NYILATKOZAT

Név: Mikulás Zsófia Blanka

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: CWM8OZ

Szakdolgozat címe:
Hatékony vírustesztelési algoritmusok

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2022.05.31

Mikulás Zsófia

a hallgató aláírása

Eötvös Loránd Tudományegyetem

Természettudományi Kar

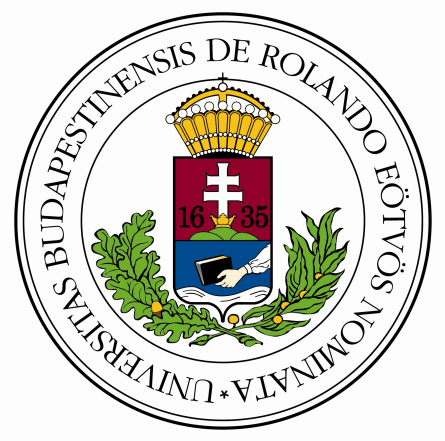
Hatékony vírustesztelési algoritmusok

Mikulás Zsófia

BSc Szakdolgozat

Témavezető: Csóka Endre, tudományos főmunkatárs

Rényi Alfréd Matematikai Kutatóintézet



Budapest, 2022.

Tartalomjegyzék

Köszönetnyilvánítás

Elsősorban szeretném megköszönni témavezetőmnek, Csóka Endrének a rendszeres konzultációkat és a szakdoldolgozat megírása során nyújtott segítségét.

Továbbá hálás vagyok a családomnak, az évfolyamtársaimnak, a Bolyai Kollégium közösségének, és K. Ákosnak az egyetemi évek alatt nyújtott támogatásukért és jó pillanatokért.

1. fejezet

Áttekintés

Az utóbbi két évben, a koronavírus jelenlétének köszönhetően, ismét komoly figyelem irányult a hatékony vírustesztelési algoritmusok keresésére a tömeges tesztelés időigényessége és költségessége miatt.

A „csoporttesztelési probléma” a II. világháború alatt, az Egyesült Államokban merült fel, amikor a háború miatt nagyon sok férfit soroztak be a hadseregbe, és szükség volt a szifilisz szűrésére. A szifiliszre már akkor is létezett pontos vérvizsgálat, a Wassermann-teszt, mint ahogy a koronavírus szűrésére is létezik most a PCR-teszt, vagy az antigén-teszt például, így elég volt mindenkitől mintát venni, és arról a műszerek egy teszt elvégzésével megmondták, fertőzött-e az illető, vagy sem. Azonban, mivel a szifilisz ritka betegség, és a koronavírus-tesztek számának aránya a mintavételek számához képest szintén alacsony, ezért az egyenkénti tesztelés nem hatékony. Ez jól látható például információelméleti eszközökkel (3.fejezet).

Robert Dorfman 1943-as cikkében [4] mutatta be a csoporttesztelési eljárást, amely összeöntött minták tesztelésével drasztikusan lecsökkentheti a szükséges tesztszámot. A csoporttesztelés azt jelenti, hogy a tesztelendő személyek egy csoportjától levett mintákat összekeverjük, és ezt rakjuk be a gépbe. A gép ugyanis azt méri, hogy a mintában jelen van-e a kórokozó, tehát ha elég pontos (és a használt

műszerek azok), a teszt eredménye $\oplus$ lesz, hogyha van a minták között legalább egy fertőzött. Ha viszont az eredmény $\ominus$, akkor biztosak lehetünk benne, hogy a minták egyike sem származik megfertőződött személytől. Ilyen tesztelések sorozatával a minták számánál lényegesen kevesebb teszttel meg tudjuk állapítani, kik fertőzöttek.

A probléma matematikai modelljei a kényelmesebb vizsgálhatóság miatt idealizáltak, tehát általában nem veszik figyelembe többek között azt, hogy a gépi és vegyi meghibásodások, valamint az emberi pontatlanság miatt olykor hibás eredményt kaphatunk. A gyakorlatban érdemes azt is figyelembe venni, hogy a minták túlzott felhígítása miatt a vírus jelenléte kimutathatatlanná válhat, illetve a műszerek nem biztos, hogy bináris, azaz egyértelműen $\oplus$, vagy $\ominus$ eredményt adnak. Fontos tényező lehet, hogy az emberek, akiktől mintát vettünk, nem azonos valószínűséggel fertőzöttek. Például a koronavírus esetében, ha egy már ismerten fertőzött személlyel egy háztartásban élők mintáját teszteljük, nagyobb valószínűséggel kapunk pozitív eredményt, mintha egy nem kontaktszemély mintáját teszteljük. A gyakorlatban ezek jelentős faktorok, ezért ha alkalmazni akarjuk az algoritmusokat, bele kell őket számolni azok megalkotásánál. Azonban ezt a sokténye zős modellt elsőre nehéz kezelni, így célszerű először az idealizált, úgynevezett „standard zajmentes” csoporttesztelési modellt vizsgálni.

Az algoritmusoknak egy fontos megkülönböztetését tudjuk tenni az alapján, hogy a következőként tesztelni kívánt csoport kiválasztásához figyelembe vesszüke a korábbi tesztek eredményeit. Ez alapján az algoritmus lehet adaptív (figyelembe vesszük), illetve nemadaptív (nem vesszük figyelembe).

Egy másik fontos különbségtétel, hogy a tesztelés során kapott eredmény bináris (azaz egyértelműen $\ominus$, vagy $\oplus$), vagy nem bináris, például mert a műszer a mintában lévő vírustestek számát, vagy arányát állapítja meg. A szakirodalomban először, érthető módon, a bináris modellt vizsgálták (pl. [4], [12], [11], [10]). A nem

bináris modellt (5.fejezet) először Charles G. Pfeifer és Peter Enis írták le [8] a következő módon. Ha az eredmény $\ominus$, akkor a 0 számot kapjuk vissza, ha viszont $\oplus$, egy, a $(0, \infty)$ halmazba eső számot kapunk eredményül, ami a $\oplus$ mértékét jelöli. Csoporttesztelési eljárás során egy csoport teszteredménye a csoportban lévő minták eredményeinek összege, tehát $\oplus$ mintát nem tartalmazó csoporté 0, a $\oplus$ mintákat tartalmazóé pedig ezen minták eredményeinek összege. Hogy könnyebb legyen kezelni a számokat, ezt az összeget lenormálhatjuk, hogy egy 0 és 1 közé eső számot kapjunk, és ez nyilvánvalóan nem változtat azon, hogy milyen algoritmusok hatékonyak ebben az esetben.

A szakdolgozatomban nagyrészt azoknak az adaptív és nemadaptív algoritmusoknak a bemutatására koncentrálok, amelyek a tesztek várható értékét, azaz $E(T)$-t optimalizálják. Igyekszem körbejárni, hogy ezeknek az algoritmusoknak a gyakorlatban való alkalmazása milyen előnyökkel, illetve hátrányokkal járnak, valamint, egyáltalán alkalmazhatóak-e értelmesen.

A dolgozat kiindulási pontja Csóka Endre 2020-as cikke [3] volt, a felépítéséhez és tartalmához pedig sokat merítettem Aldridge, Johnson és Scarlett 2020-ban publikált összefoglaló cikkéből. [7]

2. fejezet

Néhány jelölés és definíció

2.0.1. Jelölések.

- $N \sim$ az összes minta száma
- $p \sim$ annak valószínűsége, hogy egy minta $\oplus$
- $q \sim$ annak valószínűsége, hogy egy minta $\ominus$, azaz $q = (1-p)$
- $d \sim$ a $\oplus$ elemek száma
- $D \sim$ a $\oplus$ elemek halmaza
- $T \sim$ tesztek száma
- $\mathbf{X} = (\mathbf{x_{ij}}) \sim$ a tesztelési mátrix, ahol $\mathbf{x_{ij}} = 1$, ha a j. minta az i. tesztben szerepel, különben 0
- $\mathbf{y} \sim$ eredményvektor, ahol az i. koordináta 1, hogyha az i. minta a tesztelési eredmények alapján $\oplus$, és 0, hogyha $\ominus$

2.0.2. Definíció. Egy algoritmust optimálisnak nevezünk adott N számú mintára és p valószínűségre nézve, hogyha minimalizálja a tesztek várható értékét, $E(T)$-t.

2.0.3. Definíció. Egy algoritmus adaptív, ha a soron következő tesztelendő csoport kijelölésekor figyelembe vehetjük a korábbi teszteredményeket, és nemadaptív, hogyha a tesztelési eljárás megkezdése előtt összeállítjuk a csoportokat, amiket már nem változtatunk a beérkező eredmények függvényében.

2.0.4. Definíció. Tegyük fel, hogy az N mintahalmazt az $\mathbf{X}$ terv szerint teszteljük, aminek az eredménye $\mathbf{y}$. Ekkor a minták egy L halmazát kielégítő halmaznak nevezzük, ha minden $\oplus$ eredményű halmazzal van közös eleme, és a $\ominus$ eredményűekben nincs L-beli elem. (Az nyilvánvaló, hogy a D halmaz kielégítő.)

2.0.5. Definíció. A Bernoulli modellben minden egyes teszt kiválasztásakor minden mintát egymástól függetlenül, egy adott $p = \frac{\nu}{d}$ valószínűséggel vesszük be a tesztelendő halmazba.

