Интернет-журнал «Отходы и ресурсы» https://resources.today
Russian Journal of Resources, Conservation and Recycling

2021, №1 Том 8 / 2021, No 1, Vol 8 https://resources.today/issue-1-2021.html

URL статьи: https://resources.today/PDF/06INOR121.pdf

DOI: 10.15862/06INOR121 (http://dx.doi.org/10.15862/06INOR121)

Ссылка для цитирования этой статьи:
Максимов В.Е., Резникова К.М., Попов Д.А. Информационные технологии для анализа данных морского флота // Интернет-журнал «Отходы и ресурсы», 2021 №1, https://resources.today/PDF/06INOR121.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI: 10.15862/06INOR121

For citation:
Maximov V.E., Reznikova K.M., Popov D.A. (2021). Data mining for marine data analysis. *Russian Journal of Resources, Conservation and Recycling*, [online] 1(8). Available at: https://resources.today/PDF/06INOR121.pdf (in Russian) DOI: 10.15862/06INOR121

Максимов Валерий Евгеньевич
ФГОУ ВО «Дальневосточный федеральный университет», Владивосток, Россия
Студент-магистрант
E-mail: valep199778@gmail.com

Резникова Ксения Михайловна
ФГОУ ВО «Дальневосточный федеральный университет», Владивосток, Россия
Студент-магистрант
E-mail: a-da_97@mail.ru

Попов Дмитрий Александрович
ФГОУ ВО «Дальневосточный федеральный университет», Владивосток, Россия
Студент-магистрант
E-mail: dmppda@gmail.com

Информационные технологии для анализа данных морского флота

Аннотация. Практически не осталось отрасли, где не использовались бы современные информационные технологии. Сегодня очень популярны подходы Data mining. Использование этой технологии позволяет преобразовывать огромные массивы данных в полезную информацию. В статье авторами представлено определение технологии Data mining и часто используемые методы. К одним из популярных методов интеллектуального анализа данных относятся классификация, кластеризация, машинное обучение и прогнозирование. Авторы уделили особое внимание такому методу кластеризации, как метод k-средних. Суть алгоритма заключается в распределении набора данных на кластеры. Готовые результаты можно визуализировать и невооруженным глазом обнаружить разбросы, которые подразумевают неоднородность данных. Исследуя далее эти разбросы, аналитик может найти ошибки и слабые места в изучаемой области согласно поставленной задаче.

В морской деятельности очень важно располагать точными и полными данными. В области судостроения анализ данных и грамотно принятые оперативные решения могут повлиять на скорость и качество постройки судов или даже снижение затрат производства. В сфере судоходства и логистики они могут быть использованы для оптимизации маршрутов и повышения безопасности моряков. Для эффективного применения data mining, как правило, требуются высококвалифицированные специалисты баз данных и программисты. В работе авторами продемонстрирован вариант применения программного средства Orange Data Mining.

Данная программа не требует от пользователя навыков программирования, что делает ее полезным инструментом для людей, далеких от написания программного кода.

В статье исследовано применение программы Orange Data Mining для автоматизированного интеллектуального анализа данных морского флота. Полученные результаты показывают, что программа может эффективно использоваться в морской деятельности.

Ключевые слова: интеллектуальный анализ данных; классификация; кластеризация; прогнозирование; метод k-средних; Orange Data Mining; морская деятельность

Введение

Сложно представить с какими объемами данных приходится взаимодействовать современным организациям и ученым. Количество данных растет с каждым днем и для преобразования их в информацию или знания необходимы эффективные инструменты и методы. Базы данных и системы управления базами данных (СУБД) обязательны для любой организации независимо от их профиля и направления, поскольку данные об оборудовании, персонале, товарах, предоставляемых услугах и т. д. обладают неизмеримой ценностью. Методом написания запросов к базе данных, например, используя язык SQL (Structured Query Language), можно преобразовывать хранящиеся данные в информацию, которая, в свою очередь, может быть использована для принятия решений различного рода. В качестве примера в [1] приводятся управление материально-техническими запасами, отслеживание тенденций развития компании, перераспределение полномочий и др.

Анализ данных основывается не только на инструментарии и технологиях, но и на самих методах вычислений и обработки информации. Использование SQL-запросов высококвалифицированным специалистом имеет большое значение, но современные информационные технологии могут предложить более мощные способы. Их эффективное применение также положительно скажется на навыках использовавшего их сотрудника. Такой технологией интеллектуального анализа данных является data mining.

Определение data mining

Data mining представляет собой слияние различных дисциплин, таких как СУБД, статистика, искусственный интеллект и машинное обучение. Зародилась эра приложений data mining в 1980 году и использовалась в основном при помощи исследовательских инструментов, ориентированных на решение конкретных задач [2].

Целью технологии data mining является выявление скрытых закономерностей или взаимосвязей между переменными в массивах необработанных данных. Как отмечено в [3], английский термин «data mining» на русский язык не имеет однозначного перехода. Если переводить дословно, то data mining это добыча/извлечение/вскрытие данных, поэтому оригинальное название используется в большинстве случаев. Иногда data mining переводят как интеллектуальный анализ данных, что является приемлемым.

