INFORMATION SYSTEMS ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 65.015

В.Н. Наумов, П.В. Наумов

ПРОГНОЗИРОВАНИЕ РИСКОВ ПРОЕКТИРОВАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Наумов Владимир Николаевич, доктор военных наук, профессор, заслуженный деятель науки РФ, окончил факультет электронной вычислительной техники Высшего военно-морского училища радиоэлектроники им. А.С. Попова. Заведующий кафедрой бизнес-информатики Северо-Западного института управления – филиала Российской Академии народного хозяйства и государственной службы при Президенте Российской Федерации, г. Санкт-Петербург. Имеет статьи, монографии, учебники в области системного анализа информационных систем, анализа данных и машинного обучения. [e-mail: naumov122@list.ru].

Наумов Павел Владимирович, окончил Санкт-Петербургский государственный политехнический университет Петра Великого. Главный специалист разработки организационно-распорядительной документации Центра технологии судостроения и судоремонта. Имеет статьи в области менеджмента рисков. [e-mail: naumov@sstc.spb.ru].

Аннотация

В статье рассмотрена задача прогнозирования рисков проектирования информационных систем на основе использования методов машинного обучения. Проанализированный ансамбль методов классификации позволил оценить качество решения задачи с помощью комплекса показателей и выбрать лучший по их значениям. Для ее решения применены язык R, статистическая платформа SPSS, а также система Orange Canvas. Машинное обучение основано на использовании синтетической обучающей выборки, которая включает категориальные факторы возможных рисков проектирования. Выборка сформирована и проверена с помощью методов кластерного анализа и метода главных компонент.

Ключевые слова: прогнозирование, проектирование информационных систем, риски, машинное обучение, кластерный анализ, метод главных компонент, методы классификации, показатели качества классификации.

doi:

FORECASTING THE RISKS OF INFORMATION SYSTEM DESIGN

Vladimir Nikolaevich Naumov, Doctor of Military Sciences, Professor, Honored Scientist of the Russian Federation; graduated from the Faculty of Electronic Computer Equipment of the A.S. Popov Higher Naval School of Radioengineering; Head of the Business-Informatics Department of the North-West Institute of Management, the Branch of the Russian Presidential Academy of National Economy and Public Administration in St. Petersburg; an author of articles, monographs, tutorials in the field of system-oriented analysis of information systems, data analysis and machine learning. e-mail: naumov122@list.ru.

Pavel Vladimirovich Naumov, graduated from Peter the Great St. Petersburg Polytechnic University; Chief Developer of organizational and executive documentation at Shipbuilding&Shiprepair Technology Center; an author of articles in the field of risk management. e-mail: naumov@sstc.spb.ru.

Abstract

The article considers the problem of forecasting the risks of designing information systems based on the use of machine learning methods. The ensemble of methods allows us to estimate the quality of the solution of a problem using a set of indicators and select the best one. To solve it, the language R, the statistical platform SPSS, and the Orange Canvas system are used. Machine learning is based on the use of a synthetic training set, which includes categorical factors of possible design risks. The sample is formed and verified using the methods of cluster analysis and the principal component method.

Key words: forecasting, information systems design, risks, machine learning, cluster analysis, principal component method, classification methods, classification quality indicators.

ВВЕДЕНИЕ

Важность и сложность современных информационных систем в условиях цифровой трансформации общества приводит к рискам их проектирования. Как указано в [1], «проект без рисков – удел неудачников». Риск может быть вызван одной или несколькими причинами. Для анализа данных причин проводится большое число исследований. В результате таких исследований определяют перечни рисков, производят их ранжирование, анализируют динамику изменения и др.

В Своде знаний по управлению проектами (PMBOK) определены основные процессы управления рисками проектирования, а также инструменты идентификации, качественного и количественного анализа рисков, планирования реагирования, мониторинга и управления ими. Несмотря на большое число инструментальных средств, указанных в данном документе, а также в стандарте менеджмента риска [2], проблемы выполнения ИТ-проектов остаются. Так в ежегодном отчете Chaos, подготавливаемом исследовательской консультативной организацией Standish Group [3], указывается, что доля неуспешных проектов, особенно для очень больших и больших проектов, велика, причем она остается высокой, несмотря на предпринимаемые меры. Так, например, в отчете за 2016 год говорится, что 71% проектов, выполненных в 2015 году, были подвержены изменениям или оказались неуспешными.

В 2015 году главнокомандующий ВВС США заявил о неустранимых недостатках в реализации ERP-проекта, который уже стоил государству 1 миллиард долларов [4]. По данным многочисленных исследований для традиционных методов управления проектами, только 44% проектов обычно завершаются вовремя. В среднем проекты занимают 222% от изначально запланированной длительности, 189% от начального бюджета. 70% проектов сокращают исходный объем работ проекта, 30% проектов закрываются досрочно [5].

