Методы детектирования аномалий. Лекция 2: Кластеризация

Иван Шанин
ivan.shanin@gmail.com

ИПИ РАН

18.02.2019

Локальный интегральный показатель корреляции (LOCI)[1]

Для объекта $\overline{X}$ определим

- $M(\overline{X}, \epsilon)$ - плотность данных в ϵ-окрестности объекта $\overline{X}$
- $AM(\overline{X}, \epsilon, \delta)$ - среднее значение $M(\overline{X}, \epsilon)$ по всем объектам δ-окрестности объекта $\overline{X}$

$$AM(\overline{X}, \epsilon, \delta) = MEAN_{\{\overline{Y}:\rho(\overline{X},\overline{Y})\leq\delta\}} M(\overline{Y}, \epsilon)$$

- Мультигранулярный показатель отклонения

$$MDEF(\overline{X}, \epsilon, \delta) = 1 - \frac{M(\overline{X}, \epsilon)}{AM(\overline{X}, \epsilon, \delta)}$$

Большое значение *MDEF* является признаком аномальности $\overline{X}$.

[1] http://www.cs.cmu.edu/~christos/PUBLICATIONS/icde03-loci-tr.pdf

Диаграмма LOCI

Пусть $\epsilon = \frac{1}{2}\delta$.

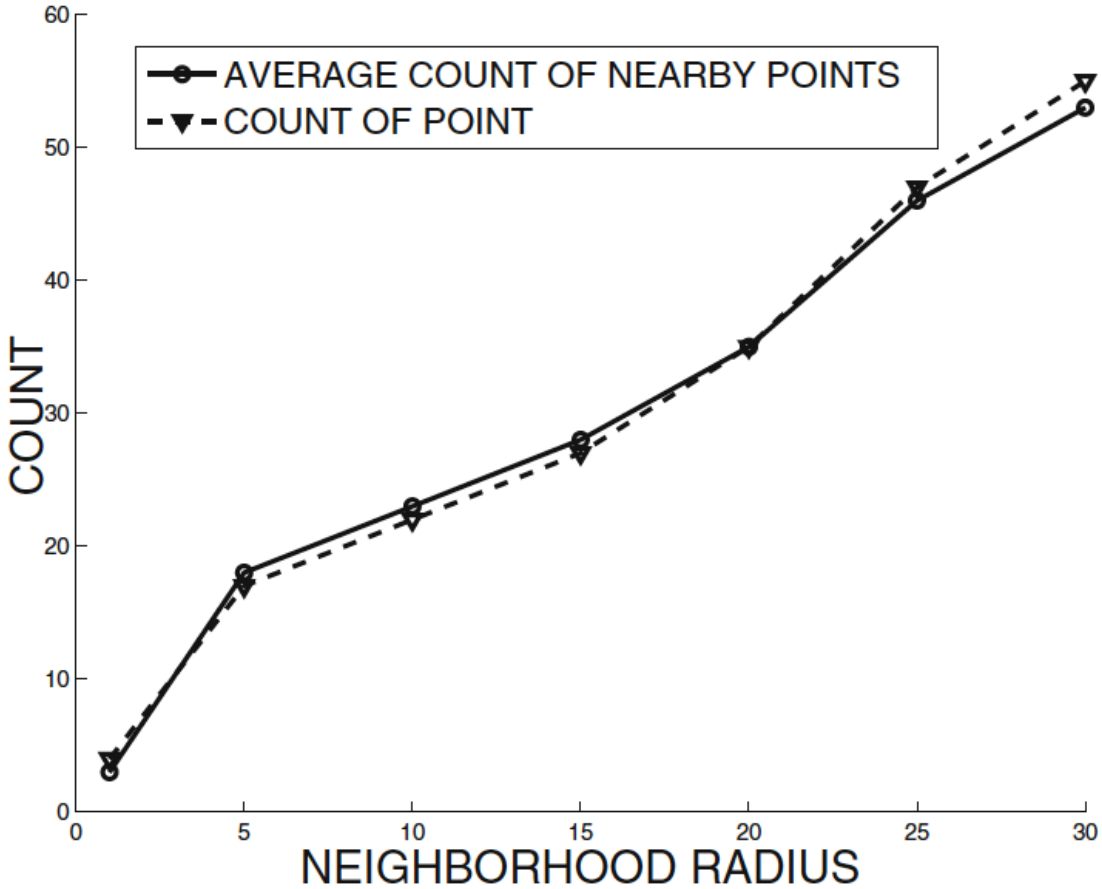


Рис. 1: Объект внутри кластера

Диаграмма LOCI

Пусть $\epsilon = \frac{1}{2}\delta$.

