УДК 004.9

**Бегутова С.В.**
Московский государственный университет экономики, статистики и информатики
E-mail: sbegutova@gmail.com

# ИСПОЛЬЗОВАНИЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ОЦЕНКИ РИСКА НЕУПЛАТЫ ТАМОЖЕННЫХ ПЛАТЕЖЕЙ

**В статье рассмотрены методы оценки риска неуплаты таможенных платежей, описано построение модели оценки риска на основе применения интеллектуального анализа данных и метода классификации. Выделены основные показатели, влияющие на оценку риска неуплаты таможенных платежей.**

**Ключевые слова: риск, методы интеллектуального анализа данных, Data Mining.**

В современных условиях таможенные платежи вносят значительный вклад в федеральный бюджет. При этом повышение уровня собираемости таможенных платежей находится в прямой зависимости не только от объема внешней торговли, но и от своевременности их поступления от участников внешнеэкономической деятельности (ВЭД).

Таможенные платежи при перемещении товаров и транспортных средств через границу уплачиваются сразу или в течение нескольких дней. Однако существует ряд случаев, вследствие которых участнику ВЭД необходимо предоставить обеспечение уплаты таможенных платежей. Возникает риск их несвоевременной уплаты по истечении срока действия обеспечения, что в свою очередь приводит к задержке поступления денежных средств в федеральный бюджет. В связи с этим необходимо до выпуска товаров и транспортных средств через таможенную границу оценить риск неуплаты таможенных платежей. Его расчет базируется на анализе данных грузовых таможенных деклараций (ГТД).

Необходимо отметить, что ГТД как источник информации имеет ряд особенностей. Во-первых, большой объем входных показателей, число которых превышает 100, существенно затрудняет процесс оценки уровня риска. К ним относятся: номер ГТД, статус, код таможни, направление перемещения товара (экспорт либо импорт), ИНН декларанта, код страны местонахождения и др. Во-вторых, данные являются разнородными (количественными, качественными). Для того чтобы оценить риск неуплаты таможенных платежей, необходимо отобрать признаки, которые коррелируют с признаком нарушения срока уплаты.

Целью данной работы является определение методов, позволяющих выявлять значимые признаки, которые используются для определения уровня риска неуплаты таможенных платежей. Для достижения цели требуется решение следующих задач:

1) обоснование необходимости использования методов интеллектуального анализа данных для оценки риска неуплаты таможенных платежей;

2) выбор метода интеллектуального анализа данных для построения модели оценки риска неуплаты таможенных платежей;

3) выявление признаков, которые позволят оценить риск неуплаты таможенных платежей, с использованием интеллектуального анализа данных.

Рассмотрим решение перечня задач более подробно.

**1. Выбор метода оценки уровня риска неуплаты таможенных платежей.** Для анализа рисков используются различные методы, среди которых можно выделить статистические, такие как расчет среднеквадратического отклонения, дисперсии, которые позволяют оценить риск, но не выявляют факторы, влияющие на этот уровень. Для их идентификации служат методы факторного, кластерного анализа данных. Многие из них в настоящее время аккумулируются в методах интеллектуального анализа данных (ИАД). ИАД позволяет не только проводить качественный анализ, но и выявлять зависимости в больших объемах данных [1].

Технологии ИАД ориентированы на выявление скрытых, неочевидных и существующих причинно-следственных взаимосвязей между различными факторами в больших объемах. Эти технологии являются основным инструментом исследования сложных процессов и обнаружения в них фрагментов с однородными свойствами или шаблонов (паттернов), от-

ражающих особенности многоаспектных отношений в данных, которые могут быть компактно выражены в понятной пользователю форме. При этом технологии ИАД позволяют осмыслить данные, оценить их как с количественной, так и с качественной точек зрения [5].

В зарубежной специальной литературе для обозначения области интеллектуального анализа данных наиболее широко используются термины Knowledge Discovery in Database (KDD, извлечение знаний из базы данных) и Data Mining (DM, извлечение данных). KDD – это процесс поиска полезных знаний в накопленных данных. Он включает в себя вопросы подготовки данных, выбора информативных признаков, очистки данных, применения методов DM, постобработки данных, интерпретации полученных результатов. Ядром всего этого процесса являются методы DM. DM – это процесс обнаружения в накопленных данных ранее неизвестных нетривиальных практически полезных и доступных интерпретаций знаний в виде зависимостей, необходимых для принятия решений. Цель технологии – нахождение в данных таких зависимостей, которые не могут быть найдены обычными статистическими методами. DM является одним из шагов KDD [4].