3. fejezet

Információelméleti határ

Kezdésként, információelméleti ismeretek segítségével adunk egy egyszerű becslést arra vonatkozóan, legalább hány teszt elvégzése szükséges a pontos eredményhez, a standard zajmentes modellben (4.fejezet). Ehhez vezessük be az entrópia fogalmát.

3.0.1. Definíció. Ha az X változó az $\{x_1, x_2, \cdots, x_n\}$ halmaz elemeit rendre $p_1, p_2, \cdots, p_n$ valószínűséggel veszi fel ($\sum p_i = 1$), akkor az $X = x_i$ esemény információtartalmát a $H(x_i) = -p_i \cdot \log_2 p_i$ értékkel, az X valószínűségi változó entrópiáját pedig a $H(X) = -\sum p_i \cdot \log_2 p_i$ értékkel definiáljuk.

A fogalmat intuitíven a következő módon vezethetjük be.

Képzeljük el, hogy olyan folyamatosan érkező, szimbólumokból álló adatsort kell rögzítenünk bitekkel, amikről előre tudjuk, hogy melyik szimbólum milyen valószínűséggel érkezhet be. Mekkora tárhelyre lesz szükségünk adott számú szimbólumból álló adatsor eltárolásához, ha a lehető leghatékonyabban, azaz várhatóan a legkevesebb tárhely felhasználásával akarjuk eltárolni?

Mindegyik szimbólumnak egyedi kódot kell találni, így csinálhatnánk azt, hogy n db szimbólum esetén mindegyikhez hozzárendelünk egy $\lceil \log_2 n \rceil$ hosszú $0-1$ so-

rozatot. Azonban, ez nem túl hatékony, a többihez képest kis valószínűséggel előforduló szimbólumok miatt nem érdemes a gyakori szimbólumokhoz is ilyen hosszú kódot rendelni. Egy egyszerű példa erre, amikor az a, b, c és d szimbólumokból álló adatsort akarjuk kódolni, és az a előfordulási valószínűsége $\frac{1}{2}$, a b-é $\frac{1}{4}$, és a c, d-é pedig $\frac{1}{8}$. Ha ekkor mindegyikhez $\lceil \log_2 4 \rceil = 2$ bit hosszúságú kódot rendelünk, egy k hosszú szimbólumsorozat számára $2k$ tárhelyet kell fenntartanunk, viszont ha a sokkal gyakrabban felbukkanó a-hoz a 0, a b-bez az 10, a maradék kettőhöz pedig az 110, illetve az 111 kódokat rendeljük, a felhasznált tárhely várható értéke $k \cdot \left(\frac{1}{2} \cdot 1 + 2 \cdot \frac{1}{4} + 2 \cdot 3 \cdot \frac{1}{8}\right) = \frac{7}{4}k$. Azt nem tehetjük meg, hogy a b, c és d közül valamelyiket 1-gyel kódoljuk, a másik kettőt pedig 10-val és 11-gyel, hiszen ha például b-t kódoljuk 1-gyel, c-t pedig 10-val, akkor az 10 kódsorozatról nem fogjuk tudni megállapítani, hogy a ba, vagy a c karaktersorozatot kódoltuk-e el vele.

Tehát a fenntartott tárhely minimalizálásához a következő feladatot kell végiggondolnunk. Adott egy adatsor, amelyben az $\{x_1, x_2, \cdots, x_n\}$ szimbólumhalmaz elemei vannak, és tudjuk, hogy a szimbólumok $p_1, p_2, \cdots, p_n$ valószínűséggel fordulnak elő (ahol $\sum p_i = 1$). Ezt akarjuk elkódolni. Követelmény, hogy az elkódolandó szimbólumsort vissza tudjuk majd fejteni a tárolt adatokból, így emellett a feltétel mellett akarjuk minimalizálni az eltárolt bitek számának várható értékét. A kérdés, hogy milyen $h_1, h_2, \cdots, h_n$ hosszúságú bitsorozatokat rendeljünk az egyes szimbólumokhoz, ha szeretnénk minimalizálni $(\sum p_i h_i)$-t, $\sum 2^{-h_i} \leq 1$ mellett?

Ha elhagyjuk azt a feltételt, hogy egész hosszú kódsorozattal kell jelölni az egyes szimbólumokat, akkor a válasz az, hogy $-\log_2 p_i$ hosszúságúnak kellene választani az egyes kódokat. Shannon megmutatta [9], hogy ez az alsó becslés aszimptotikusan éles.

Tehát az X valószínűségi változó $H(X) = -\sum p_i \cdot \log_2 p_i$-vel definiált entrópiája az adatsor egy szimbólumához szükséges tárhely átlagos mérete, ha el lehetne nem egész szám hosszú bitsorozattal is kódolni.

Az entrópia jól viselkedik az összeadásra, azaz X, Y valószínűségi változókra $H(X,Y) \leq H(X) + H(Y)$, és az egyenlőség pontosan akkor teljesül, ha X, Y függetlenek. Ha X, Y valószínűségi változók, akkor $H(Y|X)$ jelöli az Y változó X feltétellel vett feltételes entrópiáját. Ekkor teljesül, hogy $H(X,Y) = H(X) + H(Y|X)$.

A vírustesztelési modellben az $X_1, \cdots X_n$ valószínűségi változók a tesztelendő csoportokhoz tartozó eredményeket jelölik, amiknek $\oplus$ és $\ominus$ lehet az értéke, így a tesztek kimenetelétől függően egy n hosszú $\oplus - \ominus$ sorozatot jelölnek. A teszteredmények információtartalma:

$$H(X_1, \cdots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1X_2 \cdots X_{n-1})$$

Egy teszt legfeljebb 1 bit információt tud adni (hiszen a teszteredmény bináris, $\oplus$, vagy $\ominus$), és pontosan 1 bit információt ad, ha a teszt elvégzésekor $\frac{1}{2}$ valószínűléggel kapunk $\oplus$ eredményt. Tehát a tesztsorozat megadásánál érdemes figyelni arra, hogy mindig úgy adjuk meg a következő tesztelendő csoportot, hogy minden $2 \leq i \leq n$-re teljesüljön, hogy $(X_i|X_1X_2 \cdots X_{i-1})$ közel $\frac{1}{2}$, ami tehát azt is jelenti, hogy a tesztek egymástól minél inkább függetlenek legyenek, vagyis ne végezzünk olyan tesztet, amelyik már meglévő információt ad eredményül, hiszen ekkor lesz a leghatékonyabb a tesztelési eljárás.

Egy $p > 0$ valószínűséggel $\oplus$ minta információtartalma, vagy más szóval entrópiája a $-p \cdot \log_2 p - (1-p) \cdot \log_2(1-p)$ értékkel egyezik meg. Feltéve, hogy minden minta egymástól függetlenül ezzel a p valószínűséggel lesz $\oplus$, a tesztekből N-szer ennyi információt kell kinyernünk.

Tehát az információelméleti alsó határ a tesztek számára vonatkozóan:

$$(-p \cdot \log_2 p - (1-p) \cdot \log_2(1-p)) \cdot N \leq T. \qquad ([9])$$

Ahhoz, hogy alacsony p esetén ez a becslés éles legyen, nagyon sok mintát kellene összeönteni a csoporttesztekhez, azonban ha figyelembe vesszük, hogy bizonyos

számú minta összeöntése eredményezheti a fertőzöttség kimutathatatlanságát, korlátoznunk kell egy K pozitív egész számmal az összeönthető minták számát.Ekkor azonban, az alacsony fertőzöttségi arány miatt a csoporttesztek több, mint fele $\ominus$ lesz (tehát nem tudjuk elérni minden tesztelési csoport összeállításánál, hogy nagyjából $\frac{1}{2}$ valószínűséggel legyen benne fertőzött minta), ezért tudunk adni erősebb alsó becslést.

Méghozzá, $p < \frac{1}{K^2}$ esetén

$$\left(\frac{1-2p}{K} + 2p\right) \cdot N \leq T.$$

A $\frac{1}{K^2} < p < \frac{1}{K \cdot \log_2 K}$ valószínűség esetén pedig a

$$(-p \cdot \log_2 p - (1-p) \cdot \log_2 (1-p)) \leq t \cdot \left(-\frac{1-p}{Kt} \cdot \log_2 \frac{1-p}{Kt} - \left(1 - \frac{1-p}{Kt}\right) \cdot \log_2 \left(1 - \frac{1-p}{Kt}\right)\right)$$

becslést tudjuk adni, feltéve, hogy $2 \cdot \frac{1-p}{K} \geq \frac{T}{N} = t$. [3]

4. fejezet

A standard zajmentes csoporttesztelési modell

A legegyszerűbb modell, így a kiindulási probléma is, a következő: Van N darab elemünk, közülük ismeretlen számú hibás, azaz $\oplus$. A feladat, hogy minél gyorsabban (minél kevesebb lépésben) megállapítsuk, hogy pontosan mely elemek $\oplus$-ak. Tudjuk, hogy az elemek egymástól függetlenül, ugyanolyan p valószínűséggel $\oplus$-ak. Az elemek egy csoportjának a tesztelése $\oplus$ eredményt ad pontosan akkor, ha a csoportban volt legalább egy $\oplus$ elem, és $\ominus$-t ad, ha mindegyik elem $\ominus$ volt. A kérdés az, hogy mennyi a szükséges tesztek számának $E(T)$ várható értéke az összes $\oplus$ elem megtalálásához, a p és természetesen az N függvényében.

