

Алгебраические байесовские сети: точное построение канонического фрагмента знаний

А. А. Вяткин¹, М. В. Абрамов²

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

¹aav@dscs.pro, ²mva@dscs.pro

Аннотация. Алгебраические байесовские сети относятся к классу вероятностных графических моделей. Такие сети позволяют обрабатывать скалярные и интервальные оценки вероятности истинности пропозиций, соответствующих утверждениям, где все множество утверждений декомпозируется на наборы — фрагменты знаний. При практическом использовании аппарата алгебраических байесовских сетей характерным является то, что время обработки интервальных оценок на несколько порядков выше аналогичного времени в случае скалярных оценок. В целях оптимизации сложности работы за счет использования скалярных оценок ранее был предложен подход к переходу от модели с интервальными оценками к модели со скалярными, или, другими словами, поиску приближенного канонического представителя фрагмента знаний алгебраической байесовской сети. В данной работе изучается поиск точного канонического представителя фрагмента знаний, что рассматривается впервые. Такое построение позволит дать выигрыш по времени при использовании фрагментов знаний, сформированных на алфавитах небольшой мощности. В ходе исследований было обнаружено, что алгоритм точной генерации в случае фрагмента знаний мощности 1 работает почти мгновенно, в случае фрагмента знаний мощности 2 — быстрее в 30 раз, чем текущий алгоритм приближенной генерации, в случае мощности 3 — в 1.5–2 раза быстрее. Данный подход является особенно актуальным с учетом того, что в теории алгебраических байесовских сетей предполагаемые к использованию на практике фрагменты знаний будут иметь небольшие размеры.

Ключевые слова: алгебраические байесовские сети; канонический представитель; фрагмент знаний; вероятностные графические модели; машинное обучение

I. ВВЕДЕНИЕ

Вероятностные графические модели на данный момент являются объектом изучения многих исследователей и находят широкое применение в различных сферах, таких как компьютерное зрение [1] и обработка изображений [2], медицина [3–4], генетика [5], управление динамическими системами [6], общая обработка данных [7]. Среди вероятностных графических моделей отдельно можно выделить класс алгебраических байесовских сетей (АБС), позволяющих описывать знания о предметной области, представляемые в виде утверждений со степенью уверенности в их истинности.

Одной из основ, лежащих в теории АБС, является декомпозируемость знаний на фрагменты знаний (ФЗ),

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022-0003.

рассматриваемые в виде небольших наборов утверждений. Формализованные фрагменты знаний — составные элементы АБС, глобальной структуры, построенной над ними. Одним из классических математических представлений фрагмента знаний является набор квантов, где каждому кванту сопоставляется скалярная или интервальная оценка вероятности истинности. Последние позволяют формализовать степень уверенности в истинности утверждений, а также логико-вероятностно связать их. Далее, для краткости, формализованные фрагменты знаний будем также называть фрагментами знаний.

Интервальные оценки вероятности истинности, в сравнении со скалярными, позволяют более широко и полно интерпретировать знания о предметной области, однако и время работы алгоритмов, использующих интервальные оценки, становится на порядки выше [8]. Поэтому, в случае дефицита времени или вычислительных мощностей, АБС со скалярными оценками может оказаться предпочтительнее. Ранее был предложен подход к переходу от ФЗ с интервальными оценками к ФЗ со скалярными, характеризующему первый ФЗ наиболее полно, — построение канонического представителя ФЗ [9]. В этой работе рассматривалось построение приближенного канонического представителя ФЗ с тем допущением, что формирование точной структуры было бы вычислительно сложно. Однако, для ФЗ небольшого размера создание точного канонического представителя может дать выигрыш по времени, и в данной работе впервые описывается такое построение, а также сравнивается время работы алгоритмов точного и приближенного построения. Такой подход является также актуальным с той точки зрения, что ФЗ при практическом применении теории АБС предположительно будут рассматриваться над небольшими наборами утверждений (до 4–5 элементов).