Для решения задачи оценки риска неуплаты таможенных платежей достаточно применения технологии DM, которая позволит выявить значимые признаки и зависимость между ними и уровнем риска.

**2. Выбор метода интеллектуального анализа данных.** Методы Data Mining помогают решить многие задачи, из которых можно выделить основные: классификация, прогнозирование, поиск ассоциативных правил и кластеризация. Для решения перечисленных задач используются следующие методы: деревья решений, метод математических функций, правила и кластеры [4].

В основе метода выявления зависимостей лежит анализ обучающих выборок. Причем обучение модели может быть [7]:

– «с учителем», когда для каждого примера задается в явном виде значение признака его принадлежности к некоторому классу ситуаций (классообразующего признака);

– «без учителя», когда по степени близости значений признаков классификации система сама выделяет классы ситуаций.

При решении задачи оценки риска неуплаты таможенных платежей на основе данных ГТД была проведена выборка данных по признаку выпуска товара или транспортного средства через границу России под обеспечение уплаты таможенных платежей. Для каждой ГТД был проставлен признак наличия или отсутствия нарушения срока уплаты. Таким образом, был сформирован обучающий признак, который несет основную информацию для решения задачи оценки риска неуплаты таможенных платежей. Его наличие позволяет дальнейший анализ данных проводить с помощью обучения модели «с учителем».

В рамках выбранного типа модели реализованы два метода ее построения: классификация и регрессия. Для того чтобы определить наиболее эффективный метод решения поставленной задачи, приведем их краткую характеристику.

Классификация – задача разбиения множества объектов или наблюдений на группы, называемые классами. Внутри каждого класса объекты обладают схожими свойствами и признаками. Необходимо отметить, что при построении модели классификации количество классов определяется в соответствии с количеством принимаемых значений обучающего атрибута. При построении модели признаки могут быть как количественными, так и качественными. Это свойство расширяет границы применения метода классификации, так как большое количество информации о предметной области может быть качественного типа.

Метод регрессионного анализа данных позволяет анализировать значительные объемы информации с целью исследования вероятной взаимосвязи двух и более переменных. Он используется по двум причинам. Во-первых, так как описание зависимости между переменными помогает установить наличие возможной причинной связи. Во-вторых, получение аналитической зависимости между переменными дает возможность предусматривать будущие значения показателя по значениям признаков. Однако данный метод имеет ряд ограничений при применении: данные должны быть количественного типа [2].

Большинство признаков, которые подаются на вход при построении модели оценки риска неуплаты таможенных платежей, являются

качественными. Например: направление перемещения товара, код страны происхождения товара, местонахождение участника ВЭД и др. В связи с описанными выше ограничениями метод регрессионного анализа не позволит учесть такие признаки при построении модели. Классификационный метод позволяет включить в анализ качественные признаки. Таким образом, в качестве метода выберем классификацию как наиболее приемлемую для решения поставленной задачи. В данном случае имеет место бинарная классификация, так как значение классообразующего признака нарушения срока уплаты принимает два значения: «0» – нет нарушения, «1» – есть нарушение.

**3. Выявление признаков с помощью применения метода классификации.**

Для построения модели на основе метода классификации рассмотрим этапы этого процесса. На этапе выбора инструментального средства построения модели необходимо учесть, что число входных показателей превышает 100. Из множества инструментальных средств, позволяющих проводить интеллектуальный анализ данных, был выбран Oracle Data Miner (ODM). Он не ограничивает количество входных атрибутов, что позволяет подать на вход для построения модели более 100 показателей.

Первым этапом построения модели в ODM является разбивка набора исходных данных из агрегированных таблиц на два множества: обучающее и тестовое (Test). Причем тестовое множество не должно зависеть от обучающего. Обучающее множество в свою очередь делится на два подмножества: данные для построения модели (Build) и для проверки (Apply). Тестовые данные используются для проверки работоспособности модели. Таким образом, необходимо разделить все множество данных на три равные группы так, чтобы количество данных с нарушением и без нарушения срока уплаты таможенных платежей в каждой группе было одинаковое. Каждая группа будет использована последовательно на трех этапах построения модели [9].

Построение модели классификации проводится на основе данных трех агрегированных таблиц. Первая таблица представляет собой объединение данных об общих сведениях ГТД и рассчитанных таможенных платежах, подлежащих уплате (Т1). Вторая таблица включает в себя сведения о ГТД, товарах и предоставленных документах (Т2). Третья агрегированная таблица является объединением данных о ГТД, товарах и исчисленных таможенных платежах (Т3). Создание агрегированных таблиц с разным набором атрибутов позволило осуществить поиск оптимального набора показателей, влияющих на оценку риска неуплаты таможенных платежей, с наивысшей точностью построения модели. Для построения модели были отобраны ГТД за 2006-2008 гг.

