

МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Кубанский государственный аграрный
университет имени И. Т. Трубилина»

Учетно-финансовый факультет

Кафедра статистики и прикладной математики

МЕТОДЫ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

Методические рекомендации

по контактной и самостоятельной работе
для обучающихся по направлению подготовки
38.03.01 Экономика, направленность «Бизнес-аналитика»

Краснодар
КубГАУ
2021

Составители: И. А. Кацко, Н. Б. Паклин, Е. В. Кремянская

Методы принятия управленческих решений : метод. рекомендации / сост. И. А. Кацко, Н. Б. Паклин, Е. В. Кремянская. – Краснодар : КубГАУ, 2021. – 62 с.

Методические рекомендации включают краткое теоретическое изложение основных положений дисциплины «Методы принятия управленческих решений» и задания для практических занятий и самостоятельной работы.

Предназначены для бакалавров по направлению подготовки 38.03.01 Экономика, направленность «Бизнес-аналитика» очной формы обучения.

Рассмотрены и одобрены методической комиссией учетно-финансового факультета Кубанского ГАУ, протокол № 9 от 16.06.2021.

Председатель
методической комиссии

И. Н. Хромова

© Кацко И. А., Паклин Н. Б.,
Кремянская Е. В.,
составление, 2021

© ФГБОУ ВО «Кубанский
государственный аграрный
университет имени
И. Т. Трубилина», 2021

Введение

Успех функционирования любой социально-экономической системы во многом зависит от грамотного управления ею. В этой связи в последние годы все более пристальное внимание начинает уделяться интеллектуальной составляющей бизнеса, развитию методов и технологий обработки и анализа данных, доступных для специалистов любого профиля.

Данное обстоятельство, в свою очередь, требует от современного бакалавра знаний теоретических и практических основ применения программного обеспечения, позволяющего осуществлять поиск, анализ, моделирование и доставку информации, необходимой для принятия управленческих решений.

Формированию этих знаний, а также выработке у обучающихся навыков использования статистических пакетов с возможностями *Data Mining* и аналитических платформ (*Deductor*, *Loginom*) для решения конкретных прикладных задач будут способствовать материалы представленных методических рекомендаций.

1 Теоретические аспекты принятия управленческих решений

1.1 Вопросы для обсуждения

1. Понятие управленческого решения, основные этапы его разработки и принятия
2. Элементы и основные аспекты управленческого решения
3. Классификация управленческих решений
4. Классификация методов принятия управленческих решений

1.2 Теоретические сведения

Управленческое решение – это процесс сознательного выбора на основании имеющейся информации и в соответствии с выбранным (заданным) критерием одной из возможных альтернатив разрешения конкретной управленческой ситуации.

По своей сути управленческое решение представляет собой результат анализа данных, прогнозирования, оптимизации, экономического обоснования и выбора варианта достижения конкретной цели.

Основные элементы принятия решения:

- управленческая ситуация;
- проблема;
- информация;
- альтернативы;
- критерии;
- лицо, принимающее решение (ЛПР).

Чтобы возникающая ситуация была отнесена к классу управленческих, она должна отвечать ряду требований:

- в ней должно обязательно иметь место несоответствие желаемого состояния объекта управления (явления или про-

цесса) с действительным, другими словами, существовать проблема;

– обязательно должно быть лицо, принимающее решение (ЛПР);

– в ней должна присутствовать вариантность поведения участников;

– она должна быть разрешима, ибо, в противном случае, ни о каком управленческом воздействии речь идти не может.

Проблемы представляют собой системы, состоящие из элементов, связанных между собой. Проблемы могут относиться к одному из следующих четырех классов:

а) стандартные, подчиняющиеся строгим правилам поведения;

б) структурированные (количественно сформулированные), в которых существенные зависимости выявлены настолько хорошо, что могут быть выражены в числах или символах, получающих, в конце концов, численные оценки;

в) неструктурированные (качественно выраженные), содержащие лишь описание важнейших ресурсов, признаков и характеристик, количественные зависимости между которыми неизвестны;

г) слабоструктурированные (смешанные), которые содержат как качественные, так и количественные элементы, причем качественные, малоизвестные и неопределенные их стороны имеют тенденцию доминировать.

Поскольку именно наличие проблемы определяет ситуацию, то ситуации можно тоже отнести к классам с аналогичными проблемами названиями.

Успех процесса управления в значительной степени определяется полнотой и качеством информации и характером ее использования при принятии решений. Информация – это данные в любой форме и в непрерывной связи с окружающей средой, системой или контекстом.

Недостаток информации и невозможность количественной оценки ряда факторов при подготовке решений влечет за

собой неопределенность выбора, связанную с наличием нескольких альтернатив достижения поставленной цели.

Осуществление выбора альтернативы – самый сложный и ответственный шаг, определяющий, по сути, само решение.

Для оценки альтернатив с учетом всех их существенных свойств и особенностей предназначен критерий. В современной литературе не сложилось единого представления о том, следует ли применять один или несколько критериев. Существует мнение, что только при одном критерии может быть найдено лучшее решение. В противном случае элементы субъективизма исказят картину. Однако большинство исследователей считает, что необходима система критериев, то есть их должно быть несколько.

Лицо, принимающее решение (ЛПР) – человек (или группа людей), фактически осуществляющий выбор наилучшего варианта действий.

Управленческие решения могут быть классифицированы по следующим признакам:

- по степени формализации;
- надежности исходной информации (на основе надежной информации, в условиях риска, на базе неполной и неточной информации);
- длительности последствий;
- уровням планирования;
- частоте или повторяемости;
- степени охвата;
- виду процесса принятия решения (комплекс решений и выполняемые последовательно);
- числу ЛПР;
- организационному распределению;
- господствующему образу мышления;
- учету изменения данных;
- степени их независимости друг от друга (автономные и дополняющие друг друга);
- сложности (простые и сложные) и т.д.

Теория и практика накопили огромное число методов, которые применяются, в том числе, и при принятии управленческих решений. Наиболее популярные классификации такие: количественные и качественные методы; методы, используемые в условиях определенности; методы, используемые в условиях риска; методы, используемые в условиях неопределенности; методы формирования множества альтернатив; методы оценки альтернатив; методы выбора альтернативы; эмпирические (каузальные), логико-эвристические, абстрактно-логические (математические) и рефлексивные методы.

Управление сложными социальными и экономическими системами (объектами) трактуется как процесс принятия решений. В настоящее время существует ряд методик системного анализа, позволяющих принимать решения и основанных на построении моделей. Эти методики группируют в соответствии с двумя основными подходами:

1) аксиологический (целевой, или целенаправленный подход) – в его основе лежат методы, направленные на активизацию использования интуиции и опыта специалистов (МАИС (методы активизации интуиции специалистов): методы экспертных оценок, мозговая атака (штурм) и т.д.);

2) каузальный (морфологический (изучение формы), терминальный) – позволяет описать систему в терминах «перечисления» элементов и взаимосвязей между ними. Другими словами, этот подход позволяет осуществлять поиск взаимосвязей между элементами системы и представлять ее в виде «пространства состояний», что характерно для большинства методов математического моделирования. Этот подход включает методы формализованного представления систем (МФПС).

На практике целесообразно оба подхода сочетать. Кроме того, используются специальные методы моделирования систем: кибернетический, когнитивный, ситуационный, информационный, имитационного моделирования, интерактивного моделирования и др.

1.3 Тестовые задания

1. Важнейшие требования, предъявляемые к управленческим решениям:

- а) реализуемость;
- б) эффективность;
- в) оптимальность;
- г) изменчивость.

2. По длительности последствий управленческие решения могут быть:

- а) долгосрочными;
- б) среднесрочными;
- в) краткосрочными;
- г) бессрочными.

3. По уровням планирования управленческие решения могут быть:

- а) стратегическими;
- б) тактическими;
- в) оперативными;
- г) рутинными.

4. По частоте управленческие решения могут быть:

- а) одноразовыми;
- б) повторяющимися;
- в) рутинными;
- г) многократными.

5. По степени охвата управленческие решения могут быть:

- а) общими;
- б) узкоспециализированными;
- в) частными;
- г) всеохватывающими.

6. По числу ЛПР управленческие решения могут быть:

- а) индивидуальными;
- б) групповыми;
- в) совместными;
- г) общими.

7. По организационному распределению управленческие решения могут быть:

- а) централизованными;
- б) децентрализованными;
- в) индивидуальными;
- г) групповыми.

