

Методы детектирования аномалий.

Лекция 1: Плотность данных

Иван Шанин
ivan.shanin@gmail.com

ИПИ РАН

11.02.2019

Лекция 1: Плотность данных

1 Введение

- Оргвопросы
- Материалы
- План курса
- Детектирование аномалий

2 Плотность данных

- Локальный фактор аномальности (LOF)
- Оценка плотности распределения
- Параметрическая оценка плотности

Оргвопросы

- ▶ Лекции по понедельникам, ауд. 612, 17.00 - 18.20
- ▶ Три домашних задания, учет посещения, экзамен
- ▶ 55 баллов - уд., 75 баллов - хор, 85 баллов - отл
- ▶ По вопросам писать на ivan.shanin@gmail.com
- ▶ Рекомендуемая среда - Jupyter Notebook
- ▶ Слайды и домашние задания будут выкладываться на github

- ▶ Учебники и монографии:
 - ▶ Outlier Analysis - Charu C. Aggarwal (Springer, 2017, 2nd ed.)
 - ▶ Time Series Knowledge Mining - F. Mörchén (2006)
- ▶ Датасеты:
 - ▶ Outlier Detection DataSets (ODDS)
<http://odds.cs.stonybrook.edu/>
 - ▶ Numenta Anomaly Benchmark (NAB)
<https://github.com/numenta/NAB>

▶ Метрические методы:

1. Локальные методы: kNN, DBSCAN, LOF, LOCI
2. Кластеризация: k-means, иерархическая кластеризация
3. Восстановление плотности данных: непараметрические методы, EM-алгоритм
4. Классификация: наивный байесовский классификатор, одноклассовый SVM

▶ Аномалии во временных рядах:

1. Авторегрессионные модели. ARIMA.
2. Методы, основанные на расстояниях между временными рядами: евклидово расстояние, его вариации. Dynamic Time Warping.
3. Представления временных рядов: спектральное разложение, вейвлет-разложение.

План курса

- ▶ Аномалии в дискретных последовательностях:
 1. Методы, основанные на расстоянии между последовательностями: простое сравнение, наибольшая общая подпоследовательность (с учетом нормализации), compression-based dissimilarity.
 2. Оконные оценки аномальности дискретных последовательностей
 3. Моделирование дискретных последовательностей: конечные автоматы, суффиксные деревья. Скрытые марковские модели
- ▶ Поиск аномалий в графах:
 1. Методы анализа статических графов. Поиск аномальных вершин. Методы, основанные на метрике Egonet.
 2. Поиск реберных аномалий с помощью структурных генерационных моделей
 3. Применение матричной факторизации к задаче поиска аномальных ребер в графе
 4. Детектирование аномальных подграфов, методы MDL и SUBDUE.

План курса

- ▶ Аномалии в текстовых данных:
 1. Методы, основанные на частотности слов. TF-IDF. Поиск первоисточника (first story detection).
 2. Латентный семантический анализ, тематическое моделирование

Примечание

Курс находится в состоянии переработки, план может незначительно измениться в течении семестра.

Постановка задачи

Определение

Аномалия – наблюдение, отличающееся от остальных наблюдений достаточно сильно, чтобы предположить, что оно имеет иную природу происхождения.

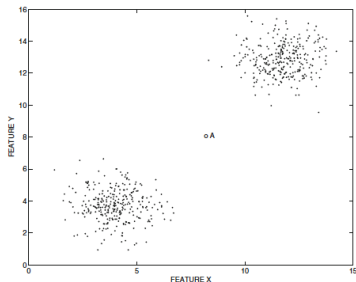
Особенности задачи детектирования аномалий:

- ▶ Характерен сильный дисбаланс классов (вплоть до полного отсутствия аномальных объектов в обучающей выборке)
- ▶ Аномалии могут быть крайне разнородны
- ▶ Аномальные объекты могут быть как шумовыми (подлежащими удалению из выборки), так и имеющими существенное значение для анализа (отражать важные изменения в данных)

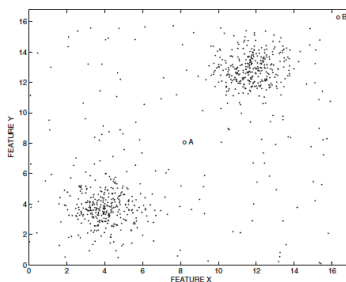
Примеры задач

- ▶ Детектирование мошеннических банковских транзакций
- ▶ Выявление аномальных сердечных сокращений по ЭКГ
- ▶ Детектирование поломок и неисправностей в работе различного оборудования
- ▶ Поведенческие пользовательские аномалии в социальных сетях
- ▶ Детектирование сетевого вторжения
- ▶ Анализ последовательностей ДНК
- ▶ Анализ химических соединений

