

УДК 004.413

Д.О. Змеев, О.А. Змеев, Л.С. Иванова, В.И. Фрейдин

Разработка подсистемы работы с байесовскими сетями для системы поддержки принятия решений в управлении проектами по разработке программного обеспечения

В настоящий момент байесовские сети являются популярным методом для решения различных задач, в том числе для создания инструментов поддержки принятия решений. Предлагается решение для создания байесовской сети и прямого вероятностного вывода. Спецификой задачи является работа с крупными сетями (более 200 вершин) с большим количеством вершин-родителей у одной вершины (15 и более). Инструмент интегрирован с Redmine и позволяет рассчитать вероятность ошибки менеджера при определении текущего состояния проекта.

Ключевые слова: байесовская сеть, SEMAT, Essence, графовая СУБД.

DOI: 10.21293/1818-0442-2022-25-3-52-56

В области управления проектами по разработке программного обеспечения достаточно сложно найти и использовать практически применимую прикладную математическую или алгоритмическую модель, которая позволила бы управлять проектами с использованием структурированных и(или) оптимизационных механизмов. Частично это связано с двойкой природой проектов по разработке программного обеспечения, в которой существуют, с одной стороны, достаточно строгий технологический процесс, а с другой стороны, творческая природа поиска наиболее подходящей конфигурации программного обеспечения для нужд стейкхолдеров.

В попытках разработать подход, который учитывал бы эту противоречивую природу, но не требовал от команд по разработке программного обеспечения существенных ресурсных затрат на поддержку, авторы статьи разрабатывают комплексное решение в форме системы поддержки принятия решений, которое построено на стандарте OMG для языка Essence [1]. Данный язык содержит структурированное описание абстрактного состояния прогресса проекта, которое впоследствии можно использовать для построения прикладных математических моделей, в частности, байесовских сетей [2].

Вершины байесовской сети соответствуют чекпоинтам (контрольным точкам) проекта, которые фиксируют истинность или ложность утверждений, соответствующих семантическому описанию достигнутого в проекте прогресса, например «Key technical risks agreed to» (ключевые технические риски согласованы). С помощью экспертной оценки определены зависимости между чекпоинтами, которые трансформируются в рёбра байесовской сети. Результатом работы системы является вероятность ошибки менеджера – вероятность ложноположительной ошибки менеджера относительно истинности утверждения о состоянии проекта. Для использования предложенной модели на практике возникла потребность в инструменте, позволяющем моделировать сеть и производить расчеты.

В результате разработки первой версии исследовательского прототипа была построена байесов-

ская сеть, включающая 646 вершин. Наибольшее число вершин-родителей для одной вершины составляет 15. Структура сети, а также значения вероятностных векторов были описаны в JSON-файле и импортировались в плагин для Redmine [3] при подготовке к вычислению прямого вероятностного вывода. Для продолжения исследований необходимо добавить возможность учета практик, используемых командой при разработке ПО, в связи с чем разработанное решение перестало удовлетворять требованиям. Возникла задача поиска нового решения.

Необходимо отметить, что спецификой задачи является отсутствие выборки данных, которую можно было бы использовать для статистических (например, непараметрический регрессионный анализ) или других, более классических подходов машинного обучения. Данный факт обычно является существенным ограничением применимости подходов, не основанных на строгом представлении проекта (например, сетевых планов графиков проекта). Однако байесовские сети как инструмент допускают использование экспертных оценок вместо или вместе с обучением на основе данных. Также в качестве дополнительных требований, следующих из общей проблематики задачи, можно выделить следующие:

1. Поддержка дискретной байесовской сети с большим количеством вершин (более 1000). Базовая сеть для чекпоинтов альфы Software System содержит 646 вершин, однако дальнейшие планы по использованию больших возможностей стандарта OMG Essence подразумевают существенный рост числа вершин в сети.

2. Поддержка работы с большим количеством у конкретной вершины сети вершин-родителей (более 20 вершин-родителей) с адекватным временем расчета вероятностей. Наибольшее число родителей у вершины базовой сети – 15, добавление новых практик влечет за собой появление новых зависимостей, т.е. увеличение максимально возможного числа вершин-родителей. Данное требование влияет и на модель хранения информации о сети, используемой инструментом, поскольку, при увеличении числа вершин-родителей соответствующий вершине веро-

ятностный вектор растет очень быстро (его длина равна 2^n , где n – число вершин-родителей), что отрицательно влияет на размер файла с данными о сети и затрудняет работу с ним.