Теоретическая значимость работы состоит в определении нового алгоритма построения точного канонического представителя ФЗ. Практическая значимость заключается в выявлении оптимального по времени работы, в зависимости от размера ФЗ, алгоритма построения канонического представителя ФЗ, что позволит расширить спектр практических задач, решаемых с помощью использования АБС.

II. ПОСТАНОВКА ЗАДАЧИ

В данном разделе дадим краткий обзор терминов, используемых в работе, а также более строго сформулируем поставленную задачу. Рассматриваемые объекты и их свойства определяются в более ранних исследованиях, посвященных теории АБС [8–10].

Как было неформально сформулировано выше, АБС рассматриваются как глобальные структуры, определяемые над наборами ФЗ. Более строго, такие сети являются моделью баз ФЗ, где связи между отдельными ФЗ можно представить в виде графа. Сами ФЗ определяются над *алфавитом* – набором атомарных позиций (*атомов*), соответствующих утверждениям. Таким алфавитом, содержащим два атома, может являться, например, алфавит $\{x_0, x_1\}$.

В данной работе, для определения модели ФЗ, также будет необходимо понятие *кванта* – конъюнкции всех элементов заданного алфавита, а также их отрицаний. Например, для алфавита $\{x_0, x_1\}$ набором квантов будет $\{\overline{x_1 x_0}, \overline{x_1 x_0}, \overline{x_1 x_0}, \overline{x_1 x_0}\}$. Итак, ФЗ формализуются в виде набора всевозможных квантов, где каждому кванту сопоставляется скалярная или интервальная оценка вероятности истинности. Оценки ФЗ определяются целостно в том смысле, что для всех квантов в ФЗ они могут быть либо все скалярные, либо все интервальные.

Стоит упомянуть, что ФЗ может определяться также через наборы конъюнктов и дизъюнктов с соответствующими оценками вероятности, однако между всеми тремя типами существуют взаимно однозначные матричные переходы, и выбор представления обуславливается требованиями конкретной задачи. В данном случае будет удобен набор ограничений, накладываемый на оценки вероятности истинности, сопоставляемые квантам. Так, для набора квантов q_1, \dots, q_n на скалярные оценки накладывается следующие ограничения:

$$\begin{cases} \sum_{i=1}^n p(q_i) = 1, \\ q_i \geq 0, \forall i \in \{1, \dots, n\}, \end{cases}$$

где $p(q)$ является оценкой вероятности истинности кванта q .

Интервальные же оценки имеют вид, например, $p(q_0) = [0.4, 0.5]$ и одновременно рассматривают всевозможные распределения, в которых оценка вероятности истинности кванта q_0 лежит в пределах заданного отрезка (в данной работе понятия «отрезок» и «интервал» в таких случаях будут определять один и тот же объект). Соответственно набору квантов с интервальными оценками сопоставляется семейство наборов квантов со скалярными оценками, где скалярные оценки лежат в интервалах, задаваемых соответствующими интервальными оценками. Таким образом, ФЗ с интервальными оценками можно рассматривать как выпуклый многогранник в N' -мерном пространстве, где $N' = 2^n$, и n – мощность алфавита, над которым построен ФЗ. Каждая точка в этой области будет сопоставляться соответствующему скалярному ФЗ, поэтому, более точно, ФЗ с интервальными оценками можно рассматривать как

выпуклый многогранник в стандартном N -мерном симплексе [11], где $N = N' - 1$.

