LEMRG – algorytm generowania pokoleń reguł decyzji dla baz danych z dużą liczbą atrybutów

Łukasz Piątek, Jerzy W. Grzymała-Busse

Katedra Systemów Ekspertowych i Sztucznej Inteligencji, Wydział Informatyki Stosowanej

Wyższa Szkoła Informatyki i Zarządzania, ul. Mjr. H. Sucharskiego 2, 35-235 Rzeszów

Geneza

- **Zastosowanie wybranych metod uczenia maszynowego (ang. *machine learnig*), umożliwiających generowanie modeli uczenia w warunkach niepewności, w szczególności w procesach:**
  - **drążenia danych *DM* (ang. *Data Mining*),**

    oraz

  - **odkrywania wiedzy w danych *KDD* (ang. *Knowledge Discovery in Databases*).**

- **Regułowa reprezentacja wiedzy *JEŻELI … TO …* .**

Cel badań

- ***Cel badań*:**

  Opracowanie *metodologii*, która powinna umożliwić usprawnienie aktualnie stosowanych metod klasyfikacji, w celu uzyskania optymalnych (lub quasi-optymalnych) wybranych modeli uczenia maszynowego (*reguł decyzji*), głównie w zakresie *zmniejszenia* (być może znacznego) ich *błędu klasyfikacji*,

- **Zestaw skutecznych *metod* (algorytmów) *klasyfikacji*, stanowiących połączenie:**

  - **metod wstępnego przygotowania (*preprocessingu*) wejściowych zbiorów danych uczących,**

    oraz

  - **algorytmów indukcji *pokoleń reguł decyzji*, umożliwiających klasyfikację *specyficznych* typów danych.**

Założenia

- **Moduł *preprocessingu* – moduł umożliwiający przygotowanie wejściowych zbiorów danych uczących, zawierających:**
  - **dane *sprzeczne*,**
  - **dane *niepełne*, oraz**
  - **dane opisane atrybutami *numerycznymi*,**

- ***Specyficzność* zbioru danych:**
  - **zbiór danych opisanych bardzo *dużą liczbą atrybutów* (np. nawet *kilka* lub *kilkadziesiąt tysięcy* atrybutów), oraz**
  - **jednocześnie zbiór danych tego typu może zawierać relatywnie *niewielką liczbę przypadków* dostępnych w zbiorze uczącym (np. mniej niż 100).**

Założenia (*c.d.*)

- **Wybrane *ograniczenia* stosowania metod uczenia maszynowego do tworzenia *reguł decyzji*:**
  - ***niestabilność* klasyfikatorów,**
  - **atrybuty *nadmiarowe*, oraz**
  - ***duża liczba* reguł,**

- ***Cechy* nowo-opracowywanych algorytmów (klasyfikatora):**
  - ***stabilność* klasyfikatora dla przestrzeni (zbioru) danych o relatywnie małej liczbie zdefiniowanych przypadków (podobnie jak np. w *boostingu* lub/oraz *baggingu*) – *iteracyjne* generowanie skumulowanego zbioru reguł decyzji (ang. *cumulative rule set*),**
  - ***klasyfikator optymalny* (lub *quasi-optymalny*) – *mniejszy* (lub co najwyżej analogiczny) *błąd klasyfikacji* w stosunku do klasyfikatorów generowanych metodami standardowymi.**

Założenia (*c.d.2*)

- *Cechy* nowo-opracowywanych algorytmów (klasyfikatora) (*c.d.*):

- Generowanie poszczególnych *pokoleń* zbiorów reguł decyzji z zastosowaniem algorytmu indukcji reguł decyzji *MLEM2* (ang. *Modified Learning from Examples Module, version 2*) :

  - obsługa zbiorów danych wejściowych zawierających (również) przypadki (*i*) *sprzeczne* ,(*ii*) *niezupełne* oraz (*iii*) opisane atrybutami *numerycznymi*,

  - wbudowane wybrane operacje weryfikacji zbioru reguł, odpowiedzialne za usuwanie ze zbioru reguł (*i*) *nadmiarowych warunków* oraz (*ii*) *nadmiarowych reguł*, oraz

  - tworzenie minimalnego opisu dyskryminującego przybliżenie danej klasy decyzyjnej za pomocą specyficznej zasady generowania kolejnych pokryć – reguły w części warunkowej zawierają pary *atrybut-wartość* (tzw. *pary dominujące*), które – w odróżnieniu od *atrybutów nadmiarowych* – charakteryzują się wysokim (bądź najwyższym) stopniem korelacji z poszczególnymi wartościami zmiennej decyzyjnej.

Schemat działania

- Schemat działania zestawu algorytmów indukcji *pokoleń reguł*:

Walidacja

- *Walidacja* algorytmów generowania *klasyfikatorów* (dla każdej iteracji *pokoleń reguł decyzji*) z zastosowaniem różnych metod:
  - *leave-one-out* – zbiory danych uczących zawierające mniej niż 100 przypadków,
  - *n-krotna walidacji skrośna* – zbiory danych uczących zawierające od 100 do 1000 przypadków,

    oraz
  - *holdout* – zbiory danych uczących zawierające powyżej 1000 przypadków.

