Indukcja konstruktywna, redukcja danych

Marcin Szeląg

Zakład ISWD, Instytut Informatyki, Politechnika Poznańska

23.10.2019

Plan prezentacji

Wprowadzenie

Źródła trudności w eksploracji danych

- Bardzo duża liczba obserwacji / przykładów
- Zbyt duża liczba atrybutów / zmiennych
- Nieistotność części zmiennych dla klasyfikacji obiektów
- Wzajemne współzależności zmiennych warunkowych
- Równoczesna obecność zmiennych różnego typu
- Występowanie niezdefiniowanych wartości zmiennych
- Występowanie błędnych wartości zmiennych
- Nierównomierny rozkład kategorii dla zmiennej celu

Indukcja konstruktywna

Indukcja konstruktywna

Indukcja konstruktywna [R. Michalski] – przekształcanie przestrzeni hipotez uczenia w ten sposób, aby pojęcie docelowe lub wysokiej jakości grupowanie mogło być w niej reprezentowane dokładnie i oszczędnie oraz aby możliwe było efektywne nauczenie się go za pomocą stosowanego algorytmu uczenia się.

Przestrzeń hipotez - zbiór możliwych hipotez dotyczących pojęcia docelowego, które można skonstruować w oparciu o wybrany algorytm uczenia się oraz sposób reprezentacji danych.

Najczęściej rozważa się przekształcenie języka reprezentacji w odniesieniu do atrybutów – usuwanie zbędnych atrybutów i/lub tworzenie nowych atrybutów, zależnych funkcjonalnie od atrybutów istniejących dotychczas.

Celem indukcji konstruktywnej jest "poprawienie" opisu przykładów, tak aby wykorzystywany algorytm odkrywania wiedzy mógł uzyskiwać lepsze wyniki.

Rodzaje indukcji konstruktywnej

Ze względu na źródło informacji decydujących o sposobach przekształcania przestrzeni hipotez wyróżnia się:

- *indukcję konstruktywną sterowaną danymi* (ang. data-driven constructive induction – DCI) – sposób przekształcania przestrzeni hipotez określa się na podstawie analizy danych w zbiorze uczącym,
- *indukcję konstruktywną sterowaną wiedzą* (ang. knowledge-driven constructive induction – KCI) – sposób przekształcania przestrzeni hipotez określa się na podstawie wiedzy dziedzinowej o zbiorze uczącym,
- *indukcję konstruktywną sterowaną hipotezami* (ang. hypothesis-driven constructive induction – HCI) – sposób przekształcania przestrzeni hipotez określa się na podstawie analizy hipotez wygenerowanych dla dotychczasowego zbioru zmiennych.

Indukcja konstruktywna sterowana danymi (DCI)

Indukcja konstruktywna na podstawie danych polega na analizowaniu rozkładów wartości zmiennych i wzajemnych korelacji tych zmiennych. Za szczególny przypadek indukcji konstruktywnej sterowanej danymi można uznać *dyskretyzację zmiennych ciągłych* oraz *agregację zmiennych porządkowych*, gdyż metody te dobierają sposób podziału dziedziny zmiennej na podstawie analizy zbioru danych.

Indukcja konstruktywna sterowana wiedzą (KCI)

Akcentowanie roli wiedzy eksperta / analityka. Wiedza ta, opisując pewne aspekty dziedziny problemu, może umożliwić określenie właściwych przekształceń przestrzeni zmiennych. Często jednak ekspert najłatwiej może wyrazić swoją wiedzę o możliwych przekształceniach zmiennych już na etapie tworzenia zbioru danych, odpowiednio dobierając zmienne wejściowe.

Indukcja konstruktywna sterowana hipotezami (HCI)

W oparciu o analizę modelu danych generowanego przez wybrany algorytm eksploracji danych w kolejnych iteracjach można:

- usuwać zmienne niewykorzystywane w modelu lub takie, które nie mają zasadniczego znaczenia dla jego dokładności,
- tworzyć nowe zmienne w oparciu o pojawiające się w modelu kombinacje warunków co do wartości użytych zmiennych (wzorce), o ile kombinacje te są spełniane przez odpowiednio dużo obiektów uczących.

