АЛГЕБРАИЧЕСКИЕ БАЙЕСОВСКИЕ СЕТИ: СТАТИСТИЧЕСКИЙ АНАЛИЗ СЛОЖНОСТИ АЛГОРИТМОВ СИНТЕЗА МИНИМАЛЬНОГО ГРАФА СМЕЖНОСТИ И ЕГО КОРРЕКТНОСТЬ

Зотов М.А., студент кафедры системного программирования СПбГУ, стажёр-исследователь СПИИРАН, zotov1994@mail.ru,

Иванова А.В., студентка кафедры системного программирования СПбГУ, стажёр-исследователь СПИИРАН, s.tigma@yandex.ru [1]

Аннотация

В качестве вторичной структуры алгебраических байесовских сетей может выступать минимальный граф смежности: для синтеза такого графа часто используются конкурирующие алгоритмы — прямой, жадный, инкрементальный и декрементальный. Статистические оценки сложности указанных алгоритмов были получены ранее: во время получения таких оценок, вычислялись логарифмы отношений времен работы двух алгоритмов. Вычислительные эксперименты строились на предположении, что логарифмы таких отношений распределены нормально. В одной из опубликованных ранее статей приводились графики полученного распределения, однако, построенные графики имели ряд недостатков: несмотря на то, что полученные эмпирические кривые были схожи с теоретическими, все же оставались вопросы к некой «скачкообразности» эмпирических кривых. В настоящей статье учтены недостатки предыдущего исследования и корректность предположения о нормальном распределении проверена не только с помощью визуального анализа эмпирических и теоретических кривых, но и с помощью формального критерия –

[1] Часть публикуемых материалов получена в рамках проекта, выполненного при финансовой поддержке Российского фонда фундаментальных исследований (грант № 15-01-09001-а).

критерия Шапиро-Уилка. Критерий показал, что предположение действительно верно, и первоначальную гипотезу отвергать нельзя. Также, визуальный анализ графиков подтверждает, что распределение отношений нормальное смещенное. Указанные выкладки проделаны для прямого и жадного, прямого и инкрементального, прямого и декрементального алгоритмов.

Введение

В теории принятия решений, когда эксперту приходится оперировать со значительным числом фактов и вероятностными оценками достоверности тех или иных суждений, человек часто прибегает к помощи ЭВМ. С увеличением вычислительного потенциала ЭВМ расширяется и потенциальная область их применения, которая может затрагивать как вопросы экономического характера — игра на биржах, прогнозирование падений и роста индексов валют, — так и вопросы медицинского характера, касающиеся развития или деградации болезней, эпидемий. С помощью автоматизированных компьютерных программ можно учитывать сотни и тысячи факторов, высказываний, объединяя их определенными логическими связками, строить выводы о поведении тех или иных сложных объектов и, наконец, анализировать результаты агрегации входных данных, делая вывод о поведении системы в целом. Такие входные данные часто называют *знаниями или фрагментами знаний (ФЗ)*. Соотношения и связи между ФЗ образуют, в свою очередь, *сеть фрагментов знаний*. Существует множество структур данных, которые позволяют хранить и обрабатывать ФЗ, к ним относятся, например, массивы, деревья, нейросети и другие.

В теории машинного обучения и интеллектуальных системах, в качестве структуры, позволяющей хранить, обрабатывать и впоследствии отображать данные, часто рассматривают байесовские сети доверия (БСД), введенные J. Pearl[16, 18, 19]. Байесовские сети доверия находят широкое применение в отрасли логистики при построении цепи поставок, системах принятия решений, прогнозировании (например, метеорологических прогнозах) и многих других отраслях науки и жизни. Вместе с тем, байесовские сети доверия, которые относятся к классу вероятностно-графических моделей[9], работают со скалярными значениями вероятностей в каждом узле построенной сети[13, 11, 14], а в качестве значения узла они содержат фрагменты знаний[19, 10]. Другими словами, БСД формализуют данные о предметной области и

позволяют сопоставлять их с вероятностью их выполнения.

В качестве так называемой вторичной структуры в АБС могут выступать *графы смежности*[12, 6, 8, 15] и минимальные графы смежности (МГС), как частный случай. В работах[7, 6] были предложены прямой и жадный алгоритмы синтеза МГС, в работах [5, 17] — инкрементальный и декрементальный соответственно. Для конкурирующих алгоритмов были предложены относительные статистические оценки сложностей[3, 5, 17], т.к. формальные оценки оказались неприменимыми на практике, т.к. базировались на латентных свойствах графов смежности, которые становятся известными только после синтеза самой структуры МГС.

