Teszteset válogatás távolság metrikával

KOVÁCS GÁBOR

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
kovacsg@tmit.bme.hu

Kulcsszavak: teszteset válogatás, sztring szerkesztési távolság, kommunikációs protokollok

E cikk a szerkesztési távolság metrika alapú tesztválogatási módszert [1,2] bővítve egy polinom idejű eljárást mutat be, amelyekkel meghatározható egy kommunikációs protokoll tesztkészletének legkisebb számosságú, legnagyobb belső távolsággal bíró lefedő tesztkészlete. Az eljárás először meghatározza a megtartandó tesztesetek minimális számát, majd második lépésként kiválasztja azt a készletet a legkisebb méretűek közül, amely teszteseteinek távolságösszege a legnagyobb, tehát tesztesetei a legjobban szétszóródnak a protokoll állapotterén.

1. Bevezetés

Az automatikus tesztgeneráló algoritmusok használatának elterjedését leginkább akadályozó tényező az, hogy jelenleg nem megoldható a – formális – modellből származtatott óriási méretű tesztkészletben található redundanciák felderítése. A szükségessé váló tesztválogatás célja a teszteset-halmaz számosságának minimalizálása úgy, hogy a leszűkített készlet még mindig képes legyen kielégíteni olyan kritériumokat, mint például a tesztesetek által definiált útvonalak lefedése [1,2], a hibamodellel megvalósított kódlefedés [3], vagy a teszteset részcélok kielégítése [4].

Tekintettel az automatikusan generált tesztkészletek gigantikus méretére, valamint arra, hogy a válogatási problémák rendszerint NP-teljesek, hatékony közelítések megfogalmazása szükséges. Ilyen közelítés a tesztesetek közötti sztring szerkesztési távolságot figyelembe vevő módszer [1], amely akkor mond két szekvenciát (tesztesetet) hasonlónak, redundánsnak, ha azok szerkesztési távolsága nem nagyobb egy adott ε küszöbszámnál. Ezt a módszert kiegészítendő e cikk megmutatja, hogy a távolság alapú teszteset válogatás polinom időben megoldható.

A probléma első része az adott küszöbszámhoz tartozó legkisebb számosságú tesztkészlet méretének meghatározása, aminek a célja a tesztkészlet méretének lehető legnagyobb mértékű csökkentése. A második részprobléma a tesztesetek szerkesztési távolsága alapján megmutatja a tesztkészletet azon minimális elemszámú részhalmazát, amelyben a tesztesetek leginkább eltérnek egymástól.

A cikk második szakasza összefoglalja az eseményszekvenciák (tesztesetek) közötti távolsághoz kapcsolódó fogalmakat, jelöléseket, definíciókat. A harmadik, illetve a negyedik szakasz egy-egy átalakítást mutat be, mellyel a minimális számosságú, illetve a maximális belső távolságú lefedő tesztkészlet meghatározható, végül pedig áttekinti a bemutatott eljárás általánosíthatóságát.

2. Szekvenciák távolsága

A cikk az egyszerűség kedvéért tesztesetek helyett eseménysorozatokra definiálja az algoritmusokat, amelyek később változtatás nélkül kiterjeszthetők az általánosabb teszteset problémára.

Egy *t* eseménysorozat, amelyet – például – véletlen sétával állíthatunk elő egy kommunikációs protokoll viselkedését definiáló véges állapotgép alapján, bemeneti (*I*) és kimeneti (*O*) események sorozatából áll: $t = x_1x_2...x_N$, ahol x_i egy esemény az $I \cup O$ halmazból.

Az eseményszekvenciák által definiált módon leképezhetők tömör, a protokoll állapotgépéről is információt hordozó karaktersztringekké úgy, hogy egy-egy karakter tetszőleges számú, de legalább egy, egymás utáni be-, illetve kimeneti eseményt reprezentál [2].

Példa. Legyen az eseménykészlet $I \cup O = \{a, b, c\}$, és legyen a karakter ábécé $C = \{V, W, X, Y, Z\}$. A leképezések halmaza legyen: $\{(a, V), (b, W), (c, X), (aa, Y), (bc, Z)\}$. E leképezés szerint az *aa*, illetve a *bc* eseménysorozat lehet egy-egy hurok a protokoll állapotgépének egy-egy állapota körül. Ha a *T* szekvenciakészlet (tesztkészlet) a következő elemekből áll: $T = \{baab, bcbc, aabc, abab\}$, akkor a megfelelő sztringkészlet $S = \{WYW, ZZ, YZ, VWVW\}$ lesz.

Két sztring közötti szerkesztési távolság [5] az egyik a másikba alakításához szükséges karakterbeszúrás, karaktertörlés, illetve karakterátírás műveletek minimális száma, szerkesztési költsége, ha e műveletek nem egységnyiek. A számítás komplexitása a két sztring hosszával arányos. A távolság egy metrika, azaz nemnegatív, szimmetrikus, és teljesül rá a háromszög egyenlőtlenség.

