ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2024, том 517, с. 66–73

МАТЕМАТИКА

УДК 519.876

КОНДЕНСАЦИЯ ГРАФОВ ДЛЯ БОЛЬШИХ ФАКТОРНЫХ МОДЕЛЕЙ

© 2024 г. Академик РАН Б. Н. Четверушкин[1, *], В. А. Судаков[1, **], Ю. П. Титов[2, *]**

Поступило 17.01.2024 г.
После доработки 25.04.2024 г.
Принято к публикации 29.05.2024 г.

В работе разработан оригинальный метод обработки больших факторных моделей на основе конденсации графа с применением моделей машинного обучения и искусственных нейронных сетей. Созданный математический аппарат может быть использован в задачах планирования и управления сложными организационно-техническими системами, при оптимизации крупных социально-экономических объектов масштаба отраслей государства, для решения задач здоровьесбережения нации (поиск совместимостей при приеме лекарственных средств, оптимизация ресурсного обеспечения здравоохранения).

Ключевые слова: факторная модель, конденсация графа, кластеризация, собственный вектор, собственные значения

DOI: 10.31857/S2686954324030119, **EDN:** YAYMIS

1. ВВЕДЕНИЕ

В работе [1] был предложен подход к моделированию сложных систем на основе ориентированного факторного графа. Его вершины – это факторы-показатели, определяющие состояние описываемой системы, а дуги – это взаимосвязи между различными факторами. Степень влияния задавалась экспертами как вес дуги. Аналогично устроены модели системной динамики [2]. Задание веса дуги возможно не только на основе экспертных суждений, но и с использованием методов машинного обучения и нейросетей на базе эмпирических данных [3]. Классическое машинное обучение сталкивается с проблемами, в том числе обусловленными рядом причин, которые не могут быть решены без создания новых методов их обработки:

1) сверхбольшой размер факторных графов (более 1 млн.вершин), приводящий к комбинаторному взрыву возможных взаимосвязей;

2) отсутствием достаточных наборов данных для машинного обучения на столь больших выборках.

Цель работы – это разработка моделей и методов позволяющих находить значения факторов в вершинах графа в случае его большой размерности. Метод решения должен позволять проводить обработку групп вершин в параллельном режиме, за счет этого должно достигаться повышение скорости решения задачи в случае 10 тыс. и более вершин.

Пусть $X = (x_1, x_2, x_3, \ldots, x_n)$ – вектор факторов влияния. $A = \|a_{ij}\|$ – матрица взаимного влияния факторов размерности n x n. Причем основные сложности существующих подходов возникают если $n > 10^6$. Часть элементов матрицы A и вектора X неизвестны.

Требуется найти значения таких факторов x_j, которые связаны хотя бы с одним другим фактором x_i известной дугой: $\exists i\left(a_{ij} \neq 0\right)$.

Для достижения поставленной цели предполагается решение следующих задач:

- разработка методов позволяющих объединять вершины графа в группы;
- разработка процедуры нахождения значений соответствующим вершинам в конденсированном графе;
- разработка методов позволяющих по значениям в конденсированном графе находить значения первичных факторов.

[1] *Федеральное государственное учреждение "Федеральный исследовательский центр Институт прикладной математики имени М.В. Келдыша Российской академии наук", Москва, Россия*

[2] *Федеральное государственное бюджетное образовательное учреждение высшего образования "Московский авиационный институт (национальный исследовательский университет)", Москва, Россия*

[*] *E-mail: office@keldysh.ru*

[**] *E-mail: sudakov@ws-dss.com*

[***] *E-mail: kalengul@mail.ru*

Для решения указанных задач создана оригинальная методика, основанная на гибридизации подходов линейной алгебры и машинного обучения:

• конденсация графов методами машинного обучения без учителя, самоорганизующимися искусственными нейронными сетями;

• поиск неподвижных точек в полученных конденсированных графах пониженной размерности методами линейной алгебры;

• параллельная обработка конденсированных вершин.