В связи с тем, что для анализа данных используется три агрегированных таблицы, которые состоят из большого числа атрибутов, построение классификационной модели может занять длительное время. К тому же избыточность атрибутов приводит к погрешностям при построении модели. С помощью ODM, в котором реализован алгоритм поиска существенных атрибутов (Attribute Importance), был определен состав значимых и незначимых показателей.

Разработка модели в ODM проходит несколько этапов: дискретизация и расщепление анализируемых данных, построение модели и расчет тестовых метрик. Тестирование используется для расчета показателя, который характеризует точность предсказания на новых данных. В ODM используются следующие правила интерпретации: если процент верно предсказанных значений находится в промежутке [0;30], то такой показатель относится к низкому уровню прогнозирования. Если результат тестирования входит в промежуток [30;70], такая ситуация говорит о хорошем уровне прогнозирования на новых данных. Если промежуток составляет [70;100], то в этом случае модель построена с наивысшей точностью для прогнозирования [9].

Для того чтобы построить модель методом классификации и получить набор правил, необходимо выбрать алгоритм. ODM включает в себя три наиболее мощных алгоритма классификации (supervised learning) [9]:

– опорных векторов (Support Vector Machine), который относится к группе граничных методов и использует теорию вычисления близости векторов для классификации объектов. При помощи данного метода решаются задачи бинарной классификации;

– байесовских сетей (Adaptive Bayes Network), который имеет в основе два предположения: все признаки являются одинаково

Таблица 1. Результаты тестирования моделей

| Таблицы агрегированных данных | Название алгоритма классификации | Точность прогнозирования, % |
|---|---|---|
| T1 (агрегация данных ГТД – общие сведения и ГТД – таможенные платежи к уплате) | Support Vector Machine | 29,78 |
| | Adaptive Bayes Network | 36,42 |
| | Decision Tree | 0 |
| T2 (агрегация данных ГТД – общие сведения, ГТД – товары и ГТД – предоставляемые документы) | Support Vector Machine | 52,94 |
| | Adaptive Bayes Network | 74,75 |
| | Decision Tree | 5,03 |
| T3 (агрегация данных ГТД – общие сведения, ГТД – товары и ГТД – исчисление таможенных платежей) | Support Vector Machine | 50,59 |
| | Adaptive Bayes Network | 51,59 |
| | Decision Tree | 0,91 |

Таблица 2. Признаки, выявленные на основе агрегированной таблицы T2

| Условное обозначение | Признак | Описание |
|---|---|---|
| x1 | Таможенная процедура | Совокупность положений, предусматривающих порядок совершения таможенных операций и определяющих статус товаров и транспортных средств для таможенных целей |
| x2 | Направление перемещения товара | Принимает значение «Экспорт» или «Импорт» |
| x3 | Признак корректировки таможенной стоимости | Стоимость корректируется при выявлении технических ошибок и других нарушений, описанных в нормативных документах |
| x4 | Код, определяющий почтовую зону декларанта | Почтовая зона регистрации декларанта |

важными и являются статистически независимыми, т. е. значение одной переменной ничего не говорит о значении другой. Использует теорию вероятности для классификации объектов;

– деревьев решений (Decision Tree), который классифицирует объекты путем построения деревьев решений.

Для задачи оценки риска неуплаты таможенных платежей критерием эффективности является точность построенной модели. Этот показатель рассчитывается как отношение количества правильных предсказаний к общему количеству предсказаний [8].

Для того чтобы определить, какой из алгоритмов классификации обладает наибольшей точностью выявления риска нарушения сроков уплаты таможенных платежей, построим модель оценки риска неуплаты таможенных платежей последовательно каждым из описанных алгоритмов.

В таблице 1 приведены результаты тестирования моделей.

Наилучшим алгоритмом является тот, у которого точность прогнозирования наибольшая. На основании полученных результатов выберем для каждой агрегированной таблицы алгоритм построения модели классификации. Для первой агрегированной таблицы с данными наилучшим алгоритмом является Adaptive Bayes Network с результатом точности 74,75%, для второй таблицы – алгоритм Adaptive Bayes Network с точностью 51,59%. Так как для третьей таблицы наибольшее значение точности составляет 36,42%, что относится к прогнозированию низкой точности, то она не будет участвовать в дельнейшем анализе данных.

