БДЗ (Теория решеток для анализа данных)

Дворчик Максим М05-114д

Характеристика датесета

Ссылка:
https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Датасет был собран с помощью прямых анкет пациентов, которые обращались в Sylhet Diabetes Hospital в Силхете, Бангладеш и одобрен врачом. Датасет содержит 520 кортежей и он охарактеризован 17-ю признаками:

- Age 1.20-65
- Sex 1. Male, 2.Female
- Polyuria 1.Yes, 2.No.
- Polydipsia 1.Yes, 2.No.
- sudden weight loss 1.Yes, 2.No.
- weakness 1.Yes, 2.No.
- Polyphagia 1.Yes, 2.No.
- Genital thrush 1.Yes, 2.No.
- visual blurring 1.Yes, 2.No.
- Itching 1.Yes, 2.No.
- Irritability 1.Yes, 2.No.
- delayed healing 1.Yes, 2.No.
- partial paresis 1.Yes, 2.No.
- muscle stiffness 1.Yes, 2.No.
- Alopecia 1.Yes, 2.No.
- Obesity 1.Yes, 2.No.
- Class 1.Positive, 2.Negative.

Целевой признак: наличие диабета (да, нет)

Задача классификации

Рассмотрим задачу бинарной классификации. В качестве положительных примеров будем использовать те кортежи, у которых номер целевого признака=1(являться мужчиной, да, позитивный), а в качестве отрицательных - 2(являться женщиной, нет, отрицательный).

В кортежах есть небинарный признак (возраст), который мы оставим без изменений (просто число), и будем обрабатывать отдельно в одном из алгоритмов.

Формирование решетки понятий

Перед началом работы строим решетку понятий наивным алгоритмом, опираясь на бинарные признаки:

1. Итерируемся по всем признакам и объектам
2. На каждой итерации делаем двойное замыкание
3. Если $G_i = G_i''$ или $M_i = M_i''$, то формируем понятие[1] из (G_i и M_i), если нет, то продолжаем замыкать

1. Формирование понятия:

Из множества объектов, входящих в понятие агрегируем следующие признаки:

- $class$ (целое число), который считается по такому правилу: если целевой признак объекта положительный, то прибавляем 1, если отрицательный, то отнимаем 1
- age, который считаем как среднее арифметическое возрастов объектов

Из получившихся понятий формируем решетку понятий

Алгоритмы работы с решеткой понятий

Алгоритм 1.

Алгоритм основан на суммировании классификатора с учетом количества пересечений признаков:

$$class_t = \sum_{i=0}^{n} \frac{class_i}{d(M_i, M_t) + 1}$$

Где $d(M_i, M_t)$ – это расстояние Хэмминга между множеством признаков понятия и классифицируемого объекта.

Если $class_t > 0$, то считаем объект положительным, нет - отрицательным

Алгоритм 2.

Алгоритм основан на суммировании классификатора с учетом количества пересечений признаков и возраста:

$$class_t = \sum_{i=0}^{n} \frac{class_i}{d(M_i, M_t) + \frac{|age_i - age_t|}{5} + 1}$$

Где $d(M_i, M_t)$ – это расстояние Хэмминга между множеством признаков понятия и классифицируемого объекта,

age_i, age_t – агрегированный возраст в понятии и возраст в объекте соответственно.

Если $class_t > 0$, то считаем объект положительным, нет - отрицательным

Алгоритм 3.

Алгоритм основан на суммировании классификатора с учетом квадрата количества пересечений признаков:

$$class_t = \sum_{i=0}^{n} \frac{class_i}{d(M_i, M_t)^2 + 1}$$

Где $d(M_i, M_t)$ – это расстояние Хэмминга между множеством признаков понятия и классифицируемого объекта.

Если $class_t > 0$, то считаем объект положительным, нет - отрицательным

Алгоритм 4.

Алгоритм основан на суммировании классификатора с учетом квадрата количества пересечений признаков и возраста:

$$class_t = \sum_{i=0}^{n} \frac{class_i}{d(M_i, M_t)^2 + \frac{|age_i - age_t|}{5} + 1}$$

Где $d(M_i, M_t)$ – это расстояние Хэмминга между множеством признаков понятия и классифицируемого объекта,

age_i, age_t – агрегированный возраст в понятии и возраст в объекте соответственно.

Если $class_t > 0$, то считаем объект положительным, нет - отрицательным

Проверка валидности

Для обучения из всей выборки случайным образом отбирал 75% положительных и 75% отрицательных примеров (в самой выборке ~61,5% положительных примеров).

В качестве метрик были использованы:

- True Positive Rate
- True Negative Rate
- Negative Predictive Value
- False Positive Rate
- False Discovery Rate
- Accuracy
- Precision
- Recall

Сравнение качества производилось по усредненным (средним арифметическим) метрикам на основании следующего метода: объекты для

обучения каждый раз выбираем случайно, объекты для классификации – из оставшихся, и выполняем каждый алгоритм 50 раз.

Результаты:

В таблице приведены значения метрик алгоритмов 1, 2, 3 и 4.

Метрика	Алгоритм 1	Алгоритм 2	Алгоритм 3	Алгоритм 4
True Positive Rate	0.951	0.96774983	0.96100014	0.9657499
True Negative Rate	0.6820001	0.38719997	0.76519996	0.52879995
False Positive Rate	0.318	0.6128	0.23479995	0.47119996
False Negative Rate	0.048999995	0.03225	0.039	0.034250002
False Discovery Rate	0.17286366	0.2835462	0.1324757	0.23368378
Accuracy	0.84753853	0.7444614	0.8856924	0.79769224
Precision	0.82713634	0.7164538	0.8675243	0.76631624
Recall	0.95100003	0.96774995	0.96099997	0.96575004

Дополнительные наблюдения

В случае положительного диабета чаще всего наблюдались следующие симптомы:

- polyuria
- polydipsia
- weakness

Что подтверждается данными из википедии:
https://ru.wikipedia.org/wiki/Сахарный_диабет

Итоги

Все алгоритмы довольно точно выявляют положительные случаи, но дают относительно много ложноположительных результатов, при этом там, где учитывается возраст точность выявления положительных случаев незначительно возрастает, а точность выявления отрицательных случаев значительно падает. Скорее всего это связано с небольшим объемом выборки относительно количества признаков для таких алгоритмов.

Таким образом, получившуюся систему можно применять в индивидуальных случаях из-за относительно большого Recall, но для массовой диагностики она не годится, так как создаст дополнительную нагрузку на здравоохранение из-за относительно высокого False Discovery Rate, так как много людей просто так будут отправлены на дообследование.