Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической
кибернетики и компьютерных наук

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РЕЛЯЦИОННЫХ И МНОГОМЕРНЫХ РЕШЕНИЙ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Гориной Наталии Николаевны

Научный руководитель
старший преподаватель _____ М. И. Сафрончик

Заведующий кафедрой
доцент,к.ф.-м.н. _____ С. В. Миронов

Саратов 2017

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

С развитием информационных технологий возникла проблема анализа больших объёмов данных, когда невозможно обработать вручную большие массивы данных и принять решение. Решением этой проблемы становится технология Data Mining — извлечение из данных скрытых знаний при помощи различных математических алгоритмов.

Крупнейшими разработчиками в этой области, включившими Data Mining решения в функциональность СУБД, являются Microsoft, Oracle, IBM [1]. В данной работе будут использованы Data Mining решения, поставляемые Microsoft.

Целью данной работы является сравнительный анализ реляционных и многомерных решений для интеллектуального анализа данных. В ходе работы реализуются следующие задачи:

— создание хранилища для конкретной предметной области;
— построение OLAP-куба на основе хранилища данных;
— создание приложения для интеллектуального анализа данных, использующего алгоритмы дерева принятия решений, кластеризации и Байеса на основе хранилища данных и OLAP-куба;
— создание простых прогнозов с использованием конструктора запросов;
— создание простых прогнозов с использованием языка запросов DMX;
— сравнительный анализ двух способов создания моделей интеллектуального анализа данных.

В качестве программных инструментов для реализации проекта выступают СУБД SQL Server 2012 с интегрированной средой конфигурирования и управления Management Studio, а также среда разработки SQL Server Data Tools.

Бакалаврская работа содержит следующие разделы:

— введение;
— теоретическая часть;
— создание источников данных: хранилища и куба;
— создание моделей интеллектуального анализа данных на основе каждого из источников и сравнений многомерных и реляционных решений;
— заключение.

1 Основное содержание работы

В первом разделе работы было дано определение следующим понятиям:

— хранилища данных и такие их свойства, как предметная ориентация, интеграция, поддержка хронологии и неизменяемость;
— OLAP-технология, OLAP-кубы, меры, измерения;
— Data Mining (интеллектуальный анализ данных) и его задачи;
— алгоритмы кластеризации, дерева принятия решения и упрощённый алгоритм Байеса.

В следующем разделе были созданы источники данных для дальнейшего интеллектуального анализа. Средствами SQL-запросов [2–4] в среде SQL Server Management Studio было построено хранилище данных, состоящее из 10 таблиц измерений и 2 таблиц фактов, отражающих продажи абонементов и посещения клиентами клубов сети. Общий вид схемы хранилища показан на рисунке 1.

По построенному хранилищу средствами SSAS в среде разработки SQL Server Data Tools [5, 6] был разработан OLAP-куб.

В качестве представления источника данных была выбрана совокупность из 6 таблиц: dimDate, dimClients, miniDimClientBodyParameters, dimDiscounts, dimSubscriptions, FactSales, две из которых (FactSales, dimSubscriptions) впоследствии были выбраны таблицами групп мер. Из таблицы dimSubscriptions мерами были выбраны Price и Length, а из FactSales — TotalAmount.

Далее, было выбрано 4 измерения: dimClients, dimDate, dimDiscounts, dimSubscriptions. На рисунках 2, 3, 4, 5 приведены атрибуты каждого измерения.

Готовый куб был обработан и развернут.

В третьем разделе дипломной работы была построена структура интеллектуального анализа данных на основе реляционного источника и на основе OLAP-куба. Для каждой из структур было создано по три модели интеллектуального анализа.

Перед тем, как добавить представление источника данных в структуру интеллектуального анализа на основе реляционного источника, в среде SQL Server Management Studio были построены 2 представления, где первое представляет собой уже совершившуюся историю покупок абонементов

