ДОДАТОК А

Графічний матеріал атестаційної роботи

Харківський національний університет радіоелектроніки
Кафедра ЕОМ

Методи ймовірнісного висновку
в мережах Байєса

Атестаційна робота
Другий (магістерський) рівень

Автор:
Крікун А.О.
студ. гр. КСМм-19-1

Керівник:
Ільїна І.В.
к.т.н доц. кафедри ЕОМ

Мета і задачі роботи

Мета: дослідження мереж Байєса, методів ймовірнісного висновку в даних мережах, зокрема, огляд на LS-метод, алгоритм його роботи та його практичне застосування.

Задачі:

- аналіз методів ймовірнісного висновку;
- опис алгоритму LS-методу;
- поетапна реалізація методу на прикладі;
- порівняння результатів роботи методу з програмним засобом;
- оцінка ефективності методу з точки зору використання пам’яті.

Що таке мережа Байєса

Мережа Байєса являє собою пару <G,B>, де
G – це направлений ациклічний граф,
P– множина таблиць умовних ймовірностей вершин.

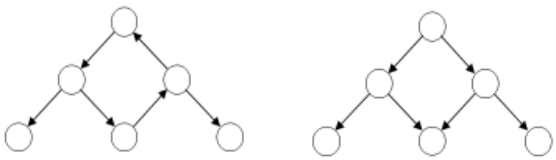
Сама назва мережі Байєса пов'язана з правилом Байєса, яке використовується при побудові ймовірнісного висновку в цій мережі.

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

3

Графове представлення мереж Байєса

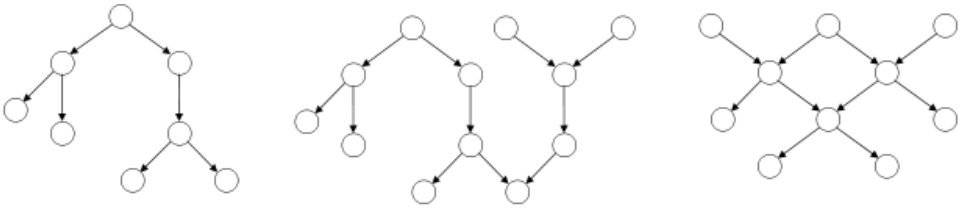
Граф – це сукупність вузлів, що з'єднані між собою дугами. Дуга між двома вершинами свідчить про наявність залежності між ними, а направлена дуга вказує напрям цієї залежності – від причини до наслідку.
В мережі Байєса всі дуги є направленими, тобто вона являється направленим графом.
Направлений граф може бути як циклічним, так і ациклічним, тобто, якщо він не містить направлених циклів.



4

Структури мереж Байєса

За типами структури графів мережі Байєса бувають деревами, однозв'язними мережами та багатозв'язними мережами .



Дерево | Однозв'язна мережа | Багатозв'язна мережа

Дерево – це направлений ациклічний граф, де кожна вершина може мати не більше одного батька.

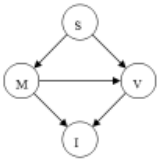
Однозв'язна мережа або – це направлений ациклічний граф, в якому кожна вершина може мати більше одного батька, але існує тільки один шлях між будь-якими двома вершинами .

В багатозв'язних мережах між будь-якими двома вершинами може існувати декілька шляхів.

Види мереж Байєса

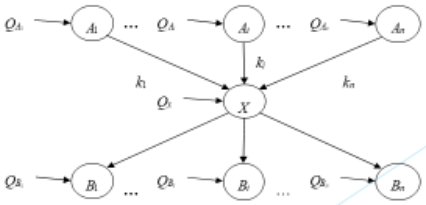
- Дискретні – це мережі, в яких змінні вершин є дискретними, тобто мають скінченну кількість станів.

Для опису ймовірнісного розподілу дискретної випадкової величини використовується ряд розподілу. В мережі Байєса він записується у вигляді таблиць умовних ймовірностей.



