Реалізація методики побудови гібридних Байєсівських мереж

Виконала:

студентка групи КА-25

Ніколенко Вікторія Олександрівна

Науковий керівник:

д.т.н., проф. Бідюк П.І.

Вступ

Ідея впровадження БМ полягає у представленні причинно-наслідкових зв’язків процесу у вигляді графа.

За допомогою побудови моделей у формі Байєсівських мереж можна враховувати невизначеності статистичного, структурного і параметричного характерів.

Завдяки коректному представленню взаємозв’язків у вигляді графів, БМ дуже зручні для розв’язання прикладних задач.

Об’єкт дослідження: гібридні Байєсівські мережі.

Предмет дослідження: методи побудови структури і оцінювання параметрів Байєсівської мережі, методи формування висновку.

Мета роботи: реалізація методики побудови гібридних Байєсівських мереж. Створення програмного продукту для побудови гібридних Байєсівських мереж. Порівняння розробленого програмного продукту з вже існуючими комерційним продуктом-аналогом.

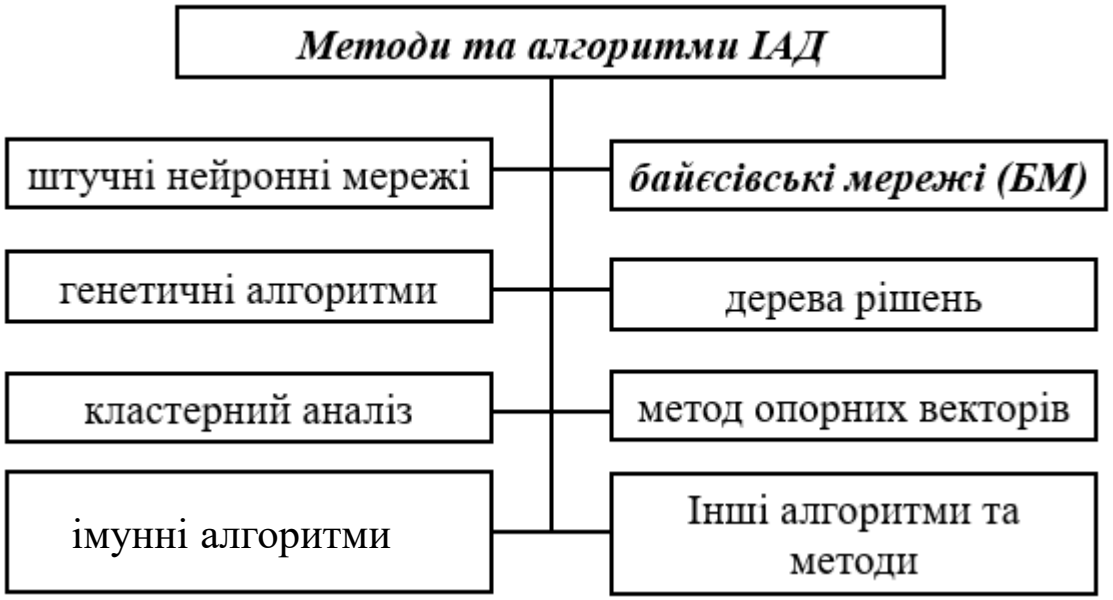
Постановка задачі

1. Аналіз методів побудови БМ за літературними джерелами.
2. Розробити методику побудови гібридної БМ.
3. Розробити програмний продукт для реалізації методики побудови гібридних БМ.
4. Порівняти результати функціонування програми на прикладах побудованих мереж (з GeNie).
5. Формулювання висновків.

Інтелектуальний аналіз даних

ІАД – це дослідження та виявлення «машиною» (алгоритмами, засобами штучного інтелекту) в сирих даних, раніше невідомих, нетривіальних, практично корисних та доступних для інтерпретації людиною знань, необхідних для прийняття рішень у різних сферах людської діяльності.

В рамках технології ІАД використовують різноманітні методи й алгоритми. Досліджувані Байєсівські мережі – один із них.

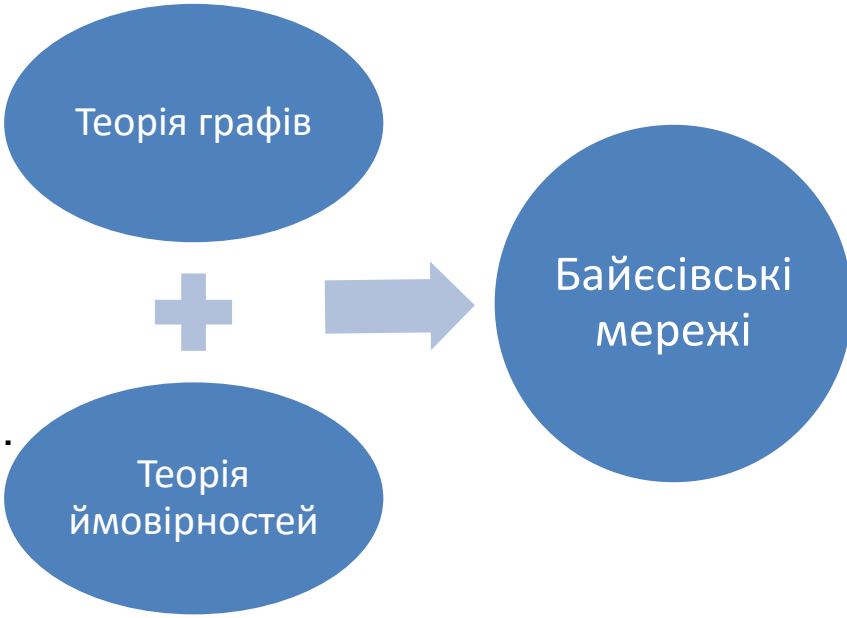


Поняття Байєсівської мережі

Термін “***байєсівська мережа***” (**Bayesian Network**) запропонований **Джуді Перлом** в 1985 році. В основу покладена ідея дослідження причинно-наслідкових зв’язків, яка запропонована в 1763 році **Томасом Байєсом** (Член Королівського Наукового Товариства): “*Опис методу розв'язання задачі в рамках доктрини випадкових подій*”.

