Прогнозирование временного ряда на основе штрафного сплайна и алгоритма кластерного анализа DBScan в реальном времени

Е.Ю. Репина[1], Е.А. Кочегурова[1]

[1]Томский политехнический университет, пр. Ленина 2, Томск, Россия, 634028

Аннотация. Прогнозирование в режиме реального времени представляет особую ценность: с помощью данных, полученных при таком прогнозе можно оперативно и объективно отражать постоянно обновляемую картину любого процесса. Модель прогнозирования, предлагаемая в данной работе, основана на рекуррентном P-сплайне. Один из основных параметров модели, существенно влияющих на результат прогноза - число измерений звена сплайна. Настройка данного параметра осуществляется с помощью плотностного алгоритма кластеризации DBScan. Данный алгоритм был модифицирован под решаемую задачу: задание верхней границы количества элементов в кластере, решение проблемы перекрытия кластеров, создание правил для применения алгоритма в режиме реального времени.

1. Введение

В настоящее время задача прогнозирования является актуальной практически для всех предметных областей и сфер деятельности. Частным случаем задачи прогнозирования является прогнозирование временных рядов, как одной из популярных форм представления информации. Разработано множество моделей для решения задачи прогнозирования временного ряда, среди которых наиболее часто используют авторегрессионные и нейросетевые модели. При этом особую ценность представляют методы интерпретации данных, поступающих в режиме реального времени (РРВ). После поступления текущие данные обрабатываются и система формирует отклик еще до прибытия новой порции данных.

Один из отличительных подходов к задаче прогнозирования является использование в качестве модели прогнозирования аппроксимирующих кусочно-полиноминальных функций, заменяющих регрессионные соотношения при моделировании и анализе процессов. С помощью подобного аппарата одновременно выполняется фильтрация и сглаживание данных с шумом, что значительно повышает точность любой задачи интерпретации данных.

В данной работе модель прогнозирования представлена в виде штрафного сплайна, параметры которого настраиваются с помощью алгоритма кластерного анализа DBScan.

2. Описание модели

В данной работе предлагается комбинированный подход к прогнозированию временного ряда, включающий следующие этапы

- На первом этапе производится настройка параметров сглаживающего P-сплайна с помощью кластеризации;

- На втором этапе будет получен штрафной P-сплайн с оптимальными параметрами (полученными на предыдущем этапе) и с прогнозом на 1 или несколько дискретных значений вперед.

Пересчет всего набора данных для настройки параметров при поступлении каждой новой точки в режиме реального времени является слишком ресурсоемким и длительным процессом. Для адаптации данной модели прогнозирования к режиму реального времени необходимо выделить некоторое число последних кластеров (LC), данные которых совместно с новоприбывшим значением будут являться начальным набором данных для повторного запуска работы алгоритма кластеризации.

3. Описание математического аппарата

Одной из форм представления цифровых данных являются временные ряды. Сглаживание временных рядов лежит в основе многих прикладных задач.

Применение сплайна позволяет получить искомое гладкое решение [1]. Сглаживающий сплайн S(t) основан на оптимизации специального вида функционала и представлен для РРВ следующей формулой:

$$J(S) = (1-\rho)(h\Delta t)^2 \int_{t_0^i}^{t_h^i} [S''(t)]^2 dt + \rho \sum_{j=0}^{h} [S(t_j^i) - y(t_j^i)]^2 \qquad (1)$$

где ρ∈ [0, 1] — весовой коэффициент, устанавливающий баланс между сглаживающими и интерполяционными свойствами сплайна $S(t)$; Δt — интервал дискретизации наблюдаемого процесса; h — количество измерений внутри i-го звена сплайна, далее h = const для всех звеньев сплайна на интервале наблюдения данных.

Все параметры сплайна оказывают определенное влияние на его свойства. Наиболее значимый вклад в точность аппроксимации вносит параметр h. Современные методы оптимизации, включая биоинсперированные и другие метаэвристики, позволяют получить только постоянное значение параметра h для всего интервала наблюдения. Однако часто динамика наблюдений меняется, и постоянные значения настроек приводят к снижению качества прогноза, а соответственно к повышению погрешности.

Одним из возможных способов улучшения прогноза является повторяющийся кластерный анализ, разбивающий временной ряд в РРВ на сегменты переменной длины h.

4. О методах кластеризации

Кластеризацией называют разбиение множества объектов на группы (кластеры). По данным, получаемым на выходе, все алгоритмы кластеризации принято делить на иерархические и неиерархические. Также, существует классификация алгоритмов кластеризации по принципам кластеризации: итеративные, плотностные, сетевые, модельные и концептуальные. Наиболее применимы итеративные и плотностные.

