PRIMENA ALGORITMA *SUPPORT VECTOR MACHINE* U SOCIJALNIM MREŽAMA

APPLICATION OF SUPPORT VECTOR MACHINE ALGORITHM IN SOCIAL NETWORKS

Petar Mitrović,
Univerzitet u Beogradu, Elektrotehnički fakultet

Sadržaj – *Pronalaženje skrivenog znanja je oblast u koju se poslednjih godina intenzivno ulaže s ciljem da se ogromne količine podataka koje su dostupne iskoriste na pravi način za dobijanje novih saznanja. Socijalne mreže savršeno se uklapaju u ovaj koncept kao nepresušni izvori informacija do kojih bi inače bilo jako teško doći. Raznolikost sadržaja i redovno osvežavanje čine socijalne mreže idealnim mestom za mašinsko učenje. Postoji veliki broj algoritama koji se danas koriste u najraznovrsnije svrhe, a posebno mesto među njima zauzima Support Vector Machine algoritam.*

Abstract – *Data mining is an area that is heavily invested in over the last years with a goal to use vast amounts of data currently available online in the right way for gaining new knowledge. Social networks fit this concept perfectly as an inexhaustible source of information that would usually be hard to get. Variety of content and regular refresh of data make social networks an ideal environment for machine learning. There are a number of algorithms used for different purpose, but the special place belongs to Support Vector Machine algorithm.*

1. UVOD

Večna težnja da se unapred otkrije sledeći korak ili rezultat ogleda završila se potragom za alternativnim rešenjima. Ubrzo se shvatilo da je to nemoguće, ali je isto tako postalo jasno da se poznavanjem rezultata prethodnih koraka (ogleda) sa manje ili više uspešnosti može "predvideti" i sledeći rezultat. Tako su nastali *data mining* i *machine learning*.

Tokom godina razvijeni su besprekorni algoritmi za pronalaženje znanja i treniranje, pa je u današnje vreme najteži poduhvat pribaviti kvalitetan skup podataka kako bi se algoritam što kvalitetnije "naučio" kako da deluje u određenim situacijama. Tu na scenu stupaju socijalne mreže. Kvalitetnom analizom ovaj skup naizgled nesređenih i nepovezanih podataka može predstavljati pravo bogatstvo i idealnu osnovu za treniranje algoritama. Tekstualni, video i fotografski sadržaji pogoduju najrazličitijim algoritmima za pronalaženje skrivenog znanja. Kako je ipak tekstualni sadržaj onaj koji preovlađuje na socijalnim mrežama, odabran je *Support Vector Machine* algoritam kao odličan klasifikator teksta.

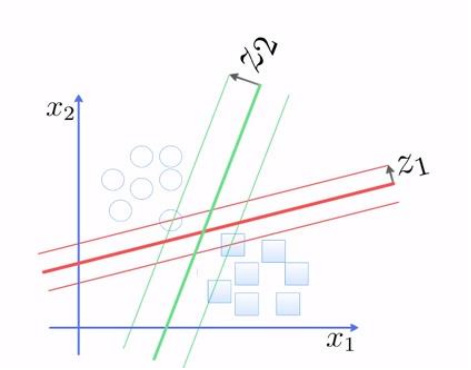
2. SVM ALGORITAM

SVM algoritam definisan je pre 20 godina od strane Vladimira Vapnika (ruski naučnik), a objavljen je 1995. godine uz asistenciju Korine Kortes (danska naučnica). Nalazi se na listi TOP 10

algoritama za pronalaženje skrivenog znanja koja je napravljena na Univerzitetu u Vermontu. Namenjen je pre svega za klasifikaciju, pa će i ovde biti korišćen u tom kontekstu. Pored klasifikacije teksta, pogodan je i za klasifikaciju slika, proteina, kao i rukopisa.

3. KAKO RADI SVM ALGORITAM

Algoritam je definisan na jednostavnoj ideji – definisati hiperravan (ili skup hiperravni) koja klasifikuje sve vektore *feature*-a iz trening skupa u dve (ili više) klasa. Potrebno je ispuniti još samo jedan uslov – ukoliko postoji više takvih hiperravni (ili skupova hiperravni), potrebno je odabrati onu sa maksimalnom marginom (maksimalnim rastojanjem do najbliže tački u svakoj od klasa). Na slici 1 ilustrovan je jednostavan primer problema ovog algoritma.



Slika 1. Definicija problema SVM algoritma

Ako x_1 i x_2 predstavljaju dva *feature*-a i želimo da sve tačke iz skupa podelimo u dve klase (kvadrati i krugovi na slici), potrebno je definisati pravu koja će to učiniti. Kako takvih pravih ima beskonačno mnogo, potrebno je odrediti onu sa najvećim rastojanjem do obe klase. To je, očigledno, "zelena" prava. Na osnovu njene jednačine

$$g(\vec{x}) = \vec{\omega}^T \vec{x} + \omega_0$$

svi vektori mogu se klasifikovati na sledeći način:

$$g(\vec{x}) \geq 1 \; \forall \vec{x} \; \epsilon \; class \; 1$$

$$g(\vec{x}) \leq -1 \; \forall \vec{x} \; \epsilon \; class \; 2$$

Ovako dobijeni uslovi na jednoznačan način određuju pripadnost svake tačke nekoj od klasa. Jednačina predstavlja *support vector machine*.

4. PRIMERI PRIMENE SVM ALGORITMA NA SOCIJALNE MREŽE

SVM algoritam do sada je najčešće primenjivan na *Twitter* socijalnoj mreži, zbog njenog pretežno tekstualnog karaktera. Analizom korisničkih *tweet*-ova moguće je otkriti puno informacija o njihovom raspoloženju i namerama. Jednostavna implementacija SVM algoritima dovoljna je za efikasnu klasifikaciju i pouzdane rezultate.