4.1. Mikor hatékonyabb a csoportos tesztelés?

Az elsőként természetesen felvetülő kérdések egyike, hogy milyen paraméterek mellett hatékonyabb a csoporttesztelés, mint az egyéni tesztelést. Ezt a kérdést Ungár vizsgálta, és válaszolta meg 1960-as cikkében [13], mégpedig a következő állítás belátásával:

4.1.1. Tétel. *Pontosan a* $0 \leq p < \frac{3-\sqrt{5}}{2}$ *valószínűségek mellett van olyan tesztelési algoritmus, amelynél a tesztek várható száma kisebb, mint* N*.*

Bizonyítás. Az, hogy ezeknél a p értékeknél optimálisabb a csoportos tesztelés, könnyen látható.

1. észrevétel:

$$\Phi_p(2) = min\{2, 1-q-q^2\}$$

Ugyanis, két lehetőségünk van. Az egyik, hogy egyesével leteszteljük a két mintát, ekkor $E(T) = 2$. A másik lehetőség, hogy először együtt teszteljük őket, és ha $\oplus$ eredményt kapunk, akkor egyesével is. Ekkor $E(T) = 3-q-q^2$, hiszen q valószínűséggel lesz $\ominus$ a második teszt, azaz nem kell elvégezni a harmadikat, és q^2 valószínűséggel már a második tesztre sincs szükség, mert a csoportos tesztelésnél $\ominus$ eredményt kaptunk.

2. észrevétel:

$$\Phi_p(n+m) \leq \Phi_p(n) + \Phi_p(m)$$

Ez nyilvánvaló.

Ebből következik, hogy ez a csoportos tesztelési eljárás a $1-q-q^2 < 2$ egyenlőtlenséget teljesítő $p = 1-q$ valószínűségek, azaz $p < \frac{3-\sqrt{5}}{2}$ esetén, ha kettes csoportokban teszteljük a mintákat, $E(T) < N$, azaz az egyesével tesztelés nem optimális.

Most lássuk be, hogy a $p > \frac{3-\sqrt{5}}{2}$ valószínűségekre az egyesével tesztelés az optimális.

Minden algoritmust reprezentálhatunk egy fagráffal, ahol a csúcsok minták csoportjait (G_i) jelölik, az élek pedig azt mutatják, hogy $\oplus$, illetve $\ominus$ eredmény esetén melyik csoportot teszteljük következőnek.

Ezekről a gráfokról feltételezhetjük a következőket, mivel mindegyik ilyen tulajdonságokkal rendelkezhetővé redukálható anélkül, hogy ennek következtében az $E(T)$ érték növekedne.

- nincs olyan G_i, amely kétszer is előfordul ugyanazon az ágon (semmi értelme olyan csoport tesztelésére időt pazarolni, aminek az eredményét már tudjuk)
- ha egy csoport teszteredménye $\ominus$, akkor az utána következő csoportok egyikében sem szerepel egyik eleme sem. (hiszen azokat a mintákat, amikről már tudjuk, hogy $\ominus$-ok, nincs értelme bármilyen csoportban újra tesztelni, semmilyen módon nem befolyásolják a teszteredményt)
- ha a gráfban van egy G_i csoport, akkor a belőle kiinduló ágakon nincs olyan csoport, ami tartalmazza (ha $G_i \oplus$, akkor az őt tartalmazó csoport is biztosan az, így annak tesztelésével nem nyerünk semmilyen új információt, ha $\ominus$, akkor ugyanúgy kivehető lenne az őt tartalmazó csoportból, mint az előző pontban)
- nincs olyan csúcs, amihez tartozó csoport teszteredménye már következik az előző tesztek eredményéből

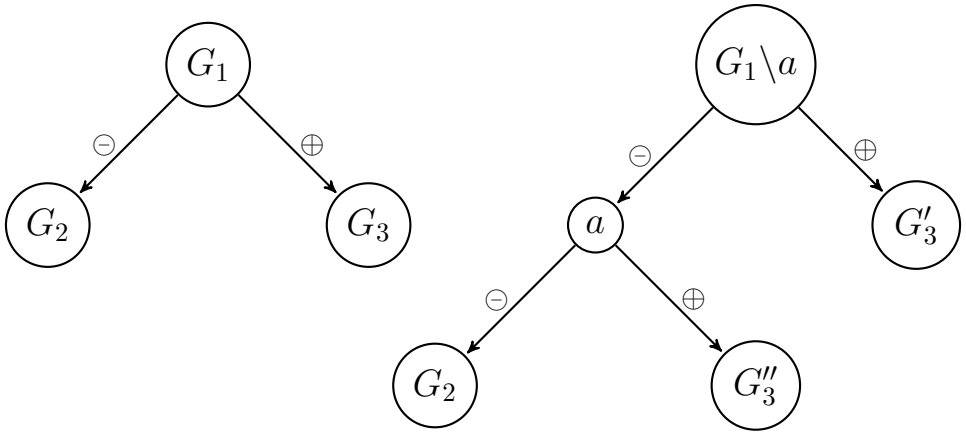
Válasszuk ki a gráf egyik csúcsát, aminek a G_1 csoport felel meg, ahol $|G_1| := m \geq 2$, és legyen G_2 az a csoport, amit a G_1 $\ominus$ eredménye esetén tesztelünk, G_3 pedig, amit $\oplus$ eredmény után.

Módosítsuk a tesztelési gráfot a következőképpen: A G_1 csoport helyett teszteljük a $G_1' = G_1 \backslash a$ csoportot, ahol $a \in G_1$ egy tetszőleges minta. Ha ennek a csoportnak a teszteredménye $\oplus$, akkor ha ennek az ágnak a tesztelését változatlanul az eredeti terv szerint folytatjuk, ugyanúgy pontosan meg tudjuk határozni a $\oplus$ elemeket, hiszen ebben az esetben a G_1 teszteredménye is $\oplus$ lett volna. Ha viszont a G_1' eredménye $\ominus$, akkor végezzünk egyéni tesztet a-n.

- ha $\ominus$, akkor azzal az ággal folytatjuk, ami az eredetiben a G_2-vel kezdődik
- ha $\oplus$, akkor ha az eredetiben G_3-mal kezdődő ággal folytatjuk, pontos eredményt kapunk, hiszen ebben az esetben G_1 is $\oplus$ lett volna

A gráf többi részét érintetlenül hagyjuk.

Vegyük észre, hogy a módosított eljárás esetén a G_3 ágon illetve mikor $G_1 \backslash a$ $\ominus$ és a $\oplus$ volt, akkor extra információnk van az eredeti eljáráshoz képest, aminek birtokában lehet hogy végre tudunk hajtani néhányat a fentebbi egyszerűsítő lépésekből. Ha igen, akkor ezeket hajtsuk végre, így kapjuk a G_3' (ha G_1' $\oplus$), és G_3'' (amikor G_1' $\ominus$ lett) csúcsokat.



Az is nyilvánvaló, hogy az új gráfban minden tesztelési ág legfeljebb eggyel hosszabb.

A bizonyítás befejezéséhez azokhoz a tesztelési szituációkhoz, amikhez a módosítással 1-gyel nőtt a tesztek száma, rendeljünk két olyat, amiknél 1-gyel csökkent.

4.1.2. Állítás. *Csak akkor kellhet több teszt, ha az eredeti gráfban eljutottunk a G_1 csoport teszteléséig, és a teszteredmény $\ominus$ lett.*

Bizonyítás. Ha $G_1 \backslash a$ eredménye $\oplus$, akkor az új gráf szerint maximum annyi teszt kell, mint az eredeti szerint, hiszen ugyanazt csinálnánk, mint amit az eredeti szerint, kivéve, ha volt egy csak a-ból álló teszt, mert akkor azt elhagyhatjuk.

Fontos észrevétel, hogy egy $\ominus$ minta megállapításához kell egy olyan őt tartalmazó csoportot tesztelnünk, aminek az eredménye $\ominus$, egy $\oplus$ minta megállapításához pedig kell legalább egy olyan teszt, ami csak ennek a mintának a megállapí-

tásához járul hozzá, tehát ennek a tesztnek az elhagyásával ezen a mintán kívül mindegyiknek az eredményét tudni fogjuk, ezét nem.

Ha $G_1 \backslash a$ $\ominus$, és a $\oplus$, akkor eggyel több tesztet végeztünk, mint az eredetiben, és a G_3'' ággal folytatjuk. Azonban az eredeti algoritmusban a $G_1 \backslash a$ csoport $(m-1)$ elemének eredményét ezután $(m-1)$ egyedi teszttel kellett megállapítani, az új tervben ezekre nincs szükség, tehát megtakarítunk legalább $(m-1) \geq 1$ tesztet, vagyis semmiképp sem csináltunk több tesztet.

Emiatt az eredeti tervhez képest csak akkor csinálhatunk több tesztet, ha $G_1 \backslash a$ és a tesztje is $\ominus$, azaz G_1 negatív.

