Подбор параметров в алгоритмах детектирования аномалий

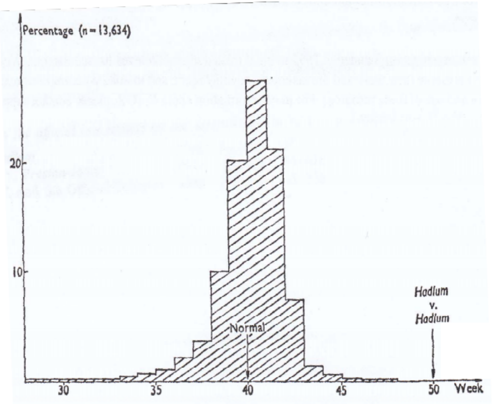
Смоляков Дмитрий

МФТИ, ИППИ РАН

Конференция МФТИ 2014

Пример аномалии

Аномалия – измерение, которое не соответствует некоторому ожидаемому поведению.



Distribution of human gestation periods.

Области применения

- Поиск мошенничества на основе необычной активности
- В медицине необычные результаты обследований могут свидетельствовать о потенциальных проблемах со здоровьем
- Аномальные значения могут быть результатом ошибок в измерении
- Раннее обнаружение аномальной работы двигателя может предотвратить неожиданную поломку

Определение аномалии [Howkins, 1980]

- Пусть Q некоторый механизм порождения данных
- $X_1 \dots X_n$ порождены механизмом Q
- $\tilde{X}_1 \dots \tilde{X}_m$ отличаются от измерений $X_1, \dots, X_n$, настолько, что можно предположить, что они порождены другим механизмом.

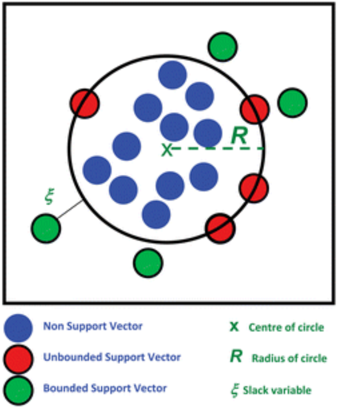
Постановка задачи

- $X_1, \ldots, X_N \quad X_i \in S$ – не размеченная выборка
- Хотим построить $f\colon S \to \{-1, 1\}$
- $f(X_i) = -1$, если X_i – аномалия
- $f(X_i) = 1$, если X_i – нормальное измерение

Support Vector Data Description

$$\begin{cases} R + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i \to \min\limits_{R,x,\xi} \\ \|\phi(x_i) - x\|^2 \leqslant R + \xi_i & i = 1, \dots, N \\ \xi_i \geqslant 0 & i = 1, \dots, N \\ R \geqslant 0 \end{cases}$$

- ν – верхняя граница количества аномалий и нижняя граница количества опорных векторов
- $\phi(x_i)$ – отображение в пространство более высокой размерности



Двойственная задача

Обычно решают двойственную задачу:

$$\begin{cases} \sum\limits_{i=1}^{N} \alpha_i \phi(x_i) \cdot \phi(x_i) - \sum\limits_{i,j=1}^{N} \alpha_i \alpha_j \phi(x_i) \cdot \phi(x_j) \to \max\limits_{\alpha} \\ \sum\limits_{i=1}^{N} \alpha_i = 1 \\ 0 \leqslant \alpha_i \leqslant \frac{1}{N\nu} \quad i = 1, \ldots, N \end{cases}$$

Достаточно знать скалярное произведение.

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Kernel Trick

Необязательно знать явный вид спрямляющего пространства, чтобы использовать $K(x, y)$. Достаточно потребовать:

- $K(x, y) = K(y, x)$ – симметричность
- $\int_X \int_X K(x, x')g(x)g(x')\mathrm{d}x\mathrm{d}x'\ \forall g(x)\colon X \to \mathbb{R}$ – положительная определённость

