Klaster analiza u funkciji diverzifikacije rizika investicionih ulaganja

Jelena Brdar
Departman za finansije, bankarstvo i osiguranje
Ekonomski fakultet Subotica
Subotica, Srbija
Naftna industrija Srbije
Novi Sad, Srbija
jelena_brdar@yahoo.com

Zita Bošnjak
Departman za poslovnu informatiku i kvantitativne metode
Ekonomski fakultet Subotica
Subotica, Srbija
bzita@ef.uns.ac.rs

Sadržaj— **Klasterovanje je najčešće prvi korak u istraživanju obimnih podataka (eng. big data). Veliki značaj i primenu pronalazi u obradi složenih podataka koji su karakteristični za finansijske berze. U ovom radu je izvršena klaster analiza na desetogodišnjim istorijskim kretanjima cena akcija 100 kompanija iz 10 privrednih sektora. Na odabrani skup podataka primenjen je algoritam K-means, kojim su kompanije sa sličnim osobinama grupisane u klastere. Rezultati klasterovanja upoređeni su sa pripadnošću kompanija po sektorima kako bi se utvrdilo da li se kompanije iz istog sektora slično ponašaju na tržištu. Drugi deo eksperimentalne analize posvećen je primeni rezultata istraživanja u diverzifikaciji rizika. Pokazano je da se raspoređivanjem investicija po klasterima može ostvariti viši prinos.**

Ključne riječi:- klaster analiza; k - means; privredni sektori; diverzifikacija rizika;

I. UVOD

Istraživanje podataka je proces otkrivanja korisnih informacija. Tehnike za istraživanja pretražuju baze podataka kako bi pronašle neobične i korisne obrasce koji bi inače ostali nepoznati. Proces otkrivanja znanja u podacima je osnovna uloga inteligentih tehnika. Savremene metode obrade podataka - data mining [1] obuhvataju tehnike nadgledanog i nenadgledanog učenja. Klasterovanje ili segmentacija je zadatak koji se ne nadzire tj. ni jedan atribut se ne koristi za vođenje trening procesa već se svi atributi tretiraju kao ulazni. Većina algoritama za klasifikaciju gradi model podataka kroz određen broj iteracija i zaustavlja iteracije kada model konvergira tj. kada se granice ovih segmenata stabilizuju.

Klaster analiza vrši se raznim algoritmima koji se razlikuju značajno u njihovoj ideji šta predstavlja klaster i kako efikasno da ih pronađemo. U radu je prikazan algoritam za klaster analizu: K-means [2]. Osnovni skup podataka za eksperiment su kompanije iz 10 različitih privrednih grana (energetika, osnovni materijali, industrija, potrošna dobra, nepotrošne usluge, finansije, zdravstvo, tehnologija, telekomunikacije, usluge) . U okviru svake privredne grane posmatrano je po 10 vodećih kompanija. Odabir je izvršen prema prinosu. Podaci za svaku kompaniju su desetogodišnje vremenske serije kretanja cena akcija vodećih kompanija u privrednoj grani.

Cilj primene klaster analize u prvom delu istraživačkog rada je utvrđivanje stepena sličnosti između kompanija koje pripadaju istom sektoru. Analizirali smo koje kompanije su se grupisale u klasterima. Drugi deo eksperimenta sproveli smo na osnovu izvršene podele kompanija u klastere – grupe. Vršili smo izbor akcija koje kreiraju portfolio. Donošenje investicionih odluka nije ni malo jednostavan posao i zbog toga naučnici istražuju i razvijaju nove metode. Nastoji se primenom tehnika za klasterovanje ostvariti veći stepen uspešnosti investicionih odluka uz manja odstupanja, tj. greške modela. Polazi se od hipoteze da se raspodelom ulaganja u akcije kompanija po klasterima u odnosu na modele sa slučajnim tehnikama za izbor ulaganja, ostvaruje diverzifikacija rizika i kreira optimalni portfolio. Eksperimentalni rezultati dobijeni su korišćenjem alata: Python skript jezik [3] za procesiranje podataka i K-means algoritma [2] za klasifikaciju kompanija u klastere.

