СИСТЕМНЫЙ АНАЛИЗ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

УДК 004.9

ОПЕРАТИВНАЯ АНАЛИТИЧЕСКАЯ ОБРАБОТКА И ИССЛЕДОВАНИЕ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ ФАКТОРОВ ЗАБОЛЕВАЕМОСТИ НАСЕЛЕНИЯ

М. А. Артемов, Д. О. Кириченко, В. Г. Рудалев, Н. П. Сереженко

Воронежский государственный университет

Поступила в редакцию 20.09.2012 г.

Аннотация. Рассмотрена проблема построения программного комплекса для сбора, хранения и отображения данных о заболеваемости населения. Описана реализация программного комплекса, математические методы обработки, приводятся результаты применения этих методах на реальных данных.
Ключевые слова: OLAP, корреляционный анализ, медицинские данные.
Annotation. The problem of constructing a software system for collecting, storing and analyzing data of morbidity was investigated. The implementation of a software system and mathematical methods were described. Investigation of result was presented.
Keywords: OLAP, correlation analysis, medical data.

ВВЕДЕНИЕ

Перспективным направлением математических исследований в области здравоохранения представляется соединение современных методов оперативной аналитической обработки данных (OLAP) и методов статистического анализа, базирующееся на широком классе различных алгоритмов и покрывающее значительное число разных типов медицинских задач.

В статье рассмотрены общая схема работы программного комплекса для сбора и обработки данных медицинской статистики с уже существующих и работающих в лечебно-профилактических учреждениях информационных систем, а также некоторые методы их анализа, включающие метод исследования таблиц сопряженности. В частности, рассматривается метод сведения таблиц большой размерности к таблицам малой размерности 2×2, а также некоторые свойства такого сведения.

Структура программного комплекса

В настоящее время задача информатизации медико-социальных служб является одной из актуальных проблем современной IT-отрасли. В силу того, что в данной работе ставится задача сбора информации с использованием уже существующих в лечебно-профилактических учреждениях информационных систем, требуемый программный комплекс можно разделить на две подсистемы: подсистему сбора информации и подсистему обработки и анализа данных.

В задачу подсистемы сбора данных входит извлечение требуемой информации из оперативного источника, находящегося на неком клиентском хосте, и пересылка на сервер, где приходящие из сети данные подвергаются первичной обработке, очистке и лишь потом добавляются в базу данных (называемую хранилищем данных в терминах OLAP[4]).

В задачу подсистемы анализа данных входит извлечения требуемой аналитиком информации из построенного на основе хранилища данных куба данных, а также вычисление статистических или иных характеристик полученной таким образом выборки.

Более подробно требования изложены в [1].

На основании этих требований авторами был спроектирован и реализован программный комплекс, упрощенная схема которого представлена на рисунке 1.

СТРУКТУРА ХРАНИЛИЩА ДАННЫХ

Одним из ключевых элементов аналитической системы, базирующейся на технологии

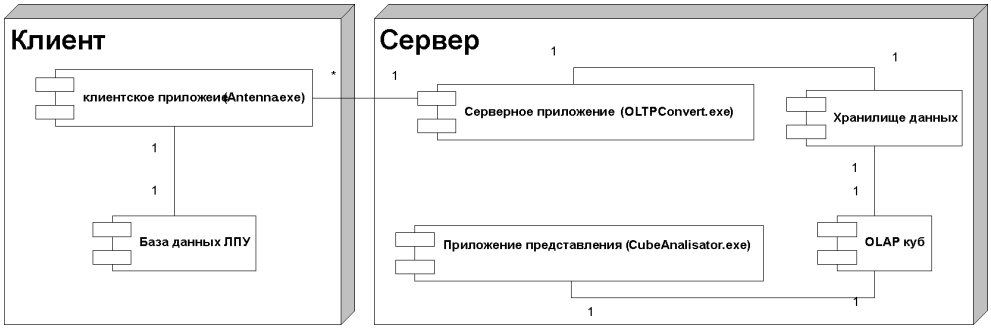

Рис. 1. Упрощенная схема работы программного комплекса

OLAP, является хранилище данных. Именно от его структуры, типов данных зависят и многомерный куб данных, который строится на основе хранилища, и измерения, по которым будет проводиться анализ.

