УДК 539.122.164.074.3

П.Ю. Войлов, А.А. Ильченко, В.Д. Шаповалов

НЕКОТОРЫЕ ВОПРОСЫ ФОРМИРОВАНИЯ БАЗЫ ЗНАНИЙ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ РАДИОИЗОТОПНОГО СОСТАВА ИСТОЧНИКОВ ИОНИЗИРУЮЩЕГО ИЗЛУЧЕНИЯ

Анотація. У статті розглянуті деякі питання побудови системи ідентифікації радіонуклідів по їх спектру методом прямого порівняння. Сформульований принцип роботи ієрархічної системи ідентифікації. Запропоновано спосіб обчислення граничної кількості моделей в базі, а також спосіб оцінки ефективності вибраної міри близькості. Виконаний ряд експериментів, на базі яких створений макет системи ідентифікації.

Введение

Задача идентификации радиоизотопного состава источников ионизирующего излучения занимает одно из центральных мест в разработке и эксплуатации средств радиационного мониторинга окружающей среды. Особый интерес при этом представляют методы, позволяющие идентифицировать источники смешанного состава с двумя и более компонентами с неизвестным заранее соотношением удельных активностей. Задача существенно усложняется при использовании сцинтилляционных кристаллов с малым энергетическим разрешением (NaI(Tl), CsI(Tl)) и значительными вариациями геометрии измерений.

Для решения этой задачи разработано большое количество подходов и алгоритмов [1]. Описанное в [2] успешное решение задачи при помощи нейронных сетей получено для малого числа изотопов; кроме того, нейросетевой подход сопряжен с известными трудностями при обучении. Методы идентификации по фотопикам (например, [3]) дают удовлетворительные результаты, однако, при неизвестной геометрии измерения, оценка удельных активностей источников может вызывать трудности. В [4] описан подход, близкий к процедурам идентификации голоса; он предполагает использование энергетических спектров функций, причем идентификация выполняется классификаторами, основанными на вычислении расстояния Кульбака-Лейблера между исследуемым и библиотечными

спектрами, авторами [4] показана высокая эффективность разработанной ими модели, однако вопрос идентификации в условиях экранирования (особенно смесевых изотопов) ими на данный момент не решен.

При проведении полевых измерений в условиях неопределенности, типичных для радиационного мониторинга на местности, эффективным подходом может являться метод прямого сравнения, в котором исследуемый спектр сравнивается с библиотечным набором, полученным для широкого набора геометрий (включая меняющиеся состав и толщину защиты, расстояние до источника и т.п.) Этот набор может быть весьма велик за счет моделирования синтетических спектров для сотен тысяч комбинаций геометрий защит и комбинаций источников, и охватывать весь диапазон геометрий и ситуаций, предусмотренных техническим заданием.

Такая организация позволит свести процесс идентификации к поиску n ближайших соседей среди библиотечных спектров, с последующим анализом качественных признаков, свойственных этим спектрам (поиск общностей в составе источников, геометриях и т.п.), что позволит сделать окончательный вывод о радиоизотопном составе и геометрии измерения. Очевидно, такая система должна включать в себя значительный объем реальных и синтетических данных. Поиск в такой системе может быть разделен на несколько уровней, т.е. база спектров будет иметь иерархическую структуру, которая может быть организована при помощи методов кластерного анализа. Для организации поиска в такой большой по объему системе должны использоваться методы и алгоритмы разведки данных (так, для организации поиска могут быть использованы древовидные структуры, индексирование, локально-чувствительные хеширующие функции и т.п. [5])

Целью данной работы является исследование некоторых основных требований к библиотеке данных для системы идентификации источников ионизирующего излучения по спектрам, регистрируемым сцинтилляционными детекторами; решаются задачи определения количества хранимых моделей, а также предлагается проверка эффективности меры близости, используемой при

идентификации, при помощи критерия десятикратной стратифицированной кросс-проверки.