Sledeće poglavlje posvećeno je pregledu relevantne literature. U okviru trećeg poglavlja opisana je klaster analiza i mere za normalizaciju podataka. Eksperimentalni rezultati izloženi su u četvrtom poglavlju.

.

II. PREGLED RELEVANTNE LITERATURE

Istorijski razvoj klaster analize započet je 1939. godine. Kao pionir razvoja navodi se Trion, koji je prvi put upotrebio klasterovanje u analizama podataka [4]. Termin klaster analiza obuhvata niz različitih algoritama i metoda za grupisanje objekata sličnog tipa u odgovarajuće kategorije. Stalni razvoj nauke, informatike, značaja klasifikacije u istraživanjima doprineli su razvoju i porastu značaja ove metode. Značajnija literatura se razvija šezdesetih godina. Brojne publikacije [5-10] imaju značajan doprinos u razvoju tehnika klasterovanja i dobijaju široku razmeru u naučnim krugovima, u statistici, analizama podataka i praktičnoj primeni.

Postoje različiti načini primene klaster analiza i algoritama za klasterovanje. U svom radu [2] Hartigan je detaljno opisao k-means algoritam. Mogućnosti koje ove tehnike pružaju u analizama su ogromne. Naročito su primenu pronašle statistici, ekonometriji i drugim ekonomskim disciplinama. Naučnik

Nanda, Mahnaty i Tiwari [11] predstavili su mogućnosti primene data mininga-a u oblasti ekonomije. Izvršili su grupisanje akcija u klastere. Nakon formiranih klastera akcije su mogle biti odabrane za portfolio. Cilj rada je bila diverzifikacija rizika. Rezultati analize su pokazali da K-means klaster analiza gradi kompaktnije skupove u odnosu na Kohonenove samoorganizujuće mreže (SOM) [12] i Fuzzy C-means algoritam [13] za klasterovanje akcija.

Po uzoru na prethodno navedene radove, urađeno je istraživanje opisano u ovom radu. U ekonomiji u oblasti investicionih ulaganja postoji velika količina složenih podataka. Rizik je neizostavni deo berzanskog posla. S tim razlogom se razvija i nauka u pravcu povećanja sigurnosti ulaganja i mogućnosti ostvarenja višeg prinosa od ulaganja. U našem radu proučavamo dve metode - klasterovanje i slučajni izbor. Na osnovu dve metode se vrši odabir akcija koje će formirati portfolio hartija od vrednosti i vrši se njihovo poređenje sa stanovišta ostvarenog prosečnog prinosa od divdende. Cilj je da se analizom primene klasterovanja dokaže polazna hipoteza o diverzifikaciji rizika berzanskih ulaganja i kreiranju optimalnog portfolia.

III. K-MEANS KLASTER ANALIZA

Kao što smo istakli u uvodnim napomenama, korišćen algoritam u radu za grupisanje industrijskih grana u klastere, je tzv. K-means. K–means je tehnika particionisanja tj. nehijerarhijskog klasterovanja. Osnovna prednost algoritma je pogodnost za rad s velikim brojem objekata (u našem radu sa 100 kompanija različitih privrednih sektora). U odnosu na hijerarhijske klaster analize, koja rezultira sukcesivnim spajanjem objekata u sve veće klastere, kod k-means potrebno je unapred proceniti optimalan broj klastera. Svaki klaster predstavljen je centroidom i svaki objekat se pridružuje najbližem centroidu. Početni izbor centroida je slučajan, a u narednim iteracijama oni se računaju kao aritmetička sredina pripadajućih objekata. Broj klastera se zadaje kao ulazni podatak. Za meru rastojanja korišćeno je Euklidsko rastojanje [14]. U formuli (1), standardna euklidska udaljenost dva objekta X i Y se računa kao kvadratni koren iz sume kvadratnih razlika za sva obeležja Xi, Yi respektivno. Što je manje Euklidsko rastojanje veća je i sličnost posmatranih obeležja – cena akcija kompanija.