В качестве базовой схемы, позволяющей не только разработку на абстрактном уровне (независимо от куба): проектирование, мониторинг, развитие математической модели, но и отражающей реальные и интересующие пользователя – специалиста в области медицинской статистики или медицинского менеджмента, и предоставляющей возможность на основе математических расчетов проводить исследование, была выбрана схема с шестью измерениями, отражающих наиболее общие и часто используемые в медицинском анализе факторы. В качестве таких измерений были взяты пол, место жительства пациента, возраст, дата обращения, установленный диагноз (на основе международной классификации болезней: МКБ-10), и лечебно-профилактическое учреждение, куда обратился пациент. Упрощенная схема представлена на рисунке 2.

МЕТОДЫ ОБРАБОТКИ

Технология OLAP предоставляет широкие возможности для исследования данных. Результат запроса к OLAP системе представляется собой n-мерную таблицу, в которой на каждом из измерений расположены интересующие аналитика значения атрибутов или совокупности таких значений (кортежи).

В силу того что, двухмерные таблицы наиболее наглядны для аналитика и математический аппарат для таблиц именно такой размерности наиболее развит, следует отталкиваться именно от двухмерных выборок.

Необходимо отметить, что классические методы статистического анализа не всегда применимы в связи со спецификой данных. В частности, зачастую они базируются на том факте, что значения атрибутов, имеют числовую форму (или хотя бы сводимую к ней, как например булевские значения). Но на практике данные не обязательно имеют подобную форму (например, населенный пункт проживания). Поэтому было принято решение использования методов анализа категоризованных данных [3].

Двухмерную таблицу (таблицу сопряженности) можно представить в виде (1):

$$\begin{array}{cccc|c} n_{1,1} & n_{1,2} & . & n_{1,c} & N_1 \\ n_{2,1} & n_{2,2} & . & n_{2,c} & N_2 \\ . & . & . & . & . \\ n_{r,1} & n_{r,2} & . & n_{r,c} & N_r \\ \hline M_1 & M_2 & . & M_c & n \end{array}, \qquad (1)$$

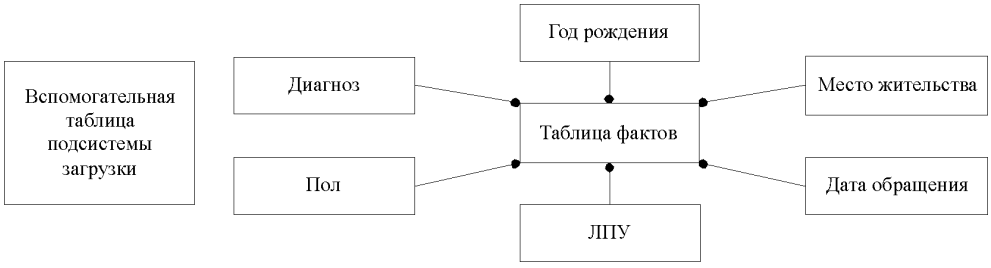

Рис. 2. Упрощенная схема хранилища данных

где r и c размеры таблицы: r – число строк, c – чисто столбцов.