Зашумленные данные

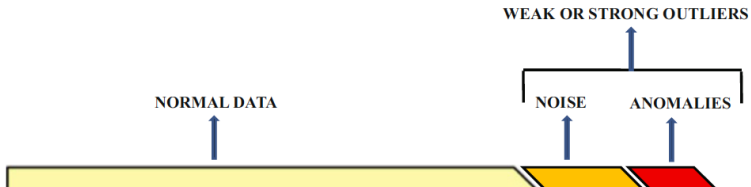


(a) No noise

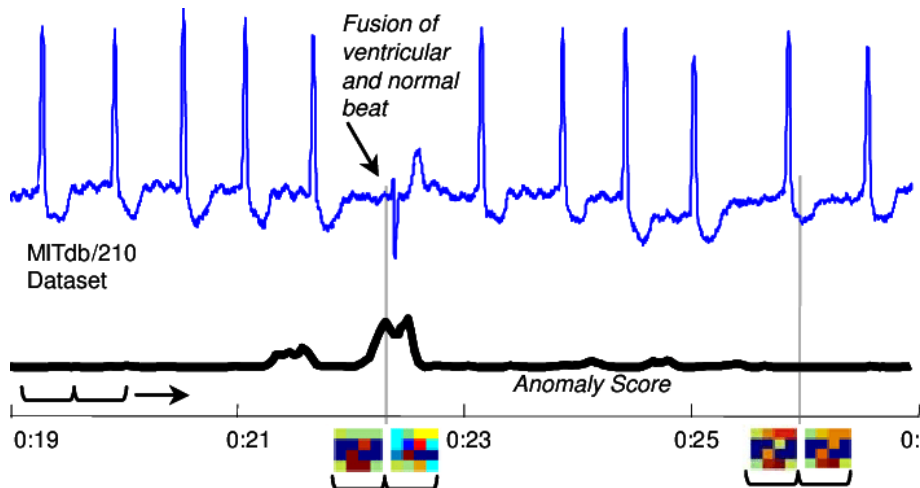


(b) With noise

Figure 1.1: The difference between noise and anomalies



Анализ ЭКГ



Обучение с учителем и без учителя

Supervised Model	Unsupervised Analog
k-nearest neighbor	k-NN distance, LOF, LOCI
Linear Regression	Principal Component Analysis
Naive Bayes	Expectation-maximization
Decision Trees, Random Forest	Isolation Trees, Isolation Forest
Rule-based	FP-outlier
Support-vector machines	One-class SVM
Neural Networks	Replicator neural network
Matrix factorization	Principal Component analysis

Лекция 1: Плотность данных

1 Введение

- Оргвопросы
- Материалы
- План курса
- Детектирование аномалий

2 Плотность данных

- Локальный фактор аномальности (LOF)
- Оценка плотности распределения
- Параметрическая оценка плотности

LOF: Local Outlier Factor

Для объекта \bar{X} определим:

- ▶ $L_k(\bar{X})$ - множество k ближайших соседей объекта \bar{X}
- ▶ $D^k(\bar{X})$ - расстояние от объекта \bar{X} до k -го ближайшего соседа
- ▶ $R_k(\bar{X}, \bar{Y})$ - относительная доступность объекта \bar{X} относительно объекта \bar{Y} :

$$R_k(\bar{X}, \bar{Y}) = \max\{\text{dist}(\bar{X}, \bar{Y}), D^k(\bar{Y})\}$$

- ▶ $AR_k(\bar{X})$ - средняя доступность объекта \bar{X} в окрестности k ближайших соседей:

$$AR_k(\bar{X}) = \frac{1}{|L_k(\bar{X})|} \sum_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

LOF: Local Outlier Factor

Получим, что значения локального фактора выбросов

$$LOF_k(\bar{X}) = \frac{1}{|L_k(\bar{X})|} \sum_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$

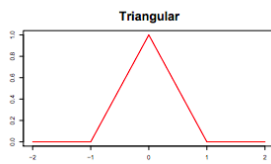
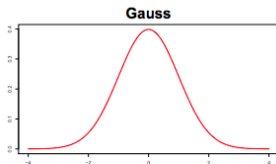
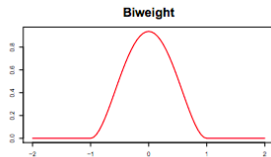
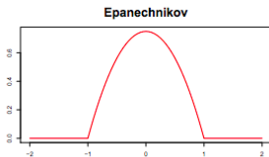
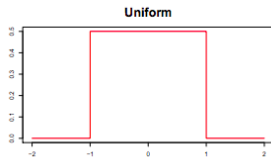
для объектов, находящихся внутри кластера близки к 1, вне зависимости от плотности кластера, тогда как значения фактора для объектов-аномалий будет значительно выше.

Ядерная оценка плотности (KDE)

- ▶ Локальная непараметрическая оценка плотности Парзена – Розенблатта:

$$\hat{\rho}_h(x) = \frac{1}{NV_h} \sum_{i=1}^N K\left(\frac{\rho(x, x_i)}{h}\right)$$

- ▶ $K(z)$ - произвольная четная функция, называемая функцией ядра



Восстановление плотности нормального распределения

Рассмотрим упрощенную ситуацию: признаки объектов распределены нормально и **независимо**: $x \in \mathbb{R}^{(n)}$, $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Оценим параметры μ и σ^2 :

$$\hat{\mu}_i = \frac{1}{m} \sum_{j=1}^m x_i^{(j)}, \quad \hat{\sigma}_i^2 = \frac{1}{m-1} \sum_{j=1}^m \left(x_i^{(j)} - \hat{\mu}_i \right)^2$$

Подставим полученные оценки в формулу плотности нормального распределения:

$$\hat{p}(x) = \prod_{i=1}^n p(x_i; \hat{\mu}_i, \hat{\sigma}_i^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(x_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right).$$