**Primena K-Means algoritma za grupisanje korisnika u socijalnim mrežama**

**Sadržaj** - Razvoj tehnologije, uslovio je masovno korišćenje pametnih telefona i socijalnih mreža. Veliki broj ljudi poseduje profile na socijalnim mrežama, tako da je moguće prikupiti mnogo informacija o njihovom ponašanju ili kretanju. Ovaj rad ima za cilj da opiše kako informacije koje možemo prikupiti sa mobilnih uređaja možemo upotrebiti u svrhu pronalaženja šablona u socijalnim mrežama. Takve informacije su vrlo značajne jer poznavanje takvih šablona jako mnogo pomaže u daljim analizama kao što su markentiške svrhe, svrha procene biračkog tela, sablon po kojem bi eventualno mogao da se širi neki novi virus itd.

**Uvod** – Razvoj interneta prouzrokovao je masovno korišćenje socijalnih mreža i mobilnih uređaja. Ljudi svakodnevno pristupaju socijalnim mrežama i ostavljalju veliku količinu informacija o sebi. Očigledno je da sve da sve te informacije mogu biti veoma korisne u različitim analizama, međutim potrebno je iz ogromne količine nestruktuiranih podataka izvući relevantne podatke. Pored toga jako je bitno dobro klasifikovati podatke koji se kasnije koriste u daljim analizama.

Glavna ideja je korišćenje prostorne i vremenske korelacije u svrhu dobre kalsifikacije korisnika. Takođe su predstavljeni pojedini predlozi kako bi proces klasifikacije mogao da se usavrši i pruži više korisnih informacija u procesu pronalaženja skrivenog znanja. Iako postoji veliki broj algoritama za klasifikaciju, ovde je korišćen K-Means algoritam. On ima neke mane, ali se ipak jako često koristi zbog svoje brze konvergencije i zbog toga što može postići dobre rezultate pažljivim izborom atributa za klasifikaciju.

**Opis korišćenog algoritma** – K-Means je algoritam čija glavna uloga je klasifikacija n objekata u k grupa. Cilj je da se u okviru iste grupe (engl. cluster) nađu objekti koji su slični. Na prvi pogled čini se da je taj zadatak jednostavan i da se ne razlikuje od proste klasifikacije. Ali glavna razlika K-Means od ostalih algoritama koji vrše prostu klasifikaciju jeste njegova sposobnost da pre izvršavanja klasifikacije nema već postojeće grupe u koje treba da rasporedi objekte već u toku svog rada „uči“ kako da izvrši klasifikaciju.

Kao meru sličnosti objekata ovaj algoritam najčešće koristi njihovu euklidsku udaljenost u p-dimenzionalnom prostoru, gde p predstavlja broj atributa objekata koji učestvuju u klasifikaciji. Formula za izračunavanje udaljenosti u   
p-dimenzionalnom prostoru je:

(1)

Grupisanje se vrši minimizacijom rastojanja između objekata tako da na kraju u jednoj grupi ostaju samo oni objekti koji su prostorno najbliži.

Koraci K-means algoritma prikazani su u nastavku:

* Korak 1. Odrediti broj grupa za kalsifikaciju.
* Korak 2. Izabrati k objekata iz skupa nad kojim treba izvršiti klasifikacju, koji će biti predstavnici svoje grupe (eng. centroids).
* Korak 3. Izračunati rastojanje ostalih (n-k) objekata do svih centroida i dodeliti ih najbiližem centroidu.  
  Nakon dodela objekata centroidima ponovo reklakulisati nove vrednosti centroida.
* Korak 4. Ponavljati korak 3 sve dok ne dođe do kovergencije, tj. dok objekti menjaju pripadnost klasteru.

Prednosti ovog algoritma su skalabilnost kao i brza konvergencija koja u prosečnom slučaju iznosi O(k\*n\*I) gde je k broj grupa za klasifikaciju, n broj objekata za klasifikaciju, dok je I broj iteracija algoritma (obično je I mnogo manje od n).

