ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF EDUCATION AND SCIENCE
OF THE REPUBLIC OF KAZAKHSTAN

Л.Н. ГУМИЛЕВ АТЫНДАҒЫ
ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ
УНИВЕРСИТЕТ ИМ. Л.Н. ГУМИЛЕВА
GUMILYOV EURASIAN
NATIONAL UNIVERSITY

ЖАС ҒАЛЫМДАР КЕҢЕСІ

Студенттер мен жас ғалымдардың «Ғылым және білім - 2015» атты X Халықаралық ғылыми конференциясының БАЯНДАМАЛАР ЖИНАҒЫ

СБОРНИК МАТЕРИАЛОВ X Международной научной конференции студентов и молодых ученых «Наука и образование - 2015»

PROCEEDINGS of the X International Scientific Conference for students and young scholars «Science and education - 2015»

УДК 001:37.0
ББК72+74.04
Ғ 96

Ғ96
«Ғылым және білім – 2015» атты студенттер мен жас ғалымдардың X Халық. ғыл. конф. = X Межд. науч. конф. студентов и молодых ученых «Наука и образование - 2015» = The X International Scientific Conference for students and young scholars «Science and education - 2015». – Астана: http://www.enu.kz/ru/nauka/nauka-i-obrazovanie-2015/, 2015. – 7419 стр. қазақша, орысша, ағылшынша.

ISBN 978-9965-31-695-1

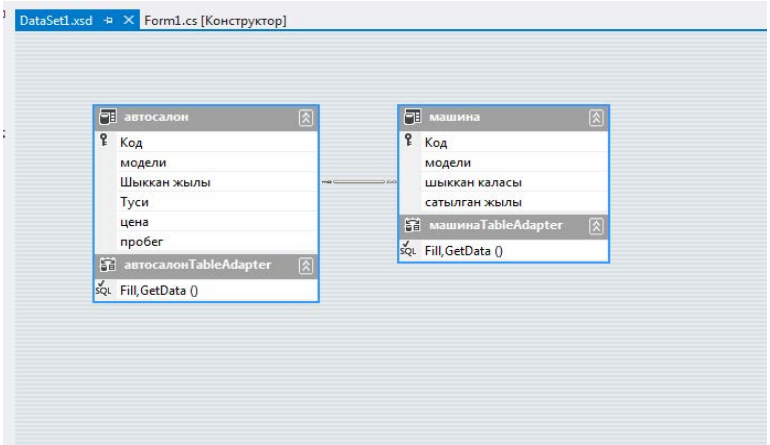
Жинаққа студенттердің, магистранттардың, докоранттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов и молодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 001:37.0
ББК 72+74.04

ISBN 978-9965-31-695-1



Сурет 2. ДҚ схемасы

Қорытындылай келе ADO.NET бұл .NET платформасында ажыратылған жүйелердің құрудың жаңа технологиясы деп айтуға болады.

Қолданылған әдебиет

1.Эндрю Троелсен. «C# и платформа .NET». Библиотека программиста. — СПб.: Питер, 2005 г.

2. Под ред. А.Е.Соловченко. Разработка Windows-приложений на Microsoft Visual Basic .NET и Microsoft Visual C# .NET. Учебный курс MCAD/MCSD/Пер. с англ. — М.:2003

УДК 004.89

ПРИМЕНЕНИЕ МЕТОДОВ АНАЛИЗА ДЛЯ БОЛЬШИХ ДАННЫХ

Исенгалиева Данагуль Бердигалиевна
Студент ЕНУ им. Л.Н.Гумилева, Астана, Казахстан
Научный руководитель –Гарифуллина Ж.Р.

Современный этап развития информационных технологий характеризуется переходом от первого этапа - накопления и структуризации данных – ко второму, который характеризуется их интенсивным использованием на практике. В результате прохождения первого этапа появились большие или очень большие базы данных и, так называемые, хранилища данных. Для эффективной работы с огромными постоянно обновляющимися базами данных в режиме реального времени используются технологии извлечения знаний из баз данных, в частности, Data Mining.

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации). Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей:

- неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем;
- объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным;
- практически полезных - это значит, что выводы имеют конкретное значение, которому можно найти практическое применение [1].

В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные подвыборкам данных, которые могут быть выражены в форме, понятной человеку. По сути, инструмент Data Mining – это пакет

исследования статистических и экономических зависимостей между различными показателями.

