УДК 681.3.06

Д.В. Коломиец

Одесский национальный университет им. И.И. Мечникова, Одесса

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА MICROSOFT CLUSTERING ПРИ АНАЛИЗЕ МЕДИЦИНСКИХ СТАТИСТИЧЕСКИХ ДАННЫХ

В статье рассматривается проблема анализа большого массива накопленных медицинских статистических данных. Представлена математическая постановка задачи кластеризации. Для решения задачи предлагается использование алгоритма Microsoft Clustering, входящего в пакет Microsoft SQL Server Analysis Services. В завершении приведены полученные результаты и возможные варианты их дальнейшего использования.

Ключевые слова: *кластеризация, медицина, статистика, алгоритм, модель, анализ.*

Введение

С ростом объемов накопленных данных и вычислительных мощностей современных компьютеров задачи интеллектуального анализа данных становятся все более и более популярными и востребованными. Одной из отраслей, в которой успешно могут быть применены методы компьютерного анализа данных и извлечения скрытых знаний, является медицина. Известно много экспертных систем установки медицинских диагнозов, построенных главным образом на основе правил, описывающих сочетания симптомов отдельных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как его лечить [1]. Основой для такого рода систем, как правило, является так называемое «хранилище данных». Согласно определению Инмона, хранилище данных – предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений [2]. В работе рассматривается многомерное хранилище данных, представленное в виде гиперкуба, базирующегося на фактах обращения пациентов в медицинское учреждение и измерениях по классам заболеваний, временному измерению, организационному измерению, возрасту и полу. Применение методов «добычи» данных (Data Mining) к такому кубу, в частности, может позволить установить или подтвердить факт наличия подверженности заболевания определенного класса некоторых возрастно-половых групп [3].

Основной материал

В процессе первичного анализа информации, когда о данных мало что известно, часто прибегают к решению задачи кластеризации исследуемых данных. Решение этой задачи помогает лучше понять природу возникновения событий (заболеваний), а также существенно сузить огромный объем информации до набора характерных групп или кластеров, обладающих схожими свойствами. Задача кластеризации состоит в разделении исследуемого множества объектов на группы «схожих» объектов, называемых кластерами (cluster). Одним из достоинств кластерного анализа является также тот факт, что он дает возможность производить разбиение объектов не по одному параметру, а по целому набору признаков и атрибутов.

Формально задача кластеризации статистических данных медицинского учреждения может быть описана следующим образом. Дано множество объектов данных I (факты обращения пациента в поликлинику), каждый из которых характеризуется набором атрибутов. Требуется построить множество кластеров C и отображение F множества I на множество C, т.е. $F: I \rightarrow C$. Отображение F задает модель данных, являющуюся решением задачи. Множество I определим следующим образом:

$$I = \{i_1, i_2, \ldots, i_n\},$$

где i_j – исследуемый объект.

Каждый из объектов характеризуется набором параметров:

$$i_j = \{x_a, x_s, x_d, x_i\},$$

где x_a – возраст; x_s – пол; x_d – признак диспансеризации; x_i – заболевание.

Каждая переменная x_k ($i = \{a, s, d, i\}$) может принимать значения из некоторого множества:

$$x_k = \{v_k^1, v_k^2, \ldots\}.$$

В нашем случае, для параметра x_s множество состоит из двух элементов {«мужской», «женский»}. Для атрибута возраст можно использовать либо все допустимые значения, представленные в таблице фактов, либо можно сформировать возрастные группы, а именно: {«дети», «подростки», «взрослые», «пенсионеры»}.

Задача кластеризации состоит в построении множества:

$$C = \{c_1, c_2, \ldots, c_k, \ldots, c_g\},$$

где c_k – кластер, содержащий «схожие» друг с другом объекты из множества I:

$$c_k=\{i_j, i_k \mid i_j \in I, i_k \in I \text{ и } d(i_j, i_k) < \sigma\};$$

где σ – величина, определяющая меру близости для включения объектов в один кластер; $d(i_j, i_k)$ – мера близости между объектами, называемая расстоянием [4].

Неотрицательное значение $d(i_j, i_k)$ называется расстоянием между элементами i_j и i_k , если выполняются следующие условия:

1) $d(i_j, i_k) \geq 0$, для всех i_j и i_k;
2) $d(i_j, i_k) = 0$, тогда и только тогда, когда $i_j = i_k$;
3) $d(i_j, i_k) = d(i_k, i_j)$;
4) $d(i_j, i_k) \leq d(i_k, i_m) + d(i_m, i_k)$.

