УДК 004.89

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

А.С. Веприков[1, 2], А.Л. Степикин[1, 2], Е.Д. Романенкова[2], А.А. Зайцев[2]

[1]Московский физико-технический институт (национальный исследовательский университет)

[2]Сколковский институт науки и технологий

Разладка во временном ряде – это момент резкой смены вероятностного распределения составляющих его данных. 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 данных бурения скважин [1], при видеонаблюдении [2] или мониторинге медицинских показателей [3]. Часто разметка доступна только для части обнаруженных аварий или инцидентов, снятых камерой.

Целью работы является создание подхода, который эффективно использует и размеченные, и неразмеченные данные. Разработанный подход позволяет получать высокое качество детектирования разладок в случае, когда доступна лишь небольшая выборка размеченных данных при наличии достаточно большого набора неразмеченных.

В качестве основного подхода к построению представлений для детектирования разладок был использован контрастивный метод TS-CP^2 [4]. Благодаря использованию специальной функции потерь модель отображает «позитивные» пары подпоследовательностей (то есть пары из одного распределения) в пространство представлений так, чтобы расстояние между образами этих пар было небольшим, а «негативные» пары (то есть пары из разных распределений) – в представления, расстояние между которыми велико. Для предсказания разладок измеряется косинусное расстояние между двумя последовательными окнами наблюдений, скользящими по последовательностям. Оригинальный метод предполагает выбор «позитивных» и «негативных» пар окон без учета разметки данных. В основе алгоритма сэмплирования лежит предположение о том, что «позитивные» пары находятся в одной последовательности на небольшом расстоянии друг от друга, а негативные – в разных последовательностях обучающей выборки. Идея предложенного подхода состоит в том, чтобы кроме этого рассматривать небольшое количество «истинно позитивных» и «истинно негативных» пар, полученных с учетом разметки.

В работе рассматриваются 2 способа обучения модели. В первом случае итоговая модель обучается с помощью комбинированной функции потерь: $L = \alpha L_s + (1 - \alpha) L_u$, где L_s – значение функции потерь, полученное для пар из размеченных данных, L_u – из неразмеченных, α – весовой коэффициент. Во втором случае модель сначала полноценно обучается на неразмеченных данных, а затем дообучается (fine-tuning) на размеченных в течение небольшого числа эпох.

В рамках данной работы были проведены эксперименты с данными различной сложности и структуры: от синтетических последовательностей нормальных случайных величин со скачком среднего и последовательностей изображений MNIST с плавными переходами между цифрами до выборки USC-HAD [5], которая содержит реальные измерения датчиков устройств, носимых человеком. В последнем случае разладки соответствуют изменениям типов человеческой активности: бег, ходьба, сон и т.д. В качестве основного критерия оценки качества детектирования была использована классификационная метрика F1-score, адаптированная для задачи детектирования разладок [6]. На Рис. 1 и 2 приведены

результаты экспериментов для обоих сценариев обучения моделей и различных соотношений размеров используемых размеченной и неразмеченной выборок. Видно, что увеличение количества доступных неразмеченных данных позволяет получить высокие результаты даже для размеченной выборки данных размером 10 точек.

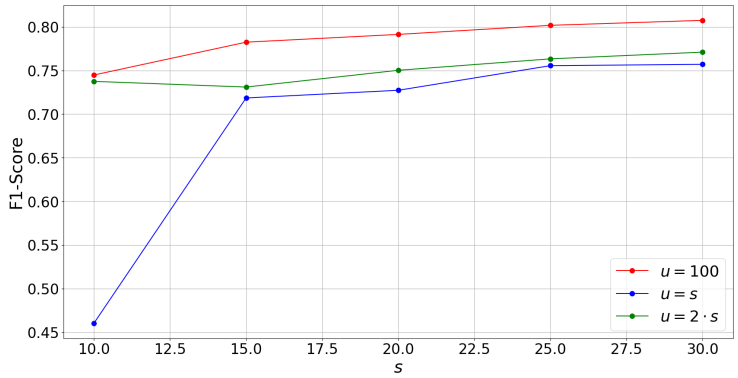

Рис. 1. График зависимости F1-score от размера размеченной обучающей выборки (s) при различных размерах неразмеченной обучающей выборки (u). Результаты для выборки USC-HAD [5]. Модель обучается с помощью комбинированной функции потерь с $\alpha = 0.5$. Чем выше F1-score, тем лучше модель.

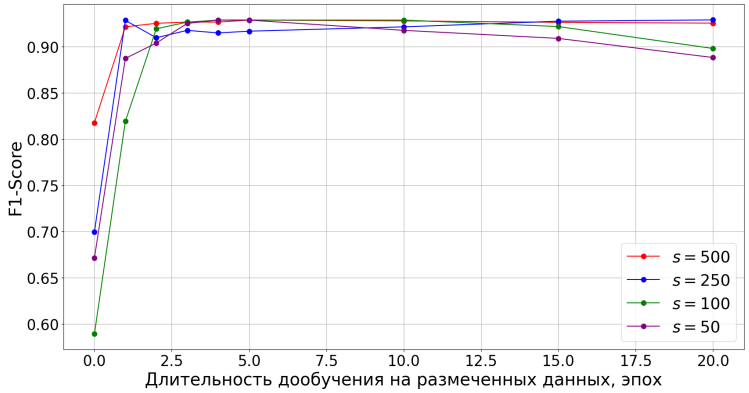

Рис. 2. График зависимости F1-score от длительности дообучения (число эпох) на размеченных данных при различных размерах размеченной выборки (s). Результаты для синтетической выборки. Чем выше F1-score, тем лучше модель.

Таким образом, представленные методы детектирования разладок в многомерных временных рядах позволяют значительно улучшить качество работы контрастивной модели TS-CP^2 за счет использования размеченных данных. Эксперименты доказывают, что для значительного улучшения оригинальной модели требуется лишь небольшая размеченная обучающая выборка.

Работа поддержана грантом РНФ (проект 20-71-10135).

Литература

1. *Romanenkova E. et al.* Real-time data-driven detection of the rock type alteration during a directional drilling. 2019. In: IEEE Geoscience and Remote Sensing Letters 17.11, C. 1861 – 1865.
2. *Sultani W. et al.* Real-world anomaly detection in surveillance videos. 2018. In: Proceedings of the IEEE conference on computer vision and pattern recognition, C. 6479-6488.
3. *J. Chen and A. K. Gupta.* Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance. 2011.
4. *Deldari S. et al.* Time series change point detection with self-supervised contrastive predictive coding // Proceedings of the Web Conference 2021. – 2021. – C. 3124-3135.
5. *Zhang M., Sawchuk A.* USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors // Proceedings of the 2012 ACM conference on ubiquitous computing. – 2012. – C. 1036-1043.
6. *Romanenkova E. et al.* InDiD: Instant Disorder Detection via a Principled Neural Network // Proceedings of the 30th ACM International Conference on Multimedia 2022. – 2022. – C. 3152-3161.