Интеллектуальный анализ данных (ИАД) определяется как «извлечение зёрен знаний из гор данных» или «разработка данных – по аналогии с разработкой полезных ископаемых».

Построение модели ИАД является составной частью более масштабного процесса. Этот процесс может быть разделён на шесть базовых этапов [2]. На рисунке 1 представлена диаграмма, отражающая последовательность этапов и технологии Microsoft SQL Server, используемые при ИАД.

Как видно из приведённой диаграммы, создание модели ИАД представляет собой динамический итеративный процесс. Первым этапом процесса ИАД является определение и постановка решаемой задачи. Этот этап включает анализ требований, определение масштаба проблемы, критериев оценки модели и определение цели интеллектуального анализа данных.

На втором этапе процесса ИАД выполняется объединение и очистка данных, определенных на первом этапе.

Третий этап процесса ИАД связан с просмотром и исследованием подготовленных данных. Методы исследования включают в себя расчет минимальных и максимальных значений, расчет средних и стандартных отклонений и изучение распределения данных.

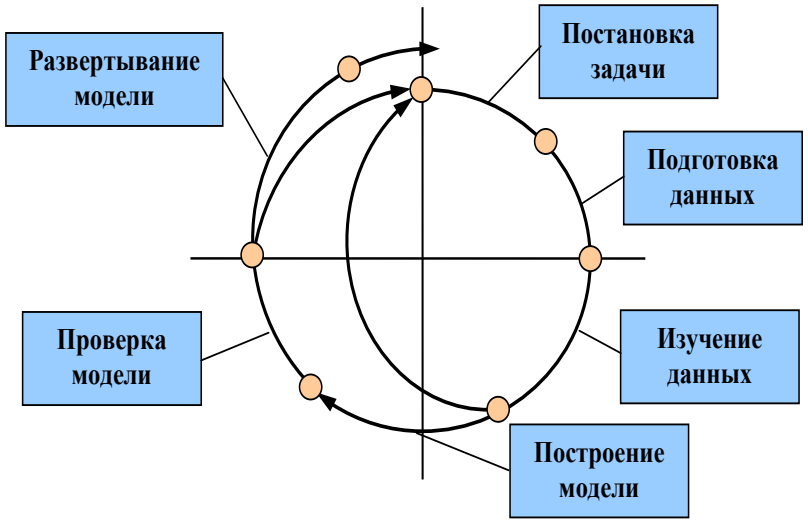


Рисунок 1 – Диаграмма анализа данных [2]

Четвертым этапом процесса ИАД является построение модели. Перед построением модели, рекомендуется случайным образом разделить подготовленные данные в отдельные наборы обучающих и контрольных данных. Набор обучающих данных используется для построения модели, а контрольный набор данных — для проверки точности модели путем создания прогнозирующих запросов. После определения структуры модели интеллектуального анализа данных выполняется ее обработка и наполнение пустой структуры шаблонами, описывающими модель. Данный процесс известен как обучение модели. Шаблоны выявляются путем применения в отношении исходных данных математического алгоритма.

На пятом этапе процесса ИАД осуществляется исследование построенных моделей и проверка их эффективности. Последним шагом процесса ИАД является развертывание в рабочей среде наиболее эффективных моделей.

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data

Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных [3].

Основные математические алгоритмы и методы технологии Data Mining.

1) Алгоритм деревьев принятия решений (Decision Trees) часто является начальной точкой исследования данных. В своей основе он является алгоритмом классификации и хорошо работает при прогнозировании как дискретных, так и непрерывных атрибутов. При построении модели алгоритм учитывает, как каждый входной атрибут в наборе данных влияет на значение прогнозируемого атрибута. Целью является нахождение комбинации входных атрибутов и их значений, которая позволит наилучшим образом прогнозировать значение целевого атрибута.

2) Алгоритм Naive Bayes позволяет быстро строить модели Data Mining, которые можно использовать для классификации и прогнозирования. Алгоритм рассчитывает вероятность, с которой каждое возможное состояние входного атрибута приводит к каждому возможному состоянию прогнозируемого атрибута.

3) Алгоритм кластеризации (Clustering) использует итеративный процесс для группировки записей из набора данных в кластеры, содержащие объекты со сходными характеристиками. Используя кластеры, можно исследовать исходные данные для нахождения в них взаимосвязей. Также на основе кластерной модели можно строить прогнозы.

