УДК 621.382

Г.Ю. Щербакова, В.Н. Крылов, Р.А. Писаренко, О.В. Логвинов

Одесский национальный политехнический университет, Одесса

ИССЛЕДОВАНИЕ АВТОМАТИЗИРОВАННОЙ КЛАСТЕРИЗАЦИИ С ИСПОЛЬЗОВАНИЕМ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Предложен метод кластеризации при обработке данных, который позволяет определять диапазоны изменения координат центров кластеров с использованием вейвлет-преобразования. Метод может быть применен при выборе параметров классификатора с учетом требуемого уровня достоверности в автоматизированных системах обработки данных. При исследовании метода определены диапазоны изменения координат центров кластера для тестовых данных.

***Ключевые слова**: классификация, кластеризация, вейвлет-преобразование, автоматизированные системы, обработка данных, мультистартовая оптимизация, диапазон.*

Введение

В автоматизированных системах (АС), например, для технической или медицинской диагностики, важной процедурой является классификация. Классификация с самообучением состоит из двух процедур: кластеризации и классификации [1]. При кластеризации может быть не известно количество кластеров, кластеры могут иметь сложную форму и пересекаться, различаться по размерам и плотности. При классификации с самообучением может быть не известно количество групп параметров в пространстве признаков и число образов в группах. Это может происходить, например, при исследовании новых технологических процессов и лекарственных средств. В ряде задач технической диагностики среди изделий необходимо выделить группы с общими свойствами. Количество таких групп (кластеров) и число изделий в них, как правило, неизвестно. Так, например, в случае смены паяльной пасты могут измениться параметры качественных и некачественных соединений и количество групп, которые они образуют в пространстве признаков, при производстве интегральных схем может меняться количество групп дефектов и их расположение на пластинах в различных партиях.

В подобных условиях существующие методы кластеризации, как иерархические, так и итеративные отличаются низким качеством [1 – 4]. При отладке указанных АС возможна ситуация, когда необходимо выделить не только классы (кластеры), а и некоторую пограничную область между ними [5], в которой перемешиваются образы различных кластеров. Такая необходимость может возникнуть при обучении АС – с целью выбора параметров классификатора.

Для указанных приложений авторами разработан метод кластеризации с оценкой положения центров кластеров на базе вейвлет-преобразования (ВП) [2, 3]. В данной работе предложен метод кластеризации, который может позволить определять набор вложенных диапазонов изменения координат центров кластеров, на основе известного свойства ВП проводить пространственную обработку с регулируемой детальностью. Такая кластеризация позволяет определить образы, находящиеся в пограничной области между кластерами. Результаты такой кластеризации могут быть применены на следующем этапе – при классификации - для выбора достоверности классификатора, требуемой с точки зрения прагматической достаточности [6].

Цель работы. Для автоматизации обработки данных измерений в АС с обработкой визуальной информации разработан метод кластеризации, позволяющий проводить определение набора диапазонов координат центров кластеров с помощью мультистартовой оптимизации с ВП (МОВП) [2, 3]. Такая кластеризация позволяет определить образы, находящиеся в области пересечения кластеров.

Кластеризация на основе вейвлет-преобразования

При кластеризации определяют количество и состав групп параметров-признаков, при классификации – строят поверхности, разделяющие эти группы в признаковом пространстве. Проведенный анализ показал, что существующие методы кластеризации (иерархические и итеративные) отличаются низким качеством кластеризации при малых наборах зашумленных данных, с далеко отстоящими подгруппами образов в пространстве признаков, отсутствием априорной информации о форме кластеров. Для снижения влияния указанных недостатков разработан метод кластеризации на базе мультистартовой оптимизации с ВФ (вейвлет-функцией) [2, 3]. Метод МОВП реализуется по схеме

$$\mathbf{c}[n] = \mathbf{c}[n-1] - \gamma[n]\mathrm{WT}_k(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])), \quad (1)$$

где $\mathbf{Q}(\mathbf{x}, \mathbf{c})$ – функционал, зависит от вектора коэффициентов $\mathbf{c} = (c_1, \ldots, c_N)$ и от $\mathbf{x} = (x_1, \ldots, x_M)$; $\gamma[n]$ – шаг; n– номер итерации; k– номер старта;

$$WT_k(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])) = \{G_{1k}, G_{2k}, \ldots, G_{Nk}\} \quad (2)$$

определяет направление движения к экстремуму;

$$\sum_{i=-s_k/2, i\neq 0}^{s_k/2} \left(Q(\mathbf{x}[n], \mathbf{c}_j + ia)\cdot \Psi_k(i)\right) / s_k \ , \quad (3)$$