В [4] data mining определяется как логический процесс, который используется для поиска полезной информации в большом объеме данных. Data mining включает в себя три основных этапа:

1. Исследование.
2. Построение модели.

3. Развертывание.

Первый этап начинается с подготовки данных. Исследование может включать в себя очистку данных, их преобразование, отбор подмножеств, анализ свойств переменных. В зависимости от поставленной задачи далее может происходить как выбор предикторов для регрессионной модели, так и разведочный анализ данных с применением графических и статистических методов. После того, как данные исследованы, уточнены и определены для конкретных переменных, вторым шагом является построение модели.

На этапе построения модели рассматриваются различные модели и выбирается наилучшая, исходя из их характеристик. Простыми словами, на данном этапе определяются и выбираются закономерности, дающие наилучший прогноз.

На завершающем этапе развертывания ранее определенная наилучшая модель применяется к новым данным для прогнозирования или оценки ожидаемых результатов.

Методы data mining

В связи со спросом на различные типы взаимодействия с данными, с их несовершенствами и возможностями оборудования и инструментария существует множество методов анализа данных. В data mining используются различные алгоритмы и методы, к которым относятся: искусственный интеллект, регрессия, классификация, кластеризация, деревья решений, правила ассоциативности, нейронные сети, генетические алгоритмы, метод k-средних, метод ближайших соседей и множество других.

Большой популярностью пользуются статистические методы и методы регрессии. К статистическим относятся диаграммы, гистограммы, средние значения и процентные ставки. Методы регрессии используют математические функции для моделирования и прогнозирования различных процессов и аппроксимации результатов.

К часто используемым также относятся ансамблевые методы: бустинг и бэггинг. Цель ансамблевых методов в том, чтобы улучшить прогнозирующие способности метода подбора модели. Бустинг является процедурой последовательного построения композиции алгоритмов машинного обучения в условиях, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов [5]. Таким образом, бустинг дает возможность вычислить более достоверные результаты, при этом уменьшив число ошибок. Такой эффект достигается при определении конкретного алгоритма для каждой части выборки. Как отмечают авторы в [6], с точки зрения качества классификации бустинг считается одним из наиболее эффективных методов. По крайней мере, применительно к классификации над деревом решений. Второй ансамблевый метод – бэггинг. Используя процедуры усреднения, метод позволяет определить класс объекта. Бэггинг основан на построении случайных подвыборок данных и наиболее эффективен применительно к нейронным сетям и деревьям решений.

В случае, когда данные неполны, применяется метод ближайших соседей. Недостающие значения свойств оцениваются на основе значений этих же свойств для других элементов. Эти элементы должны находиться в том же диапазоне либо в тех же условиях, что и искомые. Также элементы могут быть в одной группе по некоторым общим признакам. Лучшими «соседями» считаются те элементы, которые показывают ближайшее сходство по заданному классу паттернов.

Другая группа методов и алгоритмов data mining – методы агрегирования, позволяющие автоматизировать процессы классификации данных. Эта группа методов включает в себя классификацию, ассоциацию и кластеризацию. Классификация – это группировки записей по

заданным критериям, которые указываются пользователями или аналитиками. В процессе классификации могут использоваться деревья решений и семантические сети. Для нахождения элементов, связанных или похожих на другие элементы, применяется ассоциация. Часто используется следующее правило: если элемент *A* является компонентом какого-либо события, то элемент *B* также является компонентом того же события в *X %* случаев. Под кластеризацией понимается группировка записей по некоторым критериям, которые генерируются автоматически. Кластеризация позволяет обнаружить скрытые нарушения в базе данных, выявляя записи, не соответствующие ни одной группе [1].

В данной работе подробнее рассматриваются методы кластеризации для дальнейшего изучения их использования в анализе данных морского флота.

Методы кластеризации

По наиболее распространенной классификации методы кластеризации подразделяются на:

- масштабируемые и немасштабируемые;
- четкие и нечеткие;
- иерархические и плоские.

Масштабируемые алгоритмы обладают возможностью адаптации к ограниченным ресурсам вычислительной техники. С увеличением числа исследуемых записей масштабируемые алгоритмы обеспечивают линейный рост времени работы, в то время как немасштабируемые позволяют работать со всеми данными сразу, но крайне чувствительные к объему вычислительных ресурсов.

Четкие алгоритмы отличаются от нечетких степенью принадлежности. Четкие алгоритмы соотносят объекту всего один кластер, когда нечеткие ставят объекту в соответствие разные кластеры с разной степенью принадлежности. Применение в задачах кластеризации нечетких алгоритмов может содержать некоторые сомнения и неопределенности касательно принадлежности объекта кластеру.

Иерархические алгоритмы базируются на построении вложенных разбиений, в результате чего меньшие кластеры группируются в большие или же наоборот большие делятся на кластеры поменьше. В результате иерархической кластеризации образовывается дерево кластера, где корень дерева – это выборка, а листья – более мелкие кластеры. Цель плоских алгоритмов в общем разбиении объектов на кластеры [7].