В теории АБС основными операциями, рассматриваемыми над ФЗ и АБС, являются *априорный* и *апостериорный* выводы, позволяющие соответственно определять оценки вероятности истинности новых пропозициональных формул, построенных над атомами-пропозициями, а также моделировать учет в сети новой информации о предметной области. Для проведения подобных алгоритмов в случае интервальных оценок необходимо решать задачи линейного программирования, в случае же скалярных – выполнять более простые матричные операции. Исходя из этого, время работы алгоритмов, использующих интервальные оценки, может превышать время аналогичной работы в случае скалярных оценок на несколько порядков [8]. Данное свойство работы алгоритмов может оказаться критичным в случае дефицита времени или вычислительных мощностей, где АБС со скалярными оценками были бы предпочтительнее. Таким образом, в данной ситуации было бы актуально построение *канонического представителя* АБС и каждого ФЗ — переход от объектов с интервальными оценками к объектам со скалярными, где выбранный объект со скалярными оценками должен максимально полно представлять объект с интервальными. Таким объектом является центр масс области, лежащей в N -мерном симплексе и задаваемой ФЗ с интервальными оценками.

Ранее был предложен подход к поиску приближенного канонического представителя ФЗ [9], где точка, приближенная к центру масс, искалась с помощью метода Монте-Карло путем взятия среднего от набора, случайно сгенерированных точек, попавших в область, задаваемую интервальным ФЗ. В работе было указано, что достаточно 100–1000 генерации точек. Такой способ представляется удобным в плане скорости построения канонического представителя для ФЗ, определенным над большим числом атомов. Для небольшого числа атомов может быть более эффективным нахождение точного центра масс, чему посвящена данная работа и что будет рассматриваться далее. Такой подход является особенно актуальным с учетом того, что при практическом использовании АБС предполагается рассмотрение небольшого числа атомов, а также того, что при построении точного канонического представителя полученный ФЗ будет полнее описывать изначальный ФЗ.

III. ПОСТРОЕНИЕ ТОЧНОГО КАНОНИЧЕСКОГО ПРЕДСТАВИТЕЛЯ ФЗ

В данном разделе предлагается новый алгоритм построения точного канонического представителя ФЗ, опишем его:

1. В первую очередь необходимо определить вершины выпуклого многогранника H , соответствующего ФЗ с интервальными оценками и лежащего в N -мерном симплексе. Для их определения:
 - а. Рассматриваем все возможные комбинации $n-1$ границы отрезков, взятых по одному из каждого интервала, где n – количество

интервальных оценок ($N = n - 1$). Например, для оценок $p(q_0) = [0.1, 0.5]$, $p(q_1) = [0.1, 0.3]$, $p(q_2) = [0.1, 0.3]$, $p(q_3) = [0.2, 0.3]$, комбинациями были бы $(0.1, 0.1, 0.1)$, $(0.1, 0.1, 0.3)$, $(0.3, 0.3, 0.2)$, $(0.5, 0.3, 0.3)$ и так далее, всего $4 \cdot 2^3 = 32$ комбинации (пример А).

- б. Для каждой комбинации, во-первых, считаем сумму оценок в комбинации и вычитаем ее из 1 (получаем q'), и, во-вторых, смотрим на то, лежит ли q' в интервале, который не был взят для построения комбинации. Если q' лежит в интервале – дополняем комбинацию оценкой q' и считаем полученную точку вершиной многогранника, в противном случае комбинацию отбрасываем. Например, в примере А для первой комбинации $q'_1 = 1 - (0.1 + 0.1 + 0.1) = 0.7$. Для этой комбинации не была взята оценка $p(q_3)$, так что сравниваем ее с $q'_1 \Rightarrow q'_1 \notin p(q_3)$, поэтому такую комбинацию отбрасываем. Для 3-ей же комбинации $q'_3 = 1 - (0.3 + 0.3 + 0.2) = 0.2$ и $q'_3 \in p(q_0)$, поэтому считаем точку $(q'_3, 0.3, 0.3, 0.2) = (0.2, 0.3, 0.3, 0.2)$ одной из вершин многогранника.

- Находим триангуляцию выпуклого многогранника (при реализации был использован метод Делоне [12]).
- Для каждого симплекса из триангуляции считаем центр масс — среднее всех точек, а также находим объем каждого из симплексов. Например, если размерность полученного многогранника равняется $n - 1$, то в этом случае можно использовать следующую формулу [13]:

$$V = \frac{1}{(n-1)!} \det(v_2 - v_1, \dots, v_n - v_1),$$

где \det – определитель, v_1, \dots, v_n – вершины симплекса.