где s_k – длина носителя ВФ на k-м старте (s_k– четное число); a – шаг дискретизации; $\Psi_k(i)$ – ВФ на k-м старте (табл. 1); j = 1, …, N – размерность вектора параметров. Для оценки направления поиска оптимума в (2) выбраны симметричные и нестационарные ВФ [7], на первом этапе – ВФ вида

$$\Psi_1(i) = \begin{cases} 1, & \text{если } i = 1,\ldots, s_1/2; \\ -1, & \text{если } i = -1,\ldots, -s_1/2, \end{cases} \quad (4)$$

– на следующих этапах $\Psi_k(i) =$:

$$= \begin{cases} 1/\alpha_k(|i|+1), \text{ если } i > 0, \\ -1/\alpha_k(|i|+1), \text{ если } i < 0; \end{cases} i \in [-\frac{s_k}{2}, +\frac{s_k}{2}], i \neq 0; \quad (5)$$

(табл. 1), на седьмом старте –

$$\Psi_7(i) = \begin{cases} 1, \text{если } i = 1, \\ -1, \text{если } i = -1. \end{cases} \quad (6)$$

Таблица 1

Параметры ВФ для МОВП

Параметр	Значение					
k	2	3	4	5	6	7
Масштаб ВФ σ_k	1	2	3	4	5	–
s_k	20	10	6	4	4	2
$\Psi_k(i)$	(5)					(6)

При итеративном подходе к кластеризации определяют оптимальный вектор координат центров кластеров $\mathbf{c} = \mathbf{c}_{opt}$, который, удовлетворяя ограничениям, доставлял бы экстремальное значение $Q(\mathbf{x}, \mathbf{c})$ – функционалу вектора $\mathbf{c} = (c_1, \ldots, c_N)$, зависящему от вектора случайных последовательностей $\mathbf{x} = (x_1, \ldots, x_M)$. По образам $x \in X$ определяются центры множеств X_k и их границы. При этом

$$Q(\mathbf{x}, c_1, \ldots, c_M) = \sum_{k=1}^{M} \varepsilon_k(\mathbf{x}, c_1, \ldots, c_M) F_k(\mathbf{x}, c_1, \ldots, c_M) \ –$$

реализация функционала качества; $F_k(x, c_1, \ldots, c_M)$ – функция расстояния элементов $\mathbf{x}$ множества X от «центров» $\mathbf{c}_k$ подмножеств X_k (кластеров); $\varepsilon_k(\cdot)$ – характеристические функции,

$$\varepsilon_k(x, c_1, \ldots, c_M) = \begin{cases} 1, \text{когда } \mathbf{x} \in X_k, \\ 0, \text{когда } \mathbf{x} \notin X_k. \end{cases} \quad (7)$$

При малых наборах зашумленных данных и отсутствии априорной информации о форме кластеров существующие методы отличаются низким качеством кластеризации. Для снижения влияния указанных недостатков на основе метода Я.З. Цыпкина и Г.К. Кельманса [1] разработан метод кластеризации по дисперсионному признаку на базе МОВП с основными этапами (для 2 кластеров (r = 1,2)).

Этап 1. Формирование обучающей выборки.

Этап 2. Определение числа кластеров одним из известных методов [8,9].

Этап 3. Оценка функционала качества $Q(\mathbf{x}, \mathbf{c})$.

Этап 4. Задаются параметры метода МОВП: δ_1 – погрешность поиска оптимума старта, которая определяется на этапе априорных исследований функционала качества; δ_2 – погрешность, соответствующая требованиям прикладной задачи, которая определяется по показателю семантической (прагматической) достаточности в соответствии с адаптивно-критериальной системой показателей качества и эффективности [6]; $\delta\delta_j$, $j = \overline{1, jj}$ - погрешности определения диапазона, определются на этапе априорных исследований функционала качества; jj - максимальное количество диапазонов; k_{max} – максимальное количество стартов; вид ВП и ВФ; $\mathbf{c}_1[0]$ и $\mathbf{c}_2[0]$ – начальные значения координат центров кластеров; γ – шаг; a – шаг дискретизации ВФ; s_1 – длина носителя ВФ первого старта; $k = 1$; $n = 1$.

