UDC 621.391

A.O. Feklistov

Kharkiv Air Force University, Kharkiv

THE METHOD OF AUTOMATIC GENERATION OF HYPOTHESES WITH A MULTIDIMENSIONAL ASSOCIATION'S QUANTIFIER FOR DECISION SUPPORT SYSTEM

The article considers an approach of automatic generation of statistical hypotheses with multidimensional association's quantifiers in a form of logical expressions based on attributed binary tree.

Keywords: *decision support system, statistical hypothesis, quantifiers of multidimensional associations.*

Introduction

Formulation of the problem in general

One of the essential components of decision support systems are different kind of knowledge and data mining subsystems. And, one of them are subsystems aimed to discovery associations rules in a form of logical expressions of statistical hypotheses. The methods of automatic generation of multidimensional association relations between object's parameters are unknow now unlike of well-know methods of discovery multidimensional correlations.

The purpose of the article is to define an approach of automatic generation of statistical hypotheses with a multidimensional association's quantifiers in a form of logical expressions based on attributed binary tree. It suggests be useful for practical decision support systems.

Statement of Materials Research

The correlations methods are most widely used to discovery a dependencies between objects' parameters. [1 – 3]. The association rules methods are presented in a many works. Some of them have been interesting for statistical hypothesis with an association's quantifiers. For hypotheses with a quantifiers of binary associations were developed works [4, 5]. The calculation of quantitive measure of association similarity for data set with missing values was described in works [6 – 9]. And, there are other approaches for association rules algorithms [10 – 12].

The one well-known approach in data mining is generation of logical expressions for statistical hypotheses with a associations quantifiers [4, 5].

The term "binary association" means a relationship between two parameters x_i and x_j $(i \neq j)$ in which the "coincidence is stronger than differences". It means that the number of observation's cases when set of parameters have got or haven't got some kind of values simutaneously are bigger than the rest of other cases.

Object's parameters are presented as a set $X=\{x_1,\ldots,x_{n_x}\}$. Each element of this set has an appropriate function of unary predicate $\varphi(\tau)=\{\varphi_1(\tau),\ldots,\varphi_n(\tau)\}$. This function has three logical values: 1 – "true", 0 – "false" and "x" - "undefined".

The logical value is written before the predicate in parentheses. For example, the $(1)\varphi(\tau)$ indicates that unary predicate $\varphi(\tau)$ has a logical value "true". The variable "τ" is an index of observation case. Let's introduce a function $|\{M\}|$, which determines the cardinality of set $\{M\}$.

The coefficients that describe a number of observations where predicates have a certain values are calculated as follows:

$$a_{11}=|\{\tau\,|(1)\varphi_i(\tau)\wedge(1)\varphi_j(\tau)\}|;$$

$$a_{10}=|\{\tau\,|(1)\varphi_i(\tau)\wedge(0)\varphi_j(\tau)\}|;$$

$$a_{01}=|\{\tau\,|(0)\varphi_i(\tau)\wedge(1)\varphi_j(\tau)\}|;$$

$$a_{00}=|\{\tau\,|(0)\varphi_i(\tau)\wedge(0)\varphi_j(\tau)\}|.$$

The value of "reverse contingency" r, which is calculated as follows: $r=\dfrac{p_{11}\cdot p_{00}}{p_{10}\cdot p_{01}}$, where $p_{ij}=a_{ij}/m$ – the probability of formula execution $(c)\varphi_i(\tau)\ \wedge\ (d)\varphi_j(\tau)$, $c, d \in \{0, 1\}$, $i \neq j$; m – the number of observations [4].

It is considers that every reasonable measure of dependence is a strictly monotonic function r. At the same time, for the logarithmic contingency δ, then the $\delta>0$ - positive relationship, at $\delta<0$ - negative relationship, and at $\delta=0$ - no relationship [5].