Glavni nedostaci su:

* potreba da za izborom broja k pre početka rada algoritma.
* problem kod grupisanja nekonveksnih grupa objekata.
* problem lokalnih minimuma jer postoji verovatnoća da je dostignut uslov konvergencije, a da to nije I najbolje rešenje.
* Rezulatat algoritma zavisi od početnog izbora centroida, pa je vrlo bitno na pažljiv način inicijalizovati algoritam da bi se pružili što bolji rezultati klasifikacije.

Nakon opisa samog algoritma postavlja se pitanje kako je moguće proceniti kvalitet klasifikacije koju je algoritam izvršio. U tu svrhu najčešće se koristi Rosseuow funkcija za procenu tačnosti klasifikacije koja je prikazana u nastavku:

(2)

Članovi formule broj 2 su:

* b(i) je prosečna udaljenost i-tog objekta od ostalih objekata iz drugih grupa.
* a(i) je prosečna udaljenost i-tog objekta od ostalih objekata u okviru iste grupe.

Kao što se vidi vrednost formule 1 može biti u intervalu [-1, 1]. Ukoliko je vrednost bliska -1 to znači da klasifikacija nije dobro izvršena tj. I-ti element ne pripada svom klasteru. Ukoliko je vrednost bliska 1 znači da je i-ti objekat dobro klasifikovan i da se stvarno nalazi u okviru odgovarajućeg klastera. Ako se za neki objekat dobije vrednost 0 to znači da se on nalazi na granici između dve grupe.

Smatra se da je klasifikacija doro izvršena ako je prosečna vrednost formule 2 za klasifikovane objekte veća od 0,5.

**Postojeća poboljšanja algoritma** – Kao što je već navedeno za rad K-Means algoritma potrebno je izabrati broj klastera k pre pokretanje algoritma. Da bi se što bolje procenio broj k, ovaj algoritam se pokreće za različite vrednosti k, a zatim se koristeći Rosseuow funkciju procenjuje tačnost klasifikacije. Zatim za ono k koje je dalo najbolju vrednost Rosseuow funkcije pokreće se algoritam veći broj puta, jer čak I za isti broj k, algoritam može imati različite rezultate.   
To se dešava usled nasumičnog izbora inicijalnih centroida. Kao posledica toga jeste da inicijalni centroidi često pripadaju jednoj logičkoj grupi, što usporava samu konvergenciju, a često dovodi i do toga da rezultat klasifikacije ne bude optimalno rešenje već bude rezultovano nekim lokalnim minimumom. Međutim čak i veliki broj pokretanja algoritma ne vodi nužno do najboljeg rešenja.

Ovo se donekle može prevazići pažljivijim izborom inicijalnih centroida. Osnovna ideja jesta da se centroidi biraju na taj način da sa velikom verovatnoćom pripadaju različitim grupama koje će biti rezultat klasifikacije.

Logika izbora centroida koja ispunjava dati zahtev jeste da se inicijalni centroid bira proizvoljnom metodom, dok se svaki sledeći bira tako da bude najudaljeniji od svih već izabranih. Ovaj pristup omogućava dobru I brzu konvergenciju.

**Predložena ideja za poboljšanje algoritma** –Najveći nedostatak K-Means algoritma jeste nemogućnost detektovanja nekonveksnih logičkih grupa tačaka. Ideja za prevazliaženje ovog problema jeste konstrukcija minimalnog obuhvatnog stabla nad grafom čiji su čvorovi objekti za grupisanje, a euklidska rastojanja među čvorovima grane tog grafa. Nakon konstrukcije takvog stabla lako bi mogla da se odredi frekvencija ponavljanja grana sličnih dužina. Na taj način bi tačke koje sačinjavaju logičku grupu mogle lako da se grupišu u celine, jer su grane između tačaka unutar iste logičke grupe slične dužine. Na ovaj način bi se zaista grupisale tačke koje pripadaju grupama koje ne čine konveksne oblike. Ovaj algoritam se može donekle ubrzati na taj način što se obuhvatno stablo pravi od centroida, nakon pokretanja K-Means algoritma na setu podataka. Naravno u tom slučaju bi k trebalo da se izabere tako da je mnogo veće od stvarnog broja logičkih grupa.