2. МЕТОД

Решение поставленной задачи разбивается на ряд этапов, разработанных в настоящей работе:

Этап 1. Разбиваем множество номеров j факторов $X = (x_1, x_2, x_3, \ldots, x_n)$ на подмножества. Каждому подмножеству соответствует новый фактор u_i. Назовем его конденсированный фактор. А процедуру поиска таких подмножеств – конденсацией. Вектор конденсированных факторов обозначим:

$$U = (u_1, u_2, \ldots, u_m), m < n. \quad (1)$$

Разбиение может выполняться различными способами:

1) по смысловому признаку (например, u_i – это группа факторов одного компонента моделируемой системы);

2) на основании метрик связности (в u_i включаются сильносвязанные факторы из X). При большом числе зависимостей факторов в разреженной матрице A для определения сильных компонент целесообразно использовать не все зависимости. Вычисляется среднее (по всем факторам) влияние фактора на другой фактор в разреженной матрице А и при вычислении сильных компонент не учитывается влияние больше среднего. Такой подход близок к “Алгоритму выделения связанных компонент” [4].

Из матрицы A необходимо получить матрицу факторов влияния по новым конденсированным факторам. Обозначим ее $B = b_{kl}$. Обозначим $I(u_k, u_l)$ – множество пар номеров (i, j) факторов X, включенных в конденсированные факторы u_k и u_l для которых известен a_{ij}. Элементы b_{kl} определяют взаимовлияние конденсированных факторов и могут быть определены с использованием одного из следующих правил:

• расстояние ближнего соседа:

$$b_{kl} = \min_{(i,j)\in I(u_k,u_l)} a_{ij}; \quad (2)$$

• расстояние дальнего соседа:

$$b_{kl} = \max_{(i,j)\in I(u_k,u_l)} a_{ij}; \quad (3)$$

• групповое среднее расстояние:

$$b_{kl} = \sum\nolimits_{(i,j)\in I(u_k,u_l)} a_{ij} \Big/ \left|I(u_k,u_l)\right|. \quad (4)$$

Далее будет показано, что использование данных альтернативных правил в текущих тестовых примерах не показало существенной разницы по критериям производительности и точности. В работе [5] для кластеризации применяется метод Уорда, но он требует знания (или предположения) функциональной зависимости между факторами, поэтому его применение не всегда возможно. При наличии такой зависимости поиск значений всех факторов не требует методов машинного обучения. Формулы (2), (3) и (4) можно рассматривать как разновидности функций, отображающих матрицу А на элементы матрицы В. Для выбора конкретного вида функциональной зависимости нет единого математического обоснования. Форма зависимости может диктоваться спецификой решаемых задач. Возможно применение и других правил вычисления b_{kl}, отличных от формул (2), (3) и (4). Правило должно допускать эффективную компьютерную реализацию. Альтернативный вариант выбора правила – это поиск таких b_{kl}, чтобы отклонение от исходных зависимостей было минимальным:

$$\sum_{k=1}^{m}\sum_{l=1}^{m} \sum_{(i,j)\in I(u_k,u_l)} \left|b_{kl} - a_{ij}\right| \to \min. \quad (5)$$

Этап 2. Мощность множества конденсированных факторов m все еще может быть велика. Поэтому определим кластеры конденсированных факторов $V = \|v_k\|$:

$$c_k = \{u_i\}, k = \overline{1,q}, q < m. \quad (6)$$

Кластеризация выполняется либо по смысловому признаку, либо с использованием методов машинного обучения без учителя: k-средних, искусственная нейронная сеть Кохонена [6]. Метрика близости определяются на основании “похожести” столбцов в матрице B используя метрику Минковского:

$$\rho(u_k, u_l) = \left(\sum_{j=1}^{m} \left|b_{jl} - b_{jk}\right|^p\right)^{1/p} \quad (7)$$

В результате получится обобщённая матрица влияния $D = d_{kl}$ размерности q x q. Все элемен-

ты d_{kl} должны быть определены. Для этого применяются аналогичные вышеописанные правила для матрицы B [5]:

• средневзвешенное по всем известным элементам b_{ij} включённым в различные конденсированные вершины:

$$d_{kl} = \sum_{(i,j)\in I(u_k,u_l)} b_{ij} \Big/ \left|S\left(v_k, v_l\right)\right|; \quad (8)$$

• максимум по всем известным элементам b_{ij} включённым в различные конденсированные вершины:

$$d_{kl} = \max_{(i,j)\in S(v_k,v_l)} b_{ij}, \quad (9)$$

где $S\left(v_k, v_l\right)$ – множество пар номеров (i, j) факторов U, включенных в конденсированные факторы v_k и v_l для которых известен b_{ij}.