Теорія БМ створена на стику двох наук: ***теорії імовірностей*** і ***теорії графів***.

БМ – модель подання існуючих ймовірнісних залежностей між змінними.



Представлення БМ

БМ - це пара компонентів **<*G,B*>,** де:

G – це спрямований нециклічний граф.

B – множина значень умовних ймовірностей мережі для кожної вершини:

$$B = \{P\left(X^{(i)} \middle| pa\left(X^{(i)}\right)\right), \ \forall x^{(i)} \in X^{(i)}\}$$

де $x^{(i)}$ – значення, а $pa(X^{(i)})$ – набір батьків вершини $X^{(i)}$.

Повна спільна ймовірність БМ обчислюється за формулою:

$$P_B(X^{(1)},...,X^{(N)}) = \prod_{i=1}^{N} P_B(X^{(i)} \middle| Pa(X^{(i)}))$$

Кожна змінна (фактор, атрибут) процесу, що досліджується, представляється в мережі Байєса як ***вершина***, а існуючий причинно-наслідковий зв'язок між змінними процесу - у вигляді спрямованих ***дуг***, що з'єднують відповідні вершини.

Типи БМ

- Дискретні
- Неперервні
- Гібридні

Етапи побудови БМ

Етап 1: Методи побудови структури (топології) БМ

На основі оціночних функцій (search & scoring)	Із застосуванням тесту на умовну незалежність (ТУН) (dependency analysis)	Методи для роботи з неповними даними
• Чу і Ліу • Рібана і Перла • Кутато • К2 • та інші. .	• Вермута і Лоуренса • Перла • SRA • “Конструктор” • та інші.	• Алгоритми максимізації математичного сподівання • Алгоритм стиснення границь • та інші.

Етап 1: Методи побудови структури (топології) БМ

1 підетап

Побудова впорядкованої множини вершин (експерт)

Побудова тесту на умовну незалежність (визначення міри зв'язку між вершинами мережі). Значення
- кореляції,
- асоціації,
-спряженості,
- значення взаємної інформації.

2 підетап

Пошук/побудова кращої структури на основі оціночних функцій:
1. Купера-Герсковича
2. Опису мінімальної довжини

Обчислення значення взаємної інформації

Чу і Ліу запропонували ЗВІ (mutual information) в 1968 р. для оцінки ступеня залежності двох довільних змінних.

$$MI(x^i, x^j) = \sum_{x^i, x^j} P(x^i, x^j) \cdot \log\left(\frac{P(x^i, x^j)}{P(x^i) \cdot P(x^j)}\right)$$

$$MI(x^i, x^j) = MI(x^j, x^i)$$

За своєю суттю значення взаємної інформації – це аналог кореляції, але за своїм змістом – це оцінка кількості інформації, яка наявна у змінній x^i про змінну x^j. Значення взаємної інформації є невід'ємні значення, тобто $MI(x^i, x^j) = 0$, а у випадку якщо вершини x^i та x^j повністю незалежні одна від одної, то $MI(x^i, x^j) = 0$, тому що $P(x^i, x^j) = P(x^i) \cdot P(x^j)$

Функція опису мінімальної довжини

Значення функції ОМД довільної структури g при заданій послідовності з n спостережень $x^n = d_1\ d_2\ \ldots\ d_n$ обчислюється за формулою:

$$L(g, x^n) = H(g, x^n) + \frac{k(g)}{2} \cdot \log(n)$$

де $H(g, x^n)$ - значення емпіричної ентропії, а $k(g)$ кількість незалежних умовних імовірностей в структурі g.

$$H(g, x^n) = \sum_{j \in J} H(j, g, x^n) \qquad k(g) = \sum_{j \in J} k(j, g)$$

де ОМД j-ї вершини обчислюється за формулою:

$$L(j, g, x^n) = H(j, g, x^n) + \frac{k(j, g)}{2} \cdot \log(n)$$

а кількість незалежних умовних імовірностей j-ї вершини:

$$k(j, g) = (\alpha^{(j)} - 1) \cdot \prod_{k \in \phi(j)} \alpha^k$$

Функція Купера-Герсковича

Функція КГ структури g при заданій послідовності з n спостережень $x^n = d_1 d_2 ... d_n$:

$$P(g, x^n) = P(g) \cdot \prod_{j \in J} \left(\prod_{s \in S(j,g)} \frac{\left(\alpha^{(j)} - 1\right)! \cdot \prod_{q \in A^{(j)}} \left(n[q, s, j, g]!\right)}{\left(n[s, j, g] + \alpha^{(j)} - 1\right)!} \right)$$

$P(g)$ – апріорна ймовірність структури g, але її часто опускають при обчисленнях, вважаючи що всі структури рівноймовірні; запис $j \in J = \{1, ..., N\}$ означає перебір всіх вершин структури мережі g, а $s \in S(j, g)$ означає перебір множини всіх наборів значень, які приймають батьківські вершини для j-ї вершини.

$$n(s, j, g) = \sum_{i=1}^{n} I(\pi_i^{(j)} = s) \qquad n[q, s, j, g] = \sum_{i=1}^{n} I(x_i = q, \pi_i^{(j)} = s)$$

де $\pi^{(j)} = \Pi^{(j)}$ означає $X^{(k)} = x^{(k)}, \forall k \in \phi^{(j)}$, а функція $I(E)=1$, коли предикат $E=true$, інакше $I(E)=0$.

Алгоритм К2

Вхідні дані: впорядкована множина вершин $\{X_1 < X_2 < \ldots X_n\}$, де вершина X_1 - головна коренева вершина, котра не має предків, а запис $X_k < X_j$ означає, що вершина X_j не може бути предком вершини X_k. Дуги можуть йти тільки з X_k в X_j.

