косрочных прогнозов поведения финансовых рынков с помощью моделей, учитывающих их фрактальные свойства.

СПИСОК ЛИТЕРАТУРЫ

1. *Мандельброт Б.* (Не)послушные рынки: фрактальная революция в финансах. М. : Издательский дом «Вильямс», 2006. 408 с.

2. *Прудский М. В.* Фрактальный анализ финансовых рынков // Информационные системы и математические методы в экономике: сб. науч. тр. / общ. ред. М. В. Радионовой; Пермь : Перм. гос. нац. иссл. ун-т, 2012. Вып. 5. С. 109-120.

3. *Кривоносова Е. К.* Сравнение фрактальных характеристик временных рядов экономических показателей // Современные проблемы науки и образования. 2014. № 6. [Электронный ресурс]. URL: http://www.science-education.ru/ru/article/view?id=15974 (дата обращения: 05.08.2018).

4. *Мансуров А. К.* Прогнозирование валютных кризисов с помощью методов фрактального анализа // Проблемы прогнозирования. 2008. № 1. С. 145–158.

5. *Дубовиков М. М.* Эконофизика и фрактальный анализ финансовых временных рядов // Успехи физических наук. 2011. № 181 (7). С. 779–786.

6. *Белолипцев И. И.* Предсказание финансовых временных рядов на основе индекса фрактальности // Мир Науки. 2014. № 3. [Электронный ресурс]. URL: https://mir-nauki.com/issue-3-2014.html (дата обращения: 05.08.2018).

ПРИМЕНЕНИЕ АЛГОРИТМОВ DATA MINING ДЛЯ АНАЛИЗА ДАННЫХ В СФЕРЕ КРЕДИТОВАНИЯ

И. А. Задворная, О. М. Ромакина

Саратовский государственный университет, Россия
E-mail: zadvornayaia@gmail.com, akrifina1@gmail.com

Статья посвящена применению алгоритмов Data Mining для изучения данных по кредитным картам клиентов банка. Рассматривается решение проблемы хранения и обработки информации большого объёма. Автором ставится задача выявления характеристик клиентов, влияющих на вероятность задолженности по кредитным картам. Исходя из поставленной задачи, к многомерной структуре данных применяются различные алгоритмы. Для анализа данных, рассматриваемых в данной статье, используются деревья решений, кластеризация и нейронная сеть.

APPLICATION OF DATA MINING ALGORITHMS FOR THE ANALYSIS OF DATA IN LENDING

I. A. Zadvornaya, O. M. Romakina

The article is devoted to the application of Data Mining algorithms for studying data on credit cards of bank customers. It examines the solution of the problem of storing and processing information of a large volume. The author puts the task of identifying the characteristics of customers that affect the probability of debt on credit cards. Based on the task, different algorithms are applied to the multi-dimensional data structure. Decision trees, clustering, and a neural network are used to analyze the data discussed in this article.

Современным миром правит информация. Традиционно аналитики решали задачу извлечения полезной информации из наборов данных. Но их растущий объем требует новых подходов к решению задач анализа в современных исследованиях. Одной из наиболее успешно применяемых технологий является Data Mining (интеллектуальный анализ данных). Эта технология, используя междисциплинарный подход, позволяет выявлять потенциальные взаимосвязи и шаблоны в исследуемых данных. Область применения знаний, извлечённых из первоначального набора большого объёма информации, обширна. Технология интеллектуального анализа данных особенная актуальна в банковской сфере и бизнесе, поскольку именно в этих областях важна структуризация и возможность прогнозирования поведения клиентов и возможных конкурентов на основе имеющихся данных. Это объясняет актуальность выбранной мной темы работы.

Существуют различные программные продукты, в состав которых входят библиотеки, реализующие алгоритмы Data Mining. Одним из таких средств является разработка компании Microsoft (Microsoft SQL Server with Analysis Services). Данное программное средство позволяет создавать многомерные модели с различными вариантами хранения данных и предоставляет расширенные возможности для анализа [1].

