Национальный исследовательский университет – Высшая Школа Экономики

Упорядоченные множества в анализе данных. Отчет по заданию «создание алгоритма бинарной классификации».

Набор данных: “Выбор метода контрацепции”.

Студент: Герман Соколов

1 курс магистратуры
«Науки о данных»

GitHub link: https://github.com/german3d/FCA-lattices

Москва, 2015

Описание данных

Для тестирования работы алгоритма использовался набор данных «Contraceptive Method Choice Data Set» из репозитория UCI Machine Learning Repository. Данные является выборкой из совокупности результатов национального опроса, который проводился среди замужних женщин разного возраста в Индонезии в 1987 году.

Задача – определение метода контрацепции, которым пользуется респондент, исходя из его демографических и социо-экономических характеристик.

Количество объектов – 1 473

Количество признаков (исходное) – 9:

- Возраст (количественный)
- Образование (категориальный, 1=low, 2, 3, 4=high)
- Образование супруга (категориальный, 1=low, 2, 3, 4=high)
- Число детей в семье (количественный)
- Религия (бинарный, 0=Non-Islam, 1=Islam)
- Наличие работы (бинарный, 0=Yes, 1=No)
- Работа супруга (категориальный, 1, 2, 3, 4)
- Уровень жизни (категориальный, 1=low, 2, 3, 4=high)
- Внимание СМИ к респонденту (бинарный, 0=Good, 1=Not good)
- Метод контрацепции (категориальный, 1=No-use, 2=Long-term, 3=Short-term)

Так как задача – создание *бинарного* классификатора, то атрибут класса «метод контрацепции» был преобразован в бинарный признак «Пользование контрацептивами» (“x”=Да, “o”=Нет). Так как количество наблюдений в обоих классах примерно одинаково, то такое разбиение возможно.

Все остальные категориальные признаки были преобразованы в набор бинарных признаков, например, атрибут «Образование» был переписан в виде четырех признаков в зависимости от уровня. Бинарные признаки остались без изменения. Для количественных признаков применялось шкалирование с последующей конвертаций интервалов в набор бинарных признаков в соответствии со следующей логикой:

Бинарные признаки для атрибута «Возраст»:

- 21 год и младше
- 22 – 30 лет
- Старше 30 лет

Бинарные признаки для атрибута «Число детей в семье»:

- 0 (бинарный)
- 1-3
- 3-5
- Больше 5

Именно такие интервалы для признака «Возраст» в среднем определяют основные этапы жизни человека, когда может измениться система взглядов и ценностей

Выделение отдельного бинарного признака «нет детей» обосновано, так как очевидно сильно связано с принадлежностью к тому или иному классу.

Описание алгоритма

Для проведения классификации было использовано 3 различных lazy-метода, результаты применения которых сравнивались между собой для нахождения наилучшего алгоритма:

1. **Исходный** алгоритм на основе «генераторов» с использованием критерия поддержки. Данный метод вместе с кодом был описан в руководстве по созданию бинарного классификатора.
2. **Наивный байесовский классификатор** – активно и повсеместно используемый метод классификации.
3. **Предлагаемый** алгоритм. Принцип его работы:
    a. Классифицируемый объект пересекается по множеству своих признаков с каждым из объектов из положительного и отрицательного классов.
    b. Находится наибольшее по мощности пересечение с положительным и отрицательным классом. Если мощность первого больше второго – объекту присваивается положительный класс, и наоборот.
    c. Если найденные пересечения одинаковы по мощности, то используется критерий *количества* найденных пересечений такой мощности в каждом из классов. Если число пересечений данной (максимальной) мощности больше в положительном классе, то объект классифицируется как положительный
    d. Если количество пересечений максимальной мощности одинаково в обоих классах, то новый критерий – количество пересечений мощности, равной максимальной минус 1.
    e. И так далее, пока для какого-либо из критериев число элементов в одном классе превысит число элементов в другом

Результаты применения классификаторов

В качестве метрик качества были использованы все предлагаемые в руководстве метрики. Сравнение качества производилось по усредненным (средним арифметическим) метрикам на основании использования метода кросс-валидации (K=5).

В качестве проверки алгоритма предлагаемого №3 с другими (альтернативными) параметрами предлагается использовать не абсолютный, а относительный критерий количества пересечений в том и другом классе.

Таблица с результатами представлена ниже:

Таблица 1. Метрики качества

Метрика	Исходный	Наивный Байес	Предло-женный (абс.)	Предло-женный (отн.)
True Positive	639	605	798	741
True Negative	359	366	449	506
False Positive	261	263	179	123
False Negative	218	239	45	103
Negative Predictive Value	62%	61%	91%	83%
False Positive Rate	42%	42%	29%	20%
False Discovery Rate	30%	31%	21%	15%
Accuracy	68%	66%	85%	85%
Precision	71%	70%	82%	86%
Recall (True Positive Rate)	74%	72%	95%	88%
Specificity (True Negative Rate)	58%	58%	71%	80%

Можно сделать вывод, что исходный и наивный байесовский классификаторы обладают недостаточным качеством предсказания класса объекта, причем первый в среднем совсем немного лучше второго.

Предложенный классификатор на базе относительного критерия в среднем лучше распознает объекты из «отрицательного» класса (респондент не использует контрацептивы) – specificity значительно выше, при этом у данного классификатора хуже метрика recall (угадывание положительного класса). В целом, у данного классификатора значения метрик распределены более равномерно. Тем не менее, конечный выбор классификатора целиком зависит от решаемой задачи.

GitHub link: https://github.com/german3d/FCA-lattices