8. По господствующему образу мышления управленческие решения могут быть:

- а) дискурсивными;
- б) интуитивными;
- в) жесткими;
- г) гибкими.

9. По учету изменения данных управленческие решения подразделяются на:

- а) жесткие;
- б) гибкие;
- в) дискурсивными;
- г) интуитивными.

10. В основе аксиологического подхода при принятии управленческих решений лежат методы:

- а) экспертных оценок;
- б) мозговой атаки (штурма);
- в) аналитические;
- г) статистические.

11. В основе каузального подхода при принятии управленческих решений лежат методы:

- а) аналитические;
- б) статистические;
- в) экспертных оценок;
- г) мозговой атаки (штурма).

12. Проблема является слабоструктурированной, если имеет:

- а) не только количественные, но и качественные характеристики;
- б) только количественные характеристики;
- в) только качественные характеристики;
- г) статические характеристики.

13. Проблема является хорошо структурированной, если имеет:

- а) только количественные характеристики;
- б) не только количественные, но и качественные характеристики;
- в) только качественные характеристики;
- г) статические характеристики.

14. Для решения стандартных проблем, возникающих в управленческой деятельности, применяются(ются):

- а) инструкции и установленные руководителем правила;
- б) экономико-математические методы;
- в) экспертные оценки;
- г) системный подход.

15. Для решения неструктурированных проблем, возникающих в управленческой деятельности, применяются(ются):

- а) экспертные оценки;
- б) инструкции и установленные руководителем правила;
- в) экономико-математические методы;
- г) системный подход.

1.4 Вопросы для самоподготовки

1. Что понимается под управленческим решением?
2. Каковы основные признаки управленческого решения?
3. Каким требованиям должна отвечать управленческая ситуация?
4. На какие группы можно подразделить проблемы, возникающие в управленческой деятельности?
5. Какие этапы необходимо пройти в ходе решения слабо структурированных проблем?
6. Какую роль играет информация в принятии управленческих решений?
7. Какие факторы влияют на выбор альтернатив?
8. Какие факторы считаются объективными?
9. Какую роль в принятии управленческого решения играют условия внешней среды?
10. Всегда ли управленческие решения принимаются условиях неопределенности?
11. Какие методы принятия управленческих решений существуют?

2 Управление на основе данных

2.1 Вопросы для обсуждения

1. Сбор данных. Основные требования к качеству данных
2. Отчетность как процесс организации данных. Анализ данных
3. Критерии управления на основе данных
4. Зрелость аналитических данных для обоснования управленческих решений

2.2 Теоретические сведения

Все возможные виды деятельности человека можно разбить на два типа: познание мира и преобразование его. При этом любая деятельность субъекта, любое его взаимодействие с внешним миром происходят посредством моделей. Отличительная особенность моделей от других систем состоит в их предназначении – отображать моделируемый оригинал, заменять его в определенном отношении, т. е. содержать и представлять информацию об оригинале. Таким образом, модель есть системное отображение оригинала.

При построении любой модели всегда возникает задача поиска необходимых данных. Данные – это основа принятия решений – наиболее сложного и ответственного этапа деятельности человека в различных организационных структурах. Поэтому необходимо, чтобы набор данных соответствовал вопросу, который требуется решить. Помимо этого, данные должны быть своевременными, точными, чистыми, объективными, и, что важнее всего, они должны заслуживать доверия.

Итак, к качеству данных предъявляются следующие требования:

- доступность;
- точность;
- взаимосвязанность;
- непротиворечивость;
- однозначность;
- релевантность;
- надежность;
- своевременность.

Необходимые факторы управления на основе данных – составление отчетов и получение оповещений.

Отчетность — процесс организации данных в информационные сводки для отслеживания того, как функционируют разные сферы бизнеса. Информация отчетов показывает, что

произошло в прошлом. Кроме того, данные отчетности могут быть тем фундаментом, с которого можно наблюдать за изменениями и тенденциями. Они могут представлять интерес для инвесторов и акционеров, но в целом это ретроспективный взгляд на ситуацию.

Отчет показывает, что произошло. Однако для управления на основе данных необходимо прогнозировать развитие ситуации, стараться понять, почему меняются показатели, и, где возможно, экспериментальным путем обосновывать причины этих изменений. Эти задачи решает анализ данных – преобразование данных в выводы, на основе которых будут приниматься управленческие решения и осуществляться действия с помощью людей, процессов и технологий. Анализ показывает, почему это произошло. Таким образом, отчеты ретроспективны, а анализ дает рекомендации.

Для компаний с управлением на основе данных характерны виды деятельности (критерии), перечисленные ниже:

- постоянное проведение различных тестирований, направленных на совершенствование деятельности компании и ее сотрудников;
- прогнозное моделирование, прогнозирование объема продаж, курса акций или выручки;
- использование собственных прогнозных ошибок для улучшения своих моделей;
- осуществление выбора среди будущих вариантов или действий на основе комплекса взвешенных показателей.

В таких компаниях именно данные — основной фактор, обуславливающий стратегию и влияющий на нее. В ней формируется конструктивная корпоративная культура, при которой данным доверяют, а результаты анализа бывают высокозначимыми, информативными и используются для определения последующих управленческих шагов.

2.3 Тестовые задания

1. Не существует следующего типа анализа данных:

- а) описательного;
- б) разведочного;
- в) прогностического;
- г) индуктивного;
- д) каузального;
- е) экспедиционного.

2. Процесс организации данных в информационные сводки – это

- а) отчетность;
- б) статистика;
- в) группировка;
- г) бюллетень.

3. Обеспечивает количественное описание данных ... анализ

- а) описательный;
- б) разведочный;
- в) индуктивный;
- г) каузальный.

4. Изучение основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий обеспечивает ... анализ:

- а) описательный;
- б) разведочный;
- в) индуктивный;
- г) каузальный.

5. Логическое извлечение информации о генеральной совокупности (ее параметрах, типе распределения), из которой был взят набор данных, обеспечивает ... анализ

- а) описательный;

- б) разведочный;
- в) индуктивный;
- г) каузальный.

6. Нахождение причинно-следственной связи между явлениями обеспечивает ... анализ

- а) описательный;
- б) разведочный;
- в) индуктивный;
- г) каузальный.

7. К классическим описательным методам анализа данных не относится(ятся):

- а) методы поиска и анализа зависимостей;
- б) графическое представление;
- в) проверка статистических гипотез;
- г) оценка параметров распределения.

8. К классическим аналитическим методам анализа данных не относится(ятся):

- а) имитационное моделирование;
- б) методы поиска и анализа зависимостей;
- в) методы классификации;
- г) методы снижения размерности.

9. В зависимости от шкалы измерения различают данные:

- а) количественные;
- б) качественные;
- в) ранжированные;
- г) статистические.

10. К качественным шкалам относят:

- а) номинальную;
- б) порядковую;
- в) интервальную;
- г) шкалу отношений.

11. К количественным шкалам относят:

- а) интервальную;
- б) шкалу отношений;
- в) номинальную;
- г) порядковую.

12. В разведочном анализе данных используются следующие методы:

- а) визуализация данных и манипуляция с данными на основе графического изображения;
- б) использование активных и иллюстративных переменных;
- в) линейризация связей;
- г) упорядочение связей.

2.4 Вопросы для самоподготовки

1. Что понимается под данными?
2. Что является источниками данных?
3. Какие методы анализа данных существуют?
4. Какие инструменты относятся к классическим описательным методам?
5. Какие инструменты относятся к классическим аналитическим методам?
6. Какие требования предъявляются к качеству бизнес-данных, используемых для построения моделей принятия управленческих решений?
7. Что такое «отчетность»?
8. Какие виды деятельности характерны для компаний с управлением на основе данных?

3 Методология анализа бизнес-данных

3.1 Вопросы для обсуждения

1. Аналитический и информационный подходы к моделированию
2. Методы решения аналитических задач реализации управляющих воздействий
3. Формы представления, типы и виды данных, используемых для построения моделей принятия управленческих решений
4. Подготовка бизнес-данных к анализу
5. Технологии *KDD* и *Data Mining*. Аналитические платформы

3.2 Теоретические сведения

Анализ данных нельзя рассматривать только как обработку информации после ее сбора. Анализ данных – это, прежде всего, средство проверки гипотез и решения задач исследователя. Он тесно связан с моделированием, поскольку ограниченность познавательных способностей человека заставляет исследователей упрощать изучение интересующих их объектов, явлений и систем.

