Dragan Azdejković
Univerzitet u Beogradu –
Ekonomski fakultet

POBOLJŠANJE EFIKASNOSTI NEKIH TESTOVA FORENZIČKE ANALITIKE

IMPROVING THE EFFICIENCY OF SOME FORENSIC ANALYTICS TESTS

Apstrakt

Klasični testovi forenzičke analitike za pronalaženje netipičnih vrednosti u podacima bazirani su na pretrazi zapisa sa apsolutnim poklapanjem po odabranim poljima. Cilj ovog rada je da se u te testove uključe i zapisi sa sličnim sadržajem u odabranim poljima i da se pokaže da su novi testovi efikasniji od klasičnih. Testiranje je obavljeno na javno dostupnim podacima potrošnje na karticama službenika Distrikta Kolumbija za period 2009-2014.

***Ključne reči**: Forenzička analitika, Netipične vrednosti, Slični sadržaji, Distrikt Kolumbija, Efikasnost*

***JEL klasifikacija**: M42, C63, C80*

Abstract

Classic tests of forensic analytics for finding atypical values in data are based on searching records with an absolute match by selected fields. The aim of this work is to include records with similar content in selected fields in those tests and to show that the new tests are more effective than the classic ones. Testing was done on publicly available spending data on District of Columbia employee cards for the period 2009-2014.

***Keywords:** Forensic analytics, Outliers, Similar content, District of Columbia, Efficiency*

***JEL classification:** M42, C63, C80*

Uvod

Forenzička analitika se intenzivno koristi u različitim aplikacijama za otkrivanje nepravilnosti u finansijskim izveštajima (Nigrini, 1999) i nezakonitih finansijskih tokova. Ona kombinuje naprednu analitiku sa forenzičkim računovodstvenim i istražnim tehnikama da bi se identifikovali potencijalni retki događaji sa posledicama — što je poput traženja igle u ogromnim skupovima podataka i informacija koje mogu signalizirati nevolje.

Detekcija anomalija je zanimljiva jer uključuje automatsko otkrivanje zanimljivih i retkih obrazaca iz skupova podataka (Patcha & Park, 2007). Detekcija anomalija je široko proučavana u statistici i mašinskom učenju (Markos & Singh, 2003), gde je poznata i kao otkrivanje odstupanja, detekcija odstupanja, detekcija novina i rudarenje izuzetaka. Anomalije su važne jer ukazuju na značajne, ali retke događaje i mogu da podstaknu preduzimanje kritičnih radnji u širokom spektru domena primene

Testovi zasnovani na identifikovanju neprirodnih ponavljanja su efikasno sredstvo za otkrivanje jedne najčešćih finansijskih prevara kojim se organizacije suočavaju, a to je manipulacija evidencijama ili računima i preusmeravanje sredstava. Duplirana plaćanja su dobre šeme za pronevere koje sprovode finansijski direktor, blagajnik ili neko sa računovodstvenim ovlašćenjima.

Testovi zasnovani na ponavljanju se baziraju na pretpostavci da su prekomerna ponavljanja u nekom nizu višedimenzionalnih finansijskih podataka indikatori greške, prevare ili neke druge anomalije. U literaturi vezanoj za forenzičku analitiku koristi se termin neprirodna ponavljanja u podskupu, što je matematički nekorektno, jer po definiciji skupa on sadrži samo različite objekte, odnosno samo objekte koji poseduju precizno definisanu osobinu. Jasno je na šta su autori mislili, međutim, s obzirom da je skup svih podataka jedan niz, korektnije je reći da se radi o ponavljanjima u podnizovima. Tri testa ovog tipa su prikazana u 12. poglavlju knjige Nigrinija (2020). Takođe, razne varijante ovih testova, implementirane su u različitim analitičkim programima poput IDEA i TABLOU. Autor nije pronašao druge relevatne izvore u kojima se pominje ova tema, to je verovatno stoga jer je tema nova i zato što su već postojeći testovi uglavnom radili dobro.

Imajući u vidu da su ovi testovi bazirani na pojmovima jednako, odnosno različito, u ovom radu se relaksira definicija jednakosti dve stringovne promenljive. Pretpostavka je da će prevarni zapisi biti vrlo slični ispravnim kako bi se teže uočili. Takođe, neki dobavljači su u osnovi ćerke neke veće kompanije, te treba istražiti neprirodna ponavljanja ka u suštini istom dobavljaču.

Kod ovih testova potrebno je sortirati podnizove, kako bi se izveli zaključci. Obično se sortira po veličini podniza. U ovom radu je pokazano kako se i sortiranjem po drugim kriterijumima mogu izvesti relevantni zaključci.

