Metody Odkrywania Wiedzy (MOW)
Projekt analityczny klasyfikacja
Założenia wstępne

Andrzej Smyk
Błażej Szum

13 kwietnia 2015

1 Cel projektu

Celem projektu jest wnikliwa i szeroko zakrojona analiza danych z wykorzystaniem dostępnych pakietów i algorytmów zaimplementowanych w środowisku R. Za pomocą dostępnych w R narzędzi stworzymy odpowiednie modele pozwalające na przydzielenie obserwacji do jednej z kategorii na podstawie wartości jej atrybutów. Działania jakie zostaną podjęte przy realizacji zadania projektowego to:

- wstępne przetworzenie/transformacja danych,
- statystyczny opis danych,
- wybór parametrów algorytmów klasyfikacji wymagających strojenia,
- ocena jakości modeli,
- opis wyników i przedstawienie wniosków.

Celem analizy jest zbudowanie modeli określających typ lasu na postawie pozycji na mapie, nachylenia terenu, typu gleby, odległości od zbiorników wodnych czy dróg.

2 Opis danych

Dane do projektu pochodzą ze strony UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/machine-learning-databases/covtype`. Mają one charakter surowych (nieskalowanych) danych geograficznych i geologicznych opisujących działki o powierzchni $30m \times 30m$ znajdujących się w Roosevelt National Forest na Północy stanu Kolorado, USA. Każdy z przykładów opisany jest za pomocą 12 atrybutów - z których 10 to atrybuty numeryczne, a 2 binarne - zawartych w sumie w 54 kolumnach. 4 z kolumn opisują obszar, na którym znajduje się dany parcel, a 40 rodzaj występującej tam gleby. Pozostałe atrybuty numeryczne charakteryzują takie dane geograficzne jak wysokość nad poziomem morza, nachylenie terenu, odległości od dróg i zbiorników wodnych oraz inne.

Atrybut, pełniący rolę pojęcia domyślnego, tj. klasy do której przykłady będą przypisywane, to rodzaj lasu, składający się z kategorii przedstawionych w tablicy poniżej.

Klasa	Nazwa
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz

Pełen zbiór zawiera 581,012 przykładów.

3 Wstępne przygotowanie danych

Wszystkie dane, kategoryczne (binarne) oraz nominalne, przedstawione są w postaci liczbowej. W zbiorze nie występują niepełne obserwacje.

Dodane zostaną dwa nowe atrybuty, powstałe poprzez scalenie podobnych istniejących atrybutów. Rodzaje gleby, które są reprezentowane przez wiele atrybutów zostaną zastąpione pojedynczym atrybutem opisującym kategorię gleby. Również atrybuty Wilderness_Area, zostaną zamienione na jeden atrybut kategoryczny.

4 Wybór i strojenie algorytmów

Do rozwiązania zadania użyte będą algorytmy (z pakietów):

1. Support Vector Machines (*e1071* oraz *kernlab*): strojenie poprzez dobranie odpowiedniego parametru sigma podczas tworzenia radial basis function (RBF);

2. random forest (*randomForest*): wymagana parametryzacja drzewa użytego w algorytmie.

Nie wykluczamy również wykorzystania algorytmów z innych pakietów: *gbm* lub/oraz *extraTree*. Istotnym elementem będzie strojenie wykorzystywanych algorytmów do analizowanego zbioru danych. Strojone będą tylko te parametry które w znaczący sposób wpłyną na wyniki predykcji algorytmu. W celu określania efektywności strojonego algorytmu ze zbioru danych, wykorzystamy prostą lub k–krotną krzyżową walidację. W pierwszej metodzie wybrany zostanie niewielki podzbiór, około 10% rekordów który będzie służył do testowania wytrenowanych modeli. Modele będą trenowane na pozostałym zbiorze 90% rekordów. W drugiej opcji zbiór podzielimy na 10 lub 5 zestawów, z których każdy w kolejnych iteracjach będzie służył jako zbiór testowy, a pozostałe jako zbiór trenujący. Jeden ze zbiorów nie będzie wykorzystywany w walidacji, lecz zostanie użyty do sprawdzenia dokładności modelu.

Decyzję o wyborze algorytmów do analizy oraz metody strojenia, podejmiemy na etapie realizacji analizy, uzależniając ją od stopnia skomplikowania zadania projektowego.

5 Ocena jakości modeli

Do oceny jakości oraz dokładności modeli modele zostaną przetestowane na zbiorze testowym. Otrzymane wyniki porównane z rzeczywistą kategorią pozwolą na określenie dokładności modelu. Na ich podstawie będziemy również w stanie stworzyć macierz pomyłek oraz wyznaczyć powiązane z nią współczynniki. Na ich podstawie będziemy w stanie wykreślić krzywą ROC oraz wyznaczyć AUC.