Моделирование – универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности.

Построение моделей позволяет обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других управленческих задач. Важные свойства любой модели:

- упрощенность;
- конечность;
- приближенность;
- адекватность;
- целостность;

- замкнутость;
- управляемость.

По виду моделирования модели делят на:

- эмпирические;
- теоретические;
- смешанные.

Аналитический подход к моделированию базируется на том, что исследователь при изучении системы отталкивается от модели. В этом случае выбирается подходящая модель (закон, зависимость, уравнение и пр.), а варьирование входных параметров на выходе дает результат, который моделирует поведение системы в различных условиях. При аналитическом подходе не модель «подстраивается» под действительность, подбирается существующая аналитическая модель таким образом, чтобы она адекватно отражала реальность.

Информационный подход к моделированию ориентирован на использование данных. При этом реальный объект рассматривается как «черный ящик», имеющий ряд входов и выходов, между которыми моделируются некоторые связи. Таким образом, при информационном подходе отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность.

Выделяют две основные группы методов решения аналитических задач:

- 1) извлечение и визуализация данных;
- 2) построение и использование моделей.

По степени структурированности выделяют следующие формы представления данных:

- неструктурированные;
- структурированные;
- слабоструктурированные.

Все структурированные данные делятся на пять типов:

- целый;
- вещественный;
- строковый;

- логический;
- дата/время.

По виду данные подразделяются на непрерывные и дискретные.

При сборе и подготовке данных к анализу необходимо придерживаться следующих принципов:

- 1) абстрагироваться от существующих информационных систем и имеющихся в наличии данных;
- 2) описать все факторы, потенциально влияющие на анализируемый процесс/объект;
- 3) экспертно оценить значимость каждого фактора;
- 4) определить способ представления информации;
- 5) собрать все легкодоступные факторы;
- 6) собрать наиболее значимые, с точки зрения экспертов, факторы;
- 7) оценить сложность и стоимость сбора средних и наименее важных по значимости факторов.

Информационный подход к анализу получил распространение в таких методиках извлечения знаний как *KDD* и *Data Mining*.

KDD – извлечение знаний из баз данных. Методика описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для обнаружения полезного знания.

KDD включает в себя этапы подготовки данных, выбора информативных признаков, очистки, построения моделей, постобработки и интерпретации полученных результатов. Ядром этого процесса являются методы *Data Mining* (добыча данных), позволяющие обнаруживать закономерности и знания.

Аналитической платформой класса *KDD* (*Knowledge Discovery in Databases*) и *Data Mining* и основой для создания законченных прикладных решений в области анализа данных является *Deductor*.

Реализованные в *Deductor* технологии позволяют на базе единой архитектуры пройти все этапы построения аналитиче-

ской системы от создания хранилища данных до автоматического подбора моделей и визуализации полученных результатов, в частности, в виде *OLAP* кубов, таблиц, диаграмм, гистограмм, карт, графов и т.д.

Аналитической платформой, пришедшей на смену АП *Deductor*, является *Loginom*. Она предоставляет возможности глубокой аналитики и позволяет принимать управленческие решения, основанные на точной и достоверной информации.

3.3 Тестовые задания

1. Модель – это

- а) результат отображения системы;
- б) экспертное заключение;
- в) совокупность входов, выходов и состояний системы;
- г) комплекс знаний о системе (объекте).

2. Не существует моделей (по виду моделирования)

- а) эмпирических;
- б) теоретических;
- в) смешанных;
- г) мультипликативных.

3. Не является свойством модели:

- а) упрощенность;
- б) конечность;
- в) адекватность;
- г) целостность;
- д) точность.

4. Различают следующие подходы к моделированию:

- а) аналитический;
- б) информационный;
- в) синергетический;
- г) интегральный.

5. Не является способом визуализации данных:

- а) многомерный куб;
- б) таблица;
- в) диаграмма;
- г) система индикаторов.

6. Формы представления данных по степени структурированности:

- а) неструктурированные;
- б) структурированные;
- в) слабоструктурированные;
- г) гиперструктурированные.

7. Не существует следующего типа структурированных данных:

- а) целый;
- б) вещественный;
- в) строковый;
- г) логический;
- д) универсальный.

8. Установление зависимости непрерывной выходной переменной от входных переменных – это

- а) регрессия;
- б) классификация;
- в) кластеризация;
- г) ассоциация.

9. Установление зависимости дискретной выходной переменной от входных переменных – это

- а) регрессия;
- б) классификация;
- в) кластеризация;
- г) ассоциация.

10. Группировка объектов на основе данных, описывающих их свойства – это

- а) регрессия;
- б) классификация;
- в) кластеризация;
- г) ассоциация.

11. Выявление закономерностей между связанными событиями – это

- а) регрессия;
- б) классификация;
- в) кластеризация;
- г) ассоциация.

12. Специализированное программное решение, содержащее в себе все инструменты для извлечения закономерностей из сырых данных – это

- а) аналитическая платформа;
- б) синтетическая платформа;
- в) *Data Mining*;
- г) *KDD*.

13. Информативными являются признаки, которые:

- а) имеют слабую корреляцию между собой;
- б) содержат только одно значение;
- в) содержат в основном одно значение;
- г) имеют уникальные значения;
- д) имеют сильную корреляцию между собой.

14. Непрерывные данные – это данные, которые

- а) могут принимать какое угодно значение в некотором интервале;
- б) могут принимать строго определенные значения в некотором интервале;

- в) принимают любое случайное значение;
- г) принимают фиксированные значения.

15. По виду данные делятся на:

- а) непрерывные;
- б) дискретные;
- в) основные;
- г) вспомогательные.

3.4 Вопросы для самоподготовки

1. В чем заключается аналитический подход к моделированию?

2. В чем заключается информационный подход к моделированию?

3. Что понимают под моделью предметной области?

4. Какие методы построения моделей существуют?

5. Какие свойства присущи моделям?

6. Какие различают формы представления данных по степени структурированности?

7. На какие типы делятся структурированные данные?

8. В чем заключается суть *KDD* и *Data Mining*?

9. Какие возможности, позволяющие принимать управленческие решения, предоставляет аналитическая платформа *Loginom*?

10. Что такое «пакет» в АП *Loginom*?

11. Что включает модуль в АП *Loginom*?

4 Консолидация бизнес-данных

4.1 Вопросы для обсуждения

1. Источники и основные задачи консолидации данных

2. Хранилища данных (ХД), их роль в обеспечении принятия управленческих решений

3. Основные требования к ХД. Использование концепции ХД в системах поддержки принятия решений (СППР)

4. Способы использования ХД

5. Архитектура корпоративного ХД

6. Многомерные ХД

4.2 Теоретические сведения

Консолидация – это комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

Ключевым понятием консолидации является источник данных – объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи.

В процессе консолидации данных решаются следующие задачи:

- 1) выбор источников данных;
- 2) разработка стратегии консолидации;
- 3) оценка качества данных;
- 4) обогащение;
- 5) очистка;
- 6) перенос в хранилище данных.

В свою очередь, хранилище данных (ХД) – это специально организованная база данных, ориентированная на решение задач анализа данных и поддержки принятия решений, обеспечивающая максимально быстрый и удобный доступ к информации.

Чтобы ХД выполняло функции, соответствующие его основной задаче – поддержке процесса принятия управленческих решений, оно должно удовлетворять следующим требованиям:

- высокая скорость получения данных из хранилища;
- автоматическая поддержка внутренней непротиворечивости данных;
- возможность получения и сравнения срезов данных;
- наличие удобных средств для просмотра данных в хранилище;
- обеспечение целостности и достоверности хранящихся данных.

Основными задачами, которые требуется решить в процессе разработки ХД, являются:

- 1) выбор структуры хранения данных, обеспечивающей высокую скорость выполнения запросов и минимизацию объема оперативной памяти;
- 2) первоначальное заполнение и последующее пополнение хранилища;
- 3) обеспечение единой методики работы с разнородными данными и создание удобного интерфейса пользователя.

Выделяют три основных подхода к использованию ХД:

- регулярные отчеты;
- нерегламентированные запросы;
- интеллектуальный анализ данных.

Чтобы приблизить ХД к условиям и специфике конкретной организации, в настоящее время разработано несколько архитектур хранилищ: реляционные, многомерные, гибридные и виртуальные.