Число кластеров q выбирается, как и в стандартных задачах машинного обучения, на основе “метода локтя” (elbow method), метрики силуэт [7]. Если для кластеризации используется смесь нормальных распределений (Gaussian Mixture Model), то используется байесовский информационный критерий (BIC) [8].

Этап 3. Обозначим неизвестные значения факторов полученной обобщенной модели:

$$V = \left(v_1, v_2, v_3 \ldots v_q\right). \quad (10)$$

Их значения находятся из решения матричного уравнения:

$$V = DV + F, \quad (11)$$

где F – вектор значений, характеризующий внешнее влияние.

Значения факторов V могут быть найдены с помощью итерационной процедуры:

$$V_{k+1} = DV_k + F. \quad (12)$$

Скорость сходимости этой процедуры (и сама возможность сходимости) определяются собственными значениями матрицы D:

$$\lambda V = DV. \quad (13)$$

Необходимо найти значения конденсированных факторов V как собственный вектор, соответствующий максимальному собственному значению.

Этап 4. Так как $u_i \cong v_k$ для $u_i \in C_k$ по смыслу образования кластеров, необходимо построить модели $v_k \cong \varphi_k\left(X_{ki}\right)$ для всех элементов кластера чтобы отклонение отклика от значений конденсированных признаков было минимальным. Здесь $\cong$ обозначает приблизительное равенство, то есть $\left|u_i - v_k\right| < \varepsilon$, где $\varepsilon > 0$ – это небольшая допустимая погрешность в определении точного значения фактора, φ_k – это новая функция для определения значений для v_k исходя из значений допустимых значений факторов X_{ki} первичной модели из конденсированной вершины i попавших в кластер k.

Решение зависит от мощности кластера и того все ли значения X_{ki} известны:

• для случая кластеров с одной конденсированной вершиной целесообразно использовать линейную модель и решение уравнения аналогичного формуле (11);

• для кластеров с 20–100 конденсированными вершинами целесообразно оценить модель на базе классических методов машинного обучения: случайного леса, градиентного бустинга;

• для случая большого числа неопределенных факторов в векторе X_{ki} целесообразно построить $\varphi_k\left(X_{ki}\right)$ как вероятностную графовую модель;

• для кластеров с более 100 вершинами целесообразно провести обучение нейронных сетей.

Этап 4 выполняется в массовом параллельном режиме на суперЭВМ.

Полученную модель можно использовать для решения обратных задач математического моделирования в сложных системах. Изменяя некоторые значения в векторе первичных факторов X методика позволяет рассчитывать изменения значений в конденсированных вершинах u_i и через обратные функции φ_k^{-1} находить как изменятся значения остальных факторов. Решение может быть неоднозначным, если среди φ_k существуют не биективные функции, и это в большинстве случаем так. Однако метод позволит определить некие граничные (худшие/лучшие) возможные значения неизвестных при данной вариации факторов. такие граничные значения факторов позволяют перейти к нечетким оценкам и проводить многокритериальный анализ с использованием теории мягких вычислений. Такая задача в частном случае решалась в работе [9]. Новый метод, созданный в настоящей работе, позволяет получать решения для задач, сформулированных в общем виде.

Применение данных этапов решения задачи ориентированно на ускорение вычислений без существенной потери точности и будет продемонстрировано ниже.

3. АПРОБАЦИЯ МОДЕЛИ

Для тестирования разработанных методов разработано программное обеспечение на язы-

ке программирования Python. Тестирование производится на множестве различных случайных графах. Для генерации вектора факторов X в программе задается размерность вектора факторов X и определяются статистические характеристики: вероятность наличия значения фактора. Для тестового примера фактор – это действительное число. Для факторов с известными значениями они генерируются по нормальному закону. Таким образом, производится исследование работы программы с различной наполненностью вектора факторов. Следует отметить, что в реальных задачах факторы могут быть представлены с другим типом значений параметров и свойств системы.