С учетом сказанного актуальность решения задач разработки и совершенствования инструментальных средств управления рисками проектирования информационных систем не вызывает сомнения. В уже указанных отчетах Standish Group выделены три класса

проектов: успешные; проекты, выполненные частично, с превышением сроков и средств; проекты, завершившиеся неудачей. Распределение проектов на данные классы для различных сфер деятельности по данным отчета за 2015 год приведено в таблице 1. Из таблицы видно, что наибольшая доля неуспешных проектов приходится на государственные информационные системы и телекоммуникационные системы. На такие системы также приходится минимальная доля успешных проектов. Большое число существующих государственных информационных систем (например, в Санкт-Петербурге на момент написания статьи находились в эксплуатации более 60 информационных систем [6]. Граф инфраструктуры включал более 165 дуг) указывает на ее актуальность указанной проблемы для таких систем.

Таблица 1
Распределение проектов по их успешности

Сферы деятельности	Успешно завершены	Счастливчиками изменениями	Неуспешно
Банковское дело	30	55	15
Финансовая сфера	29	56	15
Правительство	21	55	24
Здравоохранение	29	53	18
Производство	28	53	19
Розничная торговля	35	49	16
Услуги	29	52	19
Телеком	24	53	23
Другое	29	48	23

1 СОСТОЯНИЕ ПРОБЛЕМЫ

Профессор Б. Бозм, американский разработчик программного обеспечения, ведущий специалист по инженерии программного обеспечения, в восьмиде-

Рис. 1. Интеллект-карта Б. Боэма

сятых годах прошлого века определил состав риск-менеджмента, основные риски программного проекта. На рисунке 1 показан фрагмент предложенной Боэмом интеллект-карты, которая включает методы анализа рисков. Данная карта показывает наличие количественных методов анализа рисков и целесообразность их совершенствования.

В стандарте по менеджменту рисков [2] приведены основные методы оценки рисков: мозгового штурма, Дельфи, контрольных листов, трехточечной оценки, сценариев, марковских сетей, сетей Байеса, деревьев решений, деревьев событий, причинно-следственного анализа и др. Приведенная в стандарте характеристика их применимости для различных процессов оценки рисков показывает, что наиболее применимыми являются метод, основанный на использовании матрицы вероятностей-воздействий, а также метод анализа видов и последствий отказов. При этом первый из них позволяет получить количественные выходные данные. При его описании в своде РМВОК [7] указаны три класса ситуаций: высокого, среднего и низкого риска. Однако правила отнесения текущей ситуации к соответствующему классу не определены. Предложенное перемножение вероятности реализации угрозы (возможности) на возможный эффект не может быть применено вследствие использования качественных шкал оценки сомножителей.

Одним из основных методов количественной оценки рисков является метод имитационного моделирования. Так в [1] предложена имитационная модель, в которой рассматриваются 10 различных факторов риска. Для каждого из них задается случайная величина – коэффициент умножения времени выполнения проекта, а также категория – фатальный или нефатальный риск. Методика, предложенная авторами, позволяет, оценив данные риски, построить гистограмму времени выполнения проекта, а также определить оценку вероятности его неуспешного завершения. Разработанный программный модуль позволяет задать до десяти основных факторов риска и сформировать пространственную выборку, которая может быть использована для решения задач оценки и прогнозирования рисков.

В модели Expert COSYSMO – советника по моделированию затрат и системному проектированию [8, 9] на основе заданного размера программного обеспечения, а также заданных значений драйверов системной стоимо-

сти, количественно оценивается риск программного проекта, а также его составляющих: рисков платформы, персонала, процесса, продукта. Модель была разработана в центре по системной и программной инженерии Южно-Калифорнийского университета под руководством Б. Боэма и входит в линейку других программных средств оценки программного обеспечения, таких как COSOMO, COCOTS, COSOSIMO, COQUALMO и др., составляющих инструментальные средства инженерии программного обеспечения. В ней производится расчет драйверов риска, которые используются для расчета трудозатрат на проектирование, вырабатываются рекомендации по управлению рисками. Приведенные в [10] выборки, характеризующие выполненные проекты, могут быть использованы для решения других задач оценки рисков.

Модель Expert COSYSMO основана на расчете показателя риска проекта как суммы произведений частных уровней риска на значения драйверов системной стоимости. Результаты произведения категоризуются на классы: низкий, средний, высокий и очень высокий уровни риска на основании таблицы 2.