1. Osnovne karakteristike testova zasnovanih na ponavljanju

Svrha testa isto-isto-isto (III) je da identifikuje neprirodna dupliranja kao potencijalne indikatore grešaka ili prevare. Primena ovog testa pomaže u otkrivanju dupliranih potraživanja troškova, grešaka u vezi sa istim isplatama dobavljačima, višestrukim potraživanjima u vezi sa garancijom ili dupliranim naknadama za usluge koje plaćaju privatni ili državni zdravstveni planovi, prevarama sa povraćajem novca kupcima, lažiranjem zaliha, trajne imovine i manipulacije sa platnim spiskom.

Test isto-isto-različito (IIR) se koristi za identifikaciju zapisa sa skoro duplikatima za polja koja je odabrao revizor. Revizor može izabrati nekoliko polja za podudaranje i bar jedno polje koje je isključeno iz uparivanja. Mark Nigrini kaže: "Isto-isto-različito test je moćan test za greške i prevaru. Ovaj test treba uzeti u obzir za svaki projekat forenzičke analitike." Njegovo iskustvo je pokazalo da „ovaj test uvek otkriva greške u podacima o obavezama" i „Što je duži vremenski period, veće su šanse da IIR otkrije greške."

Testom ponavljanja brojeva u podnizu (PBUP) identifikuju se podnizovi sa prekomernim ponavljanjem broja ili brojeva. Za test se koristi faktor učestalosti brojeva (FUB) kojim se meri obim ponavljanja brojeva za svaki podniz. Formula je prikazana u jednačini koja sledi:

$$FUB = \frac{\sum c_i^2}{n^2}$$

gde je c_i frekvencija ponavljanja broja u podnizu, ako je ona veća od 1, a n broje elemenata u podnizu. Na primer, pretpostavimo da su vrednosti podniza

23, 23, 23, 23, 18, 18, 38, 23

tada bi FUB za taj niz iznosio $(5^2 + 2^2)/8^2$

Tokom faze planiranja ovih testova, potrebna je doza kreativnosti kako bi se definisalo šta je neprirodno za dati skup podataka.

2. Podaci i metodologija

Kako bi ocenili efikasnost korigovanih testova analiziraćemo platne kartice službenika Distrikta Kolumbija. Podaci su dobijeni pristupom sajtu: https://opendata.dc.gov/datasets. U vreme kada je pisan ovaj rad, informacije do kojih smo došli pokazale su da skup podataka sadrži 563.027 zapisa.

Analizirani su podaci za Distrikt Kolumbija i period 2009 - 2024. Godina. Uz pomoć korigovanog testa III, kao i sortiranjem po novim kriterijumima proizašli su neki interesantni nalazi.

U teoriji informacija, lingvistici i računarstvu, rastojanje je metrika stringova kojom merimo razliku između dva stringa. Postoji nekoliko načina kako se može definisati to rastojanje. U ovom radu je izabrano Levenštajnovo rastojanje, jer ono po autorovom mišljenju najpogodnije da se relaksira jednakost stringova u kontekstu forenzičke analitike. Sovjetski matematičar Vladimir Levenštajn (1965) je rastojanje između dva stringa definisao kao minimalni broj izmena jednog karaktera (umetanja, brisanja ili zamena) potrebnih da se jedna reč promeni u drugu. Levenštajnovo rastojanje između dva stringa *a* i *b* (sa dužinama |*a*| i |*b*| respektivno) je dato sa *lev(a,b)*, gde je

$$lev(a,b) = \begin{cases} |a| & ako\ je\ |b| = 0, \\ |b| & ako\ je\ |a| = 0, \\ lev(tail(a), tail(b)) & ako\ je\ head(a) = head(b), \\ 1 + min\{lev(\text{tail(a), b}), lev(\text{a, tail(b)}), lev(\text{tail(a), tail(b)}\} & u\ ostalim\ slučajevima \end{cases}$$

gde je *tail*(*x*) string koji se od stringa *x* dobija brisanjem prvog karakter (odnosno, $tail(x_0x_1 \cdots x_n) = x_1x_2 \cdots x_n$, a *head*(*x*) je prvi karakter stringa (odnosno, $head(x_0x_1 \cdots x_n) = x_0$).

Prvi element u funkciji *min* korespondira brisanju karaktera (od stringa *a* ka stringu *b*), drugi umetanju, a treći zameni.

