МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА РАСПРЕДЕЛЕННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ АНАЛИЗА ФУНКЦИОНИРОВАНИЯ ИКТ-КОМПАНИЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Акмаева Виктора Сергеевича

Научный руководитель
доцент, к.э.н. _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2021

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Решение задач анализа прогнозирования развития социально-экономических объектов ориентировано на использование накопленных статистических данных. Однако наличие такой информации не гарантирует, что все факторы приняты во внимание. Принципы и назначение систем, обеспечивающих поддержку управленческих решений в условиях неопределенности, характерных для экономической деятельности, наиболее соответствуют методам интеллектуального анализа данных, которые позволяют обрабатывать неточную, трудно формулируемую информацию, выявлять закономерности и обобщать знания, моделируя экспертную деятельность.

Data Mining является важной технологией для анализа деятельности большого числа компаний, предприятий и организаций. Такой подход способен обеспечить поддержку принятия решений в сферах, связанных с функционированием этих организаций. Использование данной технологии позволит оценить эффективность различных компаний и отрасли в целом.

Целью выпускной квалификационной работы является разработка информационно-аналитической системы (ИАС) для анализа компаний ИКТ-сектора. Для достижения поставленной цели должны быть выполнены следующие задачи:

- выбор инструментальных средств для разработки ИАС;
- проектирование и разработка архитектуры ИАС, разработка интерфейса приложения;
- сравнительный анализ методов Data Mining;
- применение алгоритмов кластеризации и классификации;
- апробация информационно-аналитической системы для компаний Приволжского федерального округа.

Предметом исследования является реализация математических моделей для решения прогностических задач в рамках информационно-аналитической системы.

Объектом исследования в рамках магистерской работы является моделирование с помощью разработанных инструментальных средств функционирования ИКТ-сектора.

Элементы новизны в данном исследовании заключаются в практической реализации методики анализа показателей социально-экономического

развития, которая реализует кластеризацию и классификацию данных, связанных с экономическим развитием ИКТ-сектора.

В качестве научной значимости работы можно указать возможность использования актуальных методов Data Mining для прикладной задачи, связанной с возможностью дополнительной оценки деятельности ИКТ-компаний.

В первом разделе выпускной квалификационной работы описываются особенности использования информационно-аналитических систем, их назначение, а также произведен анализ показателей деятельности российских ИКТ-компаний. Во втором разделе описываются методы интеллектуального анализа данных, а именно, кластеризационная и классификационная задачи. Также в этом разделе рассмотрена реализация данных моделей на языке программирования Python. В третьем разделе представлена разработанная ИАС и ее применение для оценки деятельности ИКТ-компаний Приволжского федерального округа.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Особенности использования информационно-аналитических систем» описаны понятие ИАС, анализ Российского рынка ИАС, предложен набор показателей для оценки деятельности ИКТ-компаний Российской Федерации.

Информационно-аналитические системы становятся важным механизмом для анализа деятельности большого числа компаний, предприятий, отрасли в целом. Подобные системы позволяют реализовать создание аналитической отчетности на основе собранных данных. Такой подход способен обеспечить поддержку принятия решений в сферах, связанных с функционированием экономических объектов. Создание информационно-аналитической системы на основе информации о деятельности компаний в ИКТ-секторе позволит оценить их эффективность, а также оказать руководителям различного уровня помощь в принятии управленческих решений.

ИАС призваны обеспечить первичный сбор информационных потоков и систематизировать формирование исходных баз данных, выполнить быстрый отбор тематической информации из большого объема данных и выдачу ее аналитику по запросу или в постоянном режиме по заданному алгоритму [1].

Анализ существующих ИАС показывает, что они реализуют методы интеллектуального анализа данных в ограниченном объеме. Традиционные ИАС в недостаточной мере включают в себя средства бизнес-аналитики, поэтому разрабатываемая ИАС представляет собой попытку совместить простоту использования и возможность применения различных методик интеллектуального анализа данных.

Во втором разделе «Методика оценки ИКТ-компаний с помощью Data Mining» дана сравнительная характеристика кластеризационных и классификационных методов интеллектуального анализов данных. Представлены возможности платформы Python по реализации методов Data Mining.