Если расстояние $d(i_k, i_k)$ меньше некоторого σ, то говорят, что элементы близки и помещаются в один кластер. В противном случае считается, что элементы отличны друг от друга и их помещают в разные кластеры. Расстояния между объектами предполагают их представление в виде точек m-мерного пространства R^m (в нашем случае m = 4). Чаще всего в качестве меры может быть использовано Евклидово расстояние. Это расстояние вычисляется следующим образом:

$$d(i_j, i_k) = \sqrt{\sum_{t=1}^{m} (x_{jt} - x_{kt})^2} .$$

Иногда, чтобы придать большие веса более отдаленным друг от друга объектам, стандартное евклидово расстояние возводят в квадрат.

Для решения поставленной задачи кластеризации воспользуемся пакетом Microsoft Analysis Services, являющегося частью системы управления базами данных Microsoft SQL Server, и позволяющего хранить данные и выполнять бизнес анализ за счет средств работы с OLAP (OnLine Analytical Processing) и интеллектуальным анализом данных. В пакете реализован метод Microsoft Clustering, базирующийся на двух основных алгоритмах кластеризации: алгоритм k-средних (K-Means) и EM-алгоритме (Expectation Maximization). Оба алгоритма относятся к классу так называемых неиерархических алгоритмов, основанных на оптимизации некоторой целевой функции, определяющей оптимальное разбиение множества данных на группы. Первый метод, алгоритм k-средних, – метод жесткой кластеризации. Это значит, что точка данных может принадлежать только одному кластеру и для принадлежности каждой точки данных этому кластеру вычисляется одно значение вероятности. Второй метод, максимизация ожиданий (EM), – это метод мягкой кластеризации. Это значит, что точка данных всегда принадлежит к нескольким кластерам и для всех возможных сочетаний точек данных с кластерами вычисляются вероятности. Рассмотрим каждый из этих алгоритмов подробнее.

Кластеризация методом k-средних – хорошо известный метод определения принадлежности элементов кластерам с помощью минимизации разницы между элементами кластера и максимизации расстояния между кластерами. Слово «средние» в названии метода относится к центроидам кластеров. Центроид – точка данных, которая выбирается произвольно, а затем итеративно уточняется, пока не начинает представлять собой истинное среднее всех точек данных кластера. «k» означает произвольное количество точек, используемых для формирования начальных значений процесса кластеризации. Алгоритм k-средних вычисляет квадраты евклидовых расстояний между записями данных в кластере и вектор, представляющий собой среднее данного кластера. Метод сходится, выдавая окончательный набор из k кластеров, когда упомянутая сумма минимизирована.

При кластеризации методом EM алгоритм итеративно уточняет начальную модель кластеризации, подгоняя ее к данным, и определяет вероятность принадлежности точки данных кластеру. Этот алгоритм заканчивает работу, когда вероятностная модель соответствует данным. Функция, используемая для установления соответствия, – логарифм правдоподобия данных, вводимых в модель. Если в процессе формируются пустые кластеры или количество элементов в одном или нескольких кластерах оказывается меньше заданного минимального значения, малочисленные кластеры заполняются повторно с помощью новых точек и алгоритм EM запускается снова. Результаты метода масштабируемой максимизации ожидания являются вероятностными. Это значит, что каждая точка данных принадлежит всем кластерам, но с разной вероятностью. Поскольку метод допускает перекрытие кластеров, сумма элементов всех кластеров может превышать число элементов обучающего набора. Поэтому в результирующей модели интеллектуального анализа данных в показатели, выражающие мощность несущего множества, вносится соответствующая поправка.

По умолчанию модели кластеризации Microsoft используют алгоритм масштабируемой максимизации ожидания. Этот алгоритм используется по умолчанию, поскольку он обладает несколькими преимуществами в сравнении с методом кластеризации k-средних:

- не требует больше одного просмотра базы данных;
- работает даже при ограниченном объеме оперативной памяти;
- может использовать однопроходный курсор;
- по производительности опережает методы, требующие выборки.