К плоским алгоритмам кластеризации относятся алгоритмы квадратичной ошибки. Самым распространенным в этой категории алгоритмов является метод k-средних.

Метод k-средних

Метод k-средних строит заданное количество кластеров, расположенных как можно дальше друг от друга.

Математическая формула метода выглядит следующим образом:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2,$$

где k – количество кластеров, S_i – полученные кластеры, $i = 1, 2, \ldots, k$ и μ – центры масс векторов x_j принадлежат S_i [8].

Работа алгоритма состоит из следующих последовательных этапов:

1. Случайно выбирается количество k точек, являющихся исходными центрами масс кластеров.
2. Каждый объект относится к кластеру с ближайшим центром масс.
3. Пересчитываются центры масс кластеров согласно их текущего состава.
4. В случае неудовлетворения критерию остановки алгоритма повторяется этап 2.

Критерием остановки алгоритма, как правило, выбирается наименьшее изменение среднеквадратической ошибки. Если на этапе 2 не обнаружено объектов, переместившихся из одного кластера в другой, работу алгоритма можно остановить.

Главным недостатком метода k-средних является необходимости задавать количество кластеров для их последующего разбиения [9].

Для проведения кластеризации могут использоваться различные инструменты. Это могут быть языки программирования (ЯП) и интегрированные средства разработки. В таком случае программисту необходимо изучить необходимый для поставленной задачи алгоритм и написать программу самостоятельно. Также могут использоваться уже готовые программные средства, предназначенные для различного рода анализов данных. В рамках данной статьи метод k-средних применяется в программе Orange Data Mining.

Orange Data Mining

Данное программное средство с открытым исходным кодом, написанное на ЯП Python, позволяет проводить интеллектуальный анализ данных и машинное обучение при помощи интерфейса визуального программирования на основе компонентов. Компонентами служат виджеты и ранжируются от простой визуализации данных, выбора подмножества и предварительной обработки до эмпирической оценки алгоритмов обучения и прогнозного моделирования.

Программное обеспечение предоставляет пользователю следующие основные функции:

- загрузка данных из различных источников (файлы, базы данных, веб-ресурсы) и их представление в табличном виде;
- получение информации об атрибутах данных;
- изменение данных и параметров на любом шаге, позволяя отслеживать изменения в режиме реального времени;
- визуализация результатов с помощью различных графиков;
- сохранение созданной модели и ее дальнейшее использование [10–12].

К преимуществам Orange Data Mining простоту в использовании, наличие множества методов для анализа и машинного обучения, высокую скорость вычислений и доступность. Программа предоставляется бесплатно с официального сайта разработчика.

Data mining для анализа данных морского флота

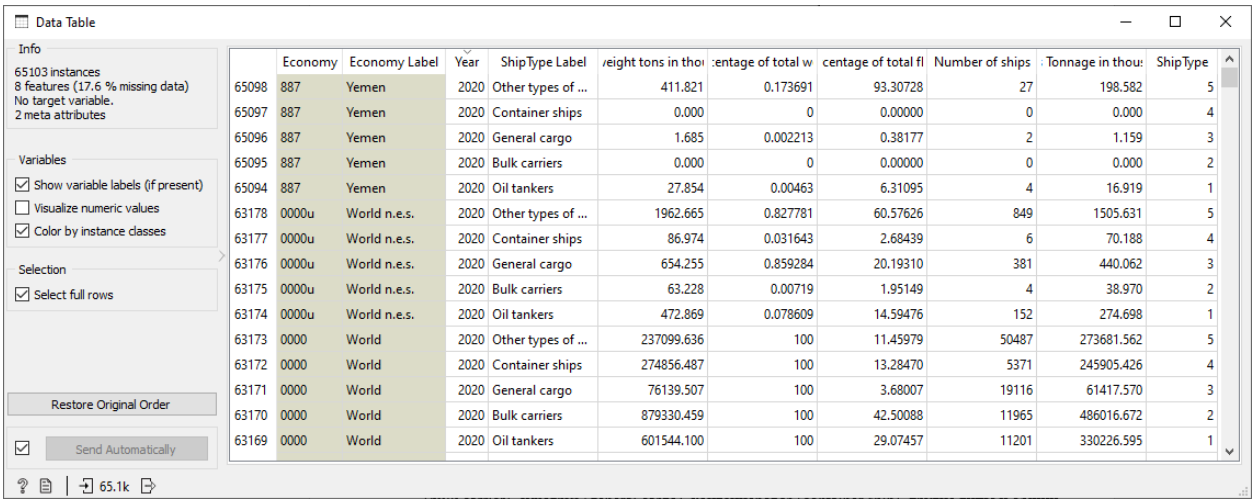
Существует огромное количество областей, где требуется анализ огромного массива данных, и морская деятельность не исключение. Данные, связанные с морским транспортом, можно преобразовывать в полезную информацию и использовать ее для принятия решений. В

качестве исходных данных исследования по теме данной статьи авторами использовался находящийся в открытом доступе датасет[1] (dataset – набор данных). Датасет представляет данные торгового флота по флагу регистрации и типу судна с 1980 по 2020 год, а именно: год постройки, флаг регистрации, тип судна, количество построенных судов, дедвейт (в тысячах), валовой тоннаж (в тысячах), процент от всего мира и процент от всего флота.