Badania wstępne

- Dane wejściowe – zbiór danych typu *MicroRNA*, określających stopień zagrożenia wystąpienia raka u ludzi[1)],
- Zbiór zawierający 68 przypadków, opisanych 217 atrybutami,
- 11 klas decyzyjnych (*konceptów*), w tym:
  - *BLDR* (6 przypadków),
  - *BRST* (6 przypadków),
  - *COLON* (7 przypadków),
  - *KID* (4 przypadki),
  - *LUNG* (5 przypadków),
  - *MELA* (3 przypadki),
  - *MESO* (8 przypadków),
  - *OVARY* (5 przypadków),
  - *PAN* (8 przypadków),
  - *PROST* (6 przypadków), oraz
  - *UT* (10 przypadków),
- Najlepsze wyniki dla 2 klas (*BRST* oraz *OVARY*).

[1)] Fang J., Grzymała-Busse J.W.: *Mining of MicroRNA Expression Data – A Rough Set Approach*, In: Proc of the 1st International Conference on Rough Sets and Knowledge Technology (RSKT'06), Springer-Verlag, Berlin, Heidelberg, 2006, pp.758-765.

Badania wstępne (*c.d.*)

- ***Pierwsze*** **pokolenie reguł:**
  - **zbiór wejściowy opisany *217* atrybutami,**
  - **11 reguł decyzji, po jednej regule dla każdej klasy decyzji,**
  - **reguły dla klas *BRST* oraz *OVARY*:**

3, 6, 6
(EAM335, 5.3581..7.30918) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 5, 5
(EAM335, 7.87172..11.003) & (EAM159, 7.95896..10.6737) &
(EAM233, 5..6.87862) –> (Label, OVARY)

- **Usunięcie *atrybutów dominujących* (2 *podejścia*):**
  - ~~**wszystkich atrybutów, które zastosowano w regułach (błąd *100%*),**~~
    oraz
  - **jedynie atrybutów z początkowego (pierwszego) warunku dla każdej reguły (usunięcie 11 atrybutów dominujących).**

Badania wstępne (*c.d.2*)

- ***Drugie*** **pokolenie reguł:**
  - **zbiór wejściowy opisany *206* atrybutami,**
  - **reguły dla klas *BRST* oraz *OVARY*:**

```
3, 6, 6
(EAM159, 5..7.56154) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
2, 1, 1
(EAM159, 7.95896..8.14026) & (EAM233, 5..6.87862) –> (Label, OVARY)
4, 4, 4
(EAM159, 8.14026..10.6737) & (EAM317, 5..5.1027) &
(EAM186, 7.88734..10.8281) & (EAM233, 5..6.87862) –> (Label, OVARY)
```

  - **usunięcie kolejnych 7 atrybutów dominujących,**

- ***Trzecie*** **pokolenie reguł:**
  - **Zbiór wejściowy opisany 199 atrybutami,**
  - **reguły dla klas *BRST* oraz *OVARY*:**

```
2, 1, 1
(EAM304, 9.39729..9.59664) & (EAM261, 6.8182..9.79852) –>
(Label, BRST)
4, 5, 5
(EAM304, 9.59664..11.8053) & (EAM261, 6.8182..9.79852) &
(EAM238, 5..5.01569) & (EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 5, 5
(EAM225, 5.125..9.1908) & (EAM317, 5..5.1027) &
(EAM233, 5..6.87862) –> (Label, OVARY)
```

  - **zwiększenie błędu klasyfikacji (brak kolejnych pokoleń reguł).**

Badania wstępne (*c.d.3*)

- **Wyniki klasyfikacji dla poszczególnych *pokoleń reguł*:**

Rule Set	Number of correctly classified cases	
	BRST	Ovary
First rule generation	2	3
Second rule generation	4	2
Third rule generation	0	2
Combined rule set (first and second rule generations)	4	3

- **Finalny zbiór reguł decyzji (ang. *cumulative rule set*):**
  - **reguły z 1-go oraz 2-go pokolenia,**
  - **różna *siła* (ang. *strength*) *reguł* z poszczególnych pokoleń:**

3, 2, 2
(EAM335, 5.3581..7.30918) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 1, 1
(EAM159, 5..7.56154) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 2, 2
(EAM335, 7.87172..11.003) & (EAM159, 7.95896..10.6737) &
(EAM233, 5..6.87862) –> (Label, OVARY)
4, 1, 1
(EAM159, 8.14026..10.6737) & (EAM317, 5..5.1027) &
(EAM186, 7.88734..10.8281) & (EAM233, 5..6.87862) –>
(Label, OVARY)

Algorytm *LEMRG*

- Autorski algorytm indukcji *pokoleń* reguł decyzji *LEMRG* (ang. *Learning from Examples Module based on Rules Generations*),