Суть алгоритму: до вершини $X^{(N)}$ по черзі, послідовно перебираючи впорядковану множину вершин, додають предків від $X^{(1)}$ до $X^{(N-1)}$ та обчислюють значення функції Купера-Герсковича (КГ) для кожної побудованої таким чином мережі. В якості батьківської для дочірньої вершини $X^{(N)}$ залишають вершину $X^{(i)}$, за якої функція Купера-Герсковича набуває максимального значення. Після цього до вершин $X^{(N)}$ і $X^{(N-1)}$ по черзі додають батьківські від $X^{(1)}$ до $X^{(N-2)}$, обчислюючи значення функції Купера-Герсковича відповідних мереж. У результаті утворюється структура мережі, для якої функція Купера-Герсковича набуває максимального значення.

Етап 2: Побудова імовірнісного висновку в БМ

Імовірнісний висновок в БМ – процес обчислення стану вершини на основі апріорної ймовірності про стани інших вершин.

Суть імовірнісного висновку полягає у визначенні значень ймовірностей станів вершин на основі побудованої структури БМ і наявних навчальних даних.

Обчислювальна складність імовірнісного висновку насамперед залежить від кількості вершин мережі та кількості дуг які з'єднують ці вершини мережі між собою.

Неоднозначність імовірнісного висновку полягає в тому, що різні методи імовірнісного висновку показують різні результати.

Саме механізм побудови імовірнісного висновку в мережі перетворює будь-яку БМ, яка описує відповідний процес, в **експертну систему**.

Етап 2: Побудова імовірнісного висновку в БМ

Алгоритми точного висновку	Апроксимаційні алгоритми
1. Алгоритм Перла розповсюдження повідомлення для однозв'язних мереж (полі-дерев)	1. Часткового або неповного висновку (exact inference partially)
2. Кластеризації дерева клік (clique tree clustering)	2. Варіаційні методи (variational algorithms) використовуються для обчислення середніх ознак великих мереж
3. Визначаючого перетину (cutest conditioning)	
4. Виключення змінних (variable elimination algorithm)	3. Методи стохастичного вибору (stochastic sampling)
5. Символьного ймовірнісного висновку (symbolic probabilistic inference)	4. Пошукові методи (search-based), основані на евристичних алгоритмах пошуку, які використовують при переході від задачі обчислення ймовірнісного висновку до оптимізаційної задачі
6. Диференціальний підхід (differential method)	

Етап 2: Побудова імовірнісного висновку в БМ

Метод складається з двох кроків.

Крок-1 обчислення матриці емпіричних значень спільного розподілу ймовірностей всієї мережі.

Крок-2 обчислюються значення ймовірностей всіх можливих станів неінстанційованих вершин.

Переваги запропонованого методу

1. швидкість роботи залежить не від кількості дуг мережі, а від розміру навчальної вибірки.
2. простота його застосування та можливість швидкої реалізації на будь-якій мові програмування
3. Відсутність необхідності в попередні перетворення структури БМ
4. Відсутня потреба в наявності ВМВ

ВХІДНІ ДІНІ:

1. Множина навчальних даних

$$D = \{d_1, ..., d_n\} \qquad d_i = \{x_i^{(1)} x_i^{(2)} ... x_i^{(N)}\}$$

де (нижній індекс – номер спостереження, а верхній – номер змінної), *n* – кількість спостережень.

2. Структура БС *g* представлена множиною з *N* батьківських вершин.

3. Множина інстанційованних вершин

$$\{X^{(P_1)} = x^{(P_1)}, ..., X^{(P_v)} = x^{(P_v)}\}$$

тобто вершин, що перебувають у деякому певному стані з одиничною ймовірністю. Якщо множина інстанційованих вершин порожня, то потрібно використати імовірнісний висновок, оснований на класичній теоремі Байєса.

Крок 1. По множині навчальних даних обчислюється матриця емпіричних значень спільного розподілу ймовірностей всієї мережі $P(X^{(1)},...,X^{(N)})$, за формулою:

$$P_{matrix}(X^{(1)}=x^{(1)},...,X^{(N)}=x^{(N)})=\frac{n[X^{(1)}=x^{(1)},...,X^{(N)}=x^{(N)}]}{n}$$

де n – кількість навчальних спостережень.

Крок 2. Перебираємо послідовно всі вершини БМ. Якщо вершина не інстанційована, то потрібно обчислити значення ймовірностей всіх можливих станів цієї вершини. Для цього робиться послідовний перебір всіх рядків матриці емпіричних значень спільного розподілу ймовірностей всієї мережі. Якщо значення вершин рядка збігаються зі значеннями інстанційованих вершин і станом аналізованої вершини, то відповідне значення $P_{matrix}(X^{(1)},...,X^{(N)})$ додається до значення ймовірності відповідного стану аналізованої вершини. Після цього нормуються значення ймовірностей станів аналізованої вершини.

Вихідні дані.

Вихідними даними являються значення ймовірностей всіх можливих станів всіх неінстанційованих вершин.

Особливості побудови гібридних мереж Байєса

Гібридна мережа Байєса B = (X , D, P)
визначається через спрямований ациклічний граф G = (X , E)
і його функції $P_i = \{P\ (x_i|pa_i)\}$,
де pa_i — набір батьківських вузлів x_i . X — набір змінних, розділених на дискретні і неперервні.

Тобто, гібридні БМ – це мережі, котрі містять вузли з дискретними і неперервними змінними.

Аглоритм Greedy Thick Thinning заснований на методі К2, який грунтується на максимізації функції правдоподібності.

Особливості побудови гібридних мереж Байєса

Одним із підходів до формування висновку у гібридній мережі Байєса може бути попередня дискретизація неперервних змінних (тобто зведення гібридної мережі до дискретної) і використання відомих алгоритмів формування висновку для простих МБ.