$$\text{Distance}(X, Y) = \sqrt{\sum (X_i - Y_i)^2} \qquad (1)$$

Objekti posmatranja su kompanije, a njihove osobine, tj. posmatrana obeležja su cene akcija. Kompanije u radu birali smo prema godišnjem prinosu u tekućoj godini. Iz svakog sektora odabrali smo po 10 vodećih kompanija - Npr. u energetskom sektoru su među najuspešnijim Petro Brasileus, British Petrol, Chevron, Exxon, Gazprom; među vodećim kompanijama finansijskog sektora su American Bank, JP Morgan, CityGroup; u sektoru tehnologije Apple, IBM, Sap, Microsoft. U tabeli 1. prikazane su izabrane kompanije. Radi jednostavnijeg prikaza u tabeli smo kompanije označili sa zvaničnim skraćenim nazivima kompanija na berzama.

TABELA 1. Privredni sektori i vodeće kompanije

Energy	***BMaterials***	***Industrials***	***CCGoods***	***NCCGoods***
Ptr	Aa	Ge	Amzn	Abc
Bp	Ach	Ba	Cmsca	Adm
Cvx	Bhp	Abb	Cost	Cah
E	Dow	Cat	Dis	Cvs
Pbr	Ip	Fdx	F	Ko
Snp	Mt	Hon	Hd	Kr
Sto	Pkx	Lmt	Hmc	Mck
Tot	Rio	Mmm	Tgt	Pep
Ogzpy	Srh	Ups	Tm	Pg
Xom	Vale	Utx	Wmt	Weg
Financials	***HealthCare***	***Technology***	***Telecom***	***Utilties***
Aig	Gsk	Apple	Amx	Aes
Bac	Ily	Csco	Cha	Duk
C	Hum	Googl	Chl	Ecx
Gs	Jnj	Sap	Chu	Etp
Jpm	Mdt	Hpg	Ntt	Hnp
Ifc	Mrk	Ibm	Oran	Kep
Met	Nvs	Intc	Tat	Ngg
Puk	Pfe	Msft	Vod	Pcg
San	Pgh	Orcl	Vz	So
Wlp	Sny	Sne	Ti	Ve

Istorijsko kretanje cena akcija od deset godina (2004-2013. godine) su obeležja na osnovu kojih smo poredili stepen sličnosti kompanija i izvršili klasterizaciju. Cene akcija su reprezentant uspeha neke kompanije. Da li pad ili rast kretanja cena akcije kompanije jednog sektora znači isto kretanje i za ostale članove sektora ili ta veza u kretanju može biti slična sa kompanijama nekog drugog sektora osnovno je pitanje prvog dela eksperimentalnog istraživanja.

Da bi analize mogle da se rade, potrebno je prvo izvršiti uređivanje (preprocesiranje) podataka. Transformacija podataka predstavlja normalizaciju koja ima za cilj da omogući upotrebu ulaznih podataka. Mere za normalizaciju koje su korišćene u radu su: Zero-one, Mean i Init.

a) ***Zero-one (0-1) normalizacija:*** Za rešavanje zadataka potrebno je izvršiti normalizaciju vrednosti atributa, odnosno izvršiti "ujednačavanje" ili "učiniti atribute bezdimenzionalnim", što znači da se vrednosti atributa svedu

na interval 0-1. Analizirane cene akcija se preskaliraju, tako da minimumu odgovara 0, a maksimumu 1. Ostale vrednosti cena akcija nalaze se u rastojanju između graničnih vrednosti 0-1.

b) ***Mean normalizacija:*** Za sve cene akcija računa se srednja vrednost, zatim se pojedinačne cene normalizuju, deljenjem njihove vrednosti sa srednjom.

c) ***Init normalizacija:*** Odredi se prva početna cena, a zatim se vrši preskaliranje prema ceni u prvom momentu posmatranja i na taj način određuju druge cene. U radu je početna cena ona iz 2004 godine.