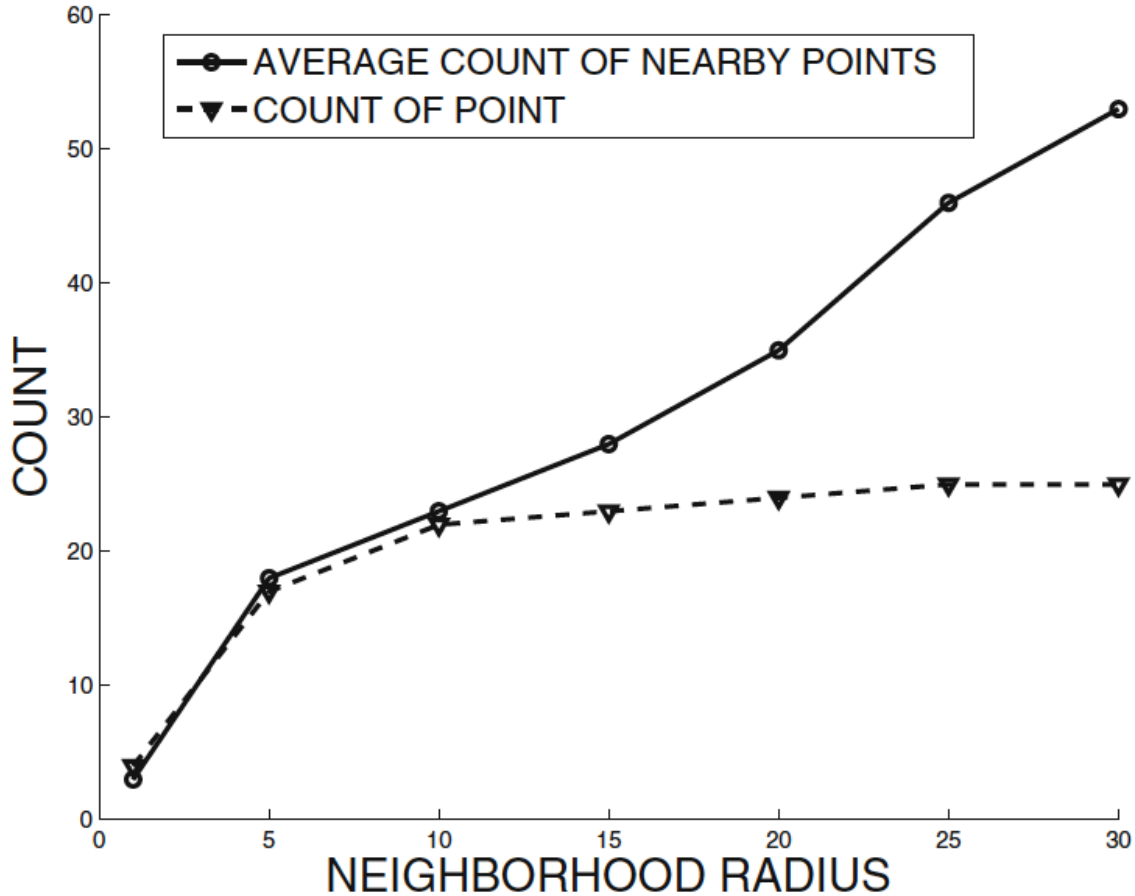


Рис. 2: Небольшой кластер аномальных объектов

Диаграмма LOCI

Пусть $\epsilon = \frac{1}{2}\delta$.