Иерархическая библиотека спектров

Формирование уровней системы

Как уже отмечено выше, система идентификации радиоизотопов может иметь несколько уровней, в простейшем случае рассмотрим вариант с двумя уровнями. На первом шаге в такой системе исследуемый спектр сравнивается, используя некоторую метрику Lx, со множеством спектров, представляющих собой центроиды кластеров, создаваемых по каждому целевому классу. Таким образом, на первом шаге отбираются k ближайших центроид и соответствующих классов, которые составлены моделями, соответствующими определенному радионуклидному составу источника, и отличающимися лишь геометрией эксперимента. Для расчета z-го канала центроиды v может использоваться выражение (1).

$$v_z = \frac{u_z \sum_{i=1}^{a} p_i s_i^{(z)}}{\sum_{i=1}^{a} p_i \sum_{ij=1}^{b} u_j} \qquad (1)$$

где uj – канальные весовые коэффициенты, определяющие вклад j-го канала в сиистеме, pi – модельные весовые коэффициенты, определяющие вклад i-й модели в центроиду класса, $s_i^{(z)}$ - значение z-го энергетического канала в модели si.

На втором шаге целевой спектр, используя эту же или иную метрику, сравнивается непосредственно со всеми моделями, входящими в отобранные классы, и отбираются модели, ближайшие к исследуемому спектру.

На третьем шаге выполняется интерпретация набора отобранных ближайших моделей (при малых расстояниях, соответствующих высокой степени близости исследуемого и библиотечного спектров), которая позволит выполнить более точную оценку активностей в смесевых изотопах и т.п. В простейшем случае (или при высокой степени неопределенности) на третьем шаге может отбираться лишь одна ближайшая модель.

Разумеется, для большого количества моделей система может и должна быть масштабирована на более, чем два шага поиска, и будет

представлять собой иерархическую кластерную структуру, на каждом уровне которой может использоваться своя метрика, свои наборы коэффициентов или пространства признаков и т.п. Вопрос организации таких структур требует отдельного рассмотрения.

Требуемое количество моделей

Практический интерес представляет количество моделей, потребных для формирования такой системы. Пусть система представляет собой предельный случай, соответствующий полному перебору, при котором имеется база знаний, содержащая модели для всех возможных ситуаций, включающих наличие преград, множество библиотечных изотопов и возможность идентификации смесевых источников с различными удельными активностями компонентов. Тогда для расчета числа моделей, полностью описывающих все возможные ситуации (т.е. полный набор), можно использовать выражение (2).

$$M(n, l, g, d) = g\sum_{i=1}^{l} \frac{n!}{i!(n-i)!} \cdot f(i, d) \qquad (2)$$

где n – общее число идентифицируемых радионуклидов в библиотеке, l – максимально возможное число радионуклидов в смеси, g – количество геометрий («фактор размытия»), d – число дискретных уровней значений удельных активностей Ai изотопов в смеси, таких, что:

$$\begin{aligned} &\sum_{i=1}^{q} A_i = 1; q \in \{1, 2 \,..\, l\}; \\ &A_i = k \cdot \frac{1}{d}; k \in \{1, 2 \,..\, d-1\} \end{aligned} \qquad (3)$$

и f(i, j) – рекурсивная функция, описывающая число сочетаний таких удельных активностей в смеси, и равна:

$$f(i, j) = \begin{cases} f(1, j) = 1 \\ f(i, j) = \sum_{k=1}^{j-i+1} f(i-1, j-k) \end{cases} \qquad (4)$$

Очевидно, что при l = 1, число m определяется лишь числом изотопов и числом геометрий (и равно их произведению). В частном же случае, достаточно близком к реальной задаче, для библиотеки из 20 изотопов, с возможностью идентифицировать смеси до 5 изотопов с удельными активностями, принимающими значениями из ряда Ai = {0,1, 0,2 .. 0,9}, на расстояниях 1,0 .. 5,0 м с шагом 1,0 м, с тремя

вариантами преград (без преграды, сталь 3 мм и сталь 5 мм) количество моделей будет равно M(20, 5, 15, 10) ~ 36·106. На рис.1 представлены зависимости полного числа моделей от числа изотопов n для g = 15 и d = 10 при l = 5, 4, 3, 2 и 1 (сверху вниз). Как видно из графика, влияние l оказывается, в целом, более выраженным, нежели влияние n.