Для реализации методов интеллектуального анализа были взяты данные по компаниям ИКТ-сектора Приволжского федерального округа. Компании и организации, по которым были собраны данные, взяты из базы данных информационно-аналитической системы FIRA. Количество составило 11619 компаний за 2016, 2017, 2018 и 2019 года. Количественная информация пред-

ставляет собой результаты работы компаний за год, по которым производился анализ деятельности компаний, а именно: результаты исследований и разработок, тыс руб; внеоборотные активы, тыс руб; оборотные активы, тыс руб; уставный капитал, тыс руб; капитал и резервы, тыс руб; долгосрочные обязательства, тыс руб; краткосрочные обязательства, тыс руб; совокупная величина чистых активов, тыс руб; доходы, тыс руб; выручка (нетто) от продажи, тыс руб; расходы по обычной деятельности, тыс руб; себестоимость проданных товаров, продукции, работ, услуг, тыс руб; валовая прибыль, тыс руб; прибыль от продаж, тыс руб; доходы от участия в других организациях, тыс руб; прибыль до налогообложения, тыс руб; чистая прибыль, тыс руб; совокупный финансовый результат периода, тыс руб; поступления – всего от текущих операций, тыс руб.

Целью ИАС является формирование исходных баз данных, обеспечение первичного анализа данных по компаниям, формирование запросов, применение отдельных алгоритмов Data Mining.

Предлагаемая методика заключается в проведении кластерного анализа и дальнейшего использования меток кластеров для формирования обучающей выборки, на основе которой будет осуществлено построение классификационной модели. Кластеризация также используется для разделения объектов на группы, но изначально классы объектов не предопределены. Классификация используется, когда целью работы является разбиение множества объектов или наблюдений на заданные группы. Решение получается на основе анализа значений атрибутов [2].

Для реализации кластерного анализа был использован метод k-средних – один из наиболее широко используемых методов кластеризации из-за его простоты и скорости. Он разбивает данные на k кластеров, назначая каждому объекту его ближайший центроид кластера (среднее значение переменных для всех объектов в этом конкретном кластере) на основе используемой меры расстояния. Он более устойчив к различным типам переменных. Кроме того, удобен для обработки больших наборов данных, которые часто используются при сегментации [3].

Поскольку кластеризация используется в основном неконтролируемым образом, необходимо иметь меру для оценки качества кластеров, предоставляемых конкретным алгоритмом. Существует два вида оценки кластерного

анализа: внутренние и внешние. Внутренние меры отображают качество кластеризации только по информации в данных. Внешние меры основаны на сравнении результата кластеризации с известным разделением на классы.

Существует множество методов классификации, которые используют различный математический аппарат и различные подходы при реализации [4]. Можно выделить следующие типы методов классификации: вероятностные, метрические, логические, линейные, логическая регрессия.

Метод построения деревьев решений является эффективным инструментом интеллектуального анализа данных и предсказательной аналитики [5].

Для того, чтобы определить качество построенной классификационной модели существуют множество видов численной оценки алгоритма:

— confusion matrix
— accuracy;
— precision;
— recall;
— F-мера.

Для реализации классификации было принято решение использовать метод деревьев решений. В качестве обобщающей характеристики точности классификационной модели предлагается использовать Accuracy.

Для того, чтобы реализовать методы кластерного и классификационного анализа был использован язык программирования Python 3.9.5 [6]. Были использованы библиотеки sklearn и SciPy, которые позволяют применять методы и алгоритмы Data Mining. Для визуализации данных была использована библиотека pandas.

Разработанная информационно-аналитическая система должна реализовывать ряд стандартных этапов анализа данных. Перед этим необходимо произвести начальную предобработку данных, а именно удаление записей о компаниях, которые имеют все нулевые показатели по категориям баланса, с целью улучшения качества анализируемой выборки.

Для составления более точной кластерной модели необходимо определить парные корреляции, на основе которых будет строиться отбор показателей модели. Результатом данного действия стало уменьшение числа оцениваемых показателей, путем включения в модель слабокоррелированные по-

казателей между собой и коррелирующие с зависимой переменной [7].