Na primer, Levenštajnovo rastojanje između „*Marina*“ i „*Marjan*“ je 3 jer da bi smo prešli sa prvog stringa na drugi neophodno je jedna zamena (*i* u *j*) i jedno brisanje (*n*) i jedno umetanje (*n*).

U ovom radu za test III modifikujemo pojam jednakosti za polja koja ukazuju na dobavljača, a zatim istražujemo neprirodna ponavljanja. Naime, za dva polja ćemo reći da su jednaka ako je Levenštajnovo rastojanje manje od neke pogodno izabrane konstante.

Kod ovih testova obično se analiziraju podnizovi sa velikom frekvencijom ponavljanja, međutim i sortiranjem podnizova po nekim drugim kritarijumima možemo dobiti neke relevatne informacije koje, kao što ćemo videti, mogu ukazati na grešku ili manipulaciju.

Analiza podataka napravljena je u okruženju programskog jezika Pajton, koje se pokazao bržim i fleksiblinijim od gotovih komercijalnih softverskih paketa odnosno eksela.

3. Rezultati i diskusija

Modifikovani test III izdvojio je veći broj neprirodnih ponavljanja nego uobičajeni test. U tabeli koja sledi se nalazi 10 najvećih stavki po frekvenciji ponavljanja koje su plaćene istog dana i glase na isti iznos. Dobavljač u jednom redu je predstavnik klastera svih dobavljača koji se za najviše 4 karaktera razlikuju po imenu od imena klastera dobavljača.

Tabela 1. Rezultati modifkovanog testa III - ponavljanja

Dobavljač	*Iznos*	*Frekvencija*
UPS*0000X4R07806042011	2.59	195
LOWES #03256	11.56	127
AMZN MKTP US	-7.39	127
FRAUD CREDIT	-9.99	108
0662 EXTRA SPACE STORA	10	95
NATIONAL EMERGENCY TRA	145.86	84
H&M0319	135.94	74
STANDARD OFFICE SUPPLY	117.75	74
NETWORK TOOL WAREHOUSE	340.19	73
AMZN MKTP US	326.47	68

Izvor: Autor

Podaci iz Tabele 1 ukazuju na moguću neefikasnost službe za nabavku, a to dalje vodi do neefikasnosti službe za knjigovodstvo. Forenzičari mogu da istraže kompletnu tabelu po dubini. Takođe, lako se mogu dobiti i ostale relevatne informacije o izdvojenim plaćanjima kao što su datum i službe koje su plaćanja izvršile.

U Tabeli 2 su rezultati testa filtrirani uz uslov da je iznos veći od 2000 i da postoje bar dva ponavljanja. U tabeli se nalaze prvih deset predstavnika klastera sortiranih po broju ponavljanja, pa onda po iznosu. Sada vidimo jednu drugačiju sliku i očigledno je da bi forenzičari svakako trebalo da ispitaju ceo klaster i ostala polja ovih transakcija, ali i klasteri koje se nalaze dublje u tabeli.

Tabela 2. Rezultati modifkovanog testa III – ponavljanja koja treba ispitati

Dobavljač	*Iznos*	*Frekvencija*
LRP PUBLICATIONS	4942	29
AWL*PEARSON EDUCATION	4276.25	23
FLIK GALLCONF 16128290	4906.5	20
KIMPTON GLOVER PARK	16467.91	19
PAYPAL	7296	16
UNITED SITE SERVICE	24301.6	12
SQU*SQ *A DIGITAL SOLU	2800	12
FAIRFIELD INN & SUITES	5970	10
GALLIHER & HUGUELY ASS	2356.7	10
DISNEY RESORT-WDTC	2164.76	9

Izvor: Autor

Konačno, Tabela 3 je dobijena na sledeći način: prvo je u bazi podataka pomoću rezultata testa formirana kolona vrednost, odnosno pomnoženi su rezultati ponavljanja sa iznosom, a zatim je izvršeno sortiranje po toj koloni. U tabeli su prvih deset rezultata testa isfiltriranih na prethodno opisan način.

Tabela 3. Vrednosti isplaćene istom dobavljaču istog dana na isti iznos

Dobavljač	*Vrednost*	*Frekvencija*
KIMPTON GLOVER PARK	312890.29	19
UNITED SITE SERVICE	291619.2	12
LRP PUBLICATIONS	143318	29
SQ *CAPITAL LAND COMPA	118800	3
PAYPAL	116736	16
C & D TREE SERVICE I	99500	1
AWL*PEARSON EDUCATION	98353.75	23
FLIK GALLCONF 16128290	98130	20
FAIRFIELD INN & SUITES	59700	10
RESIDENCE INN CAPITOL	48880	8

Izvor: Autor

Poslednja tabela možda daje najjače svetlo ka zapisima koje treba proveriti. Kao i kod prethode tabele treba imati na umu da su u tabeli data imena prestavnika klastera dobavljača te bi forenzičar pri daljoj analizi trebao da iz-

dvoji ceo klaster i pogleda ostale atribute svih zapisa u klasteru. Autor veruje da bi forenzičarima bio zanimljiv svaki zapis iz Tabele 3.