Indukcja konstruktywna – przykład

Dany jest następujący zbiór przykładów uczących:

nr.	wysokość	długość	szerokość	decyzja
1	2	12	2	1
2	6	4	2	1
3	3	8	2	1
4	4	4	3	1
5	12	4	2	2
6	4	12	2	2
7	8	6	2	2
8	6	8	3	2

Indukcja konstruktywna – przykład

Zbiór reguł decyzyjnych otrzymanych za pomocą algorytmu LEM2:

1. (długość = 4) & (wysokość =6) $\Rightarrow$ (decyzja = 1) {2}
2. (wysokość = 4) & (długość =4) $\Rightarrow$ (decyzja = 1) {4}
3. (wysokość = 3) $\Rightarrow$ (decyzja = 1) {3}
4. (wysokość = 2) $\Rightarrow$ (decyzja = 1) {1}
5. (długość = 8) & (wysokość =6) $\Rightarrow$ (decyzja = 2) {8}
6. (długość = 12) & (wysokość =4) $\Rightarrow$ (decyzja = 2) {6}
7. (wysokość = 12) $\Rightarrow$ (decyzja = 2) {5}
8. (wysokość = 8) $\Rightarrow$ (decyzja = 2) {7}

Indukcja konstruktywna – przykład

Wprowadzenie nowej zmiennej – wysokość $\times$ długość:

nr.	wysokość	długość	szerokość	**wys. $\times$ dł.**	decyzja
1	2	12	2	**24**	1
2	6	4	2	**24**	1
3	3	8	2	**24**	1
4	4	4	3	**16**	1
5	12	4	2	**48**	2
6	4	12	2	**48**	2
7	8	6	2	**48**	2
8	6	8	3	**48**	2

Indukcja konstruktywna – przykład

Zbiór reguł decyzyjnych otrzymanych za pomocą algorytmu LEM2:

1. (wysokość $\times$ długość $= 24$) $\Rightarrow$ (decyzja $= 1$), $\{1, 2, 3\}$
2. (wysokość $\times$ długość $= 16$) $\Rightarrow$ (decyzja $= 1$), $\{4\}$
3. (wysokość $\times$ długość $= 48$) $\Rightarrow$ (decyzja $= 2$), $\{5,6,7,8\}$

Redukcja liczby zmiennych

Motywacje redukcji liczby zmiennych (selekcji atrybutów):

- zmniejszenie przestrzeni hipotez $\Rightarrow$ szybsze i efektywniejsze działanie algorytmów eksploracji danych,
- czasami poprawa jakości przyszłej predykcji (unikanie nadmiernego dopasowania się modelu wiedzy do danych),
- łatwiejsza interpretacja stworzonego modelu wiedzy, np. ze względu na mniejszy rozmiar tego modelu.

Redukcja liczby zmiennych – c.d.

- W przestrzeni o dużej liczbie wymiarów prawdopodobnie nie wszystkie zmienne (atrybuty), które zostały zmierzone, będą konieczne do dokładnego rozróżnienia obiektów.
- Przykładowo, podczas rozróżniania obrazów twarzy mężczyzn i kobiet zmienne takie jak kolor oczu, włosów czy skóry będą prawdopodobnie mało przydatne do klasyfikacji.
- Możliwość wykorzystania wiedzy a priori o istotności zmiennych dla konkretnego zadania odkrywania wiedzy.
- W problemie klasyfikacji możliwe jest, iż niektóre zmienne wejściowe $X_1, X_2, \ldots X_n$ są niezwiązane ze zmienną wynikową Y oraz że dwie lub więcej zmiennych wejściowych zawiera w zasadzie te same informacje przydatne do przewidywania (korelacja!).
- Niektóre algorytmy eksploracji danych cechują się wewnętrznym doborem istotnych cech (np. drzewa decyzyjne; ryzyko nadmiernego dopasowania!), inne nie (np. kNN).

