02450 Project 1

Report

by Karol Dzitkowski
Marco Becattini

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Introduction to Machine Learning and Data Mining
Tue Herlau
24th September 2014

Abstract

The objective of this report is to apply the methods we learned in the first section of the course on "Data: Feature extraction, and visualization" on letter recognition data set to get a basic understanding of the data prior to the further analysis. This report includes a description of the dataset, detailed explanation of the attributes of the data, data visualizations and a conclusion what we found out about the data set.

Description

Selected data set

Title Letter Image Recognition Data

Characteristics Multivariate

Number of Instances 20000

Attribute Characteristics Integer

Number of Attributes 16

Missing Values? No

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

Data was obtained from a website:
`http://archive.ics.uci.edu/ml/datasets/Letter+Recognition`
and was a part of an article "Letter Recognition Using Holland-style Adaptive Classifier" available on website:
`http://link.springer.com/article/10.1007/BF00114162`

The research for this article investigated the ability of several variations of Holland-style adaptive classifier systems to learn to correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters. The best accuracy obtained was a little over 80%. We typically train on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000. See the article cited above for more details.

Data analysis

- The aim of our future machine learning algorithm working on data set is classification, that is to assign a set of values (our data record) to a capital letter in the English alphabet. It will be probably implemented as some kind of neural network. Classification is a hard problem for a letter recognition, because there is no direct algorithm of doing that, what is the reason for implementing artificial intelligence based on neural network.

- Clustering groups similar characters together, what can be implemented on our data using some sort of N dimensional metric, for example Euclidean or another appropriate similarity measure for records describing letters.

- Association rule discovery will be considered here as a way of predicting a probable value of other attribute values basing on some which are known. We will try to find a set of rules which can tell us a range the other values will be in or the most likely value. For example:

$$\{x = 5, y = 6\} \Rightarrow z \in [-1, 1]$$

or

$$\{x = 1, y = 4, z = 5\} \Rightarrow letter = \text{"}a\text{"}or\text{"}e\text{"}$$

- Anomaly detection is used to identify records corresponding to a particular letter with a significant deviation in some or all attributes. It will be helpful to identify unusual patterns of some letters.

Algorytm

Na wejściu algorytmu zostanie podany graf w formacie przedstawionym w pkt. 4.
Następujący pseudokod prezentuje przebieg algorytmu:

Pseudokod

1. Wczytaj graf G
2. Utwórz algorytmem Kruskala minimalne drzewo rozpinające:
   A. Utwórz las L z wierzchołków grafu G – każdy wierzchołek jest na początku osobnym drzewem.
   B. Utwórz zbiór S zawierający wszystkie krawędzie grafu G.
   C. Uporządkuj zbiór S w kolejności rosnącej.
   D. Dopóki S nie jest pusty:
      a. Pobierz krawędź o minimalnej wadze z S i przypisz do e.
      b. Jeśli e łączy dwa różne drzewa:
         i. dodaj e do lasu L, tak aby połączyła dwa odpowiadające drzewa w jedno.
         ii. Jeśli L jest drzewem rozpinającym idź do kroku 3.
3. przejdź drzewo L i utwórz z niego cykl Hamiltona
   A. $root$:= wybierz korzeń drzewa L.
   B. H = MetodaA(L, root).
   C. dodaj krawędź od ostatniego wierzchołka do korzenia grafu H.
4. zwróć rozwiązanie H

Opis funkcji pomocniczych

Rozwiązaniem będziemy nazywać listę wierzchołków generowaną przez metody A i B, która wskazuje kolejność przechodzenia wierzchołków w drzewie.

Metoda A

Funkcja przechodzi przez podrzewo zaczynając w korzeniu w i kończąc na jego dziecku.
Rozwiązanie MetodaA(Wierzchołek w):

1. Jeśli drzewo o wierzchołku w ma <= 3 wierzchołki:
   A. zwróć przejście metodą A podstawowego grafu i zakończ.
2. Utwórz puste rozwiązanie r.
3. Dla każdego dziecka d wierzchołka w:
   A. Dodaj do r rozwiązanie znalezione przez $MetodaB(d)$.
   B. Wierzchołek n = pobierz następne dziecko wierzchołka w.

C. Jeśli n nie jest puste:
   a. Do r dodaj pierwsze dziecko wierzchołka n jeśli istnieje lub wierzchołek n.

4. Zwróć r.

Metoda B

Funkcja przechodzi przez podrzewo o korzeniu w zaczynając na jego dziecka i kończąc na nim.
Rozwiązanie MetodaB(Wierzchołek w):

1. Jeśli drzewo o wierzchołku w ma <= 3 wierzchołki:
   A. zwróć przejście metodą B podstawowego grafu i zakończ.
2. Utwórz puste rozwiązanie r.
3. Dla każdego dziecka d wierzchołka w:
   a. Dodaj do r rozwiązanie znalezione przez $MetodaA(d)$.
   b. Wierzchołek n = pobierz następne dziecko wierzchołka w
   c. Jeśli n nie jest puste:
      i. Do r dodaj wierzchołek n jeśli istnieje i idź do pkt. 4.
   d. Do r dodaj wierzchołek w.
4. Zwróć r.

Dowód poprawności

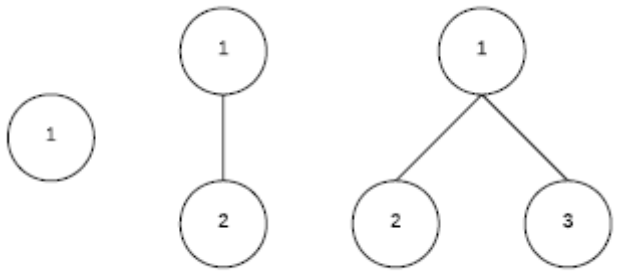
Twierdzenie 1.
Każde drzewo można przejść wracajac do korzenia przeskakując maksymalnie dwa wierzchołki, tak aby każdy wierzchołek oprócz korzenia odwiedzić dokładnie raz.