Определяющая задача перед началом анализа – это подготовка данных и создание подходящей структуры хранения. Для этой цели построим многомерную базу данных. Основное преимущество её построения – увеличение возможности процесса извлечения знаний из данных [2].

Многомерная структура данных состоит из таблиц фактов и измерений. Компонентами таблицы измерений являются несколько полей с названиями и одно целочисленное для идентификации (ключ). Основная сущность представляет собой таблицу фактов. Она содержит имена фактов или мер, ключевые поля для каждого измерения и несколько числовых полей [3].

Многомерная база данных в данном исследовании построена на основе информации по использованию кредитных карт банка [4]. Информация в наборе относится к различным временным промежуткам взаимодействия клиентов с кредитными услугами банковского учреждения. Многомерная база данных, используемая в системе, построена по схеме «Звезда» [5]. В составе системы реализованы следующие таблицы измерений: Пол, Возраст, Образование, Семья, Кредит, Счёт, История платежей, Время.

Определим основные характеристики клиентов, влияющие на вероятность неуплаты средств по кредитным картам, с помощью алгоритмов Data Mining (дерево решений, кластеризация, нейронная сеть), представленных в рамках программного продукта Microsoft [6], и применим их для анализа данных, хранящихся в многомерной структуре.

Проведём анализ данных с помощью алгоритма Data Mining «Дерево решений». Он применяется во многих реальных приложениях для решения проблемы классификации. Она представляет собой процесс обучения функции, которая отображает элемент данных в один из нескольких предопределенных классов.

Одним из распространённых алгоритмов, реализующих деревья решений, является CART. Аббревиатура CART (Classification and Regression Trees) расшифровывается как «Деревья классификации и регрессии». Алгоритм вырабатывает бинарные деревья и продолжает расщепление до тех пор, пока могут быть найдены новые, которые улучшают решение [7]. Каждый корневой узел модели CART представляет собой входную переменную x и точку разделения для этой переменной. Листовые узлы дерева содержат выходную переменную y, которая используется для прогноза значений.

Для классификации используется индексная функция Gini, которая дает представление о том, какие узлы стоит выбрать:

$$G(T) = 1 - \sum_{i=1}^{n} p_i^2 ,$$

где набор данных T содержит данные n классов, G – индекс Gini, p_i – доля обучающих экземпляров с классом i в интересующем разбиении.

Алгоритм CART обеспечивает основу для работы других алгоритмов. Регрессионный подход к построению деревьев решений, похожий на данный базовый метод используется для анализа в данном исследовании.

Установим, существует ли взаимосвязь между личными характеристиками (измерения «Образование», «Семья», «Пол», «Возраст») и возвратностью кредита по карте. В качестве выходного поля выбран основной показатель долговых обязательств у клиента перед банком в течение полугода обслуживания.

Для построенного дерева решений из 4 параметров, выбранных для анализа, на результат влияют только 3 (характеристики клиентов, связанные с образованием, полом и возрастом).

Корень дерева показывает весь стартовый набор данных перед их разделением. Первоначальное разбиение проводится по параметру образование. Для узла, определяющего класс 4 (другие виды образования) дальнейшего разделения не производится. Для класса 3 в образовании разбиение проводится по параметру гендерной принадлежности. Согласно полученным данным среди людей со средним общим образованием вероятность того, что женщина станет должником по кредитной карте, составляет около 63%, мужчина – 24%. Согласно классу 2 для людей с высшим образованием в возрасте до 27 или старше 45 склонность к задолженности составляет меньше 5%. А среди клиентов от 27 до 45 лет среди мужчин вероятность составляет около 55%, среди женщин этот показатель на порядок ниже. Среди людей с высшим образованием (аспирантура) в целом наблюдается наименьшая склонность к задолженности по рассматриваемым данным.

Перейдём к анализу данных с помощью алгоритма «Кластеризация». Кластерный анализ представляет собой набор методологий автоматической классификации в несколько групп с использованием меры ассоциации таким образом, что выборки только внутри одной группы схожи.