Таблица 2

Категории проектного риска

Значение показателя проектного риска	Категория риска
0–5	низкий
5–15	средний
15–30	высокий
30–50	очень высокий

В диссертации Р. Валерди рассмотрена выборка, содержащая программные проекты NASA, выполненные в период с 1971 по 1989 годы. Данные проекты разбиты на три категории рисков: низкий риск – 21 проект; средний риск – 55 проектов; высокий риск – 17 проектов. В другом наборе данных, связанном с выполнением проектов в промышленности, приведены 13 проектов с низким риском и 5 проектов со средним риском. Однако размер этих выборок сравнительно невелик и не может быть использован напрямую при решении других задач оценки рисков.

Предложенная в [11] модель нечеткой логики, реализованная в MatLab, основана на использовании фазификации, нечетких правил и дефазификации, позволяющих оценить категорию риска. Данная модель основана на множестве продукционных правил, формируемых экспертами, и проверена на трех наборах данных. Сформированные матрицы рисков для каждой пары анализируемых факторов рисков могут быть использованы для решения других задач оценки рисков.

В схеме последовательности выполнения функций риск-менеджмента, приведенной в [12], указано, что идентификация рисков, под которой понимается процесс выявления и фиксации рисков, является первым этапом управления рисками. От его успешности зависит весь процесс проектирования. Здесь же указано, что качество идентификации может быть улучшено при использовании различных инструментальных методов и средств. Считается, что наиболее ценным в своде знаний РМВОК является описание различных инструментальных средств, в том числе и средств управления рисками. В [13] приведена их характеристика, показано, что, несмотря на их большое число, они в основном ориентируются на традиционные подходы менеджмента качества. В них не используются современные методы машинного обучения, науки о данных, многомерной статистики. Целью настоящей статьи является решение задачи оценки и прогнозирования рисков проектирования информационных систем на основе использования методов машинного обучения. В качестве гипотезы исследования выдвинуто предположение, что применение данных методов повысит качество решения задач оценки и прогнозирования рисков. Данное предположение основано на двух факторах: возможности сформировать обучающую выборку и наличии большого числа инструментальных средств, позволяющих решить такие задачи.

2 Методология

2.1 Постановка задачи прогнозирования рисков

Сформулируем задачу прогнозирования рисков программного проекта как задачу классификации, основанную на методах машинного обучения. Определим факторы возможных рисков выполнения информационных проектов. Отметим, например, риски, определенные в монографии Т. Листера и Т. Демарко [1]. К ним относятся: внутренние изъяны календарного планирования, раздувание требований (изменение требований), текучесть кадров, нарушение спецификаций, низкая производительность. В книге К. Джонса [14] определены факторы, наблюдаемые в удачных и неудачных проектах. В частности, неудачи проектов возникают при некачественном персонале, недостатке практики у менеджеров проектов, недостаточном внутрикомандном взаимодействии и др. В монографии С. Макконелла [15] говорится, что хаотические процессы проектирования возникают вследствие поверхностного анализа требований, отсутствия участия конечного пользователя в постановке требований, недостаточной квалификации

персонала, неполного или неумелого планирования проекта и др.

В отчете Chaos к факторам рисков успешности проектов относят: вовлечение заинтересованных сторон в проект, поддержку руководства и спонсоров, однозначный список требований, правильное планирование, реалистичные ожидания, большое количество ключевых точек, компетентную команду проекта, наличие ресурсов, четкое видение и измеримые цели, нацеленность на результат. Опрошенные фирмой Standish Group менеджеры проектов указали, что причинами неудач проектов являются: неучастие пользователей, неполные требования и характеристики, изменение требований и спецификаций, отсутствие исполнительной поддержки, технологическая некомпетентность, недостаток ресурсов, нереалистичные ожидания, непонятные цели, нереальные временные рамки, новые технологии.

В книге [5] указаны 16 технологий неудачных проектов, среди которых можно выделить неудачи в области управления проектами, а также неудачи в области неправильной практики руководства. Автор данной книги К. Джонс формулирует также социальные факторы, наблюдаемые в проектах, к которым относятся: неопытные старшие администраторы, некачественный персонал, серьезные трения с клиентами и др. Таким образом, число факторов велико, и оно может быть уточнено для проектов конкретных классов или для проектов, выполняемых конкретной организацией.

Для решения задачи классификации рисков определим следующие десять факторов:

- сложность проекта;
- степень полноты требований к проекту;
- степень участия пользователей, спонсоров, стейкхолдеров при выполнении проекта;
- степень обеспеченности ресурсами;
- наличие поддержки руководства при выполнении проекта. Заинтересованность руководства в его результатах;
- стабильность или изменчивость требований.