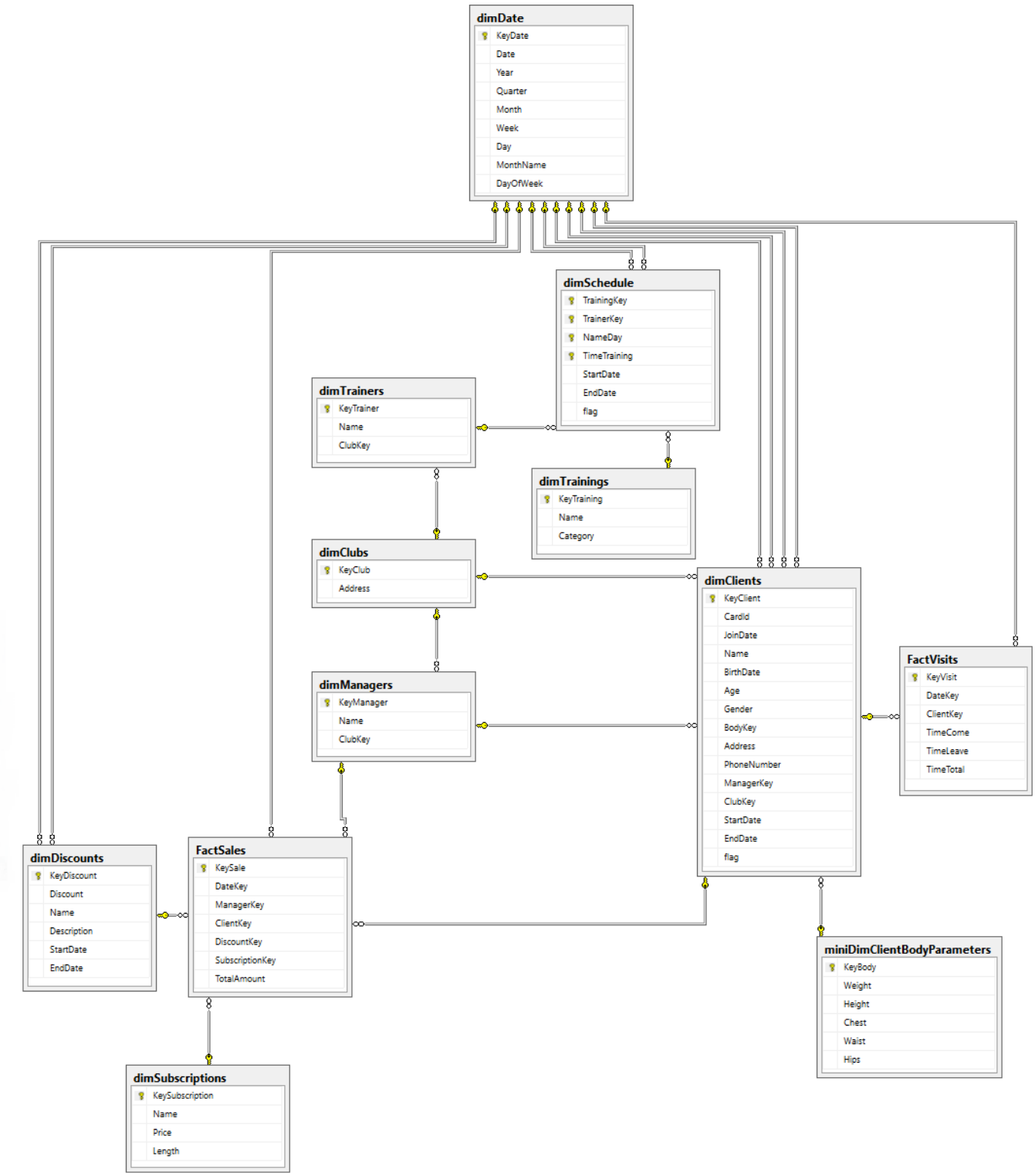


Рисунок 1 – Диаграмма базы данных сети фитнес-клубов

в фитнес-клубы с учетом скидки на каждую покупку и с учетом различных характеристик покупателей. На основании этой истории будет обучаться приложение интеллектуального анализа данных, а второе будет использовать информацию из первого для прогноза. Создание представлений существенно упрощает аналитику и прогнозы для реляционного источника.

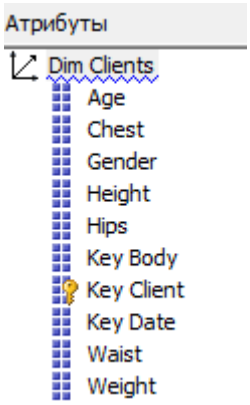


Рисунок 2 – Атрибуты измерения «Клиенты»

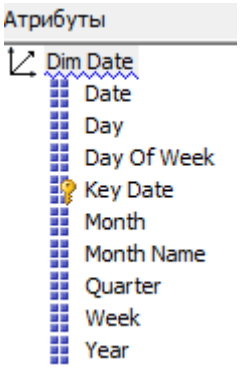


Рисунок 3 – Атрибуты измерения «Дата»

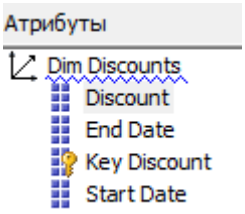


Рисунок 4 – Атрибуты измерения «Скидки»

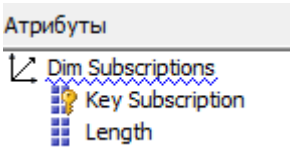


Рисунок 5 – Атрибуты измерения «Абонементы»

В среде SQL Server Data Tools эти представления были добавлены в представление источника данных. Далее, были созданы 2 структуры интеллектуального анализа данных — на основе реляционного источника и на основе куба.

Построение структур начинается с окна выбора метода определения структуры интеллектуального анализа данных. В первом случае выбирется

метод на основе реляционного источника, во втором — на основе куба.

Далее, необходимо определить, нужно ли создавать модель интеллектуального анализа данных и если да, то какой метод она будет использовать. Для обоих методов сначала был выбран алгоритм дерева принятия решений.

На следующем этапе в случае создания структуры интеллектуального анализа данных на основе реляционного хранилища необходимо указать таблицу вариантов, т.е. таблицу, данные которой будут проанализированы средствами Microsoft с целью найти закономерности и построить прогнозы на их основе. Было выбрано одно из созданных ранее представлений. В случае создания структуры интеллектуального анализа данных на основе OLAP-куба нужно выбрать единственное измерение, которое будет использоваться для выбора столбцов уровня вариантов. Было выбрано измерение Client.

После определения таблицы вариантов в случае разработки необходимо отметить ключевой столбец в этой таблице, однозначно определяющий каждую строку, прогнозируемый столбец и входные столбцы, от которых будут выявляться зависимости прогнозируемого столбца. В качестве ключевого столбца был взят номер продажи, прогнозируемого — длительность абонемента, а за входные столбцы взяты различные атрибуты, характеризующие клиентов. В случае разработки на основе OLAP-куба далее необходимо выбрать единственное измерение куба, атрибуты которого будут использоваться в качестве столбцов уровня вариантов. Было выбрано измерение Client. Так так в этом измерении нет прогнозируемого столбца Length и еще некоторых характеристик для анализа, необходимо добавить еще вложенные измерения-таблицы. Итоговый набор столбцов для интеллектуального анализа данных на основе куба приведён на рисунке 6.

На следующем этапе в обоих случаях необходимо для каждого столбца задать тип данных (число, текст и т.д.) и тип содержимого (непрерывное или дискретное значения). Для обеих моделей были указаны идентичные типы для каждого столбца.

В случае разработки структуры интеллектуального анализа данных необходимо выполнить срез куба по одному или нескольким измерениям.

