1 Постановка задачи

Решаем задачу бинарной классификации. Есть обучающая выборка, для которой известны метки классов. Есть тестовая выборка, для которой по зачениям других признаков надо восстановить значение класса.

2 Общая схема работы алгоритма

Пусть G- множество объектов,
G_+ - множество положительных объектов,
G_- - множество отрицательных объектов,
G_t - тестовые объекты;

Для каждого отдельного элемента(g) из G составляем новую таблицу признаков. Для начала удаляем g из G (а также из G_+ или G_-).

Для g выполняем:

1) Для каждого объекта из плюс контекста рассмотрим пересечение его описания (g_i^+) и описания g', и проверим вкладывается ли оно в описание какого-либо из минус примеров (g_i^-)

2) Для каждого объекта из минус контекста рассмотрим пересечение его описания (g_i^-) и описания g', и проверим вкладывается ли оно в описание какого-либо из плюс примеров (g_i^+)

Для каждого g рассмотрим новые признаки:
1) $min\sum_{i\in i^+} |g' \cap g_i^+|$
2) $mean\sum_{i\in i^+} |g' \cap g_i^+|$
3) $max\sum_{i\in i^+} |g' \cap g_i^+|$
4) $min\sum_{i\in i^+} |(g' \cap g_i^+)^+|$
5) $mean\sum_{i\in i^+} |(g' \cap g_i^+)^+|$
6) $max\sum_{i\in i^+} |(g' \cap g_i^+)^+|$
7) $min\sum_{i\in i^+} |(g' \cap g_i^+)^-|$
8) $mean\sum_{i\in i^+} |(g' \cap g_i^+)^-|$
9) $max\sum_{i\in i^+} |(g' \cap g_i^+)^-|$
10) $min\sum_{i\in i^-} |g' \cap g_i^-|$
11) $mean\sum_{i\in i^-} |g' \cap g_i^-|$
12) $max\sum_{i\in i^-} |g' \cap g_i^-|$
13) $min\sum_{i\in i^-} |(g' \cap g_i^-)^-|$
14) $mean\sum_{i\in i^-} |(g' \cap g_i^-)^-|$
15) $max\sum_{i\in i^-} |(g' \cap g_i^-)^-|$
16) $min\sum_{i\in i^-} |(g' \cap g_i^-)^+|$
17) $mean\sum_{i\in i^-} |(g' \cap g_i^-)^+|$
18) $max\sum_{i\in i^-} |(g' \cap g_i^-)^+|$

Строим такой же набор признаков для тестовых объектов.
Чтобы получить оптимальные пороговые функции классификации, по обу-

чающей выборке строим дерево принятия решений и используем его на тестовых объектах.

3 Пример использования алгоритма

Проверим на двух наборах данных:

1)Congressional Voting Records Data Set
http://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data
Данные имеют бинарную природу, шкалирование не требуется.
Используем 5-fold cross validation.
Результаты: Precision - 0.94590 Recall - 0.872165 Accuracy - 0.91739
На этом наборе данных основная функция классификации: $mean\sum_{i\in i^+}|(g'\cap g_i^+)^-|$
. Пороговое значение - 1.8452

2)Mushroom Database
https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data Number of Attributes: 22 (all nominally valued) - необходимо выполнить шкалирование.
Используем 5-fold cross validation.
Результаты: Precision - 0.8 Recall - 0.7511 Accuracy - 0.756
На этом наборе данных основная функция классификации: $max\sum_{i\in i^+}|g'\cap g_i^+|$
Пороговое значение - 20