Przygotowanie danych do klasyfikacji — analiza danych z logów

dr inż. Joanna Kołodziejczyk

14.11.2016

1 Cel laboratoriów

Poznanie procesu przygotowania danych do wykorzystania w tworzeniu modelu z wykorzystaniem metod maszynowego uczenia się.

2 Na czym polega wstępne przygotowanie danych (Data preprocessing and transformation)

2.1 Przygotowanie danych

Na tym etapie głównie skupiamy się na rozwiązaniu dwóch problemów:

1. poradzić sobie z danymi zaszumionymi (błędnymi) w tym:
    - odnalezieniu rekordów powtarzających się
    - odnalezieniu niewłaściwych wartości atrybutów
    - smoothing data (wygładzenie danych)
2. poradzić sobie z danymi brakującymi, wykorzystując jedną z metod:
    - usunąć rekordy z tymi danymi
    - uzupełnić średnią z wartości atrybutu rekordów w tej samej klasie
    - uzupełnić średnią z wartości atrybutu rekordów najbardziej podobnych

2.2 Transformacja danych

Na tym etapie dokonuje się przekształcenia danych polegających na:

- normalizacji
- konwersji typów
- wyborze atrybutów i rekordów

3 Dane do wykorzystania

W ramach zadania należy wykorzystać dwa zbiory danych:

1. KDDCup99.arff - plik z konkurstu KDD Cup 99
2. NSL-KDDTrain.arff poprawione dane KDD (usunięcie redundancji).

4 Zadania do wykonania

1. Opisać jak powstał zbiór NSL-KDDTrain.arff z pliku KDDCup99.arff. Pomocny będzie artykuł `http://www.ee.ryerson.ca/~bagheri/papers/cisda.pdf`
2. **Opisać atrybuty** z obu plików wg wzoru (może być tabla):
   (a) Nazwa atrybutu,
   (b) Opis (rozpoznanie poprawnych wartości),
   (c) Typ danej
      i. jeżeli dana jest numeryczna: podać minimum, maksimum i średnią
      ii. jeżeli dana jest wyliczeniowa, to podać ile jest rekordów w każdej wartości,
   (d) Ile jest brakujących danych,
   (e) Liczba różnych wartości,
   (f) Ile jest wartości unikalnych,

   Uwaga: wszystkie dane podane są w programie WEKA.
3. Napisać w sprawozdaniu, czy zauważyć można coś ciekawego (np. nietypowe wartości, dane unikalne, anomalia) w danych. Czy da się wskazać, atrybuty niepotrzebne, np. bo mają taką samą wartość w takich rekordach, albo nie mogą wpływać na modelowanie?
4. Przejrzeć dane pod kątem **brakujących wartości**. Poniższe zadania wykonać tylko, jeżeli w zbiorze są brakujące wartości:
   (a) Usunąć rekordy z brakującymi wartościami.
   (b) Zastosować filtry WEKA:
      i. *ReplaceMissingValues*
      ii. *EMImputation*

   Dla każdego podejścia:
   (a) Zapisać plik po transformacji.
   (b) Odpowiedzieć na pytania: Jaka została liczba rekordów? O ile procent zmniejszyła się liczba? Czy sądzisz, że będzie to miało wpływ na tworzenie modelu?
   (c) Porównać podejścia i skonstruować wnioski.

5. Przejrzeć dane pod kątem powtarzających się rekordów i błędnych wartości atrybutów. Wykonać na każdym zbiorze zadania:

   (a) Usunąć powtarzające się rekordy stosując filtr *RemoveDuplicates*.

   (b) *RemoveFrequentValues* - Dla podanego atrybutu oznacza rekordy częste i nie częste atrybutu wyliczeniowego i pozostawienie tylko rekordy z taką liczbą najczęstszych wartości, która zostanie wskazana w parametrze *-N*.

   (c) *RemoveUseless* - filtr usuwa atrybuty, które nie różnią się w ogóle lub które różnią zbyt wiele.

   Dla każdego podejścia:

   (a) Zapisać plik po transformacji.

   (b) Odpowiedzieć na pytania: Jaka została liczba rekordów/atrybutów? O ile procent zmniejszyła się liczba? Czy sądzisz, że będzie to miało wpływ na tworzenie modelu?

   (c) Porównać podejścia i skonstruować wnioski.

6. Dokonać różnego rodzaju transformacji. W sprawozdaniu krótko opisać na czym transformacja polega i co, jeżeli w ogóle, się po jej zastosowaniu zmieniło w danych.

   (a) *Normalize* ujednolicanie.

   (b) *Standarize* ujednolicanie.

   (c) *NominalToBinary* Zamienia wskazane w parametrze *-R, attribute Indices* atrybut wyliczeniowy o k wartościach na k atrybutów binarnych (przydatne dla atrybutów o bardzo licznych dziedzinach). Działa tylko dla atrybutów wyliczeniowych!

   (d) *Discretize (supervised)* Zamienia wskazane (w polu *-R, attribute Indices*) atrybuty rzeczywistoliczbowe na dyskretne dzieląc na przedziały.

   (e) *PKIDiscretize* Zamienia wskazane (w polu *-R, attribute Indices*) na dyskretne na podstawie częstotliwości atrybutów.

   (f) *Attribute Selection* Duża grupa metod do wyboru podzbioru atrybutów możliwie silnie skorelowanych ze zmienną wyjściową i możliwie słabo skorelowanych między sobą. Na selekcję składają się zawsze dwa elementy: sposób oceniania i sposób przeszukiwania (sposób przebiegania podzbiorów zbioru atrybutów, ekonomiczniejszy niż zachłanny).

      Dla każdego filtru/transformacji:

      i. Zapisać plik po transformacji.

      ii. Odpowiedzieć na pytania: Jaka została liczba rekordów/atrybutów? O ile procent zmniejszyła się liczba? Czy sądzisz, że będzie to miało wpływ na tworzenie modelu?

      iii. Porównać podejścia i skonstruować wnioski.

7. W KDDCup99.arff korzystając z tabeli poniżej połącz atrybuty wskazujące na ten sam typ ataku:

(a) DOS: denial-of-service, e.g. syn flood;

(b) R2L: unauthorized access from a remote machine, e.g. guessing password;

(c) U2R: unauthorized access to local superuser (root) privileges, e.g., various "buffer overflow" attacks;

(d) probing: surveillance and other probing, e.g., port scanning.

atrybut	rodzaj ataku
back	dos
buffer_overflow	u2r
ftp_write	r2l
guess_passwd	r2l
imap	r2l
ipsweep	probe
land	dos
loadmodule	u2r
multihop	r2l
neptune	dos
nmap	probe
perl	u2r
phf	r2l
pod	dos
portsweep	probe
rootkit	u2r
satan	probe
smurf	dos
spy	r2l
teardrop	dos
warezclient	r2l
warezmaster	r2l

Innymi słowy należy podmienić we wszystkich rekordach wartość ostatniego atrybutu, która jest w tabeli powyżej w lewej kolumnie na wartość w prawej kolumnie. Można wykorzystać metodę w javie opisaną w `http://weka.wikispaces.com/Rename+Attribute+Values`.

4.1 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_nazwisko.pdf* należy przesłać na adres jkolodziejczyk[at]ajp.edu.pl. Tytuł maila: Sprawozdanie 1 z ISPAS.