veloper, CASE-средство для проектирования и разработки баз данных ERWin Modeller, среда моделирования программного обеспечения на языке UML Visual Paradigm. Разработка программного кода велась на Java.

В настоящее время система сдана в эксплуатацию, ОАО «НИЦ СПб ЭТУ» проводит ее авторское сопровождение. Результаты опытной эксплуатации показали, что функциональные возможности системы позволяют проводить круглосуточный мониторинг работоспособности модернизированной автоматизированной системы таможенных органов, анализировать причины нештатных ситуаций и помогают выбирать действия, необходимые для устранения их последствий.

СПИСОК ЛИТЕРАТУРЫ

О таможенном регулировании в Рос. Федерации: Федеральный закон Рос. Федерации от 27.11.2010 № 311-ФЗ // Собр. законодательства Рос. Федерации. 2010. № 48. Ст. 6252.

A. A. Liss, A. S. Skripnikova

AUTOMATED SYSTEM FOR ARCHIVING AND AUDIT OF ELECTRONIC DATA INTERCHANGE IN THE CUSTOMS BUSINESS PROCESSES

The article deals with the automated system developed in the framework of the modernization of the informational and technical Russian Customs support. The processes of archiving and audit of electronic messages, forming in the course of customs activities, are discussed. The possibility of using this system to identify patterns of electronic exchange and tactics of response to emergency situations is underlined.

The automated system of customs authorities, electronic submission, electronic message, information exchange, emergency situation, data recovery

УДК: 20.53.19, 28.23.13

И. И. Холод

Применение методов Data Mining для оценки выполнения программных мероприятий предприятиями ОПК

Описываются подходы применения методов Data Mining для решения различных задач, связанных с оценкой выполнимости программных мероприятий предприятиями оборонно-промышленного комплекса.

Интеллектуальный анализ данных, Data Mining, риски, оценка рисков

Неопределенности финансового, экономического и технологического характера, которые в настоящее время сопутствуют деятельности предприятий оборонно-промышленного комплекса (ОПК), предопределяют возможности появления соответствующих рисков невыполнения ими мероприятий федеральных целевых программ (ФЦП) и государственного оборонного заказа (ГОЗ). В этих условиях появляется необходимость разработки методов оценки финансово-экономических и технологических рисков, возникающих в процессе создания предприятиями отрасли продукции военного, двойного и гражданского назначений, позволяющих прогнозировать состояние предприятий промышленности и оценивать риски невыполнения ими программных мероприятий.

В настоящее время существуют различные информационные системы для сбора и хранения информации о финансово-экономических параметрах предприятия, его технических возможностях, кадровом потенциале, а также среды его функционирования. В качестве примеров можно привести системы, созданные по заказу Министерства промышленности и торговли РФ:

– автоматизированный реестр организаций ОПК;

– единая информационная система ОПК.

Помимо информационных систем существует базы данных о выполнении предприятиями государственных контрактов:

– единый реестр государственных и муниципальных контрактов;

– реестр организаций;

– реестр недобросовестных поставщиков и др.

Перечисленные и им подобные системы формируют распределенное (поскольку системы расположены в различных ведомствах) информационное пространство, хранящее историю выполнения государственных контрактов. Вся информация может быть однозначно связана по уникальным идентификаторам: ИНН, расчетные счета и др. Объем и разнородность информации в таком хранилище очень велики, в связи с чем применение обычных статистических методов анализа неэффективно как с точки зрения получаемой с их помощью информации, так и с точки зрения времени их работы. Альтернативой могут служить методы интеллектуального анализа данных (в зарубежной литературе эта область имеет название Data Mining).

Методы интеллектуального анализа данных могут быть применены для решения различных задач в области анализа рисков невыполнения мероприятий ФЦП и ГОЗ:

– выявление потенциально недобросовестных поставщиков (исполнителей);

– выявление неисполнимых мероприятий;

– определение параметров отбора исполнителей;

– прогнозирование результатов выполнения предприятиями контракта и др.

Задачи выявления потенциально недобросовестных поставщиков (исполнителей) и неисполнимых мероприятий могут решаться методами классификации на основании информации о ранее не выполненных контрактах и характеристиках их исполнителей. Классическое описание задачи классификации имеет следующий вид [см. лит.]: имеется множество объектов

$$I = \{i_1, i_2, ..., i_j, ..., i_n\},$$

где i_j – исследуемый объект.

Такими объектами в рассматриваемых задачах являются выполняемые мероприятия и их исполнители.