Основное назначение многомерных хранилищ данных – поддержка систем, ориентированных на аналитическую обработку данных, поскольку такие хранилища лучше справляются с выполнением сложных нерегламентированных запросов.

В основе построения многомерных хранилищ данных лежит многомерная модель данных, опирающаяся на концепцию многомерных кубов (OLAP-кубов).

4.3 Тестовые задания

1. В процессе консолидации данных не решается задача

- а) выбора источников данных;
- б) оценки качества данных;
- в) очистки данных;
- г) переноса в хранилище данных;
- д) анализа зависимостей между данными.

2. Не является критерием оптимальности с точки зрения консолидации данных:

- а) высокая скорость доступа к данным;
- б) компактность хранения;
- в) автоматическая поддержка целостности структуры данных;
- г) контроль непротиворечивости данных;
- д) оперативность обработки данных.

3. Не существует ... архитектуры ХД:

- а) реляционной;
- б) многомерной;
- в) гибридной;
- г) виртуальной;
- д) реальной.

4. Начальный этап реализации любой аналитической задачи – это ... данных:

- а) консолидация;
- б) трансформация;
- в) визуализация;
- г) фильтрация.

5. Механизм, позволяющий аналитику оперировать данными из ХД посредством бизнес-терминов предметной области – это

- а) семантический слой;
- б) операционная система;
- в) пользовательский интерфейс;
- г) надстройка.

6. Основная задача ХД –

- а) поддержка процесса анализа данных;
- б) получение и сравнение срезов данных;
- в) обеспечение целостности данных;
- г) визуализация данных.

7. Хранилища данных, построенные на основе HОLAP, называются:

- а) гибридными;
- б) реляционными;
- в) многомерными;
- г) виртуальными.

8. Формирует устойчивые и непротиворечивые опорные точки для многомерного хранилища ... структура:

- а) реляционная;
- б) гибридная;
- в) многомерная;
- г) виртуальная.

9. Все ХД можно разделить на:

- а) одноплатформенные;
- б) кросс-платформенные;
- в) двухплатформенные;
- г) мультиплатформенные.

10. Отличие ХД от обычных систем хранения данных –

- а) данные из ХД не удаляются, а пополнение происходит в соответствии с определенным регламентом;
- б) данные из ХД удаляются и не пополняются;

- в) данные из ХД не удаляются и не пополняются;
- г) данные из ХД удаляются, а пополнение происходит без определенного регламента.

4.4 Вопросы для самоподготовки

1. Что такое «консолидация бизнес-данных»?
2. Что такое «источник данных»?
3. Что такое «хранилище данных»?
4. Какую основную задачу решают ХД?
5. Каким требованиям должно удовлетворять ХД?
6. Какие задачи решаются в процессе разработки ХД?
7. Какие существуют подходы к использованию ХД?
8. Что лежит в основе построения многомерных ХД?
9. Что такое «метаданные»?
10. Какие уровни метаданных можно выделить?

4.5 Задача для практических занятий

Задача 4.5 Создать в *Deductor* (или *Loginom*) пустое хранилище данных *Фармация*, спроектировать структуру ХД и загрузить в него информацию из следующих текстовых файлов: *groups.txt*, *produces.txt*, *stores.txt*, *sales.txt* (база данных кафедры статистики и прикладной математики). Результатом работы должен стать сценарий загрузки *load.ded*. Убедиться, что в хранилище загружена вся информация о продажах. Импортировать информацию о продажах из ХД, включая атрибуты товара. Установить следующие срезы:

- «кроме последнего периода 1 месяц от имеющихся данных»;
- срез только по одной любой товарной группе.

5 Трансформация бизнес-данных

5.1 Вопросы для обсуждения

1. Сущность трансформации данных

2. Основные методы трансформации данных
3. Трансформация упорядоченных данных
4. Группировка как инструмент подготовки бизнес-данных к анализу. Разгруппировка данных
5. Способы слияния данных
6. Квантование данных
7. Нормализация и кодирование данных

5.2 Теоретические сведения

Трансформация бизнес-данных – комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа.

Типичные средства трансформации данных:

- преобразование упорядоченных данных;
- квантование;
- сортировка;
- слияние;
- группировка и разгруппировка;
- настройка набора данных;
- табличная подстановка значений;
- вычисляемые значения;
- нормализация.

5.3 Тестовые задания

1. Позволяет оптимизировать представление данных с целью обеспечения дальнейшего анализа:

- а) преобразование упорядоченных данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

2. Позволяет разбить диапазон возможных значений числового признака на заданное количество интервалов и присвоить номера интервалов или иные метки попавшим в них значениям:

- а) преобразование упорядоченных данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

3. Позволяет изменить порядок следования записей исходной выборки данных в соответствии с алгоритмом, определенным пользователем:

- а) преобразование упорядоченных данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

4. Позволяет объединить две таблицы по одноименным полям или дополнить одну таблицу записями из другой:

- а) преобразование упорядоченных данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

5. Позволяет обобщить нужную информацию, объединить ее в минимально необходимое количество полей и значений:

- а) преобразование упорядоченных данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

6. Позволяет изменять номера, типы, метки и назначения полей исходной выборки данных:

- а) настройка набора данных;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

7. Позволяет производить замену значений в исходной выборке данных:

- а) табличная подстановка значений;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) слияние.

8. Позволяет преобразовать диапазон изменения значений числового признака в другой диапазон, более удобный для применения к данным тех или иных аналитических алгоритмов:

- а) табличная подстановка значений;
- б) квантование;
- в) сортировка;
- г) группировка и разгруппировка;
- д) нормализация.

9. Данные зависящие от времени, называют

- а) упорядоченными;
- б) зависимыми;
- в) нормализованными;
- г) сгруппированными.

10. Наиболее типичные функции агрегации, используемые при группировке данных:

- а) сумма;

- б) среднее;
- в) количество;
- г) максимум;
- д) минимум;
- е) медиана;
- ж) индекс.

5.4 Вопросы для самоподготовки

1. Что такое «трансформация данных»?
2. Каковы основные цели трансформации данных на этапе процесса ETL?
3. Какими типичными средствами трансформации оснащается большинство аналитических платформ?
4. Что происходит в процессе группировки данных?
5. В чем заключается суть разгруппировки данных?
6. Для чего выполняется процедура слияния данных?
7. Какие существуют способы слияния данных?
8. Что такое «квантование»?
9. С какой целью используется квантование?
10. Какие методы квантования существуют?
11. Что подразумевается под нормализацией и кодированием данных?

5.5 Задачи для практических занятий

Задача 5.5.1 По данным таблицы 1 методом статистических группировок с помощью возможностей табличного процессора *Excel* выявить влияние фондообеспеченности на эффективность использования земли и труда.

Методические указания.

1. Создать Лист 1 с исходными показателями для группировки (таблица 1). Сохранить рабочую книгу под названием *Группировка*.

Таблица 1 – Исходные данные для решения задачи 5.5.1

№ п/п	Площадь сельскохозяйственных угодий, га	Среднегодовая численность персонала, чел.	Годовой фонд оплаты труда, тыс. руб.	Среднегодовая стоимость основных средств, тыс. руб.	Стоимость валовой продукции в текущих ценах, тыс. руб.
1	12391	372	103758	753553,5	558689
2	4501	105	33703	199128	254031
3	4932	150	53320	261377	231644
4	12338	578	245991	1068768	1025296
5	6837	437	146384	454935	682126
6	9754	320	92317	742377	576507
7	7332	142	49117	317310,5	363084
8	10182	216	71868	450772,5	536524
9	5915	135	44324	138721,5	328857
10	23451	473	223349	809569,5	1139644
11	5781	136	21012	150262,5	140460
12	12239	449	153438	629206,0	790034
13	9202	477	175469	769089,0	950778
14	8628	134	49722	202426,0	341963
15	6464	205	84006	414533,5	590983
16	14105	372	129788	504474,0	979681
17	16371	311	78285	651135,0	772453
18	4305	165	45356	234008,5	249567
19	23531	315	127529	861259,0	864652
20	7860	200	62477	540528,0	570017
21	13055	550	186793	1198567,0	1095077
22	8188	292	101907	388400,0	610358
23	5078	220	76249	307163,0	328275
24	20989	751	259737	1453140,0	1476711
25	7628	455	143900	551778,0	745489

2. Выбрать команду *Вставка – Сводная таблица – ОК*. Если курсор мыши находится в поле таблицы, то диапазон задается автоматически.