На следующем этапе в обоих методах разработки необходимо указать процент проверочных данных (по умолчанию 30), а также задать максимальное количество вариантов в наборе проверочных данных (было указано зна-

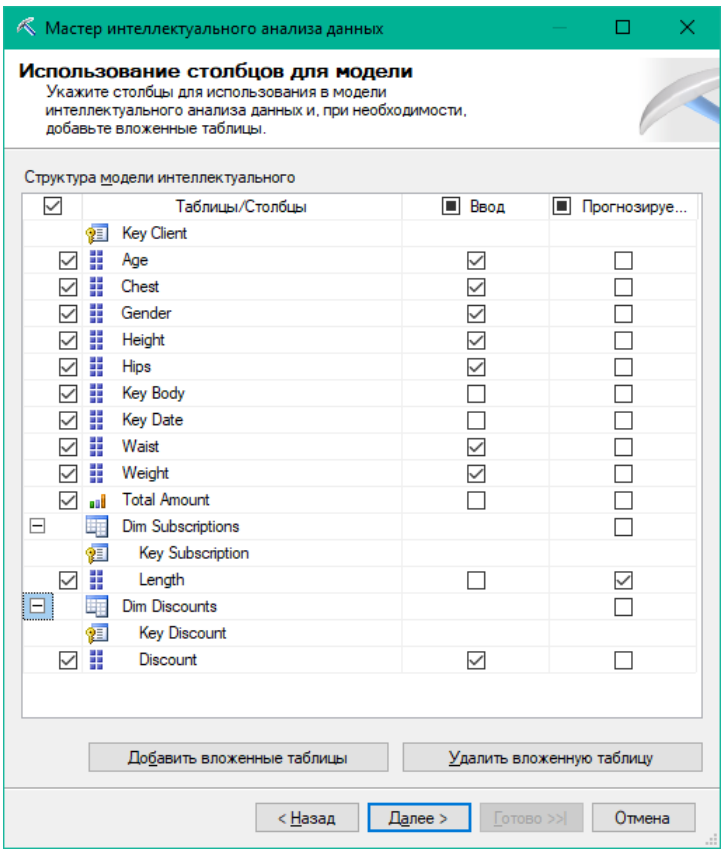


Рисунок 6 – Набор столбцов для интеллектуального анализа данных

чение 10000).

Затем даётся название структуре и модели интеллектуального анализа данных. После чего созданная структура обрабатывается и отдельно разворачивается модель данных. В соответствии с этим для модели, основанной на хранилище данных, было построено дерево разбиений на основе обучающих данных (рисунок 7). На картинке самым темным синим цветом выделен узел, характеристики которого являются наивероятнейшим условием покупки годового абонемента, т.е. в период скидки в размере 20 процентов человек старше 39 лет наиболее вероятно купит абонемент на год, и чуть менее вероятно — на три месяца. В результате анализа этого узла зафиксировано 79 из 140 случаев именно такой покупки. И напротив, наименьшая вероятность купить абонемент на 12 месяцев у людей моложе 29 лет вне зависимости от скидки — 69 из 482 молодых людей купили абонемент на год.

Для структуры интеллектуального анализа данных на основе куба дерево разбиений строится для каждого значения прогнозируемого атрибута (рисунок 8).

