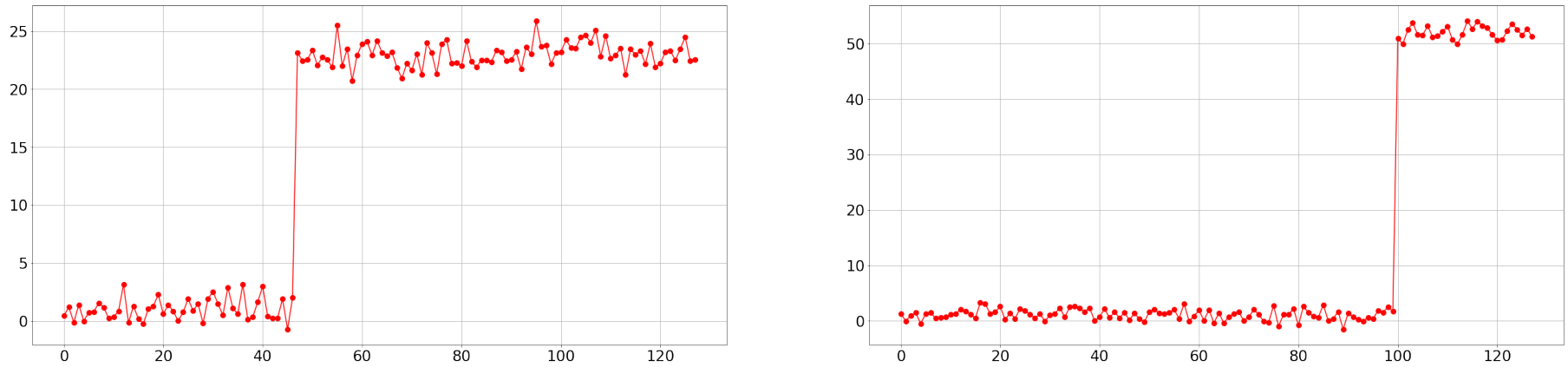
Детектирование разладок на основе обучения представлений с использованием размеченных и неразмеченных данных

Андрей Веприков Александр Степикин Евгения Романенкова Алексей Зайцев

Весна 2023

Введение

Разладка во временном ряде – это момент резкой смены вероятностного распределения составляющих его данных.



Примеры разладок во временных рядах

Причиной разладок в реальной жизни часто становятся чрезвычайные происшествия, которые необходимо быстро и точно детектировать. Кроме того, обнаруженные изменения дают ценную информацию для построения моделей в дальнейшем. Разладки, например, встречаются в

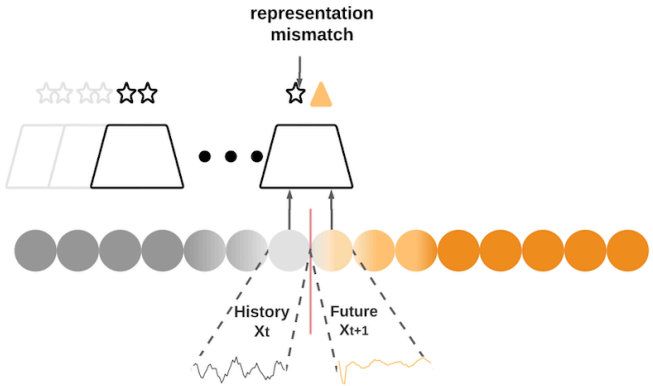
1. данных бурения скважин [3]
2. видеонаблюдении [4]
3. мониторинге медицинских показателей [1]

Введение

1. проблема: размеченных данных не хватает для построения качественной модели
2. решение: использовать semi-supervised методы
3. цель работы: получение semi-supervised подхода для детектирования разладок

Метод TS-CP2

В качестве основного подхода к детектированию был использован контрастивный unsupervised метод TS-CP2 [2].



