УДК 519.6:004.93

Г.Ю. ЩЕРБАКОВА, В.Н. КРЫЛОВ

Одесский национальный политехнический университет, Украина

АДАПТИВНАЯ КЛАСТЕРИЗАЦИЯ В ПРОСТРАНСТВЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Предложен субградиентный итеративный метод адаптивной кластеризации в пространстве вейвлет-преобразования, который позволяет повысить помехоустойчивость в процессе кластеризации. Разработана процедура реализации этого метода кластеризации, проведены экспериментальные исследования для оценки повышения помехоустойчивости метода и снижения его погрешности. Установлена помехоустойчивость метода. Относительная погрешность определения минимума при кластеризации для тестовой функции при отношении сигнал/шум по амплитуде 1,17 составила 8,32%. Разработанный метод рекомендуется к применению в широком круге задач классификации и кластеризации в случае зависимых, меняющихся с течением времени параметров, высоком уровне помех, малых объемах исследуемых выборок.

Ключевые слова: *адаптивная кластеризация, гиперболическое вейвлет-преобразование, шум, электронная аппаратура, контроль.*

Введение

Решение задач анализа растущего объема данных необходимо при контроле состояния электронной аппаратуры (ЭА) в процессе ее производства и эксплуатации. Состояние ЭА при контроле в процессе производства или эксплуатации описывается значением контролируемых параметров $X^n(t)$, которые под воздействием различных факторов изменяются от объекта к объекту и во времени

$$X^n(t) \in S_p^{(n)},$$

где $X^n(t) = \{X_1(t), X_2(t), \ldots, X_n(t)\}$ – модель ЭА, представленная n – мерным случайным вектором; $S_p^{(n)}$ – n – мерная допусковая область.

В связи с необходимостью оперативных мер при контроле по совокупности параметров необходим автоматизированный подход. Поэтому контроль проводится посредством наиболее отлаженного для такого случая метода – классификации при распознавании образов [1].

В любой фиксированный момент времени в группе контролируемой ЭА можно выделить компактные подгруппы (кластеры) с общими свойствами. Например, из-за сборки из комплектующих разных производителей, таким кластерам присуща разная степень монотонности зависимости контролируемых параметров от наработки в процессе эксплуатации. С течением времени может образоваться группа объектов, параметры которых приближаются к границе поля допуска.

Далее при контроле в процессе эксплуатации производится оценка параметров дрейфа центров кластеров во времени и их близости к границе поля допуска с целью прогноза работоспособности.

Оценить скорость изменения параметров ЭС, объединенных в один кластер, возможно, применяя для кластеризации адаптивный подход. В этом случае начальные параметры центров кластеров для последующего момента времени определяются из анализа в предыдущий момент времени.

Если нет указаний о том, к какому кластеру относится та или иная точка в признаковом пространстве (образ), но необходимо определить границы между кластерами, применяется классификация с самообучением, которая включает две процедуры: кластеризация и собственно классификация.

При кластеризации данные разделяются на кластеры по признаку компактности, так, чтобы был оптимизирован некоторый функционал качества.

Метод оптимизации выбирают в зависимости от особенностей формирования и свойств этого функционала качества, который может быть явно не известен, может обладать поверхностью многоэкстремальной, зашумленной, поскольку анализ производится по малым выборкам.

Существующие методы кластеризации в этих условиях работают плохо.

Для решения задач оптимизации при таких условиях разработан субградиентный итеративный метод оптимизации в пространстве вейвлет преобразования, который отличается повышенной помехоустойчивостью [2].

Для снижения влияния указанных выше недостатков и повышения помехоустойчивости в работе предлагается субградиентный итеративный метод адаптивной кластеризации в пространстве вейвлет преобразования.

Целью данной работы является разработка и исследование субградиентного итеративного метода адаптивной кластеризации в пространстве вейвлет-преобразования (ВП) для повышения помехоустойчивости.

Для достижения поставленной цели решены задачи:

– анализа современных методов кластеризации;

– разработки и обоснования субградиентного итеративного метода адаптивной кластеризации в пространстве ВП и процедуры реализации этого метода кластеризации;

– проведения экспериментальных исследований при оценке повышения помехоустойчивости метода.

1. Анализ современных методов кластеризации

Различают две группы методов кластеризации в зависимости от того, определено количество кластеров заранее или нет.