Обоснование корректности проведения вычислительных экспериментов

Для получения некоторых относительных статистических оценок сложности, было сделано предположение о том, что логарифмы отношений времени работы конкурирующих алгоритмов распределены нормально. Для подтверждения этого необходимо также провести некоторые вычисления, которые смогут подтвердить изначальное предположение. В [1] была показана корректность такого предположения, однако корректность была показана лишь визуально — никаких численных методов проверки на нормальность не было применено.

Для более корректной проверки, необходимо построить гистограмму частоты повторения статистик для отношений времени работы алгоритмов, а также построить аналогичную гистограмму для *логарифмированных* относительных времен работы алгоритмов. Далее необходимо построить кривую нормального распределения. После получения необходимых графиков нужно будет наложить друг на друга кривую нормального распределения и гистограмму, полученную для логарифмированных отношений времени работы конкурирующих алгоритмов.

Построение частотных гистограмм

Как было сказано ранее, при вычислении таких статистик, как верхняя и нижняя границы доверительного интервала, делалось предположение, что логарифмы отношений времени работы конкурирующих алгоритмов подчиняются нормальному распределению. Для того, чтобы

убедиться, что это так, можно использовать критерии проверки данных на нормальность, кроме того, логарифмированные данные можно визуализировать в виде гистограммы, дополнительно указав кривую эмпирического и теоретического распределений.

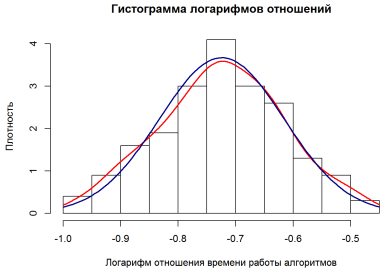
Визуализация распределений

В качестве конкурирующих алгоритмов были выбраны прямой и жадный алгоритмы. С помощью инструментов языка R[21] и визуальной среды разработки *RStudio*[20] были построены гистограммы для выходных данных алгоритма, синтезирующего на выход логарифмы отношений времени работы конкурирующих алгоритмов синтеза МГС для графов мощностью 5, 20, 35, 50, 65. Гистограммы, соответствующая кривая эмпирического (красная кривая), а также теоретическая кривая (синяя кривая), представлены на Рис. 1a — 2e.

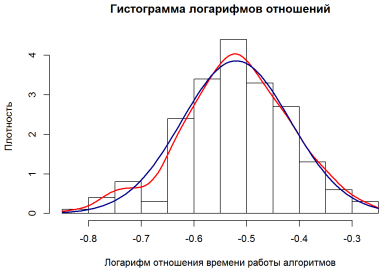
Проверка распределения с помощью критерия Шапиро-Уилка. Критерий Шапиро-Уилка[4] используется для проверки выборки на нормальность. При работе с критерием выделяют две гипотезы: H_0 — *случайная величина X распределена нормально,* H_1 — *случайная величина X не подчиняется нормальному распределению.* Для проверки гипотез, воспользуемся встроенным в R критерием Шапиро-Уилка. Результаты работы критерия подписаны под каждым из Рис. 1a — 2e.

Анализ рисунков позволяет сделать вывод о том, что, несмотря на некоторые скачки и «неровности» в эмпирических кривых, они действительно похожи на кривые нормального распределения — они имеют схожий вид и поведение. С другой стороны, стоит отметить, что нередки случаи выброса значений в эмпирическом распределении — это объясняется тем, что, возможно, стоило выбрать более мелкий шаг для построения гистограммы или увеличить число экспериментов для достижения большей плавности в графике. Однако в целом, результаты вычислений подтверждают корректность предположения о том, что логарифмы указанных отношений имеют нормальное распределение. Другой актуальной задачей является приведение статистик к нормальному виду, базируясь на других способах и критериях такого преобразования [2].

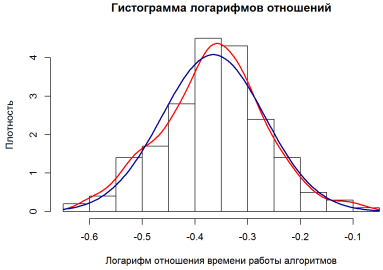
Что касается формального критерия проверки выборки на нормальность, то при полученных значениях *p-value*, по критерию Шапиро-Уилка, нельзя отвергнуть гипотезу H_0: гипотеза H_0 отвергается только при значении $p\text{-}value < 0.05$, что соответствует степени доверия в 95%.