Так, для модели интеллектуального анализа данных, использующей ал-

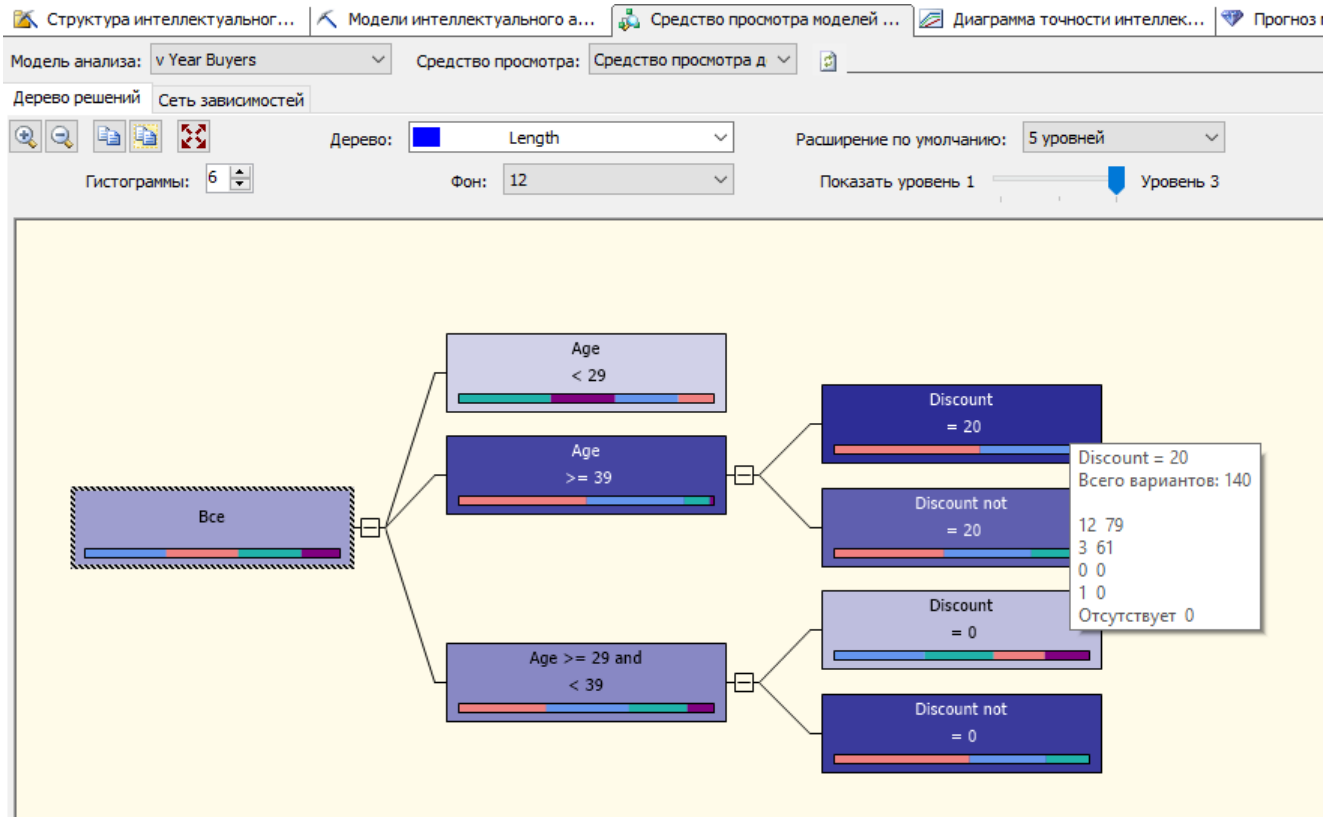


Рисунок 7 – Дерево разбиений

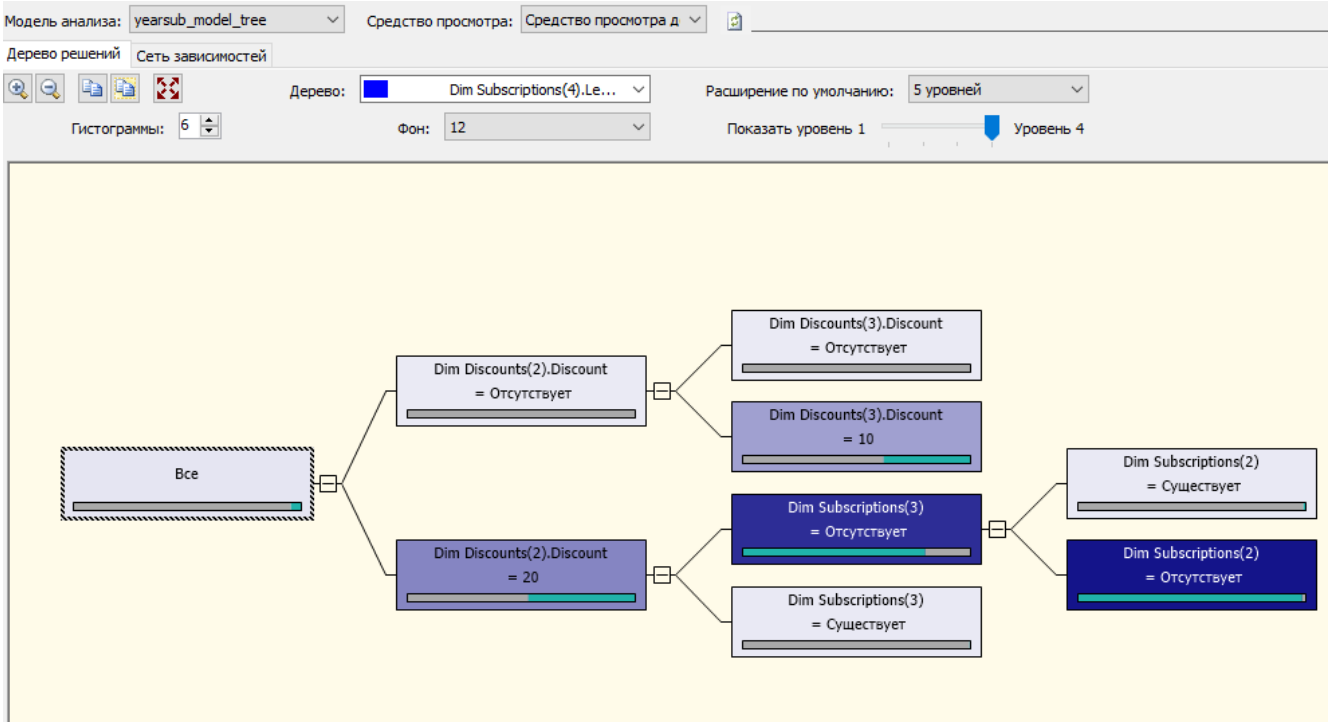


Рисунок 8 – Дерево решений для покупок годового абонемента

горитм кластеризации Вкладка «Профили кластеров» содержит общие сведения о модели в случае разработки на основе хранилища (рисунок 9) и на основе куба, где также показаны зависимости от каждого из членов измере-

ний 10. На вкладке «Профили кластеров» есть столбец для каждого кластера модели. В первом столбце перечислены атрибуты, связанные по крайней мере с одним кластером. В оставшейся области средства просмотра отображается распределение состояний атрибута для каждого из кластеров. Распределение дискретной переменной показано цветным столбцом, при этом максимальное количество видимых столбцов задается в списке «Столбцы гистограммы». Непрерывные атрибуты отображаются в ромбовидной диаграмме, отражающей среднее и стандартное отклонение в каждом из кластеров.