Zaključak

Testovi kojima se ispituju neprirodna ponavljanja mogu se unaprediti relaksiranjem pojma jednakosti stringova. To se može postići uvođenjem metrike među stringovima. U radu je pokazano kako se to može uraditi primenom Levenštajnovog rastojanja. Unapređenje se ogleda u izdvajanju većih podnizova sa karakteristikama koje su od interesa.

U radu je pokazano kako se rezultati testova ponavljanja mogu sagledati i na drugačiji način primenom filtriranja i sortiranja i na taj način doći do novih uvida o bazi podataka koji mogu biti interesantni forenzičarima.

Takođe u radu je pokazana jasna prednost okruženja jupyter i Pajtona kao okvira za primenu testova forenzičke analitike. Ta prednost se ogleda u brzini, fleksibilnosti i ceni jer ovo okruženje je besplatno. Kao prilog tome priložen je programski kod kojim su dobijeni svi nalazi u radu. Možda će nekom ovo okruženje izgledati komplikovano, ali kao što Nigrini kaže, nema forenzičke analitike bez veština iz programiranja, statistike i računovodstva.

Literatura

Friedlob, G.T. & Lydia L.F. Schleifer 1999, "Fuzzy logic: application for audit risk and uncertainty", Managerial Auditing Journal, vol. 14, no. 3, pp. 127-135

В. И. Левенштейн (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов [Binary codes capable of correcting deletions, insertions, and reversals]. Доклады Академии Наук СССР (in Russian)*, 163(4)*, 845–848. Appeared in English as: S*oviet Physics Doklady*, *10 (8)*, 707–710.

Markos M., Singh S., (2003). Novelty detection: a reiew—part 2:: neural network based approaches, *Signal Processing, Volume 83(12)*, 2499-2521.,

Patcha A., Park J., (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks, 51(12)*, 3448-3470,

Nigrini M. J. (1999). "I've Got Your Number." *Journal of Accountancy,* May (147): 79–83.

Nigrini, M. J. (2020). *Forensic analytics: Methods and techniques for forensic accounting investigations*. Wiley.

Dodatak (Programski kod korišćen za analizu podataka)

```
pip install Levenshtein

import pandas as pd
import numpy as np
from Levenshtein import distance

df=pd.read_excel('PCT.xlsx', names=['Agencija', 'Datum',
'Iznos','Dobavljač','Opis'] )

df.sort_values(by=['Datum', 'Iznos', 'Dobavljač'], inplace=True)
df.index = np.arange(0, len(df))
df.head(10)

df['Dobavljač'] = df['Dobavljač'].astype(str)
df['Datum'] = df['Datum'].astype(str)

df['Grupa'] = 0
k=0
r=df.loc[0]
c=0
df.sort_values(by=['Datum', 'Iznos', 'Dobavljač'], inplace=True)
for i in range(len(df)):
   p=df.loc[i]
   if p['Datum']==r['Datum'] and p['Iznos']==r['Iznos']  and distance(p['Dobavlja
č'],r['Dobavljač'])<= 5:
      c+=1
   else:
      df.loc[i,'G_brojač']=c
      c=0
      r=df.loc[i]
      df.loc[i,'Grupa']=k
      k=k+1

df.sort_values(by=['G_brojač','Iznos'], ascending=[False, False],inplace=True)
df1=df[['Dobavljač', 'Iznos', 'G_brojač']].head(10)

df1.to_excel('output1.xlsx')

rezultat_1=df.query('Iznos > 2000 and G_brojač>2')
rezultat_1[['Dobavljač', 'Iznos', 'G_brojač']].head(10)

rezultat_1[['Dobavljač', 'Iznos', 'G_brojač']].head(10).to_excel('output2.xlsx')
```

```
df['P_vrednost']=df['Iznos']*df['G_brojač']

df_3=df.sort_values(by='P_vrednost',ascending=False).head(10)
df_3[['Dobavljač', 'P_vrednost','G_brojač']]

df_3[['Dobavljač', 'P_vrednost','G_brojač']].to_excel('output3.xlsx')
```