Redukcja liczby zmiennych (DCI)

- Niezależność Y od X_i: dla wszystkich wartości y i x_i mamy $p(Y = y|X_i = x_i) = p(Y = y)$ – problem skończoności próbki, brak ilościowej informacji o stopniu zależności.
- W przypadku gdy interesuje nas ilościowe oszacowanie zależności pomiędzy dowolną zmienną wejściową X_i a zmienną wynikową Y, możemy np. wyznaczyć współczynnik korelacji na każdej pary zmiennych (X_i, Y) – szacowanie tylko zależności liniowej dla zmiennych ilościowych.
- (Średnia) Informacja wzajemna (ang. mutual information):
  - dla zmiennych dyskretnych (jakościowych):
    $I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$,
  - dla zmiennych ciągłych (ilościowych):
    $I(X,Y) = \int_Y \int_X p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$.
- Informacja wzajemna jest równa zero $\Leftrightarrow$ zmienne są niezależne.

Redukcja liczby zmiennych (DCI) – c.d.

- Test χ^2 – badanie zgodności wzajemnej zmiennych jakościowych (nominalnych); współczynnik V-Cramera.
- Miary wykorzystujące entropię warunkową zmiennej wynikowej Y, pod warunkiem znajomości zmiennej wejściowej X_i:

  Entropia dyskretnej zmiennej Y:

  $$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y).$$

  Entropia warunkowa dyskretnej zmiennej Y, pod warunkiem znajomości dyskretnej zmiennej X_i:

  $$H(Y|X_i) = -\sum_{x_i \in X_i} p(x_i) \sum_{y \in Y} p(y|x_i) \log_2 p(y|x_i).$$

  Zysk informacyjny (ang. information gain), Quinlan 1993:

  $$InfoGain(Y, X_i) = H(Y) - H(Y|X_i).$$

Redukcja liczby zmiennych (DCI) – c.d.

- Problem – interakcje pomiędzy pojedynczymi zmiennymi X_i a zmienną Y niekoniecznie mówią cokolwiek o tym, jaki zbiór zmiennych oddziaływuje na Y.
- "Złośliwy" przykład – Y jako funkcja parzystości zmiennych binarnych (0-1) – Y nie zależy od żadnej pojedynczej zmiennej, ale jest deterministyczną funkcją całego zbioru zmiennych (interakcje nieliniowe i nieaddytywne!).
- Wniosek: w przypadku ogólnym zbiór k najlepszych pojedynczych zmiennych X_i (dobranych np. na podstawie korelacji ze zmienną wynikową Y) nie jest tym samym co najlepszy zbiór zmiennych o rozmiarze k.
- Dla zbioru zmiennych o rozmiarze n, mamy $2^n - 1$ możliwych niepustych podzbiorów zmiennych $\Rightarrow$ wyczerpujący przegląd wszystkich możliwych podzbiorów często nie jest możliwy.
- W praktyce stosuje się heurystyczne metody wybierania podzbiorów zmiennych, polegające np. na dodawaniu lub usuwaniu pojedynczej zmiennej w aktualnym kroku metody.

Redukcja liczby zmiennych (DCI) – c.d.

Metody oceny jakości podzbioru zmiennych wejściowych $F \subseteq \{X_1, X_2, \ldots, X_n\}$, $|F| = k$, w kontekście predykcji wartości zmiennej wynikowej Y:

- podejścia typu *filter*, np:
  - selekcja podzbioru zmiennych w oparciu o korelację (ang. Correlation-based Filter Approach),
  - selekcja podzbioru zmiennych w oparciu o jakość (przybliżenia) klasyfikacji w sensie teorii zbiorów przybliżonych (ang. Rough Set Theory – RST),
- podejścia typu *wrapper*.

Redukcja liczby zmiennych (DCI) – c.d.

Model *filter* – redukcja liczby zmiennych niezależnie od przyjętego algorytmu eksploracji danych (wstępne przetwarzanie zbioru zmiennych wejściowych). Wykorzystanie całego dostępnego zbioru danych przy selekcji zmiennych. Zaletą jest szybkość działania i skalowalność wraz ze wzrostem liczby zmiennych i obiektów.