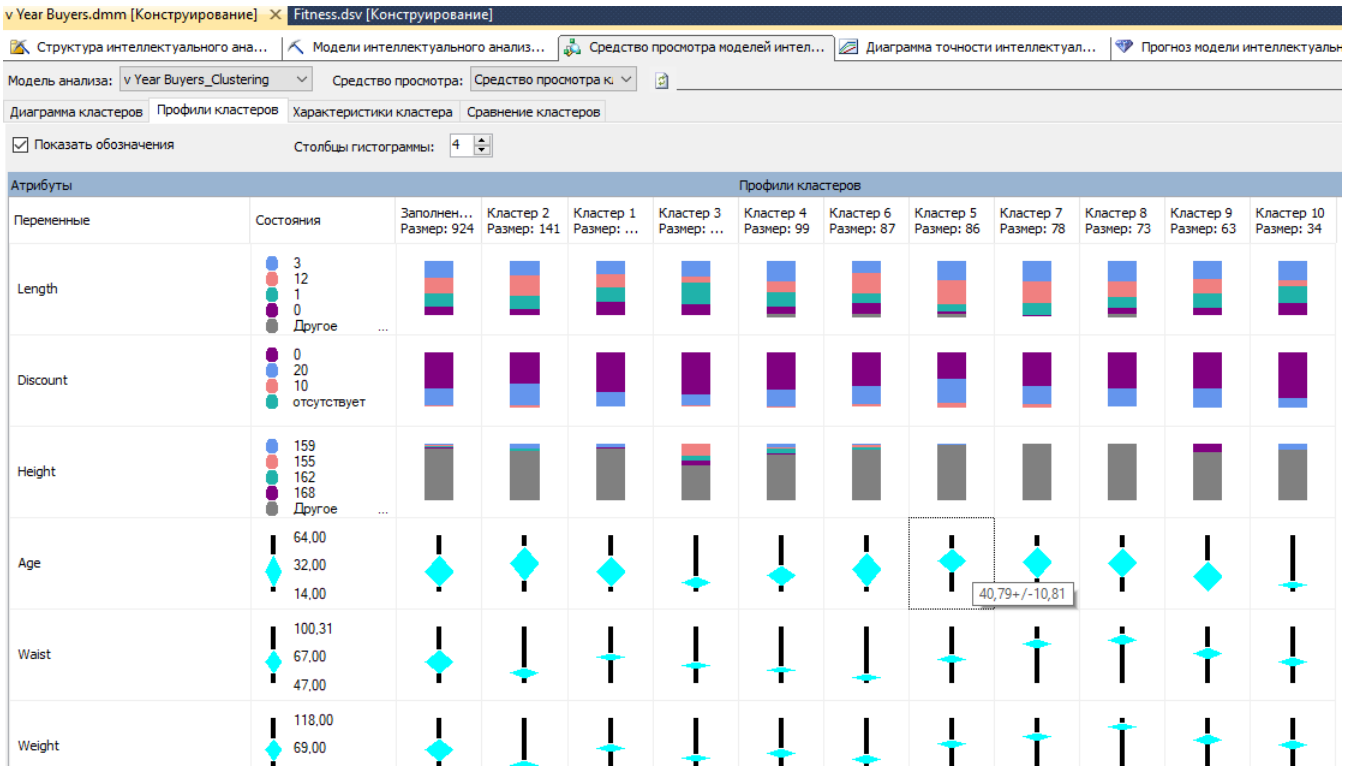


Рисунок 9 – Профили кластеров

Также была построена модель, использующая упрощенный алгоритм Байеса. Поскольку этот алгоритм не поддерживает непрерывные значения, то в сети зависимостей оказались только 2 узла — прогнозируемый Length и Discount, который содержит дискретное значение. Аналогично, в сети зависимостей для структуры на основе куба оказались атрибуты измерений Dim Discounts и Dim Subscriptions — Discount и Length.

На вкладке «Профили атрибутов» модели на основе хранилища показано, как различные состояния единственного входного атрибута влияют на результат прогнозируемого атрибута (рисунок 11). Из рисунка видно, что в большинстве случае покупка годового абонемента имела место во время проведения акций с 20-процентной скидкой.