После получения вектора факторов генерируется матрица взаимовлияния факторов A. Для генерации адекватных зависимостей между факторами, факторы моделируются как множество точек на двумерной плоскости. Точки разделяются на некоторое количество кластеров, группы факторов. Для каждого кластера (группы факторов) определяются статистические характеристики кластера в виде математических ожиданий и дисперсий значений каждой из координат. Далее случайным образом определяются значения координат для каждой точки (каждого фактора) в соответствии с нормальным распределением и параметрами для кластера данного фактора. Количество факторов в кластере не одинаковое, а задается дискретной случайной величиной, определяющей к какому кластеру (группе факторов) принадлежит фактор. Генерация может проводиться в цикле, создавая различные матрицы A для одинаковых параметров кластеризации факторов и размещения их на плоскости. Пример матрицы A с точками, расположенными на плоскости приведен на рис. 1. Для целей тестирования зависимости между факторами определяются как евклидово расстояние между точками.

Для выявления сильных компонент рассматриваются только дуги, длиннее средней длинны дуги. В результате создается новая разреженная матрица A с соответствующими дугами. Путем вычисления сильносвязанных компонент новой разреженной матрицы A определяются конденсированные компоненты. Вычисление взаимовлияния конденсированных компонент матрицы B (рис. 2) находится по формулам (2–4) в параллельном режиме. При этом общие исходные данные матрицы A позволяют сравнивать алгоритмы рассматриваемые алгоритмы. Аналогично конденсация находится для матриц C и D. На рис.2 приведен пример графа для 10^4 факторов.

Кластеризация k-means и построение искусственной нейронной сети Кохонена занимает основное время работы процессора даже при конденсированных факторах. Также длительное время занимает вычисление взаимовлияний конденсированных факторов. Для данных алгоритмов вычислительная сложность растет экспоненциально росту размерности вектора факторов. В примере известно количество кластеров из-за особенностей генерации тестовой матрицы A. Из-

