

Uczenie modeli kanonicznych w sieciach Bayesowskich - efektywne uczenie modelu Noisy OR/MAX z danych.

Krzysztof Nowak

Politechnika Białostocka

23.10.2012

Sieć Bayesowska - struktura służąca do przedstawiania zależności pomiędzy zdarzeniami bazując na rachunku prawdopodobieństwa.

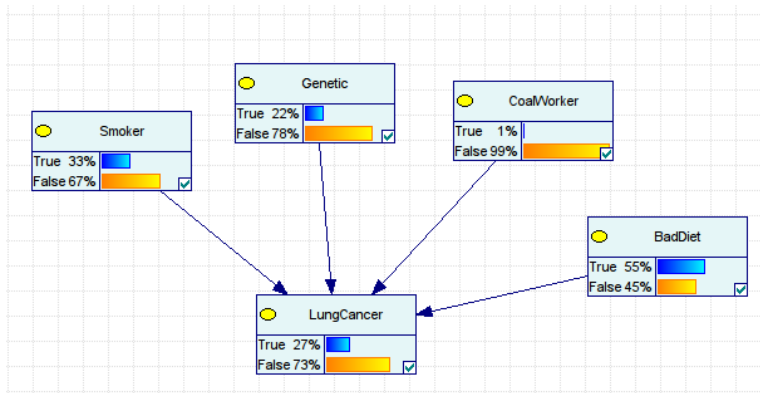
Sieć Bayesowska - struktura służąca do przedstawiania zależności pomiędzy zdarzeniami bazując na rachunku prawdopodobieństwa. W sieciach bayesowskich można wyróżnić:

Sieć Bayesowska - struktura służąca do przedstawiania zależności pomiędzy zdarzeniami bazując na rachunku prawdopodobieństwa. W sieciach bayesowskich można wyróżnić:

- Strukturę sieci - skierowany, acykliczny graf

Sieć Bayesowska - struktura służąca do przedstawiania zależności pomiędzy zdarzeniami bazując na rachunku prawdopodobieństwa. W sieciach bayesowskich można wyróżnić:

- Strukturę sieci - skierowany, acykliczny graf
- Parametry sieci - wartości liczbowe określające prawdopodobieństwo poszczególnych zdarzeń



Rysunek: Przykładowa sieć bayesowska - Genie 2.0

CPT - Conditional Probability Table

Node properties: LungCancer

General Definition Format User properties Value

Add Insert

| | | | | | | | | |
|------------|--------------------------|-----------|------------|----------|----------|-----------|----------|----------|
| Smoker | <input type="checkbox"/> | True | | | | | | |
| Genetic | <input type="checkbox"/> | True | | | False | | | |
| CoalWorker | <input type="checkbox"/> | True | | False | | True | | False |
| BadDiet | | True | False | True | False | True | False | True |
| True | | 0.7637068 | 0.75386125 | 0.722008 | 0.710425 | 0.6849424 | 0.671815 | 0.629344 |
| False | | 0.2362932 | 0.24613875 | 0.277992 | 0.289575 | 0.3150576 | 0.328185 | 0.370656 |

Rysunek: CPT - Genie 2.0

- Wykładniczy przyrost parametrów ze względu na ilość węzłów rodzicielskich.

Parametry sieci

| Parent | Smoker | Genetic | CoalWorker | BadDiet | LEAK |
|--------|--------|---------|------------|---------|------|
| True | 0.61 | 0.25 | 0.15 | 0.04 | 0.01 |
| False | 0.39 | 0.75 | 0.85 | 0.96 | 0.99 |

Rysunek: Noisy MAX/OR - Genie 2.0

- Liniowy przyrost parametrów ze względu na ilość węzłów rodzicielskich.
- Bramka Noisy OR jest szczególnym przypadkiem bramki Noisy MAX.

Modele kanoniczne - Noisy OR

- Najprostszy i najbardziej intuicyjny z modeli kanonicznych.