Критерії оцінювання якості моделі: обчислення структурної різниці

Для обчислення структурної різниці використовують формулу симетричної різниці структур

$$\delta = \sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} card\left(\Pi^{(i)}(B) \Delta \Pi^{(i)}(A)\right) = \sum_{i=1}^{n} card\left(\left(\Pi^{(i)}(B) \setminus \Pi^{(i)}(A)\right) \cup \left(\Pi^{(i)}(A) \setminus \Pi^{(i)}(B)\right)\right),$$

де B – побудована мережа Байєса; A– оригінальна мережа Байєса; n– кількість вершин мережі; $\Pi^{(i)}(B)$ – множина предків i–ї вершини побудованої мережі B; $\Pi^{(i)}(A)$ – множина предків i–ї вершини оригінальної мережі A; $card(\xi)$ – потужність скінченої множини ξ, що визначається кількістю елементів, які належать множині ξ.

Критерії оцінювання якості моделі: перехресна ентропія

Перехресна ентропія – це відстань між розподілом побудованої і оригінальної мережі Байєса. Нехай $p(v)$ – спільний розподіл оригінальної мережі Байєса, а $q(v)$ – спільний розподіл побудованої мережі Байєса.

Тоді перехресна ентропія обчислюється так

$$H(p,q) = \sum_{v} p(v) \cdot \log \frac{p(v)}{q(v)} =$$

$$= \sum_{j \in J} \sum_{s \in S(j,g)} \sum_{a \in A^{(j)}} p(X^{(j)} = a \mid \Pi^{(j)} = s) \cdot \log \frac{p(X^{(j)} = a \mid \Pi^{(j)} = s)}{q(X^{(j)} = a \mid \Pi^{(j)} = s)} .$$

Приклад побудови гібридних Байєсівських мереж

ВВП Китаю та США обчислюється:

$$ВВП = \sum ВДВ + \sum ПДВ,$$

де ВДВ – валова додана вартість, ПДВ – податок на додану вартість.

Модель економіки США вважається ідеальною для досліджень, бо вона найкоректніше відображає структуру ВВП.

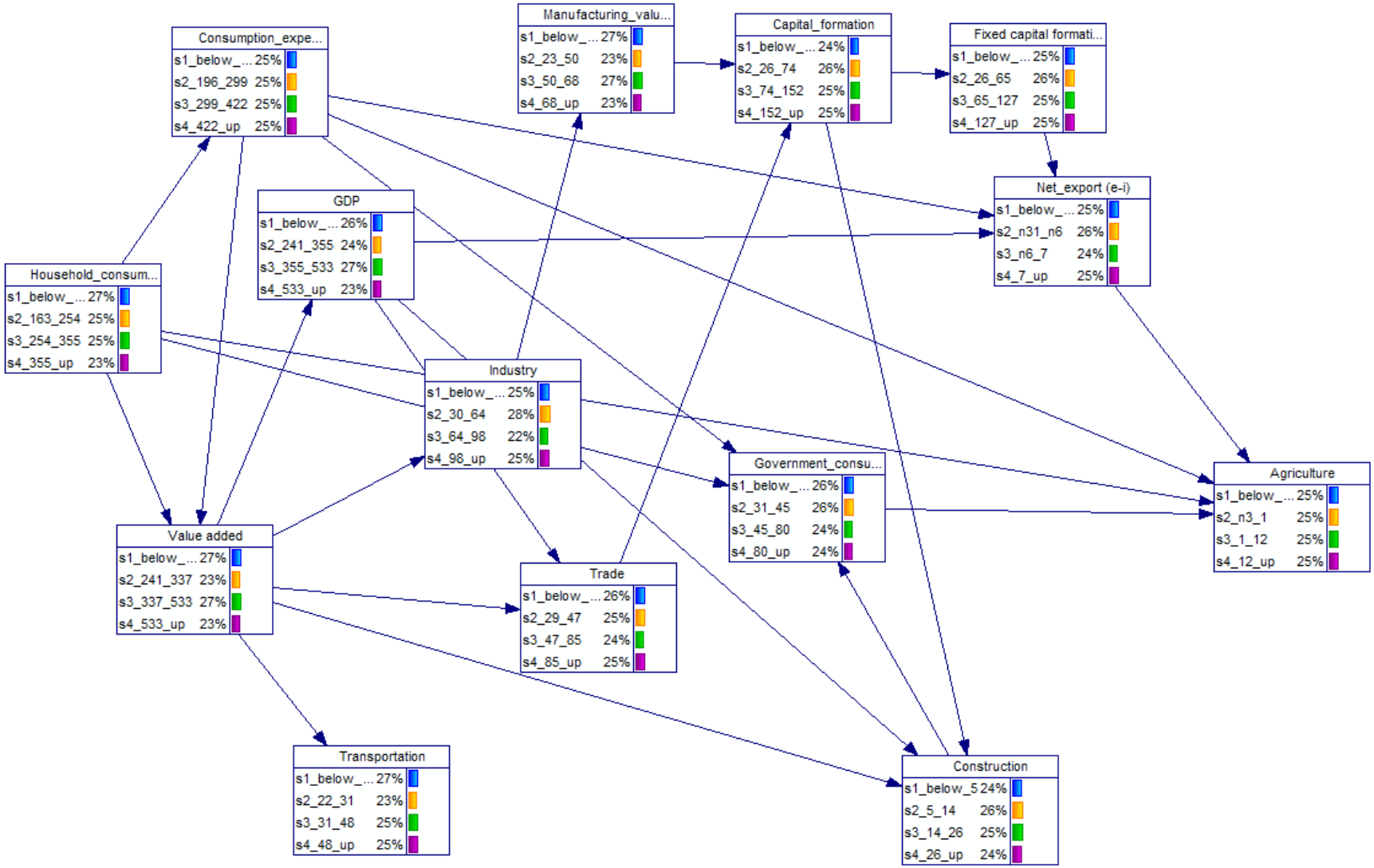
Економіка Китаю довгий час відносилася до економіки перехідного періоду, а також світовим економічним дивом.

Відповідно до теорії ВВП для раціонального аналізу необхідно формалізувати дані за арифметичними різницями (для розрахунку впливу факторів на відхилення результативного показника).