Dowód:
Dowód będzie polegał na zasadzie indukcji matematycznej.
Najpierw udowodnimy że dla podstawowych drzew 1,2,3 wierzchołkowych twierdzenie zachodzi.
Następnie udowodnimy, że można takie grafy przejść na dwa sposoby:

a. Zaczynając od korzenia, można przejść przez wszystkie wierzchołki odwiedzając je dokładnie raz i kończąc na dziecku korzenia (oczywiście z dziecka korzenia zawsze można przejść do korzenia jako ostatni ruch zamykając cykl)

b. Zaczynając od dziecka korzenia, można przejść przez wszystkie wierzchołki odwiedzając je dokładnie raz i kończąc na korzeniu

Z tym, że jeśli graf jest jednowierzchołkowy, nie trzeba go oczywiście dalej przechodzić, jeśli już w niego weszliśmy. Natomiast grafu 0 wierzchołkowego nie trzeba przechodzić wogóle, więc napewno można go przejść na te dwa sposoby.



Rys. 1 Podstawowe konstrukcje

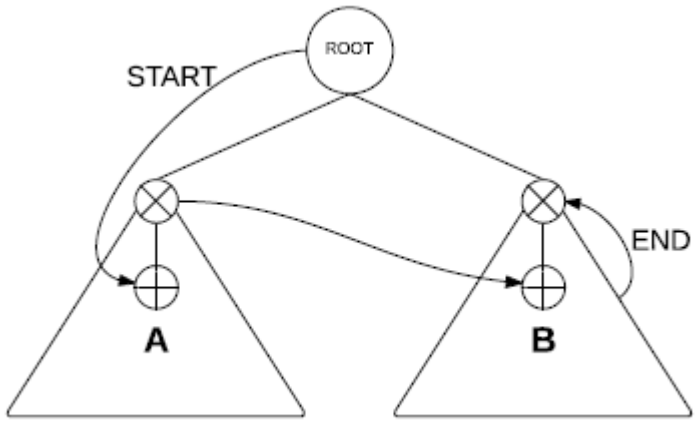
Łatwo zauważyć, że dla tych podstawowych konstrukcji udowodnienie założeń jest banalne. Np. dla grafu z 2 wierzchołkami można przejść z 1 do 2 i z 2 do 1 albo odwrotnie.

Następnie zakładamy, że umiemy przejść na te dwa sposoby drzewa A i B. Udowodnimy, że można przejść na te dwa sposoby większe drzewo C powstałe

poprzez połączenie drzew A i B z nowym wierzchołkiem (jako korzeń). Ten sposób konstrukcji pozwala stworzyć dowolne drzewo. Jeśli udowodnimy, że z możliwości przejścia w te sposoby drzew A i B wynika, że mozna przejść drzewo C, co znaczy, że można przejść na te sposoby każde drzewo.

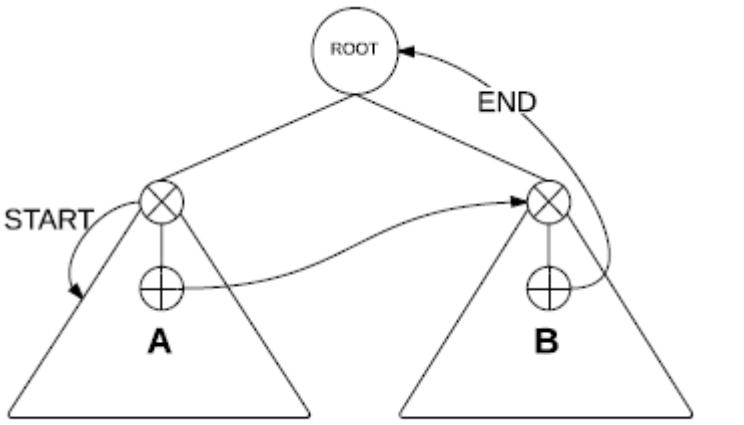
Mamy dwa przypadki:

1 Z korzenia przechodzimy do dziecka korzenia grafu A i przechodzimy go sposobem a), kończąc na korzeniu drzewa A. Następnie przechodzimy do dziecka korzenia drzewa B i przechodzimy go sposobem a), kończąc w korzeniu grafu B. W ten sposób skończyliśmy w dziecku powstałego drzewa C. W ten sposób udało się przejść drzewo C w sposób b).



Rys. 2 Sytuacja w której zaczynamy od korzenia i kończymy na jego dziecku

2 Zaczynamy z dziecka grafu C, zatem z korzenia grafu np. A. Następnie przechodzimy graf A sposobem b) kończąc w dziecku korzenia drzewa A. Następnie przechodzimy przeskakując 2 wierzchołki (korzeń drzewa A i C) do korzenia drzewa B i znowu przechodzimy go sposobem b). Na koniec skaczemy przez korzeń drzewa B i kończymy w korzeniu drzewa C. W ten sposób przeszliśmy drzewo C na sposób a).



Rys. 3 Sytuacja w której zaczynamy od dziecka korzenia i kończymy na korzeniu

Z indukcji matematycznej wynika, że każde drzewo da się przejść na sposób a) i b). Natomiast z możliwości przejścia na oba sposoby każdego drzewa wynika, że twierdzenie jest prawdziwe.