УДК 519.673

А.Ю. Варлыгина, А.Ю. Дмитренко, А.Э. Заволодько, А.Г. Ющенко

Национальный технический университет «ХПИ», Харьков

АНАЛИЗ СРЕДСТВ ПОСТРОЕНИЯ МАТЕМАТИЧЕСКОГО ОПИСАНИЯ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ В СУБД SQL SERVER

Статья посвящена анализу возможностей математического описания статистических закономерностей, предназначенных для упрощения процесса моделирования предметной области в пакете Data Mining службы Analysis Services Microsoft SQL Server 2005.

статистические закономерности, процесс моделирования, предметная область, Data Mining

Введение

Постановка проблемы. Одной из важнейших задач на современном этапе развития науки и техники является автоматизация обработки информации и анализа данных, который подразумевает их классификацию, кластеризацию, регрессионный анализ и т.п. Во многих случаях требуется проанализировать данные, многоаспектные связи в которых неочевидны, нетривиальны или не были ранее известны, однако представляют практический интерес. Для решения такого рода задач предложено множество алгоритмов и методов интеллектуального анализа данных, которые, в частности, реализуются технологией Data Mining.

Анализ литературы. Предметной областью исследования является динамика социальных процессов в Украине; рассматриваются проблемы влияния СМИ на отношение населения к НАТО. Динамика мнений различных страт населения и факторы, влияющие на её изменение, приведены в [1].

В настоящее время предложено множество алгоритмов и методов интеллектуального анализа данных: нейронные сети, алгоритм Байеса, кластеризация и т.д., которым посвящено ряд работ [5, 6]. Решения задачи построения оптимальной математической модели можно свести к регрессионному анализу данных, методология которого рассмотрена в [7, 8].

В [2 – 4, 9] рассмотрены алгоритмы анализа данных, которые реализуются технологией Data Mining служб Analysis Services Microsoft SQL Server 2005.

Цель статьи. Целью исследования является анализ возможностей существующих технологий математического описания статистических закономерностей, предназначенных для упрощения процесса моделирования предметной области. Под моделированием понимается построение оптимальной логико-математической модели для прогнозирования зависимостей в сложных статистических данных.

1. Описание объекта исследования

Предметной областью рассматриваемой в работе является: проблема влияния СМИ на отношения населения к НАТО. В качестве исходных данных служили базы, описывающие как динамику мнений населения Украины (54 социальные группы), по которым известны социологические данные за 1997, 1998, 1999 и 2000, так и влияющие факторы (24 коэффициента контент- анализа) информационного пространства (10 изданий) . Для сравнения были использованы данные одной из групп – молодые люди, место жительство которых сельская местность, язык общения – украинский, возраст – от 18 до 29 [1]. Определение модели влияния СМИ и нахождение скрытых закономерностей производился посредством интеллектуального анализа данных технологии Data Mining.

2. Алгоритмы Data Mining, используемые в пакете Analysis Services Microsoft SQL Server 2005

В работе [2] предлагается определить Data Mining, как исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в «сырых данных» скрытых знаний, которые ранее: не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Современная технология Data Mining построена на выделении фрагментов, описывающих многоаспектные взаимосвязи в данных, при поиске которых используются методы, отражающие их неочевидность и регулярность. Методы Data Mining, реализованы средствами службы Analysis Services Microsoft SQL Server 2005. В работах [3, 4] описаны типы алгоритмов, которые реализованы данной службой, они построены на основе общеизвестных методов решения задач анализа данных и прогнозирования [5, 6]:

алгоритм взаимосвязей - это алгоритм анализа данных транзакций, который группирует элементы в наборы и собирает для них статистику;

алгоритм кластеризации представляет собой алгоритм сегментации, который использует итерационные методы для группировки вариантов в наборы данных в кластерах, содержащих подобные характеристики;