IV. EKSPERIMENTALNA ANALIZA

REZULTAT KLASTEROVANJA

Nakon izvršene normalizacije, odabrali smo za početni broj klastera deset. Svaka tačka u skupu podataka pridružena je klasteru s najbližom početnom tačkom (zasnovanom na Euklidskoj udaljenosti). Ukoliko klaster ima više od jednog člana, početna tačka klastera zamenjuje se njegovim centroidom. Nakon što su svi članovi pridruženi klasterima za svaki član se proverava da li je bliži centroidu nekog drugog klastera nego centroidu vlastitog klastera. Ako jeste, premešta se u novi klaster, a centroid klastera se ponovo preračunava. Postupak se nastavlja sve dok nova poboljšanja više nisu moguća. Za ocenu kvaliteta grupisanja koristi se kvantitativna mera - klaster validacija. U radu je primenjena eksterna metoda NMI (eng. Normalized Mutual Information) mera [15]. Mera se zasniva na teoriji informacija i računa se pomoću entropija. NMI mera često se koristi za procenu rezultata klasterovanja, pronalaženje informacija, ili izbor funkcija. U evaluaciji klastera koristi stvarne labele i poredi ih sa labelama koje su rezultat klasterovanja. U našoj analizi, NMI mera je primenjena na sledeće vrednosti:

K-means: zero – one

[7, 9, 1, 9, 9, 7, 7, 9, 0, 1, 5, 0, 8, 2, 2, 0, 7, 7, 9, 0, 2, 3, 8, 1, 2, 3, 3, 3, 2, 3, 4, 4, 4, 4, 2, 4, 8, 2, 6, 4, 4, 8, 2, 4, 4, 3, 4, 3, 1, 3, 5, 5, 5, 8, 8, 7, 2, 3, 9, 3, 2, 2, 3, 3, 2, 6, 3, 2, 9, 2, 4, 9, 3, 4, 8, 4, 2, 6, 4, 9, 8, 7, 7, 7, 3, 9, 6, 6, 6, 5, 9, 2, 9, 8, 6, 9, 6, 8, 1, 0] – rezultat klasterovanja

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9] – stvarne labele prema pripadnosti sektorima

NMI mera validacije dobija vrednost iz opsega [0, 1], pri čemu vrednosti bliže nuli označavaju neslaganje između stvarnih labela i rezultata klasterovanja. Ova eksterna mera je korisna za razvoj i verifikaciju klastera. Izmerena je NMI za sve tri mere normalizacije. Rezultati su dati u tabeli 2.

TABELA 2. Mere normalizacije

Algoritam/Normalizacija	*Zero-one*	*Mean*	*Init 1*
k-means	**0.341**	**0.284**	**0.304**

Najbolji rezultat grupisanja po sektorima je imala normalizacija cena primenom zero – one mere (0,341). Rezultat pokazuje da postoji izvesno slaganje između dobijenih klastera i podele kompanija po sektorima, ali da to slaganje nije jako izraženo.

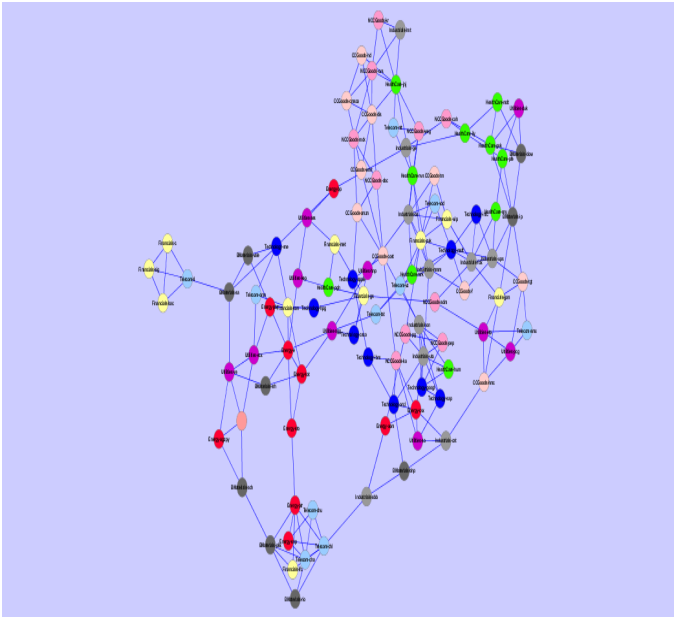
Daljom analizom rezultata K-means algoritma vidimo na koji način su se kompanije grupisale u 10 klastera kao rezultat primene K-means algoritma (tabela 3).