Каждый объект характеризуется набором переменных:

$$I_j = \{x_1, x_2, ..., x_h, ..., x_m, y\},$$

где x_h – независимая переменная, значения которой известны и на их основании определяется значение зависимой переменной y. В задаче выявления потенциально недобросовестных поставщиков независимыми переменными являются финансово-экономические, технологические, кадровые и другие характеристики предприятий ОПК. Примерами таких характеристик могут служить:

– отношение кредиторской/дебиторской задолженностей;

– рентабельность производственных фондов;

– среднесписочная численность работников;

– средний возраст сотрудников;

– отрасль предприятия;

– образование генерального директора;

– средняя заработная плата;

– регион дислокации предприятия и др.

В задаче выявления неисполнимых мероприятий независимыми переменными будут характеристики самих мероприятий (срок, цена, технические характеристики разрабатываемых изделий и т. п.).

В Data Mining набор независимых переменных часто обозначают в виде вектора:

$$\boldsymbol{X} = \{x_1, x_2, ..., x_h, ..., x_m\}.$$

Каждая переменная x_h может принимать значения из некоторого множества

$$C_h = \{c_{h1}, c_{h2}, ...\}.$$

Если значениями переменной являются элементы конечного множества, то говорят, что она имеет категориальный тип. Например, переменная «отрасль предприятия» принимает значения на множестве значений (авиационная, судостроительная, радиоэлектронная…). Независимая переменная может принимать числовые значения в некотором диапазоне (значением переменной «средний возраст сотрудников» может быть целое число в диапазоне от 18 до 70). Возможны типы числовых переменных «без ограничений» или «иметь ограничения снизу или сверху». Например, переменная «среднесписочная численность работников» имеет ограничение снизу и может изменяться в диапазоне от 0 до +∞.

Если множество значений $C = \{c_1, c_2, ..., c_r, ..., c_k\}$ зависимой переменной y конечное, то задача является задачей классификации. Например, в задаче выявления недобросовестных поставщиков зависимой переменной является «исполнение контракта», а ее значениями – {выполнен в срок, выполнен с задержкой, не выполнен}. Если пере-

менная y принимает значения на множестве действительных чисел R, то задача называется задачей регрессии. В рассматриваемых задачах такой переменной может быть «срок задержки выполнения», чьи значения меняются в диапазоне от 0 до $+\infty$.

Приведение указанных задач к задаче классификации или регрессии позволяет применить к ним известные методы Data Mining: построения деревьев решений (ID3, C4.5 и др.), аппроксимации (метод наименьших квадратов, SVM и др.), построения классификационных правил (1R, Naïve Bayes и др.). Все перечисленные методы должны применяться к накопленной информации о выполнении контрактов, их исполнителей и значениях показателей на период выполнения контракта.

В результате применения данных методов будут построены классификаторы, позволяющие по параметрам предприятия и программного мероприятия определять добросовестность исполнителя и/или выполнимость данного мероприятия. Полученные классификаторы могут уточняться как применением других методов Data Mining, так и их обучением на вновь поступивших данных.

Задача определения параметров для отбора исполнителей может решаться методами кластеризации на основании анализа характеристик исполнителей, имеющих разные результаты в выполнении контрактов.

Формально задача кластеризации описывается следующим образом [1]. Дано множество объектов данных I, каждый из которых представлен набором атрибутов. Требуется построить множество кластеров C и отображение F множества I на множество C, т. е. $F : I \to C$. Отображение F задает модель данных, являющуюся решением задачи.

Множество I определим следующим образом:

$$I = \left\{ i_1, i_2, ..., i_j, ..., i_n \right\},$$

где i_j – исследуемый объект.

В задаче определения параметров для отбора исполнителей объектами множества I также будут предприятия, атрибутами – их характеристики. Для определения наиболее значимых атрибутов, опираясь на которые можно выбирать предприятия для выполнения программных мероприятий, необходимо выполнять кластеризацию на множестве объектов I с использованием различных комбинаций атрибутов. Комбинация, которая приведет к разбиению на кластеры наилучшим образом (будет максимально соответствовать группам выполненных и невыполненных контрактов), и будет наиболее значимой.

Кластеризация может осуществляться различными методами Data Mining: построением центроидных кластеров (K-Means, Fazzy K-Means и др.), иерархическими методами (дивизимными и агломеративными) и др.

Выбирать комбинации можно простым перебором или привлекая эксперта. Первый вариант наиболее прост, может быть автоматизирован, но требует большего времени. Второй вариант может быть выполнен быстрее, но требует привлечения эксперта, который, кроме того, может пропустить комбинацию атрибутов, представляющую интерес.