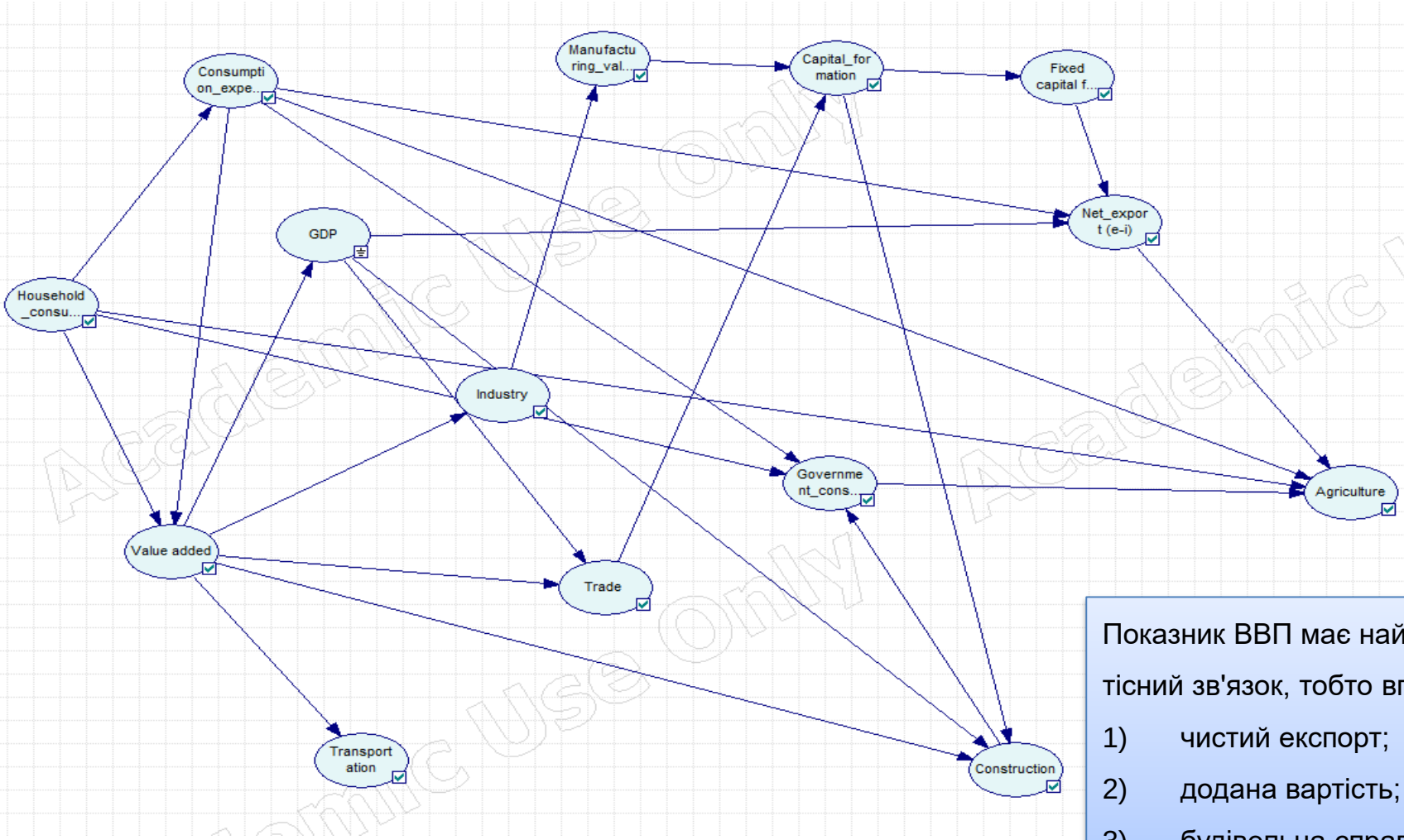
При розрахунку ВВП враховують наступні показники

1) чистий експорт (Net_export);
2) споживацькі витрати (Consumption_expenditure);
3) витрати домогосподарств (Household_consumption_expenditure);
4) урядові витрати (Government_consumption_expenditure);
5) внутрішні інвестиції (Capital_formation);
6) капітальні внутрішні інвестиції(Fixed_capital_formation);
7) додана вартість (Value_added);
8) індекс сільського господарства (Agriculture);
9) промисловість (Industry);
10) переробна промисловість (Manufacturing_value_added);
11) будівельна справа (Construction);
12) торгівля (Trade);
13) транспорт (Transportation).

ВВП США (програмний продукт)