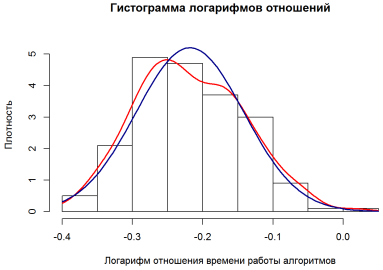
(a) 20 вершин, $p\text{-}value = 0.6212$



(b) 35 вершин, $p\text{-}value = 0.4905$

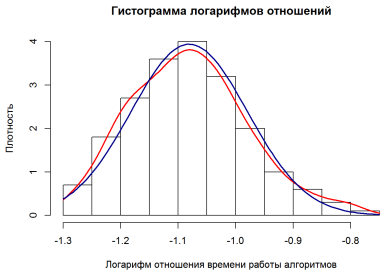


(c) 50 вершин, $p\text{-}value = 0.5937$

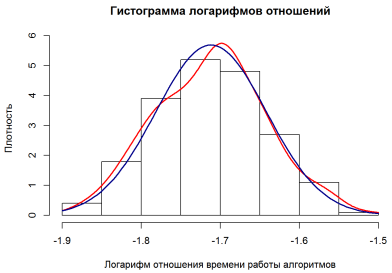


(d) 65 вершин, $p\text{-}value = 0.3853$

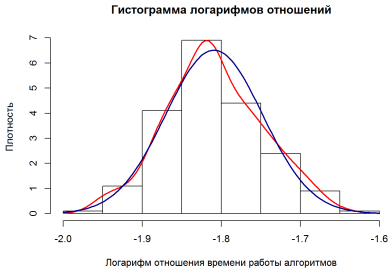
Рис. 1: 49% вершин — 2–4 порядков, 25% — 5–7 порядков, 15% — 8–10 порядков, 11% — 11–13 порядков



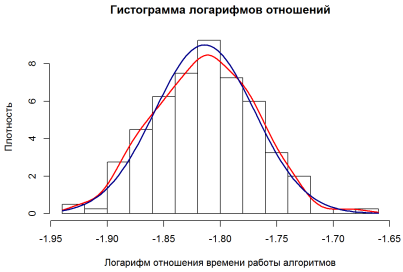
(a) 5 вершин, $p\text{-}value = 0.06428$



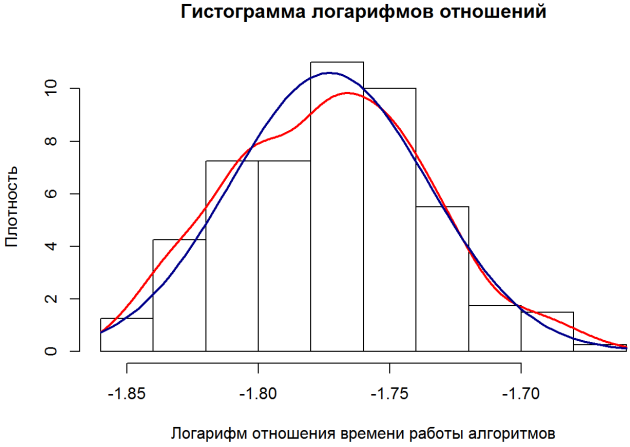
(b) 20 вершин, $p\text{-}value = 0.8403$



(c) 35 вершин, $p\text{-}value = 0.4415$



(d) 50 вершин, $p\text{-}value = 0.812$



(e) 65 вершин, $p\text{-}value = 0.365$

Рис. 2: 0% вершин — 2–4 порядков, 3% — 5–7 порядков, 7% — 8–10 порядков, 90% — 11–13 порядков

Резюмируя, отметим, что предположение о том, что выборка логарифмов отношений времени работы конкурирующих алгоритмов имеет нормальное распределение, верно: на это указывают как графики и кривые распределений, так и численные критерии.

Заключение

Показана корректность предположения о том, что логарифмы отношений времени работы имеют нормальное распределение. Теоретическая кривая и эмпирическая кривая представлены на одном графике, а также проведен тест Шапиро-Уилка, который показал, что нельзя отвергнуть гипотезу о том, что полученные после логарифмирования данные, распределены нормально.

Литература

[1] Зотов М.А. Синтез вторичной структуры алгебраических байесовских сетей: визуализация распределений относительных сложностей прямого и жадного алгоритмов // СПИСОК-2016: Материалы всероссийской научной конференции по проблемам информатики. Санкт-Петербург, 2016. С. 501–509.