Model *wrapper* – wykorzystanie docelowego algorytmu eksploracji danych w celu oszacowania jakości każdego rozważanego podzbioru zmiennych F, za pomocą wybranej statystycznej techniki próbkowania, np. k-krotnej walidacji krzyżowej (ang. k-fold cross-validation). Wadą tego podejścia jest duży nakład obliczeniowy wynikający z wielokrotnego wywoływania algorytmu uczącego. Z kolei konsekwencją dużego nakładu obliczeniowego jest problem skalowalności modelu wraz ze wzrostem liczby zmiennych i obiektów.

Selekcja zmiennych w oparciu o korelację (DCI)

Wykorzystanie macierzy korelacji dla wszystkich par zmiennych do szacowania "dobroci" (ang. merit) zbioru k zmiennych F (ang. correlation-based merit measure):

$$Merit(F) = \frac{k\overline{r_{fd}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}},$$

gdzie: $\overline{r_{fd}}$ – średnia korelacja zmiennej $X_i \in F$ ze zmienną Y, $\overline{r_{ff}}$ – średnia wzajemna korelacja zmiennych ze zbioru F.

Intuicja stojąca za przedstawionym podejściem: dobry zbiór zmiennych zawiera zmienne wysoce skorelowane ze zmienną wynikową Y i nieskorelowane nawzajem.

Selekcja zmiennych w oparciu o korelację (DCI) – c.d.

W przypadku wyznaczania korelacji pomiędzy zmienną ciągłą X_i a zmienną dyskretną X_j o wartościach ze zbioru $\{v_1, v_2, \ldots, v_t\}$, zmienna dyskretna podlega zamianie na t zmiennych binarnych $X_{j_1}, X_{j_2}, \ldots, X_{j_t}$, w oparciu o zależność:

$$X_{j_k} = \begin{cases} 1 & \Leftrightarrow X_j = v_k \\ 0 & \text{w przeciwnym przypadku,} \end{cases}$$

gdzie $k \in \{1, 2, \ldots, t\}$.

Wówczas korelacja zmiennej ciągłej X_i ze zmienną dyskretną X_j wyraża się wzorem:

$$r_{X_i X_j} = \sum_{k=1}^{t} p(X_j = v_k) r_{X_i X_{j_k}}.$$

Selekcja zmiennych w oparciu o jakość klasyfikacji (DCI)

Jakość (*przybliżenia*) *klasyfikacji* w sensie teorii zbiorów przybliżonych (Y – zmienna jakościowa):

$$\gamma_F(Cl) = \frac{\sum_{d \in Cl} |\underline{F}(d)|}{|U|},$$

gdzie: $F \subseteq \{X_1, X_2, \ldots, X_n\}$, $\underline{F}(d)$ – F-dolne przybliżenie klasy decyzyjnej $d \in Cl$, Cl – zbiór klas decyzyjnych, klasa dec. – zbiór obiektów o tej samej wartości zmiennej Y, U – zbiór wszystkich obiektów.

Interesują nas wszystkie redukty zbioru zmiennych wejściowych $X = \{X_1, X_2, \ldots, X_n\}$, czyli takie *minimalne* podzbiory $F \subseteq X$, dla których:

$$\gamma_F(Y) = \gamma_X(Y).$$

Przekształcanie zmiennych

- Celem jest przekształcenie przestrzeni zmiennych $X_1, X_2, \ldots, X_n$ w przestrzeń $Z_1, Z_2, \ldots, Z_{n'}$, gdzie zazwyczaj $n' << n$ a zmienne Z_i są zdefiniowane jako funkcje oryginalnych zmiennych X_i.
- Podejścia do redukcji wymiarowości przestrzeni:
  - *Regresja poszukiwania rzutowania* (ang. Projection Pursuit Regression – PPR), J.H. Freidman, J.W. Tukey, 1974 – poszukiwanie optymalnego liniowego rzutowania danych na przestrzeń o mniejszej liczbie wymiarów (często tylko jeden / dwa wymiary)
  - *Analiza składowych głównych* (ang. Principal Components Analysis – PCA)
  - Inne – np. analiza czynników, analiza składowych niezależnych, ...

Przekształcanie zmiennych – c.d.