A quantitative measure for the quantifier of binary association $\gamma_{\approx_2}$ defines as the reciprocal value of r, as follows [6]:

$$\gamma_{\approx_2}=\frac{1}{r}=\frac{p_{10}\cdot p_{01}}{p_{11}\cdot p_{00}}=\frac{a_{10}\cdot a_{01}}{a_{11}\cdot a_{00}}.$$

The value of $\gamma_{\approx_2} \in [0..+\infty]$ and not define in a two cases: numerator and denominator are equal ($\gamma_{\approx_2}=1$); the calculation of numerator or denominator is impossible because one of the coefficients is equal to zero.

The hypothesis with quantifier of binary association has a following form: $\approx_{\gamma\approx_n}(\varphi_i(\tau),\varphi_j(\tau))$. The hypothesis with a quantifier of multidimensional association has a following form: $\approx_{\gamma\approx_n}(\varphi_1(\tau),\ \varphi_2(\tau),\ \ldots,\varphi_n(\tau))$. It describes an relationships of association between more that two parameters $(x_1, x_2, \ldots, x_n)$ in a same way as associations between two parameters.

The calculation of quantitive measure of association for many parameters might be implemented by generation a hypotheses with a quantifier of multidimensional association $\gamma_{\approx_n}$. The coefficients of associations $\{a\}$ are presented as a nodes of binary tree with attributes (fig. 1).

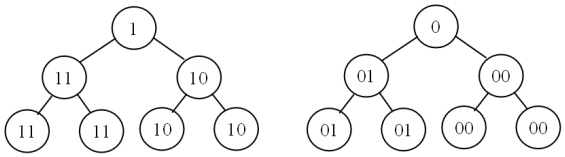

Fig. 1. A presentation of coefficients for calculation quantifier with a multidimensional association as nodes of binary tree with attributes

Each node of this tree has an attribute with a name "code" which is a sequence of zeros and ones that correspond to logical values of unary predicates. For example, the node with a code "1101" has an abbreviation a_{1101}. The code of tree node is used to make a classification of tree nodes on two equal sets: set $\{a^+\}$ which is characterise an association similarities of parameters; set $\{a^-\}$ which is characterise their association differencies.

The quantitive measure of quantifier of multidimensional association $\gamma_{\approx_n}$ is defined as:

$$\gamma_{\approx_n}=\prod_{i=1}^{n^-}a_i^-\Big/\prod_{j=1}^{n^+}a_j^+,$$

where $n=|\{a\}|$ – the total number of tree nodes; $n^+=|\{a^+\}|$ – the number of tree nodes which characterised an association similarities; $n^-=|\{a^-\}|$ – the number of tree nodes which characterised an association differencies; $\{a\}=\{a^+\}\cup\{a^-\}$; $n=n^++n^-$; $n^+=n^-=1/2\cdot n$.

The analysis of these formulas shows that it's sufficient for classification of tree nodes to have one of set $\{a^+\}$ or $\{a^-\}$. Let's consider a process of processing a set $\{a^+\}$ bear in mind that $\{a^-\}=\{a\}\setminus\{a^+\}$. The determination of set $\{a^+\}$ is based on the introduction of a set of attributes for the node and related coefficients.

The combining of all the nodes are presented as follows:

$$\{a\}=\{a_{=1}\}\cup\{a_{=0}\}\cup\{a_{>1}\}\cup\{a_{>0}\}\cup\{a_{0=1}\}.$$

The coefficient n is equal:

$$n=n_{=1}+n_{=0}+n_{>1}+n_{>0}+n_{0=1}.$$

Let's denote a set of nodes as follows $\{a_{=1}\}$, $\{a_{=0}\}$, $\{a_{>1}\}$, $\{a_{>0}\}$ and, in general as $\{a_\Sigma\}$:

$$\{a_\Sigma\}=\{a_{=1}\}\cup\{a_{=0}\}\cup\{a_{>1}\}\cup\{a_{>0}\}.$$

The cardinality of set $\{a_\Sigma\}$ is equal:

$$n_\Sigma=n_{=1}+n_{=0}+n_{>1}+n_{>0}.$$

The classification of tree nodes is based on the analysis of two conditions: $n_\Sigma = 1/2\,n$ or $n_\Sigma > 1/2\,n$. If coefficient $n_\Sigma = 1/2\,n$, then the nodes of set $\{a_{=1}\}$, $\{a_{=0}\}$, $\{a_{>1}\}$, $\{a_{>0}\}$ are nodes of association similarities, and $\{a_{0=1}\}$ – nodes of association differencies:

$$\{a^+\}=\{a_{=1}, a_{=0}, a_{>1}, a_{>0}\};\ \{a^-\}=\{a_{0=1}\}.$$

If coefficient $n_\Sigma > 1/2\,n$, then the following operations are performed.