Итеративные алгоритмы кластеризации – предполагают пошаговое перераспределение объектов между классами. Одним из популярных представителей данного семейства алгоритмов является алгоритм k-means. Основная идея данного алгоритма - пошаговая минимизация расстояний между объектами в кластерах. Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Главным недостатком данного алгоритма является необходимость заранее задавать желаемое количество кластеров, что недопустимо в РРВ. Плотностные алгоритмы кластеризации - алгоритмы, определяющие кластер как группу объектов, расположенных кучно, т.е. так, чтобы в окрестности объекта находилось минимально заданное число других объектов (соседей). Представителем этого класса алгоритмов кластеризации является алгоритмы DBScan, OPTICS.

5. Описание алгоритма DBSCAN

Алгоритм **DBScan (**Density-based spatial clustering of applications with noise) допускает кластеризацию пространственных данных в присутствии шума. Алгоритм был предложен в

1996 году М. Эстером и его коллегами для разбиения данных на кластеры произвольной формы. Для работы алгоритма используются два входных параметра: ε-окрестность, в которой будет требоваться наличие минимального количества объектов Minpts. Данный параметр ограничен минимально возможным числом измерений для построения звена сплайна, и, следовательно, не может быть уменьшен. Увеличение этого параметра в задачах РРВ нелогично.

Классическая реализация алгоритма является неприменимой для поставленной задачи прогнозирования по ряду причин. Алгоритм не реализует ограничение количества элементов сверху, что приводит к неприемлемо большим для точного сглаживания кластерам. Также, алгоритм выделяет и отбрасывает "шума", что недопустимо для задач прогнозирования, т.к. это может привести к потере важной информации.

Данные аспекты были учтены при модификации алгоритма: реализована возможность задания верхней границы числа элементов в кластере, реализовано выделение "шума" в отдельные кластеры (при условии обнаружения неразрывного участка данных, содержащего более 3-х элементов внутри).

Было проведено исследование данного алгоритма в задаче оценки параметра h в условиях работы временного ряда при наличии шума, как и в случае реальных временных рядов в РРВ.

После реализации алгоритма DBScan, была проведена его апробация на модельных данных. В качестве тестовых примеров выбраны следующие функции:

$$f_1(t) = 10 \cdot \sin\left(\frac{2\pi \cdot i}{100}\right) \quad (2)$$

$$f_2(t) = \sin\left(\frac{\pi \cdot i}{20}\right) \cdot \mathrm{e}^{0.02 \cdot t} + 3 \quad (3)$$

На данные, подготовленные для указанных функций, был наложен случайный шум в (10-20) % от диапазона полезной функции.

Результаты кластеризации при варьировании параметров представлены в следующих таблицах 1 и 2.

Таблица 1. Результаты кластеризации для функции (2).

ε	Количество кластеров, шт	Процент выброса, %
(0.1 -2)	0	100
2.1	6	37,5
2.2	4	12,5
2.3	3	1
(2.4-2.5)	2-1	1

Таблица 2. Результаты кластеризации для функции (3).

ε	Количество кластеров, шт	Процент выброса, %
(0.1 -1.9)	0	100
2	9	40,5
2.1	6	24
2.2	3	17
2.3	3	8
(2.4 - 2.6)	2	6-5
2.7 - …	1	1 – 2

Анализ параметров алгоритма показал, что при малых окрестностях все данные воспринимаются алгоритмом, как выбросы (шум). При увеличении данного параметра происходит уменьшение числа кластеров. Задание верхней границы числа элементов в кластере решает эту проблему.

Результаты кластеризации при максимальном количестве кластеров представлены на рисунках 1 и 2.

В целом графические результаты показывают, что алгоритм DBScan разбивает временной ряд логически правильно для построения звеньев сплайна.

