики.

ЛИТЕРАТУРА

1. Плахотник В. Ю. «Гониометр» – гамма-сканер кругового обзора с кодирующей маской / Плахотник В. Ю., Кочергин А. В. // Вісник ВНУ ім. В.Даля, № 9(127), ч.1/2008, с.162-166
2. Кочергин А. В, Пивоварцев С.С. Нейронная сеть для идентификации нуклидов по гамма спектру / «Искусственный интеллект» – Донецк, 2008 г., №4, с.600-604.
3. Кочергин А. В. Идентификация радионуклидов в сцинтилляционной гамма спектрометрии методом разложения / А. В. Кочергин // ВНУ (электрон. издание) – №5Е/2009.
4. Owsley, L. M. D. J. J. McLaughlin, L. G. Cazzanti and S. R. Salaymeh. Using Speech Technology to Enhance Isotope ID and Classification. – Prc. IEEE Nuclear Science Symposium, Orlando, FL, October 2009.
5. Burr T., M. Hamada, «Radio-Isotope Identification Algorithms for NaI γ Spectra» – Algorithms, Vol.2, No.1, 2009.
6. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989.
7. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Selection and Model Estimation. – Prc. of the 14th International Joint Conference on Artificial Intelligence 2 (12), 1995.
8. Zezula P., G. Amato, V. Dohnal, M. Batko. Similarity Search. The Metric Space Approach. – Springer Science+Business Media, Inc. 2006.

Рецензенти: к.т.н. Малахов О. В.;
к.т.н. Дубровкина М. В.

Стаття надійшла до редколегії 20.09.10 р.