Bramka Noisy-OR wymaga podania prawdopodobieństwa wystąpienia danego zjawiska dla poszczególnych wartości węzłów rodzicielskich, niezależnie od pozostałych:

$$p_k = P(y|\bar{x}_1, \bar{x}_2, \dots, x_k, \dots, \bar{x}_{n-1}, \bar{x}_n). \quad (1)$$

Prawdopodobieństwo w bramce Noisy-OR przy wektorze wejściowym \mathbf{X} wyliczamy następująco:

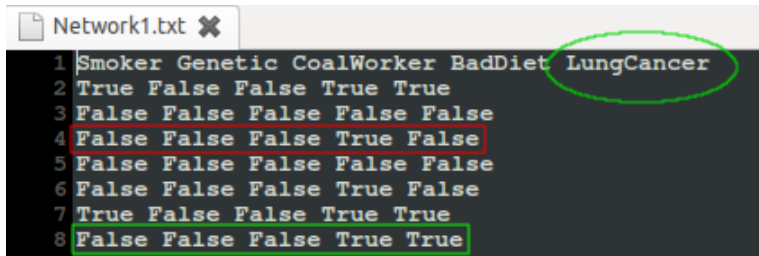
$$P(y|\mathbf{X}) = 1 - \prod_{i: x_i \in \mathbf{X}} (1 - p_i) \quad (2)$$

Parametr “LEAK” oznacza prawdopodobieństwo wystąpienia danego zjawiska, pomimo braku wystąpienia jakiegokolwiek jawnej przyczyny. Służy on do uwzględnienia tzw. niewymodelowanych przypadków:

$$p_{leak} = P(y|\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n). \quad (3)$$

Wyliczanie parametrów z danych - CPT

Standardowy węzeł w sieci bayesowskiej wymaga podania całej tabeli prawdopodobieństw warunkowych. Zliczamy poszczególne wystąpienia danych kombinacji parametrów w pliku z danymi i na ich podstawie wyliczamy prawdopodobieństwo.



```
Network1.txt X
1 Smoker Genetic CoalWorker BadDiet LungCancer
2 True False False True True
3 False False False False False
4 False False False True False
5 False False False False False
6 False False False True False
7 True False False True True
8 False False False True True
```

Rysunek: Przykładowy plik z danymi

$$\frac{27}{27 + 311} = 0.079 \quad (4)$$

Wyliczanie parametrów z danych - Noisy-OR/MAX

- Węzeł typu Noisy-OR/MAX nie wymaga podania prawdopodobieństwa dla każdej możliwej kombinacji parametrów, a jedynie dla prawdopodobieństwa wystąpienia każdego z parametrów z osobna (niezależnie od innych).
- Sposób wyliczania jest identyczny, jednak ze względu na charakter bramki Noisy-OR/MAX potrzebujemy znacznie mniej parametrów.

Usprawnienie wyliczania parametrów z danych - Noisy-OR/MAX

- W podanej wcześniej sieci, ilość rekordów składających się na wyliczenie wszystkich parametrów dla bramki Noisy-OR to około **47%**.
- Można to interpretować w taki sposób: Przy określaniu parametrów dla bramki Noisy-OR, pomijamy **53%** informacji zawartych w pliku z danymi.
- Dla porównania, określenie parametrów (CPT) dla bramki standardowej wykorzystuje cały plik z danymi.

Usprawnienie wyliczania parametrów z danych - Noisy-OR/MAX

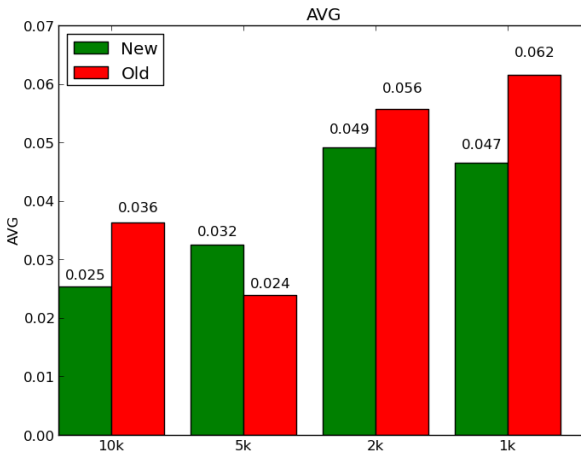
- W podanej wcześniej sieci, ilość rekordów składających się na wyliczenie wszystkich parametrów dla bramki Noisy-OR to około **47%**.
- Można to interpretować w taki sposób: Przy określaniu parametrów dla bramki Noisy-OR, pomijamy **53%** informacji zawartych w pliku z danymi.
- Dla porównania, określenie parametrów (CPT) dla bramki standardowej wykorzystuje cały plik z danymi.
- Czy da się lepiej uzyskać poszczególne parametry sieci Noisy-OR ?

Usprawnienie wyliczania parametrów z danych.