3. Для группировки исходных данных по признаку *фондообеспеченность на 100 га сельхозугодий* предварительно заполнить макет сводной таблицы.

4. Щелкнуть правой кнопкой мыши по полю «Фондообеспеченность» и выбрать из контекстного меню команду «Группировать». Шаг группирования определяется предварительно, число групп принять равным трем. В результате будет получена сводная таблица.

5. Скопировать сводную таблицу на новый лист («Группировочная таблица»), используя команду *Правка–Специальная вставка–Значения*, чтобы иметь возможность редактирования полей и итогов. Достроить группировочную таблицу, введя показатели, которые необходимо рассчитать на основании итогов сводной таблицы. Ввести соответствующие формулы и отформатировать.

Задача 5.5.2 Провести сегментацию клиентов по их поведению, создав подмодель «RFM-анализ» в *Loginom* (база данных кафедры статистики и прикладной математики). На входе подмодели использовать набор данных со следующей структурой:

Поле	Тип
ID клиента	Строковый
Дата	Дата/Время
Сумма	Вещественный

На выходе необходимо получить следующий набор данных:

Поле	Тип
ID клиента	Строковый
Recensu	Строковый
Frequency	Строковый
Monetary	Строковый
Код RFM	Строковый

6 Визуализация данных

6.1 Вопросы для обсуждения

1. Цели и задачи визуализации данных
2. Методы визуализации данных
3. Визуализаторы общего назначения
4. OLAP-анализ
5. Визуализаторы для оценки качества моделей
6. Визуализаторы, применяемые для интерпретации результатов бизнес-анализа

6.2 Теоретические сведения

Визуализация – это представление данных в виде, который обеспечивает наиболее эффективную работу пользователя.

Визуальный анализ источника данных позволяет:

- увидеть, соответствуют ли данные ожидаемым;
- оценить степень пригодности данных к анализу;
- выдвинуть гипотезы о закономерностях процессов, описываемых данными;
- определить, какие виды очистки и преобработки необходимо применить к данным.

Основные методы визуализации:

- табличные и графические;
- одномерные и многомерные;
- общего назначения и специализированные.

К визуализаторам общего назначения относятся:

- 1) графики;
- 2) диаграммы;
- 3) гистограммы;
- 4) статистика.

Для визуализации многомерных табличных данных используется метод OLAP-анализа. В OLAP-системах разрозненная информация представляется в виде многомерного ку-

ба, которым можно легко манипулировать, извлекая срезами нужную информацию (рисунок 1).

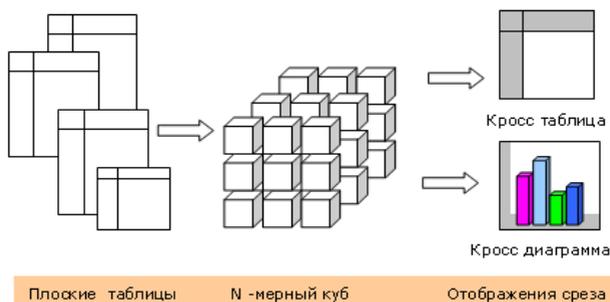


Рисунок 1 – Технология OLAP

Срезы исходного куба представляются на экране в виде кросс-таблицы и кросс-диаграммы. В отличие от обычной (плоской) таблицы кросс-таблица имеет уровни вложенности (например, разбиение строк на подстроки).

Для оценки качества моделей используется следующий типичный набор визуализаторов:

- матрица классификации;
- диаграмма рассеяния;
- ретропрогноз;
- графики контроля хода обучения.

Для интерпретации результатов анализа применяют следующие, наиболее распространенные, типы визуализаторов:

- 1) древовидные визуализаторы;
- 2) визуализаторы связей;
- 3) карты.

6.3 Тестовые задания

1. Не являются визуализаторами общего назначения:

- а) графики;
- б) диаграммы;
- в) гистограммы;

г) матрицы классификации.

2. Различают две составляющие качества моделей:

- а) адекватность;
- б) корректность;
- в) компактность;
- г) универсальность.

3. Не является визуализатором для оценки качества моделей:

- а) матрица классификации;
- б) диаграмма рассеяния;
- в) ретропрогноз;
- г) график.

4. С помощью гистограммы можно изобразить графически данные:

- а) только количественные;
- б) только категориальные;
- в) количественные и категориальные;
- г) временные.

5. Определить, является ли гистограмма распределения признака скошенной в какую-либо сторону, можно с помощью следующей числовой характеристики:

- а) дисперсия;
- б) коэффициент вариации;
- в) коэффициент асимметрии;
- г) коэффициент эксцесса.

6.4 Вопросы для самоподготовки

1. Что такое «визуализация»?
2. Какие цели преследует визуализация?
3. Какие задачи решает визуализация?

4. Какие действия позволяет совершать визуальный анализ источника данных?
5. Какие основные методы визуализации существуют?
6. Что представляют собой графики?
7. Для чего используются диаграммы?
8. Что показывает гистограмма?
9. Что такое «OLAP-анализ»?
10. Какие визуализаторы применяются для оценки качества моделей?

6.5 Задачи для практических занятий

Задача 6.5.1 В системе *Statistica* создать файл с именем *Reklama.sta*. Сохранить файл в своей папке. Внести исходную информацию о длине и ширине рекламных баннеров (таблица 2). Провести анализ зависимости цены рекламы от длины при фиксированной ширине. Представить данные в виде диаграммы рассеяния с соответствующими заголовками осей и уравнением.

Таблица 2 – Данные для нахождения площади рекламных баннеров и цены рекламы

Ширина	Длина	Площадь	Цена
47	35		
47	73		
47	111		
47	149		
47	187		
47	225		
47	263		
47	301		

Методические указания.

1. Для нахождения площади рекламы ввести в нижнем поле свойств переменной «площадь» – *Длинное имя (метка или формула с Функциями)* формулу: $=v1*v2$. Аналогично

определить цену, учитывая, что единица площади стоит 0,9 у.е.

2. Выполнить команду: *Графика – 2М Графики– вкладка Дополнительно – Тип графика Простой – Подгонка Линейная – Отметить в группе статистики все элементы – ОК (Graphs – 2D Graphs–Scatterplots–Advanced– Graphtype Regular – FitLinear–Statistics–OK).*

Задача 6.5.2 Из базы данных кафедры статистики и прикладной математики загрузить в *Statistica* файл *Excel Nedvig.xls* (<https://kubsau.ru/education/chairs/statistics/publications/>). Просмотреть информацию о переменных, предварительно выбрав их с помощью кнопки *Переменные (Variables)*. Задать условие выбора переменных в соответствии с таблицей 3. Описать переменные с помощью *Основных статистик (Basic Statistics)* и диаграмм. Результаты вычисления описательных статистик просмотреть в рабочей книге и отобразить в отчете.

Таблица 3 – Исходная информация для решения задачи 6.5.2

№ задачи	Общая площадь, м ²	Число комнат
1	Менее 54	1
2	Менее 40	1
3	Менее 45	1
4	38–48	1
5	Не менее 35	1
6	45–70	2
7	33–75	2
8	37–98	2
9	30–130	3
10	65–110	3
11	От 50 до 70	3

Задача 6.5.3 В результате тестирования группа из 24 человек набрала следующие баллы: 3; 3; 1; 0; 1; 1; 4; 4; 1; 2; 2; 2; 1; 1; 4; 2; 3; 4; 0; 4; 1; 4; 4; 2. Используя возможности табличного процессора *Excel*, проранжировать исходный ряд, под-

считать частоту и частость вариант. Выполнить команду *Данные – Анализ данных – Гистограмма*. Заполнить диалоговое окно, получить таблицу с частотами вариантов и соответствующий график.

Задача 6.5.4 На основе созданного при решении задачи 4.5 ХД *Фармация* разработать систему аналитической отчетности. Для этого написать в *Deductor Studio* сценарий обработки данных и назвать его *olap.ded*. Выбрать любые 5-7 отчетов из списка, приведенного ниже. Кроме кросс-таблицы, самостоятельно изучить кросс-диаграмму.