Отметим, что одно из положений меморандума Agile-технологий указывает, что готовность к изменениям важнее следования первоначальному плану;

- качество организации планирования при выполнении проекта;
 - используемые технологии. Применение гибких технологий;
 - способности команды проекта;
 - организация управления выполнением проекта.
- Наличие системы менеджмента качества, ключевых показателей эффективности, применение показателей модели оценки уровней совершенства СММ и др.

Задачу прогнозирования рисков можно представить как отображение $F: P \rightarrow K$, где P – множество классификационных признаков, K – множество классов проектов. К классификационным признакам будем относить указанные факторы рисков проектов, а к классам отнесем три, определенных в отчетах Chaos: успешное

выполнение проекта, выполнение проекта с частичными изменениями, неуспешное завершение проекта. Данным классам проектов поставим в соответствие три уровня рисков: низкий, средний и высокий.

2.2 Алгоритм решения задачи прогнозирования рисков

На рисунке 2 приведена блок-схема алгоритма решения задачи прогнозирования рисков. Алгоритм со-

Рис. 2. Алгоритм построения классификатора рисков проектирования информационных систем

стоит из двух этапов. На первом этапе формируется обучающая выборка, столбцами которой являются факторы рисков, а строками – ситуации риска. Данная выборка является синтетической и генерируется с помощью метода кластерного анализа и метода главных компонент. На втором этапе осуществляются формирование ансамбля алгоритмов классификации и оценка его качества с помощью тестовой выборки. Формирование ансамбля основано на том, что существует большое разнообразие методов классификации, например, наивный байесовский классификатор, деревья решений, опорные векторы, случайный лес, нейронные сети и др. Все они относятся к машинному обучению и основываются на построении обучающей выборки и ее использовании при построении алгоритма выбранного метода классификации. Выполнение алгоритма основано на применении языка R, статистической платформы SPSS, а также системы Orange Canvas.

2.3 Формирование обучающей выборки

Для формирования обучающей выборки предлагается решение задачи кластерного анализа. При этом в отличие от общепринятого подхода предлагается использовать обучение с учителем. Это позволит не только группировать наблюдения в кластеры, но и проверять правильность их отнесения к сформированному кластеру. Каждый кластер относится к одному классу, одному уровню риска. Таким образом, комплексироваться возможности двух групп задач машинного обучения: кластерного анализа и классификации.

С целью оценки значений каждого из факторов была выбрана категорическая шкала. Левому полюсу данной шкалы соответствуют значения плохого, слабого, низкого значений фактора. Правому значению шкалы соответствует отличное, совершенное, полное его значение. Так, например, для фактора, определяющего степень полноты требований к проекту, предложены следующие значения: «очень мало», «мало», «недостаточное число требований и полноты их описания», «недостаточно полно определены требования к проекту», «полные требования». Для фактора «сложность» предложены следующие категории: «очень сложный проект», «сложный проект», «проект средней сложности», «простой проект», «очень простой проект». Такое шкалирование позволяет в дальнейшем перейти к количественной шкале (в самом простом случае к линейной шкале, где левому полюсу соответствует единичное значение признака, а правому полюсу – значение признака, равное пяти). В случае необходимости число градаций категорической шкалы можно увеличить. От линейной шкалы можно перейти к нелинейной, используя, например, подходы к построению функции полезности.

В состав обучающей выборки включим два экземпляра данных: первый – с максимальным значением всех факторов, второй – с минимальным их значением. Первый экземпляр будем классифицировать как проект, который успешно завершен. Второй – как проект, который не был выполнен. Все остальные экземпляры

Рис. 3. VAT-диаграмма склонности к кластеризации. Строками и столбцами матрицы на диаграмме являются отдельные наблюдения

были сформированы с учетом практического опыта авторов, а также мнений их коллег, участвующих в проектировании информационных систем.

Полученная обучающая выборка должна быть проверена на ее склонность к решению задач кластерного анализа. На рисунке 3 приведена диаграмма оценки визуальной тенденции к кластеризации (VAT-диаграмма), элементы которой показывают на близость наблюдений. Число темных квадратов на ее главной диагонали показывает рекомендуемое число кластеров. Судя по диаграмме набор данных свидетельствует об их склонности к кластеризации. Известно, что любой алгоритм кластеризации считается результативным, если выполняется гипотеза компактности, согласно которой можно найти такое разбиение объектов на кластеры, что расстояния между объектами, принадлежащими одному кластеру будут меньше некоторого значения,

Рис. 5. Ординационная диаграмма данных в осях двух главных компонент

Рис. 4. Доля объясненной дисперсии переменных главными компонентами

а расстояние между объектами разных кластеров будет больше данного значения [16]. Это позволяет внутрикластерные и межкластерные дисперсии использовать для оценки качества решения задачи кластеризации.