3. Возможность быстрого перестроения сети с добавлением новых вершин. Поскольку важным функциональным требованием разрабатываемой системы является возможность учета дополнительных практик, используемых командой в процессе разработки ПО, каждая новая практика и проводимые в ее рамках активности могут быть причиной добавления новых вершин и ребер, что требует перестроения байесовской сети.

4. Возможность интеграции с внешней системой. Разработка инструмента поддержки принятия решений осуществляется в форме плагина для системы управления проектами Redmine. Используемый для построения сети инструмент должен иметь возможность обмена данными с внешней системой – плагином для Redmine.

5. Наличие бесплатной версии. Поскольку данный проект носит исследовательский характер, а также из-за отсутствия возможности и разумного обоснования для оплаты лицензии на ПО без уверенности в том, что рассматриваемый инструмент удовлетворяет критериям, вопрос наличия бесплатной версии был ключевым.

Наличие графического редактора и/или инструмента визуализации сети не является необходимым требованием, но существенно облегчает работу с сетью.

Целью представленного в данной работе исследования является создание инструмента для работы с байесовской сетью в рамках задачи определения вероятности ошибки менеджера проекта по разработке программного обеспечения.

При создании инструмента были решены следующие задачи:

1. Выбор способа хранения информации о структуре байесовской сети.

2. Реализация взаимодействия между Redmine и инструментом.

Ниже представлено сравнение различных инструментов по работе с байесовскими сетями, дана информация по хранению данных о структуре байесовской сети с помощью графовой СУБД, описана общая архитектура разработанного инструмента. А также представлены результаты реализации и характеристики разработанного инструмента.

Сравнительный анализ инструментов по работе с байесовскими сетями

На основе вышеуказанных требований были сформулированы критерии сравнения инструментов для работы с байесовскими сетями: построение крупных сетей (более 1000 вершин), быстрое перестроение сети, возможность интеграции с внешней системой, возможность расчета прямого вероятностного вывода для вершины с 20 родительскими вершинами, возможность импорта созданной ранее сети, наличие бесплатной или условно-бесплатной версии.

В табл. 1 приведены результаты анализа наиболее популярных на момент написания статьи инструментов.

Помимо указанных в табл. 1 критериев, также рассматривалась возможность вычисления прямого вероятностного вывода для узла с 20 родительскими вершинами, но ни один из рассмотренных инструментов не предоставляет такой функционал.

Авторы статьи [11] также столкнулись с необходимостью создания байесовской сети большого размера (десятки тысяч вершин).

Таблица 1

Сравнение инструментов для работы с байесовскими сетями

Критерий	Инструменты						
	AgenaRisk [4]	Bayes Fusion [5]	Netica [6]	BayesiaLab [7]	BayesAML [8]	MSBNx [9]	Bayes Server [10]
Построение крупных сетей (более 1000 вершин)	+	+	+(в платной версии)	+(в платной версии)	+	+	+
Интеграция с внешней системой	+	–	–	+	+	–	+
Импорт сети	+(в платной версии)	+	+(не более 15 вершин для бесплатной версии)	+(не более 15 вершин для бесплатной версии)	+	+	+
Стоимость	Desktop £2,100 Developer £3,100	Бесплатно	Commercial \$685 Educational / Personal \$285	Professional \$42,100 (в год) Academic Edition \$700 (в год)	\$2070	Бесплатно	£675 (в год)

В результате проведенного ими анализа было принято решение разработать собственный инструмент [12], главными особенностями которого являются возможность создания и обучения крупных байесовских сетей и ориентация на работу с геномными данными. В табл. 1 данный инструмент не вошел,

так как анализ инструмента показал невозможность его применения для поставленной задачи, поскольку для его использования требуется набор обучающих данных.