- В итоге, для определения центра масс многогранника, и, таким образом, канонического представителя, далее необходимо найти сумму центров масс симплексов, взвешенную нормализованными значениями их объемов.

Утверждение 1. Вышеописанный алгоритм поиска канонического представителя действительно найдет его точное представление.

Доказательство. Стоит сразу отметить, шаги 2–4 являются стандартным способом определения центра

масс выпуклого N -мерного многогранника, который и необходимо найти для получения канонического представителя. Таким образом, единственное, что остается показать – то, что вершины, получаемые на шаге 1, действительно соответствуют всем вершинам выпуклого многогранника H , задаваемого интервальным ФЗ.

Действительно, так как многогранник H образован «отсечением» гиперплоскостей вида $p_i = p_i^+$ или $p_i = p_i^-$, где p_i – скалярная оценка вероятности кванта q_i , лежащая в интервале $[p_i^-, p_i^+]$, то вершины v_j многогранника H могут иметь только вид $(p_0^{(+,-)}, \dots, p_{k-1}^{(+,-)}, p'_k, p_{k+1}^{(+,-)}, \dots, p_s^{(+,-)})$, где s – количество вершин, $p_i^{(+,-)}$ – верхняя или нижняя оценка кванта q_i , $p'_k = 1 - \sum_{i, i \neq k} p_i^{(+,-)}$. Последнее условие определяется ограничением, накладываемым на скалярные оценки квантов. Если бы было две оценки p'_k и p'_l , не равные никакому $p_i^{(+,-)}$, то v_j , образованная этими координатами, лежала бы внутри отрезка между двумя точками, полученными добавлением и отниманием малой Δ от координат, соответствующих p'_k и p'_l , а значит не являлась бы 0-мерной гранью (вершиной) [14].

При этом любая точка, взятая на шаге 1 как вершина, будет являться вершиной. Покажем это методом от противного. Предположим, что это не так, тогда существует отрезок между двумя лежащими в H точками $X = (x_1, \dots, x_n)$ и $Y = (y_1, \dots, y_n)$ и точка P_i , взятая на шаге 1, лежит между этими точками и не равна им, то есть $P_i = tX + (1-t)Y$ для $t \in (0, 1)$. Возьмем координату p_k точки P_i , полученную из границы соответствующего интервала, тогда $p_k \geq x_k$ и $p_k \geq y_k$ или $p_k \leq x_k$ и $p_k \leq y_k$, и, пусть, н.у.о., выполнены первые неравенства. Тогда $p_k = tx_k + (1-t)y_k$. Пусть, н.у.о., $y_k > x_k$, тогда $p_k = y_k$ (так как p_k не может быть строго больше из-за равенства с t), поэтому $p_k = tx_k + (1-t)p_k \Rightarrow t(x_k - p_k) = 0 \Rightarrow x_k = p_k$, так как $t \neq 0$. Таким образом, $\forall x_j, y_j$ выполнено $x_j = y_j = p_j$ для всех p_j , полученных из границы соответствующего интервала $\Rightarrow X = Y = P_i$, так как сумма всех координат должна быть равна 1. В итоге приходим к противоречию, что, в конечном счете, приводит к доказательству всего утверждения. *Конец доказательства.*

Стоит также отметить, что перебор вершин в пункте 1, вероятно, можно ускорить. Однако, как выяснилось при проведении экспериментов, основное время уходит на построение триангуляции, и для ускорения алгоритма нужно в первую очередь ускорять именно ее построение.

Второе замечание относится к построению центра масс для ФЗ, формируемого над алфавитом из одного атома. В таком контексте выпуклая оболочка вырождается в отрезок, поэтому в этом случае было решено находить центр масс (центр отрезка) непосредственно, не прибегая к построению триангуляции.

