Отбор признаков на основе SVM-подхода

Елизавета Добрецова, группы 5040102/40101

Санкт-Петербургский Политехнический Университет

Март 2025

Преподаватель: Кадырова Н.О.

Источник

Felipe Alonso-Atienza, et al., Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection, Expert Systems with Applications, Volume 39, Issue 2, 2012, Pages 1956-1967, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2011.08.051. (https://www.sciencedirect.com/science/article/pii/S0957417411011626)

Актуальность

- Желудочковая фибрилляция (Ventricular fibrillation, VF) - вызванное сбоями электрической активности сердца нарушение его ритмов, которое может привести к внезапной смерти.
- Успех дефибрилляции зависит от быстроты выявления нарушения.
- Точность существующих методов анализа ЭКГ ограничена.
- Методы машинного обучения позволяет улучшить детекцию VF, но требует эффективного выбора признаков.

Выбор признаков (Feature Selection, FS)

- Позволяет уменьшить размерность входных данных, сохраняя информативность.
- Улучшает скорость работы модели и уменьшает переобучение
- Используются методы фильтрации (filter methods), обёртки (wrapper methods) и встроенные методы (embedded methods).

Обзор существующих методов отбора признаков

- Фильтрация (Filter Methods): независимый анализ признаков (корреляция, статистические тесты).
- Обёртка (Wrapper Methods): тестирование различных подмножеств признаков с конкретной моделью.
- Встроенные методы (Embedded Methods): FS встроен в обучение модели (например, Recursive Feature Elimination, RFE).

База данных ЭКГ

Используемые базы данных:

- AHA Arrythmia Database, MIT-BIH Malignant Ventricular Arrythmia Database
- 29 записей пациентов, каждая длительностью около 30 минут
- Включает нормальные ритмы, тахикардию и фибрилляцию желудочков

Предобработка сигналов:

- Удаление среднего значения сигнала
- Фильтрация помех (низкочастотные и сетевые наводки)
- Разделение на сегменты по 1.024 секунды для анализа

Определение бутстрап-перевыборки

- $\boldsymbol{V} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x_N}, y_N)\}$, $\boldsymbol{x_i} \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- $\alpha = [\alpha_1, \ldots, \alpha_N] = s(\boldsymbol{V}, C, \sigma)$, s - оператор оптимизации SVM
- $R_{emp} = t(\alpha, \boldsymbol{V})$, t - оператор оценки эмпирического риска

- $\boldsymbol{V}^* = \{(\boldsymbol{x}_1^*, y_1), \ldots, (\boldsymbol{x_N^*}, y_N)\}$ - *бутстрап-перевыборка*. Какие-то пары из исходного множества могли попасть в $\boldsymbol{V}^*$ больше одного раза, а какие-то не попасть в него совсем.
- $\boldsymbol{V}^* = (\boldsymbol{V_{in}^*}, \boldsymbol{V_{out}^*})$
- $\alpha^* = s(\boldsymbol{V_{in}^*}, C, \sigma)$
- $R^* = t(\alpha, \boldsymbol{V_{out}^*})$
- делаем B независимых перевыборок

Исключение одного признака

- $\boldsymbol{W_u} = \{(\boldsymbol{x}_1^{(u)}, y_1), \ldots, (\boldsymbol{x}_{\boldsymbol{N}}^{(u)}, y_N)\},\ \boldsymbol{x}_{\boldsymbol{i}}^{(u)} \in \mathbb{R}^{d-1}$
- $\boldsymbol{W}_u^* = \{(\boldsymbol{x}_1^{*,(u)}, y_1), \ldots, (\boldsymbol{x}_{\boldsymbol{N}}^{*,(u)}, y_N)\}$
- $R_u^* = t(\alpha^*, \boldsymbol{W}^*_{\boldsymbol{u,out}})$
- $\Delta R_u^*(b) = R_u^*(b) - R^*(b)$ может быть вычислена для $\forall b = \overline{1, B}$

Мера важности признака

- В качестве меры риска в задаче бинарной классификации возьмём вероятность ошибки классификации P_e^*
- $\Delta P_e = P_{e,u} - P_{e,c}$
- Рассмотрим две гипотезы:
  1. $H_0 : \Delta P_e = 0 \Rightarrow$ признак u не важен
  2. $H_1 : \Delta P_e \neq 0 \Rightarrow$ признак u важен
- Распределение P неизвестно, так как неизвестны зависимости в парах $p(\boldsymbol{x_i}, y_i)$
- $\Delta P_e^*(b) = P_{e,u}^*(b) - P_{e,c}^*(b),\ b = \overline{1, B}$
- Можем вычислить парный доверительный интервал $z_{\Delta P_e^*}$
- Принимаем H_0 если $z_{\Delta P_e^*}$ имеет отрицательные значения или содержит нулевую точку, иначе принимаем вторую гипотезу

Алгоритм SVM-BR обратного выбора

1. Начать со всех признаков входного пространства V
2. Построить B парных бутстрап-перевыборок полной модели $\boldsymbol{V}^*$ и неполной $\boldsymbol{W}_u^*$
3. Для каждого бутстрап-образца b и для каждого признака u вычислить бутстрап-статистику

$$\Delta P_e^*(b) = P_{e,u}^*(b) - P_{e,c}^*(b),\ b = \overline{1, B}$$

и построить 95% $z_{\Delta P_e^*}$

4. Если $z_{\Delta P_e^*} < 0$ для какого-то признака u :
   - устранить признак u

   Иначе, если $z_{\Delta P_e^*} \subset 0$, *в зависимости от выбранной стратегии*:
   - удалить u с большим PCI
   - удалить u с меньшим PCI
5. Если есть ещё признак u, для которого $P_{e,u}^* < P_{e,c}^*$, то вероятность ошибки полной модели считать как $P_{e,c}^* = P_{e,u}^*$
6. Условие остановки: для любого признака выполняется $z_{\Delta P_e^*} > 0$. В противном случае возвращаемся к шагу 3

Разрешение неопределённостей

Что делать, если возникла неоднозначная ситуация: $z_{\Delta P_e^*} \ni 0$, т.е. невозможно статистически доказать, что признак влияет на ошибку. Предлагаются два дополнительных критерия выбора наименее полезного признака:

1. H-PCI: u - самый нерелевантный признак, если он имеет самый высокий $z_{\Delta P_e^*}$
2. S-PCI: u - самый нерелевантный признак, если его $z_{\Delta P_e^*}$ наименьший

На практике S-PCI часто лучше, он стабильно выбирает правильный признаки и даёт оптимальные результаты, тогда как H-PCI может ошибаться при сильной коллинеарности или шуме.

Применение метода на реальной задаче. Заключение

Метод SVM-BR с критерием S-PCI показал высокую эффективность при решении реальной задачи обнаружения фибрилляции желудочков по ЭКГ-сигналам. Было достигнуто сокращение числа признаков без ухудшения точности классификации.

Результаты сопоставимы или лучше, чем при использовании всех признаков и стандартных фильтрационных методов.

Методика обеспечивает надёжный, интерпретируемый и статистически обоснованный отбор признаков, что особенно ценно в биомедицинских задачах, где важна интерпретация результатов и устойчивость к шуму в данных.