Jedan od čestih ciljeva prilikom analize korisničkih *tweet*-ova je da se na osnovu što manje podataka odredi da li je korisnik dobro (pozitivno) ili loše (negativno) raspoložen po nekom pitanju. Na osnovu malog seta podataka potrebno je trenirati SVM algoritam i analizirati rezultate. *Feature*-i se biraju na osnovu dobro poznatih fraza za izražavanje zadovoljstva ili nezadovoljstva.

Ovakvo istraživanje sprovedeno je na *Trinity College*-u i rezultati su bili odlični – oko 75% uspešnosti, iako je skup podataka bio vrlo mali.

Drugi primer je korišćenje korisnika *Twitter*-a da bi se detektovale prirodne katastrofe. Poenta u ovom slučaju nije u prevenciji, već u brzom širenju vesti da se desila katastrofa. Ovaj primer specifičan je i po tome što "meša" dve primene, na socijalne i senzorske mreže. Naime, svaki

korisnik u ovom slučaju je jedan senzor, koji "okida" tako što napiše *tweet* da se desila katastrofa (zemljotres).

Rezultati ovog istraživanja pokazuju da se u oko 96% slučajeva za samo 10 minuta može utvrditi da je došlo do katastrofe samo na osnovu pisanja korisnika. Ovakav rezultat pokazuje da ovaj algoritam predstavlja jako pouzdan izvor informacija u kriznim situacijama.

Prethodna istraživanja pokazuju da se uz relativno malo truda i pažljivo odabran skup podataka mogu postići jako precizni rezultati korišćenjem SVM algoritma. To je ohrabrilo autora da predloži nov model primene, i to na socijalnu mrežu na kojoj se do sada SVM algoritam nije koristio.

5. PRIMENA SVM ALGORITMA NA *FACEBOOK*-U

Facebook je najpopularnija socijalna mreža na svetu. Usled neverovatne brzine rasta, programeri se suočavaju sa najrazličitijim problemima. Ono što je postalo posebno problematično je činjenica da ljudi imaju previše prijatelja (u proseku preko 300) i da imaju previše akcija dnevno (u proseku preko 20). To znači da bi svaki korisnik koji želi da vidi sve akcije svojih prijatelja morao da dnevno pregleda oko 6000 različitih sadržaja, što je praktično nemoguće čak i ako bi neprekidno sedeo za računarom.

Facebook je zbog toga pribegao jednom nesavršenom algoritmu – prikazati korisniku akcije samo nekih prijatelja, a za ostale korisnik mora sam da pretraži sadržaje. Ovaj princip je nesavršen iz prostog razloga što može dovesti do toga da se korisniku prikazuju sadržaji prijatelja sa kojima ne deli interesovanja, pa stoga on zapravo dobija sadržaje koji ga ne zanimaju (upravo ovo se desilo autoru rada, pa je došao na ideju za novi algoritam).

Ideja je jednostavna – potrebno je neko vreme pratiti ponašanje korisnika na *Facebook*-u. Za svaki *like* i pozitivan komentar potrebno je beležiti sadržaj na koji se odnosi kao pozitivan *feature*, a za svaki negativan komentar (*dislike* ne postoji na *Facebook*-u, što je jedna od glavnih zamerki korisnika) sadržaj beležiti kao negativan *feature*.

Nakon određenog vremena, kada se skup prikupljenih podataka može smatrati relevantnim i dovoljnim, potrebno je istrenirati SVM algoritam tako da svaki sadržaj klasifikuje kao relevantan ili irelevantan na osnovu unapred poznatih podataka. Na taj način korisnik bi na osnovu svojih interesovanja mogao da "gradi" skup sadržaja koji mu se pojavljuju u *news feed*-u i uživa u stvarima koje ga zanimaju.

6. ZA I PROTIV OVAKVOG REŠENJA

Ovakav pristup sigurno bi unapredio kvalitet boravka korisnika na mreži. Zadovoljstvo korisnika bi bilo veće, dok bi relevantni podaci stizali do više ljudi. Takođe, posledice greške u radu algoritma nisu fatalne. Ako se prikaže irelevantan sadržaj, lako ga je preskočiti (to je scenario koji i danas imamo). Sa druge strane, ako se slučajno neki relevantan sadržaj preskoči, on nije trajno izgubljen, jer ga korisnik sa lakoćom može pronaći.

Osnovni problem bile bi performanse. Kao što je ranije rečeno, potrebno je obraditi oko 6000 najrazličitijih sadržaja, a korisnik nema puno vremena da čeka na obradu. Stoga je potrebno dodatno investirati u brzinu klasifikacije, kako sa softverske strane, tako i sa hardverske.

7. ZAKLJUČAK

Na jednostavnom primeru pokazano je kako efikasan algoritam kao što je SVM može poboljšati kvalitet usluge i doprineti boljem utisku korisnika. *Facebook* je mreža koja je unela nešto novo u svet u 21. veku, ali problemi koji su stigli kasnije moraju se rešavati u hodu, da se ne bi desila sudbina prethodnika kao što je *MySpace*. Malim idejama kao što je ova moguće je vrlo lako podstaći korisnike da više koriste mrežu i da u njoj uživaju.

LITERATURA

[1] Campbell, C., Cristianini N., Simple Learning Algorithms for Training Support Vector Machine, University of Bristol

[2] Sung R., What do you Tweet? An Analysis of Twitter using Support Vector Machines, Trinity College, 2012.

[3] Sakaki T., Makoto O., Yutaka M., Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, University of Tokio

[4] Xindong W., Top 10 algorithms in Data Mining, University of Vermont

[5] www.wikipedia.org