Этап 5. Оценивается направление поиска $WT_k(Q(\mathbf{x}[n], \mathbf{c}_1[n-1], \mathbf{c}_2[n-1]))$ для кластеров по (2) на итерации n. При первом старте и $k = 1$ для этого используется ВФ $\Psi_1(i)$ (в $\mathbf{c}[0]$ – при $n = 1$)). Длина носителя s_1 для $\Psi_1(i)$ определяется при анализе целевой функции. Интегральный характер такого ВП позволяет выделить сегмент целевой функции, где с высокой вероятностью находится глобальный оптимум, и определить диапазон изменения его координат. При проведении оценки диапазона для кластера проверяется знак (3). При изменении знака, на основании известного свойства оценок градиента менять знак при переходе через оптимум, определяется ряд вложенных диапазонов для координаты центра кластера. Максимальный диапазон определяется при $k = 1$ с $s = s_1$ - длиной носителя ВФ первого старта $\Psi_1(i)$, далее - с длиной носителя $\Psi_k(i)$, изменяющейся, согласно $s = s - sk$. Для этого используется обработка с ВФ $\Psi_1(i)$ как $\mathbf{c}[n] > \mathbf{c}^* > \mathbf{c}[n-1]$, если $\mathbf{c}[n] > \mathbf{c}[n-1]$ или $\mathbf{c}[n] < \mathbf{c}^* < \mathbf{c}[n-1]$, если $\mathbf{c}[n] > \mathbf{c}[n-1]$. При этом для каждого из i элементов (2) определяют характеристические функции $\varepsilon_r(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2)$, $r = 1,2$; если $f(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) < 0$, то $\varepsilon_1 = 1, \varepsilon_2 = 0$, если $f(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) \geq 0$, то $\varepsilon_1 = 0, \varepsilon_2 = 1$. Здесь

$$f(\mathbf{x}[n], \mathbf{c}_1[n], \mathbf{c}_2[n]) = \|\mathbf{x}[n]\text{-}\mathbf{c}_1[n]\|^2 - \|\mathbf{x}[n]\text{-}\mathbf{c}_2[n]\|^2 .$$

Этап 6. Определяются координаты центров кластеров на итерации n при $k \leq k_{max}$, иначе – останов

$$\begin{cases} \mathbf{c}_1[n] = \mathbf{c}_1[n-1] - \gamma WT_k(Q(\mathbf{x}[n], \mathbf{c}_1[n-1], \mathbf{c}_2[n-1]); \\ \mathbf{c}_2[n] = \mathbf{c}_2[n-1] - \gamma WT_k(Q(\mathbf{x}[n], \mathbf{c}_1[n-1], \mathbf{c}_2[n-1]). \end{cases} \quad (5)$$

Этап 7. Если $\|\mathbf{c}_r[n] - \mathbf{c}_r[n-1]\| \le \delta_1$, поиск на текущем старте заканчивается, иначе – $n = n + 1$ и осуществляется переход к этапу 5.

Этап 8. Если $k > 1$ и $\mathbf{c}^*_{r,k} - \mathbf{c}^*_{r,k-1} \le \delta_2$ – останов; в противном случае, если $k < k_{max}$, то увеличивается номер старта $k = k + 1$, выбирается вейвлет-функция для оценки направления поиска (см. табл. 1) и осуществляется переход к этапу 5. Здесь $\mathbf{c}^*_{r,k}$, $\mathbf{c}^*_{r,k-1}$ – результаты поиска для r-го кластера на k-м и k-1-м старте соответственно. При определении диапазонов для r-го кластера до проверки $\mathbf{c}^*_{r,k} - \mathbf{c}^*_{r,k-1} \le \delta_2$ проверяются условия перехода на следующий старт: $\mathbf{c}^*_{r,k} - \mathbf{c}^*_{r,k-1} \le \delta\delta_j$; $\mathbf{c}^*_{r,k} - \mathbf{c}^*_{r,k-1} > \delta\delta_{j+1}$.

Исследование проведено для синтезированной выборки из 25 образов с разделением на 2 кластера в двумерном пространстве признаков. При исследованиях были выбраны шаг $\gamma = 0,7$ и стартовая длина носителя ВФ $\Psi_1(i)$ $s_1 = 5$. При расчете для заданных значений погрешностей $\delta\delta_j$, $j = \overline{1,4}$ для центра кластера c_1 по x_1 получено 4 диапазона ($\Delta_j c_1$ с $j = \overline{1,4}$); первый (при обработке с ВФ $\Psi_1(i)$) $\Delta_1 c_1 = [7,52;\ 16,09]$ и четвертый (при обработке с ВФ $\Psi_6(i)$) с $s_k = 4$ $\Delta_4 c_1 = [13,005;\ 13,004]$. Далее определялась принадлежность образов к кластерам с c_2 и $c_1 = 7,52$ и с c_2 и $c_1 = 16,09$. Образов, изменяющих при этом принадлежность к кластерам – четыре. Для c_1 по x_2 и для c_2 на описанном этапе исследования диапазоны не определялись.