Let's combine sets $\{a_{>1}\}$ and $\{a_{>0}\}$ as one set $\{a_?\}$: $\{a_?\}=\{a_{>1}\}\cup\{a_{>0}\}$. Let's select from set $\{a_?\}$ all nodes that are characterised association similarities. The rest of nodes will be classified as a nodes that characterised association differencies.

Thus, selected nodes $\{a^+\}$ are nodes with the highest values of the coefficients. Let's define them as nodes of "possible association similarities" $\{a_?^+\}$. The number of these nodes is defined as follows: $n_?^+=|\{a_?^+\}|=1/2\,a - n_{=1} - n_{=0}$. The rest of nodes let's define as a nodes of "possible association differencies" $\{a_?^-\}$. The number of these nodes is defined as follows: $n_?^-=|\{a_?^-\}|=1/2\cdot a$.

All nodes $\{a_{0=1}\}$ are also belong to the set $\{a^-\}$:

$$\{a^+\}=\{a_{=1}^+, a_{=0}^+, a_?^+\};\ \{a^-\}=\{a_{0=1}, a_?^-\}.$$

It worth to note that there are nodes among the nodes $a_{>1}$ and $a_{>0}$ with a code of 5 or more elements which are more suitable for the set $\{a^+\}$ than for $\{a^-\}$. For example, a node a_{11101} has more ones that node a_{11100} and, hence, this node is more appropriate for "association similarities".

So, let's introduce an additional constraint on calculation of quantitive measures of association for coefficients with arity 5 and more at the time of selection coefficients to the $a_?^+$. Let's define a coefficients Δ_1 and Δ_0 as follows.

A coefficient Δ_1 characterizes a difference between the number of ones and the average number of signs (1 and 0) in the code of node: $\Delta_1=\text{abs}((0,5c(a) -$

$c_1(a)$), where abs(x) – a function of absolute value of x, $c(a)$ – the number of digits in the code of node, $c_1(a)$ – the number of ones in the code of node.

A coefficient Δ_0 characterizes a difference between the number of zeros and the average number of characters in the code of node: Δ_0=abs[($0,5c(a)$ – $c_0(a)$], where $c_0(a)$ – the number of zeros in the code of node. It is possible to use a total sign Δ as coefficients are equal ($\Delta_1 = \Delta_0$).

Let's consider coefficient a_{11101} and values of coefficients Δ: $c(a) = c(a_{11101}) = 5$; $c_1(a) = 4$; $c_0(a) = 1$; Δ_1 = abs(5/2 – 4)= 1,5; Δ_0 =abs(5/2-1)=1,5. The usage of coefficient Δ allows to clarifies the designation of nodes $a_{>1}$ and $a_{>0}$ as follows: $a^{\Delta}_{>1}$ and $a^{\Delta}_{>0}$. It is suggested that the higher value of Δ cause the higher quantitive measure of association similarity between parameters. Thus, the nodes with higher values of Δ should be selected to the set $a^{+}_{?}$.

Let's designate a minimal value of Δ as a $\Delta^{\min}$. For the quantifier of association with arity 4 the coefficient Δ will be Δ={0, 1}; for the quantifier of association with arity 5 it will be Δ={0,5; 1,5}. To account a difference in a number of ones and zeros in the code of nodes it will be useful to select coefficients $\Delta^{*} > \Delta^{\min}$: $\{a^{+}_{?}\} = \{a^{\Delta^{*}}_{>1}\} \cup \{a^{\Delta^{*}}_{>0}\}$.

Thus, the calculation of coefficients of the set { a } with the help of a binary tree attributes allows to numerically estimate the measure of association relationships between set of parameters for a given object.