Прежде чем применять методы кластерного анализа, необходимо произвести нормализацию выборки. На практике наиболее распространены следующие методы нормализации признаков:

— минимакс – линейное преобразование данных в диапазоне, как правило [0..1], где минимальное и максимальное масштабируемые значения соответствуют 0 и 1 соответственно;
— Z-масштабирование данных на основе среднего значения и стандартного отклонения: деление разницы между переменной и средним значением на стандартное отклонение;
— десятичное масштабирование путем удаления десятичного разделителя значения переменной [8].

В данном исследовании использовались три вида нормализации: MinMax Scaler – изменяет масштаб набора данных так, чтобы все значения функций находились в диапазоне, как правило $[0, 1]$; Normalizer – изменяет масштаб вектора для каждой выборки, чтобы иметь единичную норму, независимо от распределения выборок; StandardScaler – удаляет среднее и масштабирует данные до единичной дисперсии.

По результатам построения кластерной модели, было принято решение использовать модель, в которой данные были нормализованы при помощи метода Normalizer, и котороая состоит из трех кластеров, такой выбор обусловлен результатами равномерного распределения данных по кластерам и визуального анализа по графикам.

Для трех различных способов нормализованных данных было построены три различные кластерные модели. Количество кластеров для моделей определялось на основе метода локтя, кроме того необходимо было оценить количество объектов, попадающих в каждый из кластеров. Для кластерной модели использовался индекс Дэвиса-Болдина, индекс Калински-Харабаша и Силуэт [9].

Анализ описательной статистики по кластерам позволяет интерпретировать полученные множества предприятий следующим образом: в нулевой кластер попали компании, о которых можно сказать, что они являются развивающимися; в первом кластере находятся компании, которые стагнируют; во второй кластер попали стабильные компании, которые имеют лучшие по-

казатели балансов, среди других.

С учетом измененных параметров удалось получить подходящую модель. Точность полученной модели составляет 0.876 для обучающей выборки и 0.877 – для тестовых. Количество конечных узлов – 5, глубина дерева – 4. В результате была построена классификационная модель, которая позволяет определить, к какому классу будет относиться компания ИКТ-сектора и позволит сделать вывод о перспективах развития отдельных компаний и в целом ИКТ-сектора.

В третьем разделе «Применение ИАС для компаний ИКТ-сектора» описана разработанная ИАС для анализа деятельности компаний ИКТ-сектора, представлена апробация ИАС на примере ИКТ-компаний Приволжского федерального округа.

Разрабатываемая ИАС должна обеспечивать реализацию следующих функций: конвертация и хранение данных; добавление новых компаний; выбор методов нормализации; настройка параметров метода кластеризации; настройка параметров метода классификации; формирование описательной статистики по исходным набором данных; формирование набора данных для классификации; вывод результатов в файл; визуализация результатов; оценка динамики развития по категориям компании.

На этапе проектирования ИАС для оценки ИКТ-сектора была разработана ERD-диаграмма. Диаграмма состоит из девяти сущностей, представляющих следующие сведения: общая информация о компаниях, контактная информация, реквизиты, информация о виде деятельности и бухгалтерская отчетность. При конвертации данных были использованы возможности языка Python и его библиотек Pandas и pymysql.

Создание приложения на данном этапе подразумевает создание графического интерфейса. Разработка была осуществлена с использованием HTML5, CSS3 и JavaScript ES6. Для настройки связи приложения, написанного на языке Python существует библиотека Eel, которая позволяет настроить связь между графическим интерфейсом, написанным на HTML, CSS и JavaScript, и логической частью приложения. При разработке интерфейса приложения было достигнуто корректное согласование графических элементов с логической частью ИАС.

Графический интерфейс приложения позволяет производить настрой-

ку параметров для построения классификационной модели. У пользователя есть возможность настроить: процентное отношение количества элементов в тестовой выборке к количеству элементов в тренировочной; критерий разделения данных; максимальную глубину дерева; максимальное количество узлов.

Разработанное приложение позволит реализовать в удобной форме процедуру классификации объектов на основе предварительно построенной кластеризационной модели.

Использование данного приложения позволило произвести оценку динамики развития ИКТ-сектора. Для апробации методики классификации данных были взяты данные по компаниям Приволжского федерального округа. Для этого рассматривались структуры кластеров по компаниям за 2017, 2018, 2019 годы.