Наиболее часто используемая стратегия кластеризации основана на критерии среднеквадратичной ошибки [8]. Основная задача состоит в её минимизации. Если множество N элементов в n-мерном пространстве каким-то обра-

зом было разбито на K-кластеров $\{C_1, C_2, ..., C_k\}$ по n_k элементов, где элемент содержится только в одном кластере так, что $\sum n_k = N$, где $k = 1, ..., K$, средний вектор M_k кластера C_k определяется как центроид кластера:

$$M_k = \left(\frac{1}{n_k}\right)\sum_{i=1}^{n_k} x_{ik},$$

где x_{ik} – i-й элемент, принадлежащий кластеру C_k. Отклонение для всего кластерного пространства представляет собой:

$$E_k^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - M_k)^2.$$

Цель метода кластеризации – минимизация E_k^2 для заданного кластера K.

Алгоритм разделения кластеров K-mean (K-средних) является наиболее часто используемым алгоритмом, использующим критерий среднеквадратичной ошибки. Работа алгоритма начинается с выбора начального разбиения кластера, содержащего случайно выбранные элементы, и вычисления центроида кластеров. После этого создаётся новое разбиение, назначив каждый элемент ближайшему центру кластера. На последнем шаге вычисляют новые кластерные центры как центроиды кластеров. Шаги 2 и 3 повторяются до тех пор, пока не будет найдено оптимальное разбиение [9].

Проведение кластеризации на всём объёме выборки по всем параметрам позволяет увидеть основные особенности информации в системе. В результате построения получено 10 кластеров. Рассмотрим наиболее информативные из них.

Кластер №8 определяется тем, что данному кластеру соответствуют данные по клиентам с самым большим кредитным лимитом. Они показывают наименьшее количество задолженностей.

Кластеры №6, №7 и №10 содержат часть данных с наибольшим количеством задолженностей у клиентов.

Клиенты, относящиеся к кластеру №6, характеризуются наличием выхода за границы установленного кредитного лимита на протяжении рассматриваемого периода. При этом у кредитных карт этой группы установлен невысокий лимит. К таким клиентам с большей вероятностью относятся одинокие люди с высшим образованием независимо от пола.

У клиентов кластера № 7 на начало рассматриваемого периода кредитная история считается положительной, но к концу периода наступает резкий выход за границы лимита и превышение месячных трат с последующим увеличением уровня задолженности. При этом кредитный лимит может быть различным по величине. К этой группе с большей вероятностью относятся замужние женщины.

Для кластера №10 характерно большое движение средств на счету. При этом наблюдается периодическая задолженности по счету в связи с постоянным высоким, близким к лимиту оборотом средств. Кредитный лимит варьируется от среднего к более высокому.

Вторая модель лучше, чем предыдущая, показывает взаимосвязь между параметрами и позволяет анализировать картину в целом, выявляя взаимосвязи как между персональными, так и поведенческими характеристиками клиентов.

Теперь воспользуемся последним алгоритмом «Нейронная сеть» для анализа данных. Нейронная сеть представляет собой сетевую структуру, состоящую из нескольких узлов, соединенных через направленные каналы. Каждый узел представляет собой блок обработки, а связи между узлами определяют отношения [10].

В математических терминах искусственный нейрон определяется следующими обозначениями. Имеется несколько входов $x_i, i = 1,...,m$. Каждый вход x_i умножается на соответствующий ему вес w_{ki}, где k – индекс данного нейрона в рассматриваемой сети. Веса имитируют биологические связи в естественном нейроне. Сумма произведений $x_i * w_{ki}$ для $i = 1,...,m$ обозначается *net* и выражается следующим образом: $net_k = x_1 w_{k1} + x_2 w_{k2} + ... + x_m w_{km} + b_k$.

