УДК 004.891.3
Гоц О. П. – ст. гр. ІП-51м, ФІОТ
НТУУ "Київський політехнічний інститут ім. Ігоря Сікорського"

ДІАГНОСТИЧНА МЕДИЧНА СИСТЕМА З ІНТЕЛЕКТУАЛЬНОЮ СКЛАДОВОЮ НА ОСНОВІ БАЙЄСОВИХ МЕРЕЖ ТОЧНОГО ВИСНОВКУ

Науковий керівник: к.т.н. Селін Ю. М.

Gots O. P.
NTUU"Igor Sikorsky Kyiv Politechnic Institute"

DIAGNOSTIC MEDICAL SYSTEM WITH INTELLECTUAL WAREHOUSE ON THE BASIS OF BAYESIAN NETWORKS OF EXACT INFERENCE

Supervisor: Ph.D.Selin Yu. M.

Ключові слова: Байєсові мережі, точний висновок, метод кластеризації
Keywords:Bayesian networks, exact inference, clusterization

Вступ

Розробка діагностичного медичного забезпечення є актуальною сьогодні, адже сучасне життя стає небезпечним через появу нових вірусів та хвороб, через недбалість лікарів та неправильні діагнози. Тому розробка інтелектуальної медичної системи конче необхідна.

У роботі запропоновано підходи до створення інтелектуальної інформаційної системи діагнозу хвороб на основі БМз побудовою точного ймовірнісного висновку.

Система відрізняється тим, що дозволяє швидко адаптувати методи діагностики до появи нових показників, переналаштувати систему після нових симптомів і хвороб.

Розроблено методику дослідження статистичних показників, що характеризують реальний медичний стан пацієнтів за допомогою точного ймовірнісного висновку в БМ.

БМ як інструмент інтелектуального аналізу даних

БМефективні в інформаційних системах обробки кількісних даних, представлених часовими рядами і часовими перерізами, а також якісними даними, представленими експертними оцінками, лінгвістичними змінними, інтервальними значеннями і т. д.

Використовуються БМ зазвичай в системах класифікації даних різної природи, системах прогнозування, системах автоматичного розпізнавання мовних сигналів, маркетингу і бізнесі.

МБ можна представити у вигляді направленого ациклічного графу, вершинами якого є набір таблиць умовних ймовірностей (ТУЙ).

Змінні, що використовуються в МБ, можуть бути як дискретними, так і неперервними, а характер їх надходження при аналізі та прийнятті рішення може бути і в режимі реального часу, і у вигляді статистичних масивів інформації та баз даних.

Завдяки представленню взаємодії між факторами процесу у вигляді причино-наслідкових зв'язків в мережі досягаються максимально високий рівень візуалізації та, як наслідок, чітке розуміння суті взаємодії факторів процесу між собою. Саме це відрізняє МБ від інших методів інтелектуального аналізу даних (ІАД) [1].

Формула Байєса дозволяє «переставити причину і наслідок»: за відомим фактом події обчислити вірогідність того, що вона була викликана даною причиною.

Нехай подія *A* може відбутись тільки разом з однією із попарно несумісних подій $H_1,H_2,..,H_n$, які називаються гіпотезами і утворюють групу:

$$\sum_{i=1}^{n} P(H_i) = 1$$

Тоді, якщо відбулась подія *A*, то це означає, що відбулась одна із попарно несумісних подій $AH_1,AH_2,..,AH_n$.

Це означає:$A=H_1*A+H_2*A+..H_n*A$.