Данная задача может решаться как для всех видов контрактов сразу, так и для каждого вида в отдельности. Разделение контрактов на виды может осуществляться по разным признакам: объемам финансирования, области разработки и др. В первом случае будут выявлены общие характеристики предприятий, значимые для любого типа контракта. Во втором случае будут определены характеристики, значимые для определенного вида контракта. При использовании полученной информации (о значимости атрибутов) наиболее эффективно будет объединить характеристики, полученные разными способами.

Задача прогнозирования результатов выполнения предприятиями контракта может быть решена методами анализа временных рядов.

Временным рядом называется последовательность событий, упорядоченных по времени их наблюдения. События обычно фиксируются через равные интервалы времени T и представляются в виде последовательности

$$\{e_1, e_2, ..., e_i, ..., e_n\},$$

где e_i – событие в момент времени t_i; n – общее количество событий. В рассматриваемой задаче такими событиями могут быть: сдача этапов контракта, смена руководителей работ, выдача кредитов и другие события в деятельности предприятия.

Задачу построения прогноза по временному ряду можно сформулировать следующим образом: пусть дан временной ряд $\{e_1, e_2, ..., e_i, ..., e_n\}$, требуется на его основании определить значение e_{n-k} при $k > 0$. Например, определить, будет ли сдан очередной этап, на основании событий, произошедших в деятельности предприятия в период выполнения этапа.

Прогнозирование временных рядов осуществляется в три этапа:

– построение модели, характеризующей временной ряд. Для этого применяются различные методы классификации;

– оценка построенной модели. Имеющиеся данные разбиваются на два множества: обучающее и тестовое. Построение модели выполняется на обучающем множестве, а затем с ее помощью строят прогноз на тестовом множестве. Спрогнозированные результаты сравнивают с реальными данными и по степени ошибки оценивают модель;

– если построенная на первом этапе модель получила удовлетворительную оценку, то ее можно использовать для прогноза будущих событий.

Данная задача может решаться как методами математической статистики (экстраполяция, экспоненциальное сглаживание и др.), так и методами Data Mining, например методом скользящего окна.

Основная идея метода скользящего окна представлена гипотезой существования некого закона, по которому можно определить значение очередного члена ряда как функцию от нескольких предыдущих членов. Обычно из каких-то соображений фиксируют число k и предполагают, что только k предшествующих членов влияют на дальнейшее поведение ряда, а зависимостью от остальных пренебрегают, т. е.

$$e_{n+1} = f\left(e_n, e_{n-1}, ..., e_{n-k}\right).$$

При этом говорят об «окне» размером k, в пределах которого рассматривается ряд. Для нахождения функции f временной ряд «нарезается» на множество окон (каждое из которых сдвигается на один элемент). На полученном множестве выполняется поиск искомой функции.

Необходимо заметить, что если функция f используется для предсказания численных значений, то говорят о задаче регрессии. В случае категориальных значений ряда речь идет о классической задаче классификации. Решаются эти задачи методами Data Mining. Например, задача регрессии может быть решена методом SVM, а задача классификации – методами построения деревьев решений. При использовании метода деревьев решений полученные результаты легко представить в виде правил < если–то >. В этом случае в условной части (если) указываются уже прошедшие события, а в заключительной (то) – предсказываемые.

Описанные задачи оценки рисков невыполнения программных мероприятий не ограничиваются рассмотренным списком. Другие задачи в этой области также могут решаться методами Data Mining. Основным требованием для этого является доступность всей необходимой для анализа информации.

СПИСОК ЛИТЕРАТУРЫ

Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. СПб.: БХВ-Петербург, 2009.

I. I. Kholod

APPLICATION OF DATA MINING FOR PERFORMANCE EVALUATION PROGRAM ACTIVITIES DEFENSE COMPANIES

The article describes the application of the approaches of Data Mining methods for solving various problems associated with the assessment of the feasibility of the program activities of defense-industrial complex.

Data Mining, risks, risk assessment

УДК: 20.53.19, 28.23.13

И. И. Холод, С. В. Родионов

Построение централизованных хранилищ данных для систем управления предприятиями ОПК

Описывается построение централизованных хранилищ данных для систем поддержки принятия решений, используемых в управлении предприятий оборонно-промышленного комплекса.

Системы поддержки принятия решений, хранилища данных

В настоящее время существует множество информационных систем содержащих различную информацию о деятельности предприятий оборонно-промышленного комплекса (ОПК). Такие системы разрабатываются по заказам различных ведомств, на разных уровнях (отраслевом и феде-