Credit Card Fraud Detection

- Uvodni opis problema:

  Glavni cilj našeg problema je detekcija prevara kod transakcija napravljenim kreditnim karticama, te evaluacija različitih metoda na dobro i loše balansiranom setu podataka.

  Ovaj problem je preuzet sa kaggle-a, te se koristimo njihovim setom podataka koji je loše balansiran: 492 transakcije su prevare, dok je 284 315 valjanih transakcija.

  Osim što nas zanima koja je metoda najbolja za klasifikaciju prevare, želimo zaključiti koliko se dobro te metode provode ukoliko su trenirane na loše balansiranom i dobro balansiranom setu podataka.

- Cilj i hipoteze istraživanja problema:

  Metode koje su pogodne za klasifikaciju prevara će biti trenirane na dobro balansiranom i loše balansiranom setu podataka, te će se nakon toga gledati koliko se one dobro ili loše ponašaju na različito balansiranim setovima podataka.

  Loše balansirani set podataka je zapravo izvorni set podataka, u kojem snažno prevladavaju valjane transakcije. Dobro balansirani set podataka dobivamo balansiranjem seta izvornih podataka.

- Pregled dosadašnjih istraživanja:

  Dosadašnja istraživanja koja smo imali prilike pronaći su bila većinom sa kaggle-a, tamo se uglavnom problemu pristupalo tako da se skup podataka prvo izbalansira te se nakon toga metode treniraju na balansiranom skupu podataka. Podaci su balansirani *undersampling* metodom, no korištenjem iste se gubi veliki broj podataka. Nas zanima koliko smanjenje izvornih podataka ima utjecaj na metode u odnosu na to da se iste treniraju na loše balansiranom setu podataka.

  Prema prijašnjim istraživanjima metode koje kanimo koristiti za klasifikaciju prevara se puno bolje ponašaju ukoliko su trenirane na dobro balansiranom setu podataka. Kod nekih od njih performanse nakon učenja na loše balansiranom skupu podataka mogu se poboljšati tako da se daju manja i veća težina većinskoj odnosno manjinskoj klasi. Dakle, pogreška se kažnjava različito s obzirom na to da li je krivo klasificirana manjinska ili većinska klasa.

- Materijali, metodologija i plan istraživanja:
  - Opis podataka:

    Svi podaci su numerički te ne nedostaje niti jedan podatak.

    Podaci se sastoje od V1, …, V28, Time, Amount i Class stupaca od kojih svaki sadrži jednak broj opservacija. Broj 1 u Class stupcu označuje prevaru, dok 0 označuje valjanu transakciju.

    Zbog zaštite podataka značajke V1,…,V28 su transformirane PCA algoritmom, ostale dvije značajke Time i Amount nisu transformirane. Kako PCA algoritam svojim transformacijama normalizira podatke, sami smo normalizirali samo Time i Amount. Korelacijskom matricom smo ispitali korelacije između svih 30 značajki, te očekivano dobili da V1,…, V28 nisu

međusobno korelirane, dok je koreliranost prisutna kod parova koji uključuju Time i Amount značajke.

- Metodologija:

  Loše balansiran set podataka je podijeljen na 5 skupova za treniranje i testiranje (kasnije ćemo koristiti CV metodu) u omjeru 4:1. Prilikom podjele omjer primjera koji pripadaju klasi prevare i klasi valjane transakcije u skupovima za treniranje i testiranje je jednak omjeru dobivenom na prvobitnim podacima *(stratified split).*

  Prvo smo analizirali prvobitan loše balansiran set podatka za treniranje. Osim utvrđivanja korelacija među značajkama, procjenama funkcija gustoća smo određivali koliko se razlikuju distribucije prevara i valjanih transakcija po značajkama. Većina rezultata nije imala značajnijih razlika distribucija, osim kod nekoliko iznimaka (V12 i V14).

  Nakon analize, proveli smo nekoliko različitih metoda za određivanje bitnih značajki: metoda univarijatnog odabira značajki, FDR metoda, metoda za odabir pomoću slučajnih šuma. Zbog transformacije provedene PCA algoritmom, dobili smo da FDR i univarijatni odabir značajki daju iste rezultate dok je metoda slučajnih šuma rezultirala nešto drugačijim ishodom. No u principu su sve metode označile iste značajke kao najvažnijima ili najmanje bitnima.

  Sljedeći korak u našem projektu je na temelju dobivenih rezultata odlučiti koje su značajke nebitne te onda na bitnim značajkama provesti metode.

  Nakon što smo izbalansirali set podataka(*undersampling* metodom) i podijelili ga u omjeru 4:1 na set za treniranje i testiranje, ista analiza je primijenjena bez većih razlika u rezultatima. Zatim kao i ranije provedene su metode za određivanje bitnih značajki.

  Nakon što ćemo iz dobivenih rezultata odlučiti koje su nam značajke i dalje bitne za balansirani set podataka, koristit ćemo iste metode (kao i u loše balansiranom slučaju) za klasifikaciju prevara.

- Metode koje će se koristi za klasifikaciju prevara:
  - Logistička regresija
  - SVM
  - Bayesove mreže
  - Neuronske mreže
- Ocjena uspješnosti:
  - Recall (#true positive/#positive)
  - Precision (#true positive/#predicted positive)
  - ROC-krivulja i površina ispod ROC-krivulje (prikladna za loše balansirane skupove podataka)

- Očekivani rezultati predloženog projekta:

  Očekujemo da će se metode uspješnije provoditi na balansiranom setu za učenje, dok za logističku regresiju očekujemo otprilike jednaku uspješnost.

- Popis literature:

  Kaggle https://www.kaggle.com/mlg-ulb/creditcardfraud

PMF materijali za strojno učenje

FER materijali za strojno učenje