Реализация Microsoft предоставляет два режима: масштабируемую и немасштабируемую максимизацию ожидания. По умолчанию при масштабируемой

максимизации ожидания просматривается 50 000 записей. В случае успеха модель использует только эти данные. Если модель не удается подогнать на основании 50 000 записей, считываются еще 50 000 записей. При немасштабируемой максимизации ожидания считывается весь набор данных, независимо от его размера. Этот метод создает кластеры более точно, но предъявляет значительные требования к объему памяти. Поскольку метод масштабируемой максимизации ожидания использует локальный буфер, итерации с просмотром всех данных работают быстрее и этот алгоритм гораздо лучше использует кэш памяти процессора, чем метод немасштабируемой максимизации ожидания. Более того, метод масштабируемой максимизации ожидания втрое быстрее метода немасштабируемой максимизации ожидания, даже если все данные умещаются в оперативной памяти. В большинстве случаев выигрыш в скорости не ведет к ухудшению качества окончательной модели.

Применим алгоритм к существующему набору данных, содержащему следующие атрибуты:

- возраст пациента;
- наименование и идентификатор класса заболевания;
- признак диспансеризации заболевания;
- признак, впервые ли зафиксировано обращение пациента по данному заболеванию;
- пол пациента;
- идентификатор факта обращения пациента в поликлинику.

Первое что необходимо сделать на этапе построения модели в среде Microsoft SQL Server Business Intelligence Studio – это классифицировать атрибуты, сопоставив с каждым соответствующую категорию:

- ключевой столбец или атрибут. Каждая модель должна содержать один числовой или текстовый столбец, который однозначно идентифицирует каждую запись. В нашем наборе данных – это идентификатор факта обращения пациента в поликлинику;
- входные столбцы. Модель должна содержать по меньшей мере один входной столбец, включающий значения, которые используются для формирования кластеров. Ограничения на количество входных столбцов не налагаются, но, в зависимости от количества значений в каждом столбце, введение дополнительных столбцов может привести к увеличению времени на обучение модели. Входными атрибутами нашей модели будут: возраст, наименование класса заболевания, признак диспансеризации заболевания и пол пациента. Здесь следует отметить следующую особенность обработки значений атрибутов. При интеллектуальном анализе данных в SQL Server алгоритм обрабатывает данные из столбца в соответствии со свойством «Тип содержимого» этого столбца. Например, если в данных есть столбец «Возраст» с различными значениями, нужно указать, что столбец содержит непрерывный числовой показатель, задав для него тип содержимого «Непрерывный». Однако, интереснее получить результаты анализа в разрезе возрастных групп, таких как: «дети», «подростки», «взрослые» и «пенсионеры». Для этого можно указать, что числа в столбце «Возраст» следует сгруппировать в сегменты, задав тип содержимого «Дискретизированный». При этом явно определить количество и характеристики возрастных сегментов.
- необязательный прогнозируемый столбец. Алгоритму кластеризации не требуется прогнозируемый столбец для формирования модели, однако, предусмотрена возможность добавления прогнозируемого столбца с данными почти любого типа. Значения в прогнозируемом столбце могут рассматриваться как входные по отношению к модели кластеризации, или может быть указано, что эти данные используются только для прогноза.

Результатом работы алгоритма будет построенная модель. При просмотре модели кластеризации в службах Analysis Services, кластеры отображаются на схеме (рис. 1), которая показывает связи между кластерами, а также содержит подробный профиль каждого кластера, список атрибутов, по которым различаются кластеры, и характеристики всего набора данных для обучения.

Средство просмотра модели позволяет также проанализировать характеристики каждого кластера в отдельности. Так для «Кластера 1» таблица характеристик представлена в табл. 1:

Классы, содержащие большее количество объектов со схожими свойствами, отображены на диаграмме более темным насыщенным цветом. Кроме, собственно, кластеров, на схеме отражена также степень сходства кластеров друг с другом – линии, соединяющие кластеры. Светлая или отсутствующая заливка означает, что кластеры не очень схожи. Чем темнее становится линия, тем сильнее становится сходство связей.

Атрибуты, которые содержит кластер, перечислены в столбцах «Переменные», а их состояние – в столбце «Значения». Состояния атрибутов отсортированы в списке по важности, основанной на вероятности их появления в кластере. Вероятность отображается в одноименном столбце.

Выводы

Анализ результатов полученного разбиения позволяет сделать определенные выводы. С одной стороны, можно убедиться в том, что болезни системы кровообращения больше свойственны людям пенсионного возраста, и наоборот, наличие соответствующего кластера дает основание считать, что травмам больше подвержены дети и подростки мужского пола.