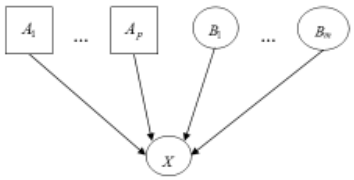
- Неперервні – це мережі, в яких змінні вершин є неперервними.

В багатьох випадках події можуть приймати будь-який стан з деякого допустимого діапазону. Тобто змінна X буде неперервною випадковою величиною, а множиною її можливих станів буде весь діапазон значень, які вона може приймати: X = {x : a≤ x≤b}

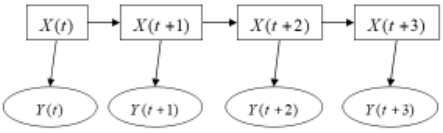


Види мереж Байєса

- Гібридні – це мережі, вершини яких містять як дискретні, так і неперервні змінні.



- Динамічні – це мережі, в яких значення вершин з часом змінюється, тому вони використовуються для моделювання часових процесів. Їх ще також називають «часовими» мережами, бо структура моделі залишається незмінною, хоч і для опису поточного стану процесу можна додавати приховані вузли.



7

Ймовірнісний висновок

Мета ймовірнісного висновку полягає в знаходженні P(X|E) – ймовірності шуканих вершин X, при деякому значенні спостережуваних вершин E. Найпростіший випадок ймовірнісного висновку – коли шукана вершина тільки одна.

За розміром вирішуваних задач виділяється два класи ймовірнісного висновку: точний та апроксимаційний. При вирішенні великих задач застосування точного ймовірнісного висновку стає неможливим через велику обчислювальну складність, і саме тоді застосовуються апроксимаційні методи, які виконують обчислення наближено.

8

Алгоритми точного висновку

- Алгоритм Перла розповсюдження повідомлення для однозв'язних мереж – ідея алгоритму полягає в тому, що нова порція спостережень розглядається як збурення, що розходиться мережею пересиланням повідомлень між поруч розташованими вершинами.
- Алгоритм визначеного перетину – ідея цього алгоритму полягає в зміні структури багатозв'язної мережі на декілька однозв'язних, шляхом інстанціювання вершин, які входять в перетин.
- Алгоритм виключення змінних – основна ідея цього алгоритму полягає в обчисленні ймовірності вершини за формулою, що основана на формулі декомпозиції сукупного розподілу ймовірностей мережі.
- Алгоритми кластеризації – в них використовуються так звані об'єднані дерева, що дозволяє використати ідею обміну повідомленнями ймовірнісного висновку Перла.

9

Апроксимаційні алгоритми

- Алгоритми стохастичної вибірки – вони генерують множину випадково обраних подій або інстанціювань в мережі згідно таблиці умовних ймовірностей моделі і потім апроксимують ймовірності шуканих змінних частотою появи подій у вибірці.
- Алгоритми неповного висновку – алгоритми неповного або часткового висновку ще називають методами спрощення моделі. Їх ідея в спрощенні мережі до такої, де можна застосовувати методи точного висновку.
- Варіаційні алгоритми – основою варіаційних алгоритмів є ідея усереднення значень ймовірностей вершин, тобто при обчисленні розглядаються тільки значущі вершини. Мережа трансформується в підграф початкового графу, в якому деякі вершини позбавляються зв'язків, поки не буде можливо застосувати точний алгоритм ймовірнісного висновку.
- Пошукові алгоритми – основою пошукових алгоритмів (search-based) є ідея переходу від задачі ймовірнісного висновку до оптимізаційної задачі пошуку найбільш ймовірного значення.

10

LS-метод ймовірнісного висновку

Ідея Lauritzen-Spiegelhalter методу ймовірнісного висновку є основоположною ідеєю методів кластеризації, де для реалізації ймовірнісного висновку необхідно спочатку привести структуру мережі Байєса до вигляду об'єднаного дерева, а потім використовувати алгоритм розповсюдження повідомлень по дереву догори та донизу і послідовно перераховувати таблиці умовних ймовірностей вершин дерева.