□

Legyen S egy olyan mintahalmaz, amelyben van olyan elem, aminek teszteredményének meghatározásához eggyel több teszt szükséges. Ha $G_1 = \{a, b, c, \cdots\}$, akkor $S_{a,b}$ legyen az a mintahalmaz, amelyet az S-ből kapunk az a és b hibássá változtatásával.

Észrevehetjük, hogy amikor $S_{a,b}$-t teszteljük, akkor szintén eljutunk a G_1 csúcshoz. Valóban, amikor elérünk S teszteléséhez, akkor G_1 egyetlen eleme sem fordulhatott elő egy korábbi $\ominus$ eredményű mintahalmazban. Így G néhány elemének hibásra változtatása nem változtatja meg a G_1-t megelőző teszteredményeket. Ugyanemiatt, amikor $S_{a,b}$-t teszteljük, a régi, vagy az új terv szerint, akkor nem az a, b minták lesznek az egyedüli hibásak a G_1-et megelőző tesztelt halmazokban. Így S, vagy $S_{a,b}$ elemeinek hibásságát a G_1, illetve a $G_1 \backslash a$, és az ezeket követő egyedi tesztelések eredményeiből kell megállapítani.

Vizsgáljuk az $m = 2$, és az $m \geq 3$ esetet külön.

1. eset: Amikor $m = 2$, legyen $G = \{a, b\}$. A fentebb meghatározott S esetében a következő két mintának eggyel kevesebb tesztre lesz szüksége.

$S_{a,b}$ esetén a fentebb írtak szerint a és b hibásságát egyedi tesztekkel kell megállapítani a régi terv szerint a G_1 tesztelése után. Az új terv szerint a G_1 tesztelése

helyett $b = G_1 \backslash a$-t teszteljük, így kihagyhatunk egy tesztet.

Ezután jelölje c az a, b minták közül azt, amelyiket először teszteljük, amikor a régi terv szerint teszteljük $S_{a,b}$-t (ez S-től függhet). Ez a G_1 első egyedileg tesztelt eleme is, amikor S_c-t a régi terv szerint teszteljük. A G_1 másik elemének hibásságát egy másik egyedi teszttel kell megállapítani. Az új terv szerint ezek közül az egyedi tesztek közül az egyik kihagyható a G_3' irányban, ha $c = b$, illetve mindkettő kihagyható G_3'' irányban, ha $c = a$, ugyanis a-t és b-t is teszteltük korábban, mielőtt G_3'' ágra léptünk. Így mindkét esetben egy tesztet spóroltunk.

A megspórolt tesztek számának várható értéke $m = 2$ esetben:

$$\sum_S P(S_{a,b}) + P(S_c) - P(S) = \sum_S P(S) \cdot \left(\left(\frac{p}{q} \right)^2 + \frac{p}{q} - 1 \right)$$

ahol S összegezve van minden mintán, ami minden korábban említett feltételnek megfelel. Csak annyit kell megjegyezni, hogy két különböző S-ből kapott $S_{a,b}$ és S_c is különböző, és a jobb oldal pozitív $p > \frac{3-\sqrt{5}}{2}$ esetén, mindaddig, amíg van olyan minta, aminek tesztelése a G_1-hez vezet. Ilyen minta létezését az a feltétel biztosítja, hogy nincs olyan csúcs, aminek eredménye következik az előző tesztekből.

2. eset: $m \geq 3$ esetén az új tesztelési tervvel legalább $m-2$ tesztet spórolunk. Valóban, a régi terv szerint $m-1$ egyedi tesztre van szükség a G_3 ágán a $G_1 \backslash a$ elemeinek hibásságának megállapításához. Ezek a tesztek az új terv alapján mind elhagyhatók. Másrészt, az új tervben van egy egyedi tesztelése a-nak, ami a régi tervben nem biztos, hogy szerepelt.

Legyen b a $G_1 \backslash a$ utolsó tesztelendő eleme, amikor a régi tervet alkalmazzuk S_a-ra. (Hogy melyik a b, az itt is S-től függhet.) Ez a $G_1 \backslash a$ utolsó tesztelendő eleme is lesz, amikor az $S_{a,b}$-t teszteljük a régi terv alapján, mivel az összes korábbi teszteredmény ugyanaz S_a-ra és $S_{a,b}$-re. Amikor $S_{a,b}$-t teszteljük az új terv alapján, akkor a b tesztelését kihagyhatjuk, amikor $G_1 \backslash a \oplus$, de minden korábbi tesztelt

eleme $\ominus$, akkor tesztelés nélkül is tudjuk, hogy b csak $\oplus$ lehet.

Tehát a megspórolt tesztek száma $m \geq 3$ esetben:

$$\sum_S P(S_{a,b}) + (m-2) \cdot P(S_a) - P(S) \geq \sum_S P(S) \cdot \left(\left(\frac{p}{q} \right)^2 + (m-2) \cdot \frac{p}{q} - 1 \right)$$

ami > 0, hogyha $p \geq \frac{3-\sqrt{5}}{2}$ □

4.2. Dorfman-típusú algoritmus

Dorfman a legegyszerűbb, standard zajmentes modellből kiindulva, a következő eljárást javasolta a témát megalapozó cikkében [4]:

4.2.1. Algoritmus. *A beérkezett mintákat rendezzük egymástól diszjunkt, k elemű csoportokba, és a k elemet összeöntve teszteljük. Ha az eredmény $\ominus$, akkor biztosak lehetünk abban, hogy a csoportban lévő minta mindegyike $\ominus$. Ha viszont $\oplus$, akkor teszteljük le mind a k elemet, egyesével.*

Ez a következők miatt lesz gyorsabb, mint az egyesével tesztelés. Mivel feltettük, hogy minden minta egymástól függetlenül p valószínűséggel $\oplus$ eredményű, ezért egy minta $q = (1-p)$ valószínűséggel $\ominus$, tehát egy k elemű mintahalmaz q^k valószínűséggel $\ominus$, így $p' := 1-q^k$ valószínűséggel $\oplus$. A minták számát N-nel jelölve a k elemű csoportok száma $\frac{N}{k}$, így a $\oplus$ eredményű csoportok várható száma $p' \cdot \frac{N}{k}$.

Tehát:

$$E(T) = \frac{N}{k} + k \cdot p' \cdot \frac{N}{k}$$

vagyis a csoportos tesztek száma, plusz a $\oplus$ eredményű csoportok elemeinek száma, amiket aztán egyesével kell tesztelni.

Ahhoz, hogy a k és p függvényében ez hatékonyabb legyen, mint az egyéni tesztelés, $E(T) \leq N$-nek kell teljesülnie, azaz

$$k \cdot q^k = k \cdot (1-p)^k \geq 1.$$

Optimális csoportméretek a p függvényében		
p	optimális k	hatékonyság $\left(\frac{E(T)}{N}\right)$
0.001	32	0.06
0.003	19	0.11
0.005	15	0.14
0.007	12	0.16
0.01	11	0.2
0.02	8	0.27
0.03	6	0.33
0.04	6	0.38
0.05	5	0.43
0.06	5	0.47
0.07	5	0.5
0.08	4	0.53
0.09	4	0.56
0.1	4	0.59
0.12	4	0.65
0.13	3	0.67
0.15	3	0.72
0.2	3	0.82
0.25	3	0.91
0.3	2	0.99

Tegyük hozzá, hogy itt a tesztek számának várható értékének kiszámításánál nem vettük figyelembe azt az esetet, amikor k elemű csoport teszteredménye $\oplus$, és az egyesével tesztelés során az első $(k-1)$ minta teszteredménye $\ominus$. Ekkor ugyanis nincs szükség az utolsó minta tesztelésére, hiszen az mindenképp $\oplus$. Tehát $p \cdot q^{k-1}$ valószínűséggel k teszt helyett $(k-1)$ teszt is elegendő az ebben a csoportban lévő

hibás elemek megtalálásához. [11]

Tehát, ha ily módon, kicsit okosabban csináljuk,

$$E(T) = \frac{N}{k} + \left(k \cdot p' - p \cdot q^{k-1}\right) \cdot \frac{N}{k}$$

Könnyen meggondolható[5], hogy ezzel a tesztelési eljárással a várható tesztszám $2\sqrt{Nd}$. Az algoritmus könnyen fejleszthető, mégpedig úgy, hogyha az első tesztelési kör után nem teszteljük le azonnal egyesével a $\ominus$ eredményű csoportokban lévő mintákat, hanem ezekre ismét elvégezzük az algoritmust. Az így kapott c-körös eljárás során a várható tesztszám $T_c = c\sqrt[c]{Nd^{c-1}}$.

Jelölje k_c azt a csoportméretet, amely optimális abban az esetben, ha $d \oplus$ elem van a halmazban, és c körben akarjuk ezeket megkeresni.

Ekkor

$$k_c = \sqrt[c]{\left(\frac{d}{N}\right)^{c-1}}$$

Az alábbi táblázat szemlélteti a c-körös tesztelési eljárás hatékonyságát, a megfelelő k_c csoportméretek választásával. A $\frac{d}{N}$-t a gyakorlatban megfeleltethetjük a p valószínűségnek.

