Dyskretyzacja atrybutów

Aleksy Barcz, Wojciech Koszołko

22 stycznia 2013

Spis treści

Rozdział 1

Cel projektu

1.1 Sformułowanie zadania.

Wstępująca i zstępująca dyskretyzacja atrybutów ciągłych z uwzględnieniem rozkładu kategorii. Badanie wpływu dyskretyzacji na jakość modeli klasyfikacji tworzonych za pomocą algorytmów dostępnych w R.

1.2 Rozwinięcie tematu.

Celem projektu była implementacj dyskretyzacji z nadzorem, globalnej. Każdy atrybut był dyskretyzowany niezależnie od pozostałych. Zostały zaimplementowane dwie metody: zstępująca i wstępująca, z których każda jest parametryzowalna poprzez podawanie odpowiednich kryteriów stopu (bądź ich kombinacji), wraz z ewentualnymi parametrami. Na potrzeby porównania jakości dyskretyzacji uzyskanej metodą *BottomUp* i *TopDown* konieczne było uzyskanie tej samej liczby przedziałów dla obu metod. W tym celu zaimplementowano dodatkowe kryterium stopu: żądaną ilość przedziałów. W związku z tym dla obu metod zaimplementowano globalne kryteria wyboru przedziałów do łączenia/podziału (ważenie wskaźników jakości przedziału licznością przedziału) - tak aby w danym momencie działania algorytmu, algorytm wybierał globalnie najlepsze przedziały do łączenia/podziału.

Rozdział 2

Dyskretyzacja zstępująca

Jako kryterium podziału dla dyskretyzacji zstępującej (*TopDown*) zaimplementowano wybór progu maksymalizującego spadek entropii dla wybranego przedziału w przypadku podzielenia przedziału na dwa nowe przedziały względem tego progu. Schemat działania kryterium podziału:

1. Dla każdego przedziału:
   (a) obliczenie progu θ maksymalizującego spadek entropii.
   (b) obliczenie ważonego (ilością próbek) spadku entropii dla wybranego progu.
2. Wybranie przedziału o maksymalnym potencjalnym ważonym spadku entropii.
3. Podział wybranego przedziału na dwa nowe, względem wyznaczonego progu.

Zaimplementowane kryteria stopu:

- żądana ilość przedziałów – algorytm stara się osiągnąć zadaną ilość przedziałów, nawet gdy kolejne podziały nie dają spadku entropii; uzyskana ilość przedziałów może być mniejsza od zadanej, gdy żadnego przedziału nie da się już podzielić (takie same wartości atrybutu a w ramach każdego przedziału lub wszystkie przedziały jednoelementowe)
- minimalny potencjalny spadek entropii (warunek globalny) – zatrzymanie algorytmu następuje gdy dla żadnego z przedziałów nie da się osiągnąć zadanego jako parametr spadku entropii (nieważonego)
- kryterium *delta* (2.1), łączące jakość kodowania z entropią [2]

$$g_{a,\theta}(P) < \frac{\log(|P|-1)}{|P|} + \frac{\Delta_{a,\theta}(P)}{|P|}, \tag{2.1}$$

$$\Delta_{a,\theta}(P) = \log(3^{|C_P|} - 2) - (|C_P|I(P) - |C_{P_{a\leqslant\theta}}|E_{a\leqslant\theta}(P) - |C_{P_{a>\theta}}|E_{a>\theta}(P)), \tag{2.2}$$

$$C_P = \{d \in C \mid (\exists x \in P)\ c(x) = d\}. \tag{2.3}$$

Rozdział 3

Dyskretyzacja wstępująca

Jako kryterium złączenia dwóch przyległych przedziałów dla dyskretyzacji wstępującej BottomUp zaimplementowano analizę statystki χ^2. Na podstawie wartości aktualnie rozpatrywanego atrybutu, tworzone są przedziały (każda unikalna wartość w oddzielnym przedziale). Następnie przedziały są łączone aż osiągną kryterium stopu. Schemat działania kryterium połączenia dwóch przyległy przedziałów:

1. Dla każdego przyległego przedziału:
   (a) Oblicz wartość χ^2
2. Połącz dwa przedziały o minimalnej wartości χ^2

Wartość statyski χ^2 pomiędzy dwoma przedziałami jest liczona na podstawie wzoru:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j}^{c} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

Gdzie:
c = liczba k
A_{ij} = liczba wartości w i-tym przedziale, j-tej klasy
R_i = liczba wartości w i-tym przedziale
C_j = liczba obiektów j-tej klasy w obu przedziałach
N = liczba wartości w obu przedziałach
$E_{ij} = (R + i * C_j)/N$
Zaimplementowane kryteria stopu:

- żądana ilość przedziałów – algorytm stara się osiągnąć zadaną ilość przedziałów; uzyskana ilość przedziałów może być mniejsza od zadanej, gdy zostanie dodane kryterium na maksymalną wartość χ^2, dla której następuje złączenie przyległych przedziałów.
- maksymalna wartość χ^2 – zatrzymanie algorytmu następuje gdy dla żadnej z pary przedziałów nie dało się osiągnąć wartości χ^2 mniejsze od zadanej w kryterium

Rozdział 4

Testy

4.1 Specyfikacja ogólna testów.

Testy zostały wykonane przy użyciu dwóch zbiorów danych - uczącego i testowego, zgodnie z następującym schematem:

1. Zbudowanie modelu dyskretyzacji na zbiorze uczącym
2. Dyskretyzacja zbioru uczącego przy użyciu modelu dyskretyzacji
3. Nauczenie klasyfikatora na zdyskretyzowanym zbiorze uczącym
4. Dyskretyzacja zbioru testowego przy użyciu modelu dyskretyzacji
5. Sprawdzenie jakości klasyfikacji na zdyskretyzowanym zbiorze uczącym

Dla uzyskania wiarygodnych wyników zastosowano we wszystkich eksperymentach walidację krzyżową. Do testów użyto zbiór danych pomiarowych opisujących procesy zachodzące na czujnikach chemicznych [4]. Zbiór danych składa się z próbek o 129 atrybutach o wartościach ciągłych, na potrzeby eksperymentów dyskretyzowano wszystkie bądź wybrane atrybuty. Ze względu na długi czas dyskretyzacji do eksperymentów wykorzystano tylko zbiór *batch1.dat*, składający się ze 445 próbek o dość równomiernym rozkładzie 6ciu klas. W celu zmniejszenia czasu eksperymentów, próbowano przeprowadzać część eksperymentów dyskretyzując tylko atrybuty najistotniejsze z punktu widzenia klasy próbek, jednakże wyniki tych eksperymentów były mało interesujące, w związku z czym skupiono się na dyskretyzowaniu całego zbioru atrybutów. Zbiór wybranych atrybutów to: $V_{imp} = \{V2, V7, V11, V18, V20, V26, V51, V67\}$. Atrybuty te zostały wybrane przy użyciu drzewa decyzyjnego z pakietu *rpart*.

```
rpart(V1 ~ ., gas1)
n= 445

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 445 347 2 (0.2 0.22 0.19 0.067 0.16 0.17)
   2) V2< 31988.61 189  91 2 (0.048 0.52 0.43 0 0 0)
     4) V51>=5.942954 89   2 2 (0 0.98 0.022 0 0 0) *
     5) V51< 5.942954 100  20 3 (0.09 0.11 0.8 0 0 0)
      10) V26< 8863.391 21  10 2 (0.43 0.52 0.048 0 0 0)
        20) V7< -1.138504 10   1 1 (0.9 0 0.1 0 0 0) *
        21) V7>=-1.138504 11   0 2 (0 1 0 0 0 0) *
      11) V26>=8863.391 79   0 3 (0 0 1 0 0 0) *
   3) V2>=31988.61 256 175 1 (0.32 0 0.0039 0.12 0.27 0.29)
     6) V2< 194291.8 177  96 1 (0.46 0 0.0056 0.062 0.056 0.42)
      12) V18>=7439.422 71   3 1 (0.96 0 0.014 0.028 0 0) *
      13) V18< 7439.422 106  32 6 (0.12 0 0 0.085 0.094 0.7)
        26) V67< 5.20329 29  16 1 (0.45 0 0 0.1 0.34 0.1)
          52) V20>=0.8470795 13   0 1 (1 0 0 0 0 0) *
          53) V20< 0.8470795 16   6 5 (0 0 0 0.19 0.62 0.19) *
        27) V67>=5.20329 77   6 6 (0 0 0 0.078 0 0.92) *
     7) V2>=194291.8 79  19 5 (0 0 0 0.24 0.76 0)
      14) V11>=9.623032 21   2 4 (0 0 0 0.9 0.095 0) *
      15) V11< 9.623032 58   0 5 (0 0 0 0 1 0) *
```