IV. ОПИСАНИЕ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА

Данный раздел посвящен сравнению двух алгоритмов генерации канонического представителя: точного и приближенного.

Для проведения эксперимента алгоритмы были реализованы на языке программирования *Python*. Всего рассматривались ФЗ построенные над 1–3 атомами, для 4-ех атомов построение точного ФЗ становится слишком долгим (превышает минуты). Случайные ФЗ формировались с использованием алгоритма, описанного в [14]. Таким образом, для всех трех типов ФЗ было сформировано $N = 500$ экземпляров, над которыми и проводился эксперимент. Приближенный алгоритм рассматривался несколько раз – в одном случае генерировалось 100 ФЗ, во втором – 1000.

Далее стоит обсудить то, как именно будут представляться результаты. Во-первых, будет показано сравнение времени работы алгоритмов, отображаемых на графиках как десятичный логарифм отношения времени работы приближенного алгоритма ко времени работы точного алгоритма. Однако, для ФЗ, построенных над 1 атомом, время работы точного алгоритма равнялось почти 0, поэтому в этом случае будет отображаться сразу логарифм от времени работы. Во-вторых, будет показан график ошибок, получаемых при использовании приближенного алгоритма. Для вычисления ошибки рассчитывалось

$$MAE = \frac{1}{N} \cdot \sum_i (|p_{exact}^i - p_{approx}^i|), \text{ где } p_{exact}^i \text{ и } p_{approx}^i \text{ —}$$

соответствующие границы в полученных интервальных оценках. В-третьих, так как долгое время генерации приближенного канонического представителя обусловлено тем, что получаемые скалярные ФЗ не попадают в ограничения для интервального ФЗ и отбрасываются, было решено также привести время генерации только 100 ФЗ и только 1000 ФЗ, чтобы показать работу алгоритма в идеальном случае. На графиках это время будет также отображаться либо напрямую, либо в виде отношения ко времени работы точного алгоритма.

Итоговые времена работы в случае ФЗ с одним атомом представлены на рис. 1. Стрелка на графике показывает, что алгоритм выполнялся, в том числе, за почти нулевое время. На рис. 2 и рис. 3 показано отношение времени работы приближенного алгоритма ко времени работы точного алгоритма для ФЗ с двумя и тремя атомами соответственно. Рис. 4 показывает полученные значения MAE приближенной генерации для ФЗ с 1–3 атомами.