алгоритм дерева принятия решений представляет собой регрессивный алгоритм и алгоритм классификации. Для дискретных атрибутов он осуществляет прогнозирование на основе связи между входными столбцами в наборе данных, используя значения или состояния этих столбцов для прогнозирования состояний столбца, который обозначается как прогнозируемый. Алгоритм идентифицирует входные столбцы, которые коррелированы с прогнозируемым столбцом;

алгоритм Байеса предназначен для решения задач классификации и прогнозирования. В процессе реализации алгоритма вычисляются вероятности состояний входных атрибутов для каждого состояния выходного атрибута;

алгоритм нейронной сети создает классификационные и регрессивные модели интеллектуального анализа данных путем построения многоуровневой перцептронной сети нейронов. При получении сведений о каждом состоянии прогнозируемого атрибута алгоритм нейронной сети рассчитывает вероятность возникновения каждого возможного состояния входного атрибута;

алгоритм кластеризации последовательностей анализирует последовательности каких-либо фактов, представляющих собой временные последовательности дискретных переменных. Алгоритм предназначен для прогнозирования наступления последующих событий на основании уже осуществленного перехода между состояниями;

алгоритм временных рядов создает модели, предназначенные для прогнозирования значений непрерывных переменных по времени;

алгоритм линейной регрессии можно использовать для определения связи между непрерывными столбцами. Связь принимает вид формулы, представляющей ряд данных;

алгоритм логистической регрессии представляет собой алгоритм регрессии являющийся частным случаем алгоритма нейронной сети, получаемый в случае удаления скрытого слоя нейросети. Алгоритм поддерживает прогнозирование значений как непрерывных, так и дискретных атрибутов.

Выбор алгоритма для построения оптимальной модели прогнозирования является сложной задачей. При анализе и интерпретации атрибутов различные алгоритмы действуют с различной степенью участия в моделировании, получая на выходе данные разных типов: вероятности возникновения возможного состояния атрибута, зависимости между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной, т.е. модели интеллектуального анализа данных могут прогнозировать значения, создавать обобщения данных и находить скрытые корреляции.

3. Построение оптимальной логико-математической модели

В поставленной задаче уделяется особое внимание возникающим в наборе данных связям, т.е. проведению регрессионного анализа, который рассматривается в [7, 8].

Рассмотрим вышеперечисленные алгоритмы с точки зрения их применимости для решения поставленной задачи:

– алгоритм взаимосвязей и алгоритм Байеса в качестве входных данных принимают только дискретные значения;

– алгоритм логистической регрессии, алгоритм нейронной сети и алгоритм кластеризации позволяют определить вероятностные оценки совместного появления некоторых конкретных значений;

– алгоритм дерева принятия решений и алгоритм линейной регрессии позволяют построить уравнение регрессии и тем самым определить наличие зависимостей в наборе данных.

Построение регрессионной формулы осуществляется с помощью алгоритма дерева принятия решений. Данный алгоритм строит модель интеллектуального анализа данных путем создания ряда разбиений, также называемых узлами, в оптимальном дереве решения. Алгоритм добавляет узел к модели каждый раз, когда выясняется, что входной столбец имеет значительную корреляцию с прогнозируемым столбцом. Способ, которым алгоритм определяет разбиение, отличается в зависимости от того, прогнозирует ли он непрерывный столбец или дискретный столбец.

Когда алгоритм дерева принятия решений строит дерево, основанное на непрерывном прогнозируемом столбце, как в нашем случае, каждый узел содержит регрессионную формулу. Разбиение осуществляется в точке нелинейности в этой регрессионной формуле [9].

В результате исследования влияния средств массовой информации на группу населения с помощью алгоритма дерева принятия решений, были выявлены: зависимость отрицательной (Neg) и нейтральной (Ambi) составляющих мнений выбранной группы от рассматриваемых факторов контент-анализа СМИ, а для положительной составляющей (Pos), при использовании методов данного пакета, зависимости найдены не были.