Приведенная зависимость позволяет оценить сложность базы знаний, но, разумеется, полный набор избыточен, и на практике необходимости в построении такого количества моделей не будет. Реальная база будет содержать значительно меньшее число моделей, что может быть связано с различными обстоятельствами. Среди общих для всех практических задач обстоятельств, которые могут на порядки уменьшить число хранимых моделей, можно выделить следующие:

- малые расстояния между некоторыми моделями, что делает нецелесообразным хранение избыточных вариантов;

- ограничения, накладываемые техническим заданием (например, необходимость одновременной идентификации лишь ограниченного набора смесевых источников).

Кроме того, некоторые факторы, влияющие на размер базы, могут быть индивидуальными для различных задач. Так, для указанного выше примера базы характерно следующее:

- малое рассеяние на воздухе излучения для высокоэнергетических источников и их комбинаций;

- эффективное рассеяние на преграде малых энергий делает невозможным выход излучения за пределы преграды.

Очевидно, хранение данных по таким геометриям и сочетаниям источников нецелесообразно.

Эффективность метрики

Большой интерес представляет вопрос выбора метрики в пространстве спектров. Известно большое число различных мер расстояния, кроме того, пространство признаков (спектров) также может быть подвергнуто преобразованиям вдоль некоторых из своих координат (выделение энергетических окон, назначение весовых коэффициентов и т.п.). Поскольку еще в [6] показано, что не существует единого алгоритма, который мог бы решать задачу кластеризации для любой конфигурации представляющих точек,

задача выбора метрики для заданного пространства признаков должна решаться заново для каждой конкретной задачи и набора моделей.

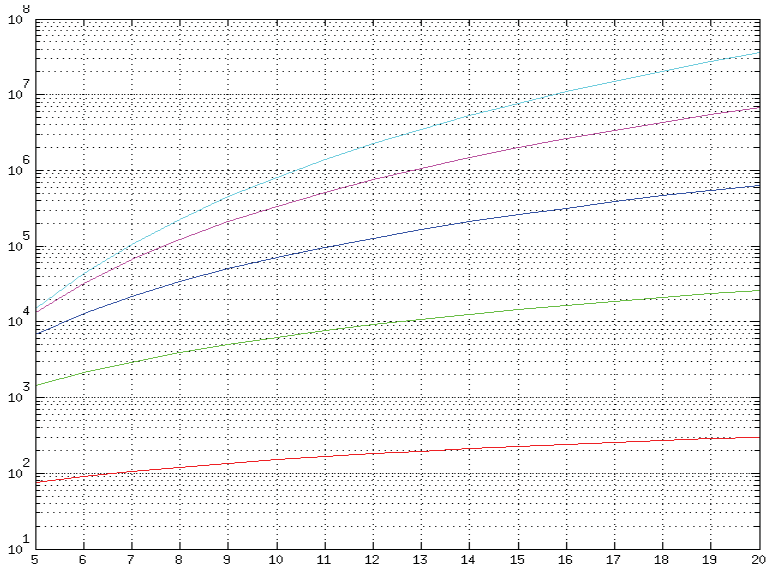

Рисунок 1 – Зависимость полного числа требуемых моделей от n для l = 5, 4, 3, 2 и 1 (сверху вниз)

Опишем простейший алгоритм такой проверки. Рассматривая метрику как часть классификатора, оценка её эффективности может быть сделана при помощи метода 10FSCV (стратифицированной десятикратной кросс-проверки) [7]. На нулевом шаге проводится стратификация обучающей выборки, которая будет заключаться в группировке множества точек обучающей выборки, соответствующих каждому классу (а таковым множеством в рассматриваемой задаче являются множество из g моделей для каждого нуклида или смесевого источника) и определении вида функции w(j, t) для выбора тестируемых моделей. Затем выполняется собственно десятикратная кросс-проверка, когда на каждом шаге из каждого класса изымается набор из g/10 моделей, проводится заново вычисление центроид, по которым выполняется идентификация изъятых моделей. Следует особо подчеркнуть, что для двухшаговой идентификации, описанной выше, под корректной идентификацией следует понимать вхождение истинного кластера в число найденных k ближайших кластеров, причем в общем случае $k > 1$. Тогда для набора из q метрик наиболее эффективной будет являться такая метрика L, для которой вероятность корректной идентификации:

$$P_x = \sum_{t=1}^{10} \frac{\sum_{c=1}^{k} \sum_{j=1}^{g/10} G_c(s_{w(t,j)}, V_t, L_x)}{M(n, l, g, d)} \rightarrow \mathbf{max}(P_x),\ x \in \{l, .. q\} \qquad (5)$$