Общая схема метода TS-CP2

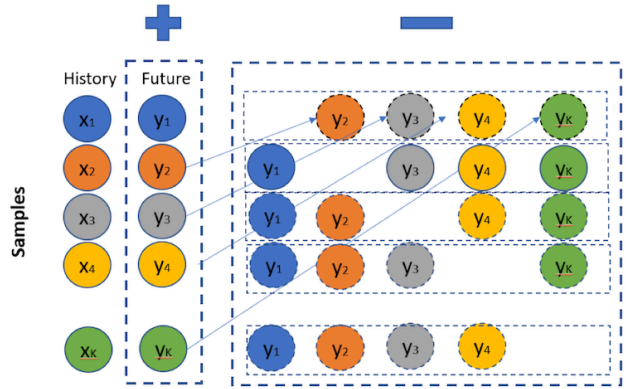


Схема формирования батча

Вероятность p_i положительной пары объектов h_i, f_i вычислялась как

$$p_i = \frac{\exp\left[Sim(h_i, f_i)/\tau\right]}{\sum_{j=1}^{K} \exp\left[Sim(h_i, f_j)/\tau\right]},$$

где $Sim(\cdot, \cdot)$ – косинусное расстояние, а τ – параметр модели.
Итоговая функция потерь вычислялась как

$$L = -\sum_{i,j}^{K} \delta_{ij} \log(p_i) + (1 - \delta_{ij}) \log(1 - p_i)$$

Методы, применяемые в работе

В работе рассматриваются 2 способа калибровки модели:

1. итоговая модель обучается с помощью комбинированной функции потерь:

$$L = \alpha \cdot L_s + (1 - \alpha) \cdot L_u,$$

где L_s – значение функции потерь, полученное для пар из размеченных данных, L_u – из неразмеченных

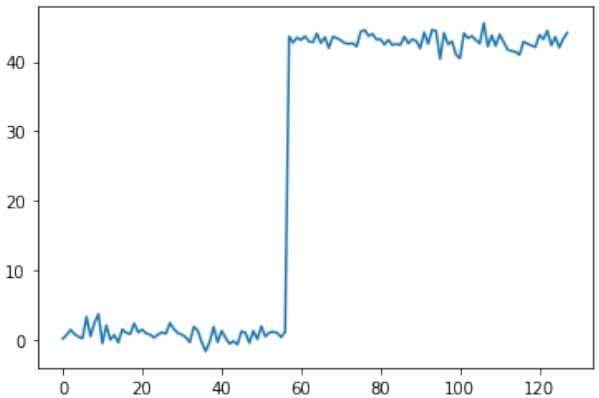
2. модель сначала полноценно обучается на неразмеченных данных, а затем дообучается (fine-tuning) на размеченных в течение небольшого числа эпох

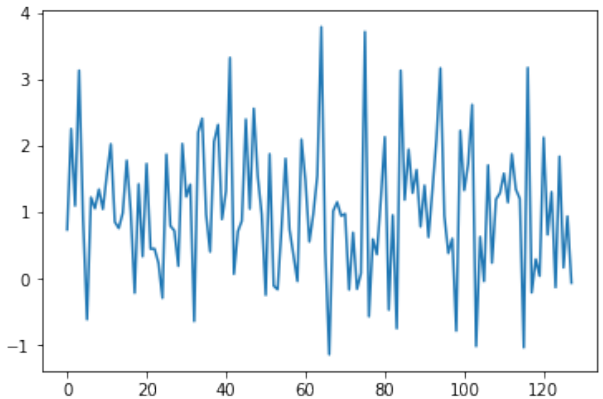
Synthetic 1D

В качестве простейшей выборки были сгенерированы последовательности длины 128, состоящие из одномерных гауссовских случайных величин.
Они имеют вид

$$x_1, ..., x_\theta \sim N(0, 1) \qquad x_{\theta+1}, ..., x_{128} \sim N(\mu, 1),$$

где $\mu \in \{1, ..., 100\}$ – случайное математическое ожидание, а $\theta \in \{2, ..., 127, 128\}$ – случайный момент разладки (если $\theta = 128$, значит, разладки нет).