Первая группа методов кластеризации – иерархические методы с двумя стратегиями инициализации первоначального разбиения агломеративной (изначально каждый кластер содержит только один элемент – точку пространства признаков) и дивизивной (вначале все точки пространства признаков принадлежат одному кластеру). Объединение (разделение) точек прерывают, получив нужное число кластеров.

Алгоритмы реализации этих методов (k ближайших соседей, дальнего соседа) верно определяют кластеры, когда те компактны и хорошо разделены. Чтобы снизить чувствительность этих алгоритмов к отклонениям данных (шуму), вводят различные меры подобия вместо евклидовых расстояний между точками кластеров (например, углы). Однако такие меры подобия для ряда алгоритмов невозможно определить.

Общим недостатком иерархических методов является трудоемкость алгоритмов при большом объеме данных, а также, в зависимости от принятой меры расстояния, чувствительность к шуму [3].

Во второй группе методов – итеративных методах – элементы перемещаются между кластерами так, чтобы был минимизирован некоторый функционал качества.

Основные недостатки такого подхода – чувствительность к начальной точке поиска, чувствительность к шуму в данных, отыскивается локальный, а не глобальный минимум.

Поскольку обе группы методов чувствительны к шуму данных, целесообразно разрабатывать методы, которые смогут позволить снизить чувствительность к шуму данных при кластеризации.

2. Гипотезы компактности

Исходными данными для процедуры кластеризации является набор (выборка) объектов, заданных векторами своих характеристик в признаковом пространстве.

В связи с тем, что выборки при кластеризации небольшие, в данных могут быть ошибки, неинформативные, шумящие признаки, результат кластеризации зависит от того, какая гипотеза принята – гипотеза компактности либо гипотеза λ-компактности [4].

Гипотеза компактности состоит в том, что реализации одного и того же образа – точки в геометрическом признаковом пространстве – образуют «компактные» сгустки.

Меры компактности при этом могут быть различны. Например, объекты считаются компактными, если евклидово расстояние между векторами их признаков не превышает заданную величину [4].

Однако если в задаче кластеризации важны не только расстояния, но и отношения между ними, применяют гипотезу λ-компактности.

Эта гипотеза учитывает нормированное расстояние между элементами множества и характеристику локальной плотности множества в их окрестности.

Сложности связанные с ростом вычислительных затрат возникают, когда требуется получить число кластеров значительно большее, чем два [4].

С учетом сказанного выше, при необходимости повышения степени оперативности контроля при адаптивной кластеризации и отсутствии априорной информации о степени регулярности распределения данных в признаковом пространстве, в этой работе принимаем гипотезу компактности исходных данных в признаковом пространстве.

3. Функционалы при кластеризации

Задача кластеризации состоит в разбиении множества образов объектов на группы (кластеры) с учетом присущего им сходства. В метрическом пространстве сходство обычно определяют через расстояние.

Расстояние может рассчитываться как между объектами образов, так и от этих объектов к центру кластера.

Обычно координаты центров кластеров заранее не известны – они находятся одновременно с разбиением данных на кластеры.

Наиболее часто при кластеризации используется функционал [3]

$$Q(\mathbf{x}, \mathbf{c}) = \sum_{k=1}^{M} \sum_{x \in X_k} \left\| \mathbf{x} - \mathbf{c}_k \right\|^2 ,$$

где $\mathbf{c}_k = \frac{1}{n_k} \sum_{x \in X_k} \mathbf{x}$ – среднее k -го кластера;

n_k – число элементов в нем.

Здесь функционал $Q(\mathbf{x}, \mathbf{c})$ измеряет общую квадратичную ошибку, вносимую при представлении данных посредством k кластеров с центрами $\mathbf{c}_k$. Такой функционал считается подходящим, когда предполагается деление на два-три кластера достаточно хорошо отделенных друг от друга.

Но в случае зашумленных данных, с отстоящими далеко подгруппами точек, применяют родственные функционалы минимума дисперсии, в частности, функционал, использующий $\overline{s}_k$ – среднеквадратическое расстояние между точками k-ого кластера, или заменяют $\overline{s}_k$ медианой или максимальным расстоянием между точками в кластере [3].

Для дальнейших исследований при отсутствии априорной информации о форме исследуемых кластеров, принимается функционал из группы, родственных функционалу минимума дисперсии [5].

В процессе кластеризации данные разделяются на кластеры так, чтобы был оптимизирован некоторый функционал качества. Этот функционал может быть явно не известен, может обладать поверхностью многоэкстремальной, зашумленной.

Для решения задач оптимизации при таких условиях разработан субградиентный итеративный метод оптимизации в пространстве ВП, который отличается повышенной помехоустойчивостью [2].

