Proc. Estonian Acad. Sci. Engin., 1995, 1, 2, 105–112
https://doi.org/10.3176/eng.1995.2.01

SOME PROBLEMS OF APPROXIMATING SYMMETRIC RELATIONS BY EQUIVALENCE RELATION

Guido VEINER

Tallinna Tehnikaülikooli Informaatikainstituut (Department of Informatics, Tallinn Technical University), Raja 15, EE-0026 Tallinn, Eesti (Estonia)

Presented by L. Mõtus

Received 24 October 1995, accepted 31 January 1996

Abstract. Some auxiliary functions are used to approximate the symmetric binary relation by the equivalence relation. The condition of approximation means minimising the number of non-coincidence edges.

Key words: binary relation, equivalence relation, approximation of binary relation, graph.

1. INTRODUCTION

Recent studies of the author [1] deal with the approximation of the symmetric reflexive binary relation R by the equivalence relation E to minimise the number of elements of the set

$$\{E-R\}\cup\{R-E\}. \tag{1}$$

On finding the relation E, the properties of reflexivity and symmetry are trivials and therefore we consider here only some problems with the property of transitivity. In the author's opinion, the notions of the graph theory can be used. The problem is solved for a special class of relations by three auxiliary functions.

2. ESSENTIAL NOTIONS AND DEFINITIONS

Let $G(V, U)$ be the graph of the binary relation R. The sets V and U are the sets of vertices and edges, respectively.

Definition 1. *The subgraph $G_c\,(V_c, U_c)$ of the graph $G(V,U)$ is a connection of the subgraphs $G_1\,(V_1, U_1)$, $G_2\,(V_2, U_2)$,..., $G_p\,(V_p, U_p)$ if:*

$$1)\ V_c = \bigcup_{i=1}^{p} V_i,$$

$$2)\ U_c = \bigcup_{i=1}^{p} U_i.$$

Assume now that a graph $G\,(V,U)$ contains the subgraph $G_a(V_a,U_a)$, which has the following structure:

1) the subgraph $G_a\,(V_a,U_a)$ contains a maximal complete subgraph $G_0\,(V_0,U_0)$;

2) the subgraph $G_0(V_0,U_0)$ intersects the maximal complete subgraphs of the graph $G\,(V,U)$.

We note here that if a connected subgraph of the graph G contains more than two vertices, then its structure can be considered as a connection of subgraphs, which corresponds to Definition 1.

Definition 2. *The connection of maximally complete subgraphs* $G_1(V_1,U_1), G_2(V_2,U_2),\ldots,G_i(V_i,U_i)$ *of the graph* $G\,(V,U)$ *is a cluster of the subgraph* $G_a(V_a,U_a)$ *if*

1) *the connection of subgraphs* $G_1, G_2,\ldots,G_i$ *is connected,*

2) $\forall_j \in \{1,2,\ldots,i\} \rightarrow V_j \cap V_0 \neq \varnothing$,

3) $\forall_j \in \{1,2,\ldots,i\} \rightarrow \{V_j - V_0\} \neq \varnothing$,

4) *there is no such maximal complete subgraph* G_q *that for the connection of subgraph* $G_1, G_2,\ldots,G_i, G_q$ *the conditions* 1, 2 *and* 3 *hold true.*

Let us denote now the numbers of elements of some essential sets such that

1) $|V_0| = v_0$,

2) $|V_i \cap V_0| = v_i$, if $G_i\,(V_i,U_i)$ is a cluster,

3) $|V_i - V_0| = t_i$, if $G_i\,(V_i,U_i)$ is a cluster too,

4) $|\{(V_i - V_0) \text{ x } (V_i \cap V_0)\} \cap U| = L_i$, it is the number of edges of the graph G, which connect the vertices of the sets $V_i - V_0$ and $V_i \cap V_0$,

5) $\sum_{i \in I} t_i = S$, if $I = \{1, 2,\ldots, n, v\}$ is the set of indices of the clusters.

It is easy to see that the edges of each cluster G_i and the edges of the U_0 form the set of non-transitive pairs of the edges. The number of such pairs is $t_i\,(v_0 - v_i)$. Consequently, we have exactly three possibilities for correcting this situation:

1) remove L_i edges of the cluster G_i;

2) remove $v_i\,(v_0 - v_i)$ edges of the subgraph G_0;

3) by new edges, connect the vertices of the sets $V_i - V_0$ and $V_0 - V_i$. We need $t_i\,(v_0 - v_i)$ edges for connection.