Количество записей (строк) в наборе данных – 79 873.

До 2011 года в исходных данных отсутствуют значения у таких атрибутов, как количество построенных судов и их валовой тоннаж.

В наборе данных следующие типы судна: нефтеналивные танкеры (oil tanker), балкеры (bulk carrier), сухогрузы (general cargo), контейнеровозы (container ship), другие типы и общий флот. В исследуемой выборке исключены данные по общему флоту. На рисунке 1 представлен фрагмент используемых данных. Стоит отметить, что количество записей уменьшилось до 65 103.



	Economy	Economy Label	Year	ShipType Label	eight tons in thou	entage of total w	centage of total fl	Number of ships	Tonnage in thou	ShipType
65098	887	Yemen	2020	Other types of ...	411.821	0.173691	93.30728	27	198.582	5
65097	887	Yemen	2020	Container ships	0.000	0	0.00000	0	0.000	4
65096	887	Yemen	2020	General cargo	1.685	0.002213	0.38177	2	1.159	3
65095	887	Yemen	2020	Bulk carriers	0.000	0	0.00000	0	0.000	2
65094	887	Yemen	2020	Oil tankers	27.854	0.00463	6.31095	4	16.919	1
63178	0000u	World n.e.s.	2020	Other types of ...	1962.665	0.827781	60.57626	849	1505.631	5
63177	0000u	World n.e.s.	2020	Container ships	86.974	0.031643	2.68439	6	70.188	4
63176	0000u	World n.e.s.	2020	General cargo	654.255	0.859284	20.19310	381	440.062	3
63175	0000u	World n.e.s.	2020	Bulk carriers	63.228	0.00719	1.95149	4	38.970	2
63174	0000u	World n.e.s.	2020	Oil tankers	472.869	0.078609	14.59476	152	274.698	1
63173	0000	World	2020	Other types of ...	237099.636	100	11.45979	50487	273681.562	5
63172	0000	World	2020	Container ships	274856.487	100	13.28470	5371	245905.426	4
63171	0000	World	2020	General cargo	76139.507	100	3.68007	19116	61417.570	3
63170	0000	World	2020	Bulk carriers	879330.459	100	42.50088	11965	486016.672	2
63169	0000	World	2020	Oil tankers	601544.100	100	29.07457	11201	330226.595	1

Рисунок 1. *Используемая выборка данных*

В реальных условиях на судостроительных заводах для сбора статистики могут использоваться дополнительные атрибуты, такие как заказы, стоимость затрат производства, прибыль, различные даты ключевых этапов постройки, данные по расходам метариалов и т. д. Для работы с таким массивом данных средства Microsoft Excel не подойдут, а прямые запросы к базе данных пользователь, как правило, не напишет и ему придется обращаться к ИТ-специалистам. В Orange Data Modeler аналитик при помощи виджетов может построить модель для анализа данных согласно поставленной ему задачи. На рисунке 2 представлена модель для проведения кластеризации методом k-средних.

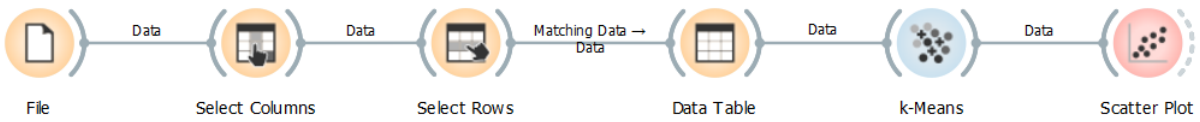


Рисунок 2. *Модель для кластеризации методом k-средних (составлено авторами)*

Демонстрируемая модель состоит из шести виджетов, последовательно соединенных друг с другом:

[1] URL: https://www.kaggle.com/maximosnikiforakis/shipping-analytics-world-merchant-fleet.

1. File. Данный виджет позволяет использовать уже готовый набор данных в виде отдельного файла. В рамках данной работы используется датасет формата .csv, но Orange Data Modeler поддерживает и другие.

2. Select Columns. Представляет настройку выборки по столбцам, если для аналитики не нужны какие-либо из исходных атрибутов.

3. Select Rows. Аналогично виджету Select Columns позволяет произвести выборку по строкам. Именно в этом виджете из типов судов исключены значения общего флота.

4. Data Table. Используется для табличного отображения данных. Исходные датасеты могут иметь структуру, отличную от табличной. Так, в используемом наборе данных первая строка состоит из перечисления всех атрибутов через запятую, а последующие строки – значения соответствующих атрибутов (тоже через запятую). Первоначальный вид таких данных сложен для восприятия, поэтому его можно преобразовать в понятную таблицу при помощи Data Table.

5. K-Means. Виджет включает в себя настройку кластеров для проведения кластеризации методом k-средних. Можно указать как строгое количество кластеров, так и воспользоваться динамической функцией «Silhouette», которая предварительно просчитывает оценку каждого кластера из указанного диапазона, указанного пользователем (при количестве записей не более 5 000).