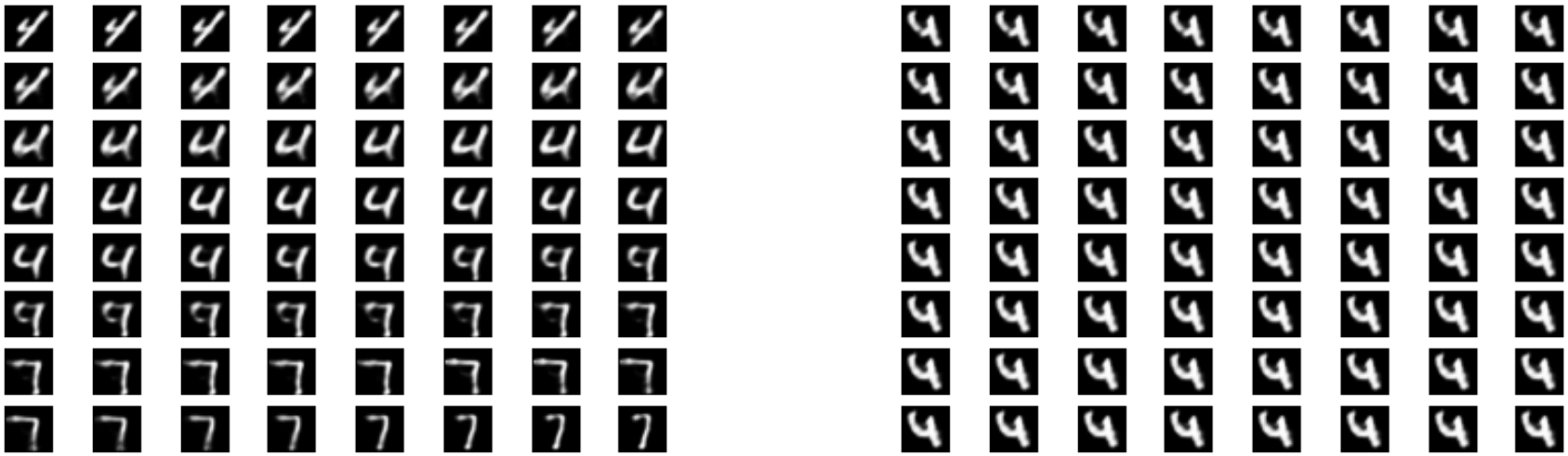




Примеры последовательностей из Synthetic 1D: с разладкой (слева) и без разладки (справа)