У загальному вигляді LS-метод передбачає виконання двох етапів. На першому етапі виконується побудова об'єднаного дерева клік з первинної структури мережі та заповнення вершин цього дерева таблицями умовних ймовірностей мережі. На другому етапі обчислюються значення ймовірностей станів вершин на основі алгоритмів розповсюдження значень ймовірності по об'єднаному дереву.

11

Перший етап LS-методу

Перший етап складається з декількох підетапів, які необхідно виконати для формування ймовірнісного висновку:

1. Моралізація графа – послідовний перебір усіх вершин мережі, у яких є батьки;
2. Приведення графу до ненаправленої форми;
3. Триангуляція графа;
4. Побудова дерева суміжності;
5. Побудова об'єднаного дерева;
6. Заповнення об'єднаного дерева таблицями

12

Другий етап LS-методу

Другий етап ще називають алгоритмом пропаганації, який містить в собі наступні кроки:

1. Процес введення спостережень у таблиці;
2. Процес сходження догори;
3. Процес сходження донизу;
4. Розрахунок ймовірностей вершини.

13

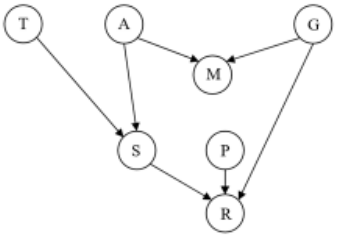
Приклад використання LS-методу

Задача оцінювання кредитоспроможності осіб для отримання кредиту у банку. Вхідні атрибути:

- T – вид контракту (англ., type of contract);
- A – вік (англ., age);
- G – стать (англ., gender);
- M – сімейний стан (англ., marital status);
- S – сума кредиту (англ., sum of credit);
- P – наявність поручителя (англ., personal guarantor);
- R – результат (англ., result).

14

Структура мережі та таблиці умовних ймовірностей вершин



Стан	Ймовірність
T1 – для робочого	0,875
T2 – для пенсіонера	0,061
T3 – для підприємця	0,064

Стан	Ймовірність
A1 – старше 40 років	0,45
A2 – молодше 40 років	0,55

Батьки	Стани батьків					
T	T1	T1	T2	T2	T3	T3
A	A1	A2	A1	A2	A1	A2
Стан	Ймовірність					
S1 – більше 10000 грн	0,55	0,6	0,63	0,8	0,4	0,46
S2 – менше 10000 грн	0,45	0,4	0,37	0,2	0,6	0,54

Таблиці умовних ймовірностей вершин

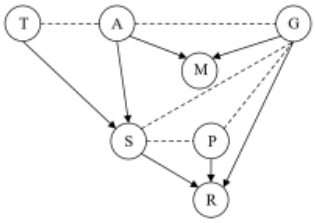
Стан	Ймовірність
P1 – нема поручителя	0,44
P2 – є поручитель	0,56

Стан	Ймовірність
G1 – жінка	0,47
G2 – чоловік	0,53

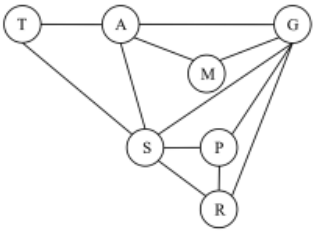
Батьки	Стани батьків			
A	A1	A2	A1	A2
G	G1	G1	G2	G2
Стан	Ймовірність			
M1 – цивільний шлюб	0,0361	0,0841	0,0321	0,0903
M2 – розлучений	0,1996	0,0771	0,0681	0,0819
M3 – одружений	0,5848	0,5455	0,8517	0,1218
M4 – неодружений	0,0792	0,2853	0,0341	0,706
M5 – вдова	0,1003	0,008	0,014	0

Батьки	Стани батьків							
G	G1	G1	G1	G1	G2	G2	G2	G2
P	P1	P1	P2	P2	P1	P1	P2	P2
S	S1	S2	S1	S2	S1	S2	S1	S2
Стан	Ймовірність							
R1 – кредит схвалено	0,96	0,91	0,99	0,99	0,86	0,76	0,985	0,9659
R2 –кредит не схвалено	0,04	0,09	0,01	0,01	0,14	0,24	0,015	0,0341