Введены следующие обозначения: $n_{i,j}$ – число сочетаний i-го атрибута одного измерения и j-го атрибута другого, N_i – сумма значений с i-й строки, M_j – сумма значений j-го столбца, n – сумма всех значений. Для дальнейшего вычисления меры связи между атрибутами стоит ввести следующий коэффициент, являющийся оценкой хи-квадрат [3]:

$$X^2 = n\left(\sum_{i,j}\frac{n_{i,j}}{N_i M_j} - 1\right). \qquad (2)$$

Для определения степени зависимости в таблице на практике используют несколько коэффициентов: коэффициент Пирсона (2), коэффициент Чупрова (3) и коэффициент Крамера (4).

$$P = \sqrt{\frac{X^2}{X^2 + n}}; \qquad (3)$$

$$T = \sqrt{\frac{X^2}{n\sqrt{(r-1)(c-1)}}}; \qquad (4)$$

$$\mathrm{C} = \sqrt{\frac{X^2}{n\min(r-1, c-1)}}. \qquad (5)$$

Все эти коэффициенты имеют свои достоинства и недостатки [2], однако все они показывают лишь общую степень взаимосвязи атрибутов (чем ближе к единице, тем связь сильнее, чем ближе к нулю, тем слабее), ничего не говоря о взаимосвязи отдельных атрибутов. Для преодоления этих ограничений в статье предлагается методика преобразования исходной таблицы размерностью $r \times c$ к $r \times c$ таблицам размерностью 2×2.

В процессе преобразования фиксируется пара индексов (i, j), где i – атрибут первого измерения, j – второго измерения. Для такой фиксированной пары индексов таблица 2×2 будет иметь вид (6).

	B	$\overline{B}$	
A	$n_{i,j}$	$N_i - n_{i,j}$	N_i
$\overline{A}$	$M_j - n_{i,j}$	$n - M_j - N_i + n_{i,j}$	$n - N_i$
	M_j	$n - M_j$	n

. (6)

Таким образом, эта таблица является таблицей соотношения i-го атрибута первого измерения и j-го атрибута второго и совокупности всех остальных атрибутов обоих измерений.

$$p = n_{i,j},\ q = N_i - n_{i,j},\ t = M_j - n_{i,j},\ s = n - M_j - N_i + n_{i,j} \qquad (7)$$

После применения подстановки (7), и обозначения i-го атрибута первого измерения как A, а j-го атрибута другого измерения как B то (6) преобразуется к виду (8).

	B	$\overline{B}$	
A	p	q	$p+q$
$\overline{A}$	t	s	$t+s$
	$p+t$	$q+s$	$p+q+t+s$

. (8)

Вид (8) является классическим видом для таблицы 2×2. Для таблиц такой размерности принято использовать коэффициенты ассоциации (9) и контингенции (10) для оценки силы взаимосвязи между атрибутами.

$$Q = \frac{ps - qt}{ps + qt}; \qquad (9)$$

$$V = \frac{ps - qt}{\sqrt{(p+t)(p+q)(s+t)(q+s)}}. \qquad (10)$$

Следует заметить, что после применения подстановки (7) в формулы (9) и (10), они преобразуются к виду (11) и (12):

$$Q = \frac{n_{i,j}n - N_i M_j}{n_{i,j}(n - M_j - N_i + n_{i,j}) + (N_i - n_{i,j})(M_j - n_{i,j})}, \qquad (11)$$

$$V = \frac{n_{i,j}n - N_i M_j}{\sqrt{N_i M_j (n - M_j)(n - N_i)}}. \qquad (12)$$

В дальнейшем обозначение $Q_{i,j}$ и $V_{i,j}$ соответствуют коэффициентам ассоциации и контингенции, вычисленным для зафиксированной пары атрибутов (i, j) после сведения исходной таблицы к таблице 2×2.

Таким образом, задача анализа таблицы сопряженности сводится к задаче анализа таблиц коэффициентов ассоциации и контингенции.

Следует заметить, что для корректного анализа необходимо, чтобы значения атрибутов в выборке принимали всевозможные допустимые значения без ограничений, либо чтобы «ненужные» значения были просуммированы в один атрибут. Действительно, такое преобразование не повлияет на значения коэффициентов ассоциации и контингенции, построенные для другой пары индексов. Значения «ненужных» атрибутов, так или иначе, войдут в суммируемые

части. Под «ненужными» значениями атрибутов, понимаются такие значения, статистические связи в которых не интересуют аналитика.

