УДК 004.891.2

РАЗРАБОТКА ИНСТРУМЕНТОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ХРАНИЛИЩА ДАННЫХ ПО ЗДРАВООХРАНЕНИЮ В РЕГИОНЕ (НА ПРИМЕРЕ ЯРОСЛАВСКОЙ ОБЛАСТИ)

Михайлов М.В.
программист, соискатель учёной степени кандидата технических наук
Ярославский государственный университет, Ярославль

Поступила в редакцию 13.07.2012,

В данной работе приводится разработка инструментов интеллектуального анализа деперсонализованного хранилища данных по здравоохранению в Ярославской области. Разработка ведётся на базе реального хранилища данных. Создаются оптимизированные для данного хранилища данных процедуры поиска ассоциативных правил.

In this article the development of tools for data mining warehouse of health in the Yaroslavl region. The development is based on real data warehouse. It creates optimized procedures for the searching for association rules.

Ключевые слова: интеллектуальный анализ данных, сиквенциальный анализ, поиск ассоциатиавных правил, кластеризация, база знаний, экспертная система.
Keywords: data mining, sequence analysis, searching for association rules, clustering, knowledge base, expert system.

1. Введение

Имеется деперсонализованное хранилище данных по предоставлению медицинской помощи гражданам Ярославской области. Ставится задача разработать специальный инструментарий для интеллектуального анализа хранилища данных. Задача разбивается на подзадачи:

1. Разработать инструментарий для кластеризации данных по произвольным мерам близости.

2. Разработать инструментарий для сиквенциального анализа данных (анализа последовательностей).

3. Разработать инструментарий для поиска ассоциативных правил.

4. Разработать инструментарий для анализа исключений.

Решение данных задач приведёт к более чёткому пониманию тенденций происходящих в системе.

В данной статье в виду объёмности задачи мы разберём только разработку инструментов поиска ассоциативных правил в деперсонализованном хранилище данных по предоставлению медицинской помощи гражданам Ярославской области.

Поиск ассоциативных правил является одной из распространённых задач интеллектуального анализа крупных хранилищ данных и заключается в определении наиболее часто встречающихся сочетаний объектов или их свойств. В здравоохранении эти процедуры можно запустить для выявления в хранилище наиболее часто встречающихся заболеваний одновременно у одного человека или для определения наиболее часто выписываемых друг с другом лекарств.

1. Процесс интеллектуального анализа данных

Применить методы интеллектуального анализа недостаточно для обнаружения знаний в данных, хотя этот этап является основным в процессе интеллектуального анализа.[1]

Весь процесс можно разбить на следующие этапы:

1. Понимание и формулировка задачи анализа.
2. Подготовка данных для автоматизированного анализа(препроцессинг).
3. Применение методов интеллектуального анализа и построение моделей.
4. Проверка построенных моделей.
5. Интерпретация модели экспертом.

Для поиска ассоциативных правил задача была сформулирована во введении данной статьи.

Препроцессинг в данном случае заключается в подготовке данных для применения процедуры поиска ассоциативных правил. Сначала необходимо разработать структуру хранилища данных. В отличие от транзакционной базы данных, целью которой является учёт транзакций, целью хранилища данных является анализ данных. Структура хранилища данных не подвергается процессу приведения к нормальным формам баз данных и имеет структуры: “звезда”, “созвездие”, “снежинка” [2].

На этапе препроцессинга необходимо выработать чёткий набор числовых и нечисловых параметров исследуемых объектов. Этот этап наименее автоматизирован в том смысле, что выбор системы параметров производится человеком, хотя конечно их значения могут вычисляться автоматически.

После выбора описывающих параметров изучаемые данные могут быть представлены в виде прямоугольной таблицы, где каждая строка представляет собой отдельный случай, объект или состояние изучаемого объекта, а каждая колонка – параметры, свойства и признаки всех исследуемых объектов. Алгоритмы поиска ассоциативных правил и других методов интеллектуального анализа работают с такими прямоугольными таблицами.