E definíció alapján az eseményszekvenciákból képezett sztringkészlethez konstruálható egy $\boldsymbol{D}$ távolságmátrix, ahol a d_{ij} elem a sztringkészlet *i*-edik és *j*-edik eleme közötti szerkesztési távolság. A távolság fogalma miatt a $\boldsymbol{D}$ mátrix szimmetrikus, és főátlójában *0* értékeket tartalmaz. Az elemek értékei e cikkben az egyszerűség kedvéért egészek, [2] alapján azonban racionális értéket is felvehetnek.

Példa. Legyenek a sztringműveletek egységnyi költségűek. Folytatva az előző példát, az *S* készlet első (*WYW*) és második (*ZZ*) eleme csak *3* sztringművelettel alakítható át egymásba. Az első (*WYW*) és harmadik (*YZ*) eleme között a távolság *2* egység, mert az első sztring első *W* karakterét törölve és a második *W*-t *Z*-re átírva a harmadik sztringet kapjuk. Az első (*WYW*) a negyedikbe (*VWVW*) az első *V* törlésével és az *Y V*-vé való átírásával alakítható át két lépésben. A második (*ZZ*) a harmadikba (*YZ*) egy átírás művelettel átvihető. A második (*ZZ*), illetve a harmadik (*YZ*) sztring távolsága a negyediktől (VWVW) *4*.

A távolságmátrix tehát:

$$\boldsymbol{D}=\begin{pmatrix} 0 & 3 & 2 & 2 \\ 3 & 0 & 1 & 4 \\ 2 & 1 & 0 & 4 \\ 2 & 4 & 4 & 0 \end{pmatrix}.$$

Ha két eseményszekvencia (teszteset) a protokoll állapotterének azonos részét járja be hasonló úton, akkor a köztük lévő szerkesztési távolság „kicsi” lesz, és fordítva, a „nagyon eltérő” eseményszekvenciák – jó eséllyel – az állapottér más-más részeit járják be. Ha két teszteset az állapottér azonos részét ellenőrzi, akkor az egyik redundánsnak tekinthető és eldobható.

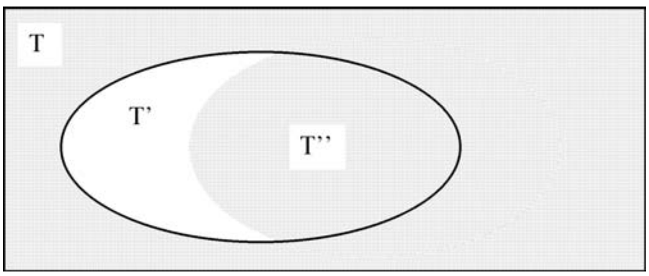
E megfontolás alapján definiálható egy ε-közelítés [1,2]: a *T'* tesztkészlet ε-lefedése a *T* tesztkészletnek, ahol *T' T* egy részhalmaza, ha *T* minden *t* eleméhez található *T'*-ben egy *t'* elem, amelyre *t* és *t'* szerkesztési távolsága nem nagyobb, mint ε. Vagyis *T* tesztkészlet tetszőleges eleme előállítható a leszűkített *T'* tesztkészlet egy eleméből maximum ε költségű szerkesztési lépéssel. A definíció következménye, hogy tetszőleges két *T'*-beli elemre azok távolsága nagyobb, mint ε.

3. A minimális számosságú lefedő készlet mérete

Az elsődleges cél a tesztválogatás során a tesztkészlet méretének lehető legnagyobb mértékű csökkentése. Ehhez az egymást ε-lefedő tesztesetek halmazából a legkisebb elemszámú *T'* részhalmazt kell kiválasztanunk, amely az összes *T*-beli elemet ε-lefedi.

A tesztkészletek viszonya az *1. ábrán* látható. A teljes *T* tesztkészletet a külső téglalap jeleníti meg, amiből keressük az ellipszissel jelölt *T'* részkészletet. A tónusozott terület azon teszteseteket jelöli, amelyek az ε távolságnál közelebb vannak egymáshoz (egymást ε-lefedő tesztesetek halmaza). *T'* egy részhalmazának elemei nem fedhetők le más tesztesettel, ezért mindenképpen elemei lesznek a kiválasztott *T'*-nek, ezt a fehér terület jelöli. Ezen kívül *T'* egy *T''* részhalmaza ε-lefedi a teljes *T* halmazt. *T''* és ezáltal *T'*, mérete akkor lesz a legkisebb, ha minden egyes *T''*-beli elemhez a lehető legtöbb független *T*-beli elemet rendeljük.