$\frac{d}{N}$	1	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
k_3	1	1.59	2.92	4.64	7.37	13.6	21.5	34.2	63	100
$\frac{T_3}{N}$	3	1.89	1.03	0.647	0.407	0.221	0.139	0.0878	1.0477	0.03
k_4	1	1.68	3.34	5.62	9.46	18.8	31.6	53.2	106	178
$\frac{T_4}{N}$	4	2.38	1.2	0.712	0.423	0.213	0.126	0.0753	0.0379	0.0225
k_5	1	1.74	3.62	6.31	11	22.9	39.8	69.3	144	251
$\frac{T_5}{N}$	5	2.87	1.38	0.793	0.454	0.219	0.125	0.0722	0.0347	0.0134
k_6	1	1.78	3.82	6.81	12.1	26	46.4	82.7	178	316
$\frac{T_6}{N}$	6	3.37	1.57	0.882	0.493	0.231	0.129	0.0726	0.0338	0.019

$\frac{d}{N}$	1	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
k_7	1	1.81	3.97	7.2	138	28.6	51.8	93.8	206	373
$\frac{T_7}{N}$	7	3.86	1.76	0.973	0.538	0.245	0.135	0.0746	0.034	0.0188
k_8	1	1.83	4.09	7.5	13.8	30.7	56.2	103	230	422
$\frac{T_8}{N}$	8	4.37	1.95	1.07	0.582	0.251	0.142	0.0776	0.0348	0.019
k_9	1	1.85	4.18	7.74	14.3	32.4	60	111	251	464
$\frac{T_9}{N}$	9	4.86	2.15	1.16	0.627	0.278	0.15	0.0811	0.0359	0.0194
k_{10}	1	1.87	4.26	7.94	14.8	33.8	63.1	118	269	501
$\frac{T_{10}}{N}$	10	5.37	2.35	1.26	0.676	0.296	0.159	0.0849	0.0372	0.02

$\frac{d}{N}$	0.0005	0.0002	0.0001	0.00005	0.00002	000001
k_3	159	292	464	737	1360	2150
$\frac{T_3}{N}$	0.0189	0.0103	0.00647	0.00407	0.00221	0.00139
k_4	299	595	1000	1680	3340	5620
$\frac{T_4}{N}$	0.0134	0.00673	0.004	0.00238	0.0012	0.000712
k_5	437	910	1590	2760	5740	10 000
$\frac{T_5}{N}$	0.0114	0.00548	0.00316	0.00181	0.000872	0.0005
k_6	564	1210	2150	3840	8240	14 700
$\frac{T_6}{N}$	0.0106	0.00497	0.0028	0.00156	0.000729	0.000408
k_7	675	1480	2680	4860	10 700	19 300
$\frac{T_7}{N}$	0.0104	0.00463	0.00252	0.00138	0.000618	0.000337
k_8	773	1720	3160	5800	12900	23700
$\frac{T_8}{N}$	0.0103	0.00463	0.00252	0.00138	0.000618	0.000337
k_9	860	1940	3590	6660	15 000	27 800
$\frac{T_9}{N}$	0.0105	0.00463	0.0025	0.00135	0.000598	0.000327
k_{10}	935	2130	3980	7430	17 000	31 600
$\frac{T_{10}}{N}$	0.0107	0.00469	0.00251	0.00135	0.000590	0.000317

A táblázat alapján jól látható az is, hogy a körök számának növelésével egy idő után már nem tudunk lényegesen javítani a hatékonyságon.

A k_c értékek csak iránymutatást adnak, hogy mely számhoz közel kell keresni a megfelelő csoportméretet, amelyik nem feltétlenül a k_c-hez legközelebbi egész számot jelenti. Például az $N = 100$, $d = 1$, $c = 4$ esetben, a táblázatban a $k_4 = 31.6$ érték szerepel, ami a gyakorlatban azt jelenti, hogy az első körben $33, 33$ és 34 mintából álló csoportokat tesztelünk.

A tesztelési terv egy másik hatékonyabbá tétele a következő.[12]

4.2.2. Algoritmus. *Első körben osszuk a mintákat k elemű csoportokba, és teszteljük le ezeket. Ha egy csoport tesztelése során $\oplus$ eredményt kapunk, kezdjük el egyesével tesztelni benne lévő mintákat, amíg nem találunk egy $\oplus$ eredményűt. Ekkor a maradékot végezzünk egy ellenőrzőtesztet.*

Kis p esetén nagy valószínűséggel lesz ennek az ellenőrző tesztnek $\ominus$ az eredménye, tehát ezeket a mintákat már nem kell egyesével letesztelni. Az algoritmus hatékonyságát a következő egyszerű állítás segítségével tesszük láthatóvá.

4.2.3. Állítás. *Jelöljük a k mintából ezzel az eljárással, az n db $\oplus$ meghatározásához szükséges tesztek várható értékét $E_k(n)$-vel. Ekkor*

$$E_k(n) = \frac{n}{n+1} \cdot (k+1) + n + 1 - 2n \cdot \frac{1}{k}$$

Bizonyítás. $E_k(0) = 1$ nyilvánvaló, és a fenti egyenlőséget is teljesíti.

Innen rekurzióval adódik az $E_k(n)$ értéke. Tegyük fel, hogy az $E_a(b)$ értékeket már minden $a \leq k$ és $b \leq n$-re meghatároztuk.

Az első körben 1 tesztet végzünk, a k mintát tartalmazó csoporton. Ezután, mivel az eredménye $n \geq 1$ esetén $\oplus$, elkezdjük egyesével tesztelni az elemeket. Ekkor az első tesztelt minta $\frac{n}{k}$ valószínűséggel lesz $\oplus$. Ebben az esetben a maradék $(n-1)$ $\oplus$ mintát $(k-1)$ elem közük kell meghatároznunk, amihez a várható tesztszámot, $E_{k-1}(n-1)$-t már tudjuk.

Tegyük fel, hogy az egyéni tesztelések során az elsőként $\oplus$ eredményű a $(j+1)$. Annak a valószínűsége, hogy az első j minta egyéni tesztelésének eredménye $\ominus$, $\prod_{i=1}^{j} \frac{k-(i+n-1)}{k-(i-1)}$, azé, hogy a $(j+1)$. $\oplus$, pedig $\frac{n}{k-j}$. Innen $k-(j+1)$ minta közül kell $(n-1)$-et megtalálnunk, ehhez várhatóan $E_{k-(j+1)}(n-1)$ tesztre lesz szükségünk.

Tehát összességében,

$$E_k(n) = 1+\frac{n}{k}\left(1+E_{k-1}(n-1)\right)+\sum_{j=1}^{k-n}\left(\left(\prod_{i=1}^{j}\frac{k-(i+n-1)}{k-(i-1)}\right)\cdot\frac{n}{k-j}\cdot\left((j+1)+E_{k-(j+1)}(n-1)\right)\right)$$

Ebből indukcióval adódik az állítás. □

Tehát, a tesztek számának várható értéke, ha az N mintát k méretű csoportokban teszteljük az első körben, megközelítőleg

$$E'_{p,N}(T) = \frac{N}{k}\cdot\sum_{i=0}^{m}\left(P_k(i)\cdot E_k(i)\right),$$

ahol $P_k(i)$ jelöli annak a valószínűségét, hogy a k minta közül pontosan i $\oplus$, illetve $m = argmin_m \sum_{i=0}^{m} P_k(i) \geq 0,99$. (Ekkor a hiba kisebb, mint $\frac{2-\frac{1}{k}}{100}\cdot N$.)

Ez az eljárás minden esetben hatékonyabb, mint Dorfman eredeti algoritmusa ha hatékonyabb csoportosan tesztelni az egyéni helyett, ezt az alábbi táblázat is jól szemlélteti.

Optimális csoportméretek a p függvényében			
p	optimális k	hatékonyság$\left(\frac{E(T)}{N}\right)$	Dorfman-féle hatékonyság
0.001	47	0.04	0.06
0.003	30	0.08	0.11
0.005	22	0.1	0.14
0.007	20	0.12	0.16
0.01	16	0.14	0.2
0.02	11	0.22	0.27
0.03	9	0.27	0.33
0.04	8	0.32	0.38
0.05	7	0.35	0.43
0.06	7	0.39	0.47
0.07	6	0.42	0.5
0.08	6	0.45	0.53
0.09	5	0.48	0.56
0.1	5	0.51	0.59
0.11	4	0.54	0.62
0.12	4	0.57	0.65
0.13	4	0.59	0.67
0.14	4	0.61	0.7
0.15	4	0.65	0.72
0.2	3	0.74	0.82
0.23	3	0.8	0.87
0.25	3	0.84	0.91
0.26	3	0.86	0.93
0.27	2	0.87	0.96
0.3	2	0.9	0.99

4.3. Bináris keresés

4.3.1. Algoritmus. *Első körben leteszteljük együtt, egy csoportban az N mintát. Ha az eredmény $\ominus$, akkor minden elem $\ominus$, tehát nem kell tovább tesztelni. Ellenkező esetben bontsuk fel a halmazt két, nagyjából $\frac{N}{2}$ elemű, diszjunkt halmazra, B_1-re és B_2-re. Ezeket teszteljük ugyanilyen módon.*

Az eljárás elképzelhető egy fagráfként, aminek minden csúcsának (a kezdőcsúcson és a végeken kívül) három szomszédja van, és minden csúcshoz tartozó halmaz a lentebbi szomszédaihoz tartozó halmazok diszjunkt uniója. Ha egy ágon eljutunk egy $\ominus$ eredményű, vagy egy egyelemű halmazhoz, azt az ágat nem folytatjuk tovább.