Рассмотрим пример регрессионной формулы, полученной в результате данного работы алгоритма, на основе исходных данных предметной области, а именно для параметров Ambi и Neg.

Алгоритм выявил следующие зависимости в порядке убывания степени влияния критериев, присутствующих в таблице контент-анализа СМИ:

– призыв (Priziv), положительные сообщения (Plus) и мнение неспециалиста (Nspec) – для нейтральной составляющей мнения группы (Ambi);

– мнение неспециалиста (Nspec), информация (Info), положительные сообщения (Plus), аналитика (Anal), отрицательные сообщения (Minus), агрессия (Aggr), призыв (Priziv) – для отрицательной составляющей мнения группы Neg.

Зависимость от других атрибутов выявлена не была. Регрессионные формулы приведены ниже:

$$Ambi = 495628{,}501 – 1502{,}258 * (Priziv\text{-}36{,}132) + + 266{,}227 *(Plus\text{-}1141{,}103)\text{-}69{,}272 *(Nspec\text{-}1421{,}878),$$

где величина Ambi определяется главным образом величиной Priziv;

$$Neg = 391216{,}148 + 27{,}036*(Minus – 1015.075) + + 335{,}051*(Priziv – 36{,}132) + 130{,}367*(Plus – – 1141{,}103)+ 26{,}059*(Aggr – 955{,}907) + 272{,}191* *(Anal – 828{,}847) – 77{,}650*(Info – 1672{,}068) – – 189{,}616*(Nspec –1421{,}878),$$

где величина Neg определяется главным образом величиной Nspec.

Рассмотрим алгоритм линейной регрессии, который представляет собой разновидность алгоритма дерева принятия решений, получаемый в случае запрета на разбиение узлов в дереве решений.

Сравним результаты, полученные на выходе данного алгоритма, с результатами, описанными выше, при одинаковых исходных значениях. Были выявлены следующие зависимости:

– визуальные эффекты (Visual), отрицательные сообщения (Minus), призыв (Priziv), положительные сообщения (Plus), аналитика (Anal), информация (Info), мнение неспециалиста (Nspec), агрессия (Aggr), зависимость (Zavis), защита (Zasch) – для нейтральной составляющей мнения группы Ambi;

– зависимость (Zavis), финансовая обеспеченность (Finan) и защита (Zasch)– для отрицательной составляющей мнения группы Neg.

Регрессионные формулы имеют вид:

$$Ambi = 495{,}628.160 - 1595{,}343*(Priziv - 36{,}132) – – 115{,}857*(Minus – 1015{,}075) – 127{,}605*(Aggr – –955{,}907) – 259{,}897*(Zasch – 326{,}958) + + 412{,}659*(Plus – 1141{,}103) – 78{,}135*(Nspec – – 1421{,}878) + 46{,}961*(Zavis – 130{,}143) + + 347{,}579*(Anal – 828{,}847) – 8{,}140*(Info – 1672{,}068) – – 52{,}679*(Visaul –1978{,}807),$$

где величина Ambi определяется главным образом величиной Priziv;

$$Neg = 391217{,}864 – 152{,}956*(Zasch – 326{,}958) + + 435{,}208*(Zavis – 130{,}143) – 650{,}582*(Finan – 96{,}760),$$

где величина Neg определяется главным образом величиной Zasch.

Регрессионные коэффициенты (численные значения в формулах) представляют вклады каждой независимой переменной в предсказание зависимой переменной. Для интерпретации направления связи между переменными учитывают знаки (плюс или минус) регрессионных коэффициентов. Если коэффициент положителен, то связь этой переменной с зависимой переменной положительна, если коэффициент отрицателен, то и связь носит отрицательный характер, и если коэффициент равен 0, – связь между переменными отсутствует.

Особо следует отметить, что если несколько столбцов установлены как прогнозируемые или если входные данные содержат вложенную таблицу, которая задана как прогнозируемая, то алгоритм строит отдельное дерево решений для каждого прогнозируемого столбца.

