# Šibko nadzorovano učenje z orodjem Snorkel

**Matjaž Kukar**[1], **Bjorn Bračko**[1]

[1]*Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana*
*E-pošta: matjaz.kukar@fri.uni-lj.si*

## Weakly supervised learning with Snorkel

*The explosion of available data and the complexity of prediction problems have increased the needs for large amounts of labelled data, posing a challenge to the supervised machine learning process. For this reason, weak supervision using noisy, inaccurately labelled training sets is an attractive alternative. We utilize the Snorkel framework which allows for easy inclusion of domain knowledge in form of label function abstractions. We train diverse predictive models and utilize them as weak classifiers — labelling functions for the Snorkel generative labeling model. We compare the performance of the models trained either with the true labels or with the Snorkel labels. We show that the latter have comparable or even better performance, especially with noise in data labels.*

## 1 Uvod

Strojno učenje je v zadnjih letih doživelo izjemen napredek, predvsem zaradi razvoja globokih nevronskih mrež. Uspešnost naučenih modelov je močno odvisna od količine in kakovosti označenih podatkov. Pridobivanje kvalitetnih oznak predstavlja ozko grlo v razvoju in uporabi modelov strojnega učenja, predvsem na področjih, kjer je le-to drago, zamudno ali celo nemogoče (npr. naknadno pridobivanje specifičnih diagnoz v medicini).

Šibki nadzor omogoča vsaj delno rešitev tega problema z uporabo šumnih, nenatančnih ali omejenih virov označevanja. Namesto zanašanja na izključno kvalitetne (a drage in težko dostopne) ročno označene podatke, šibki nadzor omogoča uporabo različnih virov informacij, kot so hevristična pravila, zunanje baze znanja, delno označeni podatki in rezultati drugih, manj zmogljivih napovednih modelov — označevalnih funkcij.

Snorkel [1] je eno izmed orodij, ki omogoča učinkovito izkoriščanje šibkega nadzora. Združuje različne šibke vire označevanja v enoten, verjetnostni model, ki ga nato uporabimo za označevanje podatkov. Osredotočili se bomo na uporabo orodja Snorkel v scenariju delno nadzorovanega (semi-supervised) učenja. Predpostavimo, da je le del podatkov označen in na njem naučimo raznolike modele kot označevalne funkcije. Le-te bomo s pomočjo orodja Snorkel uporabili za označevanje neoznačenih podatkov pri različnih stopnjah šuma in rezultate primerjali s tistimi, dobljenimi na originalnih oznakah podatkov.

## 2 Metode

### 2.1 Orodje Snorkel

Snorkel je ogrodje (framework) za šibko nadzorovano učenje, ki temelji na ideji podatkovnega programiranja [2, 3]. Uporabnikom omogoča, da namesto ročnega označevanja podatkov napišejo programske *označevalne funkcije*, ki izražajo njihovo domensko znanje. Te funkcije so lahko preprosta hevristična pravila ali pa poljubno kompleksni podprogrami, ki uporabljajo zunanje vire podatkov ali napovedne modele in se lahko tudi vzdržijo odgovora (abstain). Najpogostejše vrste označevalnih funkcij so [4]:

- **Na podlagi vzorca:** hevristike, ki temeljijo na vzorcih zajemajo pristope, kot so anotacije značilk in oblikovanje vzorcev, npr. besedni vzorci.
- **Oddaljeni viri nadzora:** nadzor na daljavo ustvarja učne oznake s hevurističnim usklajevanjem podatkov z zunanjo bazo znanja, kot sta npr. Wikipedia in DBpedia.
- **Šibki klasifikatorji:** ki ne zadostujejo v celoti za dano nalogo — npr. omejena pokritost, šumni, pristranski in/ali naučeni na drugem naboru podatkov.
- **Generatorji označevalnih funkcij:** ustvarijo več označevalnih funkcij iz enega vira, kot so množične oznake (vsak označevalec je obravnavan kot svoja označevalna funkcija).