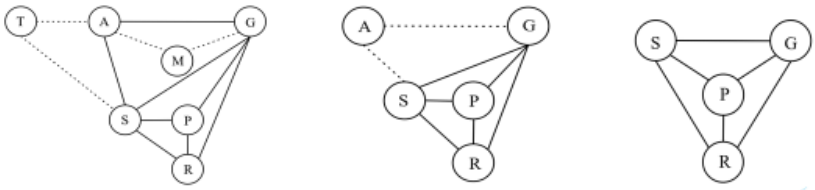
Перший етап LS-методу



Нормалізація графу

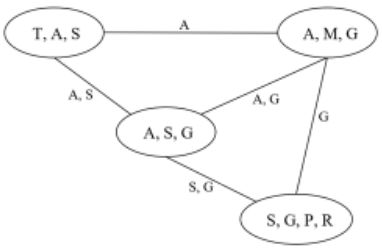


Ненаправлений граф

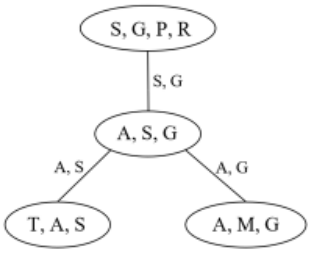


Триангуляція вершин графу

Перший етап LS-методу



Дерево клік



Об’єднане дерево

Заповнення дерева клік таблицями

Лист «T, A, S»

S1	S2	T	A
0,2165	0,1771	T1	A1
0,2887	0,1925	T1	A2
0,0172	0,0101	T2	A1
0,0268	0,0067	T2	A2
0,0115	0,0172	T3	A1
0,0161	0,019	T3	A2

Лист «A, S, G»

S1	S2	A	G
0,47	0,47	A1	G1
0,53	0,53	A1	G2
0,47	0,47	A2	G1
0,53	0,53	A2	G2

Корінь «S, G, R, P»

R1	R2	S	P	G
0,1145	0,0047	S1	P1	G1
0,1157	0,0188	S1	P1	G2
0,1503	0,0015	S1	P2	G1
0,1686	0,0025	S1	P2	G2
0,0796	0,0078	S2	P1	G1
0,0749	0,0236	S2	P1	G2
0,1102	0,0011	S2	P2	G1
0,1212	0,0042	S2	P2	G2

Лист «A, M, G»

M1	M2	M3	M4	M5	A	G
0,0361	0,1996	0,5848	0,0792	0,1003	A1	G1
0,0321	0,0681	0,8517	0,0341	0,014	A1	G2
0,0841	0,0771	0,5455	0,2853	0,008	A2	G1
0,0903	0,0819	0,1218	0,706	0	A2	G2

19

Другий етап LS-методу

Введемо спостереження виду P=P1, G=G2, тобто у клієнта чоловічої статі немає поручителя. Для цього заповнюються нулями входження P=P2 в корені об'єднаного дерева та входження G=G1 в кліці «A, S, G»

R1	R2	S	P	G
0,1145	0,0047	S1	P1	G1
0,1157	0,0188	S1	P1	G2
0	0	S1	P2	G1
0	0	S1	P2	G2
0,0796	0,0078	S2	P1	G1
0,0749	0,0236	S2	P1	G2
0	0	S2	P2	G1
0	0	S2	P2	G2

S1	S2	A	G
0	0	A1	G1
0,53	0,53	A1	G2
0	0	A2	G1
0,53	0,53	A2	G2

20

Процес сходження догори

Маргіналізація по A та S

S1	S2	A
0,2453	0,2046	A1
0,3317	0,2182	A2

Маргіналізація по A та G

	A	G
1	A1	G1
1	A1	G2
1	A2	G1
1	A2	G2

Нова таблиця кліки «T, A, S»

S1	S2	T	A
0,8825	0,8659	T1	A1
0,8703	0,8821	T1	A2
0,0704	0,0496	T2	A1
0,0808	0,0307	T2	A2
0,0469	0,0844	T3	A1
0,0488	0,0871	T3	A2

Нова таблиця кліки «A, M, G»