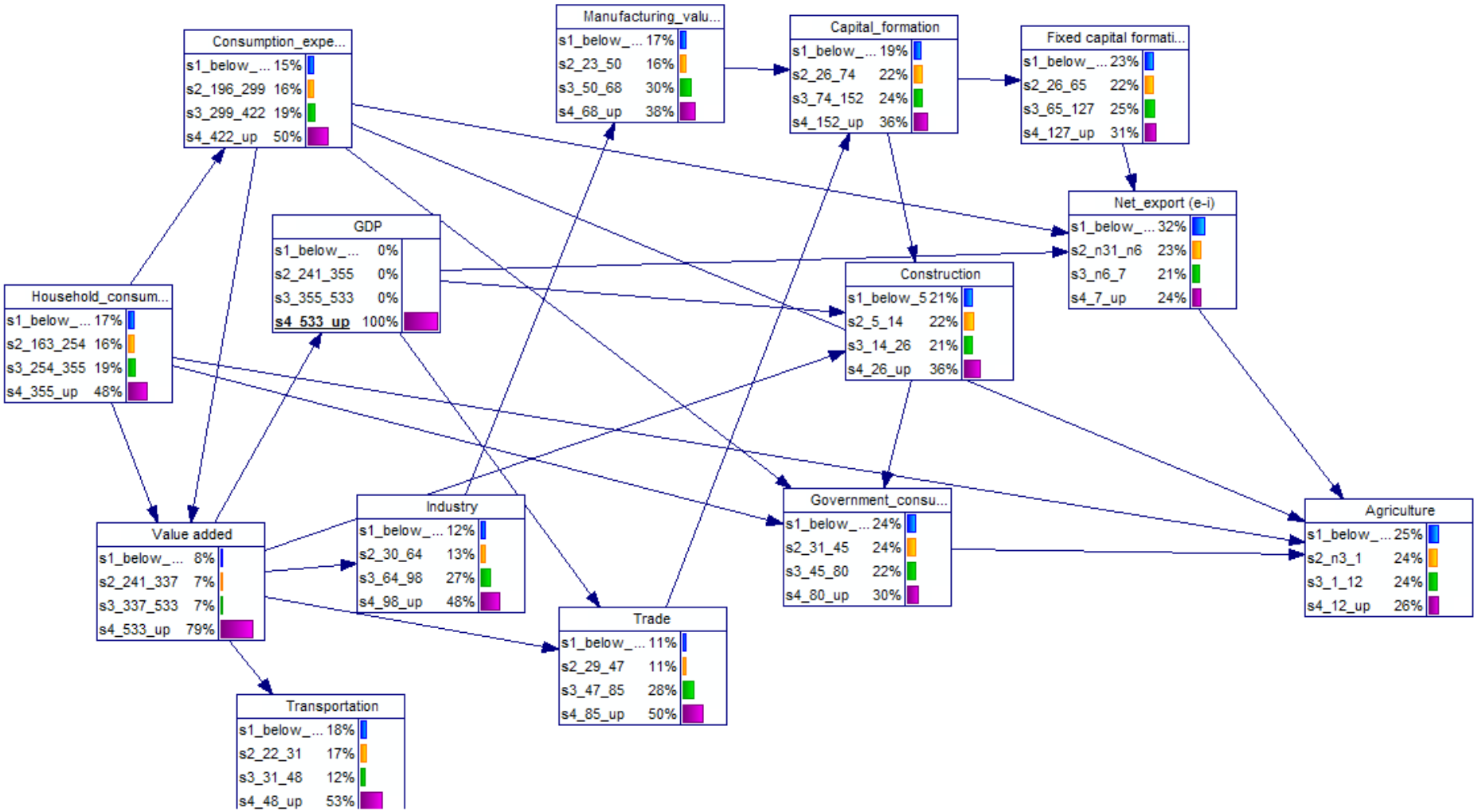
ВВП США (у GeNIe)



Показник ВВП має найбільш тісний зв'язок, тобто вплив:

1) чистий експорт;
2) додана вартість;
3) будівельна справа;
4) торгівля.

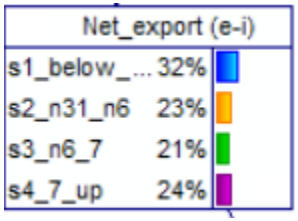
ВВП США (оновлена мережа)



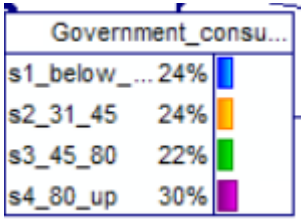
Перевірка моделі ВВП США

	2014	Інтервали Ймовірність	Результат
GDP	684,9	s4_533_up 100%	+
Net_export (e-i)	-21,60	s1_below_n31 32%	-
Consumption_expenditure	507,90	s4_422_up 50%	+
Household_consumption_expenditure	473,60	s4_355_up 48%	+
Government_consumption_expenditure	34,30	s4_80_up 30%	-
Capital_formation	198,60	s4_152_up 36%	+
Fixed_capital_formation	128,5	s4_127_up 31%	+
Value_added	685,20	s4_533_up 79%	+
Agriculture	-10	s4_12_up 26%	-
Industry	98,2	s4_98_up 48%	+
Manufacturing_value_added	73	s4_68_up 38%	+
Construction	44,10	s4_26_up 36%	+
Trade	99,10	s4_85_up 50%	+
Transportation	73,50	s4_48_up 53%	+

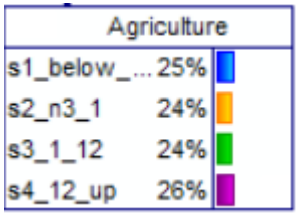
$$CA = \frac{CorrectForecast}{N} = \frac{10}{13} = 0{,}77$$