4. Субградиентный итеративный метод адаптивной кластеризации

Для повышения помехоустойчивости кластеризации при контроле в процессе производства или эксплуатации ЭА на основе субградиентного итеративного метода оптимизации разработан субградиентный итеративный метод адаптивной кластеризации в пространстве ВП.

При итеративном подходе к кластеризации определяют оптимальный вектор $\mathbf{c} = \mathbf{c}_{opt}$, который, удовлетворяя ограничениям, доставлял бы экстремальное значение $Q(\mathbf{x}, \mathbf{c})$ – функционалу вектора переменных $\mathbf{c} = (c_1, \ldots, c_N)$, зависящему от вектора случайных последовательностей $\mathbf{x} = (x_1, \ldots, x_M)$.

По показам образов $x \in X$ определяются центры множеств X_k и их границы. При этом:

$$Q(\mathbf{x}, c_1, \ldots, c_M) =$$

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

реализация функционала качества;

$$F_k(x, c_1, \ldots, c_M) \quad -$$

функция расстояния элементов $\mathbf{x}$ множества X от «центров» $\mathbf{c}_k$ подмножеств X_k (кластеров);

$\varepsilon_k(\cdot)$ – характеристические функции,

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

Для двух кластеров поисковый регулярный итеративный алгоритм кластеризации для определения значений центров кластеров c_1^* и c_2^*

$$\begin{cases} c_1[n] = \\ = c_1[n-1] - \gamma_1[n] \tilde{\nabla}_{c1+} Q(\mathbf{x}[n], c_1[n-1], c_2[n-1]); \\ c_2[n] = \\ = c_2[n-1] - \gamma_2[n] \tilde{\nabla}_{c2+} Q(\mathbf{x}[n], c_1[n-1], c_2[n-1]), \end{cases}$$

где $\gamma_k[n]$ – величина шага;

n – номер итерации;

$\tilde{\nabla}_{c1+} Q(x[n], c_1[n-1], c_2[n-1])$ – оценка градиента реализации для первого кластера;

$\tilde{\nabla}_{c2+} Q(x[n], c_1[n-1], c_2[n-1])$ – оценка градиента реализации для второго кластера;

k – номер кластера.

По реализациям функционала качества $Q(\mathbf{x}, \mathbf{c})$ оценивается градиент

$$\nabla_c Q(\mathbf{x}, \mathbf{c}) = \left(\frac{\partial Q(\mathbf{x}, \mathbf{c})}{\partial c_1}, \frac{\partial Q(\mathbf{x}, \mathbf{c})}{\partial c_2} \right).$$

Но, если в условиях помех оценка градиента проводится разностным методом [5], поисковый регулярный итеративный метод, используемый при кластеризации, также даст низкую помехоустойчивость. Это обусловлено низкой помехоустойчивостью оценки градиента разностным методом.

Адаптивный метод кластеризации в пространстве ВП заключается в следующем.

Для кластеризации в каждый момент времени

– инициализируются параметры метода;

– для каждого из i элементов взвешенной суммы с ВП определяют значение характеристических функций $\varepsilon_l(x, c_1, c_2), l = 1,2$, входящих в оценку субградиента. Для этого по методике [5] пары значений

$$c_1[n-1], c_2[n-1];\ c_1[n-1] \pm ie_1 a[n], c_2[n-1];$$

$$c_1[n-1], c_2[n-1] \pm ie_2 a[n]\ \ (i = \overline{1, N})$$

при данном $x[n]$ подставляют в

$$f(x, c_1, c_2) = \|x[n]-c_1\|^2 - \|x[n]-c_2\|^2 .$$

Здесь N – длина носителя вейвлет-функции; $a[n]$ – скаляр.

Функция $f(x, c_1, c_2)$ равна нулю на границе и имеет различные знаки в различных областях. Поэтому, если значение $f(x, c_1, c_2)$ отрицательно, $\varepsilon_1 = 1, \varepsilon_2 = 0$, если положительно, $\varepsilon_1 = 0, \varepsilon_2 = 1$ [5].

В качестве базового для оценки градиента был использован градиентный метод [6].

Исходные данные для его работы: начальное значение координаты минимума, начальное значение шага $\gamma = 1$, коэффициент, обусловливающий изменение шага γ вблизи минимума $\beta = 0,5$, точность определения оценки градиента ε, количество итераций j.