[2] Зотов М.А., Левенец Д.Г., Иванова А.В., Тулупьев А.Л. Статистическая сложность синтеза вторичной структуры алгебраических байесовских сетей: двухфакторный анализ // Гибридные и синергетические интеллектуальные системы. (Светлогорск, 6-11 июня 2016 г.). Светлогорск: Балтийского федерального университета им. Иммануила Канта, 2016. С. 405-411.

[3] Зотов М.А., Тулупьев А.Л. Синтез вторичной структуры алгебраических байесовских сетей: методика статистической оценки сложности и компаративный анализ прямого и жадного алгоритмов // Компьютерные инструменты в образовании, 2015. № 1. С. 3–16.

[4] Критерий Шапиро-Уилка [Electronic resource] // machinelearning.ru. URL: http://www.machinelearning.ru/wiki/index.php?title=Критерий_Шапиро-Уилка (accessed at 23.04.2017).

[5] Левенец Д.Г., Зотов М.А., Тулупьев А.Л. Инкрементальный алгоритм синтеза минимального графа смежности // Компьютерные инструменты в образовании. 2015. Вып. 6. С. 3–18.

[6] Опарин В.В., Тулупьев А.Л. Синтез графа смежности с минимальным числом ребер: формализация алгоритма и анализ его корректности. // Тр. СПИИРАН. 2009. №11. С. 142–157.

[7] Опарин В.В., Фильченков А.А., Сироткин А.В., Тулупьев А.Л. Матроидное представление семейства графов смежности над набором фрагментов знаний // Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. 2010. №4(68). С. 73–76.

[8] Тулупьев А.Л. Алгебраические байесовские сети: глобальный логико-вероятностный вывод в деревьях смежности: Учеб. пособие. СПб.: СПбГУ; ООО Издательство «Анатолия», 2007. 40 с. (Сер. Элементы мягких вычислений.)

[9] Тулупьев А.Л. Алгебраические байесовские сети: логико-вероятностная графическая модель баз фрагментов знаний с неопределенностью. Санкт-Петербургский институт информатики и автоматизации РАН, — Изд-во С.-Петерб. ун-та, 2009.

[10] Тулупьев А.Л. Алгебраические байесовские сети: логико-вероятностный подход к моделированию баз знаний с неопределенностью. // СПб.: СПИИРАН, 2000. С. 282.

[11] Тулупьев А.Л. Байесовские сети: логико-вероятностный вывод в циклах. СПб.: Изд-во С.-Петербургского ун-та, 2008. 140 с. (Элементы мягких вычислений.)

[12] Тулупьев А.Л. Дерево смежности с идеалами конъюнктов как ациклическая алгебраическая байесовская сеть // Тр. СПИИРАН. Вып. 3, т. 1. СПб.: Наука, 2006. С. 198–227.

[13] Тулупьев А.Л., Николенко С.И., Сироткин А.В. Байесовские сети: логико-вероятностный подход. // СПб.: Наука, 2006. С. 607.

[14] Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. СПб.: Изд-во С.-Петерб. ун-та, 2009. С. 400.

[15] Фильченков А.А. Синтез графов смежности в машинном обучении глобальных структур алгебраических байесовских сетей. Дисс.... к-та физ.-мат. н. Самара, 2013. С. 339. (Самарск. гос. аэрокосм. ун-т им. ак. С.П. Королева (нац. исслед.))

[16] Korb K.B., Nicholson A.E. Bayesian Artificial Intelligence. NY.: Chapman and Hall/CRC, 2004. 364 p.

[17] Levenets D.G., Zotov M.A., Romanov A.V., Tulupyev A.L., Zolotin A.A., Filchenkov A.A. Decremental and Incremental Reshaping of Algebraic Bayesian Networks Global Structures // Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'16). T. 2. Sochi: Springer, 2016. P. 57–67.

[18] Neapolitan R. E. Learning Bayesian Networks. Pearson Prentice Hall, 2003. 674 p.

[19] Perl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. NY etc.: Morgan Kaufmann Publ., 1994. P. 552.

[20] RStudio – Open source and enterprise-ready professional software for R [Electronic resource] // r-project.org. URL: https://www.rstudio.com/ (accessed at 23.04.2017).

[21] The R Project for Statistical Computing [Electronic resource] // r-project.org. URL: https://www.r-project.org/ (accessed at 23.04.2017).