УДК 519.6:004.93

Г.Ю. Щербакова, канд. техн. наук, В.Н. Крылов, д-р техн. наук, С.Г. Антощук, д-р техн. наук

Определение количества кластеров при прогнозировании состояния электронной аппаратуры

Предложен субградиентный иерархический метод выбора числа кластеров при прогнозировании параметров в условиях зашумленных выборок данных с учетом локальной плотности их распределения.

Hierarchical sub-gradient method for the quantity cluster choice is proposed. This method in the noised data cases with local density of this data selection is applied for the parameters prediction.

Введение

При производстве электронной аппаратуры (ЭА) важно сократить длительность производственных испытаний [1]. Одним из основных способов для этого является автоматизированное прогнозирование изменения параметров изделий во времени для оценки момента наступления отказа. Если не обязательно прогнозировать значение параметров, а достаточно прогнозировать класс, к которому они будут относиться, такое прогнозирование осуществляется с помощью статистической классификации [1]. Одно из основных преимуществ такого прогноза – прогнозирование по малым выборкам, позволяющее понизить его стоимость и повысить оперативность, является и недостатком. Этот недостаток связан с тем, что ошибки оператора, сбои оборудования при измерении обусловливают зашумленность выборок при малом их объеме. А это требует дополнительных итеративных приближений при классификации в процессе прогноза, что обусловливает снижение его оперативности и достоверности. Для проведения классификации необходимо знать количество классов. Так как на ранних этапах прогноза параметров, как правило, отсутствуют сведения о структуре данных, первым этапом классификации при прогнозировании является кластеризация – выделение среди ЭА компактных подгрупп (кластеров) с общими свойствами. Поэтому статистическая классификация при прогнозировании включает две процедуры: кластеризация и собственно классификация. Методы кластеризации используются для разбиения ЭА на кластеры, а методы классификации – для принятия решения при прогнозировании параметров [1]. При кластеризации в первый момент контроля среди ЭА выделяются кластеры с близкими параметрами. Эти параметры, как правило, медленно изменяются во времени. Поэтому оценку их изменения предлагается проводить адаптивным методом, когда начальные параметры центров кластеров для последующего момента времени определяются из анализа в предыдущий момент времени.

Изменение выходных параметров у ЭА в начальный период эксплуатации приводит к одному из трех видов рассеяния [2]: незначительное изменение математических ожиданий и почти линейное изменение во времени среднеквадратического отклонения (СКО); систематическое смещение центра рассеяния с монотонным возрастанием ширины поля рассеяния во времени за счет роста СКО; поле рассеяния с одним или несколькими экстремумами математического ожидания - из-за использования или хранения в условиях со значительными изменениями влияющих факторов. Для каждого из трех видов рассеяния при контроле из-за изменения СКО и математического ожидания параметров со временем кластеры могут сближаться, поскольку, например, из-за сборки ЭА из комплектующих разных производителей или разных условий хранения, таким кластерам присуща разная степень монотонности зависимости параметров от наработки. Поскольку прогноз проводится по малым, зашумленным выборкам, а СКО и математическое ожидание параметров изменяются со временем, что вызывает изменение границ кластеров вплоть до пересечения, повышение помехоустойчивости и снижение погрешности процедуры разбиения на кластеры в первый момент контроля в значительной степени обусловливает помехоустойчивость и погрешность прогноза в целом.

При определении количества кластеров в обеих основных группах методов кластеризации – итеративных и иерархических при подходах с четким и с нечетким разбиением - из-за зашумленности и малых выборок исходных данных возникает ряд проблем.

Количество кластеров определяют путем поиска экстремума функционалов, учитывающих компактность распределения данных в кластере и удаленность различных кластеров [3, 4]. В условиях зашумленной выборки такой функционал может быть многоэкстремальным, и при поиске

экстремума возникают проблемы, связанные с недостаточной помехоустойчивостью, высокой погрешностью, чувствительностью к локальным экстремумам и начальной точке поиска методов оптимизации. Чтобы снизить влияние этих проблем, ограничивают количество кластеров, что требует участия экспертов [4, 5].