**Primena algoritma** – Socijalnu mrežu možemo posmatrati kao n objekata koji svojim međusobnim vezama formiraju graf. Ukoiko se u svrhu klasifikacije koristi K-Means algoritam sem pažljivog izbora centroida jako je bitno koristiti dobre atribute objekata kao ulaz da bi nam tačnost klasifikacije bila sto bolja.  
Postojeća istraživanja koriste uglavnom činjenicu da veliki broj ljudi poseduje mobilne uređaje I da se na osnovu podataka sakupljenih sa mobilnih uređaja mogu otkriti šabloni u ponašanju korisnika.

Pristup u kome se kao ulaz algoritama koristila geografska pozicija ljudi u svrhu klasifikacije uglavnom su davale dobre rezultate u pogledu da se zaista u okviru iste grupe klasifikuju ljudi koji provode dosta vremena u istom prostoru, ali takav podatak ne daje dovoljno informacija o vrstama veze koje imaju ljudi u okviru iste grupe.   
Osnovna ideja za poboljsanje analize veza među ljudima jeste sem prostorne korelacije uključiti I vremensku korelaciju. Tačnije ako kao jos jedno dimenziju algoritma upotrebimo i vremensku dimenziju na bolji način možemo da sagledamo da li je neka veza među ljudima čisto poslovna ili se pored toga oni druže I van posla. Ideja jeste da ako neki ljudi dele prostor u toku radnog vremena, a ne I nakon radnog vremena postoji velika verovatnoća da među njima postoji poslovni odnos, a ne i prijateljstvo. Dok ukoliko neki ljudi dele prostor I nakon radnog vremena, velika je verovatnoća da među njima postoji prijateljski odnos. Ovakve informacije mogu biti jako korisne, jer ako dvoje ljudi radi sličan posao, I druži van posla sa nekim od kolega postoji velika verovatnoća da ukoliko budu raspoređeni u isti tim jos bolje obavljaju posao. Takođe ukoliko neki ljudi imaju sličnu prostornu korelaciju, ali se ne poznaju zasita postoji dobra šansa da oni mogu ostvariti dobar kontakt u stvarnom životu.

**Zaključak** - U ovom radu opisana je primena K-menas algoritma u klasifikaciji krisnika socijalnih mreža bazirana na geografskoj poziciji korisnika. Osim opisa samog algoritma kao i nekih njegovih nedostataka i postojećih rešenja opisan je I značaj izbora ulaznih parametera. Jer pogrešan izbor parametara klasifikacije, može da grupiše korisnike po nekom kriterijumu koja nije od kvalitativnog značaja i koja ne može pomoći u daljoj analizi podataka. Analizom vrste veza medju ljudima bitna je iz razloga da znajući vrste tih veza lako možemo otkriti na koj način pojedine informacije mogu da propagiraju od jednog do drugog korisnika.

**Reference**

1. Krebs, V. Social Network Analysis, A Brief Introduction. Dostupno na http://www.orgnet.com/sna.html, 2014.
2. http://en.wikipedia.org/, 2014.
3. http://www.onmyphd.com/?p=k-means.clustering
4. http://www.naftaliharris.com/blog/visualizing-k-means-clustering/, 2014.
5. Pang-Ning Tang, Michael Stenbach, Vipin Kumar, Introduction to Data Mining, 2005.
6. Eagle, E. MIT Medialab: Reality Mining, available at: http://reality.media.mit.edu/. Massachusetts Institute of Technology, City,

2009.