Также можно показать, что для построенных таким образом коэффициентов справедливы следующие свойства:

1) Если $Q_{i,j} \geq \alpha$ и $Q_{i,k} \geq \alpha$, то $Q_{i,(j,k)} \geq \alpha$, где α – произвольное число, а $Q_{i,(j,k)}$ – коэффициент ассоциации двух объединяемых атрибутов.

2) Если $V_{i,k} \geq \alpha$ и $V_{i,j} \geq \alpha$, то $V_{i,(j,k)} \geq \alpha$, где α – произвольное число, а $V_{i,(j,k)}$ – коэффициент контингенции двух объединяемых атрибутов.

Пусть по условию:

$$Q_{i,j} = \frac{n_{i,j}n - N_i M_j}{n_{i,j}(n - M_j - N_i + n_{i,j}) + (N_i - n_{i,j})(M_j - n_{i,j})} \geq \alpha, \quad (13)$$

$$Q_{i,k} = \frac{n_{i,k}n - N_i M_k}{n_{i,k}(n - M_k - N_i + n_{i,k}) + (N_i - n_{i,k})(M_k - n_{i,k})} \geq \alpha. \quad (14)$$

Для удобства следует ввести следующие обозначения:

$$A_{i,j} = n_{i,j}n - N_i M_j \quad (15)$$

$$A_{i,k} = n_{i,k}n - N_i M_k \quad (16)$$

$$B_{i,j} = n_{i,j}(n - M_j - N_i + n_{i,j}) + (N_i - n_{i,j})(M_j - n_{i,j}), \quad (17)$$

$$B_{i,k} = n_{i,k}(n - M_k - N_i + n_{i,k}) + (N_i - n_{i,k})(M_k - n_{i,k}). \quad (18)$$

После подстановки (15) – (18) в (13) и (14) получится:

$$\frac{A_{i,j}}{B_{i,j}} \geq \alpha, \quad (19)$$

$$\frac{A_{i,k}}{B_{i,k}} \geq \alpha. \quad (20)$$

Коэффициент ассоциации в объединенной ячейке будет равен:

$$Q_{i,(j,k)} = \frac{n(n_{i,j} + n_{i,k}) - N_i(M_j + M_k)}{(n_{i,j} + n_{i,k})(n - M_j - M_k - N_i + n_{i,j} + n_{i,k}) + (N_i - n_{i,j} - n_{i,k})(M_j + M_k - n_{i,j} - n_{i,k})} \quad (21)$$

Путем несложных преобразований получается:

$$Q_{i,(j,k)} = \frac{A_{i,j} + A_{i,k}}{B_{i,j} + B_{i,k} + C}, \quad (21)$$

где

$$C = 2n_{i,j}(-M_k + n_{i,k}) + 2n_{i,k}(-M_j + n_{i,j}). \quad (22)$$

Так как $n_{i,k} \geq 0, n_{i,j} \geq 0, n_{i,k} \leq M_k, n_{i,j} \leq M_j$, то $C \leq 0$.

Таким образом:

$$\begin{aligned} Q_{i,(j,k)} &= \frac{A_{i,j} + A_{i,k}}{B_{i,j} + B_{i,k} + C} = \\ &= \frac{A_{i,j}}{B_{i,j}}\left(\frac{B_{i,j}}{B_{i,j} + B_{i,k} + C}\right) + \\ &+ \frac{A_{i,k}}{B_{i,k}}\left(\frac{B_{i,k}}{B_{i,j} + B_{i,k} + C}\right) \geq \\ &\geq \alpha\left(\frac{B_{i,j}}{B_{i,j} + B_{i,k} + C}\right) + \alpha\left(\frac{B_{i,k}}{B_{i,j} + B_{i,k} + C}\right) = \\ &= \alpha\left(\frac{B_{i,j} + B_{i,k}}{B_{i,j} + B_{i,k} + C}\right). \end{aligned} \quad (23)$$