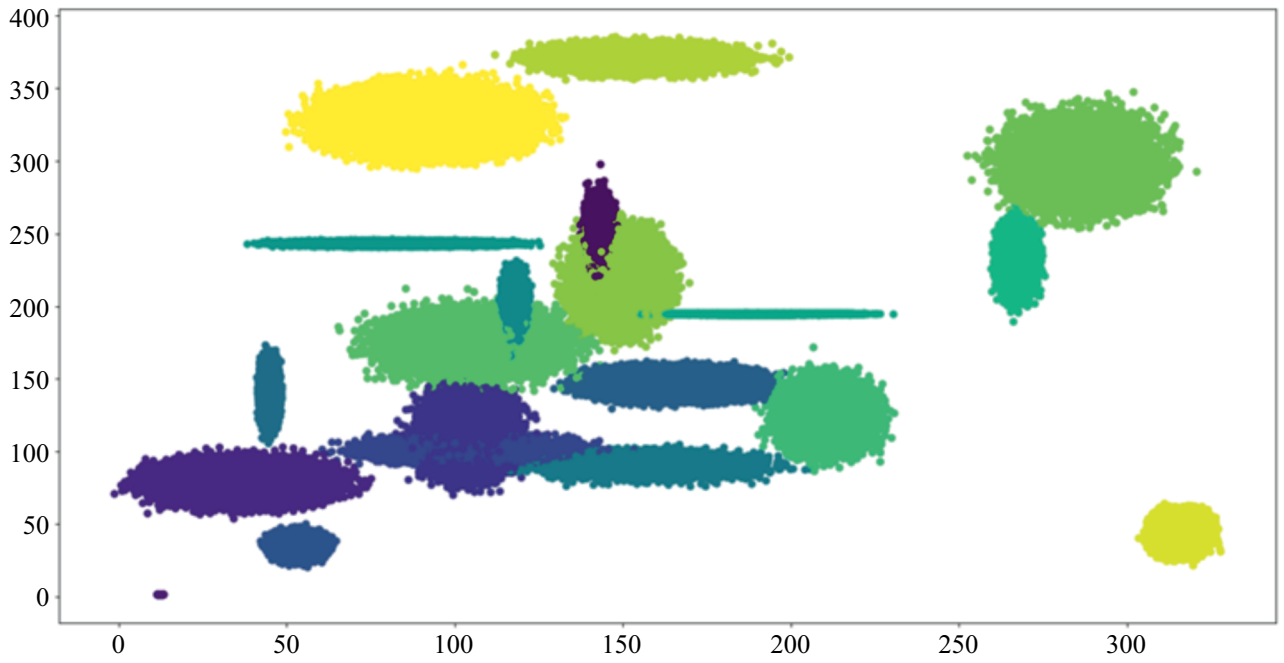

Рис. 1. Пример генерации факторов и матрицы A.

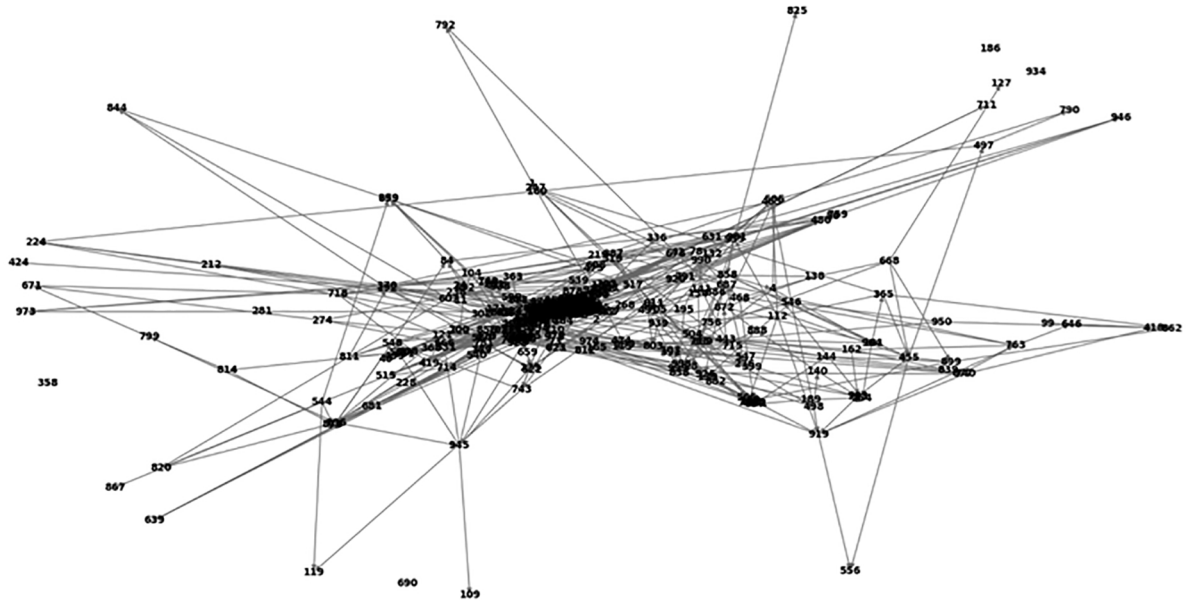

Рис. 2. Пример матрицы B, состоящей из конденсированных факторов.