6. Scatter Plot. Виджет для визуализации полученных данных. Представляет из себя диаграмму рассеивания.

На рисунке 3 изображен результат работы построенной модели в виде диаграммы рассеивания.

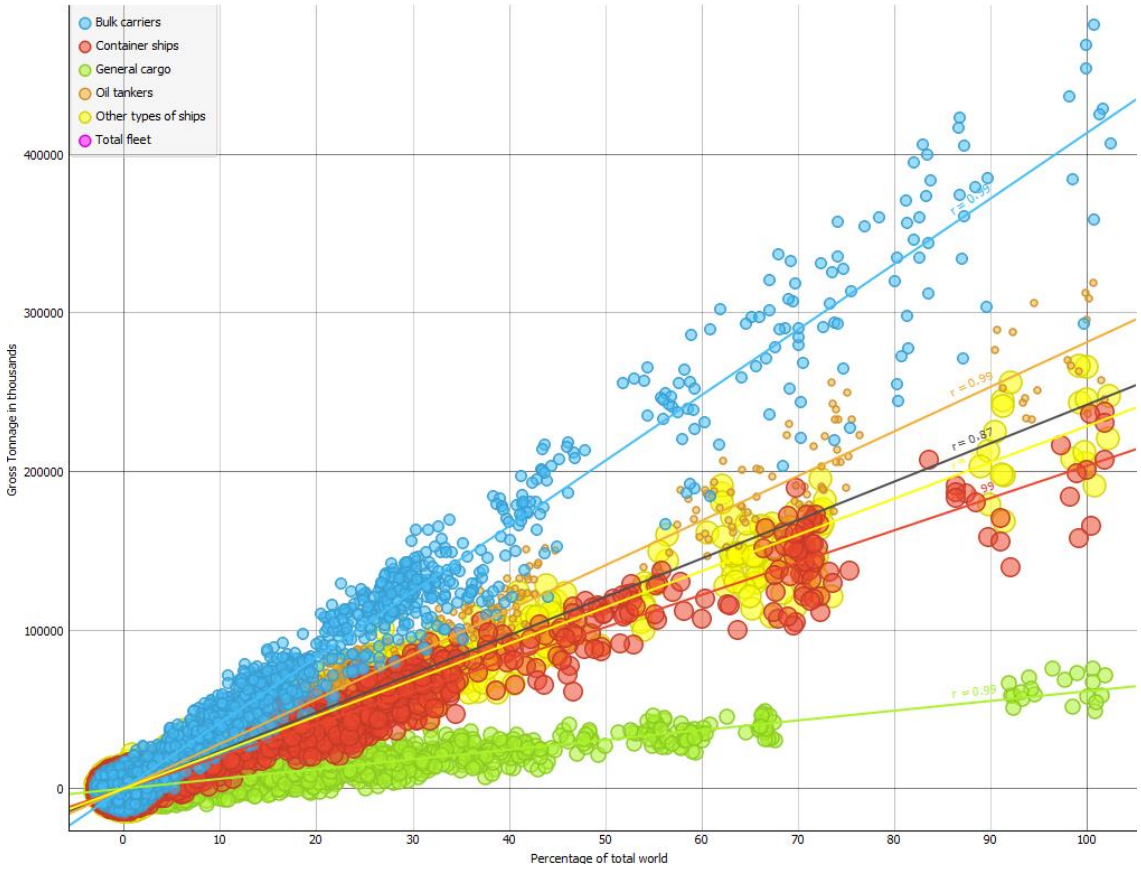


Рисунок 3. *Диаграмма рассеивания данных морского флота (составлено авторами)*

Аналитик может использовать как встроенную функцию для выбора осей, так и выбрать их самостоятельно. На рисунке 3 представлена зависимость между валовым тоннажом и процентом построенных судов по всему миру. Также в качестве третьей меры выбран тип судна. Он влияет на размер отображаемых точек, что также позволяет отразить результат аналитику более детально. В настройках виджета включена сетка и отображение линии регрессии. Точки, наиболее плотно прилегающие к линии, говорят о том, что модель верна. Усиленный разброс наблюдается у группы судов, относящихся к прочим судам (желтые точки), не указанным в наборе данных. Легенда в левом углу графика отображает исключенные из выборки данные по общему флоту (total fleet), что также полезно для минимизации потерь информации, полного представления данных.

Из полученной диаграммы рассеивания графика можно отметить, что больше всего по всему миру построено контейнеровозов и судов других типов с валовым тоннажом в пределах от 10 000 до 30 000, в то время как значения других типов судов сильно отличаются. В зависимости от поставленной задачи эту информацию можно интерпретировать по-разному и использовать в изучении причин этих цифр, выявлении потребностей в постройке тех или иных судов.

Конкретно для судостроительного комплекса или завода были бы полезны данные о финансовой составляющей производства. Добавив стоимостные значения, можно проследить зависимость между параметрами и принять меры для повышения эффективности производства или уменьшения затрат с сохранением продуктивности. Также при использовании других виджетов можно построить предиктивную (прогнозируемую) модель и грамотно ее использовать в бизнес-процессах предприятия.

