## Case study 1.12 (ZED – TPD 2009)

## Kategoryzacja emajli do folderów w skrzynce użytkownika

Ogólnym celem jest stworzenie systemu automatycznie klasyfikującego wiadomości tekstowe (typu emajle) do właściwego folderu w kliencie pocztowym danego użytkownika.

**Opis danych:**

Wiadomości tekstowe pochodzą z archiwum anglojęzycznego z pewnego projektu badania naukowego przeprowadzonego na początku tego wieku.

Początkowo zawartość każdej wiadomości zapisana jest w osobnym pliku tekstowym. Treść wszystkich wiadomości napisana jest w języku angielskim i zapisana w formacie tekstowym, bez znaczników *HTML*. Nagłówek wiadomości zawiera tylko podstawowe pola (nadawca, odbiorca, temat, data i pewne informacje kontrolne, np. sposób ew. kodowania).

Trudnością jest duża liczba folderów w skrzynce użytkownika. Nazwy tych grup tworzą etykiety klas decyzyjnych. Ponadto rozkład wiadomości w klasach jest niezrównoważony. Dodatkowo, jeśli rozpatrywać etykietę czasową i uporządkować emejle wg. tego identyfikatora można zauważyć „dryft" zmiany liczności wiadomości w poszczególnych folderach.

Dostępne są dwa zbiory danych:
Pierwszy zawiera 1736 wiadomości podzielonych na 58 folderów (najmniej liczny 10 a najbardziej 166 wiadomości).
Drugi zawiera 3651 wiadomości podzielonych na 22 foldery (najmniej liczny 10 a najbardziej 1192 wiadomości).

Oryginalne postacie plików emejli zostały poddane przetwarzaniu wstępnemu tak, aby wyłuskać informacje dotyczące nagłówka wiadomości oraz samej treści. Następnie wobec tekstu zastosowano metody przetwarzania języka naturalnego, gdzie trzeba było podjąć decyzję między innymi co do: konfiguracji parametrów takich jak analiza języka (lematyzacja,..), usunięcie stop słów.

W dalszej kolejności zbudowano zbiór atrybutów – większość z nich to częstości wystąpienia termów – na ogół pojedynczych słów. Inne to atrybuty liczbowe dotyczące występowanie określonych wartości parametrów w nagłówkach emejli – obecnie udostępniamy już przetworzony zbiór przygotowany do eksploracji (w stylu arff).

Obecnie liczba atrybutów jest dość dużą – kilka tysięcy.

Sugerowana jest selekcja atrybutów, jeśli doprowadzi to do skuteczniejszego klasyfikatora oraz ułatwi aspekty wydajności obliczeniowej.

**Uwagi do analizy:**

Ogólnym celem jest skonstruowanie jak najskuteczniejszego systemu uczącego się przydzielania wiadomości do poszczególnych folderów – klas.

W ogólności oczekujemy skuteczności na poziomie minimum 51% (pierwszy zbiór) i 65% drugi zbiór.

Przy czym oprócz globalnej jak najwyższej zdolności predykcyjnej, istotna jest analiza skuteczności rozpoznawania poszczególnych kategorii. Nie jest akceptowalna sytuacja, gdzie pomimo wysokiej globalnej trafności, któraś z klas jest prawie w ogóle rozpoznawana.

Uwaga: warto podczas oceny zdolności predykcyjnej uwzględnić porządek chronologiczny wiadomości – pierwszy atrybut pełni rolę znacznika czasowego; wtedy metodycznie uczymy się z przeszłych historycznie danych a klasyfikujemy pewną ilość następnych wiadomości.

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

Warto zwrócić uwagę na

(a) Redukcje rozmiarów – selekcję atrybutów

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

(c) ze względu na zadanie automatycznej klasyfikacji, forma symbolicznej prezentacji wyników nie jest konieczna.

(d) należy się zastanowić nad wyborem właściwych miar oceny zdolności predykcyjnej.

(e) przemyślenia sposobu przetwarzania tak dużych wolumenów danych oraz uwzględnianiu chronologii wiadomości.

(f) w zestawie wiadomości pierwszy atrybut „@attribute timestamp numeric” jest liczbą oznaczającą znacznik czasowy.

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.