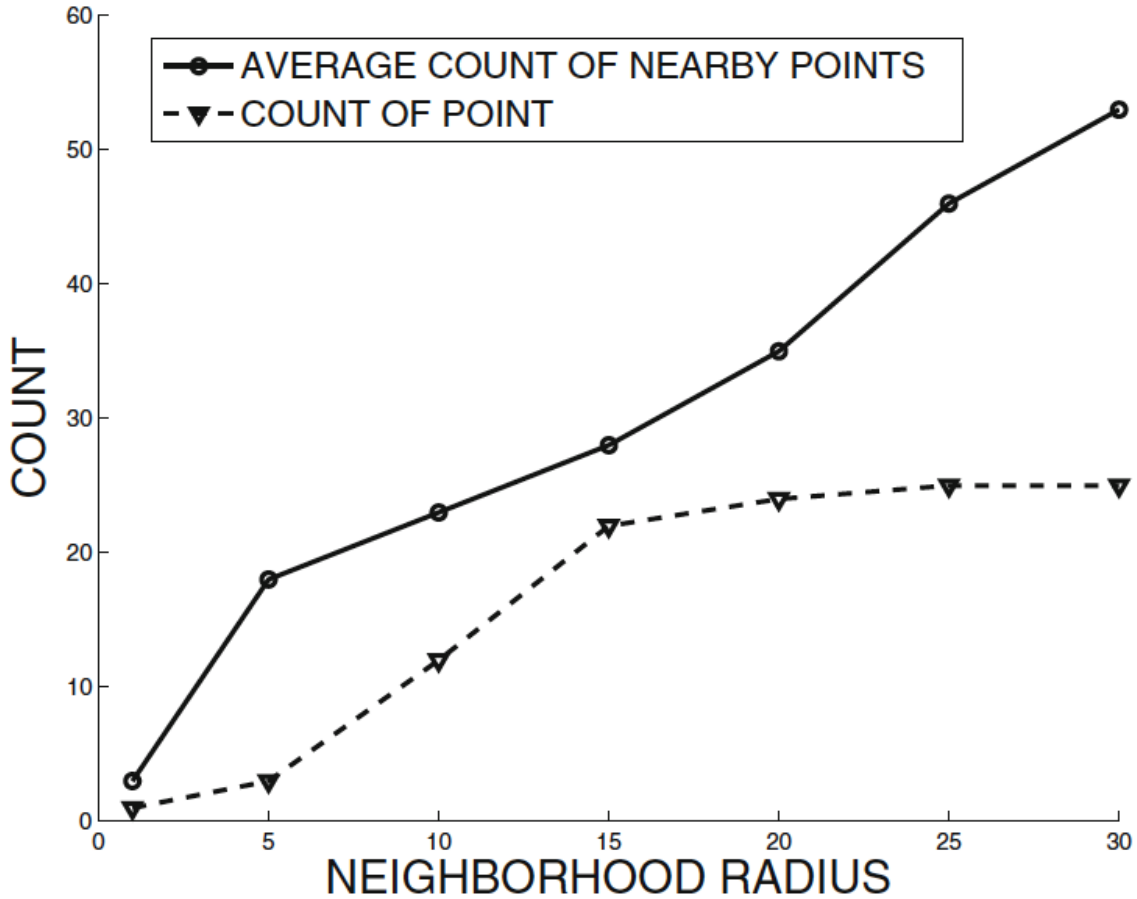
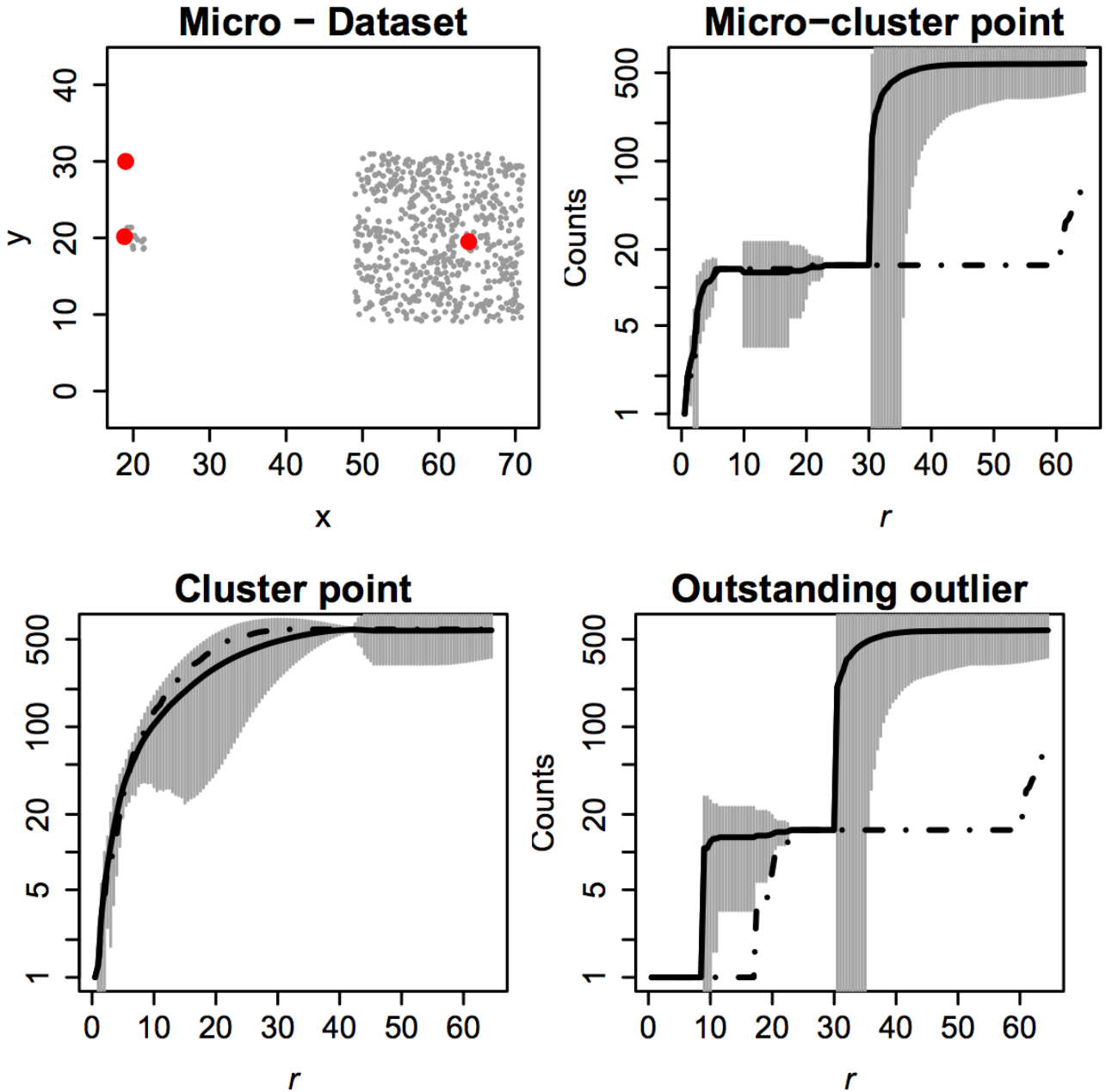