Для проверки выполнения данной гипотезы, а также для визуализации результатов кластеризации в силу большой размерности решаемой задачи целесообразно выполнить ее редукцию, например, с помощью метода главных компонент. Это позволит осуществить переход от многокоординатного пространства переменных-факторов риска в координатное пространство главных компонент. На рисунке 4 приведены результаты решения задачи редукции методом главных компонент. Отметим, что возможность и целесообразность такого решения определена тем, что существует значимая на уровне 0,05 ранговая корреляция между переменными-факторами рисков. На диаграмме в качестве координат выбраны две главные компоненты, которые объясняют более 70% общей дисперсии изменчивости факторов риска ИТ-проекта.

На приведенной диаграмме приведены графики двух зависимостей, первая показывает доли дисперсии, объясненной каждой из главных компонент, вторая показывает суммарную дисперсию. Судя по данной диаграмме с использованием критерия «каменистой осыпи» достаточно использовать две главные компоненты, что позволяет построить ординационную диаграмму, приведенную на рисунке 5. Отметим, что вычисляемые собственные числа больше единицы только для первых двух главных компонент.

Диаграмма показывает, что первая главная компонента объясняет более 58% изменчивости факторов. Вторая главная компонента – более 10%. Проведенная в результате решения задачи определения главных компонент диаграмма позволяет сделать вывод, что можно выделить три кластера. Для одного из них первая главная компонента отрицательная. Для двух других данная компонента – положительная. Однако данные кластеры различаются знаком второй главной компоненты.

Таблица 3

Влияние переменных на главные компоненты

Фактор	Dim1	Dim2	Dim3
Сложность	10,4	13,56	6,97
Неполные требования	11,35	3,73	4,24
Неучастие пользователей	6,26	38,98	10,3
Недостаток ресурсов	11,28	15,13	0
Поддержка руководства	9,65	1,24	10
Изменение требований	12,35	1,68	8,35
Отсутствие планирования	10,88	0,87	6,66
Низкие технологии	7,21	12,51	36,6
Способности команды	9,88	3,62	6,64
Отсутствие управления ИТ	10,73	8,69	10,01

В таблице 3 приведены значения значимости переменных для формирования каждой главной компоненты. В таблице 4 приведены коэффициенты корреляции между факторами и главными компонентами (факторные нагрузки главных компонент). С их помощью можно определить семантику главных компонент.

Таблица 4

Значения коэффициентов корреляции (факторных нагрузок) и показателей общности

Фактор риска	Dim1	Dim2	Общность компоненты, h^2
Сложность	0,78	-0,38	0,75
Неполные требования	0,82	-0,2	0,7
Неучастие пользователей	0,61	0,64	0,78
Недостаток ресурсов	0,81	0,40	0,82
Поддержка руководства	0,75	0,11	0,58
Изменение требований	0,85	-0,13	0,74
Отсутствие планирования	0,80	-0,1	0,65
Низкие технологии	0,65	0,36	0,55
Способности команды	0,76	-0,2	0,62
Отсутствие управления ИТ	0,79	-0,3	0,72

Положительная корреляция свидетельствует о том, что с ростом значений факторов рисков (увеличением сложности, ухудшением квалификации, уменьшением степени участия руководства, пользователей и др.) приводит в среднем к росту значений главных компонент. По значениям коэффициентов корреляции можно определить, что первая компонента определяет сам проект,

его сложность, требования к нему. Вторая компонента определяет участие пользователей, ресурсы для выполнения проекта.

Общность компоненты показывает, что с помощью двух главных компонент можно объяснить высокую долю дисперсии факторов рисков. Наиболее высокая общность характерна для рисков, связанных с ресурсами, неучастием пользователей, сложностью проекта. Невысока степень общности для учета технологии, а также для риска, связанного с поддержкой руководства, спонсоров проекта.

На основе сделанного вывода о склонности исходных данных к кластеризации решена задача кластерного анализа. На рисунке 6 приведена дендрограмма, представляющая собой результаты иерархического кластерного анализа исходного набора данных. При построении дендрограммы был выбран один из наиболее популярных методов кластерного анализа, метод Варда, суть которого заключается в построении кластеров с использованием на каждом шаге агломерации результатов дисперсионного анализа. При построении кластеров анализируются координаты x_{ijk} , означающие значение k -го фактора риска для j -го наблюдения при принадлежности данного наблюдения i -му кластеру.

С целью проверки качества построения кластеров на рисунке 7 приведена диаграмма, показывающая распределение объектов по кластерам в результате решения задачи методом k -средних, а также решения задачи выделения главных компонент. Данный рисунок подтверждает, что существует четкое разделение объектов по кластерам. Поэтому можно сделать вывод о хорошем качестве исходной обучающей выборки, ее пригодности для решения задачи классификации. Решение задачи проверки качества обучения с учителем с помощью таблицы сопряженности свидетельствует о сравнительно хорошей обучающей выборке и возможности ее использования для решения задачи классификации рисков проектирования информационных систем.