Regresja poszukiwania rzutowania używa struktury modelu postaci:

$$\hat{y}(\mathbf{x}) = \alpha_0 + \sum_{i=1}^{n'} h_i(\alpha_i^T \mathbf{x}),$$

gdzie: $\hat{y}(\mathbf{x})$ – estymowana wartość zmiennej wynikowej Y,
$\mathbf{x}$ – n-wymiarowy wektor współrzędnych punktu danych,
α_i – n-wymiarowy wektor wag,
$\alpha_i^T \mathbf{x}$ – rzut wektora $\mathbf{x}$ na i-ty wektor wag,
$h_i(\cdot)$ – funkcja nieliniowa skalarnego rzutowania.

Dla każdego z n' wymiarów nowej przestrzeni, zachodzi konieczność doboru postaci funkcji $h_i(\cdot)$ oraz "kierunku rzutowania" α_i.

Przedstawiona postać struktury modelu jest podstawą sieci neuronowych, dla których zazwyczaj $h_i(t) = \frac{1}{1+e^{-t}}$.

Przekształcanie zmiennych – c.d.

Ograniczenia modelu PPR:

- trudność interpretacji modelu gdy $n' > 1$,
- duża złożoność obliczeniowa algorytmów estymacji parametrów modelu $\Rightarrow$ niepraktyczność algorytmów dla dużych zbiorów danych.

Przekształcanie zmiennych – c.d.

Analiza składowych głównych:

- Oryginalne n zmiennych przewidujących $X_1, X_2, \ldots, X_n$ jest zastępowane nowym zbiorem n zmiennych $Z_1, Z_2, \ldots, Z_n$, powstających w wyniku kombinacji liniowych oryginalnych zmiennych.
- Wariancja zbioru danych wyrażona w terminach nowych zmiennych jest maksymalna.
- Sekwencyjne wydobywanie większości zmienności danych w przestrzeniu $X \Rightarrow$ już kilka pierwszych składowych Z_i może zawierać większość informacji tkwiących w danych.
- Składowe są ortogonalne $\Rightarrow$ łatwiejsza interpretacja.

Ocena złożoności hipotez i tworzonych zmiennych

Niezależnie od przyjętej do oceny hipotez miary złożoności, takiej jak np. liczb węzłów drzewa decyzyjnego czy liczba reguł decyzyjnych, trzeba pamiętać o tym, aby w odpowiedni sposób uwzględnić także złożoność tworzonych zmiennych!

Pominięcie tego elementu może owocować bardzo prostymi hipotezami kosztem bardzo złożonych zmiennych – cała złożoność związana z reprezentowaniem pojęcia docelowego zostanie przeniesiona do zmiennych.

Redukcja liczby obiektów

Redukcja liczby obiektów

Duża liczba obiektów, niezbędna do uzyskania dobrego i statystycznie istotnego modelu wiedzy, może być poważną przeszkodą praktyczną, ze względu na duży koszt obliczeniowy. Stąd też często zachodzi konieczność ograniczania liczby przykładów przetwarzanych przez algorytmy eksploracji danych.

Możliwe techniki ograniczania liczby przykładów obejmują:

- okienkowanie,
- redukcję liczby przykładów (próbkowanie zewnętrzne),
- próbkowanie wewnętrzne,
- dekompozycję zbioru danych.

Okienkowanie

- Uniwersalne podejście do nadzorowanego uczenia się na podstawie dużych zbiorów danych.
- Uczenie się na podstawie początkowo małych i w miarę potrzeby rosnących losowych podzbiorów W zbioru danych D, nazywanych *zbiorami roboczymi*.
- Za każdym razem model wiedzy stworzony na zbiorze $W \subseteq D$ jest testowany na pozostałej części zbioru danych, tj. na zbiorze $D - W$.
- Kolejny zbiór roboczy W' powstaje ze zbioru W poprzez dodanie niektórych, losowo wybranych, przykładów błędnie sklasyfikowanych przez aktualny model wiedzy.

Okienkowanie – c.d.

