Case study 1.9 (ZED – TPD 2009)

Analiza danych nt. rozpoznawania pisma

Ogólnym celem jest jak najlepsze rozpoznawanie cyfrowych obrazów liter w języku angielskim. Oprócz jak najwyższej trafności klasyfikowania, należy zwrócić uwagę na rozpoznawanie poszczególnych liter, w szczególności litery „H”.

Opis danych:

Dane zebrano w ramach projektu naukowego, gdzie pobrano cyfrowe wersje bardzo dużej liczby biało-czarnych obrazów każdej z 26 dużych liter (ang. capital letters) alfabetu w języku angielskim. Obraz każdego znaku zapisywany był w 20 różnych krojach czcionki (tzw. font style). Dla każdej litery oraz kroju czcionki oprócz wyglądu wzorcowego dokonano zaburzenia wyglądu i w ten sposób utworzono 20 tys. różnych zapisów powyższych liter.

Wersja cyfrowa litery (na podstawie matrycy pikseli – w poniższym opisie nazywana „box”) była podstawą do zdefiniowania 16 podstawowych atrybutów, charakteryzujących każdy obraz, które są scharakteryzowane poniżej. Należy dodać, że każdy z atrybutów jest liczbowy i został już przeskalowany do zakresu [0, 15].

Zestawienie informacji o atrybutach

1.	x-box	horizontal position of box	(integer)
2.	y-box	vertical position of box	(integer)
3.	width	width of box	(integer)
4.	high	height of box	(integer)
5.	onpix	total # on pixels	(integer)
6.	x-bar	mean x of on pixels in box	(integer)
7.	y-bar	mean y of on pixels in box	(integer)
8.	x2bar	mean x variance	(integer)
9.	y2bar	mean y variance	(integer)
10.	xybar	mean x y correlation	(integer)
11.	x2ybr	mean of x * x * y	(integer)
12.	xy2br	mean of x * y * y	(integer)
13.	x-ege	mean edge count left to right	(integer)
14.	xegvy	correlation of x-ege with y	(integer)
15.	y-ege	mean edge count bottom to top	(integer)
16.	yegvx	correlation of y-ege with x	(integer)

Zestawienie informacji o klasach : 26 capital letter (26 values from A to Z)

Rozkład liczności przykładów w klasach:

789 A	766 B	736 C	805 D	768 E	775 F	773 G
734 H	755 I	747 J	739 K	761 L	792 M	783 N
753 O	803 P	783 Q	758 R	748 S	796 T	813 U
764 V	752 W	787 X	786 Y	734 Z		

Uwagi do analizy:

Ogólnym celem jest skonstruowanie jak najskuteczniejszego systemu uczącego się rozpoznawania liter.
Przy czym oprócz globalnej jak najwyższej zdolności predykcyjnej, istotna jest analiza skuteczności rozpoznawania poszczególnych liter. Nie jest akceptowalna sytuacja, gdzie pomimo wysokiej globalnej trafności, któraś z liter jest rozpoznawana poniżej 85%. Równocześnie, z uwagi na potencjalne zakłócenie w zapisie zaburzonych czcionek nie oczekuje się w pełni 100% globalnej trafności, lecz sugeruje się osiągnięcie jak najbliższej (ale nie niższej niż około 92%).

Dodatkowo należy uwzględniać rozmiar danych i czas związany z jego przetwarzaniem. Poszukiwanie rozwiązania w czasie większym niż kilkanaście minut jest niekorzystne.

Należy zastanowić się nad metodyką oceny zdolności predykcyjnej.

Warto zwrócić uwagę na

(a) ocena jakości danych i ew. oczyszczanie pliku z niepoprawnych elementów,

(b) ocenę ważności poszczególnych atrybutów dla zdolności rozpoznawania liter,

(c) należy zachować dużą ostrożność w zakresie ew. redukcji liczby atrybutów, gdyż zostały one już poprzednio wybrane przez ekspertów,

(d) konstrukcja tzw. klasyfikatora, czyli wskazań dla podejmowania końcowej decyzji – wymaga porównania wielu sposobów konstrukcji takich klasyfikatorów opartych na zróżnicowanych algorytmach uczących.

Ponadto trzeba być świadomym ograniczeń wydajnościowo-pamięciowych związanych z wczytaniem i przetwarzaniem tak dużych zbiorów danych.
Na przykład w oprogramowaniu WEKA zamiast interfejsu graficznego można wykorzystać tryb uruchomienia poszczególnych składników z linii komend.

Inne zagadnienia przekaże prowadzący – konsultant.