На основании анализа можно сделать вывод о необходимости создания инструмента, удовлетво-

ряющего вышеперечисленным требованиям. Отметим, что основной задачей является построение, сохранение данных о сети и расчет вероятностей. Исходными данными для построения сети являются стандарт OMG Essence и экспертные оценки о наличии зависимостей между утверждениями, полученные коллективом разработчиков и проектных менеджеров компаний «Education ERP», «Kreosoft» и «Сибирские Информационные Системы».

Хранение данных о структуре байесовской сети

Поскольку байесовская сеть является направленным ациклическим графом, для хранения информации о ее структуре можно использовать графовые базы данных. Графовые СУБД существенно превосходят другие виды СУБД по производительности при выполнении операций создания, сохранения и внесения изменений в схему базы данных для задач с естественной графовой структурой данных [13]. В качестве СУБД был выбран инструмент Neo4j Graph Database [14], поскольку он предоставляет удобное API для взаимодействия с БД, имеет инструменты визуализации, обширную документацию, а также доступен для бесплатного использования.

Каждая вершина байесовской сети транслируется в вершину графа в БД, зависимости между вершинами сети аналогично становятся зависимостями между вершинами графа. Поскольку расчет вероятности с помощью прямого вероятностного вывода осуществляется для каждой итерации проекта, необходимо сохранять значение вероятности на предыдущей итерации. Для этого в БД планируется создание суррогатных вершин, связанных со всеми остальными вершинами, соответствующими вершинам байесовской сети. Вес ребра между суррогатной вершиной и конкретной вершиной графа будет равен значению вероятности для конкретной вершины на предыдущей итерации.

Каждая вершина графа в БД содержит следующую информацию: идентификатор, имя вершины, перечень родителей вершины в виде строки, соответствующий вероятностный вектор в виде строки.

Архитектура системы

Выбор способа хранения информации о структуре байесовской сети повлиял на архитектуру всей системы, поскольку использование СУБД требовало реализации ряда компонентов: компонента взаимодействия с СУБД для создания, изменения и получения данных о структуре сети, компонента взаимодействия с Redmine, компонента для математических расчетов. Разработанные компоненты и их зависимости представлены на рис. 1.

Основная логика, в том числе алгоритм подсчета вероятностей, реализованы на языке Python 3.8. BaseServer предоставляет API для взаимодействия с Redmine, MasterProbability содержит основной алгоритм расчета прямого вероятностного вывода, а модули pyNeoByse и CreateDB отвечают за взаимодействие с СУБД Neo4j. Информация об экспертных оценках, необходимая для расчета вероятностей, представлена в виде JSON-файла.

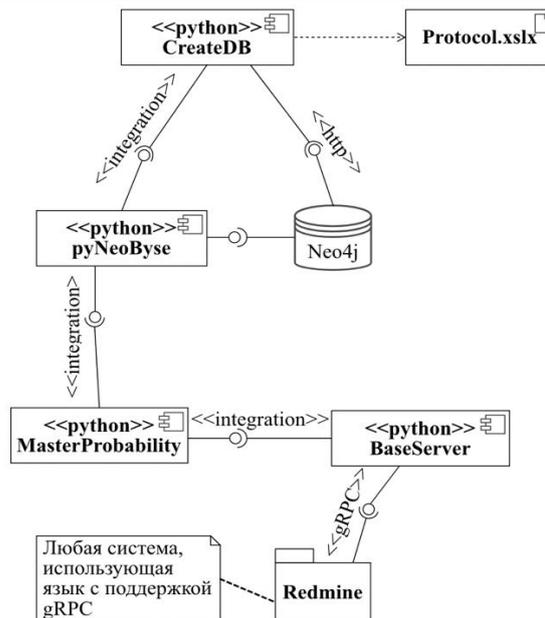


Рис. 1. Компоненты системы

На рис. 2 представлен алгоритм работы инструмента. Байесовская сеть конфигурируется заранее на основе данных об альфах проекта и их чекпоинтах. При запуске расчетов разработанный авторами плагин для Redmine осуществляет сбор информации о расставленных менеджером чекпоинтах альф проекта и отправляет на сервер с помощью API.