TABELA 3. Broj klastera – K – means alogoritam

Cluster 0	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
E-ogzpy	E-cvx	BM-dow	I-ba	CCG-amzn
BM-ach	E-xom	BM-ip	I-hon	CCG-cmsca
BM-mt	I-cat	I-ge	I-lmt	CCG-cost
BM-vale	NCCG-pg	I-fdx	I-mmm	CCG-dis
U-ve	U-so	I-ups	I-utx	CCG-hd
		CCG-f	NCCG-kr	CCG-wmt
		CCG-tgt	NCCG-pep	NCCG-abc
		NCCG-cah	NCCG-weg	NCCG-cvs
		F-met	F-puk	NCCG-ko
		HC-gsk	F – wlp	NCCG-mck
		HC-ily	HC-hum	T-apple
		HC-mdt	HC-jnj	T-sap
		HC-pfe	HC-nvs	T-ibm
		HC-sny	T-googl	T-orcl
		T-intc	TEL-ntt	
		U-duk		
Cluster 5	***Cluster 6***	***Cluster 7***	***Cluster 8***	***Cluster 9***
BM-aa	CCG-tm	E-ptr	BM-bhp	E-bp
F-aig	HC-mrk	E-snp	I-abb	E-e
F-bac	T-msft	E-sto	CCG-hmc	E-pbr
F-c	TEL-tat	BM-pkx	NCCG-adm	E-tot
TEL-ti	TEL-vod	BM-rio	F-gs	BM-srh
	TEL-vz	F-ifc	F-jpm	F-san
	U-hnp	TEL-cha	T-hpg	HC-pgh
	U-ngg	TEL-chl	TEL-amx	T-csco
		TEL-chu	U-etp	T-sne
			U-pcg	TEL-oran
				U-aes
				U-ecx
				U-kep

Na osnovu rasporeda kompanija u klasterima, ne može se zaključiti da se sve kompanije iz jednog sektora slično ponašaju. Po nekoliko kompanija iz jednog sektora može imati blisku vezu sa kompanijama iz drugih sektora. U klasteru 0 - energetska kompanija Gazprom pokazuje različitost. Grupisana je van energetskog sektora sa još dve kompanije iz sektora osnovnih materijala i jednom iz uslužnog sektora. Može se zaključiti da cene akcija ove kompanije ne kreću slično kretanjima ostalih kompanija iz energetskog sektora. Kompanije iz energetskog sektora su se podelile u klastere 7 i 9. Tri kompanije (Petro China, Statoil, China Petroleum) iz klastera 7, kao i 4 kompanije (Petro Brasileus, Total SA, Exxon, British Petrol) iz klastera 9, pokazuju sličnosti u tržišnim kretanjima. U drugom klasteru dosta sličnosti u kretanju su pokazale 5 kompanija zdravstvenog sektora. Treći klaster izdvaja grupisanje 5 kompanija industrijskog sektora. U četvrtom klasteru sličnost su pokazale kompanije iz sektora potrošnih i nepotrošnih dobara. Dosta različito se klasteruju kompanije iz finansijskog sektora pa se može zaključiti da nema neke međuzavisnosti u okviru ovog sektora. Vodeće kompanije tehnološkog sektora IBM, Oracle, Sap, Apple grupisale su se u klasteru broj 4 i pokazuju sličnosti u kretanju cena akcija. Preostale kompanije npr. Microsoft pripada klasteru 6. gde su grupisane tri kompanije telekomunikacija, dve uslužne, po jedna zdrastvena i kompanija iz sektora potrošnih dobara. Vrednost dobijena merom validacije 0,341 i ukazivala je da će kompanije biti na ovakav način grupisane. Između nekih kompanija koje pripadaju istom sektoru kao što smo videli postoji sličnost, ali ne i generalno pravilo da sve kompanije iz jednog sektora uvek imaju slično kretanje cena akcija. U daljim istraživačkim radovima pored vodećih kompanija, uključićemo i kompanije koje su sa nižim stopama prinosa i ponoviti postupak klasterovanja.