### 2.2 Označevanje podatkov z orodjem Snorkel

Orodje Snorkel označevalne funkcije uporabi za označevanje neoznačenih podatkov. Ker so te funkcije lahko šumne in med seboj protislovne, Snorkel nauči *generativni model*, ki upošteva soglasja, nesoglasja in korelacije med pari označevalnih funkcij. Model združuje izhode označevalnih funkcij v verjetnostne oznake za vsak neoznačen primer, pri čemer uporabijo metodo za samodejno izbiro odvisnosti za modeliranje brez dostopa do pravih oznak, oceno psevdo verjetnost in optimizacijo v prostoru kompromisa med napovedno zmogljivostjo in časovno kompleksnostjo [4].

Oznake uporabimo za učenje končnega modela. Ta model se uči iz trdih ali verjetnostnih oznak (če učnio algoritem to omogoča), kar omogoča posplošitev preko specifičnih hevristik, ki jih izrazijo označevalne funkcije.

Glavna prednost ogrodja Snorkel je, da omogoča hitro in učinkovito ustvarjanje velikih količin učnih podat-

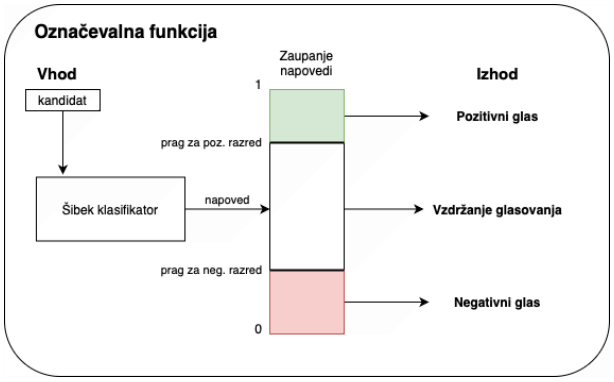
kov z minimalnim ročnim označevanjem. Modeli, naučeni na podatkih, označenih s pomočjo orodja Snorkel, so pogosto primerljivi ali celo boljši od modelov, naučenih s tradicionalnim nadzorovanim učenjem [4]. To kaže, da je šibki nadzor lahko učinkovita alternativa tradicionalnemu nadzoru, še posebej v primerih, ko je pridobivanje označenih podatkov težavno.

### 2.3 Strojno učenje in označevalne funkcije

Za učenje šibkih klasifikatorjev smo uporabili naslednje učne algoritme:

- Naivni Bayesov klasifikator
- Logistična regresija
- Naključni gozd
- Odločitveno drevo
- K-najbližjih sosedov (KNN)
- Ekstremni gradientni boosting (XGBoost)
- Nevronska mreža z dvema skritima plastema (NN)

Vsak učni algoritem smo pognali na vsakem problemu z različnimi hiper-parametri in izločili tiste klasifikatorje, ki se niso uspeli naučiti ničesar (točnost manjša od deleža večinskega razreda). Preostale smo pretvorili v označevalne funkcije tako, da smo s pomočjo interne optimizacije odločitve šibkih klasifikatorjev omejili z dvojnim pragom in vzdržanjem glasovanja, kot je predstavljen na sliki 1. V primeru, da je napovedna verjetnost klasifikatorja nad pragom za negativni razred in pod pragom pozitivnega, se označevalna funkcija vzdrži glasovanja.




Slika 1: Označevalne funkcije z dvojnim pragom in vzdržanjem glasovanja (za binarne šibke klasifikatorje). V primeru, da je zaupanje napovedi šibkega klasifikatorja nad pragom za negativni razred in pod pragom za pozitivni razred, se označevalna funkcija vzdrži glasovanja.

## 3 Materiali

Poskuse smo izvedli na v osnovi označenih podatkih in simulirali odsotnost oznak, kar nam omogoča objektivno vrednotenje rezultatov.

Uporabili smo na tri različne podatkovne nabore: METABRIC [5], Twitter sovražni govor [6] in lastna anonimizirana množica medicinskih podatkov Medic. Vse podatke smo predhodno ustrezno predpripravili, prečistili, prekodirali in atribute po potrebi ustrezno skalirali.