4.2 Naiwny klasyfikator Bayesa.

Do testów wykorzystano naiwny klasyfikator Bayesa z pakietu *e1071*. Klasyfikator Bayesa dla atrybutów o wartościach dyskretnych liczy bezpośrednio prawdopodobieństwa warunkowe, natomiast w przypadku atrybutów o wartościach ciągłych estymuje dla każdego atrybutu rozkład prawdopodobieństwa, co wiąże się z pewną niedokładnością predykcji. W związku z tym dobrze przeprowadzona dyskretyzacja atrybutu powinna polepszyć jakość klasyfikacji, dzięki uniknięciu estymacji rozkładu prawdopodobieństwa wartości atrybutu. W ramach eksperymentu przeprowadzono 5-krotną walidację krzyżową na zbiorze danych, wyniki zestawiono w tabeli 4.1. W tabeli 4.3 zestawiono wyniki dla dyskretyzacji zstępującej i kryterium stopu wykorzystującego minimalny spadek entropii w obszarze interesujących wartości parametru (zadanego minimalnego spadku) wraz ze średnią ilością utworzonych przedziałów. Wyniki 5-krotnej walidacji krzyżowej na zbiorze danych z ograniczonym zbiorem dyskretyzowanych atrybutów zestawiono w tabeli 4.2 - eksperymenty te dawały mniej interesujące wyniki w związku z czym skupiono się na eksperymentach dysretyzujących wszystkie dostępne atrybuty.

4.2.1 Porównanie działania metody wstępującej i zstępującej.

Jakość dyskretyzacji wstępującej i zstępującej porównano dla zadanej liczby przedziałów dla każdego atrybutu. W tabeli 4.4 zestawiono jakość klasyfikacji przy użyciu klasyfikatora Bayesa.

metoda	kryterium stopu	parametr	średnia dokładność	odchylenie std
–	–	–	0.7592	0.0306
TopDown	delta (2.1)	–	0.8498	0.0436
TopDown	minimalna wartość entropii	0.18	0.8315	0.0286
TopDown	zadana ilość przedziałów	6	0.8352	0.0541
BottomUp	maksymalna wartość χ^2	0.01	0.3198	0.0226
BottomUp	zadana ilość przedziałów	6	0.3423	0.0434

Tabela 4.1: $V1 \sim .$, dokładność klasyfikacji - klasyfikator Bayesa, 5-krotna walidacja krzyżowa

metoda	kryterium stopu	parametr	średnia dokładność	odchylenie std
–	–	–	0.8711	0.0435
TopDown	delta (2.1)	–	0.8835	0.0158
BottomUp	maksymalna wartość χ^2	0.01	0.3398	0.0926

Tabela 4.2: $V1 \sim V_{imp}$, dokładność klasyfikacji - klasyfikator Bayesa, 5-krotna walidacja krzyżowa

min spadek entropii	średnia dokładność	odchylenie std	średnia ilość przedziałów
0.20	0.8320	0.0514	1.82
0.19	0.8074	0.0601	3.85
0.18	0.8315	0.0286	4.03
0.17	0.8300	0.0633	10.42
0.16	0.8108	0.0681	15.46
0.15	0.8143	0.0425	19.39
0.14	0.8118	0.0510	28.46
0.13	0.7745	0.0613	44.85
0.12	0.7362	0.0682	66.90
0.11	0.7271	0.0630	92.32
0.10	0.6922	0.0385	123.52