Выбор часто используемых при кластеризации функционалов качества – с минимальной дисперсией - оправдан, если данные образуют достаточно отдаленные друг от друга группы с приблизительно одинаковым числом элементов. Однако в результате нарушений технологических процессов производства, испытаний, хранения ЭА, среди их параметров могут появиться случайные, зашумленные группы данных. Тогда число данных в разных группах может сильно разниться, и при использовании функционала с минимальной дисперсией может случиться, что большая группа будет разделена [3]. То есть разделение данных уже на первом этапе классификации – кластеризации не будет соответствовать их реальной структуре, вызывая рост погрешности при последующем прогнозировании. В случае зашумленных, с далеко отстоящими подгруппами данных может не подтверждаться и гипотеза компактности [5]. Чтобы решить эти проблемы, при кластеризации учитывают не абсолютные значения, а отношения между расстояниями в признаковом пространстве, то есть применяют гипотезу λ- компактности.

Но гипотеза λ- компактности разработана для иерархического метода кластеризации, не ориентированного на адаптивный подход. Этому методу свойственны низкая помехоустойчивость и резкое возрастание с ростом количества классифицируемых объектов и числа кластеров времени вычислений, поскольку определение числа кластеров в этом случае проводят с использованием полного перебора вариантов оценок относительных расстояний между элементами кластера [5]. С учетом того, что расчеты λ- компактности также требуют дополнительных вычислительных затрат, и необходимость проведения кластеризации для всех моментов контроля, указанные недостатки могут понизить соответственно помехоустойчивость и оперативность прогноза в целом.

При кластеризации данные разделяют на кластеры по признаку компактности или λ- компактности, так, чтобы был оптимизирован функционал качества. Метод оптимизации выбирают с учетом свойств этого функционала, который может обладать поверхностью многоэкстремальной, зашумленной, так как прогнозирование производится по малым выборкам. Методы кластеризации, основанные на градиентном поиске оптимума функционала качества, не обеспечивают достаточной помехоустойчивости, а использующие при поиске оптимума оценку субградиента, отличаются высокой погрешностью. В связи с этим для решения задач оптимизации в таких условиях разработан субградиентный итеративный метод оптимизации в пространстве вейвлет - преобразования (ВП), с повышенной помехоустойчивостью и низкой погрешностью, пониженной чувствительностью к локальным экстремумам и стартовой точке поиска [7]. На основе этого метода был разработан метод адаптивной кластеризации в пространстве ВП [8]. Для снижения влияния указанных недостатков, в [1] для сокращения длительности испытаний ЭА было предложено применить этот метод.

Целью данной работы является разработка метода выбора числа кластеров при прогнозировании параметров в условиях зашумленных выборок данных с учетом локальной плотности их распределения при помощи помехоустойчивого мультистартового субградиентного итеративного метода оптимизации в пространстве вейвлет-преобразования (ВП) с пониженной чувствительностью к локальным экстремумам и стартовой точке поиска [7]. Для достижения поставленной цели решены задачи анализа методов выбора числа кластеров и разработки метода выбора числа кластеров при прогнозировании и процедуры реализации этого метода.

Анализ методов выбора числа кластеров

Решение проблемы определения числа кластеров актуально для обеих групп методов кластеризации – иерархических и итеративных. Общим недостатком иерархических методов является трудоемкость алгоритмов при большом объеме данных, а также, в зависимости от принятой меры расстояния, низкая помехоустойчивость [3]. Во второй группе методов - итеративных методах - элементы перемещаются между кластерами так, чтобы был минимизирован функционал качества. Общие недостатки методов – чувствительность к начальной точке поиска и к шуму в данных, отыскивается локальный, а не глобальный минимум.

В связи с разнообразием методов кластеризации, зависимостью получаемых результатов от принятых гипотез и мер компактности, начальной точки поиска, чувствительности к шуму в данных и т.п. предложены различные подходы для выбора количества кластеров и оценки результата кластеризации.