Világos, hogy ezzel az eljárással $\lceil \log_2 N \rceil$, vagy $\lceil \log_2 N \rceil - 1$ lépés alatt található meg egy $\oplus$ elem (a tesztelési eljárást gráfja $\lceil \log_2 N \rceil$ szintű, és egy mintáról akkor állapítható meg, hogy $\ominus$, ha egyedül teszteltük, azaz eljutottunk a legalsó szintre).

Hogyha több gép áll rendelkezésünkre, azaz egyszerre több tesztet is tudunk végezni, és az egy teszt elvégzéséhez szükséges időt t-vel jelöljük, akkor ha a minták között d db $\oplus$ van, akkor ezt az algoritmust alkalmazva az összes $\oplus$ mintát megtaláljuk $(\lceil \log_2 N \rceil) \cdot t$ idő alatt. Ehhez $2 \cdot d$ gép már elegendő, hoszen egyszerre maximum d csoportban van $\oplus$ elem, és mindig az előző tesztkörben $\oplus$ eredményt adó halmazokat szedjük két részre, és teszteljük.

Ha egy időben csak egy csoportot tudunk tesztelni, és a minták között d db $\oplus$ van, akkor ezt az algoritmust alkalmazva összesen $d \cdot \lceil \log_2 N \rceil + O(d)$ db teszttel meg tudjuk találni mindegyik $\oplus$ elemet.

4.4. Sobell és Groll javaslata

Az előzőhöz hasonló elvű algoritmusokat vizsgált Sobel és Groll [11], [10], azzal a lényeges különbséggel, hogy a megkapott N elemű mintahalmazt egy tesztkörben

csak két részre bontották, így egy időben csak egy csoporttesztet végeztek, mert a következő tesztelendő csoport méretét és tartalmát az előző alapján választották. Ezek közül én most csak a legegyszerűbbet ismertetem.

4.4.1. Algoritmus. *Az algoritmus során mindig egy ismeretlen eredményű mintákból álló csoporttesztet végzünk, aminek méretét az alapján választjuk ki, hogy az eddigiek alapján ezzel a választással az elvégzett összes teszt várható értéke a lehető legalacsonyabb legyen.*

A tesztelési protokoll megállapítása nagyon sok számolást igényel, így csak a kis N-re való vizsgálatát mutatom be, de az elv alapján ki lehet számolni nagyobb N-ekre is.

Kezdjük néhány észrevétellel. Annak a valószínűsége, hogy az m elemű $\oplus$ halmazban pontosan y db $\oplus$ minta van, feltéve, hogy van benne legalább egy:

$$P(d_m = y | d_m \geq 1) = \frac{\binom{m}{y} p^y q^{m-y}}{1 - q^m} \tag{4.1}$$

Annak a valószínűsége, hogy az m elemű halmazból random kiválasztott x elemű halmazban nincsen $\oplus$ minta:

$$P(d_x = 0 | d_m \geq 1) = \sum_{y=1}^{m-x} \frac{\binom{m}{y} p^y q^{m-y}}{1 - q^m} \cdot \frac{\binom{m-y}{x}}{\binom{m}{x}} = \frac{q^x - q^m}{1 - q^m} \tag{4.2}$$

4.4.2. Lemma. *Legyen A és B a minták egy-egy, egymástól diszjunkt részhalmaza. Ha $A \cup B$-t, és A-t tesztelve is $\oplus$ eredményt kapunk, akkor az összes mintához tartozó feltételes eloszlás B-ben pontosan annyi, mint az eredeti binomiális eloszlás.*

Bizonyítás. Jelöljük d_A-val és d_B-vel a két halmazban lévő $\oplus$ elemek lehetséges számát. Ekkor

$$P(d_B \leq b | d_A + d_B \geq 1, d_A \geq 1) = P(d_B \leq b | d_A \geq 1) = P(d_B \leq b),$$

hiszen ha $d_A \geq 1$, akkor $d_A + d_B \geq 1$ mindenképp teljesül, illetve A és B diszjunktak, így d_A és d_B függetlenek. □

Jelöljük $G(m,n)$-nel a tesztek számának várható értékének minimumát (tehát a várható értéket, ha a következő tesztcsoportot okosan választjuk), ha az n mintából kiválasztott $\oplus$ halmaz mérete m, és a minták egymástól függetlenül p valószínűséggel $\oplus$-ak. Jelöljük el külön azt az esetet, amikor $m = 0$, méghozzá $H(n)$-nel. A protokoll megállapítása során az m és n értékek folyamatosan változnak, és a kiindulási értékek $m = 0$ és $n = N$. A tesztelési protokollt rekurzívan fogjuk megállapítani.

Ha az x jelöli a következő tesztcsoport méretét, akkor az $m = 0$ esetben

$$H(n) = 1 + \min_{1 \le x \le n} \{q^x H(n-x) + (1 - q^x) G(x,n)\} \tag{4.3}$$

az $n \ge m \ge 2$ esetben pedig a (4.2) egyenlet és a 4.4.2lemma alapján

$$G(m,n) = 1 + \min_{1 \le x \le m-1} \left\{ \left(\frac{q^x - q^m}{1 - q^m} \right) G(m-x, n-x) + \left(\frac{1 - q^x}{1 - q^m} \right) G(x,n) \right\} \tag{4.4}$$

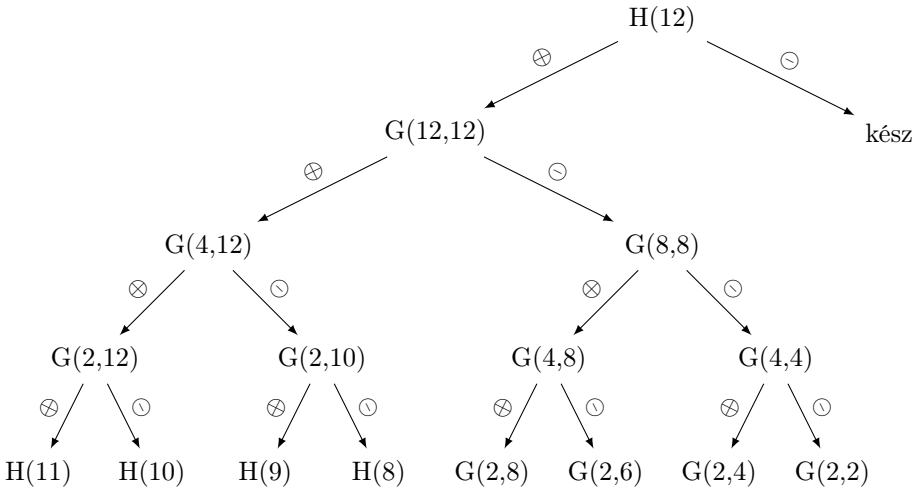
lesz a tesztek számának a várható értéke, az x méretű halmaz optimális választásával. A képletekben a konstans 1 mindig az x méretű következő halmaz tesztjét jelzi, amit úgy választunk ki, hogy minimális legyen a többi minta megállapításához szükséges tesztek számának várható értéke.

$$H(0) = 0 \text{ és } G(1,n) = H(n-1) \tag{4.5}$$

Ezekből az adatokból meg tudjuk állapítani az optimális tesztelési protokollt, hiszen a minimális várható értékek a következő sorrendben rekurzívan következnek egymásból:

$$H(0) = G(1,2), G(2,2), H(2) = G(1,3), G(2,3), G(3,3), H(3) = G(1,4), G(2,4), \cdots$$

Például, az $N = 12$, és $p = 0,02$ kiindulási értékekkel a kezdő x érték a 12, és az optimális tesztelési gráf első fele a következőképpen néz ki.



Az optimális x értékek pedig a következők:

Tesztszám várható értéke	optimális x
H(12)	12
G(12,12)	4
G(4,12)	2
G(8,8)	4
G(2,12)	1
G(2,10)	1
G(4,8)	2
G(4,4)	2
H(11)	11
H(10)	10
H(9)	9
H(8)	8
G(2,8)	1
G(2,6)	1
G(2,4)	1
G(2,2)	1

Kiszámolható, hogy ebben az esetben a tesztek várható értéke $H(12) = 2,07$, sőt, $0,7847$ valószínűséggel nincs is a minták között $\oplus$, azaz 1 teszt elég.

Az elvégzendő tesztek maximális számára vonatkozóan nagyon kis p esetén megmutatható, hogy

$$M(n) = (n+1)(\alpha(n)+1) + 1 - 2^{1+\alpha(n)}, \text{ és}$$

$$M(m,n) = \alpha(m) + n(\alpha(n-1)+1) + 1 - 2^{1+\alpha(n-1)},$$

ahol $2^{\alpha(z)} \leq z \leq 2^{\alpha(z+1)}$.

Ezt a maximumot azonban csak akkor érjük el, ha p értéke nagyon közel van a 0-hoz, és mind az n db ismeretlen minta értéke $\oplus$.