Выводы

Таким образом, в работе для автоматизации обработки данных измерений в АС разработан метод кластеризации. Этот метод позволяет проводить определение диапазонов координат центров кластера с помощью мультистартовой оптимизации с ВП. Такая кластеризация позволяет определить образы, находящиеся в области пересечения кластеров. На основании исследований сделан вывод, что такой метод может быть применен при необходимости выбора параметров классификатора с учетом требуемого уровня достоверности на этапе отладки АС и может быть рекомендован для применения в широком круге прикладных задач, отвечающих этим условиям.

Список литературы

1. Цыпкин Я.З. Адаптация и обучение в автоматических системах / Я.З. Цыпкин. – М.: Наука, 1968. – 400 с.

2. Shcherbakova G. Electronic apparatus automation inspection with adaptive clustering in wavelet domain / G. Shcherbakova, V. Krylov, S. Antoshchuk // Proc. of Int. Conf. CADSM'2009, Lviv, Ukraine. – 2009. – P. 153-154.

3. Wavelet transform domain adaptive clustering for electronic product quality inspection / G. Shcherbakova, V. Krylov, R. Pisarenko, V. Kuzmenko // Proc. of the 7th IEEE Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems, Berlin. – 2013. – V. 1. – P. 153-156.

4. Nithya N.S. A Survey on Clustering Techniques in Medical Diagnosis / N.S. Nithya, D.K. Duraiswamy, P. Gomathy // International Journal of Computer Science Trends and Technology. – 2013. – V. 1, Issue 2. – P. 17-22.

5. Коплярова Н.В. Об исследовании компьютерной системы диагностики электрорадиоизделий на основе данных испытаний / Н.В. Коплярова, В.И. Орлов // Вестник СибГАУ. – 2014. – С. 24-30.

6. Абакумов В. Распознавание визуальной информации в автоматизированных системах / В. Абакумов, С. Антощук, В. Крылов // Электроника и связь. – 2003. – № 19. – С. 46-48.

7. Krylov V.N. Contour images segmentation in space of wavelet transform with the use of lifting / V.N. Krylov, M.V. Polyakova // Optical-electronic informatively-power technologies. – 2007. – №2 (12). – P. 48-58.

8. Щербакова Г.Ю.Определение количества кластеров при прогнозировании состояния электронной аппаратуры / Г.Ю. Щербакова, В.Н. Крылов, С.Г. Антощук // Электроника и свіязь. – 2010. – № 3 (56). – С. 91 – 95.

9. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Изд-во инта математики, 1999. – 270 с.

Поступила в редколлегию 25.11.2015

Рецензент: д-р техн. наук, проф. С.Г. Антощук, Одесский национальный политехнический университет, Одесса.

ДОСЛІДЖЕННЯ АВТОМАТИЗОВАНОЇ КЛАСТЕРИЗАЦІЇ З ВИКОРИСТАННЯМ ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ

Г.Ю. Щербакова, В.Н. Крилов, Р.О. Писаренко, О.В. Логвінов

Запропонований метод кластеризації при обробці даних, який дозволяє визначити діапазони зміни координат центрів кластерів з використанням вейвлет-перетворення. Метод може бути використаний при виборі параметрів класифікатора з урахуванням потрібного рівня достовірності в автоматизованих системах обробки даних. При дослідженні методу визначені діапазони зміни координат центрів кластера для тестових даних.

Ключові слова: *класифікація, кластеризація, вейвлет-перетворення, автоматизовані системи, обробка даних, мультістартова оптимізація, діапазон.*

RESEARCH OF THE AUTOMATED CLASTERIZATION USING THE WAVELET TRANSFORM

G.Y. Shcherbakova, V.N. Krylov, R.A. Pisarenko, O.V. Logvinov

A method of clustering in time of data processing is designed. This method allow of clusters centre coordinates determination with wavelet transformation using. This method may be used for choice of classification parameters for select ranges of reliability in time of repairing of automated systems of data processing. The ranges of clusters centre coordinates for test data is determined in time of this method investigation.

Keywords*: classification, clustering, wavelet transform, automated systems, data processing, multi-start optimization, range.*