Definition 3. *Let* $i \in I$ *and*

$$f_1(G_i) = (v_0 - v_i)/v_i, \tag{2}$$

$$f_1(G_{0i}) = [t_i(\nu_0 - \nu_i) + \nu_i(S - t_i)]/[\nu_i(\nu_0 - \nu_i)], \tag{3}$$

$$f_2(G_i) = f_1(G_i)/f_1(G_{0i}), \tag{4}$$

be three auxiliary functions of the cluster $G_i\,(V_i,U_i)$.

The non-transitive pair of edges is called a fork. We can see that the second addend of the numerator is also a number of forks. Such forks are located between the vertices of $V_0 \cap V_i$ and the vertices of other clusters.

Let us call the vertices of the sets $V_i \cap V_0$ and $V_i - V_0$ the internal and external vertices, respectively. Let the set of edges, which connect the vertices of internal and external vertices, be the main edges.

Definition 4. *Let the cluster $G_i\,(V_i,U_i)$ be remove-tendentious if $\nu_0 - \nu_i > \nu_i$.*

It is easy to see that the remove-tendentious property is the sufficient condition for the condition $f_1(G_i) > 1$. If a cluster is remove-tendentious, then cutting its main edges is "better" than the connection vertices of the sets $V_0 - V_i$ and $V_i - V_0$.

Let E_0 be an equivalence relation which best approximates a given symmetric relation R in the sense of the condition of approximation (1). The E_0 is called the optimal partition here.

3. LEMMA

Lemma. *If the number of all remove-tendentious clusters G_i of the subgraph $G_a\,(V_a,U_a)$ is n and*

$$f_2\,(G_i) \geq 1, I = 1, 2,..., n, \tag{5}$$

then there exists a E_0, the class of which is V_0.

Proof. If G_a has only one cluster, then the validity of assertion follows from Definition 4.

Now assume that $n \geq 2$, $\nu_1 = \nu_2 = ... = \nu_n = \nu$ and $t_1 = t_2 = ... = t_n = t$. From the expression of the function $f_2\,(G_i)$ by condition (5) it follows

$$t \leq (\nu_0 - \nu)^2/[\nu_0 + (n-2)\nu].$$

Let us prove that

$$n\nu(\nu_0 - \nu)^2/[\nu_0 + (n-2)\nu] \leq n\nu\nu_0 - (n+1)n\nu^2/2. \tag{6}$$

Here the left side of the inequality is not less than the sum of the numbers L_i, $i = 1, 2,..., n$. The right side is the number of edges when the complete clusters G_i are removed from the subgraph G_a. We can easily transform from the inequality (6) that

$$v \leq v_0(n-1)/[(n-2)(n+1)+2]. \tag{7}$$

Note that $v \leq v_0/n$ and substitution to the inequality (7) forms an identity. The latter result proves the validity of the inequality (6). Now it is clear that the connection of the sets of external vertices $V_i - V_0$ and the set $V_0 - V_i$ is not optimal. We can conclude that by the remove-tendentious property. This completes the proof of the lemma.

Corollary 1. *If the number of clusters is $k < n$, then the inequality* (6) *is likewise true.*

4. MAIN THEOREM

To prove the theorem, we use the following proposition: if the sets of vertices of the clusters $G_1, G_2,...,G_n$ transpose to new clusters G_i' such that the new numbers of the main edges are greater than such numbers formerly, then the probability of removing new clusters G_i' by removing the edges of U_0 does not decrease.

Theorem. *If each cluster $G_i, I \in \{1, 2,..., n\}$ of the subgraph G_a has*

1) $f_2(G_i) \geq 1$,

2) is remove-tendentious,

then there exists a relation E_0, where the sets of main edges of the clusters G_i are removed.