Оценивают, насколько разбиение на кластеры соответствует данным, с помощью крите-

рия хи-квадрат и статистики Колмогорова – Смирнова [3]. Этот подход оправдан для больших наборов данных, но не для малых, зашумленных выборок. В итеративных методах, где кластеризация производится достижением экстремума критерия или функционала качества, основанного на (1) [3], повторяют процедуры кластеризации для различного числа кластеров, оценивая изменение функции критерия. Такой подход применим, если данные разделяются на компактные, хорошо разделенные кластеры.

Если известен диапазон для искомого числа кластеров, проводят кластеризацию на $c_{\min}, \mathbf{K}, c_{\max}$ кластеров. Варианты оцениваются по критерию $J_3 = J_1 - qJ_2$. $J_1 -$ функционал средней близости точек в кластерах определяют через потенциальную функцию $K(x_i, x_j)$ близости точек x_i и x_j, $J_2 = \frac{2}{c(c-1)} \sum_{i=1}^{c} \sum_{j>i} K(A_i, Aj)$ определяют с учетом

$$K(A_i, A_j) = \frac{2}{N_i N_j} \sum_{x_l \in A_i} \sum_{x_p \in A_j} K(x_l, x_p).$$

Число кластеров тогда выбирается как $c_{opt} = c_j / \max J_3(c_j)$, с учетом $c_j = c_{\min}, \mathbf{K}, c_{\max}$, и корректируется с помощью экспертов [4]. Однако не всегда возможно заранее определить диапазон $c_{\min}, \mathbf{K}, c_{\max}$ значений количества кластеров и привлечь к корректировке результата экспертов.

Общие недостатки итеративных методов определения числа кластеров - в случае малых, зашумленных выборок данных, и поскольку такой функционал может быть полимодальным, возможности такого подхода ограничены помехоустойчивостью и чувствительностью к локальным экстремумам метода поиска оптимума критерия.

Исходными данными для процедуры кластеризации является набор (выборка) объектов, заданных векторами своих характеристик в признаковом пространстве. В связи с тем, что выборки при кластеризации при классификационном прогнозировании параметров небольшие, в данных могут быть ошибки, неинформативные, шумящие признаки, результат кластеризации зависит от того, какая гипотеза принята - гипотеза компактности либо гипотеза λ - компактности [5].

Гипотеза компактности состоит в том, что реализации одного и того же образа отражаются в признаковом пространстве в геометрически близкие точки, образуя «компактные» сгустки. Мера компактности при этом может характеризоваться средним расстоянием от центра тяжести до всех точек кластера, средней длиной ребра графа кратчайшего незамкнутого пути (КНП), соединяющего точки одного образа и т.д.

При кластеризации данные разделяются на кластеры так, чтобы был оптимизирован некоторый функционал качества. Наиболее часто при кластеризации используется функционалы из группы с минимальной дисперсией, например [3]

$$Q(x,c) = \sum_{k=1}^{M} \sum_{x \in X_k} \| x - c_k \|^2, \qquad (1)$$

где n_k - число элементов в кластере k, $c_k = \frac{1}{n_k} \sum_{x \in X_k} x$ - среднее этого кластера, $Q(x, c)$ - функционал вектора переменных $c = (c_1, \mathbf{K}, c_N)$, зависящий от вектора случайных последовательностей $x = (x_1, \mathbf{K}, x_M)$.

Но в случае зашумленных данных, при большом различии между количеством элементов в кластерах, гипотеза компактности может не подтвердиться, и при использовании (1) больший кластер будет разделен [3]. При кластеризации в таких условиях учитывают не только расстояния между элементами, но и отношения между ними, то есть применяют гипотезу λ - компактности. Эта гипотеза учитывает нормированное расстояние между элементами кластера и локальную плотность множества в их окрестностях. Граница кластера при этом проходит по участку графа КНП с наибольшим значением характеристики λ [5].

