Отчет по БДЗ. Курс: Упорядоченные множества в анализе данных.
Иванова Полина

Задача бинарной (ленивой) классификации решалась с помощью двух алгоритмов.

Для первого алгоритма, при попытке классифицировать поступивший объект g, считаем величину поддержки в положительном G^+ и отрицательном G^- контекстах:

$$\frac{|\left(g' \cap g_i^+\right)^+|}{|G^+|}, \frac{|\left(g' \cap g_j^-\right)^-|}{|G^-|},$$

где g_i^+ и g_j^- — элементы G^+ и G^- соответственно. Если вычисленная поддержка в положительной контексте (в отрицательном контексте) больше некоторого минимального порога (threshold), то объекту добавляется положительная (соответственно, отрицательная) метка. Если объекту присвоились и положительная, и отрицательная метки, то классификация противоречива. В данном алгоритме возможны также случаи, когда классификация не проводится: если поддержки и для положительного, и для отрицательного контекстов меньше порогового значения (threshold).

Второй алгоритм учитывает число элементов g_i^+ из положительного контекста тренировочной выборки, таких что для поступившего элемента g_τ признаки $g'_\tau \cap g_i^+$ содержатся в некотором описании элемента положительного контекста, при этом не содержатся в описаниях элементов отрицательного контекста. То есть, если для элемента g_i^+ выполнены условия:

$$|g'_\tau \cap g_i^+| > 0 \text{ и } \frac{|\left(g'_\tau \cap g_i^+\right) \subseteq g_j^-|}{|G^-|} = 0,$$

то учитываем элемент g_i^+ в «голосовании» за положительную классификацию. Аналогично, считается доля соответствующих элементов из отрицательного контекста. Затем доля таких элементов в каждом контексте сравнивается с некоторым пороговым числом threshold: если доля больше, чем threshold, то элементу присваивается соответствующая метка.

Алгоритмы проверялись на части массива **Tic-Tac-Toe** (а именно **train2.csv**, **test2.csv**).

Для первого алгоритма после применения метода скользящего контроля (k-fold cross-validation с $k = 5$) был выбран threshold= 0.125. В качестве метрики качества использовалась точность. Результаты в таблице **??**. Метод не показал себя с хорошей стороны, так как и величина поддержки, и точность оценивания не велики.

Таблица 1: Результаты применения Алгоритма 1. 'Not classified' — доля элементов, для которых не был определен класс (не ставилась ни одна из меток 'positive' и 'negative')

K-fold cross validation: $k = 5$		
threshold	accuracy	not classified
0.1	0.02	0.03
0.125	0.19	0.22
0.13	0.16	0.33
Тестовая выборка		
threshold	accuracy	not classified
0.125	0.216	0.44

Для второго алгоритма также использоввлся метод скользящего контроля с $k = 5$ для выбора порогового значения threshold. Результаты показаны в таблице **??**.

Можем видеть, что Алгоритм 1 значительно проигрывает Алгоритму 2, как и при применении метода скользящего контроля, так и на тестовой выборке.

Также, были попытки сделать несколько модификаций. Для Алгоритма 1 можно смотреть не только на поддержку, но и на "фальсифицируемость"гипотез, то есть на величины

$$\frac{|\left(g' \cap g_i^+\right)^-|}{|G^+|}, \frac{|\left(g' \cap g_j^-\right)^+|}{|G^-|},$$

Таблица 2: Результаты применения Алгоритма 2. 'Not classified' — доля элементов, для которых не был определен класс (не ставилась ни одна из меток 'positive' и 'negative')

K-fold cross validation: $k = 5$		
threshold	accuracy	not classified
0.5	0.51	0.03
0.2	0.33	0.6
0.1	0.95	0.01
Тестовая выборка		
threshold	accuracy	not classified
0.5	0.51	0.03
0.1	0.93	0.05

и сравнивать эти величины с некоторым максимальным пороговым значением. Однако на рассмотренном примере данных модификация не дала значительных улучшений.

Для Алгоритма 2 можно сравнивать описанные выше доли для положительного и отрицательного контекстов не с некоторым пороговым значением, а между собой, присваивая ту метку ('positive' или 'negative'), для которой соответствующая доля выше (и присваивая объекту обе метки, если разница между долями меньше, например, 0.01). На тестовой выборке **test2.csv** такой метод дал точность 95%, что примерно совпало с точностью Алгоритма 2 на **test2.csv**.

В качестве другой модификации, можно разрешить небольшой уровень "фальсифицируемости"для "голосующих"элементов контекста. То есть, если для элемента g_i^+ проверяем выполнимость условий:

$$|g'_\tau \cap g_i^+| > 0 \text{ и } \frac{|\left(g'_\tau \cap g_i^+\right) \subseteq g_j^-|}{|G^-|} < \varepsilon.$$

Аналогичные двойственные условия проверяются для элементов g_j^- отрицательного контекста.

Эта модификация Алгоритма 2 проверялась для $\varepsilon = 0.05$ и для разных пороговых значений долей голосующих тренировочных элементов (от 0.4 до 0.7). Точность для каждого проверяемого порогового значения была около 50%, что значительно меньше точности Алгоритма 2.