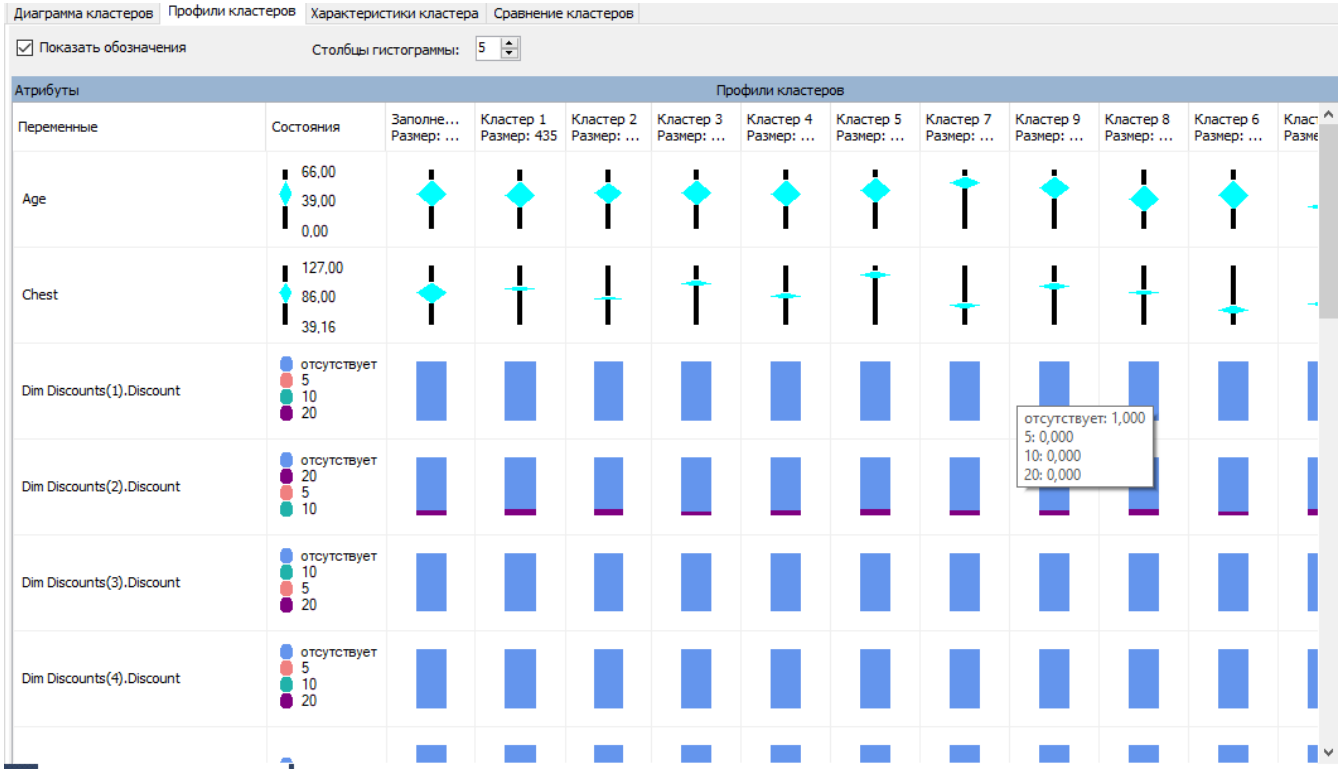


Рисунок 10 – Профили кластеров по атрибутам

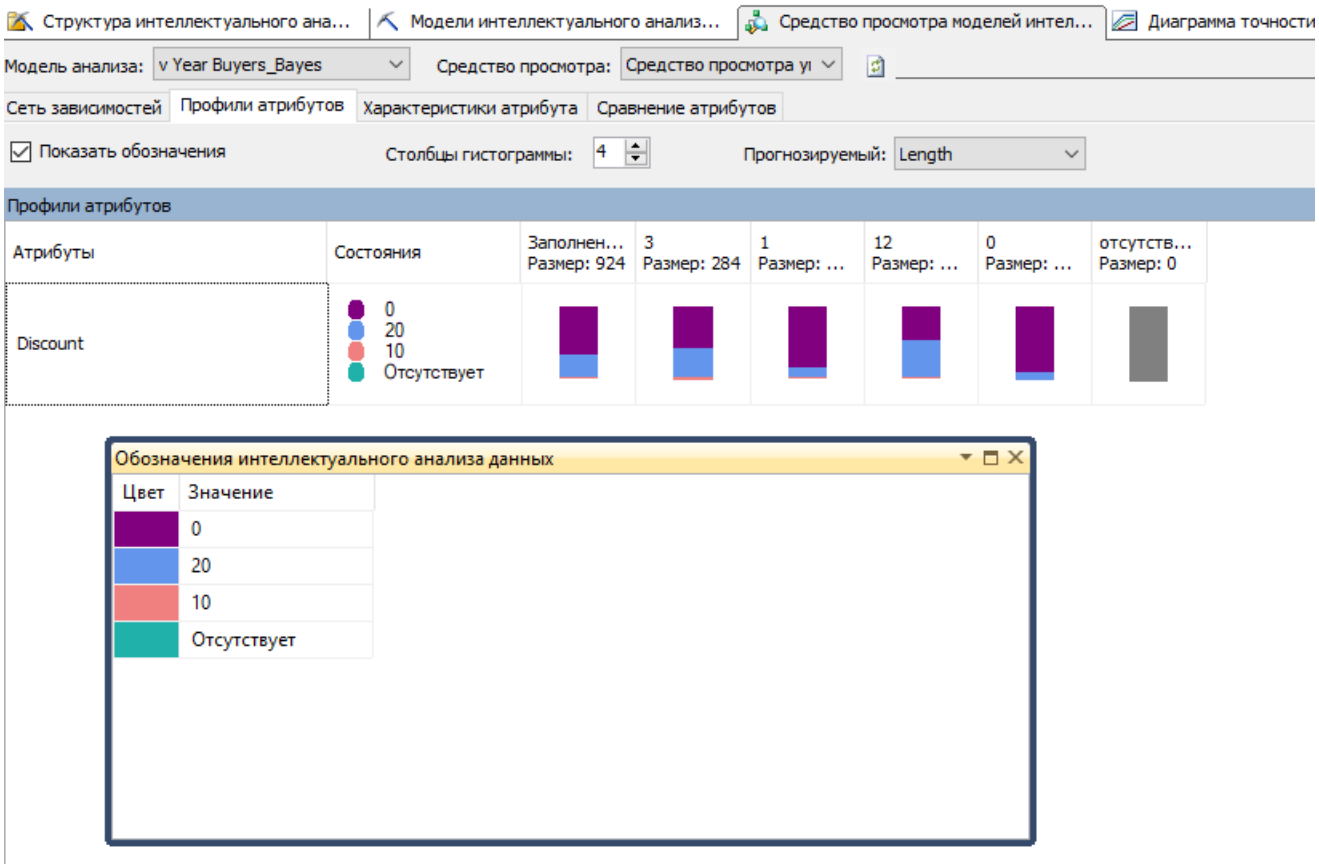


Рисунок 11 – Профили атрибутов

На рисунке 12 показан профиль атрибута измерения Dim Subscriptions(4)

для модели на основе куба.

Рисунок 12 – Профили атрибутов

На основе модели интеллектуального анализа, построенной на основе реляционного источника, использующей алгоритм кластеризации (т.к. только в этой модели за прогнозирующие атрибуты берутся не только возраст и процент скидки) в конструкторе запросов был построен прогноз (рисунки 13, 14) [7,8]. Результаты показывают, с какой вероятностью клиенты из представления `v_year_prospective_buyers` купят абонемент на год. Вероятность указана в самом левом столбце.