Використавши теорему додавання, одержимо:

$$P(A) = P(H_1 * A + H_2 * A + .. + H_n * A) = P(H_1 * A) + P(H_2 * A) + .. + P(H_n * A)$$

З теореми множення ймовірностей:

$$P(H_i) = P(H_i) * P_{H_i}(A),$$

де *i =1, 2, 3, …, n.*

$$P(A) = P(H_1) * P_{H_1}(A) + P(H_2) * P_{H_2}(A) + .. + P(H_n) * P_{H_n}(A). \quad (1.1)$$

Одержана формула (1.1) називається *формулою повної імовірності.*

Після цього нас цікавить питання про те, як зміняться ймовірності гіпотез H_i, де i=1, 2…, n, якщо подія *A* відбулась. Тобто, як обчислити $P_A(H_i)$. Справедливі рівності:

$$P(H_iA) = P(A) * P_A(H_i) = P(H_i) * P_{H_i}(A), \text{ звідки}$$

$$P_{H_i}(A) = \frac{P(H_i)}{P(A)} \quad (1.2)$$

Ця формула (1.2) називається *формулою Байєса*[2].

Ймовірнісний висновок в БМ

Існує два основних типи знаходження висновків в мережах Байєса:

- ймовірнісний висновок (probabilistic inference або belief updating)
- максимальне апостеріорне пояснення (Maximum a Posteriori - MAP explanation або belief revision).

Метою ймовірнісного висновку є знаходження *P(X|E)* - апостеріорної ймовірності шуканих вершин *X* , при деякому значенні спостережуваних вершин *E*.

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

Проте, коли задача структурована та важлива точність імовірнісного висновку, алгоритмиточного ймовірнісного висновку будуть доцільними.Важливим етапом при використанні алгоритмів точного висновку є правильна побудова моделі, при великій кількості вузлів, таку модель потрібно сегментувати чи розбити на кілька під моделей, які працюватимуть незалежно одна від одної. Лише за таких умов задача вирішуватиметься ефективно.

Основоположний алгоритм побудови точного ймовірнісного висновку в мережах Байєса – алгоритм передачі повідомлення між вузлами мережі (алгоритм Перла). З часом з'явилися алгоритми, побудовані на основі ідеї алгоритму Перла, які є більш ефективними і можуть бути використаними для обчислення точного ймовірнісного висновку у значно складніших системах. До них належать:

- алгоритмcutsetcondition(визначеного перетину);
- алгоритм variable elimination (виключення змінних);
- алгоритм bucketelimination (поглинаючого виключення);
- алгоритм clusterization (кластеризації).

До найбільш ефективних алгоритмів апроксимаційного висновку належать:

- алгоритм stochasticsampling (стохастичної вибірки);
- алгоритм model simplification (спрощення моделі);
- алгоритм search-based (пошукові);

- варіаційні алгоритми.

Алгоритм кластеризації

Джуді Перл був першим, хто запропонував побудову точного ймовірнісного висновку у мережах Байєса, що базується на основі ідеї обміну повідомленнями між вершинами-батьками і вершинами-нащадками у направлених ациклічних графах для обчислення значення їх ймовірностей. Ним був розроблений алгоритм передачі повідомлень між вершинами в МБ.

Ключовою особливістю є те, що мережа повинна бути однозв'язною, тобто представлятися у вигляді ациклічного направленого графу, у якому між двома будь-якими вершинами існує лише один шлях. Однозв'язні мережі також називають полідеревами.

З часом було запропоновано алгоритм кластеризації (LS – алгоритм, від винахідників Lauritzen та Spiegelhalter)

Алгоритм оперує об'єднаними деревами, кожна вершина якого містить деякий набір змінних і ТУЙ, це дозволяє використовувати ідею обміну повідомленнями ймовірнісного висновку на основі ідеї Перла.

Алгоритм побудови об'єднаного дерева представлений блок-схемою на рис. 1.

Рис. 1. Алгоритм побудови об'єднаного дерева в МБ

Об'єднане дерево формується з доменного графа МБ - так називається граф, вершинами якого є вузли МБ, а ребрами з'єднуються ті вершини, які в мережі були залежні один від одного, тобто імовірності наслідків однієї вершинами залежать від результатів інших(ої) вершин (и). Доменний граф не містить в собі таблиць умовних ймовірностей (ТУЙ) БМ, тому несе в собі інформацію лише про якісні, а не кількісні характеристики залежностей змінних в мережі.

На етапах моралізації і триангуляції в доменний граф додаються додаткові ребра, необхідні для подальшого перетворення в деревовидний граф.