Построим модель классификации для выявления значимых признаков с помощью алгоритма байесовских сетей (Adaptive Bayes Network) на основе агрегированной таблицы T2, на которой результаты тестирования показали наивысшую точность (74,75%). В таблице 2 представлены значимые признаки, выявленные в результате построения модели классификации.

Формализованная модель имеет следующий вид:

$$\mathrm{Risk} = F(x_1..x_4),$$

$$\mathrm{Risk} \geq 0, \mathrm{Risk} \leq 1$$

$$\mathrm{Risk} \to 0,$$

где $\mathrm{Risk}$ – уровень риска неуплаты таможенных платежей,

$x_1..x_4$ – значения признаков в ГТД.

В результате построения модели сформировано 39 правил. Приведем пример значений

атрибутов класса, в который входят участники ВЭД с нарушением срока уплаты таможенных платежей.

$$\text{Risk} = F(x_1..x_4) = F(400011, ИМ3, 198035) = 0.994.$$

Данная запись означает, что если производится выпуск для внутреннего потребления товаров, перемещаемых в виде отдельных компонентов и оформленных по единому коду товарной номенклатуры ВЭД России, с направлением перемещения «Импорт» и при этом таможенная стоимость не корректировалась и принята в сроки выпуска товаров, то риск несвоевременной оплаты таможенных платежей составляет 0,994 и в этом случае необходима дополнительная проверка платежеспособности участника ВЭД.

**Выводы.** В статье, посвященной вопросу оценки уровня риска неуплаты таможенных платежей, проведен анализ методов оценки, который показал обоснованность применения метода интеллектуального анализа данных, и в частности технологии Data Mining.

Для построения модели оценки риска неуплаты таможенных платежей был выбран метод классификации, который относится к группе методов «обучения с учителем». В качестве обучающего атрибута принадлежности классу выступает «Признак наличия нарушения срока уплаты таможенных платежей».

Для построения модели на основе метода классификации использовано инструментальное средство Oracle Data Miner. В результате проведенного эксперимента выбран алгоритм байесовских сетей (Adaptive Bayes Network), который обеспечивает наибольшую точность построенной модели (74,75%). В результате построения модели были определены признаки, влияющие на оценку риска неуплаты таможенных платежей: таможенная процедура; направление перемещения товара; признак корректировки таможенной стоимости; код, определяющий почтовую зону декларанта.

Поставленная цель статьи была достигнута. Полученные результаты исследования были представлены для изучения и практического применения в таможенные органы. Применение полученной модели позволит проводить оценку риска неуплаты таможенных платежей и сократить число не поступивших платежей от участников ВЭД.

**Список использованной литературы:**

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998.– 5 с.
2. Баргесян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP.– BHV-СПб., 2008.
3. Бершадский А.М., Бождай А.С. Разработка методов информационно-аналитического обеспечения процесса подготовки и переподготовки государственных и муниципальных служащих в области информационных технологий с учетом социально-экономической специфики района//Открытое образование, 2008 – с. 23-33.
4. Дюк В.А., Самойленко А.П. Data Mining: учебный курс // СПб.: Питер, 2001. -53 с.
5. Евсюков В.В. Интеллектуальный анализ данных в банковской деятельности// «Банковское дело» №7 2006г. с. 42-46.
6. Лагоша Б.А. Моделирование рисковых ситуаций в экономике и бизнесе.-М.: Финансы и статистика, 2003.– 5 с.
7. Тельнов Ю.Ф. Интеллектуальные информационные системы в экономике. Учебное пособие. — М.: СИНТЕГ, 2006.-5 с.
8. Inmon W. H. Building the Data Warehouse. – Wiley, 2006.
9. Oracle Database Documentation Library 10g Release 2. Data Mining Application Developer's Guide.

Сведения об авторе: Бегутова Светлана Владимировна, сотрудник кафедры прикладной информатики в экономике Московского государственного университета экономики, статистики и информатики (МЭСИ), 119501, Москва, ул. Нежинская, 7, тел.: (495) 4116633, 4427755, 4421203, (916) 7395958, e-mail: sbegutova@gmail.com

**Begutova S.V.**

**USING OF METHODS OF INTELLECTUAL ANALYSIS OF DATA FOR APPRAISAL OF RISKS OF CUSTOMS PAYMENTS NON-PAYMENT**

In this article the author regards the methods of appraisal of risks of customs payments non-payment, describes the construction of the model of risks appraisal on the base of using of intellectual analysis of data and method of classification. General indexes influenced on the appraisal of risks of customs payments non-payment are distinguished here.

Key words: risk, methods of intellectual analysis of data, Data Mining.