MNIST

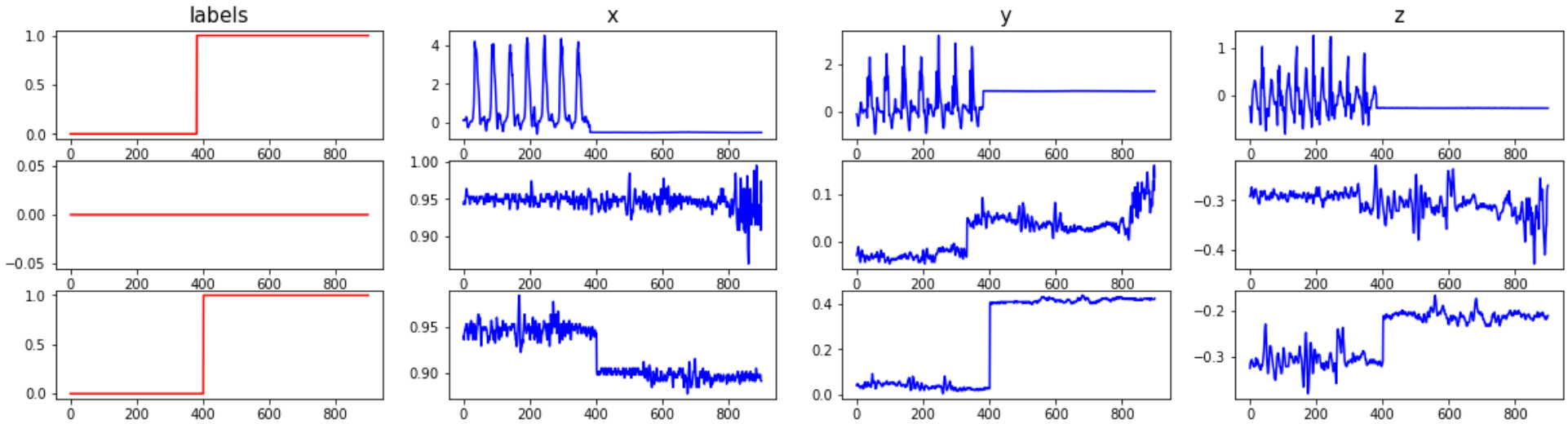
Данная выборка была получена на основе базы данных MNIST, которая представляет собой набор изображений рукописных арабских цифр размера 28×28 пикселей.
Последовательности с разладкой содержат плавный переход от одной цифры в к другой, нормальные последовательности состоят из изображений одной и той же цифры.



Примеры последовательностей из MNIST: с разладкой (слева) и без разладки (справа)

USC-HAD

Набор данных USC-HAD [5] включает последовательности измерений датчиков устройств, носимых человеком. Разладкой в таких последовательностях является момент смены одного из 12 видов человеческой деятельности на другой, например «бег», «ходьба», «сон» и так далее.
Размерность каждого отдельного наблюдения равна 6. Эта выборка является несбалансированной: примерно 85% последовательностей содержат разладку.



Примеры последовательностей из USC-HAD: с разладкой (1 и 3 строки) и без разладки (2 строка)

Варьирование α и supervised num

В первом эксперименте мы одновременно варьировали весовой коэффициент α и s – размер размеченной выборки. Размер неразмеченного датасета был фиксирован – 500 последовательностей для Synthetic 1D и MNIST, 250 и 500 для USC-HAD.

В качестве метрики качества была использована метрика F1-Score. В таблицах жирным выделен лучший результат, при использовании неразмеченных данных, то есть при $\alpha \neq 0$.

	s = 5	s = 10	s = 15	**s = 25**	s = 50	s = 75
α = 0	0.8950 ± 0.0108	0.9056 ± 0.0225	0.9239 ± 0.0091	0.9369 ± 0.0192	0.9406 ± 0.0142	0.9322 ± 0.0219
α = 0.3	0.8642 ± 0.0404	0.9087 ± 0.0097	0.9045 ± 0.0100	0.9154 ± 0.0117	0.9245 ± 0.0017	0.9183 ± 0.0084
α = 0.5	0.8914 ± 0.0193	0.8988 ± 0.0125	0.8900 ± 0.0118	**0.9198 ± 0.0032**	0.8525 ± 0.0420	0.8775 ± 0.0340
α = 0.7	0.7989 ± 0.1254	0.8882 ± 0.0423	0.7502 ± 0.1500	0.7602 ± 0.0953	0.7336 ± 0.0276	0.7831 ± 0.0626

Synthetic 1D

	s = 15	s = 20	s = 25	**s = 30**
α = 0	0.8620 ± 0.0482	0.8610 ± 0.0699	0.9538 ± 0.0119	0.9612 ± 0.0193
α = 0.3	0.6758 ± 0.0882	0.7126 ± 0.0102	0.7123 ± 0.0584	0.6801 ± 0.0340
α = 0.5	0.6495 ± 0.0408	0.7152 ± 0.0279	0.7253 ± 0.0195	**0.7193 ± 0.0301**
α = 0.7	0.6845 ± 0.0260	0.6665 ± 0.0278	0.7043 ± 0.0079	0.7107 ± 0.0158

