

*Дуля И.С.,
студент магистратуры
1 курс, Институт прикладной математики
и компьютерных наук
Томский государственный университет
Россия, г. Томск*

ПРИМЕНЕНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ К ЗАДАЧЕ КЛАСТЕРИЗАЦИИ ВРЕМЕННЫХ РЯДОВ

***Аннотация:** Появление новых открытых наборов данных подталкивает исследователей к созданию новых алгоритмов и подходов к решению задачи кластеризации временных рядов. Достижения последних десятилетий в области глубокого обучения и компьютерного зрения служат мотивацией к применению подходов глубокого обучения к задаче кластеризации временных рядов. В статье на примере одной из задач рассмотрена эффективность применения свёрточного автоэнкодера к задаче кластеризации временных рядов. Также проведено сравнение рассматриваемого подхода с базовыми методами решения задачи.*

***Ключевые слова:** глубокое обучение, кластеризация, извлечение признаков, автоэнкодер, обучение без учителя.*

***Annotation:** New open datasets are appeared and push researchers to creation new algorithms and approaches to solving the time-series clusterization problem. Achievements of recent years in deep learning and computer vision motivate using deep learning approaches to time-series clusterization problem. The article discusses the efficiency of applying a convolutional autoencoder to the time-series clusterization problem. We also compare the deep learning approach with the basic methods for solving the problem.*

Key words: *deep learning, clusterization, feature extraction, autoencoder, unsupervised learning.*

Активное внедрение концепции четвертой промышленной революции, где все производство завязано на сборе и анализе данных, подогревает интерес к теме анализа временных рядов и задаче кластеризации временных рядов, в частности. Исследователи активно разрабатывают новые подходы и алгоритмы решения этой задачи. Примечательно, что исследователи не уделяют должного внимания подходам на основе глубокого обучения [1]. Однако достижения в этой области могут быть эффективно применены и к задаче кластеризации временных рядов.

Подход на основе глубокого обучения предполагает использование автоэнкодера в связке с преобразованием, позволяющем перейти от временной области к двумерной дискретной функции или матрице. В качестве такого преобразования может использоваться непрерывное вейвлет преобразование [2]. Вейвлет-преобразование (англ. Wavelet transform) — интегральное преобразование, которое представляет собой свертку вейвлет-функции с сигналом. Непрерывное вейвлет преобразование предполагает, что используемая вейвлет функция является непрерывной, что позволяет её масштабировать на любой действительный коэффициент. Непрерывное вейвлет преобразование осуществляется по следующей формуле:

$$cwt(\tau, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \psi\left(\frac{t - \tau}{s}\right) dt,$$

где s – коэффициент масштабирования, τ – параметр сдвига, $x(t)$ – исходный сигнал, $\psi(\cdot)$ – вейвлет-функция. Описанное преобразование основано на концепциях масштабирования и смещения. Масштабирование представляет собой растяжение или сжатие вейвлета на коэффициент масштабирования. Сдвиг представляет перемещение вейвлетов с различным масштабом от

начала до конца сигнала. Сжатые вейвлеты помогают фиксировать резкие изменения сигнала, а растянутые – медленные изменения.

Рассмотрим задачу определения дефекта двигателя по его звуку в качестве задачи, на которой будет сравниваться эффективность подхода на основе глубокого обучения с базовыми подходами. Набор данных FordB [3] был создан для проведения конкурса на Всемирном конгрессе по искусственному интеллекту в 2008 году. Набор содержит 3636 наблюдение, длина каждого временного ряда составляет 500 значений.

Далее на этапе подготовки данных к временным рядам было применено непрерывное вейвлет преобразование. В качестве вейвлета был выбран вейвлет Морле, что касается диапазона масштабов, то было рассмотрено три различных диапазона: 28,54 и 122. На основе анализа вейвлет-скалограмм было решено взять диапазон масштабов от 1 до 28.

Автоэнкодеры — это нейронные сети прямого распространения, которые восстанавливают входной сигнал на выходе. Внутри у них имеется скрытый слой, который представляет собой код, описывающий модель. Свёрточные автоэнкодеры имеют свёрточные слои и слои субдискретизации. В таблице 1. представлено описание слоёв выбранного свёрточного автоэнкодера.

Таблица 1.

Описание слоёв свёрточного автоэнкодера

Номер	Тип слоя	Размер	Число	Номер	Тип слоя	Размер	Число
1	InputLayer	(28,28,1)	0	8	Conv2D	(4,4,8)	584
2	Conv2D	(28,28,16)	160	9	UpSampling	(8, 8, 8)	0
3	MaxPooling	(14,14,16)	0	10	Conv2D	(8, 8, 8)	584
4	Conv2D	(14,14,8)	1160	11	UpSampling	(16,16, 8)	0
5	MaxPooling	(7,7,8)	0	12	Conv2D	(14, 14,	1168
6	Conv2D	(7,7,8)	584	13	UpSampling	(28, 28,	0
7	MaxPooling	(4,4,8)	0	14	Conv2D	(28, 28, 1)	145

Далее была обучена модель автоэнкодера на 15 эпохах. Обученный энкодер был использован для перехода от матриц вейвлет-коэффициентов в пространство латентных признаков, на основе которых далее была обучена модель k-means для двух кластеров (есть дефект, нет дефекта). Полученная матрица несоответствий представлена на рисунке 1. Точность составила 70,27%, что достаточно неплохо для обучения без учителя.