По данным результатам можно сказать, что динамика не наблюдается, так как значительных изменений по кластерам не произошло. В 2018 году наблюдается увеличение числа ИКТ-компаний, которые демонстрируют устойчивое развитие по ряду базовых показателей баланса. В 2019 году количество таких компаний резко уменьшилось.

ЗАКЛЮЧЕНИЕ

В ходе разработки ИАС был осуществлен сравнительный анализ и выбор инструментальных средств, в результате которого для создания базы данных использован MySQL. Для разработки логической части информационно-аналитической системы использовался язык Python, для реализации методов интеллектуального анализа данных использовались существующие библиотеки Python. Для разработки интерфейса была применена библиотека Eel, которая используется для настройки связи логической части приложения, написанной на языке Python, с графическим интерфейсом, написанным на HTML, CSS и JavaScript.

Была спроектирована и разработана архитектура информационно-аналитической системы, которая позволяет:

— конвертировать и хранить данные о компаниях;
— добавлять новые компании;
— выбирать методы нормализации;
— настраивать параметры кластеризации;
— настраивать параметры классификации;
— формировать описательную статистики по исходным наборам данных;
— формировать обучающие выборки для классификации на основе кластерной модели;
— выводить результаты анализа в файл;
— визуализировать результаты;
— оценивать динамику развития ИКТ-компаний.

Был произведен сравнительный анализ методов Data Mining, в результате которого метод k-средних использован для построения кластерной модели, метод деревьев решений – для построения классификационной модели.

Разработанное приложение позволяет применять алгоритмы кластеризации и классификации для дальнейшего отслеживания динамики развития компаний ИКТ-сектора. Интерфейс приложения позволяет совместить простоту использования и возможность применения различных методик интеллектуального анализа данных.

Для компаний ИКТ-сектора Приволжского федерального округа были построены кластерная и классификационная модели, что позволило оценить динамику развития в этой отрасли для регионов Приволжского федераль-

ного округа. Результаты показали, что динамика не наблюдается, так как значительных изменений по кластерам не произошло. Дальнейшее направление исследования может включать анализ различных групп показателей с учетом рассмотрения компаний не только в отдельных регионах, но и по Российской Федерации в целом.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Business Intelligence (BI) Tools Reviews 2021 | Gartner Peer Insights [Электронный ресурс] // https://www.gartner.com/ – Global Research and Advisory Company | Gartner URL: https://www.gartner.com/reviews/market/analytics-business-intelligence-platforms (дата обращения: 26.05.2021). Загл. с экрана. Яз. англ.

2 Барсегян, А.А. Технологии анализа данных: Data Mining, Text Mining, Visual Mining, OLAP. 2 изд / –Санкт-Петербург: БХВ-Петербург, 2008. –384с.

3 Dean, J. Big Data, Data Mining, and Machine Learning (Wiley and SAS Business Series) / J. Dean. –New York: Wiley, 2014. –288 p.

4 Aggarwal, C. Data Classification: Algorithms and Applications / C. Aggerwal. –New York: Yorktown Heights, 2020. –703p.

5 Aggarwal, C. Data Classification: Algorithms and Applications / C. Aggerwal. –New York: Yorktown Heights, 2020. –703p.

6 About Python™ | Python.org [Электронный ресурс] // https://www.python.org/ – Welcome to Python.org URL: https://www.python.org/about/ (дата обращения: 13.05.2021). Загл. с экрана. Яз. англ.

7 Описательная статистика [Электронный ресурс] // http://www.machinelearning.ru/ – Заглавная страница URL: http://www.machinelearning.ru/wiki/index.php?title=Описательная_статистика (дата обращения: 13.05.2021). Загл. с экрана. Яз. рус.

8 sklearn.preprocessing.MinMaxScaler — scikit-learn 0.24.1 documentation [Электронный ресурс] // https://scikit-learn.org/stable/ – scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (дата обращения: 10.05.2021). Загл. с экрана. Яз. англ.

9 Davies, D.; Bouldin, D. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence // PAMI-1 (2), 1979. –P.224-227.