### 3.1 METABRIC

Podatkovna zbirka METABRIC [5], je zbirka podatkov, namenjena raziskovanju raka dojke. Vsebuje genomske in klinične podatke o približno 2000 bolnicah z rakom dojke. Napovedujemo atribut "splošno preživetje". Podatkovni nabor je relativno uravnotežen, z 58% razreda 0 in 42% razreda 1. Zbirka je zanimiva, saj je sestavljena iz treh modalitet (klinični, RNA in mutacijski atributi), kar pomeni, da lahko izkoristimo sposobnost orodja Snorkel, da za učenje označevalnega modela uporabimo dodatne modalitete, medtem ko za končni model uporabimo le eno (klinični atributi).

### 3.2 Twitter sovražni govor

Twitter sovražni govor [6] je zbirka 32000 tvitov, ki so klasificirani kot sovražni ali nesovražni. Vsebuje pa tudi okoli 17000 neoznačenih tvitov. Nabor označenih tvitov je neuravnotežen, sovražni govor predstavlja le 7% tvitov, zato ga s podvzorčenjem uravnotežimo. Na voljo imamo tudi pet označevalnih funkcij na podlagi vzorcev in hevristik.

### 3.3 Medic

Zadnji nabor je Medic, lastna anonimizirana množica medicinskih podatkov. Nabor sestavlja 4000 učnih primerov z 19 številskimi atributi, ki predstavljajo rezultate medicinskih preiskav. Oznake so binarne in predstavljajo prisotnost ali odsotnost bolezenskega stanje. 58% oznak pripada pozitivnemu razredu, 42% negativnemu razredu, na voljo pa imamo pa na voljo tudi označevalno funkcijo, ki temelji na medicinskem znanju.

## 4 Rezultati

### 4.1 Testna metodologija

Označene podatke smo razdelili na tri glavne dele [7]:

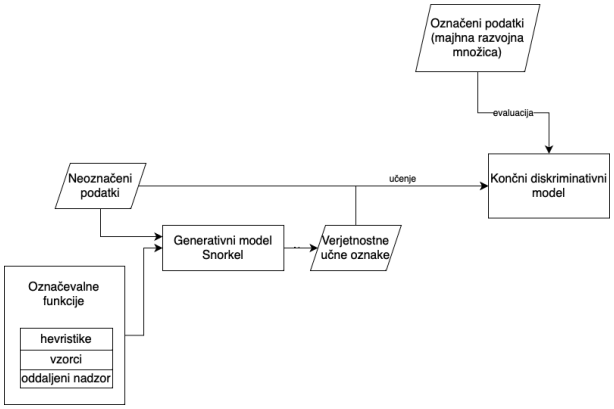
- **A1**: Podatki za učenje šibkih klasifikatorjev, ki služijo kot osnova za označevalne funkcije.
- **B1**: Neoznačeni podatki za učenje generativnega modela in končnega klasifikatorja.
- **B2**: Podatki za ocenjevanje končnih klasifikatorjev.

Z orodjem Snorkel in dobljenimi označevalnimi funkcijami označimo drugi del (B1) podatkovne množice. Oznake uporabimo pri učenju končnega modela, ki ga ovrednotimo na množici B2.

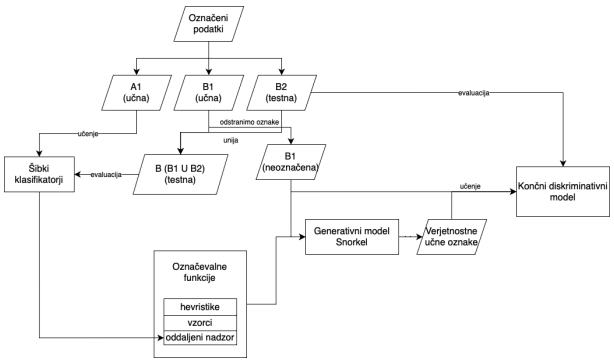
Množico $\mathbf{B} = \mathbf{B1} \cup \mathbf{B2}$ lahko uporabimo tudi za vrednotenje šibkih klasifikatorjev, nikakor pa ne za kakršno koli optimizacijo hiper-parametrov. V vseh poskusih smo zagotovili podobno velikost in porazdelitev razredov v vseh treh množicah. Slika 2 prikazuje diagram uporabe orodja Snorkel v realnem, slika 3 pa v našem testnem scenariju. Celoten eksperimentalni scenarij ponovimo desetkrat z naključnim razbitjem podatkov, rezultate F1-mere pa predstavimo s povprečjem in standardnim odklonom.