Нейрон вычисляет значение выхода y_k как определенную функцию значения net_k: $y_k = f(net_k)$. Функция f называется функцией активации. Для слоя скрытых нейронов это функция гиперболического тангенса:

$$y = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}}.$$

Сигмоидальную функцию будем использовать как функцию активации для слоя выходных нейронов:

$$y = \frac{1}{(1 + e^{-net})}.$$

Построенная нейронная сеть состоит из десяти входных нейронов. В качестве входных параметров используются персональные характеристики клиентов и показатели проводимых операций по кредитной карте. В модели также используется один скрытый слой с пятью полностью связанными нейронами и два выходных нейрона. Выходы предназначены для прогнозирования вероятности положительной/отрицательной кредитной истории в дальнейшем.

С помощью построенной нейронной сети были подтверждены рассматриваемые далее предположения, что доказывает корректность построенной модели. Первое предположение состоит в том, что клиенты с высоким кредитным лимитом в целом наименее склонны к невозвратности кредита в связи с возможностью свободно оперировать в рамках большого объёма средств. Второе предположение состоит в том, что клиенты со средним кредитным лимитом и склонностью к большому обороту средств по карте склонны к невозвратности кредита в связи с постоянным превышением разрешённого объема используемых средств. Рассмотрена также ситуация с резким уходом в отрицательный баланс средств на конец рассматриваемого периода. По результатам анализа подобное клиентское поведение говорит о резком скачке в сторону большой единовременной задолженности с последующим возвращением средств в течение не менее 4 месяцев (при дополнительных ограничениях и рассмотрении скачков на более ранних периодах).

Благодаря высокой скорости получения результатов работы модели, терпимости к сложности данных нейронных сетей, относительной зависимости системы на этапе построения от разработчика, а также гибкости и компактности удалось разработать систему, основанную на изучаемых данных, которая способна на основе краткосрочных наблюдений за поведением клиента предсказать вероятность положительной кредитной истории в дальнейшем.

СПИСОК ЛИТЕРАТУРЫ

1. *MacLennan J., Crivat B.* Data Mining with Microsoft SQL Server 2008 / John Wiley and Sons. 2009. 672 p.
2. *Inmon W. H.* Building the Data Warehouse, Third Edition / Wiley Computer Publishing, 2002. 576 p.
3. *Hawkins H.* Data Warehousing Architecture and Implementation / New Jersey : Prentice Hall PTR, 1998. 362 p.
4. *Lichman M.* UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science [Electronic resource]. URL: http://archive.ics.uci.edu/ml (accessed: 03.05.2018).
5. *Berson A., Smith S.* Data Warehousing, Data Mining, and OLAP / N.Y. : McGraw Hill Education, 2017. 638 p.
6. *Zaki M.* Data mining and analysis: Fundamental concepts and algorithms / Cambridge : Cambridge India, 2016. 567 p.
7. *Han J., Kamber M.* Data Mining: Concepts and Techniques / Waltham : Elsevier, 2006. 772 p.
8. *Kantardzic M.* Data Mining: Concepts, Models, Methods, and Algorithms / N.Y. : John Wiley and Sons, 2003. 343 p.
9. *Kovalerchuk B., Vityaev E.* Data Mining in Finance: Advances in Relational and Hybrid Methods / N.Y. : Kluwer Academic Publishers, 2002. 322 p.
10. *Witten I. H., Frank E., Hall M. A.* Data Mining: Practical Machine Learning Tools and Techniques / Waltham : Morgan Kaufmann, 2016. 654 p.

МЕТОДЫ АППРОКСИМАЦИИ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

Д. А. Зайнышева

Саратовский государственный университет, Россия
E-mail: d.zaynysheva@gmail.com

В регрессионном анализе нередко приходится сталкиваться с ситуацией, когда среди объясняющих переменных наблюдается зависимость. Тогда говорят о наличии мультиколлинеарности. В таких ситуациях МНК-оценки формально существуют, но обладают «плохими» статистическими свойствами. Регуляризация — метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу. Эта информация часто имеет вид штрафа за сложность модели. Методы регрессии Ridge и Lasso осуществляют регуляризацию параметров и позволяют преодолеть некоторые недостатки метода наименьших квадратов.

Методы регуляризации зачастую позволяют добиться уменьшения дисперсии прогноза за счет незначительного увеличения его смещения. В результате точность прогноза растет.