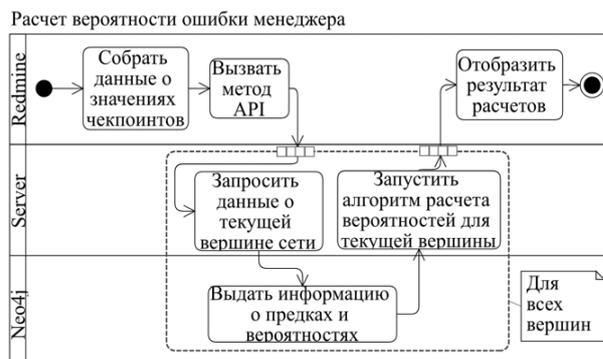


Рис. 2. Взаимодействие частей системы

Полученные результаты

В результате с помощью разработанного инструмента была смоделирована сеть с 646 вершинами, 1783 связями (наибольшее число родителей для одной вершины – 15). Для сети с такими параметрами вычисление прямого вероятностного вывода для одной итерации осуществляется за 6,4 с (запуск осуществлялся на оборудовании со следующими характеристиками: процессор Intel(R) Core(TM) i7-9750H, 8 Gb RAM DDR4). Файл со структурированной байесовской сетью и подсчитанными вероятностями доступен по ссылке [15].

В табл. 2 представлен пример результатов вычислений прямого вероятностного вывода, полученный с помощью разработанного инструмента. Подробный разбор алгоритма вычисления вероятностей представлен в работе [16]. Идентификаторы

чекпоинтов формируются по схеме «первая буква названия альфы, номер состояния, номер чекпоинта». На нулевой итерации вероятность ошибки считается нулевой.

Т а б л и ц а 2

Результаты работы алгоритма – вероятности истинности утверждений о состоянии проекта (фрагмент)

Вершина сети	Итерация 0	Итерация 1
O11	0,0	0,0
O12	0,0	0,0422
O13	0,0	0,2202
O21	0,0	0,6387
O22	0,0	0,2554
O23	0,0	0,5233
O24	0,0	0,68
O25	0,0	0,6935
O31	0,0	0,0457
O32	0,0	0,3261
O33	0,0	0,2774
O34	0,0	0,1207
O35	0,0	0,1207
O41	0,0	0,0808
O42	0,0	0,1162
O43	0,0	0,1255
O44	0,0	0,0829

Выводы

Применение графовой СУБД убирает ограничения на количество вершин в сети, позволяя инструменту удовлетворять требованию № 1. Также снимается ограничение на количество вершин-родителей (с точки зрения проблемы хранения) и выполняется требование о возможности быстрого перестроения сети. Требование об интеграции удовлетворено с помощью API и библиотек для взаимодействия с Neo4j, предоставляемых разработчиком СУБД.

Дальнейшие исследования в области создания инструмента поддержки принятия решений для менеджеров проектов требуют существенного расширения байесовской сети. В текущей версии системы узким местом является вычислительная часть: получение вероятностного вектора и вычисление условной вероятности являются достаточно трудоемкими по времени. Следовательно, возникает задача оптимизации вычислений. Кроме того, с ростом числа вершин сети и количества связей между вершинами неизбежно произойдет увеличение размера вероятностных векторов, что существенно затрудняет их хранение. Возникает вторая задача – решить вопрос хранения вероятностных векторов. Обе задачи являются темой будущей работы авторов статьи.

Литература

1. The Essence of Software Engineering: Applying the SEMAT Kernel / I. Jacobson, P.-W. Ng, P.E. McMahon, I. Spence, S. Lidman. – Addison-Wesley, 2013. – 224 p.
2. Змеев Д.О. О представлении прогресса проекта по разработке программного обеспечения в форме динамической байесовской сети / Д.О. Змеев, Л.С. Иванова, Р.Р. Рафикова // Информационные технологии и математическое моделирование (ИТММ–2020): матер. XIX Междунар. конф. им. А.Ф. Терпугова. – Томск, 2021. – С. 291–297.

3. Redmine [Электронный ресурс]. – Режим доступа: <https://www.redmine.org/>, свободный (дата обращения: 10.12.2021).

4. AgenaRisk: Bayesian Network Software for Risk Analysis and Decision Making [Электронный ресурс]. – Режим доступа: <https://www.agenarisk.com>, свободный (дата обращения: 02.02.2022).