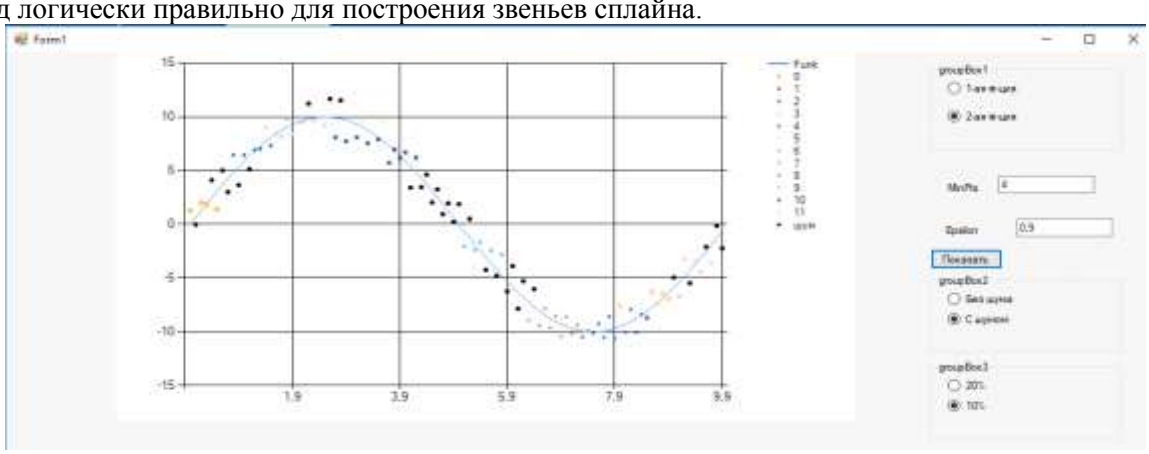

Рисунок 1. Результат работы алгоритма DBScan для функции (2).

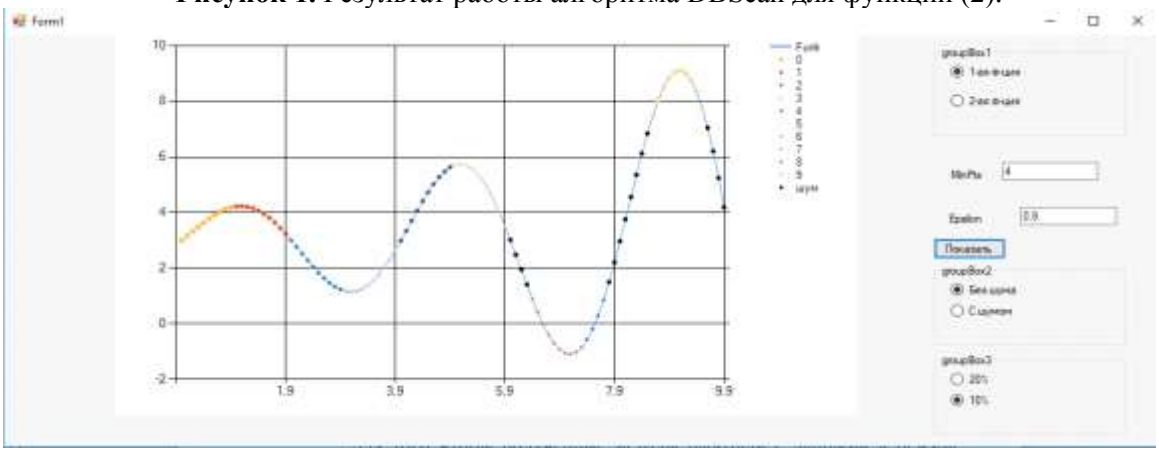

Рисунок 2. Результат работы алгоритма DBScan для функции (3).

6. Результаты прогноза

Результаты проверки полученного метода на модельных данных представлены далее:

Таблица 3. Результаты прогнозирования.

$\sigma_\xi, \%$	0%	5%	10%	15%	20%
$y_1(t) = 10 \cdot \sin\left(\frac{2\pi \cdot t}{100}\right)$					
h_{opt}	3	4	6	3	4
$\mathrm{RMSPE}_{\min}$	0.26	1.26	1.69	2.76	3.22
$y_2(t) = \sin\left(\frac{\pi \cdot t}{20}\right) \cdot e^{0.02 \cdot t} + 3$					
h_{opt}	9	3	3	3	3
$\mathrm{RMSPE}_{\min}$	0.56	2.72	4.24	5.31	5.92

7. Заключение

На данном этапе исследований получено удовлетворительное качество прогноза при использовании описанной модели прогнозирования (среднее значение погрешности 7%). Далее предполагается модернизация данной модели для улучшения результатов прогноза.