M1	M2	M3	M4	M5	A	G
0,0361	0,1996	0,5848	0,0792	0,1003	A1	G1
0,0321	0,0681	0,8517	0,0341	0,014	A1	G2
0,0841	0,0771	0,5455	0,2853	0,008	A2	G1
0,0903	0,0819	0,1218	0,706	0	A2	G2

21

Процес сходження догори

Нова таблиця кліки «A, S, G»

S1	S2	A	G
0	0	A1	G1
0,13	0,1084	A1	G2
0	0	A2	G1
0,1758	0,1156	A2	G2

Маргіналізація по S та G

S1	S2	G
0	0	G1
0,3058	0,2241	G2

Нова таблиця кліки «A, S, G»

S1	S2	A	G
0	0	A1	G1
0,4251	0,4839	A1	G2
0	0	A2	G1
0,5748	0,516	A2	G2

Нова таблиця кореня «S, G, R, P»

R1	R2	S	P	G
0	0	S1	P1	G1
0,0353	0,0057	S1	P1	G2
0	0	S1	P2	G1
0	0	S1	P2	G2
0	0	S2	P1	G1
0,0168	0,0053	S2	P1	G2
0	0	S2	P2	G1
0	0	S2	P2	G2

22

Процес сходження донизу

Маргіналізація по S та G

S1	S2	G
0,1193	0,0874	G1
0,1345	0,0986	G2

Маргіналізація по A та G

	A	G
0	A1	G1
0,1049	A1	G2
0	A2	G1
0,1282	A2	G2

Нова таблиця кліки «A, S, G»

S1	S2	A	G
0	0	A1	G1
0,0572	0,40477	A1	G2
0	0	A2	G1
0,0773	0,0509	A2	G2

Нова таблиця кліки «A, M, G»

M1	M2	M3	M4	M5	A	G
0	0	0	0	0	A1	G1
0,0033	0,0071	0,0893	0,0035	0,0014	A1	G2
0	0	0	0	0	A2	G1
0,0115	0,0105	0,0156	0,0905	0	A2	G2

Процес сходження донизу

Маргіналізація по A та S

S1	S2	A
0,0572	0,0477	A1
0,0773	0,0509	A2

Нова таблиця кліки «T, A, S»

S1	S2	T	A
0,0504	0,0413	T1	A1
0,0673	0,0449	T1	A2
0,004	0,0023	T2	A1
0,0062	0,0015	T2	A2
0,0026	0,004	T3	A1
0,0037	0,0044	T3	A2

Нормалізація таблиць

Нормалізована таблиця кліки «T, A, S»

S1	S2	T	A
0,2252	0,1126	T1	A1
0,4061	0,1305	T1	A2
0,0179	0,0064	T2	A1
0,0377	0,0045	T2	A2
0,0119	0,0109	T3	A1
0,0227	0,0128	T3	A2

Нормалізована таблиця кліки «A, S, G»

S1	S2	A	G
0	0	A1	G1
0,2765	0,169	A1	G2
0	0	A2	G1
0,3739	0,1803	A2	G2

Нормалізована таблиця кліки «A, M, G»

M1	M2	M3	M4	M5	A	G
0	0	0	0	0	A1	G1
0,0143	0,0303	0,3795	0,0151	0,0062	A1	G2
0	0	0	0	0	A2	G1
0,05	0,0453	0,0675	0,3913	0	A2	G2

Нова таблиця кореня «S, G, R, P»

R1	R2	S	P	G
0	0	S1	P1	G1
0,5594	0,091	S1	P1	G2
0	0	S1	P2	G1
0	0	S1	P2	G2
0	0	S2	P1	G1
0,2655	0,0838	S2	P1	G2
0	0	S2	P2	G1
0	0	S2	P2	G2

25

Розрахунок ймовірностей стану вершин

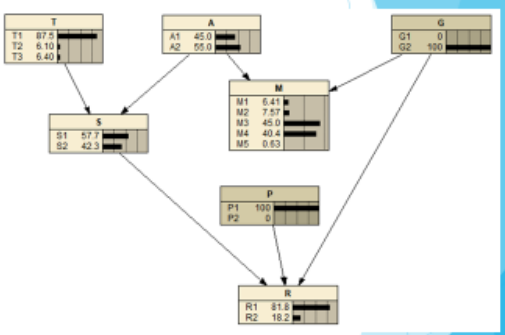
Для вершини T сумуємо значення з кліки «T, A, S», для вершини A – з кліки «A, S, G», для вершини M – з кліки «A, M, G», для вершини S – з кліки «A, S, G» та для вершини R – з кореня дерева.

