А. В. Пашук

ИСПОЛЬЗОВАНИЕ ЯЗЫКА R ДЛЯ АНАЛИЗА И ВИЗУАЛИЗАЦИИ ДАННЫХ

Рассматривается пример использования языка программирования R для анализа и визуализации больших объемов данных на примере распознавания спам-писем с использованием модели, обученной методом Random forest.

Введение

R - это бесплатный язык программирование и программная среда для статистических вычислений и визуализации. Язык широко используется статистиками и учеными, работающих с большими объемами данных для разработки приложений, сбора, преобразования и анализа данных. Главным преимуществом языка является тот факт, что он разработан и оптимизирован для обработки больших объемов данных.

I. Обучение модели методом Random forest

В рамках примера рассмотрим модель, учитывающую всего один фактор - количество заглавных букв в тексте письма. Очевидно, что в большинстве случаев спам-письма буду иметь значительно большее количество таких символов (см.рис.1.).

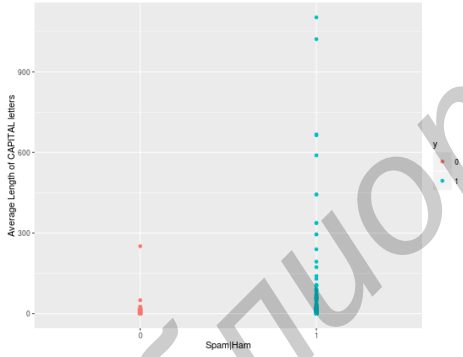

Рис. 1 – Количество заглавных букв в письмах

Для получения модели машинного обучения будем использовать алгоритм Random Forest. Также перед построением модели разобьем наш датасет на тестовый (30% от всех писем) и обучающий (70%). Для обучения модели используется функция train из пакета caret. Также следует отметить, что функция train может принимать множество различных параметров, например, параметры кросс-валидации.

```
inTrain <- createDataPartition(y=dataset$y,
                                    p=0.7,
                                    list=FALSE)
training <- dataset2[inTrain,]
testing <- dataset2[-inTrain,]
// train the model
modFit<-train(y~.,data=training,method="rf",
        trControl=trainControl(method="cv",
                                    number=5))

// predict
pred <- predict(modFit, testing)
```

Результат работы модели представлен на рисунке 2.

```
> testing$predRight <- pred == testing$y
> table(pred, testing$y)

pred   0   1
   0 757  21
   1  79 522
```

Рис. 2 – Результат работы модели

Точность работы алгоритма составляет 0.93. Стоит отметить, что при правильной настройке модели можно получить значительно лучшие результаты.

II. Выводы

Язык R является мощным универсальным средством, подходящим как для разведочного анализа данных, так и для программирования сложных систем анализа больших объемов данных. Позволяет в полуавтоматическом режиме генерировать отчеты с добавлением результатов анализа, графиков и диаграмм.

1. The caret Package [Electronic resource]. – Mode of access: http://topepo.github.io/caret/index.html. – Date of access: 13.03.2016.
2. Package doParallel [Electronic resource]. – Mode of access: https://cran.r-project.org/web/packages/doParallel/doParallel.pdf. – Date of access: 13.03.2016.

Пашук Александр Владимирович, магистр технических наук, ассистент кафедры информатики, aliaksandr.pashuk@gmail.com.

Научный руководитель: Гуринович Алевтина Борисовна, кандидат физико-математических наук, доцент, gurinovich@bsuir.by.