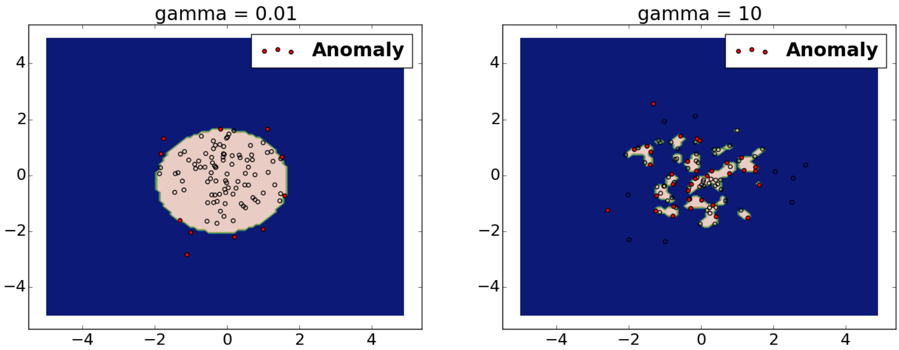
Тогда $K(x, y)$ – скалярное произведение в некотором пространстве.

Примеры ядер

Название	Формула	Параметры
Линейное	$x \cdot y$	—
Полиномиальное	$(\gamma x \cdot y + d)^k$	$\gamma,\ d,\ k$
Гауссово	$\exp(-\gamma \|x - y\|^2)$	γ
Сигмоидное	$\operatorname{th}(\gamma x \cdot y + d)$	$\gamma,\ d$

Выбор ядерной функции

Параметры ядра могут существенно влиять на результат



Хочется получить функцию риска для эффективного подбора параметров ядра.

Задача подбора параметров ядра

- $X_1, \dots, X_n$ - обучающая выборка
- γ - параметризация ядра
- $F(X_1, \dots, X_n, \gamma)$ – алгоритм, который по заданной выборке строит классификатор $f(\cdot)$
- $\tilde{X}_1, \dots, \tilde{X}_m$ - произвольная тестовая выборка
- $\tilde{y}_1, \dots, \tilde{y}_m$ - метки тестовой выборки
- $\mathcal{R}(f(\cdot), \tilde{X}_1, \dots, \tilde{X}_m, \tilde{y}_1, \dots, \tilde{y}_m)$ – риск ошибки на тестовой выборке
- Хотим научиться подбирать γ, не зная тестовую выборку

Оптимизация доли опорных векторов и аномалий

В приложениях часто пытаются достигнуть границы значений количества аномалий и опорных векторов [Lukashevich et al 2009]

- $f_{anomaly}$ – доля объектов отмеченных, как аномалии
- $f_{support}$ – доля объектов используемых, как опорные векторы

$$r = (\nu - f_{anomaly})^2 + (\nu - f_{support})^2$$

Оптимизация функции от ядерной матрицы

Мерой риска может послужить отношение квадрата стандартного отклонения к среднему значению [Evangelista et al 2007]

$$r = \frac{s^2}{\bar{k} + \varepsilon}$$

$$\bar{k} = \frac{\sum_{i=1}^{N} \sum_{j=i}^{N} K(x_i, x_j)}{N(N-1)}; \quad s^2 = \frac{\sum_{i=1}^{N} \sum_{j=i}^{N} (K(x_i, x_j) - \bar{k})^2}{N(N-1) - 1}$$

Учет распределения аномалий

Функция риска может в явном виде учитывать распределения аномальных точек [Steinwart et al 2005]:

$$r = \frac{1}{(1+\rho)|S|}\sum_{i\in S} l(1, sign(f(x_i))) + \frac{\rho}{1+\rho}\mathbb{E}_\mu l(-1, sign(f(x_i)))$$

$$l(y,t) = \max\{0, 1-yt\}; \quad \mu \text{ -- распределение аномалий}$$

Здесь ρ подбирается из априорных соображений.

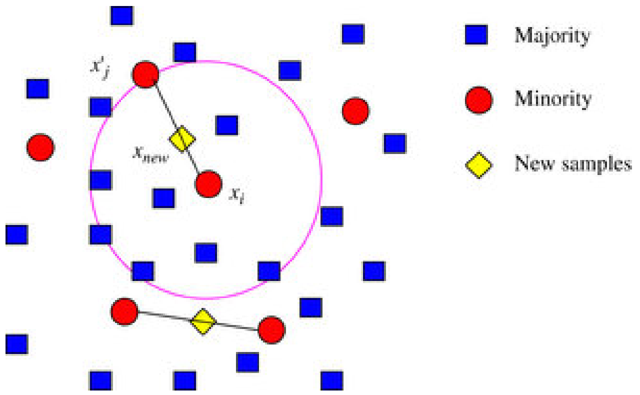
Генерация тестовой выборки

Искусственно сгенерируем примеры аномальных и нормальных точек.