1. Построить кросс-таблицу и кросс-диаграмму по трем измерениям (отдел, месяц года, товарная группа), в ячейках которого отображается сумма и объем (количество проданных единиц продукции) продаж за все периоды, имеющиеся в ХД. Какая торговая точка приносит наибольшую сумму продаж? Какая товарная группа имеет максимальную сумму продаж? Построить кросс-диаграмму сумм продаж: общие продажи, продажи по торговым точкам, продажи по товарным группам.

2. То же, что в п.1, но за последние три месяца от имеющихся данных.

3. То же, что в п.1, но за последние три недели от имеющихся данных.

4. Найти сумму максимальной и средней стоимости покупки за последний месяц от имеющихся данных.

5. Сформировать многомерный отчет и график загруженности торговых точек по времени суток и торговым точкам. На какие часы приходятся пики продаж?

6. То же, что в п. 5, но за три месяца от имеющихся данных.

7. Сформировать многомерный отчет и график загруженности торговых точек по дням недели.

8. То же, что в п. 7, но за последний месяц от имеющихся данных.

9. Сформировать многомерный отчет и график загруженности торговых точек по дням месяца. Построить линию тренда.

10. То же, что в п. 9, но за последние три месяца от имеющихся данных.

11. 20 самых продаваемых товаров.

12. То же, что в п. 11, но за последние три недели от имеющихся данных.

13. 10 самых продаваемых товаров по воскресеньям.

14. 5 самых популярных товаров в каждой товарной группе.

15. То же, что и п. 14, но за последнюю неделю.

16. Товары, дающие 50 % объема продаж.

17. То же, что и п. 16, но за последние 3 месяца от имеющихся данных.

18. То же, что и п. 16, но за последнюю неделю.

19. 10 самых продаваемых товаров с 18 до 21 часа.

20. 10 товаров, пользующихся наименьшим спросом осенью.

21. Товары, дающие 50 % объема продаж в летние месяцы.

Задача 6.5.5 В АП *Loginom* создать отчеты по результатам анализа данных рынка недвижимости (база данных кафедры статистики и прикладной математики (<https://kubsau.ru/education/chairs/statistics/publications/>)). Построить диаграмму рассеяния стоимости квартир.

7 Очистка и предобработка бизнес-данных

7.1 Вопросы для обсуждения

1. Технологии и методы оценки качества данных
2. Понятие чистоты и предобработки данных. Инструменты предобработки в аналитическом приложении

3. Фильтрация данных
4. Обработка дубликатов и противоречий
5. Выявление аномальных значений
6. Восстановление пропущенных значений и сокращение размерности
7. Сокращение числа признаков
8. Сокращение числа значений признаков и записей

7.2 Теоретические сведения

Качество данных – совокупность их свойств и характеристик, определяющих степень пригодности для анализа.

Различают следующие уровни качества данных:

- технический;
- аналитический;
- концептуальный.

При оценке качества данных и разработке стратегии его повышения необходимо уделять внимание всем трем уровням.

С целью повышения качества данных используется комплекс методов и алгоритмов, получивших название «очистка данных».

Одним из наиболее распространенных методов оценки качества данных и выявления проблем является профайлинг. В его процессе определяется информация о некотором атрибуте (поле) источника данных и проверяется ее соответствие заданным ограничениям.

Если объем данных не слишком велик, то для оценки их качества можно применять визуальные методы. Для этого используются как встроенные средства визуализации аналитической платформы, так и сторонние программные средства, в которых можно просматривать данные в табличной и графической формах, рассчитывать статистические характеристики.

Главная задача оценки качества данных – отличить «грязные» данные от «чистых», т.е. выявить, какие данные содер-

жат ошибки, и оценить, какую долю «грязные» данные составляют в общем объеме данных.

Таким образом, чистота данных подразумевает отсутствие в них ошибок ввода, структурных нарушений, некорректных форматов и других причин, которые мешают анализу данных.

Качество данных связано с конкретными целями и задачами анализа, используемыми моделями, методами и алгоритмами. Нельзя достоверно утверждать, являются ли данные качественными, пока они не будут связаны с конкретной аналитической задачей.

Предобработка данных – комплекс методов и алгоритмов, применяемых в аналитическом приложении с целью подготовки данных к решению конкретной задачи и приведения их в соответствие с требованиями, определяемыми спецификой задачи и способами ее решения.

Предобработку данных можно рассматривать как комбинацию методов очистки и специальных методов оптимизации данных для решения конкретной задачи.

Типичный набор инструментов предобработки и подготовки данных к анализу, поставляемый с большинством аналитических платформ, содержит следующие средства:

- 1) очистка от шумов и сглаживание рядов данных;
- 2) восстановление пропущенных значений;
- 3) редактирование аномальных значений;
- 4) обработка дубликатов и противоречий;
- 5) снижение размерности входных данных;
- 6) устранение незначимых факторов.

7.3 Тестовые задания

1. Совокупность свойств и характеристик данных, определяющих степень их пригодности для анализа – это ... данных:

- а) качество;
- б) структура;

- в) форма представления;
- г) уровень.

2. Не существует уровня качества данных:

- а) технического;
- б) аналитического;
- в) концептуального;
- г) синтетического.

3. Оценка качества данных преследует две цели:

- а) контроль пригодности данных к анализу;
- б) выявление проблем в функционировании систем сбора и консолидации данных;
- в) обнаружение нарушений в структуре данных;
- г) выявление противоречий и дубликатов на уровне записей.

4. Проверка соответствия информации о некотором атрибуте источника данных заданным ограничениям осуществляется в процессе:

- а) профайлинга;
- б) визуальной оценки;
- в) фильтрации данных;
- г) консолидации данных.

5. В основе технологии фильтрации лежат три базовых понятия:

- а) условие;
- б) значение;
- в) отношение;
- г) распределение.

6. Все аномальные значения можно разделить на два класса:

- а) искусственные;

- б) естественные;
- в) количественные;
- г) качественные.

7. Наиболее простым подходом к обнаружению аномалий в одномерных рядах данных является:

- а) статистический;
- б) математический;
- в) экономический;
- г) абстрактно-логический.

8. Снижение размерности данных подразумевает их сокращение по трем направлениям:

- а) сокращение количества признаков;
- б) сокращение числа наблюдений;
- в) сокращение числа значений измерения;
- г) сокращение числа кросс-таблиц.

7.4 Вопросы для самоподготовки

1. Что понимается под качеством данных?
2. Какими способами может быть реализована оценка качества данных?
3. Какие цели преследует оценка качества данных?
4. Из каких этапов состоит задача обработки аномальных значений?
5. Как выглядит пропущенное значение?
6. Какие существуют методы восстановления пропущенных значений?
7. В чем заключается суть сокращения размерности?

8 Data Mining: задача ассоциации

8.1 Вопросы для обсуждения

1. Ассоциативные правила
2. Алгоритм *Apriori*
3. Иерархические ассоциативные правила

8.2 Теоретические сведения

Правила для количественного описания взаимной связи между двумя или более событиями называются ассоциативными правилами.

Базовым понятием в теории ассоциативных правил является транзакция – некоторое множество событий, происходящих совместно.

Ассоциативное правило состоит из двух наборов предметов, называемых условием и следствием и записываемых в виде $X \rightarrow Y$, что читается как «из X следует Y ». Таким образом, ассоциативное правило формулируется в виде: «Если условие, то следствие».

При практической реализации систем поиска ассоциативных правил используют различные методы, которые позволяют снизить пространство поиска до размеров, обеспечивающих приемлемые вычислительные и временные затраты, например, алгоритм *Apriori*.

В основе алгоритма *Apriori* лежит понятие частого набора, который также можно назвать частым предметным набором, часто встречающимся множеством.

Методика поиска ассоциативных правил с использованием частых наборов состоит из двух шагов:

- 1) следует найти частые наборы;
- 2) на их основе необходимо сгенерировать ассоциативные правила, удовлетворяющие условиям минимальной поддержки и достоверности.

Если производить поиск ассоциативных правил среди отдельных предметов (например, товаров в магазине), то можно обнаружить, что во многих случаях ассоциации с высокой поддержкой для отдельных товаров практически отсутствуют. Особенно это характерно для гипермаркетов, где ассортимент товаров каждого вида очень велик. Поэтому при поиске ассоциативных правил в данном случае целесообразно рассматривать не отдельные предметы, а их иерархию. Если на нижних иерархических уровнях интересные ассоциации отсутствуют, то на более высоких они могут иметь место. Иными словами, поддержка отдельного предмета всегда будет меньше, чем поддержка группы, в которую он входит.