Рис. 6. Дендрограмма кластерного анализа с использованием метода Варда

2.4 Решение задачи классификации рисков проектов информационных систем

Данная обучающая выборка может быть использована для решения задачи классификации с целью прогнозирования рисков выполнения ИТ-проектов. Выберем ансамбль методов классификации. В их состав включим методы: деревьев решений, опорных векторов, нейронных сетей, случайного леса, а также наивный байесовский классификатор, логистическую регрессию и метод Ада-бустинг. Такое разнообразие методов позволит более обоснованно решить задачу, используя голосование или усреднение. Решим данную задачу с помощью статистического пакета SPSS, а также с помощью приложения Orange Canvas. Использование данных инструментальных средств определено их популярностью, а также наличием в их составе большого числа методов классификации.

В качестве примера на рисунке 8 приведено дерево решений, построенное на основе обучающей выборки с использованием CHAID-алгоритма. Отметим, что сравнительно небольшая обучающая выборка потребовала использования методов повторной выборки, в частности бутстреп-метода с числом повторений, равным десяти. В случае необходимости размер обучающей выборки может быть увеличен методом Монте-Карло.

3 РЕЗУЛЬТАТЫ И ВЫВОДЫ

Качество решения задачи классификации для прогнозирования рисков проектирования информационных систем проверено с помощью показателей AUC, Accuracy (CA), Precision, Recall, F1. Данные показатели являются типовыми и используются по результатам сравнения уровней рисков, заданных в тестовой выборке, и результатов работы классификаторов. В таблице 5 приведены результаты решения задачи с помощью ансамбля методов классификации.

Отметим, что приведенные показатели чувствительны к оценке различных свойств классификаторов (ошибок первого и второго рода, общей безошибочности классификатора и др.). На основании таблицы 5 можно сделать вывод, что, несмотря на то, что все из рассматриваемых методов показывают высокие значения показателей качества классификации, существует вариация их значений. Метода, лучшего по всем показателям, нет. Возникает классическая задача многокритериальной оптимизации. В силу того, что все они нормированы и безразмерны, а также имеют одно и то же направление – максимизацию, можно

Рис. 7. Ординационная диаграмма результатов кластерного анализа методом k -средних

Рис. 8. Дерево решений задачи классификации рисков проектирования информационных систем, построенное с помощью CHAID-метода в системе SPSS

Таблица 5

Результаты анализа качества классификации

Метод классификации	Показатели качества классификации								
	AUC	CA	F1	Precision	Recall	Среднее	Стандартное отклонение	минимум	максимум
Метод деревьев решений	0,937	0,883	0,878	0,895	0,883	0,895	0,024	0,878	0,937
Метод опорных векторов	0,976	0,858	0,851	0,853	0,858	0,879	0,054	0,851	0,976
Нейронные сети	0,964	0,792	0,779	0,780	0,792	0,821	0,080	0,779	0,964
Наивный байесовский классификатор	0,973	0,875	0,876	0,883	0,875	0,896	0,043	0,875	0,973
Логистическая регрессия	0,824	0,667	0,653	0,668	0,667	0,696	0,072	0,653	0,824
Случайный лес	0,972	0,867	0,856	0,883	0,792	0,874	0,065	0,792	0,972
Ада бустинг	0,922	0,900	0,899	0,900	0,900	0,904	0,010	0,899	0,922
Среднее	0,938	0,835	0,827	0,837	0,824				
Стандартное отклонение	0,054	0,081	0,086	0,085	0,081				

Таблица 6

Матрица ошибок
для наивного байесовского классификатора

Наблюдаемое значение	Предсказанное значение уровня риска, %		
	низкий	средний	высокий
низкий	97,6	0,0	0,0
средний	2,4	71,4	9,1
высокий	0,0	28,6	90,9

использовать их свертку для выбора лучшего метода с помощью соотношения

$$1/n \sum_{i=1}^n x_{ij} \rightarrow \max_j,$$

где x_{ij} – значение i -го показателя качества классификации для j -го метода.

Для решаемой задачи классификации лучшим оказался метод наивного байесовского классификатора, несмотря на то, что данный метод считается простым. Обратим внимание на то, что и дисперсия (среднее квадратическое отклонение) для данного метода соизмерима с ее значениями других анализируемых методов классификации. Кроме того, максимальное значение показателя качества для данного метода близко к максимуму среди всех методов. В таблице 6 приведена матрица ошибок для наивного байесовского классификатора.