5. BayesFusion, LLC. Data Analytics, Modeling, Decision Support [Электронный ресурс]. – Режим доступа: <https://www.bayesfusion.com>, свободный (дата обращения: 03.02.2022).

6. Netica Application [Электронный ресурс]. – Режим доступа: <https://www.norsys.com/netica.html>, свободный (дата обращения: 03.03.2022).

7. BayesiaLab 10 [Электронный ресурс]. – Режим доступа: <https://www.bayesia.com>, свободный (дата обращения: 10.02.2022).

8. BayesAML [Электронный ресурс]. – Режим доступа: <https://www.hugin.com/bayesaml/>, свободный (дата обращения: 10.02.2022).

9. MSBNx. Bayesian Network Editor and ToolKit [Электронный ресурс]. – Режим доступа: <https://msbnx.azure-websites.net>, свободный (дата обращения: 15.02.2022).

10. Bayes Server [Электронный ресурс]. – Режим доступа: <https://www.bayesserver.com>, свободный (дата обращения: 15.02.2022).

11. Michiels M. BayeSuites: An open web framework for massive Bayesian networks focused on neuroscience / M. Michiels, P. Larrañaga, C. Bielza // Neurocomputing. – 2021. – Vol. 428. – P. 166–181.

12. Bayesian networks – BayeSuites [Электронный ресурс]. – Режим доступа: https://neurosuites.com/morpho/ml_bayesian_networks, свободный (дата обращения: 16.02.2022).

13. Robinson I. Graph Databases / I. Robinson, J. Webber, E. Eifrem. – O’Reilly Media, 2013. – 178 p.

14. Neo4j Graph Database [Электронный ресурс]. – Режим доступа: <https://neo4j.com/product/neo4j-graph-database/>, свободный (дата обращения: 15.02.2022).

15. Файл со структурированной байесовской сетью и подсчитанными вероятностями [Электронный ресурс]. – Режим доступа: https://drive.google.com/file/d/1ez6L5EybuUUpW8FTv_n3XD9ziwPMfXXAu/view, свободный (дата обращения: 16.05.2021).

16. Змеев Д.О. Прототип системы поддержки принятия решений для управления проектами на основе стандарта OMG Essence и байесовских сетей: дис. ... канд. техн. наук: 05.13.11 / Томский гос. ун-т (Томск), 2022 [Электронный ресурс]. – Режим доступа: <http://vital.lib.tsu.ru/vital/access/manager/Repository/koha:000894552>, свободный (дата обращения: 05.07.2022).

Змеев Денис Олегович

Ассистент каф. программной инженерии
Национального исследовательского
Томского государственного ун-та (НИ ТГУ)
Ленина пр-т, 36, корп. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 52-97-93
Эл. почта: denis.zmееv@accounts.tsu.ru

Змеев Олег Алексеевич

Д-р физ.-мат. наук, профессор каф. программной инженерии НИ ТГУ
Ленина пр-т, 36, корп. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 52-97-93
Эл. почта: ozmeyev@gmail.com

Иванова Лидия Сергеевна

Ассистент каф. программной инженерии НИ ТГУ
Ленина пр-т, 36, корп. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 52-97-93
Эл. почта: lidiya.ivanova@persona.tsu.ru

Фрейдин Валерий Игоревич

Студент высшей ИТ-школы НИ ТГУ
Ленина пр-т, 36, корп. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 52-97-93
Эл. почта: www.carat@list.ru

Zmееv D.O., Zmееv O.A., Ivanova L.S., Freydin V.I.
**Development of a subsystem to use Bayesian networks
in a decision support system for software development
management**

Bayesian networks are currently a popular tool for solving various problems, including creating decision support systems. This paper proposes a tool for creating a Bayesian network and direct probabilistic inference. The specificity of the problem consists in working with large networks (more than 1000 nodes) with a large number of parent nodes at one node (15 or more). The tool is integrated with Redmine and allows you to calculate the probability of a manager's error when determining the current state of the project.

Keywords: Bayesian network, SEMAT, Essence, graph DBMS.