2. Применение методов поиска ассоциативных правил

Формально опишем задачу:

$$P_i = \{p_{i,1}, p_{i,2}, p_{i,3}, ..., p_{i,j_i}, ..., p_{i,n_i}\}, i = \overline{1,m}, j_i = \overline{1,n_i} \tag{1}$$

, где P_i - это набор полей таблицы транзакций состоящий из m полей, а p_{i,j_i} - одно из n_i количества значений для поля P_i. В нашем случае:

P_1 - идентификатор рецепта

P_2 - идентификатор пациента
P_3 - дата рожденья пациента
P_4 - пол пациента
P_5 - код лечебно-профилактического учреждения
P_6 - Код врача
P_7 - Диагноз (МКБ10)
P_8 - Дата выписки рецепта
P_9 - Код МНН лекарственного средства
P_{10} - Код торгового наименования лекарственного средства
P_{11} - Номенклатурный код лекарственного средства
P_{12} - Лекарственная форма
P_{13} - Дозировка
P_{14} - Количество упаковок
P_{15} - Дата отпуска в аптеке
P_{16} - Код аптеки
P_{17} - Категория льготы пациента
P_{18} - Стоимость лекарственного средства
Объединённые наборы этих свойств называется транзакцией:

$$T_k = \{p_{i,j_i}\}, k = \overline{1,K}, i = \overline{1,m} \tag{2}$$

, где K - количество транзакций, k -номер транзакции, а j_i - номер значения из списка допустимых значений для свойства P_i

Множество транзакций для одного пациента $p_2 = x$ мы обозначим D_{p_2}

$$S_{p_2} = \{T|p_2 = x\} \tag{3}$$

Произвольный набор свойств обозначим F (для свойств P_7 - диагноз и P_9 - код МНН)

$$F = \{p_{i,j_i}|i \in \{7,9\}\} \tag{4}$$

Множество пациентов в рецептах которых есть свойства из набора F обозначим

$$D_F = \{p_{i,j_i}|i \in \{7,9\}\} \subseteq D \tag{5}$$

, где D - множество всех пациентов.

Отношение количества пациентов, у которых в транзакциях содержится набор F, к общему числу пациентов называется поддержкой (support) набора F и обозначается $Supp(F)$

$$Supp(F) = \frac{|D_F|}{|D|} \tag{6}$$

При поиске можно указать минимальное значение поддержки интересующих наборов $Supp_{min}$ Набор называется частным, если значение его поддержки больше минимального значения поддержки. Таким образом при поиске ассоциативных правил требуется найти множество всех частных наборов

$$L = \{F|Supp(F) > Supp_{min}\} \tag{7}$$

2.1 Алгоритм поиска ассоциативных правил

Для поиска ассоциативных правил в общем случае создан алгоритм Apriori, который использует одно из свойств поддержки, гласящее, что поддержка любого набора объектов не может превышать минимальной поддержки любого из его подмножеств.

$$Supp_F \leq Supp_E, E \subset F \tag{8}$$

В нашем случае под конкретную задачу нахождения наиболее часто встречающихся пар диагноз-диагноз, МНН-МНН, диагноз-МНН мы создадим специальный алгоритм, работающий с существующей структурой данных. Этот алгоритм описан ниже и ответит на конкретные вопросы: чем одновременно болеют, чем одновременно лечатся и чем чаще всего лечат то или иное заболевание.

Алгоритм 1 (Функция подсчёта паросочетаний диагнозов у одного пациента):

```
function combination ($a) // Вход - массив диагнозов
{
   $b = Array();
   $max_count = count($a);
   if ($max_count <= 1) return $b;
   $count = 0;
   for ($i = 0; $i<$max_count; $i++ )
   {
      for ($j = $i+1; $j<$max_count; $j++)
      {
         $b[$count] = Array($a[$i], $a[$j]);
         $count++;
      }
   }
   return $b; // Выход - массив паросочетаний]
}
```

Алгоритм 2 (Поиск ассоциативных правил в паросочетаниях заболеваний):

```
mysql_connect("localhost", "root", "root");
mysql_select_db('dlo');
$result = mysql_query("SELECT DISTINCT `SS`, `DS` FROM recipes ORDER BY `SS`, `DS`");
$count = 0;
while ($res = mysql_fetch_object($result)){
   mysql_query("INSERT INTO `assoc_ds`(`SS`, `DS`) VALUES ('{$res->SS}','{$res->DS}')");
}
$result = mysql_query("SELECT * FROM `assoc_ds` ORDER BY `id`");
$count = 0;
$current_ss = 0;
$current_ds = Array();
while ($res = mysql_fetch_object($result)){
   if($current_ss == $res->SS || $current_ss == 0){
      $current_ss = $res->SS;
      $current_ds[] = $res->DS;
   }else{
      $current_ss = $res->SS;
      $combinations = combination($current_ds);
      $current_ds = Array();
      foreach($combinations as $value){
         $result2 = mysql_query("SELECT * FROM `assoc_ds2` WHERE `ds1`='{$value[0]}' and `ds2`='{$value[1]}' LIMIT 1");
         if ($res2 = mysql_fetch_object($result2)){
            $q1 = "UPDATE `assoc_ds2` SET `countin` = `countin`+1 WHERE `id` = {$res2->id}";
            mysql_query($q1);
         }else
            $q2 = "INSERT INTO `assoc_ds2`(`ds1`, `ds2`, `countin`) VALUES ('{$value[0]}','{$value[1]}', '1')";
             mysql_query($q2);
         }
      }
   }
}
```