Proof. a) We assume now that the subgraph G_a contains n equal clusters G_i and each G_i has x internal and y external vertices. Let y be the maximal number of the external vertices. Hence the assertion is true by the lemma. We may transform from the expression (4) of the function $F_2(G_i)$

$$S \leq (v_0 - x - y)(v_0 - x)/x + y, \tag{8}$$

the first term of which on the right side is the upper bound of the external vertices sum if the number of clusters is $n - 1$. Now we increase the number of external vertices of any cluster by α so that the assumptions stay true. Let this new cluster be G_1'. Now the right side of the relation (8) will be

$$[v_0 - x - (y + \alpha)](v_0 - x)/x + y + \alpha$$

and we get

$$(v_0 - x - y)(v_0 - x)/x - \alpha(v_0 - x)/x + y + \alpha. \tag{9}$$

The second term of the connection shows that the sum of external vertices of the clusters $G_2, G_3,...,G_n$ decreases. Let us replace the sets of external vertices so that $k - 1$ clusters similar to G_1' arise. The number of clusters does not increase, so $k \leq n$. Note that the last cluster G_k may contain the external vertices less than $y + \alpha$, therefore the number of the main edges of G_k is not greater than the corresponding number of G_1'. By the lemma, no one cluster subset forms the optimal partition.

b) Assume again that the subgraph G_a has the structure described above. From the connection (8) and $S = ny$, it follows

$$n \leq (v_0 - x - y)(v_0 - x)/(xy) + 1. \tag{10}$$

It is evident that $nx \leq v_0$. Replacing n by the expression of the right term of inequality (10) gives

$$(v_0 - x - y)(v_0 - x)/y + x \leq v_0. \tag{11}$$

Assume that we add α internal and ß external vertices to G_1. Let the result be G_1'. We shall show that removing the cluster G_1' is not optimal.

It is clear that the sum of the external vertices of the other clusters decreases. We replace the vertices of the clusters $G_2, G_3, \ldots, G_n$ to create $k-1$ clusters, such as G_1'. Note that $k - 1 \leq n - 1$. From (10) it follows

$$k - 1 \leq [(v_0 - x - y - \alpha - ß)(v_0 - x - \alpha)]/[(x + \alpha)(y + ß)]. \tag{12}$$

Now we prove that v_0 is great enough for creating the new clusters, such as G_1'. We show that $k\,(x + \alpha) \leq v_0$. Using the multiplier $x + \alpha$, from (12) we transform

$$k\,(x + \alpha) \leq [(v_0 - x - y - \alpha - ß)(v_0 - x - \alpha)]/(y + ß) + x + \alpha.$$

Indicate that

$$[(v_0 - x - y - \alpha - ß)(v_0 - x - \alpha)]/(y + ß) + x + \alpha \leq v_0. \tag{13}$$

For this is sufficient if

$$(v_0 - x - y)(v_0 - x)/y - \alpha - 1 \geq [(v_0 - x - y - \alpha - ß)(v_0 - x - \alpha)]/(y + ß). \tag{14}$$

We obtain from the latter

$$ß(v_0 - x - y)(v_0 - x) + \alpha y\,(v_0 - x) + ßy\,(v_0 - x) + \\ + \alpha y\,(v_0 - x - y - \alpha - ß) \geq \alpha y\,(y + ß) + y\,(y + ß). \tag{15}$$

Note that

1) $\alpha y(v_0 - x) \geq \alpha y(y + ß)$ because by the remove-tendentious property $v_0 - x - \alpha \geq y + ß$, it follows that $v_0 - x > \alpha + ß$;

2) similarly, $ßy(v_0 - x) > y(y + ß)$;

3) the fourth term on the left side of the inequality (15) is positive because $v_0 - x - \alpha > y + ß$.

It follows that the expression (14) is true. Using the inequality (14) to (13), by substituting we obtain

$$(v_0 - x - y)(v_0 - x)\,/\,y + x - 1 \leq v_0,$$

which is true by the inequality (11) and so $k\,(x+\alpha) \leq v_0$ by transitivity of the expressions. From the lemma it follows that the relation E_0 does not contain the classes G_1' defined in this part.

c) If the subgraph $G_a(V_a,U_a)$ contains the cluster G_1', formed by adding the α internal vertices to one G_1 of the equal clusters, then we may discuss analogously for parts a) and b).

The described replacings of the external and internal vertices exhaust the structure cases of G_a. This completes the proof of the theorem.

Let us now consider a special case of the subgraph G_a. We assume that G_a has $p+q=n$ clusters and

1) $f_2(G_i) \geq 1, i = 1, 2, \ldots, p,$

2) $f_2(G_j) < 1, j = p+1, p+2, \ldots, p+q.$

If we remove the whole clusters G_j, $j = p+1, p+2, \ldots, p+q$, then a subgraph of G_a keeps. Let it be G_a'. Compute the new values of the function $f_2(G_i)$, $i = 1, 2, \ldots, p$, and denote those values $f_2'(G_i)$, $i = 1, 2, \ldots, p$.