MNIST

Варьирование α и supervised num

	s = 25	s = 30	s = 50	s = 100
α = 0	0.7697 ± 0.0623	0.8076 ± 0.0775	0.8030 ± 0.0514	0.8233 ± 0.0441
α = 0.3	0.8369 ± 0.0496	0.8624 ± 0.0391	0.8026 ± 0.0732	0.8543 ± 0.0309
α = 0.5	**0.8402 ± 0.0294**	0.8520 ± 0.0680	0.8385 ± 0.0594	0.8708 ± 0.0370
α = 0.7	0.7668 ± 0.0701	0.8112 ± 0.0697	0.8398 ± 0.0308	0.8196 ± 0.0745
α = 1	0.7434 ± 0.0299			

USC-HAD, unsupervised num = 250

	s = 25	**s = 30**	s = 50	s = 100
α = 0	0.6845 ± 0.0410	0.6535 ± 0.0364	0.8021 ± 0.0551	0.7593 ± 0.0308
α = 0.3	0.6994 ± 0.0204	0.7016 ± 0.0234	0.7749 ± 0.0277	0.7832 ± 0.0092
α = 0.5	0.7381 ± 0.0279	**0.7805 ± 0.0356**	0.7521 ± 0.0437	0.8038 ± 0.0596
α = 0.7	0.7307 ± 0.0235	0.7728 ± 0.0580	0.7320 ± 0.0777	0.7740 ± 0.0131
α = 1	0.7323 ± 0.0125			

USC-HAD, unsupervised num = 500

Выводы:

1. для получения высоких значений F1-Score достаточно небольшого количества размеченных данных
2. для выборки USC-HAD использование дополнительных неразмеченных данных улучшает качество детектирования

Варьирование unsupervised num

Во втором эксперименте для данных Synthetic 1D и MNIST мы меняли unsupervised num – размер неразмеченного датасета, при фиксированных α и s.

unsupervised_num	F1-Score
400	0.9147 ± 0.0033
450	0.8724 ± 0.0463
500	**0.9198 ± 0.0032**
550	0.7942 ± 0.1137
600	0.8804 ± 0.0431