Данный вывод подтверждается и с помощью ROC-кривой, приведенной на рисунке 9 для каждого из трех классов.

Анализ данной матрицы показывает, что наибольшие ошибки связаны со вторым уровнем риска. Это обусловлено тем, что в обучающей и тестовой выборках данный класс имеет наименьшую мощность.

Рис. 9. ROC-кривые для различных уровней риска: а) низкий; б) средний; в) высокий

ЗАКЛЮЧЕНИЕ

В статье рассмотрено решение задачи классификации уровней риска проектирования. Установлено, что типовые классификаторы успешно справляются с решением данной задачи по выбранным факторам оценки рисков ИТ-проекта. Используемый индуктивный подход основан на процедуре обучения алгоритмов классификации. При этом ее сходимость зависит от качества обучающей выборки, используемых классификационных признаков, выбранной шкалы оценивания, числа классов риска и других факторов. Получаемые алгоритмы, а также правила оценки рисков должны быть сформированы с учетом особенностей предметной области, в которой выполняются ИТ-проекты. Увеличение числа классов ситуации, использование ансамбля методов позволит улучшить качество классификации, возможности классификатора. Решение задачи прогнозирования рисков должно выполняться в динамике, а не только при принятии решения на выполнение проекта. Это позволит исследовать частные риски, уточнить качество исходных данных, формировать различные тренды.

В случае необходимости можно выбрать числовую шкалу для зависимой переменной – уровня риска. В этом случае от задачи классификации следует перейти к задаче регрессионного анализа. Для этого могут быть использованы методы деревьев решений, опорных векторов и другие методы машинного обучения. Регрессионная модель позволит оценить значимость различных признаков, точность их оценки. При этом также следует использовать ансамбль методов и решать задачу выбора лучшего из них.

В силу того, что для каждой организации, для каждого проекта существует своя специфика, необходимо использовать свои обучающие выборки, которые формируются с учетом накопленного опыта проектирования.

СПИСОК ЛИТЕРАТУРЫ

1. Листер Т., Демарко Т. Вальсируя с медведями: управление рисками в проектах по разработке программного обеспечения. – URL: https://royallib.com/book/demarko_tom/valsiruya_s_medvedyami.html (дата обращения: 09.08.2019).
2. ГОСТ Р ИСО/МЭК 31010-2011. Менеджмент риска. Методы оценки риска. – М.: Стандартинформ, 2012.
3. CHAOSReport2015-Final.pdf. – URL: https://www.standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf (дата обращения: 09.09.2019).
4. Давайте подискутируем на тему Agile. – URL: http://gibtech.ru/blog/discus?entry_id=111 (дата обращения: 20.09.2019).
5. Jones C. Patterns of Software System Failure and Success International Thompson Computer. – URL: <https://pdfs.semanticscholar.org/2db6/4759eb7ac5d431b7789bbc25fedb6ed4aa76.pdf> (дата обращения: 10.09.2019).
6. Наумов В.Н., Кучеренко Д.В. Исследование структурной сложности инфраструктуры государственных

информационных систем Санкт-Петербурга методами анализа социальных графов // Современные наукоемкие технологии. – 2019. – № 2. – С. 114–122.

7. РМВОК. Руководство к Своду знаний по управлению проектами. – 6-е изд. – М.: Изд. Олимп-бизнес, 2016.

8. Valerdi R., Boehm Barry W. COSYSMO: A system engineering cjst model/ – URL: <http://csse.usc.edu/TECHRPTS/2010/usc-csse-2010-508/usc-csse-2010-508.pdf> (дата обращения: 20.09.2019).

9. Valerdi, Ricardo & John E., Gaffney Jr. (2007). Reducing Risk and Uncertainty in COSYSMO Size and Cost Drivers: Some Techniques for Enhancing Accuracy. – URL: https://www.researchgate.net/publication/229034929_Reducing_Risk_and_Uncertainty_in_COSYSMO_Size_and_Cost_Drivers_Some_Techniques_for_Enhancing_Accuracy (дата обращения: 28.09.2019).

10. Software Project Risk Assessment and Efforts Contingency Model based on COCOMO Cost Factor COCOMO. – URL: <https://www.researchgate.net/publication/236679817> (дата обращения: 28.09.2019).

11. Software_Project_Risk Fuzzy Expert-COCOMO Risk Assessment and Effort Contingency Model in Software Project. – URL: <https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=2464&context=etd> (дата обращения: 10.09.2019).

12. Николаенко В.С. Анализ инструментов по обеспечению функции управления рисками в ИТ-проектах // Государственное управление. Электронный вестник. – апрель 2015. – Вып. 49. – URL: http://e-journal.spa.msu.ru/uploads/vestnik/2015/vipusk_49_aprel_2015_g._problemi_upravlenija_teorija_i_praktika/nikolaenko.pdf (дата обращения: 10.09.2019).