1. ábra
Az eredeti tesztkészlet és a lefedő készlet viszonya



Készítsünk a ***D*** mátrix alapján egy *0* vagy *1* elemeket tartalmazó ***A*** mátrixot úgy, hogy ha az *i*-edik szekvencia ε-lefedi a *j*-edik szekvenciát, vagyis ha távolságuk nem nagyobb ε-nál, akkor legyen $a_{ij} = 1$. A *T''* készlet minimális számosságát az ***A*** mátrix rangja (lineárisan független sorainak, illetve oszlopainak száma) adja meg.

Példa. Az előző szakasz példájában szereplő ***D*** mátrixhoz tartozó ***A*** mátrix a következő, ha $\varepsilon = 2$:

$$\boldsymbol{A}=\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Az ***A*** mátrix rangja *2*, tehát *T''* kételemű. Mivel nincs olyan teszteset, ami nem ε-lefedhető más tesztesettel, azaz nincs csak *0*-t tartalmazó sor, illetve oszlop ***A***-ban, a lefedő *T'* is kételemű lesz. A példa két megoldása a *{WYW, ZZ}* és a *{YZ, VWVW}* készlet, amelyek kölcsönösen lefedik egymást.

Mivel a távolság definíciója szimmetrikus, minimális számosságú ε-lefedő *T''* tesztkészletpár mindig létezik, ha $A \neq 0$, különben, azaz a legrosszabb esetben, *T* önmaga a megoldás. Előfordulhat az is, hogy nem csak egy ilyen *T''* pár létezik, ezért a megoldások között egy további kritérium alapján is különbséget tehetünk, különbséget kell tennünk.

4. Maximális belső távolságú lefedő készlet

Azt már tudjuk, hogy mekkora annak a tesztkészletnek a minimális mérete, amely képes kielégíteni az ε-lefedést. A tesztkészlet-optimalizáció második lépése ezek közül kiválasztja azt a minimális elemszámú tesztkészlet jelöltet, amelyik a lehető legnagyobb mértékben szórja szét a teszteseteket a protokoll állapotterén, vagyis megkeresi azt a megoldást, amelyre a *T'* készleten belüli szekvenciák közötti távolságok összege maximális. A megoldandó optimalizálási probléma tehát:

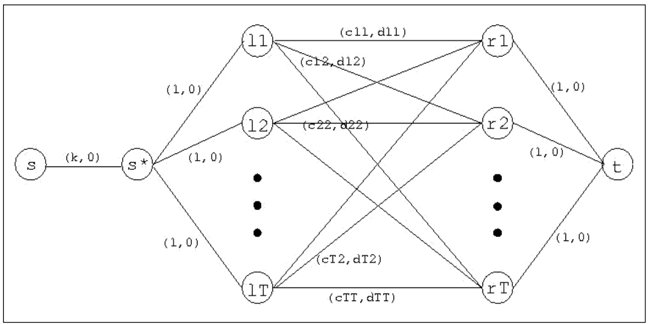
$$max \sum_{\forall t'_i, t'_j \in T'} d(t'_i, t'_j)$$

ahol *d* a távolságfüggvény, valamint t'_i és t'_j *T'*-beli elemek, azaz távolságuk nagyobb, mint ε.

A probléma ekvivalens a következő folyamproblémával [6]. Nyilvánvalóan minél több elemet tartalmaz a *T'* tesztkészlet, annál nagyobb lehet a maximális távolságösszeg, ezért legyen *T'* elemszáma pontosan *k*, a 3. szakaszban megtalált minimális számosság értéke. Legyen ***C*** egy a ***D*** távolságmátrix méretével megegyező méretű kapacitásmátrix, amely elemei *0* vagy *1* értéket vehetnek fel. Legyen *G = (V, E,* ***C***, ***D****)* egy a ***C*** és ***D*** mátrixokkal paraméterezett páros irányított gráf, ahol a csomópontok halmaza öt részre osztható $V = B \cup J \cup \{s, s^*, t\}$.

Tartozzék a ***D*** mátrix minden egyes sorához a $B = \{l_1, l_2, ..., l_T\}$ bal oldali csomóponthalmaz egy eleme, valamint minden egyes oszlopához a $J = \{r_1, r_2, ..., r_T\}$ jobb oldali csomóponthalmaz egy eleme. Vezessen a bal oldali csomóponthalmaz *i*-edik csomópontjából (l_i) c_{ij} kapacitással és d_{ij} költséggel irányított él a jobb oldali csomóponthalmaz *j*-edik csomópontjába (r_j). Legyen *s* összekötve az *s** csomóponttal egy *k* kapacitású és *0* költségű irányított él által. Legyen *s** összekötve *B* összes csomópontjával egységnyi kapacitású, *0* költségű irányított éllel. Vezessen irányított él *J* minden csomópontjából *t*-be szintén egységnyi kapacitással valamint *0* költséggel. Az így konstruált *G* gráf a *2. ábrán* látható.