- Procedura powtarzana jest tak długo, jak długo kolejny model wiedzy jest lepszy od poprzedniego, biorąc pod uwagę błąd modelu na całym zbiorze danych D, lub dopóki błąd aktualnego modelu nie spadnie poniżej zadanego poziomu.
- Zalecane jest k-krotne powtórzenie procedury okienkowania z różnymi początkowymi losowymi zbiorami roboczymi $W \subseteq D$.
- W takim przypadku, ostatecznie wybierany jest najlepszy z ostatnich, przyciętych w celu uniknięcia nadmiernego dopasowania, modeli danych ze wszystkich k powtórzeń. Jakość modelu może być tutaj mierzona zarówno jego błędem na zbiorze danych, jak i np. złożonością modelu.

Redukcja liczby przykładów (próbkowanie zewnętrzne)

- Operacja zmniejszenia rozmiaru zbioru danych, która z dużym prawdopodobieństwem pozostawia w zredukowanych danych interesujące i użyteczne zależności, występujące w pełnym zbiorze.
- Swoista technika zapobiegania nadmiernemu dopasowaniu – przycinanie nie hipotezy, a zbyt dużego zbioru danych.
- Cel – wybór podzbioru złożonego z przykładów uznanych za najbardziej reprezentatywne.
- Przykładowa realizacja – określenie miary odległości dla przykładów i dążenie do wybierania przykładów maksymalnie od siebie odległych (maksymalizacja średniej odległości pomiędzy dowolnymi dwoma przykładami ze zredukowanego zbioru, przy jednoczesnym możliwie niewielkim zróżnicowaniu tych odległości) – pozwala na w miarę "równomierne" i jednocześnie "reprezentatywne" zredukowanie danych (analogia do wzięcia tylko węzłów siatki 2D).

Próbkowanie wewnętrzne

Przeniesienie wyboru próbki przykładów do samego algorytmu eksploracji danych, stosowanego bezpośrednio do oryginalnego zbioru danych o dużym rozmiarze.

Idea polega na tym, aby algorytm ten wykonywał obliczenia zależne od liczby obiektów na podstawie losowej, każdorazowo niezależnie wybranej próbki (żaden przykład nie jest z góry skazany na pominięcie!).

Przykładowo podczas indukcji drzewa decyzyjnego można próbkować zbiór obiektów w każdym węźle na potrzeby wyboru testu w tym węźle.

Dekompozycja zbioru danych

Zależności występujące w dużych zbiorach danych mogą być zbyt złożone, aby opisujące je wzorce były czytelne i możliwe do interpretacji, co jest celem odkrywania tych zależności.

W takiej sytuacji pozostaje dekompozycja zbioru danych na mniejsze podzbiory i poszukiwanie w tych podzbiorach prostszych zależności o ograniczonej stosowalności.

Nierównomierny rozkład kategorii zmiennej celu

- Problem pojawiający się w rzeczywistych zbiorach danych, szczególnie często dla pojęć pojedynczych (tylko dwie kategorie i niewielka liczba przykładów pozytywnych lub negatywnych)
- Nierównomierny rozkład kategorii zakłóca proces tworzenia modelu – pozorna atrakcyjność "reguł większościowych"
- Przykładowe techniki niwelujące nierównomierność rozkładu:
  - próbkowanie przykładów z kategorii większościowych – wybór pewnej liczby przykładów z każdej kategorii, która jest zbyt liczna w stosunku do pozostałych,
  - techniki replikacji przykładów z kategorii mniejszościowych – multiplikowanie przykładów z kategorii zbyt słabo reprezentowanych w zbiorze danych; gdy pożądana liczba przykładów danej kategorii wynosi k', a faktyczna liczba wynosi k, to oczekiwana liczba kopii każdego przykładu tej kategorii to $\frac{k'}{k}$,
  - generowanie przykładów syntetycznych – SMOTE.

Podsumowanie

- Indukcja konstruktywna – selekcja i tworzenie atrybutów – zobacz np. rozdz. 4.4 w pracy

  `http://www.cs.put.poznan.pl/kkrawiec/pubs/phd.pdf`

- Różne techniki redukcji liczby obiektów – zobacz np.

  `https://machinelearningmastery.com/statistical-sampling-and-resampling`

- Techniki równoważenia rozkładu kategorii atrybutu decyzyjnego – zobacz np. `https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset`

Zakończenie

Dziękuję za uwagę