Рис. 3: Аномальный объект

Диаграмма LOCI



Смесь вероятностных распределений

Определим порождающую модель данных M. Предположим, что данные сгенерированы из вероятностных распределений $G_1, G_2 \ldots, G_k$ по следующему стохастическому процессу:

- с вероятностью α_r выбирается вероятностное распределение G_r
- из G_r генерируется объект выборки

где $\alpha_1, \ldots, \alpha_r$ - набор **априорных** вероятностей.
Функция плотности распределения объектов:

$$p(\overline{X}|M) = \sum_{i=1}^{k} \alpha_i p_i(\overline{X}|\theta_i)$$

Функция правдоподобия выборки D:

$$Likelihood(D|M) = \prod_{\overline{X} \in D} p(\overline{X}|M)$$

EM-алгоритм

Задача

По выборке D и заранее заданному параметру k оценить параметры модели $M(\overline{\alpha}, \overline{\theta})$.

Требуется максимизировать логарифм правдоподобия

$$L(D|M) = \log \left[\prod_{\overline{X} \in D} p(\overline{X}|M) \right] = \sum_{\overline{X} \in D} \log \left[\sum_{i=1}^{k} \alpha_i p_i(\overline{X}|\theta_i) \right]$$

Данная задача решается чередованием двух этапов:

- **E-шаг (expectation):** По текущей оценке параметров M оценить $P(I_r|\overline{X})$ – **апостериорную** вероятность того, что объект $\overline{X}$ был сгенерирован r-й компонентной (по формуле Байеса)
- **M-шаг (maximization):** По полученной разметке выборки построить новые значения параметров $M(\overline{\alpha}, \overline{\theta})$

EM-алгоритм

Приведем формулы[2] расчета скрытых и явных параметров на каждом шаге.