Za pregledniji prikaz rezultata klasterovanja u radu smo primenili i vizuelizaciju podataka 3NN grafom (eng. Three Nearest Neighbour) [16]. Po strukturi grafa na slici 1 vidimo da su neki delovi više povezani i oni predstavljaju klastere. Vizuelnim presekom grafa može se izvršiti grupisanje u klastre. Za pravljenje slike korišćen je Cytoscape [17]. Privredni sektori su obeleženi različitim bojama – Npr. crvenom bojom je označen sektor energetike, zelenom – zdravstveni, plavom – tehnološki, svetlo plava – telekomunikacije, ljubičasta – uslužni. Posmatranjem grafa vidimo koji sektori i koje kompanije su najbliži susedi.

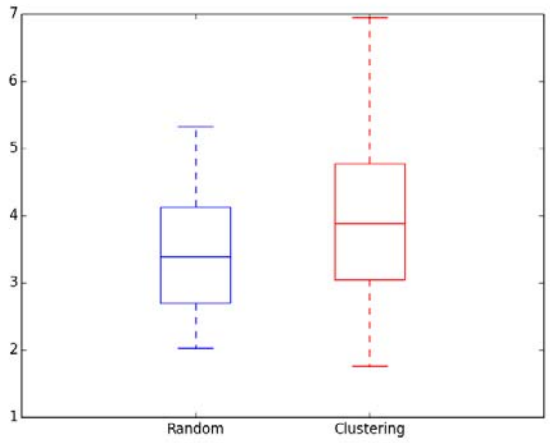


Slika 1. 3NN – tri najbliža suseda za svaku kompaniju

DIVERZIFIKACIJA RIZIKA

Ulaganje na finansijskim berzama je kompleksan i neizvestan posao. Izvršiti pravilan izbor u akcije u koje investirati zahteva detaljne analize. Mnogi investicioni analitičari se trude da ostvare optimalan portfolio i diverzifikuju rizik ulaganja. Drugi deo eksperimenta je bio kreiranje portfolia – skupa akcija u koje bi investirali.

Osnovni cilj klaster analize u radu je uspostavljanje optimalnog portfolia, idealne kombinacije akcija posmatranih kompanija sa stanovišta rizika, prinosa i proporcije u portfoliu koja obezbeđuje najveći stepen korisnosti. Ona ispunjava kriterijum minimiziranja rizika diverzifikacijom portfolia. Izbor za ulaganje izvršili smo na osnovu grupisanih kompanija u klastere. Iz svakog klastera birali smo po jednu hartiju od vrednosti na slučajan način i formirali portfolio od 10 akcija. Postupak izbora ponovili smo u 100 iteracija, kako bi bolje estimirali prinos od investicije. Druga mera kod kreiranja portfolia hartija od vrednosti je izbor 10 akcija slučajnom metodom. Poređenje dva pristupa za izbor u koje hartije od vrednosti investirati izvršili smo na osnovu prosečnog procentualnog dividendnog prinosa.



Slika 2. Diverzifikacija rizika

Rezultati analize pokazuju da K-means klaster analiza daje prosečno bolji prinos u poređenju sa metodom potpunog slučajnog izbora hartija od vrednosti (slika 2.). Prosečna zarada od dividendnog prinosa primenom metode sa klasterovanjem je 3.899%. Metoda slučajnog izbora ostvarila je zaradu od 3.479%. Viši prinos je pokazatelj doprinosa K-means klasterovanja u kreiranju optimalnog portfolia i diverzifikaciji rizika.