3. Представление и интерпретация полученных результатов

Решение задачи поиска ассоциативных правил, как и любой другой задачи сводится к обработке исходных данных и получению результатов. Обработка исходных данных выполняется по описанным выше алгоритмам. Результаты, получаемые при решении этой задачи, принято представлять в виде ассоциативных правил. В связи с этим в поиске выделяют два основных этапа:

1. Нахождение всех частых наборов (Алгоритм 2).
2. Генерация ассоциативных правил из найденных частых наборов.

Ассоциативные правила имеют следующий вид:

```
если (условие) то (результат)
```

где *условие* обычно не логическое выражение, а набор объектов из множества F, с которыми связаны (ассоциированы) объекты, включённые в *результат* данного правила. Например ассоциативное правило:

```
если I11.9(гипертония) то I20.8(стенокардия)
```

означает, что если у пациента выявлен диагноз I11.9(гипертония в кодировке МКБ10), то с определённой долей вероятности у него может быть выявлен диагноз I20.8(стенокардия).

Ассоциативное правило можно представить как импликацию $X \Rightarrow Y (X \in F, Y \in F, X \cup Y = \varphi)$

Основным достоинством ассоциативных правил является их лёгкое восприятие человеком и простая интерпретация языками программирования. Однако они не всегда полезны. Выделяют три вида правил:

полезные правила - содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение.

тривиальные правила - содержат действительную и легко объяснимую информацию, которая уже была известна.

непонятные правила - содержат информацию, которая не может быть объяснена. Такие правила могут быть получены на основе или аномальных значений, или глубоко скрытых знаний. Напрямую такие правила нельзя использовать для принятия решений, т.к. их необъяснимость может привести к непредстазуемым результатам. Для лучшего понимания требуется дополнительный анализ.

Количество ассоциативных правил может быть очень большим и трудно воспринимаемым для человека. К тому же не все из построенных правил несут в себе полезную информацию. Для оценки их полезности вводятся следующие величины:

Достоверность (confidence) – показывает вероятность того, что из наличия у пациента набора X следует наличие набора Y. Достоверностью правила $X \Rightarrow Y$ является отношение числа пациентов, содержащих наборы X и Y, к числу пациентов, содержащих набор X:

$$Conf_{X \Rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D_X|} = \frac{Supp_{X \cup Y}}{Supp_X} \tag{9}$$

К сожалению, достоверность не позволяет оценить полезность правила. Если процент наличия у пациентов набора Y при условии наличия в них набора X меньше, чем процент безусловного наличия набора Y, то вероятность случайно угадать наличие в транзакции набора Y больше, чем предсказать это с помощью правила $X \Rightarrow Y$. Для исправления такой ситуации вводится мера – *улучшение*.

Улучшение (improvement) – показывает полезнее ли правило, случайного угадывания. Улучшением правила является отношением числа пациентов, содержащих наборы X и Y к произведению количества пациентов, содержащих набор X, и количества пациентов, содержащих набор Y:

$$impr_{X \Rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D_X||D_Y|} = \frac{Supp_{X \cup Y}}{Supp_X \times Supp_Y} \tag{10}$$

Если улучшение больше единицы, то это значит, что с помощью правила предсказать наличие набора Y вероятнее чем, случайное угадывание, если меньше единицы, то наоборот. В последнем случае можно использовать отрицающее правило, т.е. правило, которое предсказывает отсутствие набора Y

Данные оценки используются при генерации правил. Аналитик при поиске ассоциативных правил задаёт минимальные значения перечисленных величин. В результате те правила, которые не удовлетворяют этим условиям, отбрасываются и не включаются в решение задачи.

Заключение

В данной статье показан процесс разработки инструментальных средств для интеллектуального анализа хранилища данных по предоставлению медицинской помощи гражданам на примере Ярославской области. Мехинизмы интелектуального анализа могут применяться как для оценки в целом объёмов помощи, так и для контроля. Интелллктуальный анализ предлагает ряд агрегированных величин для понимания тенденций происходящих в отрасли.

Список литературы

[1] Witten I.H., Frank E. Data Mining. Practical Machine Learning Tools and Techniques, Second Edition. Elsevier Inc., 2005

[2] Михайлов М.В. Инструментально-математические средства мониторинга страховой системы в режиме реального времени. Известия высших учебных заведений. Поволжский регион., №3(11), 2009, стр. 59 – 65.