Результат кластеризации методом k-средних представлен на рисунке 4. В настройках указано два кластера, на которые Orange Data Mining разбил все данные.

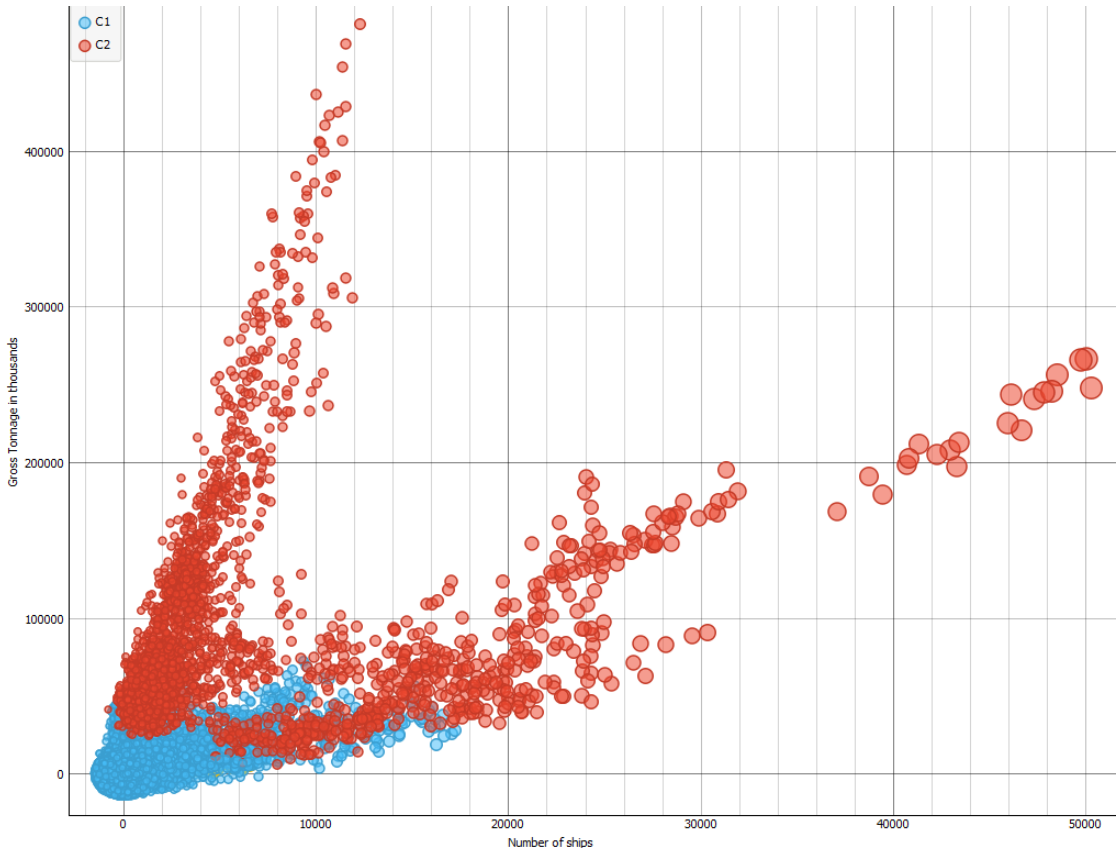


Рисунок 4. *Кластеризация методом k-средних (составлено авторами)*

На диаграмме представлена зависимость между валовым тоннажом и количеству построенных судов. Третья мера – тип судна. Согласно обозначениям структуры набора данных, большая точка обозначает другие типы. При наведении курсора мыши на маленькую точку программа выводит подсказку о том, что она принадлежит балкеру или другому судну, в зависимости от конкретной точки.

Кластеризация поделила датасет на два кластера, которые аналитик интерпретирует в исследовании по-своему, как и с примером на рисунке 3. Если кластер синего цвета хотя бы визуально находится в одной области, то кластер из красных точек имеет большой разброс и разветвление. Это говорит о том, что какие-то данные выбиваются из потенциальных взаимосвязей между параметрами. Включив отображение линии регрессии, как показано на рисунке 5, можно увидеть, что к таким «неправильным» данным относится часть красного кластера в диапазоне от 0 до 10 000 по количеству судов и примерно от 10 000 до 48 000 по валовому тоннажу.