Synthetic 1D

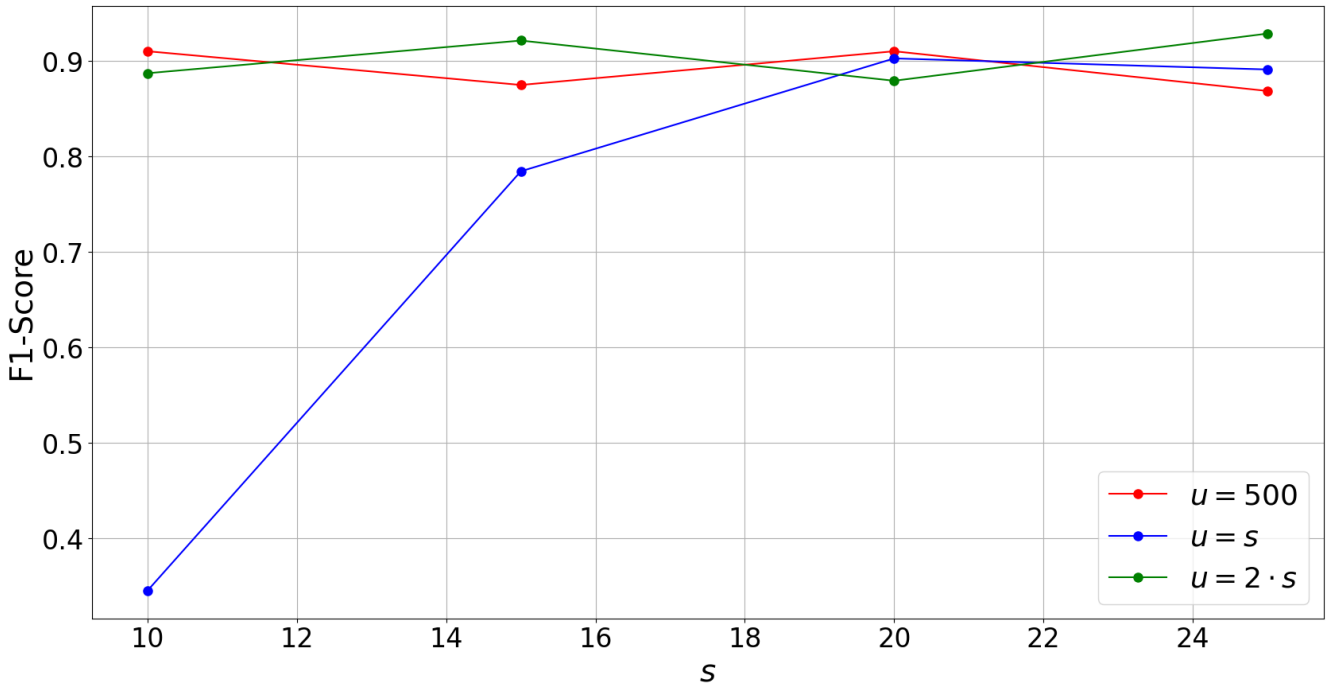
unsupervised_num	F1-Score
400	0.6814 ± 0.0299
450	0.6540 ± 0.0427
500	0.7193 ± 0.0301
550	**0.7436 ± 0.0261**
600	0.7091 ± 0.0111

MNIST

Вывод: оптимальный размер неразмеченной выборки в обоих случаях составляет 500 последовательностей.

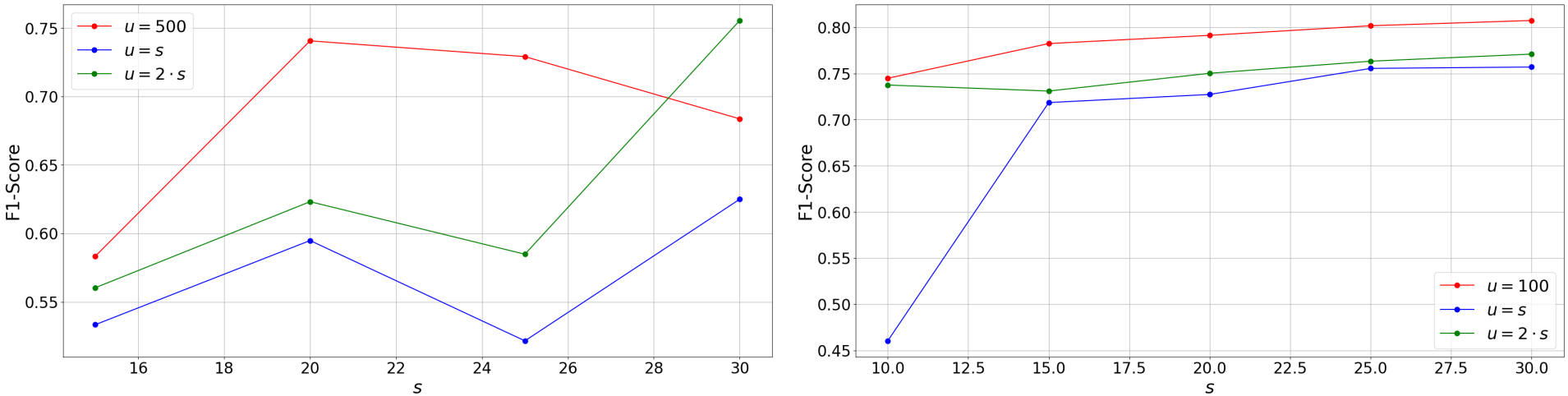
Построение графиков обучения

В рамках третьего эксперимента было исследовано качество детектирования разладок при различных соотношениях размеров размеченной выборки s и неразмеченной – u.