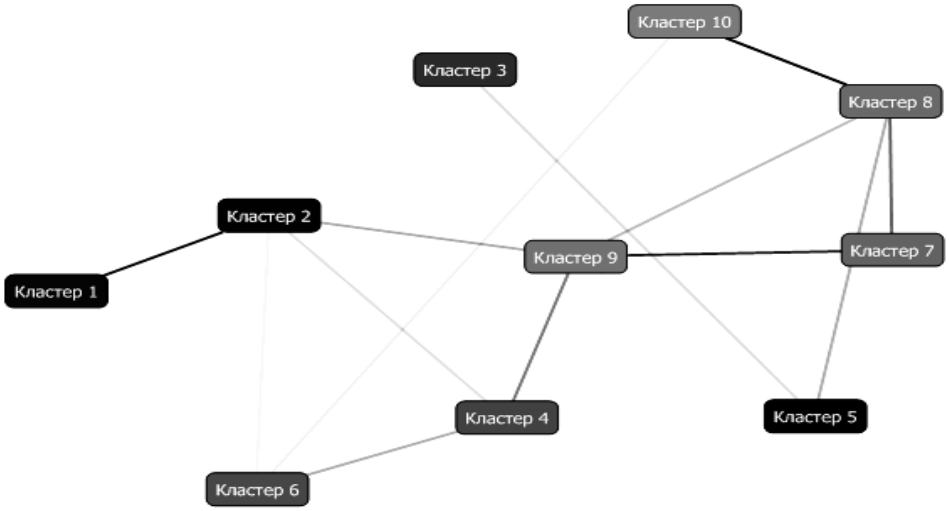

Рис. 1. Схема кластеров

Таблица 1

Таблица характеристик кластера

Переменные	Значения	Вероятность
Возрастная группа	Пенсионеры	100,00%
Категория заболевания	Болезни системы кровообращения	99,47%
Впервые	Истина	91,50%
На диспансерном учете	Ложь	83,41%
Пол	Мужской	50,82%
Пол	Женский	49,18%
На диспансерном учете	Истина	16,59%
Впервые	Ложь	8,50%

Таким образом, можно утверждать, что сформированная модель отражает естественные и логически объяснимые закономерности. Однако бо́льший исследовательский интерес представляют кластеры, содержащие неочевидные, неожиданные закономерности, составляющие так называемые скрытые знания. Т.е. полученное разбиение может служить основой для применения других методов Data Mining, для того чтобы попытаться выяснить смысл, а также разобраться в причинах, вызвавших такое разбиение.

Список литературы

1. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян. – БХБ-Петербург, 2004. – 78 с.

2. Inmon W.H. Building the Data Warehouse. Third Edition. John Wiley & Sons, Inc., 2002.

3. Коломиец Д.В. Многомерная модель медицинского учреждения / Д.В. Коломиец, Т.И. Петрушина // Тезисы докладов международной научно-практической конференции. – 2007. – С. 89-90.

4. Коломиец Д.В. Задача кластеризации статистических данных медицинского учреждения / Д.В. Коломиец // Тезисы докладов международной научно-практической конференции. – 2010. – С. 60-61.

Поступила в редколлегию 17.09.2010

Рецензент: д-р физ.-мат. наук, проф. В.Г. Шевчук, Одесский национальный университет им. И.И.Мечникова, Одесса.

ВИКОРИСТАННЯ АЛГОРИТМУ MICROSOFT CLUSTERING ПІД ЧАС АНАЛІЗУ МЕДИЧНИХ СТАТИСТИЧНИХ ДАНИХ

Д.В. Коломієць

В статті розглядається проблема аналізу великих масивів накопичених медичних даних. Запропонована математична модель задачі кластеризації. Для вирішення задачі пропонується використання алгоритму Microsoft Clustering, що входить до складу пакету Microsoft SQL Server Analysis Services. Наприкінці наведені отримані результати та можливі варіанти їх подальшого використання.

Ключові слова: *кластерізація, медицина, статистика, алгоритм, модель, аналіз.*

MEDICAL STATISTICAL DATA ANALYSIS USING MICROSOFT CLUSTERING ALGORITHM

D.V. Kolomiyets

The problem of huge data massive analysis is investigating in the article. Mathematical model of the clustering task is produced. In order to solve the task Microsoft Clustering algorithm is suggested. There are some results at the end of the article and the way these results can be used is shown.

Keywords: *clustering, medicine, statistics, algorithm, model, analysis.*