Если необходимо получить вероятности каждого варианта покупки, можно воспользоваться функцией PredictHistogram; если спрогнозировать какое-то событие исходя из конкретных значений параметров — NATURAL PREDICTION JOIN (автоматически сопоставляет имена столбцов исходного запроса, совпадающих с именами столбцов в модели) [9–11]. Ниже представлен пример запроса, который для 35-летней женщины определит вероятности покупки каждого абонемента с учетом 10-процентной скидки. Результаты запросы никуда не записываются, а однократно выводятся на экран (рисунок 15). Вероятности покупки показаны в столбце PROBABILITY.

```
SELECT
  [v Year Buyers_Clustering].[Length],
  PredictHistogram([Length])
FROM
```

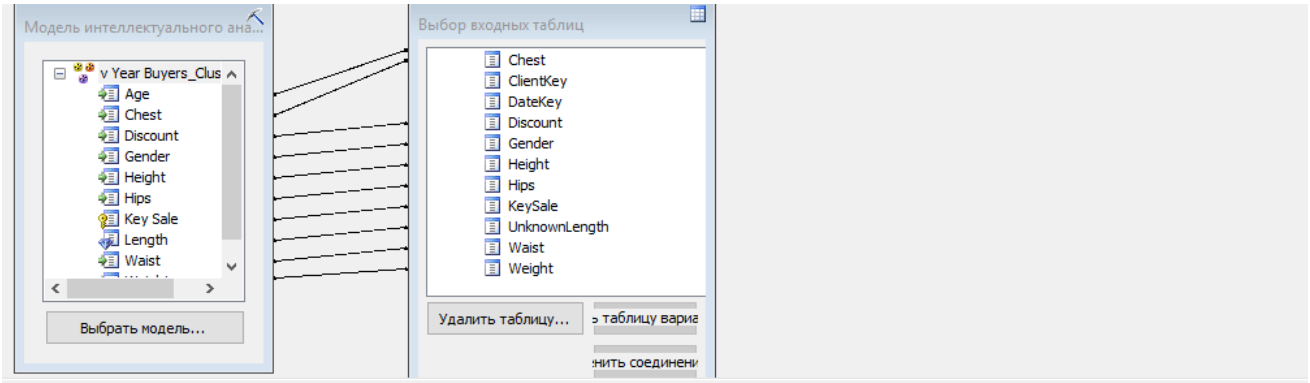


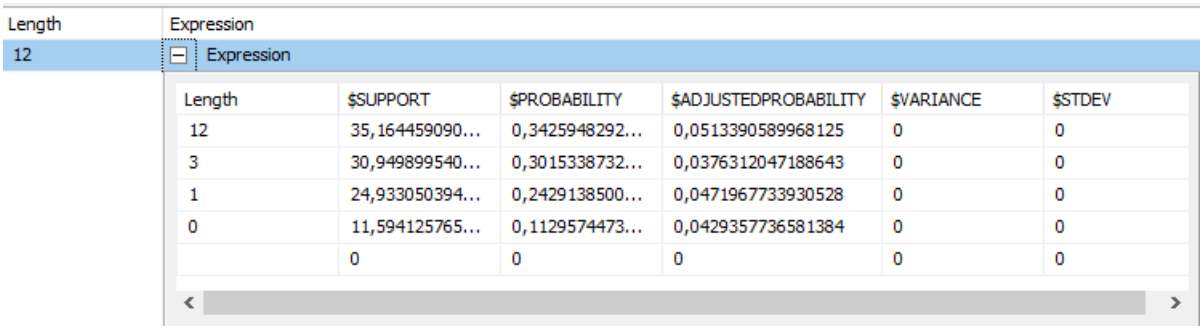
Рисунок 13 – Прогнозирующий запрос

Вероятность результата	Length	KeySale	DateKey	ClientKey	CardId	Age	Gender	Weight	Height	Chest	Waist	Hips	Discount
0,37277426626419	12	1604	20160622	2440	149	52	Ж	46	162	80	54	83	0
0,269752592865215	12	1420	20160505	2102	268	30	Ж	79	186	106	74	108	10
0,385516067456409	12	1601	20160621	2418	354	31	Ж	86	196	103	77	110	0
0,385516067456409	12	1629	20160709	2418	354	31	Ж	86	196	103	77	110	0
0,385516067456409	12	1799	20160814	2418	354	31	Ж	86	196	103	77	110	0
0,385516067456409	12	1955	20160921	2418	354	31	Ж	86	196	103	77	110	0
0,388829129325553	12	1356	20160503	2115	130	47	Ж	86	191	109	79	108	10
0,37277426626419	12	1441	20160506	2023	272	49	Ж	57	173	91	59	86	10
0,37277426626419	12	1749	20160809	2529	369	57	Ж	55	167	93	63	88	0
0,444505975953525	12	1380	20160504	1895	311	57	M	80	183	82	70	82	10
0,370340727672443	12	1329	20160503	1952	218	46	Ж	51	158	94	64	93	10
0,418878001749585	12	1510	20160525	2183	42	33	M	76	174	82	69	83	0
0,444536383781918	12	2135	20161115	3108	344	48	M	69	158	78	71	84	0
0,37277426626419	12	1362	20160503	1833	282	43	Ж	36	150	80	53	88	10
0,388829129325553	12	1794	20160813	2377	130	47	Ж	87	191	110	80	108	0
0,388829129325553	12	2354	20170129	3191	130	48	Ж	87	191	110	80	108	0
0,37277426626419	12	1498	20160516	2029	120	30	Ж	33	151	82	50	81	0
0,37277426626419	12	1590	20160618	2029	120	30	Ж	33	151	82	50	81	0
0,37277426626419	12	1609	20160626	2029	120	30	Ж	33	151	82	50	81	0
0,37277426626419	12	1674	20160729	2029	120	30	Ж	33	151	82	50	81	0
0,37277426626419	12	1450	20160507	1946	6	42	Ж	42	158	89	53	84	10
0,37277426626419	12	1750	20160810	1946	6	42	Ж	42	158	89	53	84	0
0,388582125682569	12	1533	20160604	2007	67	33	Ж	80	186	107	80	111	0
0,444536383781918	12	2621	20170505	3833	73	59	M	74	189	67	58	73	10
0,444536383781918	12	1357	20160503	1717	131	41	M	80	182	83	70	84	10
0,444536383781918	12	1762	20160811	2634	350	42	M	73	166	83	67	77	0
0,444150942704109	12	1434	20160505	1513	292	59	M	87	189	86	73	81	10
0,37277426626419	12	1445	20160506	1523	381	25	Ж	39	158	80	51	80	10
0,408408618940826	12	1376	20160504	1823	22	17	M	82	176	84	72	84	10