- **E-шаг (формула Байеса):**

$$\hat{P}(I_r|\overline{X},\overline{\alpha},\overline{\theta}) = \frac{\alpha_r p_r(\overline{X}|\theta_r)}{\sum_{i=1}^{k}\alpha_i p_r(\overline{X}|\theta_i)}$$

- **M-шаг (максимизация правдоподобия):**

$$\alpha_r = \frac{1}{|D|}\sum_{\overline{X}\in D}\hat{P}(I_r|\overline{X},\overline{\alpha},\overline{\theta});$$

$$\theta_r = \arg\max_{\theta}\sum_{\overline{X}\in D}\hat{P}(I_r|\overline{X},\overline{\alpha},\overline{\theta})\log p_r(\overline{X}|\theta)$$

[2] http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

Метод k средних

Выберем k случайных точек - нулевое приближение к центрам масс кластеров. Затем будем повторять следующие шаги:

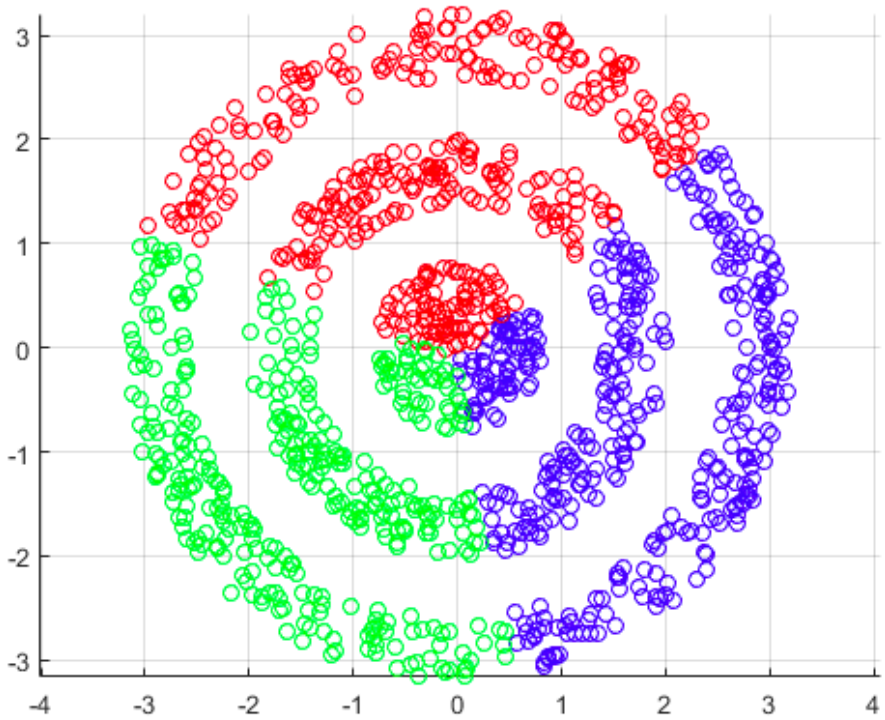
- Отнести каждый объект выборки к одному из k классов (ближайшему по некоторой функции расстояния)
- По получившимся кластерам рассчитать новые центры масс

Таким образом на каждом шаге будет монотонно убывать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Критерий останова - шаг, на котором данное значение не изменилось.

- Метод является упрощенной версией EM-алгоритма

Проблемы метода k-средних

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения, а только одного из локальных минимумов.
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
- Число кластеров надо знать заранее.



DBSCAN

Поделим выборку на следующие классы объектов:

- Элемент ядра, если хотя бы n точек находятся в ϵ-окрестности
- Объект p доступен из эелемента ядра q **напрямую**, если находится в его ϵ-окрестности
- Объект p доступен из элемента ядра q, если существует путь $p_1, \ldots, p_n$ из p в q, где все (p_i, p_{i+1}) связаны напрямую и являются элементами ядра (возможно, за исключением q)