Процедура вычисления минимума при кластеризации включает:

– вычисление оценки градиента;

– если значение оценки градиента меньше заданного значения точности ε - останов;

– вычисление величины шага: задается начальное значение величины шага $\gamma = 1$; вычисляется вспомогательное значение приращения функции Δ, если приращение функции Δ меньше нуля – $\gamma[n] = \gamma$ и переход к следующему этапу, иначе $\gamma[n] = \beta\gamma$ и переход к предыдущему этапу;

– расчет координаты минимума на n-ой итерации,

– $n = n + 1$ и переход к начальному этапу вычисления минимума при кластеризации.

При вычислении оценки градиента на каждой итерации на первом этапе вычисляется взвешенная сумма значений минимизируемого функционала $Q(x[n], c_1[n-1], c_2[n-1])$ с вейвлет-функцией Хаара. Это позволяет переместить поиск в район экстремума с погрешностью, определяемой асимметрией этого функционала.

На втором этапе оценки градиента при кластеризации вычисляется взвешенная сумма минимизируемого функционала $Q(x[n], c_1[n-1], c_2[n-1])$ с гиперболической функцией $\Psi(i) = \frac{1}{\alpha x}$ при начальном масштабе $\alpha = 0,5$:

$$HWT(c[n]) = Q(x[n], c_1[n-1], c_2[n-1]) * \Psi(i),$$

где * – операция взвешенного суммирования.

Далее, после определения оценки градиента, определяют приближение к значению координаты центра кластера, используя итеративный алгоритм в пространстве гиперболического ВП по схеме

$$c_1[n+1] = c_1[n] + \gamma[n] HWT(c[n]),$$

где $HWT(c[n])$ – значение взвешенной суммы с вейвлет-функцией в точке $c[n]$;

$\gamma[n]$ – шаг.

Если найденная на этом этапе координата оптимума отличается от координаты оптимума, найденной на предыдущем этапе не более, чем на δ, процесс поиска заканчивается. Здесь δ – заданная точность поиска координаты оптимума.

Для оценки субградиента использовано гиперболическое вейвлет-преобразование (ГВП), полученное по лифтинговой схеме [7].

На каждом уровне поиска координаты оптимума значение масштаба α увеличивается в соответствии с $\alpha = \{0,5; 1; 2; 3; 4; 5\}$. Если условие окончания поиска координаты оптимума при значении величины $\alpha = 5$ не достигается, оценка субградиента производится разностным методом. После этого поиск заканчивается.

В процессе поиска координаты оптимума осуществляется последовательный переход от поиска координаты оптимума с помощью вейвлета Хаара, способного обеспечить высокую помехоустойчивость, вплоть (с ростом α) до поиска с помощью дифференциатора, способного дать максимальную точность.

Далее проверяют вышеописанное условие точности определения центра кластера, если он достигнут – останов для заданного временного шага. Для последующих временных шагов начальные параметры центров кластеров – определяются из анализа на предыдущем шаге.

5. Эксперименты по методу кластеризации

Как отмечено выше, на первом этапе метода кластеризации для определения оценки градиента используется взвешенное суммирование значений минимизируемой функции с вейвлет-функцией Хаара.

Это позволяет переместить поиск в район экстремума с погрешностью, определяемой асимметрией целевой функции в этой области.

Для оценки зависимости относительной погрешности определения экстремума в зависимости от асимметрии была синтезирована функция

$$a(i)=\begin{cases} -i^2+100\cdot i, \text{ при } i\le 50; \\ \dfrac{a(50)\cdot(50+shag-i)}{shag}, \text{ при } 50<i\le 50+shag; \\ 0; \text{ при } i>50+shag, \end{cases}$$

где shag – параметр для изменения асимметрии функции, $i=\overline{1,100}$.

Асимметрия α функции $a(i)$ определялась как отношение оценки ее третьего центрального момента к кубу оценки среднего квадратичного отклонения для значений $i=\overline{1,50+shag}$.

Точность определения координаты экстремума функции $a(i)$ оценивалась для трех значений асимметрии $(\alpha_1=-0,557, \alpha_2=-0,4136, \alpha_3=-0,2705)$ для значений половины длины носителя вейвлет-функции от одного до десяти.

По результатам такой оценки можно сделать вывод, что относительная погрешность определения координаты экстремума асимметричной функции при взвешенном суммировании с вейвлетом Хаара прямо пропорциональна коэффициенту асимметрии целевой функции в области поиска.