Гипотеза λ - компактности разработана для иерархических методов кластеризации. Для них характерны резкое возрастание с ростом числа классифицируемых объектов и кластеров времени вычислений, поскольку разделение на k кластеров проводят путем перебора вариантов, определяемых как число сочетаний из $(m-1)$ объектов по $(k-1)$ [5] и низкая помехоустойчивость. Эта гипотеза позволяет формировать кластеры с приблизительно одинаковым количеством объектов, что может служить для повышения помехоустойчивости. Но при прогнозе состояния ЭА, когда структура данных не обязательно представляет равные группы, может быть потеряна полезная информация.

Сокращают время вычислений λ - компактности, увеличивая количество этапов кластеризации, либо вводя критерии качества, позволяющие избежать полного перебора вариантов [6].

Так в алгоритме $\lambda - KRAB - 2$ кластеризация проводится в два этапа [5]. На первом этапе в евклидовом пространстве проводят кластеризацию на $k' > k$ кластеров. Центры этих кластеров используются в качестве вершин графа КНП в λ - пространстве. При этом из-за применяемых алгоритмов кластеризации в евклидовом пространстве остаются проблемы зависимости результатов от начальной точки поиска и чувствительности к локальным экстремумам функционала качества.

В λp - алгоритме [6] граф КНП строится с учетом вероятности разрыва его ребра

$$p_{iz} = \frac{\lambda_{iz}}{\sum_{j=1}^{k} \lambda_{iz}}. \qquad (2)$$

Здесь λ_{iz} - λ-расстояние, соответствующее i - му ребру, связанному с вершиной z в λ - КНП; k - количество ребер, связанных с z. При формировании λ - КНП вершина z объединяется с той из соседних, вероятность разрыва связывающего ребра с которой минимальна. Количество кластеров выбирают путем исследования критерия $K = \max(k_i)$,

$$k_i = \frac{f(i+1)}{f(i)}, \quad i \in [1, n-1], \qquad (3)$$

где $f(i)$ - среднее λ - расстояние на $i-$ом шаге алгоритма.

Параметр k_i позволяет оценивать нарушения локальной однородности распределения элементов в λ - пространстве.

Достоинством такого подхода является существенный (по сравнению с алгоритмом $\lambda - KRAB$ [5]) выигрыш во времени из-за формирования λ - КНП за один проход алгоритма [6], однако результат кластеризации остается зависимым от помехоустойчивости и чувствительности к локальным экстремумам метода поиска оптимума критерия.

Таким образом, можно констатировать, что методы выбора количества кластеров чувствительны к шуму в исходных данных, к локальным экстремумам функционала и начальной точке поиска, недостаткам функционалов с минимальной дисперсией. Поэтому, и в связи с затратами иерархических алгоритмов кластеризации при работе в λ - пространстве, в работе предлагается повысить оперативность, помехоустойчивость, снизить погрешность, чувствительность к локальным экстремумам и начальной точке поиска этой процедуры при классификационном прогнозировании параметров.

Метод выбора числа кластеров при прогнозировании

Метод выбора числа кластеров при прогнозировании предполагает следующие этапы.

Этап 1. Отображение параметров элементов выборки контролируемой ЭА из евклидова пространства в λ - пространство по методике [5] предполагает:

1) построения полного графа в евклидовом пространстве;

2) расчет нормированного расстояния между всеми парами точек множества параметров

$$d_i = \frac{\alpha_i}{D},$$

где αi - расстояние между i - ой парой точек в полном графе; D - диаметр графа (самое длинное ребро);

3) расчет характеристики локальной плотности множества в окрестности i -го ребра

$$\tau_i = \frac{\alpha_i}{\beta_{i\min}\tau_{\max}},$$

где $\beta_{i\min}$ - длина самого короткого ребра; $\tau_{\max}$ - наибольшее в графе значение $\tau_i^* = \frac{\alpha_i}{\beta_{i\min}}$.

4) расчет длин ребер графа в λ - пространстве как $\lambda_i = \tau_i^2 \times d_i$.

