Техника исследования гравитационных волн

Лекция 14

Информационно-энтропийный анализ для обнаружения и классификации гравитационных волн

Усіпов Н.М., PhD · КазНУ имени аль-Фараби

Цель и план лекции

Цель

Изучение построения энтропийных признаков сигнала, методов обнаружения событий и их классификации, а также способов корректной оценки качества классификаторов.

1. Введение
2. Постановка задачи обнаружения
3. Энтропийные признаки
4. Алгоритмы классификации
5. Оценка качества классификатора
6. Обнаружение событий
7. Статистическая значимость и несбалансированность классов
8. Выбор и интерпретация признаков

1 Введение

- Поток данных детекторов содержит огромное количество событий: большинство из них являются шумом или глитчами, и только малая часть может оказаться астрофизическими сигналами.

2 Постановка задачи обнаружения

- Обнаружение рассматривается как проверка гипотез: H_0 – данные содержат только шум, H_1 – данные содержат шум и сигнал.
- вероятностью ложной тревоги – принятия шумового участка за сигнал;
- вероятностью обнаружения – правильной регистрации присутствующего сигнала.
- Согласно лемме Неймана–Пирсона, оптимальным является отношение правдоподобия.

3 Энтропийные признаки

Для каждого сегмента данных формируется вектор признаков. В него могут входить:

Спектральная энтропия H_f

Показывает концентрацию энергии по частоте.

Временная энтропия H_t

Показывает концентрацию энергии по времени. Малое значение указывает на коротко существующие события.

Частотно-временная энтропия

Рассчитывается по нормированной спектрограмме.

Энтропии Реньи и Тсаллиса разных порядков

Отражают хвосты распределения.

Для сигналов и глитчей значения имеют разный характер: чирп распределён во времени и частоте по определённой траектории, а короткий глитч локализован по времени.

3 Энтропийные признаки (продолжение)

Перестановочная и выборочная энтропия

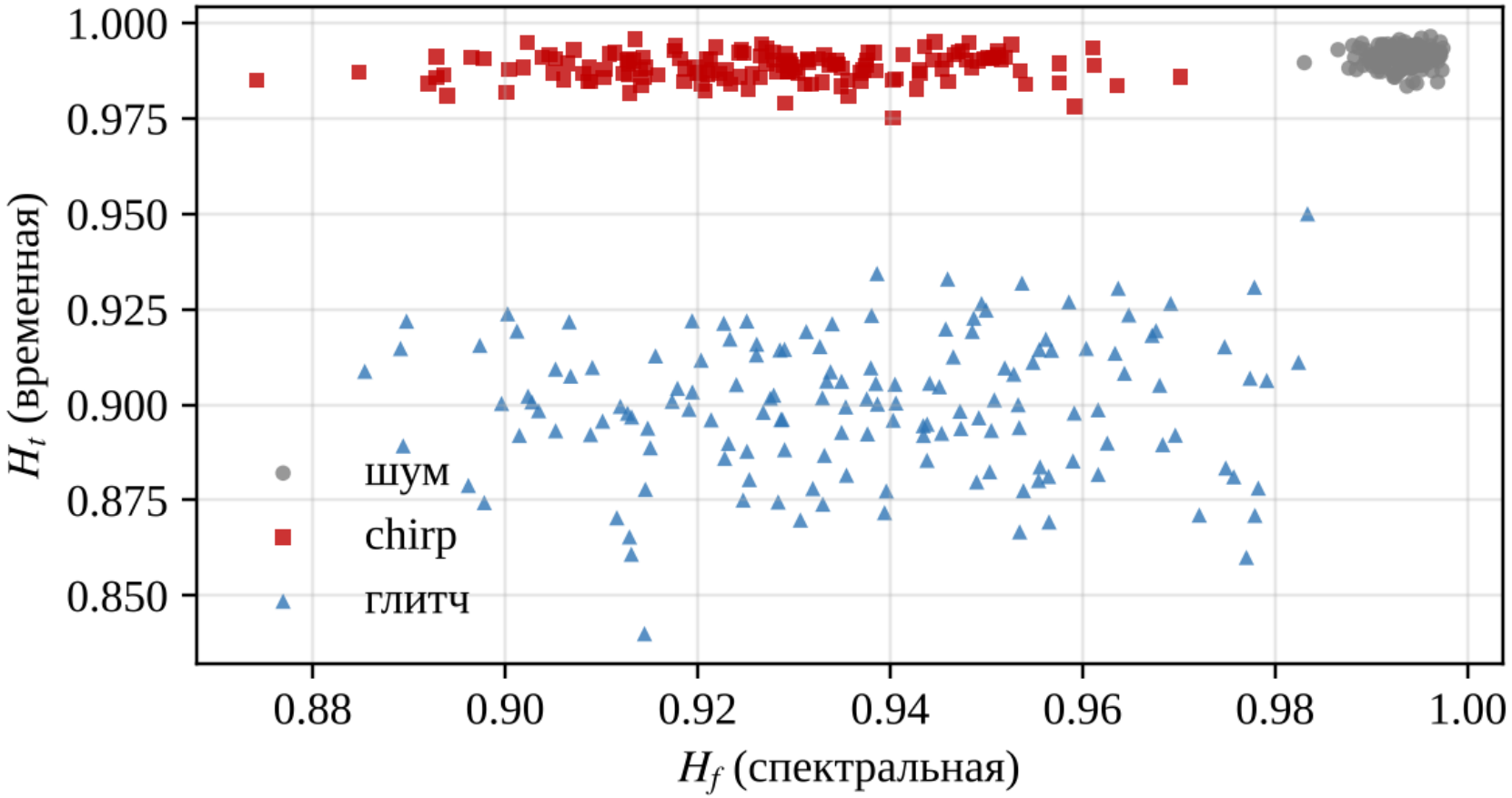
Отражают временной порядок значений.