где $G_c(s_i, V_t, L_x)$ – функция такая, что:

$$G_c(s_i, V_t, L_x) = \begin{cases} 1, \text{ центроида истинного класса} \\ \text{модели } s_i \text{ является c – м ближайшим} \\ \quad \text{соседом } s_i \text{ среди центроид} \\ \text{кластеров разбиения } V_t \text{ для метрики } R_x \\ 0, \text{ иначе} \end{cases}$$

si – i-я модель, $i \in \{l, ... M(n, l, g, d)\}$, w(t, j) – некая функция выбора из класса j-й тестируемой модели для t-го шага, удовлетворяющая условию стратификации, Vt – t-е множество $\frac{M(n, l, g, d)}{g}$ центроид, рассчитанных за вычетом t-го множества $\frac{g}{10}$ моделей.

Тогда вероятность ошибки 1-Px для метрики Lx, очевидно, есть вероятность попадания истинного кластера в число k+1 … $\frac{M}{g}$ ближайших соседей.

Экспериментальные исследования

Для проверки некоторых базовых свойств предлагаемой системы идентификации была выполнена следующая работа. Был построен макет системы идентификации из M(2, 2, 40, 10) моделей, для чего была выполнена серия из 80 базовых измерений для двух изотопов, 137Cs и 133Ba, размещаемых на малом удалении от гамма-сканера (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) Ш50Ч100 мм). Согласно (2) M(2, 2, 40, 10) = 440; недостающие 360 смесевых спектров были получены аддитивно из базовых измерений. Центроиды кластеров (усредненные спектры по всему набору геометрий для данного изотопа или смеси) представлены на рис. 2. На этом же рисунке выделен точками один из экспериментальных спектров; хорошо видны отличия от любой из полученных центроид.

 ISSN 1562-9945

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

$$L_p\left[(x_1,...,x_n),(y_1,...,y_n)\right] = \sqrt[p]{\sum_{i=1}^{n}\left|x_i - y_i\right|^p} \qquad (6)$$

$$L\left[(x_1,...,x_n),(y_1,...,y_n)\right] = \lim_{k\to\infty}\left(\sqrt[k]{\sum_{i=1}^{k}\left|x_i - y_i\right|^k}\right) \qquad (7)$$

$$L\left[(x_1,...,x_n),(y_1,...,y_n)\right] = 1 - \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sigma_x \cdot \sigma_y}\right) \qquad (8)$$

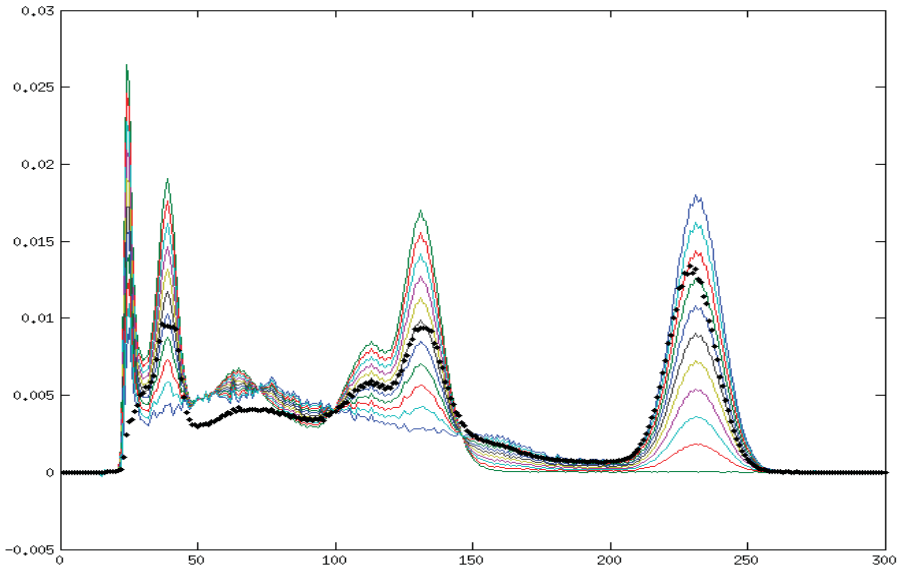

Рисунок 2 – Нормированные центроиды кластеров (по оси абсцисс – номер энергетического канала)

Мы намеренно ограничили здесь набор вариантов простейшими метриками, не рассматривая, например махаланобисову меру, меру EMD и т.п. Это связано с тем обстоятельством, что потери процессорного времени на более сложные алгоритмы (особенно EMD) будут весьма велики, поэтому их применение должно быть обосновано, а эффективность подтверждена экспериментально. Рассмотрению данного вопроса авторы планируют посвятить отдельную работу.