32%-23%=9%

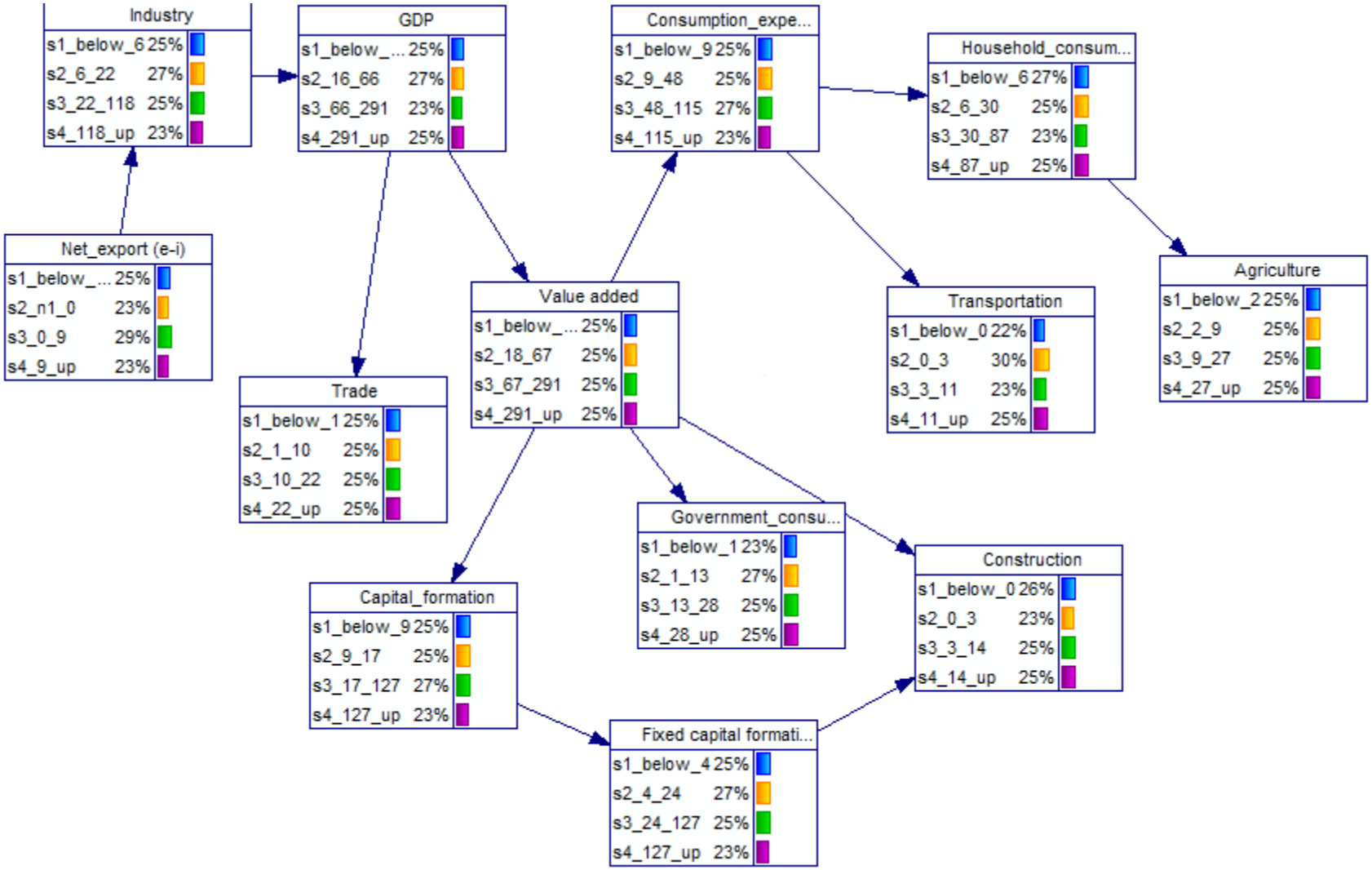


30%-24%=6%

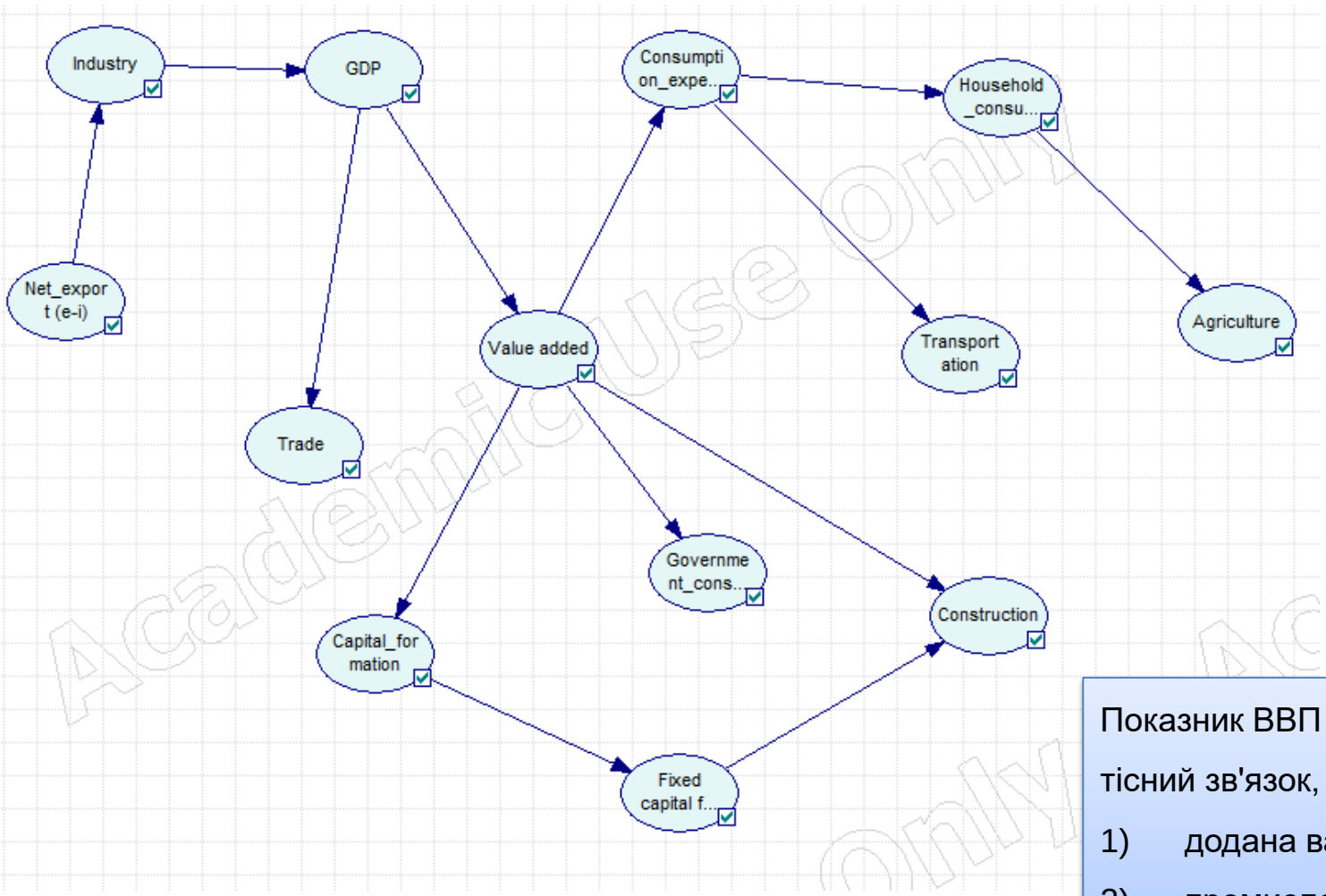


26%-25%=1%

ВВП Китаю (програмний продукт)