Вейвлет-энтропия и энтропия мод

Описывают распределение по масштабам.

Для сигналов и глитчей значения имеют разный характер: чирп распределён во времени и частоте по определённой траектории, а короткий глитч локализован по времени.

3 Энтропийные признаки (продолжение)



Распределение модельных событий в пространстве энтропийных признаков (p = 15)

4 Алгоритмы классификации

Линейный дискриминантный анализ

Простой метод, строящий линейные границы между классами. Подходит для небольшого числа признаков.

Метод опорных векторов

Ищет разделяющую поверхность с максимальным зазором.

Случайный лес и градиентный бустинг

Ансамбли деревьев решений; устойчивы к масштабу признаков.

Нейронные сети

Могут использовать как вектор признаков, так и непосредственно спектрограммы.

5 Оценка качества классификатора

- При оценке работы классификатора необходимо разделить выборку на обучающую и тестовую, чтобы не завышать результаты.
- матрица ошибок, показывающая соотношение верных и ошибочных прогнозов для каждого класса;
- точность, полнота и F1-мера;
- ROC-кривая и AUC;
- перекрёстная проверка для учёта разброса.
- Важным источником ошибок является различие распределений обучающих и реальных данных.

5 Классификация шума, чирпа и глитча

Листинг 14.1.

```
import numpy as np
from scipy import signal
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split

fs, n = 4096, 2048

def H(p):
    p = p / p.sum()
    p = p[p > 0]
    return -(p * np.log(p)).sum() / np.log(p.size)

def features(x):
    f, _, P = signal.spectrogram(x, fs, nperseg=128, noverlap=96)
    P = P[(f >= 20) & (f <= 500)]
    return H(P.sum(0)), H(P.sum(1)), H(P.ravel())
```

5 Классификация шума, чирпа и глитча (продолжение)

Листинг 14.1.

```
def chirp(Mc):
    m = Mc * 4.925e-6
    t = np.arange(-n, 0) / fs
    tau = 1 / fs - t
    f = (5 / (256 * tau))**0.375 / (np.pi * m**0.625)
    h = (np.pi * m * f)**(2 / 3) * np.cos(-2 * (5 * m)**-0.625 * tau**0.625)
    h[(f < 20) | (f > 3830 / Mc)] = 0
    return h / np.sqrt((h**2).sum())

def sample(kind, snr, r):
    x = r.normal(size=n)
    if kind == 1:
        x += snr * np.roll(chirp(r.uniform(15, 40)), -r.integers(300))
    if kind == 2:
        t = np.arange(n) / fs
        f0, Q, t0 = r.uniform(60, 300), r.uniform(4, 12), r.uniform(.1, .4)
        g = np.exp(-((t - t0) * 2 * np.pi * f0 / Q)**2)
        g = g * np.sin(2 * np.pi * f0 * (t - t0))
        x += snr * g / np.sqrt((g**2).sum())
    return x
```

5 Классификация шума, чирпа и глитча (продолжение)

Листинг 14.1.

```
rng = np.random.default_rng(0)
for snr in (8, 12, 20):
    X = [features(sample(k, snr, rng)) for k in range(3) for _ in range(250)]
    y = np.repeat(range(3), 250)
    a, b, ya, yb = train_test_split(np.array(X), y, test_size=0.4,
                                    random_state=1, stratify=y)
    model = LinearDiscriminantAnalysis().fit(a, ya)
    p = model.predict(b)
    print(f"SNR={snr}: точность {(p == yb).mean():.3f}")
    if snr == 12:
        print(confusion_matrix(yb, p))
```

5 Результат: Классификация шума, чирпа и глитча

```
SNR=8: точность 0.860
SNR=12: точность 0.970
[[100   0   0]
 [  8  92   0]
 [  0   1  99]]
SNR=20: точность 0.990
```

Данный эксперимент показывает, что точность растёт с увеличением значения ρ.