В таблице 1 приведены результаты десятикратной стратифицированной кросс - проверки (выполненной в соответствие с соображениями, изложенными выше) для q = 8 метрик, по четыре метрики для двух вариантов пространства признаков. В первом варианте признаки рассматриваются в окне, ограниченном [30; 300] каналами (фактически, для 133Ba и 137Cs – полное энергетическое окно за исключением потенциально зашумленных нижних каналов). Во втором варианте рассматривается комбинированное окно в двух диапазонах - [100; 150] и [200; 300] каналов (пики полного поглощения). Приведены вероятности выбора истинного кластера при k = 3 для первой, второй и третьей ближайших центроид, а также вероятность ошибки.

По приведенным результатам видно, что для данного макета наиболее эффективной является метрика Минковского и диапазон [30; 300] каналов. Из соображений увеличения быстродействия можно рекомендовать манхэттенскую меру.

Здесь следует отметить, однако, что авторами установлено существенное снижение эффективности манхэттенской меры при работе в пространстве признаков уменьшенной размерности (например, с применением анализа главных компонентов). Поскольку снижение размерности является одним из важнейших способов увеличения быстродействия системы идентификации, в общем случае эвклидова мера является более предпочтительной для использования. Подробное рассмотрение этого вопроса, как и исследование эффективности различных методик снижения размерности, впрочем, выходит за рамки данной работы, поскольку требует значительного расширения макета системы.

Таблица 1

Результаты применения метода 10FSCV для проверки эффективности мер близости на макете системы идентификации

Мера	Полный спектр				Работа в окнах			
	I	II	III	-	I	II	III	-
Манхэттенская	0,72	0,25	0,03	0	0,33	0,23	0,14	0,30
Эвклидова	0,68	0,30	0,02	0	0,34	0,24	0,14	0,28
Чебышевская	0,53	0,26	0,14	0,08	0,36	0,24	0,16	0,24
Корреляционная	0,54	0,30	0,13	0,03	0,09	0,09	0,09	0,73

 ISSN 1562-9945

Выводы

Рассмотрены некоторые общие соображения и принципы построения системы идентификации радионуклидов по их спектру, сформулированы основные требования к такой системе. Предложены способ вычисления граничного количества моделей в базе, который может использоваться для грубой оценки сложности разрабатываемой системы, а также способ оценки эффективности выбранной меры близости, показаны результаты на макете системы идентификации.

Целями дальнейших исследований являются:

- методы и алгоритмы кластеризации и построение иерархической поисковой системы;

- алгоритмы и методы снижения размерности данных на каждом уровне;

- методы и алгоритмы разведки данных для организации поиска на каждом уровне иерархической системы.

Разумеется, эти вопросы должны решаться для конкретного набора спектральных данных, что делает необходимым расширение тестовой базы за счет синтетических и реальных спектров, с учетом влияния вариаций калибровки и соотношения сигнал-шум в исследуемом спектре.

ЛИТЕРАТУРА

1. T. Burr, M. Hamada, “Radio-Isotope Identification Algorithms for NaI γ Spectra” - Algorithms, Vol.2, No.1, 2009.
2. Кочергин А. В., Пивоварцев С. С. Нейронная сеть для идентификации нуклидов по гамма спектру /«Искусственный интеллект» - Донецк, 2008г., №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. L. M. D. Owsley, 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. P. Zezula, G. Amato, V. Dohnal, M. Batko. Similarity Search. The Metric Space Approach. – Springer Science+Business Media, Inc. 2006.
6. Backer E. Cluster analysis by optimal decomposition of induced Fuzzy sets. Delftse Universitaire Pres, Delft, Holland, 1978.
7. R. Kohavi 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.

Получено 25.04.2010г.