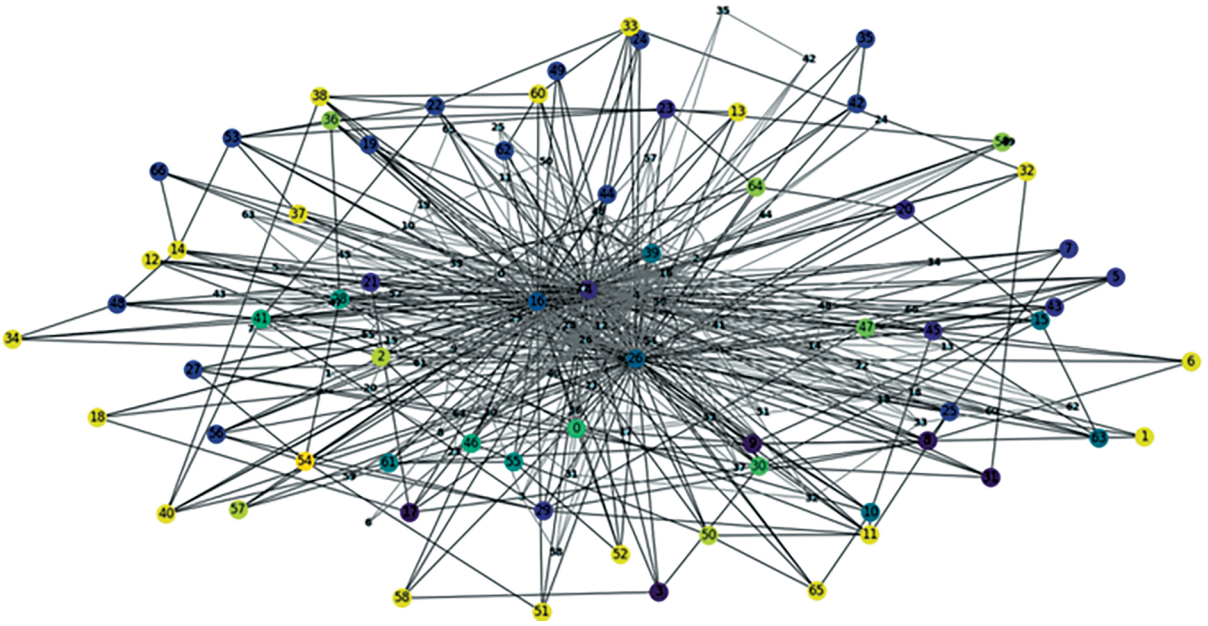

Рис. 3. Пример кластеризации конденсированных факторов методом k-means.

начально было выбрано 20 центральных точек и сгенерированы случайные точки вокруг центров. Поэтому количество кластеров было равно 20. Число центральных точек было выбрано так, чтобы произвести вычислительные эксперименты за приемлемое время. При решении реальных задач, выбор числа кластеров является компромиссом между двумя показателями – необходимыми метриками качества кластеризации (например, метрика силуэт) и приемлемым временем решения задачи. Пример конденсации графа путем кластеризации приведен на рис. 3.

Эффективность работы алгоритма оценивается по времени работы отдельных оставляющих и точности восстановленных факторов. В рассматриваемой тестовой задаче доля неизвестных факторов и неизвестных зависимостей равна 0.3.

В таблице 1 приведены результаты сравнения времени работы полной обработки графа без конденсации и 3 варианта конденсации вершин по формулам (2), (3) и (4). рассматривались 4 варианта размерности тестовой задачи. Оценки времени получены по 100 вычислительным экспериментам.

Таблица 1. Среднее время выполнения методов (в сек)

Размерность вектора факторов	**10000**	**20000**	**30000**	**40000**
Решение задачи без конденсации				
Суммарное время на восстановление неизвестных факторов	**9.234**	**77.726**	**258.076**	**640.243**
Время выполнения общих шагов в случае конденсации графа				
Поиск сильно связанных компонент в матрице A	**0.045**	**0.219**	**0.452**	**0.832**
Создание сокращенной матрицы A	**0.266**	**1.270**	**2.942**	**5.711**
Определение сильных компонент	**0.016**	**0.066**	**0.146**	**0.300**
Конденсация с учетом расстояния по формуле (2)				
Создание матрицы B путем объединения кластеров	**0.236**	**1.011**	**2.269**	**3.924**
Восстановление неизвестных факторов B	**0.003**	**0.006**	**0.004**	**0.005**
Вычисление k-means	**0.017**	**0.018**	**0.018**	**0.018**
Суммарное время на восстановление неизвестных факторов	**0.585**	**2.590**	**5.831**	**10.791**
Конденсация с учетом расстояния по формуле (3)				
Создание матрицы B путем объединения кластеров	**0.257**	**1.012**	**2.235**	**3.839**
Восстановление неизвестных факторов B	**0.004**	**0.006**	**0.006**	**0.004**
Вычисление k-means	**0.016**	**0.016**	**0.016**	**0.016**
Суммарное время на восстановление неизвестных факторов	**0.605**	**2.590**	**5.798**	**10.702**
Конденсация с учетом расстояния по формуле (3)				
Создание матрицы B путем объединения кластеров	**0.268**	**1.063**	**2.368**	**4.061**
Восстановление неизвестных факторов B	**0.003**	**0.007**	**0.004**	**0.004**
Вычисление k-means	**0.017**	**0.017**	**0.017**	**0.016**
Суммарное время на восстановление неизвестных факторов	**0.616**	**2.642**	**5.930**	**10.925**