Так как

$$B_{i,j} \geq 0, B_{i,k} \geq 0, C \leq 0, B_{i,j} + B_{i,k} + C > 0,$$

то

$$\frac{B_{i,j} + B_{i,k}}{B_{i,j} + B_{i,k} + C} \geq 1. \quad (24)$$

А значит:

$$Q_{i,(j,k)} \geq \alpha. \quad (25)$$

Свойство 2 доказывается аналогично.

Из этих свойств вытекают следующие следствия:

1) $Q_{i,(j,k)} \geq \min\{Q_{i,k}, Q_{i,j}\}$
2) $V_{i,(j,k)} \geq \min\{V_{i,k}, V_{i,j}\}$

ПРИМЕР ИСПОЛЬЗОВАНИЯ

В качестве примера использования программного комплекса представляет интерес исследование взаимосвязи возрастных категорий людей, обращающихся в медицинские учреждения за определенный период времени (например, 2007 год), и населенным пунктом, в котором они проживают. В целях обеспечения конфиденциальности истинные названия населенных пунктов не указываются. Малые и «заброшенные» населенные пункты не учитывались. Исходная выборка представлена в таблице 1.

Таблица 1

Исходная выборка данных

Год рождения	Нас. пункт 1	Нас. пункт 2	Нас. пункт 3	Нас. пункт 4	Нас. пункт 5
1910, 1915, 1920	69	0	8	8	1
1925	251	3	19	10	11
1930	408	19	23	48	50
1935	422	25	19	25	58
1940	592	26	79	66	34
1945	355	22	22	31	30
1950	780	23	64	48	37
1955	834	25	45	19	58
1960	915	21	34	37	72
1965	370	31	36	19	21
1970	439	15	28	56	41
1975	656	35	78	45	66
1980	623	1	100	9	61
1985	581	22	91	25	25
1990	920	77	71	51	101
1995	298	44	27	12	35
2000	758	20	15	46	60
2005	505	12	23	7	10

Промежуток в 5 лет выбран для разграничения людей определенных возрастных категорий.

Построенные по таблице 1 коэффициенты Пирсона, Крамера и Чупрова соответственно равны:

$P = 0.2248$;
$T = 0.0781$;
$C = 0.1153$.

Их значения говорят о существовании слабой общей связи между возрастом и местом проживания пациентов. Далее проводится преобразование исходной таблицы к $80 = 5*18$ таблицам 2×2. Затем по методике (6), (11) по полученным таблицам производится расчет коэффициентов ассоциации, и их значения записываются в следующую таблицу (таблица 2).

Отметим, что упомянутые операции выполняются программным комплексом автоматически.

При анализе таблицы можно обратить внимание, что для населенного пункта 2 коэффициенты ассоциации, соответствующие людям подросткового возраста 17 и 12 лет (1990, 1995 года рождения) указывают на связь средней силы между этими двумя атрибутами. Это позволяет в соответствии с доказанными выше свойствами объединить эти два значения атрибута в один. Коэффициент ассоциации для объединенного значения, вычисленного по формуле (21) будет равен 0.49983, что подтверждает наличие статистической взаимосвязи средней силы. Это можно объяснить, в частности, наличием неблагоприятных факторов для людей этой возрастной группы.

