

Ленивая классификация для объектов, представленных бинарными признаками на примере данных из репозитория tic-tac-toe

Было имплементировано 4 алгоритма:

- Метод ближайших соседей с метрикой Хэмминга (т.е. разное значение одинакового признака +1 к расстоянию). Параметр метода k – число соседей, которых нужно рассмотреть (knn.py)
- Метод ближайших соседей с метрикой вычисляемой следующим образом: одинаковое значение признаков +0, пары (o b) (x b) +1, пара (x o) +2. Параметр метода k – число соседей, которых нужно рассмотреть (knn2.py)
- Метод вычисляющей метрику близости классифицируемого объекта к положительному и отрицательному классам посредством вычисления суммарного пересечения с контекстами (вычисляется мощность пересечения классифицируемого объекта с каждым объектом положительного контекста отдельно, мощности складываются, делятся на мощность контекста, умноженную на кол-во признаков). Объект классифицируется к тому классу, к которому он “ближе”. Параметр метода α – пороговое отношение мощности пересечения образца с объектом из контекста к мощности признаков образца, при котором мощность этого пересечения учитывается (test_.py)
- Метод на основе генераторов: вычисляем пересечение классифицируемого объекта с объектом положительного (отрицательного) контекста, после чего подсчитываем количество объектов положительного и отрицательного контекста, если мощность положительных (отрицательных) объектов будет выше отрицательных (положительных) то считаем, что данное пересечение голосует за положительную (отрицательную) классификацию, иначе не голосует. Параметр метода α – пороговое отношение мощности пересечения образца с объектом из контекста к мощности признаков образца, при котором мощность этого пересечения учитывается (test_generators.py)

Рассмотрим полученные результаты:

Метод k ближайших соседей с метрикой Хэмминга хорошо работает при значениях $k=1-10$, показывая следующие результаты:

Accuracy of the procedure is : 99.06%,

Fall-out (FPR) of the procedure is : 2.68%,

False omission rate (FOR) of the procedure is : 0.0%,

Precision (PPV) of the procedure is : 98.58%,

Sensitivity (TPR) of the procedure is : 100.0

При $k > 10$ результаты ухудшаются (чем больше k , тем хуже результат).

Метод k ближайших соседей с метрикой, учитывающий тот эмпирический факт, что разница между "х" и "о" более категорична, нежели разница между "х" и "б" ("b" и "о"), при значении $k \leq 3$ классифицирует правильно 100% тестовых примеров:

Accuracy of the procedure is : 100.0%,

Fall-out (FPR) of the procedure is : 0.0%,

False omission rate (FOR) of the procedure is : 0.0%,

Precision (PPV) of the procedure is : 100.0%,

Sensitivity (TPR) of the procedure is : 100.0%

При значении $k > 3$ результаты ухудшаются (чем больше k , тем хуже результат).

Метод с метрикой, вычисляющийся через попарное пересечение образца с каждым из объектов, при значениях параметра $\alpha > 0.7$ не классифицирует ни одного объекта. При $\alpha = 0.7$ достигается наилучший для этого метода результат с параметрами:

Accuracy of the procedure is : 98.59%,

Fall-out (FPR) of the procedure is : 0.0%,

False omission rate (FOR) of the procedure is : 3.73%,

Precision (PPV) of the procedure is : 100.0%,

Sensitivity (TPR) of the procedure is : 97.87%

При $\alpha < 0.7$ результат ухудшается (чем меньше α , тем хуже результат)

Метод на основе генераторов при значениях параметра $\alpha > 0.7$ не классифицирует ни одного объекта. При $\alpha = 0.7$ достигается наилучший для этого метода результат с параметрами:

Accuracy of the procedure is : 99.06%,

Fall-out (FPR) of the procedure is : 2.68%,

False omission rate (FOR) of the procedure is : 0.0%,

Precision (PPV) of the procedure is : 98.58%,

Sensitivity (TPR) of the procedure is : 100.0%

При $\alpha < 0.7$ результат ухудшается (чем меньше α , тем хуже результат)

Вывод: хорошие результаты всех методов свидетельствует о том, что сама специфика dataset Tis-Tac-Toe такова, что объекты близко находящиеся к другу к другу (в смысле некоей адекватно подобранной меры близости), скорее всего, будут обладать схожими метками класса.