Az $N = 12$ esetben például $\alpha(12) = 3$, így $M(12) = 37$. Ebből is jól látszik, miért nem érdemes a tesztek maximális számára optimalizálni. Ebben a konkrét esetben például az egyéni teszteléshez tartozó maximális tesztszám (12) szignifikánsan kevesebb, mint a fentebb bemutatott algoritmus esetében, így az erre való optimalizálás esetében elvetnénk ezt a tesztelési lehetőséget, pedig várható értékben lényegesen jobban járunk vele.

A szerzőpáros egy későbbi cikkben azt is megmutatta, hogy az eljárásuk ismeretlen p mellett is jól működik.[10]

5. fejezet

A nem bináris modell

A valóságban a teszteléshez használt eszközök a kórokozók jelenlétének mértékét jelzik, tehát nem bináris eredményt adnak. Az optimális tesztelési algoritmusok keresése során ezt a tényt érdemes kihasználni, hiszen nem bináris visszajelzés esetén több információ áll a rendelkezésünkre. Ezt az alábbiakban röviden bemutatom.

5.0.1. Definíció. [8] A nem bináris modellben csoportos tesztelés során az egyes csoporttesztek eredményei $[0, \infty)$ halmazba eső számok, oly módon, hogy a csoport teszteredménye a csoportba tartozó mintákhoz tartozó értékek összege, és minden nem fertőzött mintához (a korábbi modellben $\ominus$) tartozó érték a 0.

A modell egyik nagy előnye, hogyha egy $0 \leq n$ elemű csoport teszteredménye $x_n > 0$, akkor a benne lévő mintákat elég addig tesztelni, amíg azok eredményeinek összege el nem éri az x_n-et. Így például az utolsó mintát sosem kell tesztelni, hiszen a teszteredménye $x_n - x_{n-1}$, ahol x_{n-1} a többi minta teszteredményének összege.

Most tekintsük a következő módosítását a Dorfman-típusú algoritmusnak.

5.0.2. Algoritmus. *A beérkezett N darab mintát egymástól diszjunkt, $\{n_1, \cdots n_l\}$ elemű csoportokba rendezzük, és az így kapott csoportok elemeit összeöntve tesz-*

teljük. Ha az eredmény 0*, akkor biztosak lehetünk abban, hogy a csoportban lévő minta mindegyikének* 0 *az értéke. Ha viszont egy pozitív számot kapunk vissza, akkor teszteljük le a csoport* $n_i - 1$ *elemét egyesével.*

Jelöljük T_k-val a k méretű csoportban lévő elemek értékének meghatározásához szükséges tesztszámot Ekkor

$$P(T_k = j) = \begin{cases} q^k, & \text{ha } j = 1 \\ pq^{k-j+1}, & \text{ha } 2 \leq j \leq k-1 \\ 1 - q^2, & \text{ha } j = k \end{cases}$$

Tehát

$$E(T_k) = k - \frac{q^2\left(1-q^{k-1}\right)}{p}, \tag{5.1}$$

amiből világosan látszik, hogy minden p valószínűség esetén optimálisabb ez az eljárás az egyéni tesztelésnél. Tehát ha a használt műszer kimutatja a fertőzés jelenlétének mértékét, érdemes azt kihasználni, mert több tesztet spórolhatunk vele.

Az $\{n_1, n_2 \cdots n_l\}$ optimális megválasztásához minimalizálni kell a $\sum_{i=1}^{l} n_i = 1$ és $\forall i\ n_i \geq 0$ feltételek mellett az $E(T_N) = \sum_{i=1}^{l} E(T_{n_i})$ összeget.

A következő jelölés bevezetésével rekurzívan meg tudjuk állapítani az N minta optimális felosztását.

$$H(T_k) = \min_{1 \leq n \leq k} \{H(T_{k-n}) + E(T_n)\} \text{ahol } H(T_0) = 0$$

. Összehasonlításképp, néhány konkrét érték:

p	T_k	$H_D(T_k)$	$H_M(T_k)$
0.25	25	21.2	16.8
0.1	50	28.8	22.1
0.01	100	19.5	14.3

$H_D(T_k)$: Dorfman-típusú algoritmus szerinti, $H_M(T_k)$: eszerint az algoritmus szerinti minimális várható érték.

6. fejezet

Nemadaptív algoritmusok

Ebben a fejezetben olyan zajmentes nemadaptív algoritmusokat mutatok be, amik nem feltétlenül pontosan mondják meg, mely minták $\oplus$-ak, illetve $\ominus$-ak, viszont gyorsak, és nagy valószínűséggel helyesen találják el, hogy mely minták $\oplus$-ak, illetve $\ominus$-ak.

Mindegyiknél veszünk egy $\mathbf{X}$ tesztelési mátrixot a Bernoulli modell alapján, és a kapott $\mathbf{y}$ eredményvektor alapján próbáljuk dekódolni, a K halmazt minél pontosabban meghatározni.

6.1. COMP

Ez az algoritmus volt az első gyakorlati tesztelési algoritmus [2], [1]. Az eljárás abból a megfigyelésből indul ki, hogy $\ominus$ eredmény esetén biztosak lehetünk abban, hogy a halmazbeli elemek mind $\ominus$-ok. Ekkor felteszi, hogy a többi elem $\oplus$.

6.1.1. Algoritmus. *[2] A Bernoulli modell alapján sorsolunk egy* $\mathbf{X}$ *tesztelési mátrixot, és elvégezzük a tesztköröket.*

A tesztek elvégzése előtt minden mintát „talán $\oplus$*”-ként jelölünk meg. Ha egy minta benne van egy* $\ominus$ *eredményű halmazban, a jelölést „biztosan* $\ominus$*”-ra változtat-*

juk. A tesztelési sorozat végén a „talán ⊕" jelű mintákat jelöljük meg ⊕-ként.

6.1.2. Lemma. *Az így kapott K_{COMP} halmaz kielégítő halmaz, és tartalmaz olyan mintákat, amik igazából ⊖-ok, valamint K_{COMP} tartalmaz minden kielégítő halmazt tartalmaz, így K_{COMP} a legnagyobb kielégítő halmaz.*

Bizonyítás. Mivel egy elem csak akkor lesz a K_{COMP} halmazban, ha nem szerepel egyetlen ⊖ eredményű tesztelt halmazban sem, ezért nyilván nem lesznek benne hamis ⊖-ok, illetve az is világos, hogy $K \subseteq K_{COMP}$, hiszen a K-beli elemekről nem kerülhetett le a „talán ⊕" jelölés, tehát K_{COMP} valóban kielégítő halmaz.

Vegyünk egy L kielégítő halmazt, illetve egy $i \notin K_{COMP}$ elemet. Ekkor i benne van egy olyan tesztelt halmazban, aminek ⊖ lett az eredménye, tehát $i \in L$ nem lehetséges. Azaz $L \subseteq K_{COMP}$ □

A COMP algoritmus szuboptimális, ami nem meglepő, hiszen nem használja ki a ⊕ teszteredményeket.

6.1.3. Tétel. *Tegyük fel, hogy egy nemadaptív, zajmentes csoporttesztelés eredménye adott* $\mathbf{y}$*, amit a COMP eljárással akarjuk dekódolni. A Bernoulli modellben, egy optimalizált $p = \frac{1}{d}$ paraméter, és $d = \Theta(n^\alpha)$, $\alpha \in (0,1)$, mellett az $R = \frac{1}{e \ln 2}(1-\alpha)$ arányra teljesülni fog, hogy $\forall \delta, \varepsilon$-ra létezik olyan csoporttesztelési algoritmus, amelyre*

$$\frac{\log_2 \binom{n}{d}}{T} > R - \delta$$

és a hibavalószínűség legfeljebb ε.

Bizonyítás. Annak a valószínűsége, hogy egy ⊖ minta egy ⊖ tesztben megjelenik, $p \cdot q^k$. (Hiszen annak a valószínűsége, hogy egy csoportteszt ⊖, q^k, és az adott minta p valószínűséggel bukkan fel fel. Így annak a valószínűsége, hogy egy ⊖ minta egy ⊕ eredményű csoporttesztben szerepel, $(1 - p \cdot q^k)^T$.

Az algoritmus akkor ad pontos eredményt, ha minden $\ominus$ elem előfordul egy $\ominus$ eredményű tesztelt halmazban. Tehát

$$P(\text{hiba}) = P\left(\bigcup_{i\in K^c} \{\text{i egy olyan elem, amely nem fordul elő } \ominus \text{ eredményű tesztelt halmazban}\}\right)$$

$$\leq |K^c|(1-p\cdot q^k)^T \leq Ne^{-Tp\cdot q^k}$$

A $p(1-p)^k$- nek $p=\frac{1}{k+1}\sim\frac{1}{k}$-ban van maximuma, tehát vehetjük a $p=\frac{1}{k}$-t. Tudjuk, hogy $(1-\frac{1}{k})^k\to\frac{1}{e}$. Így $T=(1+\delta)ek\ln n$-et véve $Tp\cdot q^k\sim(1+\delta)\ln n$.

Használva a $P(\text{sikeres})\leq\frac{2^T}{\binom{n}{k}}$ becslést, azt kapjuk, hogy

$$\frac{\log_2\binom{n}{k}}{T}\sim\frac{(1-\alpha)}{\ln 2}\frac{k\ln n}{T}$$

.