2. ábra Az ekvivalens folyamprobléma



Ez egy maximális költségű maximális folyamprobléma, ahol az *s*-ből *t*-be irányuló maximális folyam értéke *k*. A maximális folyam bizonyos c_{ij} értékeket *1*-re állítva kijelöli *T'*-t. Minden *B*-beli csomópontból maximum egy kimenő él kerül kiválasztásra, és minden *J*-beli csomópontra maximum egy bejövő él kerül kiválasztásra. Tehát a folyam ***D*** minden sorából, illetve oszlopából maximum egy elemet jelöl meg úgy, hogy a megjelölt elemek összege maximális.

Vagyis az optimalizálási probléma átfogalmazva:

$$\max \sum_{i=1}^{|T|} \sum_{j=1}^{|T|} c_{ij} d_{ij} \text{, ahol}$$

$$\forall j : \sum_{i=1}^{|T|} c_{ij} \leq 1,$$

$$\forall i : \sum_{j=1}^{|T|} c_{ij} \leq 1 \text{ és}$$

$$\sum_{i=1}^{|T|} \sum_{j=1}^{|T|} c_{ij} = k .$$

A problémát [7]-ben *k*-számosságú hozzárendelési problémának (k-cardinality Assingment Problem, *k*-AP) nevezték el, és bebizonyították, hogy polinom időben megoldható.

Példa. Az már az előző szakaszból tudjuk, hogy $k = 2$ az $\varepsilon = 2$ esetben. Az egyik megoldás szerint $c_{14} = c_{23} = 1$, a másik szerint $c_{32} = c_{41} = 1$, és minden más elem ***C***-ben *0*. Mivel a maximális távolságösszegre mindkét megoldás esetén *3*-at kapunk ($c_{14} {}^* d_{14} + c_{23} {}^* d_{23} = 1{}^*2+1{}^*1 = 3$, $c_{32} {}^* d_{32} + c_{41} {}^* d_{41} = 1{}^*2+1{}^*1 = 3$), két egyenértékű megoldás létezik: *{baab, bcbc}* és *{aabc, abab}*.

5. Összefoglalás

A minimális számosságú ε-lefedő tesztkészlet megtalálása a cikkben leírt módon visszavezethető lineáris függetlenség vizsgálatára. Így meghatározható a legkisebb elemszámú olyan tesztkészlet mérete, amely legfeljebb egy előre meghatározott ε szerkesztési költséggel átalakítható a teljes tesztkészletté. A legkisebb méretű és a protokoll állapotterén leginkább szétszórt lefedő tesztkészlet megtalálása ekvivalens egy polinom időben megoldható maximális költségű minimális folyamproblémával.

Míg az eseménysorozatok sztringekké alakíthatók át, a mindennapi gyakorlatban használt tesztesetek fákkal reprezentálhatók. Ez a cikkben bemutatott átalakítást az egyes adatobjektumok közötti szerkesztési távolságának számítási módja nem érinti, így az – változtatás nélkül – alkalmazható fákra is. Ugyanakkor megjegyzendő, hogy a tesztesetek leírására alkalmazható gyökérrel rendelkező, bejárási sorrend nélküli címkézett fák közötti szerkesztési távolság meghatározása szintén NP-teljes [8].

Irodalom

[1] Vuong S.T., Curgus J.:
Test coverage metrics for communication protocols.
In proceedings of the IWPTS.
Leischendam, The Netherlands, 1991.

[2] Feijs L.M.G., Goga N., Mauw S., Tretmans J.:
Test Selection, Trace Distance and Heuristics.
In proceedings of Testing Communication Systems XIV, pp.267–282., Berlin, Germany, 2002.

[3] Kovács G., Le Viet D., Wu–Hen-Chang A., Pap Z., Csopaki G.:
Applying Mutation Analysis to SDL Specifications.
In proceedings of SDL-Forum,
Stuttgart, Germany, 2003.

[4] Csöndes T., Kotnyek B., Szabó J. Z.:
Application of Heuristic Methods for Conformance Test Selection.
European Journal of Operational Research, 2001.

[5] Wagner R.A., Fischer M.J.:
The String-to-String Correction Problem.
Journal of the ACM, 21(1):168–173, 1974.

[6] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C.:
Új algoritmusok.
Scolar Kiadó, Budapest, 2003.

[7] Dell' Amico M., Martello S.:
The k-cardinality Assignment Problem.
Discrete Applied Mathematics, 76:103–121, 1997.

[8] Zhang K., Statman R., Shasha D.:
On the editing distance between unordered labeled trees.
Information Proc. Letters, 42(3):133–139, 1992.