- Аномалии генерируются из равномерного распределения
- Нормальные данные генерируются на основе тестовой выборки, например при помощи SMOTE (Synthetic Minority Over-sampling Technique)

$$r = \frac{1}{|X_{Normal}|} \sum_{x_i \in X_{Normal}} [f(x_i) = -1] + \frac{1}{|X_{Anomaly}|} \sum_{x_i \in X_{Anomaly}} [f(x_i) == 1]$$

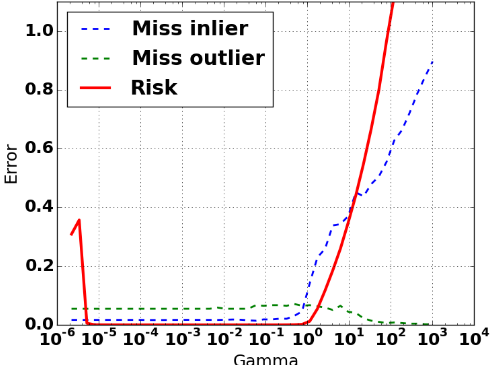
Synthetic Minority Over-sampling Technique



Постановка эксперимента

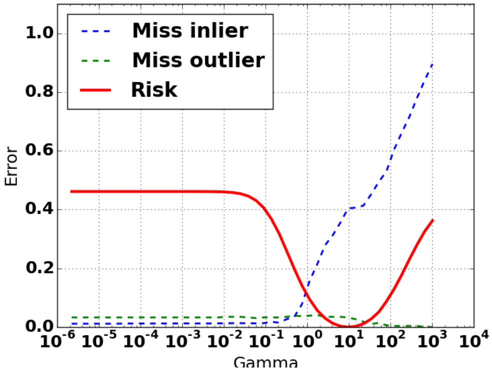
- Нормальные данные генерируются из стандартного многомерного нормального распределения
- Аномалии генерируются из равномерного распределения, причем границы подбираются так, чтобы все нормальные данные попали в заданные границы.
- Получившиеся данные делятся на обучающую и тестовую выборку. На первой происходит подбор параметров и обучение модели, на второй проверка результата.

Результаты



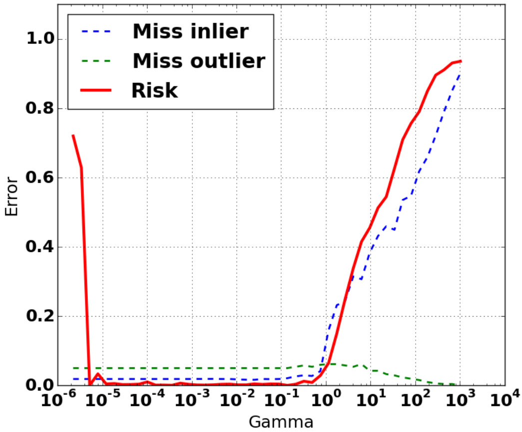
Оптимизация количества аномалий и опорных векторов

Результаты



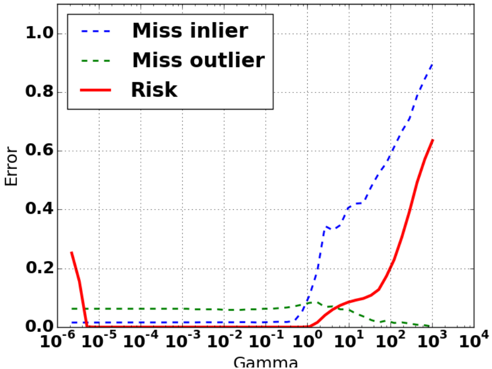
Оптимизация ядерной матрицы

Результаты



Учет распределения аномалий

Результаты



Случайная генерация тестовой выборки

Выводы

- Только в методе оптимизации ядерной матрицы существует единственный минимум
- Критерий оптимальности может быть различным в зависимости от важности ошибки на нормальных и аномальных данных
- Только критерий на основе случайной генерации тестовой выборки способен учитывать различиях ошибок

Спасибо за внимание