DOI: 10.21293/1818-0442-2022-25-3-52-56

References

1. Jacobson I., Ng P.-W., McMahon P.E., Spence I., Lidman S. *The Essence of Software Engineering: Applying the SEMAT Kernel*. Addison-Wesley, 2013. 224 p.
2. Zmееv D.O., Ivanova L.S., Rafikova R.R. *O predstavlenii progressa proekta po razrabotke programmnogo obespechenija v forme dinamicheskoy bajesovskoj seti* [Presenting the progress of a software development project in the form of a dynamic Bayesian network]. *Informacionnye tekhnologii i matematicheskoe modelirovanie (ITMM–2020)*. Materialy XIX Mezhdunarodnoj konferencii imeni A.F. Terpugova [Information technologies and mathematical modeling (ITMM–2020). Proceedings of the IX International Conference named after A.F. Terpugov]. Tomsk, 2021, pp. 291–297 (in Russ.).
3. Redmine. Available at: <https://www.redmine.org/>, free. (Accessed: December 10, 2021).
4. AgenaRisk: Bayesian Network Software for Risk Analysis and Decision Making. Available at: <https://www.agenarisk.com>, free (Accessed: February 02, 2022).
5. BayesFusion, LLC. Data Analytics, Modeling, Decision Support. Available at: <https://www.bayesfusion.com>, free (Accessed: February 03, 2022).
6. Netica Application. Available at: <https://www.nor-sys.com/netica.html>, free (Accessed: March 03, 2022).
7. BayesiaLab 10. Available at: <https://www.bayesia.com>, free (Accessed: February 10, 2022).

8. BayesAML. Available at: <https://www.hugin.com/bayesaml/>, free (Accessed: February 10, 2022).

9. MSBNx. Bayesian Network Editor and ToolKit. Available at: <https://msbnx.azurewebsites.net>, free (Accessed: February 15, 2022).

10. Bayes Server. Available at: <https://www.bayesserver.com>, free. (Accessed: February 15, 2022).

11. Michiels M., Larrañaga P., Bielza C. BayeSuites: An open web framework for massive Bayesian networks focused on neuroscience. *Neurocomputing*, 2021, vol. 428, pp. 166–181.

12. Bayesian networks – BayeSuites. Available at: https://neurosuites.com/morpho/ml_bayesian_networks, free (Accessed: February 16, 2022).

13. Robinson I., Webber J., Eifrem E. *Graph Databases*. O'Reilly Media, 2013, 178 p.

14. Neo4j Graph Database. Available at: <https://neo4j.com/product/neo4j-graph-database/>, free (Accessed: February 15, 2022).

15. Structured Bayesian network and probabilities file. Available at: https://drive.google.com/file/d/1ez6L5EyBUUpW8FTv_n3XD9ziwPMfXXAu/view, free. (Accessed: May 16, 2021).

16. Zmееv D.O. *Prototip sistemy podderzhki prinjatija reshenij dlja upravlenija proektami na osnove standartov OMG Essence i bajesovskih setej* [Decision support system prototype for project management based on OMG Essence Standard and Bayesian networks. Cand. Diss.]. Tomsk, 2022. Available at: <http://vital.lib.tsu.ru/vital/access/manager/Repository/koha:000894552>, free (Accessed: July 5, 2022) (in Russ.).

Denis O. Zmееv

Assistant, Department of Software Engineering,
National Research Tomsk State University (NR TSU)
36, Lenin pr., Bldg. 2, Tomsk, Russia, 634050
Phone: +7 (382-2) 52-97-93
Email: denis.zmееv@accounts.tsu.ru

Oleg A. Zmееv

Doctor of Science in Physics and Mathematics, Professor, NR TSU
36, Lenin pr., Bldg. 2, Tomsk, Russia, 634050
Phone: +7 (382-2) 52-97-93
Email: ozmееv@gmail.com

Lidiya S. Ivanova

Assistant,
Department of Software Engineering, NR TSU
36, Lenin pr., Bldg. 2, Tomsk, Russia, 634050
Phone: +7 (382-2) 52-97-93
Email: lidiya.ivanova@persona.tsu.ru

Valery I. Freydin

Student,
Higher IT-School, NR TSU
36, Lenin pr., Bldg. 2, Tomsk, Russia, 634050
Phone: +7 (382-2) 52-97-93
Email: www.carat@list.ru