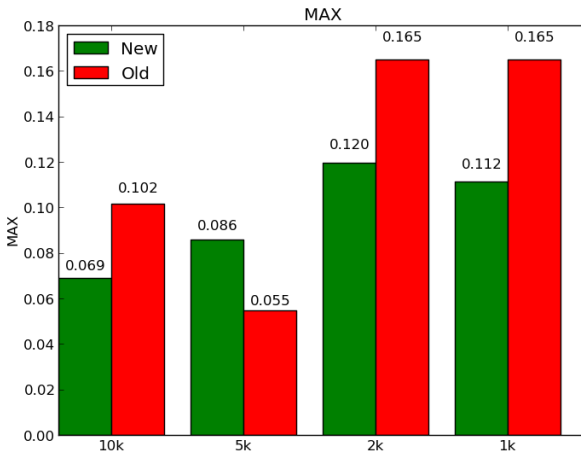
- Układy równań parametrów.

Usprawnienie wyliczania parametrów z danych.

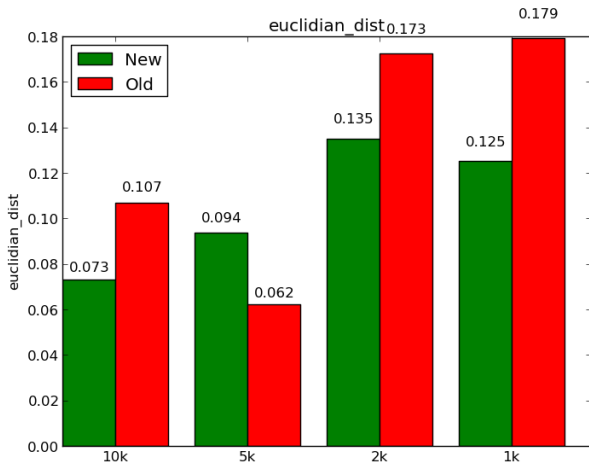
- Układy równań parametrów.
- Wybieramy układ n niewiadomych TODO_i -



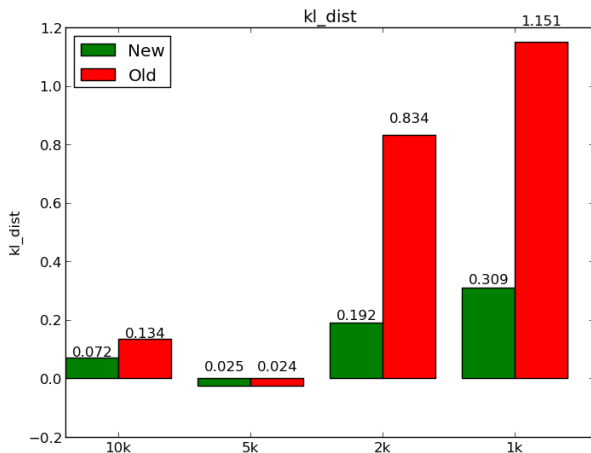
Rysunek: Średni błąd parametru



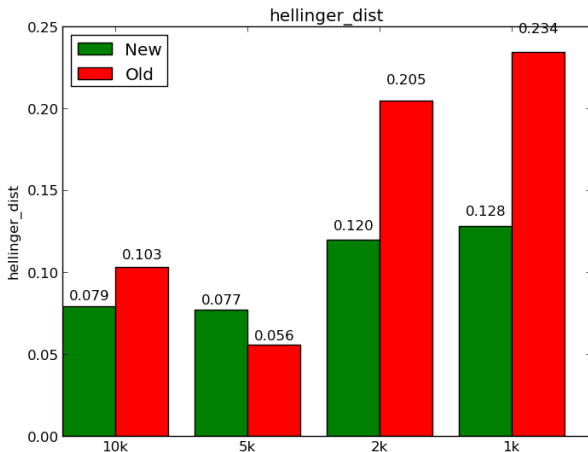
Rysunek: Maksymalny błąd parametru



Rysunek: Odległość euklidesowa wektoru prawdopodobieństw przybliżonych od wzorcowych.



Rysunek: Dywergencja Kullbacka-Leiblera.



Rysunek: Odległość Hellingera.

- Francisco J. Diez, Marek J. Drużdżel - "Canonical Probabilistic Models for Knowledge Engineering" (28.4.2007)
- Nir Friedman, Moises Goldszmidt - "Learning Bayesian networks with local structure"
- Agnieszka Oniśko, Marek J. Drużdżel, Hanna Wasyluk - "Learning Bayesian network parameters from small data sets: application of Nosi-OR gates" (1.3.2001)