Дифференциальная диагностика заболеваний по электрокардиограмме

автор задачи: *Воронцов Константин Вячеславович*, voron@forecsys.ru
консультант: *Целых Влада Руслановна*, celyh@inbox.ru

25 февраля 2014 г.

1 Постановка задачи

Предлагаемая задача является типичной задачей классификации. Объектами x являются больные. Признаками $f_1(x), \ldots, f_n(x)$ являются $n = 216$ характеристик, вычисляемых по электрокардиограмме. Способ их вычисления является секретным ноу-хау, но для решения данной задачи знать его не обязательно. Достаточно воспользоваться важной подсказкой, которую дают разработчики диагностической системы: для каждой болезни существуют признаки, имеющие, как правило, более высокие значения у больных данной болезнью. Вторая подсказка: эта задача неплохо решается! Ожидаемый уровень ошибок от 5% до 15%, в зависимости от заболевания.

Дано. В архиве `ekg-data.rar` находятся 10 файлов по 5 болезням, по 2 файла для каждой болезни. В каждом файле первый столбец содержит метки классов (0—здоров, 1—болен), следующие 216 столбцов — значения признаков.

Для каждой из 5 болезней есть два типа выборок. Эталонные — более надежные специально отобранные случаи (соответствующие файлы имеют в названии букву «Э»). Остальные — случаи, когда диагнозы устанавливались врачами менее надежно (не было всестороннего обследования), эти выборки рекомендуется использовать в качестве контрольных.

Найти. Построить алгоритм классификации, способный выдавать диагноз для произвольного вектора признаков. Провести оценку качества классификации на отложенной контрольной выборке.

Критерий. По контрольной выборке требуется рассчитать два критерия:

- доля ошибок первого рода — доля здоровых, которых алгоритм ошибочно классифицировал как больных;
- доля ошибок второго рода — доля больных, которых алгоритм ошибочно классифицировал как здоровых;

Предпочтение отдаётся минимизации ошибок второго рода.

Варианты исследования. Можно перемешивать эталонную и контрольную выборки и заново разбивать их на обучение и контроль. Это позволит увеличить объём выборки и обеспечить однородность обучения и контроля, возможно, немного пожертвовав чистотой данных.

Во всех 5 эталонных файлах выборка здоровых одна и та же, чтобы задачу можно было решать как 5 независимых двухклассовых задач. При желании можно собрать в один файл 5 выборок разных заболеваний и одну выборку здоровых (взяв её из любого из 5 файлов), чтобы решать 6-классовую задачу. Аналогичным образом устроены и файлы контрольных данных: в них также совпадает выборка здоровых.

2 Алгоритмы

Для решения задачи можно использовать любые методы машинного обучения.

Цель данного экспериментального исследования — найти модель классификации, наиболее подходящую для решения данной задачи.

2.1 Метрический алгоритм с жадным отбором признаков

Пусть x_i, $i = 1, \ldots, \ell$ — объекты обучающей выборки, u — классифицируемый объект. Определим взвешенную метрику Минковского:

$$\rho(u, x_i) = \left(\sum_{j=1}^{n} w_j \big| f_j(u) - f_j(x_i) \big|^p \right)^{\frac{1}{p}},$$

где w_j — неотрицательные веса признаков, $p > 0$.

Для произвольного $u \in X$ отсортируем объекты обучающей выборки $x_1, \ldots, x_\ell$ в порядке возрастания расстояния до u:

$$\rho(u, x_u^{(1)}) \leqslant \rho(u, x_u^{(2)}) \leqslant \cdots \leqslant \rho(u, x_u^{(\ell)}),$$

где $x_u^{(i)}$ — i-й обучающий сосед объекта u, $y_u^{(i)}$ — ответ на i-м соседе объекта u.

Метрический алгоритм классификации:

$$a(u; X^\ell) = \arg\max_{y \in Y} \sum_{i=1}^{\ell} \left[y_u^{(i)} = y \right] w(i, u),$$

где $w(i, u)$ — вес i-го соседа объекта u, неотрицательный, не возрастающий по i. Способы задания весов можно найти в [3].

Функция расстояния по j-му признаку:

$$\rho_j(u, x_i) = \big| f_j(u) - f_j(x_i) \big|.$$

Жадный алгоритм отбора признаков и оптимизации метрики основан на последовательном добавлении признаков по одному. Сначала $\rho^{(0)}(u, x_i) = 0$. На t-м шаге процесса строится метрика

$$\rho^{(t)}(u, x_i) = \rho^{(t-1)}(u, x_i) + w_j \rho_j(u, x_i),$$

при этом перебираются все признаки $j = 1, \ldots, n$, и для каждого признака перебираются значения весов w_j, например, по сетке или методом половинного деления. Определяется такая пара (j, w_j), при которой достигается минимум функционала скользящего контроля (leave-one-out):

$$\mathrm{LOO}(j, w_j) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \backslash \{x_i\}\big) \neq y_i \Big].$$

Можно также реализовать вариант алгоритма, когда на некоторых итерациях выбранный признак k заменяется другим признаком j:

$$\rho^{(t)}(u, x_i) = \rho^{(t-1)}(u, x_i) - w_k \rho_k(u, x_i) + w_j \rho_j(u, x_i).$$

Признаки добавляются, пока функционал LOO уменьшается.

На каждой итерации t предлагается выводить графики зависимости LOO от w_j для нескольких лучших признаков. Обязательно выводить два графика LOO, для ошибок первого и второго рода.

Возможна полуавтоматическая экспериментальная реализация алгоритма, когда решение о добавлении или замене признака производится на основании визуального анализа графика на каждой итерации.

2.2 Алгоритм вычисления оценок Ю. И. Журавлёва

2.3 Алгоритм голосования пороговых конъюнкций

2.4 Алгоритм решающих деревьев

2.5 Алгоритм SVM

2.6 Алгоритм логистической регрессии: LR, RLR, ElasticNet

3 Обоснования

Данные подготовлены по уникальной технологии информационного анализа электрокардиосигналов, разработанной проф. д.м.н. В. М. Успенским [1, 2].

Список литературы

[1] Успенский В. М. Информационная функция сердца // Клиническая медицина, 2008. — Т. 86, № 5. — С. 4–13.

[2] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. — М.: «Экономика и информация», 2008. — 116 с.

[3] Воронцов К. В. Метрические алгоритмы классификации. Лекции по машинному обучению. — 2014. http://www.MachineLearning.ru/wiki/images/c/c3/Voron-ML-Metric-slides.pdf