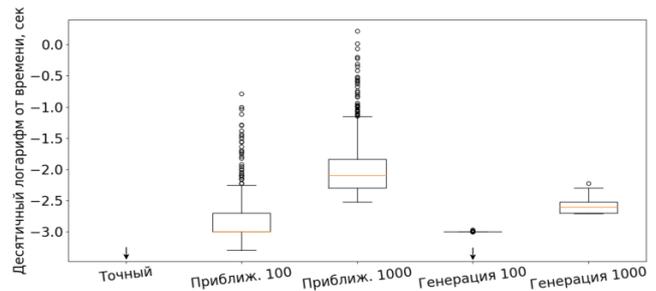


Рис. 1. Время работы алгоритмов для ФЗ с 1 атомом

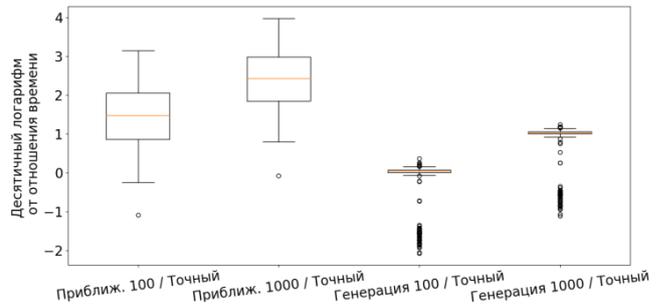


Рис. 2. Отношение времени работы алгоритмов для ФЗ с 2 атомами

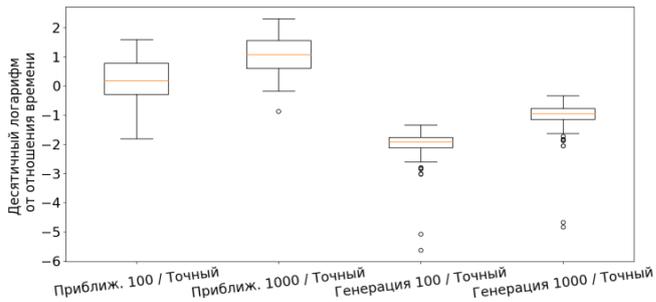


Рис. 3. Отношение времени работы алгоритмов для ФЗ с 3 атомами

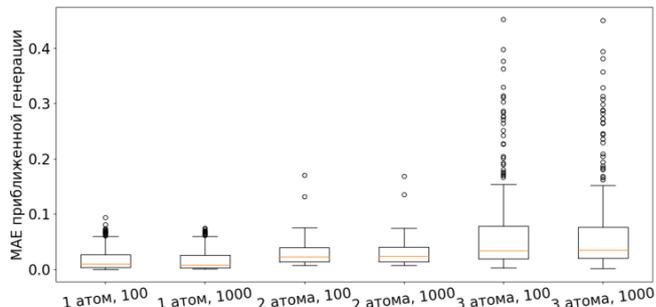


Рис. 4. MAE приближенной генерации

Итоговые результаты оформим в виде следующих утверждений:

1. Время работы приближенного алгоритма, генерирующего 1000 ФЗ на порядок больше времени работы алгоритма, генерирующего 100 ФЗ, при этом ошибка изменяется не сильно (на ~ 0.001 единицы). Таким образом, предпочтительнее использовать генерацию 100 ФЗ, с чем и будет далее сравниваться алгоритм точной генерации.
2. На текущий момент, для ФЗ, построенных над 1–3 атомами предпочтительнее использовать алгоритм точной генерации. В случае двух

атомов он работает быстрее в среднем в ~30 раз, в случае в трех атомов – в среднем в ~1.5–2 раза. В случае одного атома работа алгоритма точной генерации почти мгновенна.

3. В случае с тремя атомами алгоритм приближенной генерации потенциально может быть быстрее в 80–100 раз, на что указывает сравнение времени работы точного алгоритма и генерации 100 ФЗ. Для остальных типов ФЗ точный алгоритм будет быстрее даже в сравнении с простой генерацией ФЗ.

V. ЗАКЛЮЧЕНИЕ

В данной работе впервые рассматривается алгоритм точного построения канонического представителя ФЗ. Этот алгоритм был сравнен с алгоритмом приближенного построения, изучаемым ранее в теории АБС. В ходе исследований было обнаружено, что алгоритм точной генерации в случае ФЗ, построенного над одним атомом, работает почти мгновенно, в случае ФЗ над двумя атомами – быстрее в 30 раз, в случае ФЗ с тремя атомами – в 1.5–2 раза. Однако для ФЗ, построенного над 3 атомами, алгоритм приближенной генерации может быть ускорен и дать в этом случае выигрыш по времени. Также было установлено, что разницы между работой приближенного алгоритма, генерирующего 1000 ФЗ, и алгоритма, генерирующего 100 ФЗ, практически нет (в пределах 0.001 единицы для значения из интервальной оценки), а время работы выше на порядок, поэтому в этом случае предпочтительнее использовать генерацию 100 ФЗ.

Дальнейшим направлением для исследований может быть ускорение работы приближенного алгоритма. Помимо этого, на текущий момент описан алгоритм приближенной генерации канонического представителя всей АБС, и последующими исследованиями в данном случае могут послужить оценка такого построения с точки зрения точности. Результаты исследований также позволяют расширить спектр практических задач, в которых возможно применение АБС, и в дальнейшем могут использоваться, например, в сфере оценки защищенности систем от социоинженерных атак [16–17].