Tabela 4.3: $V1 \sim .$, dokładność klasyfikacji - klasyfikator Bayesa, 5-krotna walidacja krzyżowa, TopDown + MinEntropyDecreaseCriterion

metoda	średnia dokładność	odchylenie std
–	0.7592	0.0306
TopDown	0.8352	0.0541
BottomUp	0.3423	0.0434

Tabela 4.4: $V1 \sim .$, porównanie działania metody zstępującej i wstępującej, dokładność klasyfikacji - klasyfikator Bayesa, 5-krotna walidacja krzyżowa, zadana ilość przedziałów = 6

4.3 Wnioski.

Dyskretyzacja okazała się najbardziej korzystna w przypadku zdyskretyzowania wszystkich atrybutów zbioru danych – uzyskano dużą poprawę dokładności klasyfikacji przy użyciu dyskretyzacji zstępującej. W przypadku wybrania tylko najbardziej istotnych atrybutów skok jakościowy nie jest już tak widoczny, choć zauważalna jest pewna stabilizacja wyników – mniejsze odchylenie stan-

dardowe. Prawdopodobnie wynika to z faktu, że duża ilość nieistotnych atrybutów mocno zaburza działanie klasyfikatora Bayesa i wskazuje na konieczność selekcji atrybutów przy praktycznej klasyfikacji danych. Dla kryterium zstępującego przetestowano trzy różne kryteria stopu – wyniki wskazują, że najlepszym (a jednoczesnie najwygodniejszym w użyciu, bo bezparametrycznym) kryterium stop jest kryterium *delta*, bardziej elastyczne od dwóch pozostałych, dopasowujące się lokalnie do analizowanego przedziału. Okazało się jednak, że dwa pozostałe kryteria – minimalny próg spadku entropii oraz żądana ilość przedziałów uzyskały wyniki bardzo zbliżone do kryterium *delta*. Wymagało to jednak optymalizacji wartości ich parametrów. Dla kryterium wstępującego nie udało się uzyskać dobrych wyników, mimo sprawdzenia działania algorytmu na publicznie dostępnych przykładach oraz eksperymentów z wartością parametru χ^2_{min}. Słabe wyniki dyskretyzacji były spowodowane osiąganiem niskich wartości χ^2 dla przyległych wartości. To powodowało zbyt częste łączenie przyległych przedziałów wartości. Przypuszczalnie analizowany zbiór danych kiepsko nadaje się do stosowania tej metody, co wskazuje na konieczność dobierania metody dyskretyzacji do analizowanego zbioru danych.

Rozdział 5

Pakiet R

Projekt został zaimplementowany jako pakiet języka R, *discretize*, posiadający następujący interfejs:

- CrossValidateBayes
- RequestedIntervalsNumCriterion
- TopDown
  - predict
  - print
  - summary
- DeltaCriterion
- MinEntropyDecreaseCriterion
- BottomUp
  - predict
  - print
  - summary
- MinChiCriterion

Każda publiczna metoda została udokumentowana (polecenie help()), a do pakietu dołączono testy jednostkowe napisane przy użyciu RUnit, wykonywane podczas budowania pakietu i wywoływane poleceniem "R CMD check katalog_pakietu". Do pakietu został dołączony zestaw danych batch1.dat [4], który można załadować przy użyciu wywołania data(gas1). Pakiet został sprawdzony i zbudowany bez ostrzeżeń przy użyciu poleceń:

```
R CMD check discretize
R CMD build --resave-data discretize
```

Oraz zainstalowany poleceniem:

```
R CMD INSTALL --clean discretize_1.0.tar.gz
```

Bibliografia

[1] Google's R style guide. http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html.

[2] P. Cichosz. *Systemy uczące się*. Wydawnictwa Naukowo-Techniczne, 2000.

[3] A. Frank and A. Asuncion. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010.

[4] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A. Ryan, Margie L. Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles, sensors and actuators b: Chemical (2012) doi: 10.1016/j.snb.2012.01.074. http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset, 2012.