Если структура базы данных транзакций позволяет отражать иерархию товаров, то можно исследовать все образованные ими иерархические уровни. Ассоциативные правила, обнаруженные для предметов, расположенных на различных иерархических уровнях, получили название «иерархические ассоциативные правила».

8.3 Вопросы для самоподготовки

1. Какие правила называются ассоциативными?
2. Какие примеры приложения ассоциативных правил можно привести?
3. Что является базовым понятием в теории ассоциативных правил?
4. Из каких наборов предметов состоит ассоциативное правило?
5. Что такое «поддержка ассоциативного правила»?
6. Что такое «достоверность ассоциативного правила»?
7. Что лежит в основе алгоритма *Apriori*?
8. Из каких шагов состоит методика поиска ассоциативных правил с использованием частых наборов?
9. Какие ассоциативные правила являются иерархическими?

9 Data Mining: кластеризация

9.1 Вопросы для обсуждения

1. Сущность и задачи кластеризации
2. Алгоритмы кластеризации
3. Выбор числа кластеров

9.2 Теоретические сведения

Кластеризация – это

1) группировка объектов на основе близости их свойств; каждый кластер состоит из схожих объектов, а объекты разных кластеров существенно отличаются;

2) процедура, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$.

Кластеризацию используют, когда отсутствуют априорные сведения относительно классов, к которым можно отнести объекты исследуемого набора данных, либо когда число объектов велико, что затрудняет их ручной анализ.

Цели кластеризации в *Data Mining* могут быть различными, и зависят от конкретной решаемой задачи:

- изучение данных;
- облегчение анализа;
- сжатие данных;
- прогнозирование;
- обнаружение аномалий.

Первые две задачи наиболее популярны в бизнес-аналитике.

Существует несколько десятков алгоритмов кластеризации. Наиболее понятные и простые в использовании – алгоритм k -средних и сети Кохонена.

9.3 Тестовые задания

1. В кластер S_1 входят четыре объекта, расстояние от которых до пятого объекта составляет соответственно 2,5,6,7. Расстояние от пятого объекта до кластера S_1 , если исходить из метода «ближнего соседа», равно:

- а) 2;
- б) 5;
- в) 6;
- г) 7.

2. В кластер S_1 входят четыре объекта, расстояние от которых до пятого объекта составляет соответственно 2,5,6,7. Расстояние от пятого объекта до кластера S_1 , если исходить из принципа «дальнего соседа», равно:

- а) 2;
- б) 5;
- в) 6;
- г) 7.

3. В кластер S_1 входят четыре объекта, расстояние от которых до пятого объекта составляет соответственно 2,5,6,7. Расстояние от пятого объекта до кластера S_1 , если исходить из принципа «средней связи», равна:

- а) 2;
- б) 5;
- в) 6;
- г) 7.

4. Процедура, при которой начальным является разбиение, состоящее из n одноэлементных классов, называется:

- а) дивизимной;
- б) корреляционной;
- в) агломеративной;
- г) линейной.

5. Кластерный анализ позволяет проводить:

- а) группировку объектов;
- б) группировку объектов и группировку признаков;
- в) группировку коэффициентов корреляции;
- г) группировку дисперсий;
- д) группировку признаков.

6. Цель кластерного анализа –

- а) образование групп сходных объектов;
- б) установление по результатам наблюдений адекватной аналитической зависимости между факторами;
- в) оценка степени тесноты связи между переменными;
- г) исследование влияния случайных величин на изменчивость факторов.

9.4 Вопросы для самоподготовки

1. Что такое «кластеризация»?
2. Какие цели преследует кластеризация?
3. Какие алгоритмы кластеризации используются чаще всего?
4. Что такое «метрика»?
5. Как определяется Евклидово расстояние?
6. Как вычисляется расстояние Манхэттена?
7. Что лежит в основе построения сети Кохонена?
8. Каких правил необходимо придерживаться при использовании кластеризации?

9.5 Задача для практического занятия

Задача 9.5.1 Из базы данных кафедры статистики и прикладной математики в *Statistica* загрузить файл *Excel Nedvig.xls*. Провести иерархическую классификацию недвижимости, используя правило объединения (метод) Варда и Евклидову меру близости. В качестве переменных для анализа

выбрать только количественные переменные для однокомнатных квартир: общую площадь, жилую площадь, площадь кухни, цену (V5, V6, V7, V8).

Методические указания.

1. Выполнить команду *Анализ (Statistics) – Многомерный разведочный анализ (Multivariate Exploratory Technicues) – Кластерный анализ (Cluster Analysis) – Иерархическая классификация (Joining (treeclustering))*. Заполнить соответствующее диалоговое окно.

2. Визуализировать данные в виде вертикальной дендрограммы (*OK – Вертикальная дендрограмма (Verticalicicle plot)*).

3. Найти 3 кластера с помощью метода *K* средних (для этого вернуться в предыдущее окно и выбрать кластеризацию методом *K* средних: *Отмена (Cancel) – Кластеризация методом K средних (K-means clustering)*).

4. Определить расстояния между кластерами, средние площади и цены для кластеров. Выполнить дисперсионный анализ результатов классификации.

5. С помощью дискриминантного анализа найти функции классификации, позволяющие отнести новый объект (квартиру) к одному из классов по наибольшему значению соответствующей функции. Для этого закрыть окно кластерного анализа и выполнить команду *Анализ (Statistics) – Многомерный разведочный анализ (Multivariate Exploratory Technicues) – Дискриминантный анализ (Discriminant Analysis)* и выбрать группирующую (кластер) и независимые (общая площадь, жилая площадь, площадь кухни, цена) переменные.

10 Data Mining: классификация и регрессия.

Статистические методы принятия управленческих решений

10.1 Вопросы для обсуждения

1. Применение классификации и регрессии
2. Методы классификации и регрессии
3. Простая линейная регрессия
4. Множественная линейная регрессия
5. Регрессия с категориальными входными переменными
6. Методы отбора переменных в регрессионные модели
7. Гетероскедастичность и мультиколлинеарность
8. Логистическая регрессия

10.2 Тестовые задания

1. Зависимость, при которой каждому значению величины X соответствует единственное значение величины Y и наоборот, называется:

- а) статистической;
- б) корреляционной;
- в) функциональной;
- г) математической.

2. Зависимость, при которой каждому фиксированному значению независимой переменной X соответствует не одно, а множество значений переменной Y , называется:

- а) статистической;
- б) корреляционной;
- в) функциональной;
- г) математической.

3. Зависимость, при которой функциональной зависимостью связаны фактор X и среднее значение результативного показателя Y , называется:

- а) статистической;
- б) корреляционной;
- в) функциональной;
- г) математической.

4. Графически изображение реальных статистических данных в виде точек в декартовой системе координат называется:

- а) диаграммой рассеяния;
- б) корреляционным полем;
- в) круговой диаграммой;
- г) верификацией модели.

5. Большинство программ, реализующих алгоритм иерархической классификации, предусматривает графическое представление результатов классификации в виде:

- а) дендрограммы;
- б) блок-схемы;
- в) дерева решений;
- г) ориентированного графа.

6. Цель корреляционного анализа в узком смысле –

- а) дать оценку степени тесноты связи между переменными;
- б) установить по результатам наблюдений адекватную аналитическую зависимость между факторами;
- в) исследовать влияние факторов на изменчивость средних значений наблюдаемых случайных величин;
- г) исследовать влияние случайных величин на изменчивость факторов.

7. Цель регрессионного анализа –

- а) установить по результатам наблюдений адекватную аналитическую зависимость между факторами;
- б) дать оценку степени тесноты связи между переменными;

- в) исследовать влияние факторов на изменчивость средних значений наблюдаемых случайных величин;
- г) исследовать влияние случайных величин на изменчивость факторов.

8. Главная цель метода канонических корреляций –

- а) поиск максимальных корреляций между группами факторных и результирующих признаков;
- б) установление по результатам наблюдений адекватной аналитической зависимости между факторами;
- в) оценка степени тесноты связи между переменными;
- г) исследование влияния случайных величин на изменчивость факторов.

10.3 Вопросы для самоподготовки

1. На какие группы делятся методы решения задач классификации и регрессии?
2. Какие методы классификации и регрессии относятся к статистическим?
3. Какие методы машинного обучения существуют?
4. В чем состоит задача линейной регрессии?
5. Для чего служит множественная регрессия?
6. Какие методы предназначены для оценки значимости регрессионной модели?
7. Какие методы отбора входных переменных в регрессионные модели наиболее популярны?
8. Что такое гетероскедастичность?
9. Что называется мультиколлинеарностью?
10. Какую функцию называют логистической?