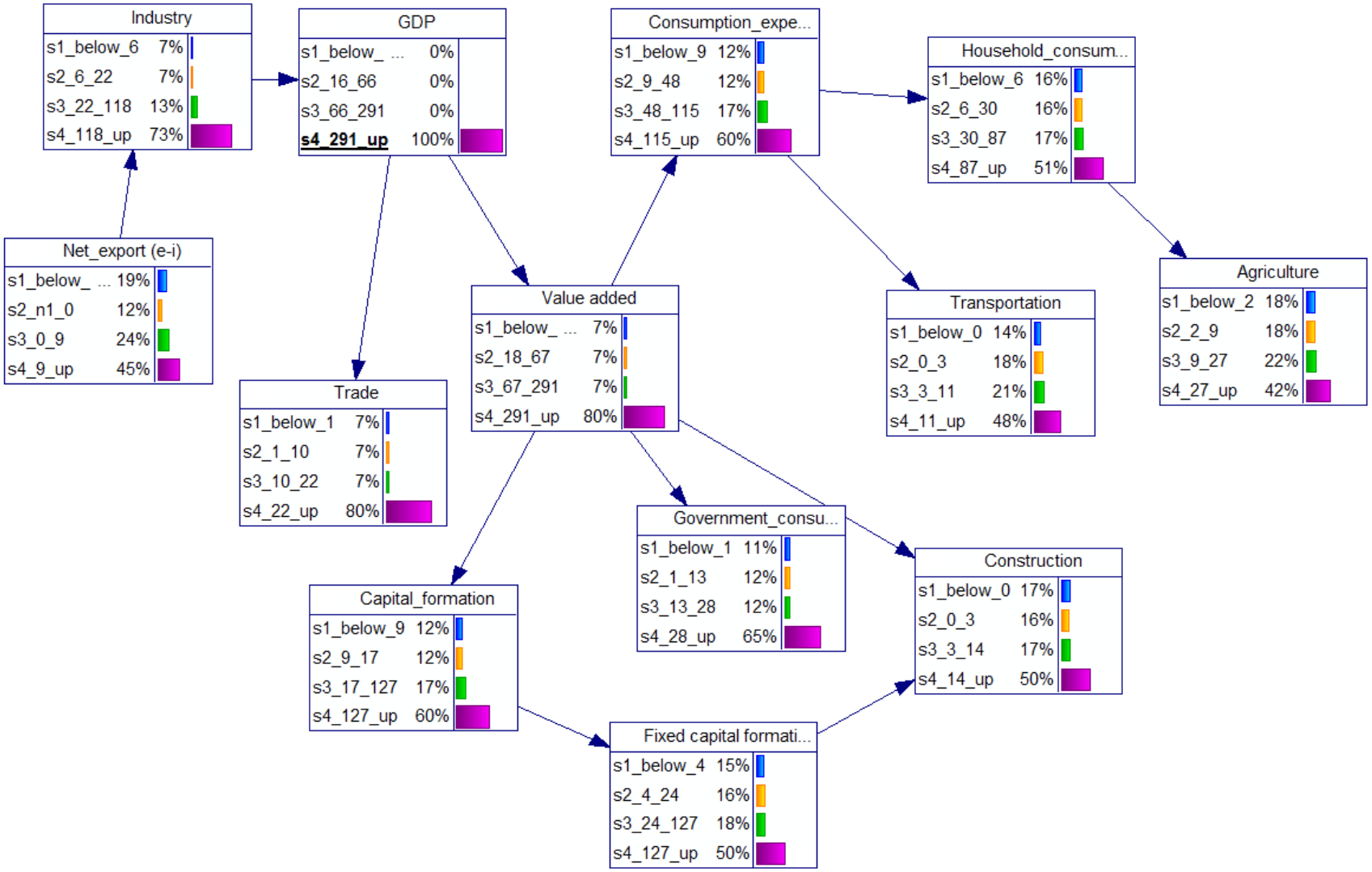
ВВП Китаю (у GeNIe)



Показник ВВП має найбільш тісний зв'язок, тобто вплив:

1) додана вартість;
2) промисловість;
3) торгівля.

ВВП Китаю (оновлена мережа)



Перевірка моделі ВВП Китаю

	2014	Інтервали Ймовірність	Результат
GDP	912,2	s4_291_up 100%	+
Net_export (e-i)	49,5	s4_9_up 45%	+
Consumption_expenditure	504,3	s4_115_up 60%	+
Household_consumption_expenditure	407,3	s4_87_up 51%	+
Government_consumption_expenditure	97,1	s4_28_up 65%	+
Capital_formation	356,9	s4_127_up 60%	+
Fixed_capital_formation	339,1	s4_127_up 50%	+
Value_added	861,9	s4_291_up 80%	+
Agriculture	59,8	s4_27_up 42%	+
Industry	204,4	s4_118_up 73%	+
Construction	70,5	s4_14_up 50%	+
Trade	121,5	s4_22_up 80%	+
Transportation	47,8	s4_11_up 48%	+

$$CA = \frac{CorrectForecast}{N} = \frac{13}{13} = 1$$

Висновки

1. Виконано аналіз поточного стану розвитку методів інтелектуального аналізу даних та обґрунтована ефективність застосування апарату байєсівських мереж (БМ).

2. Визначено етапи побудови БМ. Проаналізовано методи побудови структури (топології) БМ та методи формування імовірнісного висновку.

3. Визначено особливості побудови гібридних мереж Байєса.

4. Розроблено програмний продукт, який дозволяє будувати гібридні Байєсівські мережі.

5. Порівняно результати роботи створеного програмного продукту з комерційним продуктом-аналогом.

Наукова новизна

1. Розроблено оригінальні моделі у формі БМ, які апробовано на прогнозуванні макроекономічних процесів – ВВП Китаю та США. Досягнуто високу точність оцінок прогнозів.

2. Створено програмний продукт для побудови і використання ймовірнісних моделей у формі байєсівських мереж.

Дякую за увагу!