4) Алгоритм поиска ассоциаций (Association) обеспечивает эффективный метод нахождения зависимостей в больших наборах данных. Данный алгоритм проходит в цикле по всем записям, содержащимся в БД, с целью нахождения значений элементов, которые с наибольшей вероятностью появятся вместе. Такие элементы группируются в наборы элементов, и на их основе генерируются правила, которые затем можно использовать для прогнозирования.

5) Алгоритм нейронной сети (Neural Network), как и алгоритмы деревьев принятия решений и Naive Bayes, в основном используется для исследования данных, классификации и прогнозирования. Алгоритм нейронной сети основывается на методах искусственного интеллекта и исследует все возможные взаимосвязи между данными. Т. к. этот алгоритм исследует данные «тщательнее» других, он является самым медленным из указанных алгоритмов классификации [4].

Реализация кластерного анализа данных на медицинских показателях

Для реализации технологии Data Mining был выбран метод кластеризации. Алгоритм кластеризации применялся для классификации СВА (семейно-врачебных амбулаторий) детской поликлиники по основным, заболеваниям среди детей. Исходными данными для анализа послужили следующие показатели: инфекционные и паразитарные болезни, болезни крови, болезни эндокринной системы, болезни нервной системы, болезни глаза и его придаточного аппарата, болезни уха и сосцевидного отростка, болезни системы кровообращения, болезни органов дыхания, болезни органов пищеварения, болезни кожи и подкожной клетчатки, болезни костно-мышечной системы и соединительной ткани, болезни мочеполовой системы, отдельные состояния, возникшие в перинатальном периоде врожденные аномалии, прочие заболевания, отравления и травмы.

В качестве меры близости использовалось расстояние Евклида. Обработка проводилась при помощи пакета Statistica. Результат кластеризации врачебных участков детской поликлиники по основным заболеваниям детей приведен на рисунке 2.

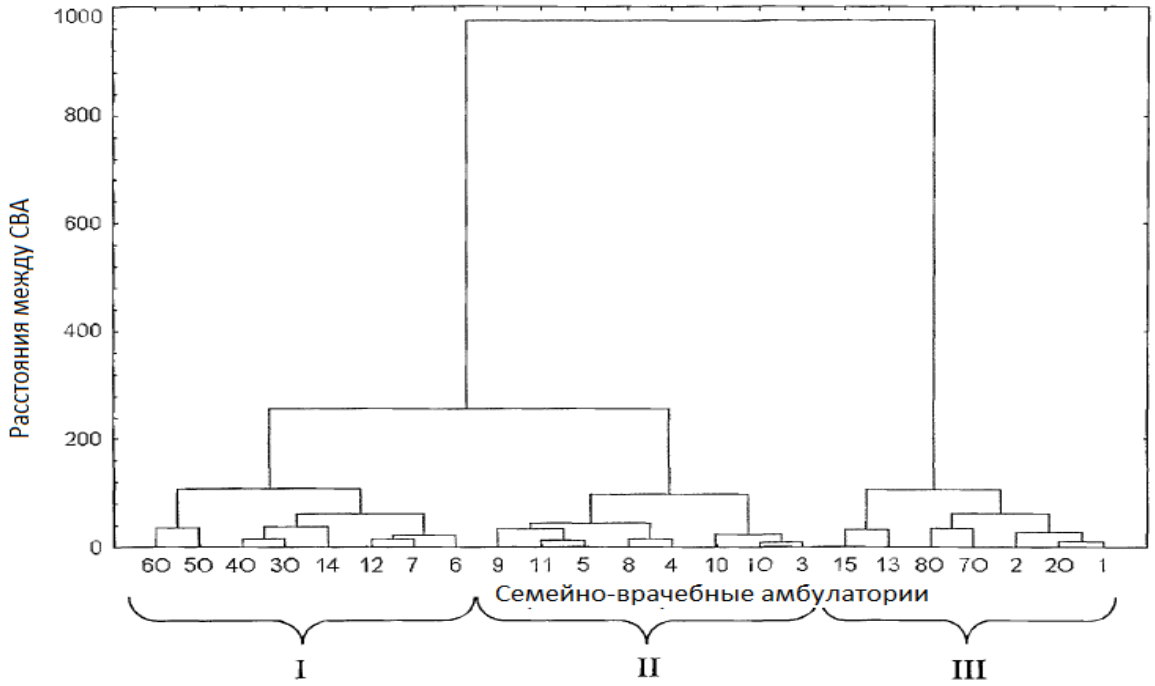


Рисунок 2 - Дендрограмма распределения семейно-врачебных амбулаторий на классы по основным заболеваниям у детей, где, I - класс участков с низким уровнем заболеваемости; II — класс участков со средним уровнем заболеваемости; III - класс участков с высоким уровнем заболеваемости

Для более точной классификации семейно-врачебных амбулаторий детской поликлиники и выделения трех классов в зависимости от уровня заболеваемости детей был применен метод k-средних, результаты которого представлены в таблице 1.