ЗАКЛЮЧЕНИЕ

В статье был рассмотрены математические методы обработки данных, использованные при реализации, разработанного авторами програм-

Таблица 2

Таблица коэффициентов ассоциации

Год рождения	Нас. пункт 1	Нас. пункт 2	Нас. пункт 3	Нас. пункт 4	Нас. пункт 5
1910, 1915, 1920	0.3715	-1	0.2054	0.3672	-0.7021
1925	0.2908	-0.5551	-0.0095	0.0844	-0.5733
1930	-0.1459	0.0075	-0.2231	0.3551	0.2119
1935	-0.0777	0.1559	-0.3185	-0.0013	0.2935
1940	-0.1541	-0.026	0.2571	0.3346	-0.2109
1945	-0.0681	0.1809	0.033	0.0568	-0.2615
1950	0.0875	-0.1891	-0.154	0.2128	0.0225
1955	0.205	-0.1614	-0.1827	-0.4364	-0.0332
1960	0.1979	-0.3008	-0.3736	-0.1599	0.0373
1965	-0.0565	0.3421	0.0964	-0.0737	-0.19
1970	-0.1082	-0.1481	-0.1495	0.4075	0.0692
1975	-0.1478	0.0848	0.1942	0.0645	0.1046
1980	-0.0302	-0.9355	0.392	-0.6296	0.1177
1985	-0.0417	-0.0794	0.3737	-0.1667	-0.3294
1990	-0.1272	0.3558	-0.0514	-0.0508	0.1681
1995	-0.2163	0.5665	0.012	-0.2401	0.1642
2000	0.1762	-0.2309	-0.6187	0.0649	0.0368
2005	0.4448	-0.2416	-0.2315	-0.5913	-0.5821

много комплекса анализа данных медицинской статистики. Предложен, подробно разобран и проиллюстрирован на практическом примере метод сведения таблиц размерности $r \times c$ к таблицам размерности 2×2.

СПИСОК ЛИТЕРАТУРЫ

1. *Есауленко И. Э.* Проектирование и программная реализация комплекса анализа данных медико-статистической информации / И. Э. Есауленко, М. А. Артемов, Д. О. Кириченко, В. Г. Рудалев, Н. П. Сереженко // Системный анализ и управление в биомедицинских системах. – 2012. – Том 11, № 2. – С. 301–305.

2. *Кендалл М.* Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М. : Издательство «Наука», 1973. – 900 с.

3. *Аптон. Г.* Анализ таблиц сопряженности / Г. Аптон. – М. : Финансы и статистика, 1982. – 143 с.

4. *Барсегян А. А.* Методы и модели анализа данных OLAP и Data mining. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. — СПб.: БХВ-Петербург, 2004. — 336 с.

Артемов Михаил Анатольевич – доктор физико-математических наук, профессор, заведующий кафедрой программного обеспечения и администрирования информационных систем ВГУ. E-mail: artemov_m_a@amm.main.vsu.ru

Artemov Mikhail A. – Head of Department of Software and Information System Administering, Voronezh State University. Sciences, Professor. E-mail: artemov_v_a@mail.ru

Кириченко Денис Олегович – аспирант факультета ПММ ВГУ. Направление научных исследований – оперативная аналитическая обработка данных. E-mail: Yalo55@yandex.ru

Рудалев Валерий Геннадьевич, кандидат физико-математических наук, доцент кафедры технической кибернетики и автоматического регулирования ф-та ПММ ВГУ. Направление научных исследований – стохастические системы в управлении и связи. E-mail: rud_wl@mail.ru

Сереженко Николай Петрович, кандидат медицинских наук, доцент кафедры нормальной анатомии человека ВГМА им. Н. Н. Бурденко. Направление научных исследований – методы математической обработки медико-биологической информации. E-mail: nps-med@rambler.ru

Kirichenko D. O. – Post-graduate student of Department of Applied Mathematics, Mechanics and Informatics, Voronezh State University.

Rudalev V. G. – Candidat of Physics-math. Sciences, Associate Professor, the dept. of the Technical Cybernetics and Automatic Control, Voronezh State University. E-mail: rud_wl@mail.ru

Serezhenko N. P. – Candidate of Medical Sciences, Associate Professor, the Department of Normal Human Anatomy VGMA them. Burdenko. E-mail: nps-med@rambler.ru