10.4 Задача для практического занятия

Задача 10.4 Из базы данных кафедры статистики и прикладной математики в *Statistica* загрузить файл *Excel*

Nedvig.xls. С использованием модуля *Множественная регрессия* (*Statistics – MultipleRegression*) построить и проанализировать модель стоимости жилья в г. Краснодаре.

Методические указания.

1. Для анализа переменной V8 (цена) выполнить команды:
а) *Графика – 2М Графики – Гистограммы*, вкладка *Дополнительно – Тип графика Простой – Распределение Нормальное – Переменные V8 – ОК* (*Graphs – 2D Graphs – Histograms – Advanced – Regular – Normal – Variables V8 – ОК*);

б) *Графика – 2М Графики – Диаграмма рассеяния* (*Graphs – 2D Graphs – Scatterplots*). При этом переменная X – порядковый номер; Y – цена, тыс. руб.;

в) с помощью диаграммы размаха проанализировать однородность данных (команда *Графики* (*Graphs*) – *2М графики* (*2D Graphs*) – *Диаграммы размаха* (*Box Plots*)). В качестве зависимой переменной выбрать цену квартиры, в качестве независимой – число комнат.

2. Проанализировать парные корреляции факторов V5 (общая площадь), V6 (жилая площадь), V7 (площадь кухни), V8, выполнив команду: *Анализ – Основные статистики и таблицы – Парные и частные корреляции – Матрица парных корреляций* (*Statistics – Basic Statistics – Correlation matrices*).

3. Провести регрессионный анализ, открыв модуль множественной регрессии *Анализ – Множественная регрессия – вкладка Дополнительно* (*Statistics – Multiple Regression – Advanced*), и выбрав *Описательные статистики, матрицы корреляций* (*Review descriptive statistics, correlation matrices*) – *ОК*. В качестве зависимой переменной указать цену квартиры (V8), в качестве независимых – V5, V6, V7;

4. Последовательно изучить результаты регрессионного анализа.

11 Анализ и прогнозирование временных рядов

11.1 Вопросы для обсуждения

1. Методы прогнозирования
2. Временной ряд. Компоненты временного ряда
3. Цели и задачи анализа временных рядов
4. Детерминированная и случайная составляющая временного ряда
5. Исследование временных рядов и автокорреляция
6. Модели прогнозирования деятельности субъектов бизнеса

11.2 Вопросы для самоподготовки

1. Что такое прогнозирование с точки зрения технологий анализа данных?
2. На какие группы делятся все методы прогнозирования?
3. Что позволяют получать формализованные методы?
4. На чем основываются эвристические методы прогнозирования?
5. Что используется при комплексном прогнозировании?
6. Что такое «временной ряд»?
7. Какие задачи решаются в ходе анализа временных рядов?
8. Что такое «детерминированная составляющая» временного ряда?
9. Что такое «случайная составляющая» временного ряда?
10. Какие компоненты можно выделить во временном ряду?
11. Когда во временном ряду имеет место автокорреляция?
12. Какие существуют модели прогнозирования?
13. В чем состоит суть отдельных моделей прогнозирования?

11.3 Задача для практического занятия

Задача 11.3 По данным таблицы 4 в *Excel* провести анализ и прогнозирование временного ряда урожайности пшеницы озимой с использованием различных моделей. Сделать вывод.

Таблица 4 – Динамика урожайности пшеницы озимой, ц/га

Год	Урожайность, ц/га	Год	Урожайность, ц/га	Год	Урожайность, ц/га
1986	37,5	1998	49,8	2010	56,7
1987	33,8	1999	49,1	2011	66
1988	37,9	2000	47	2012	46,9
1989	36,8	2001	39,1	2013	48,6
1990	39,2	2002	49,2	2014	49,4
1991	40,8	2003	51,8	2015	45,7
1992	56,2	2004	47,9	2016	43,4
1993	44,5	2005	54,8	2017	31,5
1994	39,6	2006	42,8	2018	44,4
1995	50,6	2007	58	2019	53,4
1996	30,2	2008	48,9	2020	59,5
1997	54,2	2009	48,9	–	–

Темы докладов по дисциплине

1. Современные аналитические информационные системы поддержки принятия решений
2. ETL-системы
3. Типовые процессы принятия управленческих решений
4. Деревья решений
5. Концепция искусственных нейронных сетей
6. Система как объект управления
7. Модель предметной области
8. Ансамбли моделей
9. Сравнение моделей
10. Имитационное моделирование
11. Применение экономико-математических моделей при разработке управленческих решений
12. Система управления реляционными базами данных MS SQL Server
13. Информационный поиск в текстах (Information Retrieval)
14. Система STATISTICA Data Miner
15. Система SPSS Clementine

Рекомендуемая литература

1. Андерсон, Карл Аналитическая культура. От сбора данных до бизнес-результатов / Карл Андерсон; пер. с англ. Юлии Константиновой; [науч. ред. Руслан Салахийев]. — М.: Манн, Иванов и Фербер, 2017.

2. Бражко, Е. И. Управленческие решения : учебное пособие / Е.И. Бражко, Г.В. Серебрякова, Э.А. Смирнов. — 2-е изд. — Москва : РИОР : ИНФРА-М, 2020. — 126 с. — (ВО: Бакалавриат). — DOI: <https://doi.org/10.12737/17836>. - ISBN 978-5-369-00266-7. - Текст : электронный. - URL: <https://znanium.com/catalog/product/1062387>

3. Глебова, О. В. Методы принятия управленческих решений : учебное пособие / О. В. Глебова. — Саратов : Вузовское образование, 2017. — 274 с. — ISBN 978-5-906172-20-4. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/62071.html>

4. Жукова, Г. С. Математические методы принятия управленческих решений : учебное пособие / Г.С. Жукова. — Москва : ИНФРА-М, 2021. — 212 с. — (Высшее образование: Бакалавриат). — DOI 10.12737/1084987. - ISBN 978-5-16-016169-3. - Текст : электронный. - URL: <https://znanium.com/catalog/product/1084987>

5. Литвак, Б. Г. Управленческие решения : практикум / Б. Г. Литвак. - Москва : Московская финансово-промышленная академия, 2012. - (Академия бизнеса). - ISBN 978-5-4257-0029-2. - Текст : электронный. - URL: <https://znanium.com/catalog/product/451396>

6. Методы принятия управленческих решений : учебное пособие / А. А. Рудычев, Е. Н. Чижова, С. П. Гавриловская, Р. А. Мясоедов. — Белгород : Белгородский государственный технологический университет им. В. Г. Шухова, ЭБС АСВ, 2015. — 171 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <http://www.iprbookshop.ru/66664.html>

7. Паклин, Н. Б. Бизнес-аналитика: от данных к знаниям: учебное пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.

8. Юкаева, В. С. Принятие управленческих решений : учебник / В. С. Юкаева, Е. В. Зубарева, В. В. Чувикова. - Москва : Дашков и К, 2016. - 324 с. - ISBN 978-5-394-01084-2.
- Текст : электронный. - URL:
<https://znanium.com/catalog/product/430348>

Оглавление

Введение.....	3
1 Теоретические аспекты принятия управленческих решений.....	4
2 Управление на основе данных.....	11
3 Методология анализа бизнес-данных.....	17
4 Консолидация бизнес-данных.....	23
5 Трансформация бизнес-данных.....	28
6 Визуализация данных.....	35
7 Очистка и предобработка бизнес-данных.....	41
8 Data Mining: задача ассоциации.....	46
9 Data Mining: кластеризация.....	48
10 Data Mining: классификация и регрессия. Статистические методы принятия управленческих решений.....	52
11 Анализ и прогнозирование временных рядов.....	56
Темы докладов по дисциплине.....	58
Рекомендуемая литература.....	59

МЕТОДЫ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

Методические рекомендации

Составители: **Кацко** Игорь Александрович,
Паклин Николай Борисович,
Кремянская Елена Владимировна

Подписано в печать _____ 2021. Формат 60×84 $\frac{1}{16}$.

Усл. печ. л. – 3,6. Уч.-изд. л. – 2,8.

Кубанский государственный аграрный университет.
350044, г. Краснодар, ул. Калинина, 13.