Из таблицы видно, что время решения задачи с использованием конденсации на порядок меньше времени чем без использования конденсации. При восстановлении факторов вычислялось математическое ожидание координат фактора. Выделение сильных компонент и создание матрицы B позволяет существенно сократить количество факторов, что приводит к резкому уменьшению времени поиска решения.

При этом точность найденных значений факторов при конденсации уменьшается, но незначительно. Так относительная средняя точность значения фактора без конденсации составляет 0.893 для 10000 факторов; 0.912 для 20000 факторов; 0.935 для 30000 факторов и 0.953 для 40000 факторов, а при конденсации любым из трех методов – 0.873; 0.900; 0.906 и 0.920 соответственно. Также следует отметить увеличение точности при увеличении выборки, связанное с тем, что при усреднении факторы объединяются, а размерность матрицы B растет незначительно.

В данный момент разработанные методы апробируются для задачи определения рисков полифармокотерапии, возникающих при одновременном применении нескольких лекарственных средств. Исходными данными такой задачи являются утвержденные Минздравом РФ инструкции по применению лекарственных средств. В инструкциях указаны взаимодействия между некоторыми препаратами, но такое взаимодействие имеется только для парных взаимодействий. Риски, по применению трех и более медицинских препаратов не определены в инструкциях. Кроме того, в инструкциях определены только некоторые попарные взаимодействия, так как полное множество даже попарных взаимодействий лекарственных средств велико.

Взаимодействие лекарственных средств возможно определить с помощью экспертов. Такие экспертные оценки, принятые медицинским сообществом, оформлены в клинические рекомендации. Рекомендации описывают лекарственные препараты (медицинские непатентованные названия), назначаемые при определенных заболеваниях и диагнозах. При этом ситуации, когда необходимо назначить более трех лекарственных средств, или пациент уже принимает большое количество лекарственных средств, оцениваются консилиумом врачей.

Взаимодействие лекарственных средств требует учета влияния лекарственного средства на организм человека. Из инструкций к лекар-

ственным средствам выявляется множество мишеней действия. В инструкции к лекарственному препарату указываются не все мишени, а только основные.

$$P_i = \left(p_{i,1}\, p_{i,2} \ldots p_{i,m_i}\right), \text{ для } \forall\, P_i \in S, \qquad (14)$$

где S – множество лекарственных средств в реестре Минздрава; P_i – множество мишеней для i-го лекарственного средства;

Принимаемое лекарственное средство воздействует на мишени различным образом. Степень воздействия определяется лингвистическими переменными: $p_{i,j} \in$ {“усиливает”, “ослабляет”…}. Переход от лингвистических переменных к непрерывным осуществляется с помощью частотного анализа, экспертных оценкок или через нечеткие множества. По результатам данного перехода указывается степень влияния $p_{i,j} \in [0..1]$, где 0 – отсутствие влияния на мишень, 1 – максимальное влияние.

Возможно взаимовлияние некоторых мишеней, которое может быть вычислено из инструкции к лекарственному средству. В основе вычислений взаимовлияний мишеней положено предположение, что лекарственное средство представлено одним действующим веществом, которое, в зависимости от дозировки, влияет сразу на несколько мишеней. Поэтому, если в инструкции к лекарственному средству не указано влияние на мишень, а у другого лекарственного средства с аналогичным действующим веществом это влияние найдено, то можно предположить о возможном неуказанном влиянии первого лекарственного средства на все остальные мишени. Такая информация может быть представлена матрицей взаимовлияния мишеней A.