Основное концептуальное ограничение методов регрессионного анализа состоит в том, что они позволяют обнаружить только числовые зависимости, а не лежащие в их основе причинные связи.

Для определения достоверности построенной модели были определены относительные ошибки расчетов, представленные в табл. 1 и 2 соответственно.

Проанализировав полученные результаты можно сделать вывод, что относительная погрешность вычислений на основе регрессионных формул, построенных с помощью алгоритма дерева принятия решений и алгоритма линейной регрессии, значительно различаются между собой. Оценивая достоверность построенных моделей можно заключить, что результат моделирования зависит главным образом от «природы» исходных данных, а именно от существующих в них взаимосвязей и степени их влияния на искомое значение.

Таблица 1

Анализ полученных результатов для параметра Ambi

Год	Ambi (исходные)	Ambi (дерево решений)	Ошибка, %	Ambi (линейная регрессия)	Ошибка, %
1996	411940	456490	10,8	414580	0,6
1997	647334	653500	0,95	646570	0,1
1998	657449	657680	0,04	657770	0,05
1999	390633	386940	0,94	387040	0,9
2000	335379	374220	11,6	351680	4,8
2001	531017	444950	16,2	516150	2,8
Среднее значение ошибки			6,75		1,54

Таблица 2

Анализ полученных результатов для параметра Neg

Год	Neg (исходные)	Neg (дерево решений)	Ошибка, %	Neg (линейная регрессия)	Ошибка, %
1996	386194	385350	0,2	387730	0,3
1997	202292	203030	0,3	201050	0,6
1998	328724	328770	0,01	331660	0,8
1999	659193	656340	1,1	648230	1,6
2000	372643	383060	2,7	419920	12,6
2001	398262	390740	1,9	358710	9,9
Среднее значение ошибки			1,03		4,3

Выводы

В статье были рассмотрены возможности технологии Data Mining службы Analysis Services Microsoft SQL Server 2005 для построения оптимальной логико-математической модели прогнозирования зависимостей в сложных статистических данных. В качестве предметной области рассмотрена проблема влияния СМИ на отношение к НАТО для выбранной социальной группы Украины.

Проведенное исследование позволило выявить взаимосвязи в данных, полученных в результате контент-анализа СМИ, и степень их взаимного влияния.

Авторы признательны доценту В.М. Пошта-ренко за предоставление материалов по пакету Data Mining.

Список литературы

1. Yushchenko A., Postupniy A., Nikitina L., Chernetska T., Zavolodko A., Dovgopol M. Intellectual modeling of information management of political mentality. Dynamics of social Ukrainian stratums towards the NATO, NATO-MW – 2001. – [Електрон. ресурс]. – Режим доступу: http:/ www.nato.int/acad/fellow/00-02/f00-02.htm.

2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.

3. Алгоритмы интеллектуального анализа данных. – [Электрон. ресурс]. – Режим доступа: http://msdn2.microsoft.com/ru-ru/library/ms175595.aspx.

4. Алгоритмы Data Mining. – [Электрон. ресурс]. – Режим доступа: http://www.businessdataanalytics.ru/DataMiningSQLServer2005-2.htm.

5. Люггер Джордж. Искусственный интеллект: стратегия и методы решения сложных проблем. – М.: Издательский дом «Вильямс», 2003. – 864 с.

6. Искусственный интеллект / В.Н. Бондарев, Ф.Г. Аде. – Севастополь: Изд-во СевНТУ, 2002. – 615 с.

7. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: ИНФРА-М, 2003. – 544 с.

8. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, 2003. – 686 с.

9. Алгоритм дерева принятия решений. – [Электрон. ресурс]. – Режим доступа: http://msdn2.microsoft.com/ru-ru/library/ms175312.aspx.

Поступила в редколлегию 12.04.2007

Рецензент: канд. техн. наук, проф. С.А. Соколов, Национальный технический университет «ХПИ», Харьков.