Моралізованим (moral) називається доменний граф, в якому проведені додаткові шляхи (ребра) між кожними вершинами А і В, для яких у БМ знайдеться вершина С, залежна і від А, і від В.

Граф називається тріангульованим (triangulated), якщо в ньому відсутні цикли з чотирьох і більше вершин. Цикл в даному випадку визначається як множина вершин, в якій кожна вершина з'єднана рівно з двома іншими вершинами цієї множини.

Після цього будується «заготовка» для об'єднаного дерева – дерево, що представляє собою деревовидний граф, вузлами якого є об'єднання вершин доменного графа. Далі в дерево об'єднань включаються таблиці умовних ймовірностей з розглянутої БМ - цим завершається процес побудови об'єднаного дерева.

Для реалізації алгоритму кластеризації в БМ зазвичай використовується архітектура *Hugin*.

У архітектурі Hugin маємо об'єднане дерево і відповідні таблиці для кожної кліки (така підмножина вершин, що кожні дві вершини з цієї підмножини поєднанні ребром). Сепаратор кожного ребра дерева буде також містити відповідні таблиці.

Процес пропагації (передача повідомлень) також відбувається в два етапи – сходження догори та донизу. В архітектурі Hugin на етапі сходження догори відправник не ділить свою таблицю на повідомлення, а замість цього записує його в сепаратор. Це економить підрахунки, але й потребує більшого об'єму памяті. На етапі сходження донизу сепаратор ділить нове повідомлення на те, яке він раніше зберігав і саме на це відношення множить свою таблицю отримувач повідомлення. Економія обчислень відбувається завдяки діленню таблиць сепараторів, що мають менший розмір, ніж таблиці клік.

Ефективність алгоритму

Алгоритм кластеризації для побудови точного ймовірнісного висновку в БМ гарантує точність обчислень, нехтуючи обчислювальною складністю. Він є одним із найефективніших в своєму класі.

В свою чергу апроксимаційні алгоритми більш прості в обрахунках, проте результат має певну похибку, в залежності від типу алгоритму. Існують такі апроксимаційні алгоритми, які взагалі не гарантують точність, що заперечує їхньому використанню у медичній галузі.

Використовуючи Java-бібліотеку Jayes [3], було проведено ряд експериментів для порівняння обчислювальної складності алгоритму кластеризації для побудови точного ймовірнісного висновку в БМ і алгоритму стохастичної вибірки для апроксимаційного.

Побудувавши деяку модель БМ, для кожного алгоритму було проведено по 10 випробовувань, при різній кількості вершин графу БМ для знаходження часу навчання системи. На рис. 2. приведені результати експерименту.

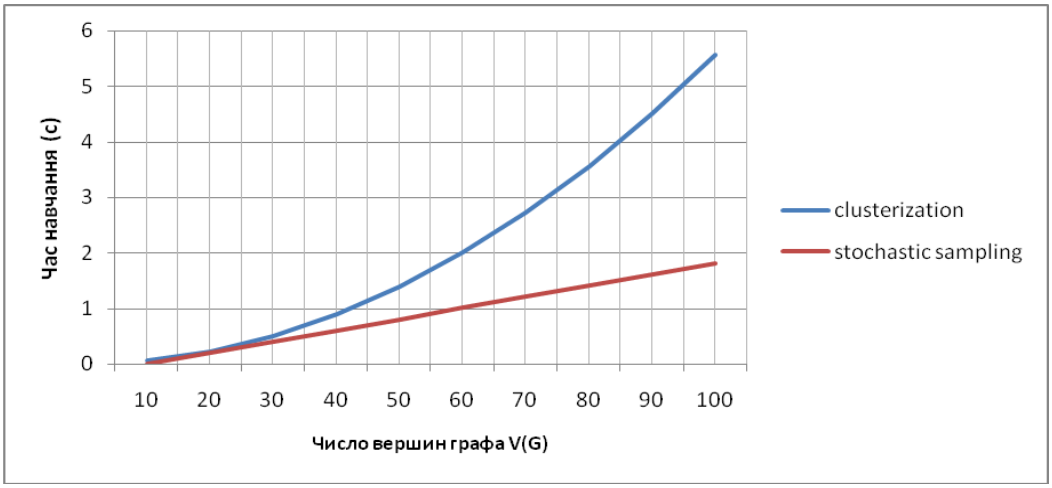

Рис. 2. Графік залежності к-сті вершин графу від часу навчання системи для алгоритмів стохастичної вибірки і кластеризації