При взаимодействии двух лекарственных средств их влияние на мишени усиливаются. Итоговое влияние комбинации принимаемых лекарственных средств на мишени определяется значениями факторов при мишенях. Так как количество комбинаций принимаемых лекарственных средств велико, то эффективно хранить только важные и интересующие врачей и пациентов комбинации, а остальные комбинации могут быть вычислены по результатам работы системы. Падение точности оценок при конденсации графа, показанное в тестовом примере, не существенно, а время вычисления взаимовлияния лекарственных средств является важным критерием, поэтому целесообразно применение конденсации графа.

4. ЗАКЛЮЧЕНИЕ

Разработанная методика построения моделей сложных систем ориентирована на решение задач учета большого числа взаимосвязных факторов. Она позволяет решать высокоразмерные задачи путем выделения сильносвязных компонент в факторном графе и кластеризации этих компонент. Методика может быть использована при решении обратных задач математического моделирования для поиска рациональных вариаций факторов, поддающихся изменениям. Создаваемые модели позволяют учесть как экспертные суждения о взаимном влиянии факторов, так и данные, содержащиеся в эмпирических наблюдениях. Целесообразно применение данного подхода в масштабных задачах здоровьясбережения нации (поиск совместимостей при приеме лекарственных средств, оптимизация ресурсного обеспечения здравоохранения). В перспективе методика позволит оценивать решения по управлению сложными организационно-техническими системами и может быть использована для оптимизации крупных социально-экономических систем государственного масштаба.

СПИСОК ЛИТЕРАТУРЫ

1. *Четверушкин Б.Н., Судаков В.А.* Факторная модель для исследования сложных процессов // Доклады Академии наук. 2019. Т. 489. № 1. С. 17–21.
2. *Forrester J.W.* Policies, decisions and information sources for modeling // European Journal of Operational Research. 1992. V. 59. № 1. P. 42–63.
3. *Honti G., Dörgő G., Abonyi J.* Network analysis dataset of system dynamics models // Data in Brief. 2019. V. 27. P. 104723.
4. *Jain A., Murty M., Flynn P.* Data Clustering: A Review // ACM Computing Surveys. 1999. V. 31. № 3.
5. *Lance G.N., Williams W.T.* A general theory of classification sorting strategies in hierarchical system // Comp. J. 1967. № 9. P. 373–380.
6. *Kohonen T.* Essentials of the self-organizing map // Neural Networks. 2013. V. 37. P. 52–65.
7. *Alam A., Ahamad M.K.* K-Means Hybridization with Enhanced Firefly Algorithm for High-Dimension Automatic Clustering // Journal of Advanced Research in Applied Sciences and Engineering Technology. 2023. V. 33. № 3. P. 137–153.
8. *Reynolds D.* Gaussian Mixture Models // Encyclopedia of Biometrics. Boston, MA: Springer, 2009.
9. *Нестеров В.А., Судаков В.А., Сыпало К.И., Титов Ю.П.* Матрица нечетких корреспонденций модели авиационных перевозок // Известия Российской академии наук. Теория и системы управления. 2022. Т. 6. № 6. С. 95–102.

GRAPH CONDENSATION FOR LARGE FACTOR MODELS

Academician of the RAS B. N. Chetverushkin[a], V. A. Sudakov[a], Yu. P. Titov[b]

[a]*Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia*

[b]*Moscow Aviation Institute (National Research University), Moscow, Russia*

The paper proposes an original method for processing large factor models based on graph condensation using machine learning models and artificial neural networks. The proposed mathematical apparatus can be used in problems of planning and managing complex organizational and technical systems, in optimizing large socio-economic objects on the scale of state sectors, to solve problems of preserving the health of the nation (searching for compatibility when taking medications, optimizing resource provision for healthcare).

Keywords: factor model, graph condensation, clustering, eigenvector, eigenvalues