Таблица 1 - Кластеры семейно-врачебных амбулаторий детской поликлиники по основным заболеваниям детей

№ класса	Участки
1	6, 7, 12, 14, 30, 40, 50
2	3,4,5,8,9,10, 11, 10,60
3	1,2, 13,15,20,70,80

Средние количественные показатели уровней рассматриваемых заболеваний для каждого кластера приведены в табл. 2.

Таблица 2 - Численные показатели основных нозологических форм для кластерных групп

Нозология	Уровень заболеваний (на 1000 детей)		
	1 кластер	2 кластер	3 кластер
Инфекционные и паразитарныеболезни	47,22±16,08	37,75±17,96	51,13±14,72
Болезни крови	6,65±0,47	7,07±0,33	8,02±0,72
Болезни эндокринной системы	15,41±0,49	16,61±0,39	19,34±0;88
Болезни нервной системы	25,63±0,83	27,38±0,69	32,01±1,39
Болезни глазаи его придаточногоаппарата	58,67±1,71	63,33±1,73	74,21±2,77

Болезни уха и сосцевидногоотростка	46,37±1,31	50,25±1,38	58,76±2,27
Болезнисистемы кровообращения	1,12±0,22	1,94±0,70	1,61±0,28
Болезни органов дыхания	621,40±17,44'	673,76±19,64	789,05±28,75
Болезни органов пищеварения	39,18±4,02	44,06±1,34	48,58±5,31
Болезни кожи и подкожнойклетчатки	19,01±0,54	20,57±0,56	24,07±0,92
Болезни костно-мышечнойсистемы и соединительнойткани	21,25±4,03	36,46±13,43	30,08±4,98
Болезни мочеполовой системы	14,80±0,42	16,04±0,43	18,75±0,73
Отдельные состояния возникшиев перинатальном периоде	32,75±0,88	35,79±1,06	41,74±1,67
Врожденные аномалии	21,16±7,19	17,05±8,07	23,01 ±6,61
Прочие заболевания	48,98±4,93	55,06±1,61	60,68±6,61
Отравления и травмы	41,88±1,48	56,17±1,49	65,61 ±2,63

Как видно из представленных результатов, третий кластер характеризуется высоким уровнем заболеваний, практически по всем нозологическим формам, за исключением заболеваний системы кровообращения и болезней костно-мышечной системы и соединительной ткани, а первый кластер, наоборот, характеризуется низким уровнем заболеваний практически по всем рассматриваемым нозологическим формам, за исключением инфекционных заболеваний и врожденных аномалий.

Таким образом, результаты кластерного анализа показали, что на 1,2, 13, 15 семейно-врачебных амбулаториях, а также на 2, 7 и 8 семейно-врачебных амбулаториях отрожки отмечается наиболее высокий уровень заболеваемости-детей практически по всем нозологическим формам.

Заключение

Таким образом, в данной статье были изучены особенности технологии Data Mining,основные математические алгоритмы и методы данной технологии, а также реализована классификация заболеваемости детей по семейно-врачебным амбулатория на основе кластерного анализа.

Список использованных источников

1. Григорий Пятецкий-Шапиро, Data Mining и перегрузка информацией // Вступительная статья к книге: Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. 3-е изд. перераб. и доп. СПб.: БХВ-Петербург, 2009. 512. – С.13.-14

2. MicrosoftSQLServer 2008: Datamining – интеллектуальный анализ данных. Пер. с англ. / Дж. Макленнен, Чж. Танг, Б. Криват. – БХВ-Петербург. 2009. – 720 с.

3. Барсегян А.А., Куприянов М.С. и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. –336 с.

4. Бергер А.Б. Microsoft SQL Server 2005 Analysis Services. OLAP и мно-гомерный анализ данных / Бергер А.Б, Горбач И.В., Меломед Э.Л, Щербинин В.А., Степаненко В.П. / Под общ. Ред. А.Б. Бергера, И.В. Горбач. – СПб.: БХВ-Петербург, 2007. – 928 с.

УДК 004.42