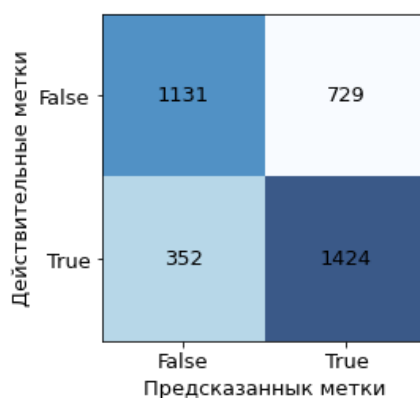


Рисунок 1. Матрица несоответствий

Для подтверждения качества рассмотренного метода необходимо его сравнить с другими моделями/методами классификации временных рядов. В качестве базовых методов кластеризации были взяты три метода: кластеризация по исходным временным рядам без извлечения признаков, кластеризация по признаковому представлению рядов и кластеризация на основе метрической модели (k-means) с метрикой расстояния DTW. В таблице 2 приведены сводные результаты для всех четырёх подходов.

Таблица 2.

Сводные результаты оценки точности моделей

	Accuracy	Precision	Recall	Fscore
WT + AE	0,703	0,720	0,678	0,698
Raw TS + k-means	0,536	0,564	0,446	0,498
Feature-based TS + k-means	0,606	0,574	0,654	0,611
k-means + DTW	0,581	0,581	0,611	0,595

Подход на основе глубокого метода обучения показал самую высокую точность, причем прирост точности по отношению к лучшему подходу из базовых составил почти 10%. Следует заметить, что полученные результаты не говорят о том, что подход на основе глубокого обучения в любой задаче будет показывать более высокий результат. Это в существенной степени зависит от данных и самой задачи. Преимущество данного подхода заключается в относительной универсальности этапа подготовки данных, он будет практически идентичен для различных задач, что снижает влияние опыта аналитика на конечный результат. Возможность работы с многомерными временными рядами еще одно преимущество моделей глубокого обучения [4]. Здесь, сигнал с каждого отдельного датчика подаётся в виде отдельного канала (по аналогии с RGB каналами цветного изображения). Соответственно, все зависимости между сигналами с различных датчиков могут учитываться одновременно, что очень важно при работе с многомерными временными рядами.

В задачах машинного обучения критическую роль играет объем обучающей выборки. В настоящих прикладных задачах существуют ограничения на её размер, возникающие по различным причинам (стоимость, небольшое число наблюдаемых объектов и тд.). Используемый автоэнкодер содержит несколько тысяч параметров, исходя из чего можно предположить, что данный подход потребует достаточно большой объем обучающей выборки. С этим связано ограничение в применении глубоких методов обучения в реальных задачах. Часто предпочтение отдают более простым моделям, которые содержат минимальное число параметров. В связи с этим, критически важно показать, что основной метод может работать и в условиях ограниченного объема обучающих данных. Результаты эксперимента приведены на рисунке 2.

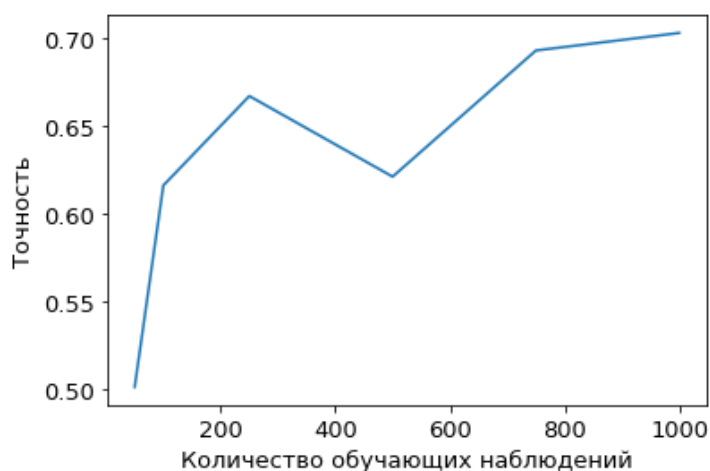


Рисунок 2. Зависимость точности от числа обучающих наблюдений

Несмотря на то, что выбранная модель имеет большое число параметров, для её обучения требуется сравнительно небольшое число наблюдений, после 250 наблюдений точность перестаёт сильно увеличиваться.

В данной работе было проведено исследование эффективности применения глубоких автоэнкодеров в связке с непрерывным вейвлет-преобразованием для решения задачи кластеризации временных рядов. Был описан алгоритм подготовки данных и обучения моделей на примере задачи определения дефекта двигателя. Было проведено сравнение подхода с базовыми алгоритмами решения задач классификации временных рядов. Полученные результаты показывают, что глубокое обучение может обеспечить существенный прирост точности по сравнению с базовыми алгоритмами решения задачи. Было показано, что подход, основанный на глубоком обучении и непрерывном вейвлет-преобразовании легко можно применять в реальных задачах, когда существенно ограничен размер обучающей выборки.

Использованные источники:

1. Lines J. HIVE-COTE: The hierarchical vote collective of transformation-based ensembles for time series classification / J. Lines, S. Taylor, A. Bagnall // ACM Transactions on knowledge discovery from data. – 2018. – № 12.
2. Hurley C. Wavelet, Analysis and Methods / C. Hurley, J. Mclean – Waltham Abbey: ED-Tech press, 2018. – 280 p.
3. Dataset: FordA [Электронный ресурс]: набор данных / Time-series classification repository. – URL: <http://www.timeseriesclassification.com/description.php?Dataset=FordA> (дата обращения 20.04.2021)
4. Bellman R. Dynamic Programming / R. Bellman. – Princeton: Princeton University Press, 2010. – 392 p.