V. ZAKLJUČAK

Rezultati prikazani u ovom radu i pregled relevantne literature jasno ukazuju da je klaster analiza metoda sa velikim potencijalom za podršku odlučivanju o ulaganju na berzi. Viši prinos K-means klasterovanja u odnosu na metodu slučajnog izbora pokazatelj je doprinosa u kreiranju optimalnog portfolia i diverzifikaciji rizika. Ipak, ova tehnika povlači odgovornost istraživača, pa je nužna određena doza opreza prilikom njenog korišćenja. Ukoliko se pravilno koristi, ova analiza ima potencijal da otkrije nova saznanja koja do tada nisu otkrivena pomoću drugih metoda. U daljem istraživačkom radu pored najuspešnijih kompanija u analizu ćemo uključiti i kompanije sa nižim stopama prinosima i ponoviti postupak klasterovanja sa proširenim uzorkom. Izbor kompanija i njihovih akcija za portfolio bi mogle biti kompanije koje su bliže klaster centrima. Na taj način bi analizirali da li se mogu ostvariti još više i sigurnije stope prinosa od ulaganja u rizičnijim uslovima. Pored algoritma k-means, analize ćemo proširiti i sa eksperimentima i primenom algoritma - Affinity Propagation. Daljim istraživanjima nastojaćemo da pronađemo što sigurnije pristupe za investiciona ulaganja.

ZAHVALNICA

Posebno se zahvaljujemo kompaniji „Naftna industrija Srbije“ i timu Sektora za treninge i razvoj koji su pružili podršku za istraživanja i učešće na naučnoj konferenciji Infoteh 2015.

LITERATURA

[1] M. Craven, J. Shavlik, „Using Neural Networks for Data Mining”, Computer Science Dep., Univ. of Wisconsin *Future Generation Computer Systems*, vol. 13, pp. 211-229, 1997.
[2] Hartigan, A. Wong, „A k-means clustering algorithm”, Applied Statistics, vol. 28, pp. 100-108, 1979.
[3] PyScripter: code.google.com/p/pyscripter/
[4] C. Tryon, „Cluster analysis”, Ann Arbor: Edwards Brothers, 1939.
[5] R. Sokal, H. Sneath, „Principles of numerical taxonomy” Freeman, San Francisco – London, 1963.
[6] N. Jardine, R. Sibson, „Mathematical taxonomy” Wiley, New York, 1971.
[7] J. Bijnen, „Cluster analysis” Tilburg University Press, Tilburg, Netherlands, 1973.
[8] R. Anderberg, „Cluster analysis for applications”, Academic Press, New York, 1973.
[9] A. Hartigan, „Clustering algorithms”, Wiley, New York, 1975.
[10] V. Tolaa, F. Lilloc, M. Gallegatia, R. Mantegnac, „Cluster analysis for portfolio optimization”, Dipartimento di Economia, Università Politecnica delle Marche, Journal of Economic Dynamics and Control, vol. 32, pp. 235-258, 2008.
[11] S. Nanda, B. Mahnaty, M. Tiwari, „Clustering Indian stock market data for portfolio management”, Department of Industrial Engineering and Management, Indian Institute of Technology, Expert Systems with Applications, vol. 37, pp. 8793-8798, 2010.
[12] T. Kohonen, „Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics, vol. 43 (1), pp. 59–69, 1982.
[13] J. Bezdek, „Pattern Recognition with Fuzzy Objective Function Algorithms”, Kluwer Academic Publishers, 1981.
[14] Euclidean distances: http://quest4rigor.com/tag/euclidean-distances/
[15] A. Pablo, A. Estévez, M. Tesmer, A. Claudio, A. Perez, J. Zurada, „ Normalized Mutual Information Feature Selection”, IEEE Transactions on Neural Networks, vol. 20 (2), 2009.
[16] S. Dudani, Hughes Research Laboratories, Malibu „The distance – Weighted k- Nearest- Neighbor Rule”, Systems, Man and Cybernetic, vol. 6, 2010.
[17] Cytoscape: http://www.cytoscape.org/

ABSTRACT

Clustering is usually the first step in big data research. It has significant importance in complex data processing which is typical for financial markets. This paper offers cluster analysis of share price fluctuation for 100 companies in 10 different industries, over 10 years. Alhorithm K-means has been applied to selected data set by which companies with similar characteristics have been assembled in clusters. Clustering results are compared to sector affiliation of given companies in order to determine if companies from the same sector utilize similar market behavior. Second part of experimental analysis has been dedicated to implementation of research results in risk diversification. It has been proven that larger yield can be achieved by cluster investment deployment.

CLUSTER ANALYSIS IN INVESTMENT RISK DIVERSIFICATION'S PURPOSE

Jelena Brdar, Zita Bošnjak