НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
Факультет компьютерных наук

Упорядоченные множества в анализе данных

Отчет о выполнении задания на тему:
«Модификация алгоритма ленивой бинарной классификации»

Выполнил:
студент группы мНоД16_ИССА
программы «Науки о данных»
Тетин Е. Н.

Москва, 2016

Введение

Целью данной работы является модификация предложенного алгоритма ленивой бинарной классификации, основанного на использовании «генераторов», с последующим анализом метрик его качества.

В качестве данных для классификации использовался массив Tic-Tac-Toe из репозитория UCI Machine Learning Repository.

Описание алгоритма

На вход алгоритму подается файл с данными, для которых известен целевой класс (в качестве аргумента принимается имя csv-файла без расширения). Используется метод скользящего контроля для оценки качества алгоритма (k-fold cross-validation). Содержимое файла делится на k частей: на i-ой части происходит «обучение», на (i+1)-ой части проводится классификация объектов (i = 1, 2, … k-1). На каждом шаге вычисляются метрики качества, которые затем усредняются по k шагам.

В результате «обучения» строятся плюс- и минус-контексты, содержащие объекты с «положительной» классификацией (принадлежностью к классу) и «отрицательной» классификацией соответственно.

Процесс классификации заключается в следующем. По очереди рассматриваются все классифицируемые объекты; для них проверяются гипотезы о «положительной» либо «отрицательной» классификации. При проверке «положительной» гипотезы для каждого объекта из плюс-контекста строится пересечение его описания с описанием классифицируемого объекта, после чего проверяется, вкладывается ли это пересечение в описание какого-либо из объектов минус-контекста. Аналогично выполняется проверка «отрицательной» гипотезы.

Решение о классификации конкретного объекта может приниматься в зависимости от значения пороговой функции. Изначально в качестве такой функции рассматривается поддержка пересечения описаний. Варианты пороговой функции рассматриваются как модификации предложенного алгоритма.

Модификации алгоритма

0. В качестве значения пороговой функции принимается поддержка пересечения описаний классифицируемого объекта и объекта из плюс(минус)-контекста, или его «вес». «Положительная» классификация объекта, таким образом, проводится при условии, что «вес» в плюс-контексте больше «веса» в минус-контексте.
1. В качестве критерия «положительной» классификации объекта принимается наибольшая мощность пересечения его описания с описаниями плюс(минус)-контекста.

Оценка качества алгоритма

Accuracy = (TP + TN)/N – доля верных классификаций.

Precision = TP/(TP + FP) – доля верных классификаций среди объектов, классифицированных как «положительные».

Recall = TP/(TP + FN) – доля верных классификаций среди истинно «положительных» объектов.

Specificity = FP/(FP + TN) – доля ошибочных «положительных» классификаций.

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Объекты	Алгоритм (мод. 0)	Алгоритм (мод. 1)
True Positive	383	479
False Positive	182	172
True Negative	210	128
False Negative	90	86

Таблица 1. Классифицированные объекты

Метрика (усредненная)	Алгоритм (мод. 0)	Алгоритм (мод. 1)
Accuracy	69%	70%
Precision	68%	74%
Recall	81%	85%
Specificity	46%	57%

Таблица 2. Метрики качества

Вывод

Алгоритм с критерием классификации, основанном на вычислении мощности пересечения описаний объектов (мод. 1), показал несколько лучший результат по точности, хотя и с увеличенной долей ошибочных «положительных» классификаций. Алгоритм, использующий критерий поддержки для классификации объектов (мод. 0), дал результат, не сильно уступающий результату другого алгоритма, но при этом работал значительно дольше. В целом алгоритм с мод. 1 является более предпочтительным.