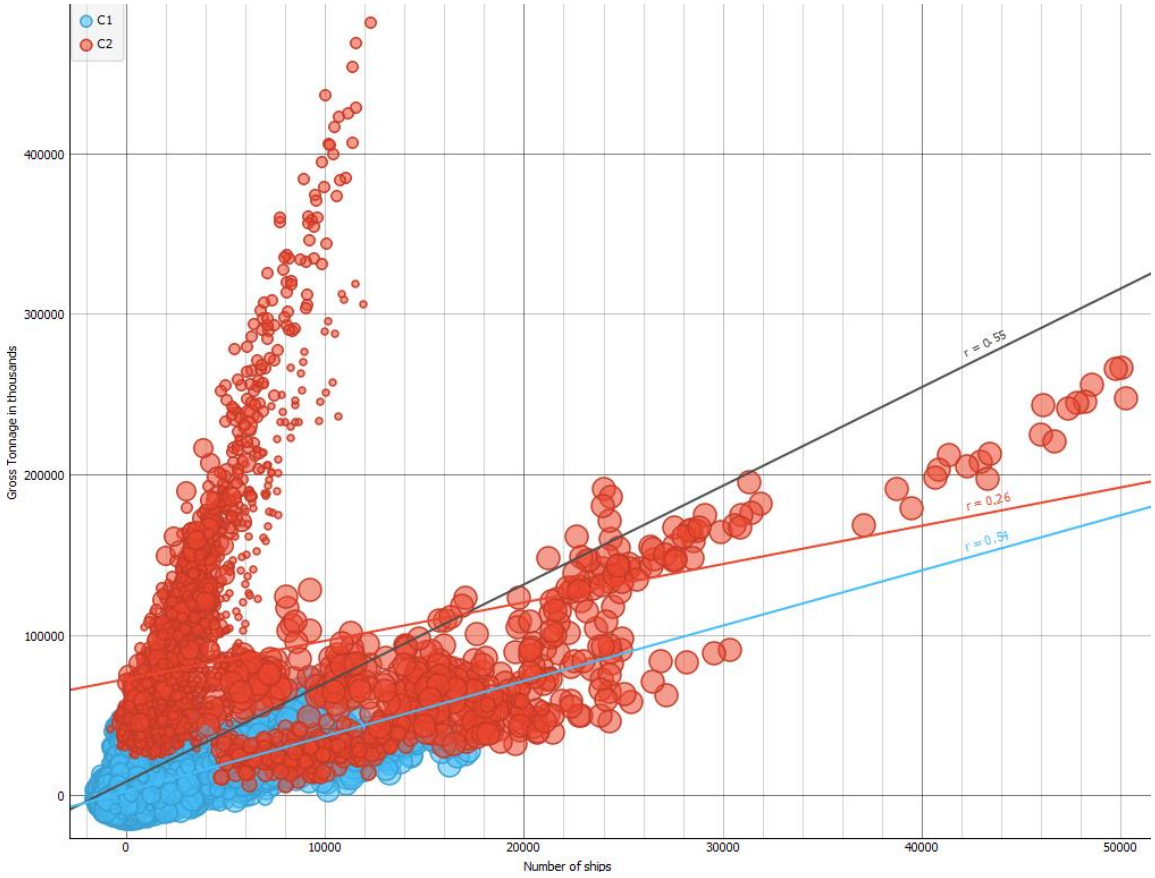


Рисунок 5. *Кластеризация методом k-средних с линиями регрессии (составлено авторами)*

Имея для анализа данные по конкретному предприятию, аналитик может эффективно сопоставить их для дальнейшего улучшения производства, повышения эффективности переработки металлолома, оптимизации логистики, контроля передвижения металлопроката или даже изменения плана по утилизации судна.

Стоит отметить, что и другие методы data mining могут применяться в морской деятельности. Так, в [13] используются самоорганизующиеся карты Кохонена для анализа эффективности методов автоматизированного неразрушающего контроля, что тоже является неотъемлемой частью в сфере морской деятельности.

Заключение

Исследование зависимостей между объектами является неотъемлемой частью для развивающегося производства. Автоматизация процессов классификации данных, их кластеризации и прогнозирования сокращают ручную работу и снижают вероятность ошибок, совершающихся человеческим фактором. На современные судостроительные комплексы внедряется новейшее передовое оборудование и, по мнению авторов, не должна отставать и ИТ-инфраструктура.

Не так редко на практике случается, что для решения каких-либо задач организация не располагает собственными силами и приходится пользоваться услугами аутсорсинга. Аналитику или другому пользователю Orange Data Mining не требуются навыки программирования для проведения своих исследований. Логично предположить, что с ростом сложности задачи возрастет и сложность модели, но квалифицированному специалисту будет под силу ее создать ровно так же, как он когда-то научился работать с другим программным обеспечением.

Изучая методы интеллектуального анализа данных и грамотно применяя их в компании, можно кардинально изменить ее текущее положение, сделав более конкурентоспособной на рынке. Особенно это касается работы с данными морского флота, поскольку производство одного только судна занимает годы, а какая-либо ошибка ведет к серьезным последствиям. Причем ошибки недопустимы не только на производстве, но и в открытом море.