Рисунок 14 – Результаты прогноза

```
  [v Year Buyers_Clustering]
NATURAL PREDICTION JOIN
```

```
(SELECT 35 AS [Age],
  'Ж' AS [Gender],
   10 AS [Discount]) AS t
```



Length | Expression

12 | Expression

Length	$SUPPORT	$PROBABILITY	$ADJUSTEDPROBABILITY	$VARIANCE	$STDEV
12	35,164459090...	0,3425948292...	0,0513390589968125	0	0
3	30,949899540...	0,3015338732...	0,0376312047188643	0	0
1	24,933050394...	0,2429138500...	0,0471967733930528	0	0
0	11,594125765...	0,1129574473...	0,0429357736581384	0	0
	0	0	0	0	0

Рисунок 15 – Результаты прогноза

Построение прогноза для структуры, которая была построена на основе куба, является невозможным в силу того, что прогнозируемый столбец не может быть столбцом вложенных таблиц.

После разработки структур интеллектуального анализа данных двух типов — на основе реляционного источника и на основе OLAP-куба, можно прийти к следующим выводам:

— разработка структуры интеллектуального анализа данных на основе реляционного источника более понятна и удобна, поддерживает добавление представления в качестве таблицы вариантов, и если требуется быстро определить скрытые зависимости и гарантированно построить прогноз, она как нельзя лучше подойдет;

— если прогнозируемый атрибут содержит малое количество дискретных значений также будет целесообразнее строить структуру на основе реляционного источника;

— если прогнозируемый атрибут входит в состав измерения, которое состоит из очень большого количества членов измерения, и сама база содержит довольно большое количество данных, лучше разрабатывать структуру интеллектуального анализа данных на основе OLAP-куба, чтобы детально рассмотреть зависимости каждого атрибута измерения отдельно.

ЗАКЛЮЧЕНИЕ

В результате данной работы был проведён сравнительный анализ реляционных и многомерных решений для интеллектуального анализа данных. В ходе работы были реализованы следующие задачи:

- создание хранилища для эмуляции работы сети фитнес-клубов;
- построение OLAP-куба на основе хранилища данных;
- создание приложения для интеллектуального анализа данных, использующего алгоритмы дерева принятия решений, кластеризации и Байеса на основе хранилища данных и OLAP-куба;
- развертывание каждой из моделей и приведение конкретных примеров работы каждого из алгоритмов;
- создание простых прогнозов с использованием конструктора запросов;
- создание простых прогнозов с использованием языка запросов DMX;
- сравнение двух способов создания моделей интеллектуального анализа данных, формулирование выводов.

Было выяснено, что более удобным способом анализа данных является построение структуры интеллектуального анализа данных на основе реляционного источника, в то время как построение такой структуры для многомерного источника целесообразно проводить в случае большого количества значений прогнозируемого атрибута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 НОУ Интуит | Лекция | Рынок инструментов Data Mining [Электронный ресурс]. — URL: http://www.intuit.ru/studies/courses/6/6/lecture/200 (Дата обращения 25.05.2017). Загл. с экр. Яз. рус.

2 *Петкович, Д.* Microsoft SQL Server 2012. Руководство для начинающих / Д. Петкович. — Санкт-Петербург: БХВ-Петербург, 2013.

3 *Сарка, Д.* Microsoft SQL Server 2012 Реализация хранилищ данных / Д. Сарка, М. Лах, Г. Йеркич. — Санкт-Петербург: Русская редакция, 2014.

4 Инструкция CREATE TABLE (Transact-SQL) [Электронный ресурс]. — URL: https://msdn.microsoft.com/ru-ru/library/ms174979.aspx (Дата обращения 25.05.2017). Загл. с экр. Яз. рус.

5 Многомерное моделирование (учебник по Adventure Works) [Электронный ресурс]. — URL: https://msdn.microsoft.com/ru-ru/library/ms170208(v=sql.110).aspx (Дата обращения 26.05.2017). Загл. с экр. Яз. рус.

6 *Елманова, Н.* Введение в OLAP-технологии Microsoft / Н. Елманова, А. Федоров. — Москва: Диалог-МИФИ, 2002.

7 Создание моделей интеллектуального анализа данных и выполнение к ним запросов с помощью расширений интеллектуального анализа данных: учебники (службы Analysis Services — интеллектуальный анализ данных) [Электронный ресурс]. — URL: https://msdn.microsoft.com/ru-ru/library/bb895168(v=sql.120).aspx (Дата обращения 26.05.2017). Загл. с экр. Яз. рус.

8 Структура и методы использования прогнозирующих запросов расширений интеллектуального анализа данных [Электронный ресурс]. — URL: https://msdn.microsoft.com/ru-ru/library/ms131992.aspx (Дата обращения 26.05.2017). Загл. с экр. Яз. рус.

9 SELECT FROM <модель> PREDICTION JOIN (расширения интеллектуального анализа данных) [Электронный ресурс]. — URL: https://msdn.microsoft.com/ru-ru/library/ms132031.aspx (Дата обращения 26.05.2017). Загл. с экр. Яз. рус.

10 НОУ Интуит | Лекция | Концепции языка DMX [Электронный ресурс]. — URL: http://www.intuit.ru/studies/courses/2312/612/lecture/13278 (Дата обращения 26.05.2017). Загл. с экр. Яз. рус.

11 Data Mining Extensions (DMX) Reference [Электронный ресурс]. — URL: https://docs.microsoft.com/en-us/sql/dmx/data-mining-extensions-dmx-reference (Дата обращения 25.05.2017). Загл. с экр. Яз. англ.