Corollary 2. *If G_a contains the subgraph G_a', and the new values are*

$$f_2'(G_i) \geq 1, i = 1, 2, \ldots, p,$$

then removing the main edges of clusters G_i, $i = 1, 2, \ldots, p$, is optimal.

5. APPLICATIONS

The auxiliary functions discussed here have two common features. First, they have local nature, because those functions are dealing with the set of non-transitive pairs of edges of the subgraph G_a only. Second, those functions create a kind of a developing process. If the set of elimination edges is determined and removed, then the auxiliary functions may be often used once more in the subgraph G_a.

If the source graph G contains such subgraph G_a that we can specify the class of E_0, then we may remove some set of rows and columns of the incidence matrix of G. Therefore the degree of matrix representation is less for the next step of valuations. When Corollary 2 is applicable, then removing the main edges of clusters G_i, $i = 1, 2, \ldots, p$, may change the values of functions such that on other clusters G_j, $j = p+1, p+2, \ldots, p+q$, the theorem will be usable. The next table contains the data of such G_a. The cluster to control the removal is G_8. It is easy to see that after removing the sets of main edges of the clusters G_i, $i = 1, 2, \ldots, 7$, we must remove the main edges of the cluster G_8, too. Finally, one class of E_0 is V_0.

Values of the functions of the subgraph G_a

Part I ($v_0 = 140$)

	1	2	3	4	5	6	7	8
v_i	1	4	7	15	25	10	20	45
t_i	9	9	9	25	15	9	5	65
L_i	9	36	40	375	375	90	100	2935
$f_1(G_i)$	139.0	34.0	29.9	8.3	4.6	13.0	6.0	2.1
$f_1(G_{0i})$	10.0	3.3	2.3	2.7	1.7	1.9	1.3	1.5
$f_2(G_i)$	13.8	10.3	12.8	3.1	2.8	6.7	4.6	0.8

Part II ($v_0 = 95$)

	1	2	3	4	5	6	7	8
$f_1'(G_i)$	94.0	22.8	19.8	5.3	2.8	8.5	3.8	
$f_1'(G_{0i})$	9.9	3.1	2.1	2.5	1.4	1.7	1.1	
$f_2'(G_i)$	9.5	7.3	9.3	2.1	2.0	4.9	3.5	

Note. Finding the subgraphs of G_a needs effective algorithms to determine the maximal complete subgraphs. Those algorithms were created by L. Võhandu and R. Kuusik [2].

REFERENCES

1. Veiner, G. Some problems on approximating symmetric relations by equivalence relation. – Trans. Tallinn Techn. Univ., 1993, **734**, 75–83.
2. Võhandu, L. and Kuusik, R. Cliques and algorithms with a hidden parallelism. – Trans. Tallinn Techn. Univ., 1993, **734**, 63–74.

MÕNEST PROBLEEMIST SÜMMEETRILISE RELATSIOONI LÄHENDAMISEL EKVIVALENTS-RELATSIOONIGA

Guido VEINER

On käsitletud niisuguse ekvivalentsrelatsiooni E_0 leidmist, mis kõige paremini lähendab sümmeetrilist binaarset relatsiooni R tingimusel, et hulga

$$\{R - E_0\} \cup \{E_0 - R\}$$

elementide arv oleks minimaalne. On üldistatud varem käsitletud teoreem. Käsitluses on kasutatud kolme põhilist abifunktsiooni, mis lahendavad probleemi ühe relatsioonide klassi puhul.

НЕКОТОРЫЕ ПРОБЛЕМЫ АППРОКСИМАЦИИ СИММЕТРИЧНЫХ ОТНОШЕНИЙ ОТНОШЕНИЕМ ЭКВИВАЛЕНТНОСТИ

Гуйдо ВЕЙНЕР

Рассмотрены проблемы аппроксимации симметричного бинарного отношения R отношением эквивалентности E_0. Условием аппроксимации является минимизация числа элементов множества

$$\{R - E_0\} \cup \{E_0 - R\}\ .$$

Обобщен ранее полученный автором результат с помощью трех вспомогательных функций. Эти функции использованы в одном классе отношений.