- Na podstawie *zbioru danych wejściowych* (w tym z brakującymi wartościami – znak *?*) oraz kolejnego *pokolenia reguł decyzji* następuje podział zbioru danych na 2 podzbiory:
  - *lokalny* – wszystkie wartości zbioru danych, które spełniają reguły są zastępowane nowymi znakami *?*, oraz
  - *globalny* – gdzie nie zwraca sie uwagi na (*i*) prawą stronę reguł ani na (*ii*) znak koniunkcji, a zastępuje wszelkie wartości spełniane przez dowolny warunek ze zbioru reguł nowymi znakami *?*,

- Generowanie kolejnego pokolenia reguł następuje dla obu zbiorów (*lokalnego* oraz *globalnego*),

---

- Powyższa metoda umożliwia wyłączanie jedynie wybranych par *atrybut-wartość*, a nie jak uprzednio (*badania początkowe*) całych atrybutów (tj. dla wszystkich dopuszczalnych wartości).

Algorytm *LEMRG* (*c.d.*)

- **Przykład modyfikacji zbioru uczącego w algorytmie *LEMRG*:**

Zbiór danych wejściowych (*początkowy*)

Lp.	Temperature	Hemoglobin	Blood_pressure	Oxygen_saturation	Comfort
1	low	fair	low	fair	low
2	?	fair	normal	poor	low
3	normal	?	low	good	low
4	normal	good	?	?	medium
5	?	good	?	?	medium
6	low	?	normal	fair	medium
7	normal	?	normal	good	medium
8	normal	?	?	good	very_low
9	high	good	?	fair	very_low
10	high	good	normal	good	medium

Reguły decyzji (*MLEM2*)

```
1, 2, 2
(Hemoglobin,fair) -> (Comfort,low)
1, 2, 2
(Blood_Pressure,low) -> (Comfort,low)
2, 2, 2
(Blood_Pressure,normal) & (Oxygen_Saturation,good)
-> (Comfort,medium)
2, 1, 1
(Oxygen_Saturation,fair) & (Blood_Pressure,normal)
-> (Comfort,medium)
2, 1, 1
(Temperature,normal) & (Hemoglobin,good)
-> (Comfort,medium)
2, 1, 1
(Temperature,high) & (Oxygen_Saturation,fair)
-> (Comfort,very_low)
```

Zbiór *lokalny*

Lp.	Temperature	Hemoglobin	Blood_pressure	Oxygen_saturation	Comfort
1	low	?	?	fair	low
2	?	?	normal	poor	low
3	normal	?	?	good	low
4	?	?	?	?	medium
5	?	good	?	?	medium
6	low	?	?	?	medium
7	normal	?	?	?	medium
8	normal	?	?	good	very_low
9	?	good	?	?	very_low
10	high	good	?	?	medium

Zbiór *globalny*

Lp.	Temperature	Hemoglobin	Blood_pressure	Oxygen_saturation	Comfort
1	low	?	?	?	low
2	?	?	?	poor	low
3	?	?	?	?	low
4	?	?	?	?	medium
5	?	?	?	?	medium
6	low	?	?	?	medium
7	?	?	?	?	medium
8	?	?	?	?	very_low
9	?	?	?	?	very_low
10	?	?	?	?	medium

Podsumowanie. Dalsze kierunki rozwoju badań

- **Generowanie pokoleń reguł z zastosowaniem innych algorytmów indukcji reguł, w tym *GTS*[2] (ang. *General-To-Specyfic*) oraz *AQ*[3]:**
  - **konieczność zaadaptowania (lub opracowania nowych) metod obsługi wejściowych zbiorów danych uczących, w celu standaryzacji sposobu klasyfikacji dla każdego stosowanego algorytmu indukcji reguł (*MLEM2*, *GTS* oraz *AQ*)**

- **Implementacja algorytmów w specjalizowanym narzędziu komputerowym,**
  - **generowania optymalnych (lub quasi-optymalnych) klasyfikatorów – zapisanych w formie połączonych zbiorów reguł decyzji (ang. *cumulative rule sets*) – cechujących się błędem klasyfikacji *mniejszym* (lub co najwyżej analogicznym) niż dla aktualnie stosowanych metod, w których wykorzystywane są pojedyncze zestawy reguł decyzji (tj. generowane w sposób tradycyjny).**

[2] Hippe Z.S.: *Uczenie maszynowe – obiecującą strategią przetwarzania informacji w biznesie?*, Informatyka 4(1997)27-31, 5(1997)29-33.

[3] Michalski R.S., Mozetic I., Hong J., Lavrac N.: *The multi-purpose incremental learning system AQ15 and its testing application to three medical domains*, In: Proc. of the AAAI'86, Philadelphia (USA), 1986, pp.1041-1045.

Wyższa Szkoła Informatyki i Zarządzania
ul. Sucharskiego 2, 35-225 Rzeszów, Polska

Dziękuję za uwagę…