6 Обнаружение событий

- Для обнаружения событий используется пороговый критерий по энтропии на скользящих окнах.

6 Обнаружение при заданной вероятности ложной тревоги

Листинг 14.2.

```
stat = lambda x: -features(x)[1]
r = np.random.default_rng(5)
s0 = np.array([stat(sample(0, 0, r)) for _ in range(1000)])
thr = np.quantile(s0, 0.99)
print(f"Порог при ложной тревоге 1%: {thr:.4f}")
for snr in (8, 12, 16, 24):
    s1 = [stat(sample(1, snr, r)) for _ in range(300)]
    pd = np.mean(np.array(s1) > thr)
    print(f"rho={snr:2}: вероятность обнаружения {pd:.2f}")
```

```
Порог при ложной тревоге 1%: -0.9860
rho= 8: вероятность обнаружения 0.66
rho=12: вероятность обнаружения 1.00
rho=16: вероятность обнаружения 1.00
rho=24: вероятность обнаружения 1.00
```

7 Статистическая значимость и несбалансированность классов

- В реальной задаче количество шумовых сегментов во много раз превышает количество сигналов.
- Практическая эффективность детектора задаётся чувствительным объёмом ⟨VT⟩ – произведением объёма пространства и времени наблюдения, в которых детектор способен зарегистрировать инъектированные сигналы при заданной частоте ложных тревог. Этот показатель позволяет напрямую сравнивать конвейеры поиска и применяется в соревнованиях алгоритмов, например Mock Data Challenge.
- Важной особенностью гравитационно-волновых поисков является необходимость экстраполяции к чрезвычайно малым FAR. Классификатор, обученный на ограниченных выборках, не обязательно обладает нужным поведением на хвосте распределения шума.

8 Выбор и интерпретация признаков

- При большом количестве признаков возрастает риск переобучения и излишних корреляций.
- Методы объяснения, например значения Шепли (SHAP), показывают вклад каждого признака в конкретное решение.
- Необходимо учитывать сдвиг распределения. Свойства шума зависят от детектора и цикла наблюдения, поэтому модель, обученная на одном периоде, может давать ухудшенные результаты на другом.

9 Каскадная схема обработки

- В практических приложениях разумно использовать каскадную структуру. На первом этапе дешёвые энтропийные признаки или простые статистики отсекают большую долю заведомо шумовых сегментов.
- Классификация глитчей важна сама по себе. Определив тип помехи, инженер может найти её источник и устранить его.
- Для поиска сигналов неизвестного типа пригодны методы обнаружения аномалий: изолирующий лес, одноклассовый метод опорных векторов и автокодировщики.

10 Разбор примера

1. Используем матрицу ошибок из листинга 14.1 при ρ = 12.
2. Общая точность равна (100 + 92 + 99)/300 = 0,97.
3. Полнота для чирпа составляет 92/100 = 0,92.
4. Точность для чирпа равна 92/93 ≈ 0,989, поскольку лишь один глитч ошибочно отнесён к этому классу.
5. Точность для класса «шум» равна 100/108 ≈ 0,926.
6. При вероятности ложной тревоги 1 % и сегментах длиной 0,5 с получается 2·0,01 = 0,02 ложных срабатывания в секунду, то есть 1728 в сутки.
7. Это показывает, почему порог 1 % непригоден для окончательного решения и подходит лишь для предварительного отбора.

11 Вопросы для обсуждения на занятии

Показатели

Почему общая точность плохо характеризует поиск редких сигналов?

Каскад

Какой порог задать на первой ступени обработки?

Перенос

Как проверить применимость модели, обученной на данных одного цикла наблюдений, к другому циклу?

Контрольные вопросы

1. Сформулируйте задачу обнаружения сигнала как проверку гипотез.
2. Что такое ROC-кривая и какой показатель используется для сравнения методов?
3. Какие энтропийные признаки можно использовать для описания сигнала?
4. Как энтропийные признаки различают чирп и глитч?
5. Какие алгоритмы применяются для классификации событий?
6. Какие показатели применяются при оценке качества классификаторов?
7. Почему важно разделение на обучающую и тестовую выборки?
8. В чём опасность обучения только на синтетических данных?

Список литературы

1. Neyman J., Pearson E.S. On the problem of the most efficient tests of statistical hypotheses // Philosophical Transactions of the Royal Society A. – 1933. – Vol. 231. – P. 289–337.
2. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. – 2006. – Vol. 27, No. 8. – P. 861–874.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – 2nd ed. – New York: Springer, 2009.
4. Pedregosa F., et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.
5. Zevin M., et al. Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science // Classical and Quantum Gravity. – 2017. – Vol. 34. – Art. 064003.
6. Abbott B.P., et al. A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals // Classical and Quantum Gravity. – 2020. – Vol. 37, No. 5. – Art. 055002. – DOI: 10.1088/1361-6382/ab685e.
7. Shannon C.E. A mathematical theory of communication // Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423, 623–656.