ВМК МГУ
Кафедра математических методов прогнозирования

Отчет о выполнении задания 6
«Краткая характеристика пакета Boruta системы R»

Студент: Исмагилов Тимур Ниязович

2012

Цель задания

Требуется написать краткое пособие для работы с одним из пакетов системы R.

Я выбрал пакет Boruta — алгоритм отбора значимых признаков. Суть алгоритма в том, что признаки, которые могут оказаться лишними, проверяют, используя вместо них случайные величины и запуская randomForest.

Этот пакет был выбран, так как задача выбора фиктивных признаков будет решаться нами в задании номер 7.

Описание интерфейса

Реализация алгоритма — функция Boruta:
Boruta(x,y,confidence=0.999,maxRuns=100,light=TRUE,doTrace=0,getImp=getImpRf,...):

- *x* — вектор признаков.
- *y* — вектор ответов.
- *confidence* — необходимая уверенность в ответе, для того чтобы его дать. Значение не следует менять, кроме как с целью ускорения работы алгоритма.
- *maxRuns* — ограничение сверху на количество выполнений алгоритма randomForest. Следует увеличить, если алгоритм не дает ответа для некоторых атрибутов.
- *doTrace* — 0, 1, или 2 — при значении 1 выводит точку после каждого выполнения randomForest, при значении 2 — еще и дополнительную отладочную информацию.
- *getImp* — функция, возвращающая значимость атрибута. Стандартная функция — *getImpRf*, использует randomForest и Z-меру уменьшения точности (при замене данных случайными).
- *light* — TRUE или FALSE — позволяет запустить алгоритм в менее точном, «облегченном» режиме, в котором признаки, оказавшиеся бесполезными, сразу же удаляются.

Для того, чтобы получить случайные значения признака, имеющиеся значения просто перемешиваются. Алгоритм сравнивает Z-меры для данных и случайных значений, атрибуты, раз за разом оказывающиеся незначимыми, получают статус Rejected, значимыми — Confirmed. Алгоритм останавливается либо когда все атрибуты будут помечены, либо когда количество итераций достигнет *maxRuns*. Во встором случае останутся признаки, про которые еще ничего не ясно, они будут помечены Tentative. Для того, чтобы разметить эти признаки, помимо перезапуска Boruta с большим *maxRuns*, можно использовать функцию *TentativeRoughFix*.

Возвращается объект класса Boruta из следующих компонент:

- *finalDecision* — вектор разметки атрибутов, элементы принимают одно из 3 значений — Confirmed, Tentative или Rejected.
- *ZScoreHistory* — хранит значения Z-мер для всех атрибутов и вызовов randomForest.
- *timeTaken* — время, которое выполнялась функция.
- *impSource* — описание функции значимости *getImp*.
- *call* — собственно строчка вызова функции.

Пример работы с пакетом

1. Для функционирования Boruta требуется пакет randomForest. Перед началом работы оба пакета требуется загрузить и установить.

2. Для тестирования я использовал данные из задания номер 3. Известно, что 2-й атрибут в нем был фиктивным. Я запустил Boruta на первых 5000 элементах:

```
t3data <- read.table("X1dot.txt", sep = ',') ;
x = t3data[1:1000, 2:10];
y = t3data[1:1000, 1];
Boruta(x, y, doTrace = 2)
```

Получаем:

Initial round 1: .....
Initial round 2: .....
Initial round 3: .....
Final round: .....
5 attributes confirmed after this test: V4 V5 V6 V7 V8

1 attributes rejected after this test: V3
.....
1 attributes rejected after this test: V2
.....
1 attributes rejected after this test: V9
.....
Boruta performed 130 randomForest runs in 3.435997 mins.
5 attributes confirmed important: V4 V5 V6 V7 V8
3 attributes confirmed unimportant: V2 V3 V9
1 tentative attributes left: V10

При этом алгоритм выдавал предупреждения по поводу того, что в этой задаче классификация идет на два класса.

Признак V2 был забракован. Сложно сказать, правильно ли были забракованы признаки V2 и V9, но, во всяком случае, графики, построенные при выполнении 3-го задания, показывают, что распределение этих признаков для класса 0 и класса 1 совпадают в большей степени, чем для других признаков.

3. Затем я в качестве целевого признака взял 2-ой, случайный:

```
t3data <- read.table("X1dot.txt", sep = ',') ;
x = t3data[1:1000, 3:10];
y = t3data[1:1000, 2];
Boruta(x, y, doTrace = 2)
```

Получаем:

Initial round 1: .....
Initial round 2: .....
1 attributes rejected after this test: V7

Initial round 3: .....
Final round: .....
2 attributes rejected after this test: V4 V10
....
1 attributes rejected after this test: V8
....
1 attributes rejected after this test: V5
.....
1 attributes rejected after this test: V9
.....
1 attributes rejected after this test: V6
.....
1 attributes rejected after this test: V3

Boruta performed 113 randomForest runs in 5.040038 mins.
No attributes has been deemed important
8 attributes confirmed unimportant: V3 V4 V5 V6 V7 V8 V9 V10

Как и следовало ожидать, все признаки были забракованы, т.к. не могли помочь предсказать значение случайного признака.

Список литературы

- Учебное пособие А. Г. Дьяконова по системе R.
- R Boruta reference Manual, Miron B. Kursa — http://cran.gis-lab.info/web/packages/Boruta/Boruta.pdf