СПИСОК ЛИТЕРАТУРЫ

- [1] Ji Q. Probabilistic Graphical Models for Computer Vision. Academic Press, 2019. 299 p.
- [2] Zhang M., Wang N., Li Y., Gao X. Neural probabilistic graphical model for face sketch synthesis // IEEE transactions on neural networks and learning systems. 2019. Vol. 31, № 7. P. 2623-2637.
- [3] Cai M.B., Shvartsman M., Wu A., Zhang H., Zhu X. Incorporating structured assumptions with probabilistic graphical models in fMRI data analysis // Neuropsychologia. 2020. Vol. 144. p. 107500.
- [4] Nguyen T.M., Poh K.L., Chong S.L., Loh S.W., Heng Y.C.K., Lee J.H. The use of probabilistic graphical models in pediatric sepsis: a feasibility and scoping review // Translational Pediatrics. 2023. Vol. 12, № 11. p. 2074.
- [5] He Z., Zhou J. 2020. Inference attacks on genomic data based on probabilistic graphical models // Big Data Mining and Analytics. 2020. Vol. 3, № 3. P. 225-233.
- [6] Kapteyn M.G., Pretorius J.V., Willcox K.E. A probabilistic graphical model foundation for enabling predictive digital twins at scale // Nature Computational Science. 2021. Vol. 1, № 5. P. 337-347.
- [7] Yang J., Yue Z., Yuan Y. Deep probabilistic graphical modeling for robust multivariate time series anomaly detection with missing data // Reliability Engineering & System Safety. 2023. Vol. 238. p. 109410.
- [8] Kharitonov N., Vyatkin A., Tulupyev A. Algebraic Bayesian Networks: The Generation of the Network Canonical Representation // Proceedings of the Seventh International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'23). IITI 2023. Lecture Notes in Networks and Systems. 2023. Vol. 777. P. 13-22. doi: 10.1007/978-3-031-43792-2_2.
- [9] Kharitonov N.A., Tulupyev A.L. Algebraic Bayesian Networks: The Generation of the Knowledge Pattern Canonical Representation // Proceedings of 2021 24th International Conference on Soft Computing and Measurements. SCM 2021. 2021. P. 144–146.
- [10] Тулупьев А.Л. Алгебраические байесовские сети: локальный логико-вероятностный вывод. СПб.: ООО Изд-во «Анатолия», 2007. 80 с.
- [11] Кострикин А.И., Манин Ю.И. Линейная алгебра и геометрия. 2-е изд. М.: Наука, 1986. 304 с.
- [12] Скворцов А.В. Обзор алгоритмов построения триангуляции Делоне // Вычислительные методы и программирование. 2002. Т. 3, № 1. С. 14-39.
- [13] Stein P. A note on the volume of a simplex // The American Mathematical Monthly. 1966. Vol. 73, № 3. P. 299-301.
- [14] Matoušek Jiří. Lectures in Discrete Geometry // Graduate Texts in Mathematics. 2002. Vol. 212. p. 86.
- [15] Vyatkin A.A., Tulupyev A.L. Statistical Comparison of the Running Times of Global Posteriori Inference Algorithms in Algebraic Bayesian Networks // 2023 XXVI International Conference on Soft Computing and Measurements, SCM 2023. St. Petersburg, Russia. 2023. P. 24-28.
- [16] Korepanova A.A., Oliseenko V.D., Abramov M.V. Applicability of Similarity Coefficients in Social Circle Matching // 2020 XXIII. International Conference on Soft Computing and Measurements, SCM 2020. St. Petersburg, Russia. 2020. pp. 41-43.
- [17] Хлобыстова А.О., Абрамов М.В. Адаптация модели многоходовых социоинженерных атак с учётом информационного влияния // Сборник докладов XXIV Международной конференции по мягким вычислениям и измерениям, SCM 2021. М. СПб: СПбГЭТУ «ЛЭТИ». 2021. С. 65–68.