Як видно з графіку, алгоритм кластеризації має більшу обчислювальну складність, про що свідчить експоненціальне зростання часу при навчанні, після збільшення кількості вершин графу БМ. В свою чергу, складність апроксимаційного алгоритму має лінійну залежність від кількості вершин графу.

Проте, при кількості вершин до 30, час навчання для обох алгоритмів майже однаковий.

Для моделювання структури медичної інтелектуальної системи БМ можна розбити на підмережі, графи якої матимуть біля відносно невелику кількість вершин. Відповідно до МКХ-10 (міжнародний класифікатор хвороб), можна виділити такі класи, для кожного з яких в подальшому побудувати БМ:

- Деякі інфекційні та паразитарні хвороби;
- Новоутворення;
- Хвороби крові і кровотворних органів та окремі порушення з залученням імунного механізму;
- Хвороби ендокринної системи, розладу харчування та порушення обміну речовин;
- Розлади психіки та поведінки;
- Хвороби нервової системи;
- Хвороби ока та придаткового апарату;
- Хвороби вуха та соскоподібного відростка;
- Хвороби системи кровообігу;
- Хвороби системи дихання;
- Хвороби органів травлення;
- Хвороби шкіри та підшкірної клітковини;
- Хвороби кістково-м'язової системи та сполучної тканини;
- Хвороби сечостатевої системи;
- Вагітність, пологи та післяпологовий період [4].

Таке розбиття допоможе зберегти відносно невелику обчислювальну складність для точного алгоритму кластеризації, яка не сильно відрізнятиметься від апроксимаційних алгоритмів, при цьому зберігатиметься виграш у точності обчислень.

Висновки

БМ чудово підходять для аналізу процесів різної природи, та мають ряд переваг серед інших інтелектуальних методів аналізу даних та прогнозування.

Раціональне використання БМ, їх швидка і надійна робота в першу чергу залежать від моделі та алгоритмупобудови висновку в мережі.

У даній роботі був описаний і проаналізований алгоритм кластеризації, який належить до класу точних алгоритмів. Цей алгоритмпоєднує в собі три переваги: точність результатів, відносно малий час роботи і універсальність.

Експериментальним шляхом проаналізована різниця в обчислювальній складності алгоритму кластеризації з апроксимаційним алгоритмом. З чого зроблені висновки, що БМ для всієї медичної системи потрібно розбити на незалежні БМ, що дозволить зберегти точність та покращити ефективність обчислень.

Сьогодні медична статистика представлена достатньо повно і з необхідним рівнем достовірності, що дозволить побудувати якісну модель системи та використовувати в ній алгоритм кластеризації.

Список літератури

1. Бідюк П. І. Інтелектуальний аналіз слабоструктурованих даних за допомогою байєсових мереж: звіт по результатам виконання робіт за грантом грант НТУУ „КПІ” № 3/5-ГР, 2006-2007р. / П. І. Бідюк, О. М. Терентьєв, Л. О. Коршевнюк. – 2007. – 85 с.
2. Тичинська Л.М. Формула повної ймовірності. Формула БАйєса // Теорія ймовірностей. - 2010. – 112 с.
3. Michael Kutschke. An Introduction to Bayesian Networks with Jayes[Електронний ресурс]. – 2013. Режим доступу: http://www.codetrails.com/blog/introduction-bayesian-networks-jayes/ (останній візит: 30.03.17).
4. МКХ-10 [Електронний ресурс]. – 2016. Режим доступу: https://mkh10.com.ua/ (останній візит: 29.03.17).