Этап 2. Построение λ - КНП - графа кратчайшего незамкнутого пути в λ - пространстве по (2) [6].

Этап 3. Рассчитывается показатель качества $k_j = \frac{f(j+1)}{f(j)}$ по (3). Здесь $j \in [1, n-1]$ - номер итерации [6].

Этап 4. Производится поиск максимума критерия $K = \max(k_i)$ с помощью помехоустойчивого мультистартового субградиентного итеративного метода оптимизации в пространстве вейвлет-преобразования (ВП) [7].

В результате исходное множество объектов разделяется на два кластера. Если, исходя из содержательной постановки задач классификации и прогнозирования этого достаточно, поиск количества кластеров (классов при дальнейшем прогнозировании) окончен. В связи с тем, что классификационный анализ проводится для разделения ЭА по классам долговечности или работоспособности, число классов, как правило, в пределах 3…5 [4]. При необходимости разделения исходных данных на количество кластеров более двух, проводится дальнейшее разделение полученных кластеров вышеописанным методом (этап 4).

Этап 5. Для каждого кластера определяются координаты его центра (как оценка математического ожидания) и с помощью субградиентного мультистартового метода кластеризации в пространстве ВП [8] уточняется граница между кластерами. Эти данные используются как исходные на этапе адаптивной кластеризации при формировании данных для прогнозирования параметров ЭА [1].

Выводы

Таким образом, в работе предложен субградиентный иерархический метод выбора числа кластеров при прогнозировании параметров в условиях зашумленных выборок данных с учетом локальной плотности их распределения. Этот метод позволил за счет включения этапа кластеризации в λ- пространстве понизить влияние недостатков применяемых при этом функционалов с минимальной дисперсией, повысить оперативность процедуры выбора за счет применения усовершенствованного λp -алгоритма [6], и повысить помехоустойчивость, снизить погрешность, чувствительность к локальным экстремумам и начальной точке поиска с помощью помехоустойчивого мультистартового субградиентного итеративного метода оптимизации в пространстве вейвлет-преобразования (ВП) [7].

Литература

1. *Shcherbakova G.* Electronic Apparatus Parameters Prediction with Adaptive Clustering Procedure / G. Shcherbakova, V. Krylov // Proc. of International Conf. CADSM'2009. – Lviv, Ukraine. - 2010.
2. *Недоступ Л.А.* Технологические методы управления качеством радиоэлектронных измерительных устройств / Л.А Недоступ, Е.Т.Удовиченко, Г.А. Шевцов.- М.: Изд. стандартов, 1976. – 124 с.
3. *Дуда Р.* Распознавание образов и анализ сцен / Р. Дуда, П. Харт. Пер. с англ. Г.Г. Вайнштейна и А.М.Васьковского. Под ред. В.Л. Стефанюка. - М. : Мир, 1976. – 511 с.
4. *Дорофеюк Ю.А.* Методы структурно-классификационного прогнозирования многомерных динамических объектов / Ю.А. Дорофеюк, А.А. Дорофеюк //Искусств. интел.-2006.- № 2. – С. 138 – 141.
5. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. - Новосибирск: Изд-во ин-та математики, 1999. – 270 с.
6. *Юдін С.А.* Метод формування образів в задачах інтелектуального аналізу даних /С.А. Юдін. Автореферат дисертації на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23. – Одеса:ОНПУ. – 2006. – 18 с.
7. *Крилов В.Н.* Субградієнтний ітеративний метод оптимізації в просторі вейвлет-перетворення/ В.Н.Крилов, Г.Ю. Щербакова // Збірн. наук. праць Військ. ін-ту Київського нац. ун-ту ім. Т. Шевченка - Вип. 12.- 2008. - С. 56-60.
8. *Щербакова Г.Ю.* Адаптивная кластеризация в пространстве вейвлет-преобразования / Г.Ю. Щербакова, В.Н.Крылов. // Радіоелектронні і комп'ютерні системи. –2009. - № 6. – С. 123 – 127.