Упорядоченные Множества в Анализе Данных
Большое Домашнее Задание

Макаров Михаил

Цель данной работы заключается в решении задачи бинарной классификации: целевая переменная принимает два значения. Дана обучающая выборка, для которой известны значения целевой переменной. То есть обучающая выборка разбивается на два класса G^+ и G^- в зависимости от целевой переменной. Есть тестовая выборка, для которой необходимо "предсказать"значения целевой переменной. Тестирование алгоритмов проводилось на двух наборах данных: Tic-Tac-Toe из UCI Machine Learning Repository и Titanic из Kaggle.

Для решения данной задачи в рамках работы используются два алгоритма.

Алгоритм 1. Имеет два параметра $card$ и ε. Для классифицируемого объекта g необходимо выполнить:

- Для каждого объекта g^+ из G^+, если $|g \cap g^+| > card$, то если $\frac{|(g \cap g^+)^-|}{|G^-|} \leq \varepsilon$, тогда объект g^+ голосует за g.
- Для каждого объекта g^- из G^-, если $|g \cap g^-| > card$, то если $\frac{|(g \cap g^-)^+|}{|G^+|} \leq \varepsilon$, тогда объект g^- голосует за g.
- Если количество положительных голосов $\geq$ количества отрицательных голосов, то классифицируем как $+$.

Алгоритм 2. Имеет два параметра sup и ε. Для классифицируемого объекта g необходимо выполнить:

- Для каждого объекта g^+ из G^+, если $|g \cap g^+| > 0$, то если $\frac{|(g \cap g^+)^-|}{|G^-|} \leq \varepsilon$ и $\frac{|(g \cap g^+)^+|}{|G^+|} > sup$, тогда классифицируем объект g как $+$.
- Для каждого объекта g^- из G^-, если $|g \cap g^-| > 0$, то если $\frac{|(g \cap g^-)^+|}{|G^+|} \leq \varepsilon$ и $\frac{|(g \cap g^-)^-|}{|G^-|} > sup$, тогда классифицируем объект g как $-$.
- По умолчанию классифицируем как $+$.

Tic-Tac-Toe. Для тестирования точности работы алгоритмов будем использовать кросс-валидацию: разобьем данные на k частей и последовательно будем обучать модель на $k-1$ частях, а тестировать точность

$card$	ε	$accuracy$
0.35	0.45	0.555
0.40	0.80	0.508
0.50	0.20	0.717
0.60	0.00	0.962

Таблица 1: Значения $accuracy$ для некоторых параметров (*Алгоритм 1*)

на оставшейся части. Для подбора оптимальных параметров будем использовать метрику точности (Accuracy).

Алгоритм 1. Для проверки каждой пары параметров используется перебор $card$ от 0 до 0.9 с шагом 0.1 и перебор ε от 0 до 0.9 с шагом 0.05 (Таблица 1). Наибольшее значение $accuracy = 0.962$ достигается при $card = 0.6$ и $\varepsilon = 0$.

Алгоритм 2. Для проверки каждой пары параметров используется перебор sup от 0 до 0.95 с шагом 0.05 и перебор ε от 0 до 0.9 с шагом 0.05 (Таблица 2). Наибольшее значение $accuracy = 0.941$ достигается при $sup = 0.1$ и $\varepsilon = 0$.

sup	ε	$accuracy$
0.05	0.40	0.640
0.10	0.00	0.717
0.10	0.00	0.941
0.15	0.05	0.752

Таблица 2: Значения $accuracy$ для некоторых параметров (*Алгоритм 2*)

По результатам тестирования можно сделать вывод, что оба алгоритма показывают высокий уровень точности при подборе соответствующих параметров. Также на данном наборе данных параметр погрешности ε не помог увеличить точность (оптимальное значение равно 0).

Titanic. В данном наборе данных присутствуют числовые (некатегориальные) признаки. Например, возраст, плата за поездку, количество родственников. Поэтому было проведено шкалирование - разбивка данных на бины и получение по ним категориальных признаков. Тестирование также проводилось с использованием кросс-валидации.

Алгоритм 1. Для проверки каждой пары параметров используется перебор $card$ от 0 до 0.9 с шагом 0.1 и перебор ε от 0 до 0.9 с шагом 0.05 (Таблица 3). Наибольшее значение $accuracy = 0.796$ достигается при $card = 0.80$ и $\varepsilon = 0.05$.

Алгоритм 2. Для проверки каждой пары параметров используется

$card$	ε	$accuracy$
0.20	0.20	0.792
0.60	0.65	0.786
0.80	0.00	0.571
0.80	0.05	0.796
0.90	0.05	0.723

Таблица 3: Значения $accuracy$ для некоторых параметров (*Алгоритм 1*)

перебор sup от 0 до 0.95 с шагом 0.05 и перебор ε от 0 до 0.9 с шагом 0.05 (Таблица 4). Наибольшее значение $accuracy = 0.795$ достигается при $sup = 0.45$ и $\varepsilon = 0.2$.

sup	ε	$accuracy$
0.15	0.10	0.705
0.25	0.10	0.746
0.45	0.2	0.795
0.65	0.35	0.783

Таблица 4: Значения $accuracy$ для некоторых параметров (*Алгоритм 2*)

Для данного набора данных оба алгоритма также показывают хорошие результаты. Более того, параметр ε помогает увеличить точность.

Нахождение глобального оптимума. Для нахождения глобального оптимума посчитаем среднее значение точности для Tic-Tac-Toe и Titanic для разных значений параметров. Для Алгоритма 1 наиболее оптимальные параметры: $card = 0.70$ и $\varepsilon = 0.10$, при которых $avg\_acc = 0.866$ (Таблица 5).

$card$	ε	$acc\_titanic$	acc_tic_tac	avg_acc
0.10	0.20	0.793	0.598	0.695
0.20	0.40	0.764	0.599	0.682
0.60	0.00	0.585	0.962	0.773
0.70	0.10	0.789	0.942	0.866

Таблица 5: Значения $accuracy$ для некоторых параметров (*Алгоритм 1*)

Для Алгоритма 2 наиболее оптимальные параметры: $sup = 0.20$ и $\varepsilon = 0.05$, при которых $avg\_acc = 0.757$ (Таблица 6).

По результатам данной работы можно сделать вывод о том, что оба рассмотренных алгоритма могут быть использованы для решения задачи бинарной классификации. Однако, следует понимать, что данные

$card$	ε	$acc\_titanic$	acc_tic_tac	avg_acc
0.05	0.80	0.384	0.640	0.512
0.15	0.05	0.755	0.752	0.753
0.20	0.05	0.762	0.752	0.757
0.90	0.30	0.782	0.657	0.719

Таблица 6: Значения *accuracy* для некоторых параметров (*Алгоритм 2*)

алгоритмы, скорее всего, не подойдут для классификации больших данных. Т.к. при классифицации каждого объекта используется весь набор данных.