ЛИТЕРАТУРА

1. Milosz M. Data mining as a modern method of data analysis // Izvestiya KGASU. – 2008. – №1(9). – Pp. 162–167. URL: https://izvestija.kgasu.ru/en/nomera-zhernala/archive?sod=sod1_2008&idizv=37 (дата обращения: 16.01.2021).
2. Venkatadri M., Reddy Lokanatha C. A Review on Data mining from Past to the Future // International Journal of Computer Applications. – 2011. – Vol.15. – №7. – Pp.19–22. URL: https://www.researchgate.net/publication/50946165_A_Review_on_Data_mining_from_Past_to_the_Future (дата обращения: 16.01.2021).
3. Вахитов А.Р. Использование KPI, технологий OLAP и data-mining при обработке данных // Известия ТПУ. – 2009. – №5. – С. 175–179. URL: https://elibrary.ru/item.asp?id=12883815 (дата обращения: 16.01.2021).
4. Bharati M. Ramageri. Data mining techniques and applications // Indian Journal of Computer Science and Engineering. – 2010. – Vol.1. – №4. – Pp. 301–305. URL: http://journaldatabase.info/articles/data_mining_techniques_applications.html (дата обращения: 16.01.2021).
5. Кравченко Ю.А., Нацкевич А.Н. Модель решения задачи кластеризации данных на основе использования бустинга алгоритмов адаптивного поведения муравьиной колонии и к-средних // Известия ЮФУ. Технические науки. – 2017. – №7 (192). – С. 90–102. – URL: https://elibrary.ru/item.asp?id=32251805 (дата обращения: 16.01.2021).
6. Федорова Е.А., Мусиенко С.О., Федоров Ф.Ю. Прогнозирование банкротства субъектов малого и среднего предпринимательства в России // Финансы и кредит. – 2018. – №11 (779). – С. 2537–2552. URL: https://cyberleninka.ru/article/n/prognozirovanie-bankrotstva-subektov-malogo-i-srednego-predprinimatelstva-v-rossii (дата обращения: 17.01.2021).

7. Ершов К.С., Романова Т.Н. Анализ и классификация алгоритмов кластеризации // Новые информационные технологии в автоматизированных системах. – 2016. – №19. – С. 274–279. URL: https://www.elibrary.ru/item.asp?id=25864070 (дата обращения: 17.01.2021).
8. Афанасьева Т.В., Сапунков А.А., Заварзин Д.В. Применение алгоритма кластеризации k-means для улучшения темпоральной статистики просмотра коммерческих предложений // Автоматизация процессов управления. – 2016. – №4 (46). – С.41–46. URL: https://www.elibrary.ru/item.asp?id=27522910 (дата обращения: 17.01.2021).
9. Прудковский Н.С. Кластеризация данных методом k-средних // Сборник трудов Восьмой всероссийской научно-технической конференции «Безопасные информационные технологии» (БИТ 2017). – 2017. – С. 347–350. URL: https://www.elibrary.ru/item.asp?id=35550003 (дата обращения: 17.01.2021).
10. Amrita Naik, Lilavati Samant. Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime, Procedia Computer Science // Procedia Computer Science. – 2016. – Vol. 85. – Pp. 662–668. URL: http://www.sciencedirect.com/science/article/pii/S1877050916306019 (дата обращения: 17.01.2021).
11. M. Peker, O. Özkaraca, A. Şaşar. Expert System Techniques in Biomedical Science Practice. – 2018. – Pp. 143–167.
12. Demšar, J., Blaž Z. Orange: Data Mining Fruitful and Fun – A Historical Perspective // Informatica. – 2013. – Vol. 37. – No 1. – P. 55–60. URL: http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/view/434 (дата обращения: 17.01.2021).
13. Овечкин М.В. Data Mining подход к вопросу анализа эффективности методов автоматизированного неразрушающего контроля // Международный научно-исследовательский журнал. – 2018. – №9–1 (75). URL: https://research-journal.org/wp-content/uploads/2018/09/9-1-75.pdf (дата обращения: 28.01.2021).

Maximov Valery Evgenievich
Far Eastern federal university, Vladivostok, Russia
E-mail: valep199778@gmail.com

Reznikova Kseniya Mikhailovna
Far Eastern federal university, Vladivostok, Russia
E-mail: a-da_97@mail.ru

Popov Dmitry Alexandrovich
Far Eastern federal university, Vladivostok, Russia
E-mail: dmppda@gmail.com

Data mining for marine data analysis

Abstract. There is practically no industry left where modern information technologies would not be used. Data mining approaches are very popular today. Using this technology allows to transform huge amounts of data into useful information. In the article, the authors present the definition of data mining technology and frequently used methods. Some of the popular data mining techniques include classification, clustering, machine learning, and prediction. The authors paid special attention to such a clustering method as the k-means. The algorithm's essence is to distribute the dataset into clusters. The finished results can be visualized and detect the scatter by naked eye, which implies heterogeneity in the data. By further investigating these variations, the analyst can find errors and weaknesses in the study area according to the task at hand.

Accurate and complete data is essential in maritime activities. In the field of shipbuilding data analysis and well-made operational decisions can affect the speed and quality of ship construction or even reduce production costs. In shipping and logistics, they can be used to optimize routes and improve the safety of seafarers. Effective use of data mining usually requires highly qualified database specialists and programmers. In this work, the authors have demonstrated a variant of using the Orange Data Mining software tool. This program does not require programming skills from the user, which makes it a useful tool for people far from writing program code.

The article explores the application of the Orange Data Mining program for automated mining of marine data. The results obtained show that the program can be effectively used in maritime activities.

Keywords: Data mining; classification; clustering; prediction; K-means; Orange Data Mining; marine activities