Зависимость F1-Score от размера размеченной выборки. Датасет Synthetic 1D.

Построение графиков обучения



Зависимость F1-Score от размера размеченной выборки. Датасеты MNIST (слева) и USC-HAD (справа).

Выводы:

1. при $s = 30$ мы получаем лучшие метрики при меньших u для датасета MNIST
2. при достаточном размере размеченной выборки ($s \geq 15$) значение F1-Score выходит на плато для датасета Synthetic 1D и USC-HAD
3. лучшие результаты получены при использовании больших размеченных выборок для всех датасетов

Finetuning

1. полное обучение модели на неразмеченных данных (unsupervised num = 500)
2. дообучение модели на небольшой размеченной выборке в течение небольшого числа эпох

В экспериментах мы также варьировали размер размеченной выборки для дообучения s.

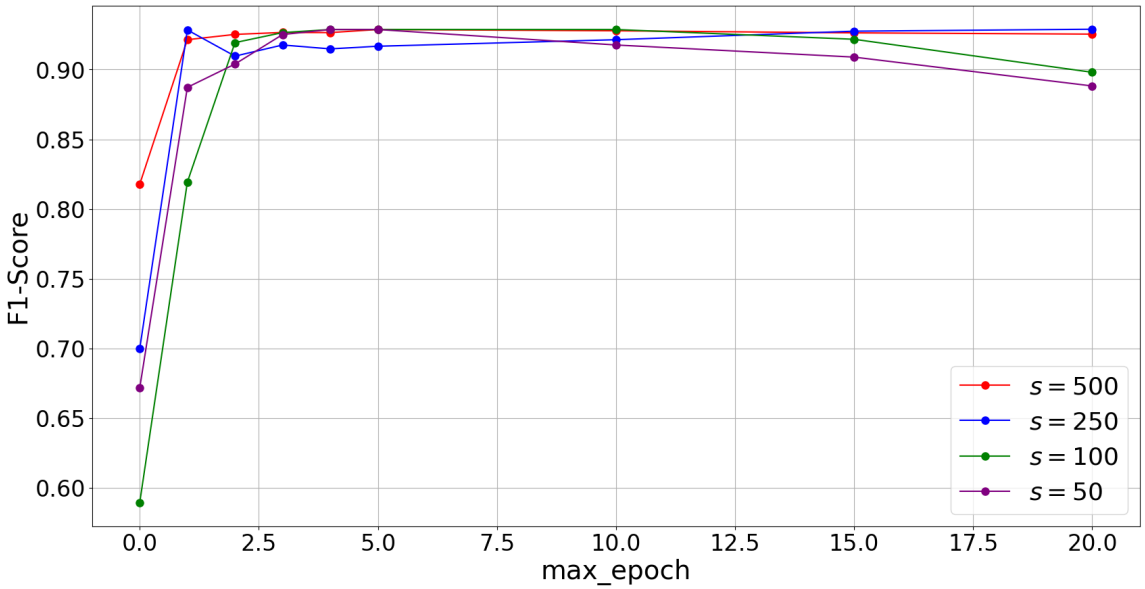


График зависимости F1-Score от длительности дообучения. Результаты приведены для различных размеров s размеченной выборки. Выборка Synthetic 1D.