Ezután $(1+\delta)ek\ln n$-et véve kapjuk, hogy R akármilyen közel kerülhet $\frac{(1-\alpha)}{e\ln 2}$-höz.

□

6.2. DD eljárás

Az eljárás azon a megfigyelésen alapszik, hogy ha egy $\oplus$ eredményű csoporttesztben csak egy olyan minta van, amelyik nem volt még $\ominus$ eredményű csoportban, akkor az a minta mindenképp $\oplus$.

6.2.1. Algoritmus. *[6] A Bernoulli modell alapján sorsolunk egy* $\mathbf{X}$ *tesztelési mátrixot, és elvégezzük a tesztköröket.*

A tesztek elvégzése előtt minden mintát a „lehet, hogy $\oplus$" halmazba tesszük. Ha egy csoportteszt $\ominus$ eredményű, az összes benne lévő mintát áttesszük a „biztosan $\ominus$" halmazba. Hogyha van egy olyan $\oplus$ eredményű csoportteszt, amiben egy kivételével

minden minta a „biztosan ⊖” halmazban van, akkor azt a mintát áthelyezzük a „biztosan ⊕” halmazba.

A tesztsorozat végén a „biztosan ⊕” halmaz elemeit jelöljük meg ⊕-ként.

6.2.2. Lemma. *Az így kapott K_{DD} halmaz nem tartalmaz olyan mintákat, amik igazából ⊖-ok, és $K_{DD} \in K$.*

Bizonyítás. Az eljárás alapján nyilvánvaló, hogy a „biztosan ⊖” halmazban csak ⊖, míg a „biztosan ⊕” halmazba csak ⊕ elemek kerülhettek. Azonban az nem biztos, hogy minden minta átkerült a „lehet, hogy ⊕” halmazból valamelyikbe, így lehet olyan ⊕ minta, amelyik nincs benne a „biztosan ⊕” halmazban. Tehát $K_{DD} \in K$. □

6.2.3. Tétel. *Tegyük fel, hogy egy nemadaptív, zajmentes csoporttesztelés eredménye adott* $\mathbf{y}$*, amit a DD eljárással akarjuk dekódolni. A Bernoulli modellben, egy optimális $p = \frac{1}{d}$ paraméterválasztással és $d = \Theta(n^\alpha)$, $\alpha \in (0,1)$, mellett az $R = \frac{1}{e \ln 2} min\{1, \frac{1-\alpha}{\alpha}\}$ arányra teljesülni fog, hogy $\forall \delta, \varepsilon$-ra létezik olyan csoporttesztelési algoritmus, amelyre*

$$\frac{\log_2 \binom{n}{d}}{T} > R - \delta,$$

és a hibavalószínűség legfeljebb ε.

6.3. SCOMP - ismételt COMP

Az eljárás [6] a „biztosan ⊕” halmazból kiindulva keres egy kielégítő halmazt. Az elnevezése a „Sequential COMP”-ból származik, mivel egyfajta egymás után ismétlése a COMP algoritmusnak.

6.3.1. Algoritmus. *A Bernoulli modell alapján sorsolunk egy* $\mathbf{X}$ *tesztelési mátrixot, és elvégezzük a tesztköröket.*

Először a DD eljárással állapítsuk meg a „biztosan ⊕” és „biztosan ⊖” halmazok elemeit. A K_{SCOMP} halmazba a „biztosan ⊕” halmaz elemeit rakjuk. Ezután minden olyan ⊕ eredményű csoporttesztet, amiben nincs K_{SCOMP}-beli elem, megjelölünk „furcsa”-ként. Azt a $k \notin K_{SCOMP}$ elemet, ami a legtöbb „furcsa” csoporttesztben szerepel, belerakjuk a K_{SCOMP} halmazba, és azokról a csoportokról, amikben benne van, levesszük a „furcsa” jelölést. (Hogyha nem egyértelmű, válasszunk ki egyet közülük.) Az utóbbi lépést addig ismételjük, amíg van olyan csoportteszt, ami megjelölhető „furcsa”-ként.

Az eljárás végén a K_{SCOMP} elemeit jelöljül meg ⊕-ként.

6.3.2. Állítás. *Minden adott* **X** *tesztelési mátrix esetén bármely olyan R arány, amely elérhető a DD eljárással, elérhető a SCOMP eljárás által is. Tehát a 6.2.3. tételben meghatározott arány a megadott feltételek mellett elérhető a SCOMP eljárással való dekódolás esetén is.*

Bizonyítás. Az állítás mögötti egyszerű gondolat, hogy amikor a $K_{DD} \in K$ halmaz csak kicsit tér el a K halmaztól, akkor a K_{SCOMP} halmaz is csak néhány elemében térhet el tőle, hiszen kielégítő halmaz, és $K_{DD} \in K_{COMP}$ □

7. fejezet

A csoportos tesztelés egyéb felhasználási területei

A csoporttesztelési problémát a szifilisz szűrésével összefüggésben fogalmazták meg, azonban az eredményeket később számos más területen is alkalmazni tudták.

Sobel és Groll korai munkájukban([11]) felsoroltak a néhány alapvető alkalmazási lehetőséget, mint például hibás tartályok, kondenzátorok, vagy karácsonyfaégők megtalálása. Az utóbbi időben javasoltak csoporttesztelésen alapuló megoldásokat a gyártási folyamatok során előforduló minőségellenőrzésekre is, például integrált áramkörök esetében és molekuláris elektronikában.

Az alkalmazások közül sok a nemadaptív algoritmusokat használja, aminek az az oka, hogy sok esetben az adaptív algoritmus nem praktikus, hiszen sok figyelmet igényel, időigényes, míg előre rögzített teszttervet futtatni egyszerűbb, sok tesztet lehet párhuzamosan elvégezni, így gyorsabb.

Tekintsünk néhány további alkalmazást, a teljesség igénye nélkül.

A probléma eredetét tekintve, nem meglepő, hogy biológiai területeken számos alkalmazási lehetőség van, így például DNS-tesztek esetében a megfelelő genomok keresése, ezzel ritka genetikai tulajdonsággal rendelkező egyedek kiszűrése során

alkalmaznak nemadaptív csoporttesztelési algoritmusokat; valamint természetesen vírusok, baktériumok, különböző fertőzések jelenlétének megállapítása során is alkalmazzák. A biológiai vonatkozású felhasználások között van a fehérje-fehérje interakciós kísérletek tervezése, a nagy áteresztőképességű gyógyszerszűrés, valamint az immunhiányos grafikonok hatékony megtanulása.

Informatikai területeken alkalmazzák például hálózati tomográfiában és anomáliák felfedezésében, adattárolás és tömörítés során, adatbázisrendszerek és kiberbiztonság kapcsán.

Végül, de nem utolsósorban a csoporttesztelést számos statisztikai és elméleti számítástudományi probléma megoldásában alkalmazták. Ilyenek például keresési problémák, a ritka következtetés és tanulási problémák (Sparse inference and learning), valamint a klasszikus mintaillesztési probléma és a rejtett páros gráfok nagyfokú csúcsainak becslése.

Irodalomjegyzék

[1] S.Jaggi C. L. Chan P. H. Che és V. Saligrama. *Non-adaptive group testing: Explicit bounds and novel algorithms*. IEEE Transactions on Information Theory, 60(5):3019–3035, 2014.

[2] S.Jaggi C. L. Chan P. H. Che és V. Saligrama. *Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms*. In 49th Annual Allerton Conference on Communication, Control, és Computing, pages 1832–1839, 2011.

[3] Endre Csóka. *Application-oriented mathematical algorithms for group testing*. arXiv: 2005.02388v1 [q-bio.QM], 2020.

[4] Robert Dorfman. *The detection of defective members of large populations*. The Annals of Mathematical Statistics, 14(4):436–440, 1943.

[5] C. H. Li. *A sequential method for screening experimental variables*. Journal of the American Statistical Association, 57(298):455–477, 1962.

[6] L. Baldassini M. Aldridge és O. T. Johnson. *Group testing algorithms: Bounds and simulations*. IEEE Transactions on Information Theory, 60(6):3671–3687, 2014.

[7] Oliver Johnson Matthew Aldridge és Jonathan Scarlett. *Group Testing: An Information Theory Perspective*. arXiv: 1902.06002v3 [cs.IT], 2020.

[8] Charles G. Pfeifer és Peter Enis. *Dorfman-Type Group Testing for a Modified Binomial Model*. Journal of the American Statistical Association, Vol. 73, No. 363, pp. 588-592, 1978.

[9] Claude E Shannon. *A mathematical theory of communication*. Bell system technical journal, 27(3):379–423, 1948.

[10] M. Sobel és P. A. Groll. *Binomial Group-Testing with an Unknown Proportion of Defectives*. Technometrics, 1966.

[11] M. Sobel és P. A. Groll. *Group Testing To All Defectives in Eliminate Efficiently a Binomial Sample*. The Bell System Technical Journal, 1958.

[12] A. Sterrett. *On the Detection of Defective Members of Large Populations*. The Annals of Mathematical Statistics, 1957.

[13] Péter Ungár. *The Cutoff Point for Group Testing*. Communications on pure és applied mathematics, vol. XIII, 49-54, 1960.