P(T1) = 0,2252 + 0,1126 + 0,4061 + 0,1305 = 0.8744;
P(T2) = 0,0179 + 0,0064 + 0,0377 + 0,0045 = 0,0665;
P(T3) = 0,0119 + 0,0109 + 0,0227 + 0,0128 = 0,0583;
P(A1) = 0,2765+ 0,169 = 0,4455;
P(A2) = 0,3739 + 0,1803 = 0,5542;
P(M1) = 0,0143 + 0,05 = 0,0643;
P(M2) = 0,0303 + 0,0453 = 0,0756;
P(M3) = 0,3795 + 0,0675 = 0,447;
P(M4) = 0,0151 + 0,3913 = 0,4064;
P(M5) = 0,0062;
P(S1) = 0,2765 + 0,3739 = 0,6504;
P(S2) = 0,169 + 0,1803 = 0,3493;
P(R1) = 0,5594 + 0,2655 =0,8249;
P(R2) = 0,091 + 0,0838 = 0,1748

26

Порівняння результатів роботи

Стан	Ймовірність	
	LS-метод	Програма Netica
T1	0,8744	0,875
T2	0,0665	0,061
T3	0,0583	0,064
A1	0,4455	0,45
A2	0,5542	0,55
G1	0	0
G2	1	1
M1	0,0643	0,0641
M2	0,0756	0,0757
M3	0,447	0,45
M4	0,4064	0,404
M5	0,0062	0,0063
S1	0,6504	0,5770
S2	0,3493	0,423
P1	1	1
P2	0	0
R1	0,8249	0,818
R2	0,1748	0,182



Розрахунок об’єму пам’яті

Для всіх трьох архітектур необхідне виділення однакового об’єму пам’яті:

- для вхідних таблиць: 3+2+2+2+20+16 = 57;
- для спостереження: 2+2 = 4;
- для вихідних таблиць: 3+2+2+5+2+2+2 = 18.

В архітектурі Lauritzen-Spiegelhalter потрібно зберігати кліки, для яких потрібно ще виділити пам’ять: 16+8+12+20 = 56.

В архітектурі Hugin також необхідно виділяти пам’ять на кліки, а ще на сепаратори: 16+8+12+20+4+4+4 = 68.

В архітектурі Shenoy-Shafer використовується дуже багато сепараторів, що значно підвищує кількість виділяїмої пам’яті, тому розрахунок її значення не є доцільним.

Отже, підсумовуючи, отримуємо наступні значення кількості пам’яті:

- Lauritzen-Spiegelhalter: 57+4+18+56 = 135;
- Hugin: 57+4+18+68 = 147;
- Shenoy-Shafer – не розраховується.

Висновки

- Розглянуто базові поняття мереж Байєса, як ймовірнісних експертних систем, в яких база знань представляється топологією мережі і таблицею умовних ймовірностей кожної вершини; наведена їх структура та описані типи існуючих мереж;
- Розглянута задача побудови ймовірнісного висновку, виконано аналіз існуючих алгоритмів ймовірнісного висновку;
- Розглянуто LS-метод ймовірнісного висновку, наведено поетапну його реалізацію на прикладі задачі оцінки кретитоспроможності населення;
- Виконано моделювання ситуації в даній задачі, результати роботи методу було порівняно з програмним засобом Netica, який в собі використовує метод виключення зв'язних дерев;
- Зроблено порівняння LS-методу та інших архітектур з точки зору використання пам'яті, після якого можна сказати, що архітектура Lauritzen-Spiegelhalter є більш ефективною за необхідними об'ємами пам'яті, ніж інші представлені архітектури.