Оценка помехоустойчивости метода кластеризации проводилась с использованием функции $f(x)=x^2$ при значениях ее аргумента $x=1,\dots 80$. Помеха была распределена по нормальному закону с нулевым средним и среднеквадратическим отклонением, изменявшимся в диапазоне от 0 до 5477, максимальное значение тестируемой функции было принято $f(x)=6400$..

При отношении сигнал/шум по амплитуде 1,17 (помеха распределена по нормальному закону с нулевым средним и среднеквадратическим отклонением 5477, максимальное значение тестируемой функции $f(x)=6400$ относительная погрешность определения минимума составила 8,32%.

Выводы

Разработан и обоснован адаптивный субградиентный итеративный метод кластеризации в пространстве вейвлет-преобразования:

– разработана процедура реализации этого метода кластеризации;

– проведены экспериментальные исследования для оценки повышения помехоустойчивости метода и снижения его погрешности.

Установлена помехоустойчивость адаптивного субградиентного итеративного метода кластеризации в пространстве вейвлет-преобразования:

– относительная погрешность определения минимума для тестовой функции при отношении сигнал/шум по амплитуде 1,17 составила 8,32%.

Эти результаты позволяют рекомендовать разработанный метод адаптивной кластеризации к применению в широком круге практически важных задач классификации и кластеризации при контроле электронной аппаратуры в процессе ее производства или эксплуатации в случае зависимых и меняющихся с течением времени параметров, при высоком уровне помех и при малых объемах исследуемых выборок.

Литература

1. Зубарев В.В. Вплив дефектів функціональних матеріалів на надійність електроніки / В.В. Зубарев, С.В. Лєнков, В.А. Мокрицький, Д.О. Перегудов. – Одеса: Друк, 2003. – 452 с.

2. Крилов В. Н. Субградієнтний ітеративний метод оптимізації в просторі вейвлет-перетворення / В.Н. Крилов, Г.Ю. Щербакова // Збірник наукових праць Військового інституту Київського національного університету ім. Т. Шевченка. – К., 2008. – Вип. 12. – С. 56-60.

3. Дуда Р. Распознавание образов и анализ сцен / Р. Дуда, П. Харт – М.:Мир,1976. – 509 с.

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

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

6. Полак Э. Численные методы оптимизации. Единый подход. – М.:Мир, 1976. – 509 с.

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.

Поступила в редакцию13.02.2009

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

АДАПТИВНА КЛАСТЕРИЗАЦІЯ В ПРОСТОРІ ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ

Г.Ю. Щербакова, В.М. Крилов

Запропонований субградієнтний ітеративний метод адаптивної кластеризації в просторі вейвлет-перетворення, який дозволяє підвищити завадостійкість в процесі кластеризації. Розроблена процедура реалізації цього метода кластеризації, проведені експериментальні дослідження для оцінки підвищення завадостійкості метода і зниження його похибки. Встановлена завадостійкість метода. Відносна похибка визначення мінімуму при кластеризації для тестової функції при відношенні сигнал/шум по амплітуді 1,17 склала 8,32%. Розроблений метод рекомендується до використання в широкому колі задач класифікації і кластеризації у випадку залежних параметрів, які змінюються з часом, при високім рівні завад, малих об'ємах вибірок, що досліджуються.

Ключові слова: адаптивна кластеризація, гіперболічне вейвлет-перетворення (ГВП), шум, електронна апаратура, контроль.

ADAPTIVE CLUSTERING IN WAVELET DOMAIN

G.Y. Shcherbakova, V.N. Krylov

The adaptive clustering in hyperbolic wavelet transforming domain is designed and justified. The implementation procedure for adaptive clustering in hyperbolic wavelet transforming domain was worked up. The experimental investigation for this clustering method noise immunity increasing and own error decreasing estimation was carried out. The method noise immunity was established: test function relative error minimum definition be equal to 8,32 % in case when amplitude signal-to-noise ratio be equal to 1,17. That results allow recommend the adaptive clustering in the wavelet domain for applying in clustering and classification tasks in case of parameters correlation and those changing in a course of time, high level of noise and small samples.

Key words: Adaptive clustering, hyperbolic wavelet transforming (HWT), noise, electronic apparatus, inspection.

Щербакова Галина Юрьевна – канд. техн. наук, доцент, доцент кафедры электронных средств и информационно-компьютерных технологий Одесского национального политехнического университета, Одесса, Украина, e-mail: asg@ics.opu.ua

Крылов Виктор Николаевич– д-р техн. наук, проф., проф. кафедры информационных систем Одесского национального политехнического университета, Одесса, Украина, e-mail: asg@ics.opu.ua