Conclusions

The article considers an approach of automatic generation of statistical hypotheses with multidimensional association's quantifiers in a form of logical expressions based on attributed binary tree. The proposed approach of using a binary tree with attributes to automatically generation of associative relationships between object parameters was presented for the first time.

The given results are the basis for further researches in a way of development of mathematical and software of intelligent decision support systems.

References

1. Корн Г. Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. – М.: Наука, 1970. – 720 с.

2. Справочник. Математический энциклопедический словарь. – М.: Сов. энциклопедия, 1988. – 847 с.

3. Тернер Д. Вероятность, статистика и исследование операций / Д. Тернер. – М.: Статистика, 1976. – 431 с.

4. Edwards A.W.F. The measure of association in 2x2 table / A.W.F. Edwards // Journal of the Royal Statistical Society, ser.A29. – P. 109-114.

5. Гаек П. Автоматическое образование гипотез: математические основы общей теории / П. Гаек, Т. Гавранек. – М.: Наука, Главная редакция физико-математической литературы, 1984. – 280 с.

6. Феклистов А.А. Автоматическое образование нечетких обобщений на основе квантора простой ассоциации / А.А. Феклистов // Сборник статей «Информационные системы». – Х: НАНУ, ХВУ, 1995. – С. 89-92.

7. Феклистов А.А. Метод нечетких обобщений для систем поддержки принятия решений / А.А. Феклистов // Інформатизація та нові технології. – 1996. – №2. – С. 6-8.

8. Феклистов А.А. Метод определения коэффициентов логической неопределенности в системах поддержки принятия решений / А.А. Феклистов // Зб. наук. пр. Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова. – К.: ІПМЕ, 2003. – Вип. 22. – С. 232-237.

9. Феклістов А.О. Метод формалізації операторів логічної невизначеності в інтелектуальних системах підтримки прийняття рішень / А.О. Феклістов, О.Я. Лазарєва, О.О. Феклістов // Системи обробки інформації. – Х.: ХВУ, 2003. – Вип. 6. – С. 22-26.

10. Amir A. A new and versatile method for association generation / A. Amir, R. Feldman, R. Kashi // Information systems. – 1997. – Vol.22, No.6/7. – P. 333-347.

11. Ситников Д.Э. Метод поиска обобщенных ассоциативных зависимостей между дискретными признаками / Д.Э. Ситников, Е.В. Титова // Системи обробки інформації. – Х.: ХВУ, 2002. – Вип. 6. – С. 194-202.

12. Титова Е.В. Сравнительная характеристика простых и расширенных ассоциативных правил для признаков объектов в базах данных / Е.В. Титова // Системи обробки інформації. – Х.: ХВУ, 2003. – Вип. 2. – С. 31-37.

Надійшла до редколегії 7.07.2015

Рецензент: д-р техн. наук, проф. О.М. Сотніков, Харківський університет Повітряних Сил ім. І. Кожедуба, Харків.

МЕТОД АВТОМАТИЧНОГО УТВОРЕННЯ ГІПОТЕЗ ІЗ КВАНТОРОМ БАГАТОВИМІРНОЇ АСОЦІАЦІЇ ДЛЯ СИСТЕМ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

А.О. Феклістов

В статті розглядається метод автоматичного утворення статистичних гіпотез із кванторами багатовимірної асоціації у формі логічних виразів, який використовує атрібутовані бінарні дерева.

Ключові слова: *система прийняття рішень, статистична гіпотеза, квантор багатовимірної асоціації.*

МЕТОД АВТОМАТИЧЕСКОГО ОБРАЗОВАНИЯ ГИПОТЕЗ С КВАНТОРОМ МНОГОМЕРНОЙ АССОЦИАЦИИ ДЛЯ СИСТЕМ ПРИНЯТИЯ РЕШЕНИЙ

А.А. Феклистов

В статье рассматривается метод автоматического образования статистических гипотез с кванторами многомерной ассоциации в форме логических выражений, который использует атрибутированные бинарные деревья.

Ключевые слова*: система принятия решений, статистическая гипотеза, квантор многомерной ассоциации.*