13. Наумов В.Н. Проектирование информационных систем. – СПб.: Изд. Центр СЗИУ, 2018.

14. History of Project Management. Process Quality Associates Inc – URL: <http://www.pqa.net/ProdServices/ccpm/W05002001.html> (дата обращения: 10.09.2019).

15. Макконелл С. Сколько стоит программный проект. – СПб.: Питер, 2007.

16. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.

REFERENCES

1. Lister T., Demarko T. *Valsiruya s medvediami: upravlenie riskami v proektakh po razrabotke programmnogo obespecheniia* [Waltzing with Bears: Managing Risk on Software Projects]. Available at: https://royallib.com/book/demarko_tom/valsiruya_s_medvedyami.html (accessed: 09.08.2019).
2. *GOST R ISO/MEK 31010-2011. Menedzhment riska. Metody otsenki riska* [National Standard. Risk Management. Risk Assessment Methods]. Moscow, Standartinform Publ., 2012.
3. *CHAOSReport2015-Final.pdf*. Available at: https://www.standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf (accessed: 09.09.2019).

4. *Davaite podiskutiruem na temu Agile* [Let Us Discuss Agile]. Available: http://gibtech.ru/blog/discus?entry_id=111 (accessed: 20.09.2019).
5. Jones C. *Patterns of Software System Failure and Success International Thompson Computer*. Available at: <https://pdfs.semanticscholar.org/2db6/4759eb7ac5d431b7789bbc25fedb6ed4aa76.pdf> (accessed: 10.09.2019).
6. Naumov V.N., Kucherenko D.V. Issledovanie strukturnoi slozhnosti infrastruktury gosudarstvennykh informatsionnykh sistem Sankt-Peterburga metodami analiza sotsialnykh grafov [Analysis of the Structural Complexity of the State Information Systems of St. Petersburg Using Methods of Analysis of Social Graphs]. *Sovremennye naukoemkie tekhnologii* [Modern High Technologies], 2019, no. 2, pp. 114–122.
7. PMBOK. *Rukovodstvo k Svodu znaniy po upravleniiu proektami. 6-e izd.* [Guide for the Body of Knowledge on Project Management. 6th Edition]. Moscow, Olimp-biznes Publ., 2016.
8. Valerdi R., Boehm Barry W. *COSYSMO: A System Engineering Cost Model*. Available at: <http://csse.usc.edu/TECHRPTS/2010/usc-csse-2010-508/usc-csse-2010-508.pdf> (accessed: 20.09.2019).
9. Valerdi, Ricardo & John E., Gaffney Jr. *Reducing Risk and Uncertainty in COSYSMO Size and Cost Drivers: Some Techniques for Enhancing Accuracy*. 2007. Available at: https://www.researchgate.net/publication/229034929_Reducing_Risk_and_Uncertainty_in_COSYSMO_Size_and_Cost_Drivers_Some_Techniques_for_Enhancing_Accuracy (accessed: 28.09.2019).
10. *Software Project Risk Assessment and Efforts Contingency Model based on COCOMO Cost Factor COCOMO*. Available at: <https://www.researchgate.net/publication/236679817> (accessed: 28.09.2019).
11. *Software Project Risk Fuzzy Expert-COCOMO Risk Assessment and Effort Contingency Model in Software Project*. Available at: <https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=2464&context=etd> (accessed: 10.09.2019).
12. Nikolaenko V.S. Analiz instrumentov po obespecheniiu funktsii upravleniya riskami v IT-proektakh [Analysis of Risk Management Tools in IT Projects]. *Gosudarstvennoe upravlenie. Elektronnyi vestnik* [Public Instrumentation. E-Journal], 2015, iis. 49. Available at: http://e-journal.spa.msu.ru/uploads/vestnik/2015/vipusk__49._aprel_2015_g._problemi_upravleniya_teorija_i_praktika/nikolaenko.pdf (accessed: 10.09.2019).
13. Naumov V.N. *Proektirovanie informatsionnykh system* [Designing Information System]. St.Petersburg, SZIU Publ., 2018.
14. *History of Project Management. Process Quality Associates Inc.* Available at: <http://www.pqa.net/ProdServices/ccpm/W05002001.html> (accessed: 10.09.2019).
15. MacConnell S. *Skolko stoit programmnyi proekt* [Software Project Survival Guide]. St. Petersburg, Piter Publ., 2007.
16. Zagoruiko N.G. *Prikladnye metody analiza dannykh i znaniy* [Applied Methods for Data Analysis and Knowledge]. Novosibirsk, IM SO RAN Publ., 1999.