Finetuning

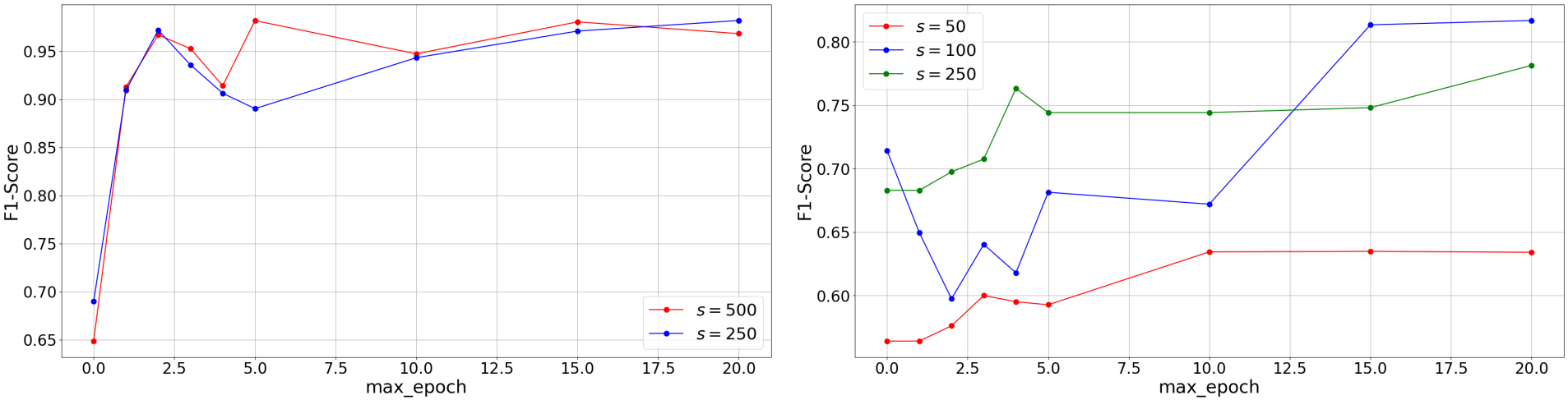


График зависимости F1-Score от длительности дообучения. Результаты приведены для различных размеров s размеченной выборки. Выборки MNIST (слева) и USC-HAD (справа)

Выводы:

1. при достаточном значении дообучающих эпох значение F1-Score выходит на плато для выборок Synthetic 1D и MNIST
2. для выборки USC-HAD значения F1-Score при количестве дообучающих эпох 15 мы получаем лучшие метрики при меньших s
3. F1-Score увеличивается с ростом количества эпох, которые обучается наша модель для всех выборок

Выводы

1. в работе представлены два метода детектирования разладок в многомерных временных рядах. Эти методы позволяют улучшить работу контрастивной модели TS-CP2 за счет использования размеченных данных
2. в рамках работы были проведены эксперименты с данными различной сложности и структуры, включая синтетические последовательности нормальных случайных величин со скачком среднего, последовательности изображений MNIST с плавными переходами между цифрами и выборку USC-HAD, содержащую реальные измерения датчиков устройств, носимых человеком
3. эксперименты демонстрируют, что для значительного улучшения оригинальной модели требуется лишь небольшая размеченная обучающая выборка

Список литературы

[1] Jie Chen and Arjun K Gupta. Parametric statistical change point analysis: with applications to genetics, medicine, and finance. 2012.

[2] Shohreh Deldari, Daniel V Smith, Hao Xue, and Flora D Salim. Time series change point detection with self-supervised contrastive predictive coding. In *Proceedings of the Web Conference 2021*, pages 3124–3135, 2021.

[3] Evgeniya Romanenkova, Alexey Zaytsev, Nikita Klyuchnikov, Arseniy Gruzdev, Ksenia Antipova, Leyla Ismailova, Evgeny Burnaev, Artyom Semenikhin, Vitaliy Koryabkin, Igor Simon, et al. Real-time data-driven detection of the rock-type alteration during a directional drilling. *IEEE Geoscience and Remote Sensing Letters*, 17(11):1861–1865, 2019.

[4] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[5] Mi Zhang and Alexander A Sawchuk. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 1036–1043, 2012.