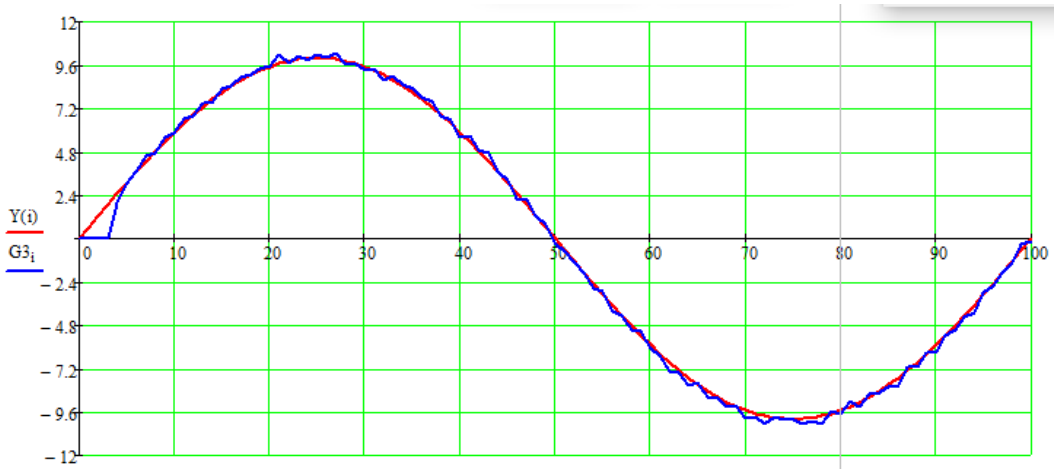

Рисунок 3. Результат прогнозирования для функции для функции (2).

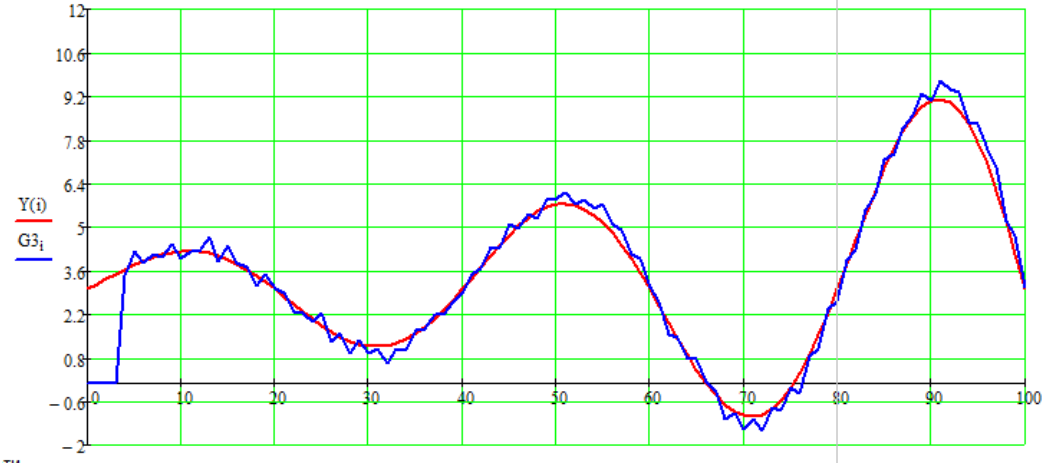

Рисунок 4. Результат прогнозирования для функции для функции (3).

8. Благодарности

Работа выполнена при поддержке РФФИ (грант № 18-07-01007).

9. Литература

[1] Кочегурова, Е.А. Текущее оценивание производной нестационарного процесса на основе рекуррентного сглаживающего сплайна / Е.А. Кочегурова, Е.С. Горохова // Автометрия. – 2016. – Т. 52, № 3. – С. 79-85.

[2] Плотностный алгоритм кластеризации пространственных данных с присутствием шума — DBSCAN [Электронный ресурс]. – Режим доступа: https://habr.com/post/143151 (12.10.2018).

[3] Iorio, C. Parsimonious time series clustering using P-splines / C. Iorio, G. Frasso, A. D'Ambrosio // Expert Systems with Applications. – 2016. – Vol. 52. – P. 26-36.

[4] Аверкин, А.Н. Прогнозирование временных рядов на основе гибридный нейронных сетей / А.Н. Аверкин, С.А. Ярушев // Наука и образование. – 2016. – № 12. – С. 233-246.

[5] Abdoos, A. Short term load forecasting using a hybrid intelligent method / A. Abdoos, M. Hemmati, A.A. Abdoos // Knowledge-Based Systems. – 2015. – Vol. 76. – P. 139-147.

Time series prediction based on the penalty spline and the real-time DBScan cluster analysis algorithm

E. Repina[1], E. Kochegurova[1]

[1]Tomsk Polytechnic University, Lenin Ave. 2, Tomsk, Russia, 634028

Abstract. Real-time forecasting is particularly valuable: with the help of the data obtained with such a forecast, it is possible to quickly and objectively reflect a constantly updated picture of any process. The prediction model described in this paper is based on the recurrent P-spline. One of the main parameters of the model, which significantly affect the result of the forecast, is the parameter h-the number of measurements of the spline link. The setting of this parameter is performed by using the density clustering algorithm DBScan. This algorithm was modified for the problem to be solved: setting the upper limit of the number of elements in the cluster, solving the problem of overlapping clusters, creating rules for applying the algorithm in real time.