### 4.2 Rezultati končnih klasifikatorjev

V tabeli 1 so predstavljeni rezultati končnih klasifikatorjev za nabor Metabric, naučenih na množici B1 na




Slika 2: Diagram uporabe orodja Snorkel v realnem scenariju.




Slika 3: Diagram uporabe orodja Snorkel v našem eksperimentalnem scenariju.

kliničnih atributih s pravimi ali Snorklovimi oznakami, ter ovrednoteni na množici B2. Modeli so primerljivi, opazimo pa večjo razliko pri nevronskih mrežah, kjer so modeli s pravimi oznakami slabši.

| Model | Prave ozn. | Ozn. Snorkel |
|---|---|---|
| Logistična regresija | **0.776 ± 0.032** | 0.773 ± 0.028 |
| Naključni gozd | 0.763 ± 0.034 | **0.770 ± 0.021** |
| Odločitveno drevo | 0.747 ± 0.027 | **0.762 ± 0.035** |
| XGBoost | **0.785 ± 0.041** | 0.768 ± 0.042 |
| NN | 0.706 ± 0.035 | **0.772 ± 0.043** |

Tabela 1: Rezultati končnih klasifikatorjev za nabor Metabric, naučenih s pravimi ali Snorklovimi oznakami na množici B1, ovrednoteni na množici B2. Rezultati so predstavljeni kot povprečje in standardni odklon F1-mere za 10 različnih ponovitev.

V tabeli 2 so predstavljeni rezultati končnih klasifikatorjev za nabor Twitter sovražni govor, naučenih na množici B1 s pravimi ali Snorklovimi oznakami, ovrednoteni na množici B2. Modeli so glede na uspešnosti med seboj primerljivi.

V tabeli 3 so predstavljeni rezultati končnih klasifikatorjev za nabor Medic, naučenih na množici B1 s pravimi ali Snorklovimi oznakami, ovrednoteni na množici B2. V vseh primerih oznake označevalnega modela Snorkel vodijo v nekoliko boljše končne modele, kot prave oznake. Za nevronske mreže pa so te razlike še bolj očitne.

Generativni model orodja Snorkel se je pri označevanju

| Model | Prave oznake | Ozn. Snorkel |
|---|---|---|
| Naivni Bayes | **0.875 ± 0.004** | 0.872 ± 0.004 |
| Logistična regresija | 0.937 ± 0.001 | **0.938 ± 0.003** |
| Naključni gozd | **0.986 ± 0.002** | 0.982 ± 0.003 |
| Odločitveno drevo | **0.966 ± 0.002** | 0.964 ± 0.003 |
| KNN | **0.938 ± 0.004** | 0.927 ± 0.007 |
| XGBoost | **0.912 ± 0.003** | 0.911 ± 0.003 |
| NN | **0.979 ± 0.001** | 0.973 ± 0.002 |

Tabela 2: Rezultati končnih klasifikatorjev za nabor Twitter, naučenih s pravimi ali Snorklovimi oznakami na množici B1, ovrednoteni na množici B2. Rezultati so predstavljeni kot povprečje in standardni odklon F1-mere za 10 različnih ponovitev.

| Model | Prave ozn. | Ozn. Snorkel |
|---|---|---|
| Naivni Bayes | 0.83 ± 0.001 | **0.85 ± 0.001** |
| Logistična regresija | 0.86 ± 0.001 | **0.87 ± 0.001** |
| Naključni gozd | 0.85 ± 0.01 | **0.86 ± 0.01** |
| Odločitveno drevo | 0.84 ± 0.01 | **0.85 ± 0.01** |
| KNN | 0.82 ± 0.01 | **0.84 ± 0.001** |
| XGBoost | 0.86 ± 0.01 | **0.87 ± 0.01** |
| NN | 0.81 ± 0.01 | **0.866 ± 0.01** |

Tabela 3: Rezultati končnih klasifikatorjev za nabor Medic, naučenih s pravimi ali Snorklovimi oznakami na množici B1, ovrednoteni na množici B2. Rezultati so predstavljeni kot povprečje in standardni odklon F1-mere za 10 različnih ponovitev.

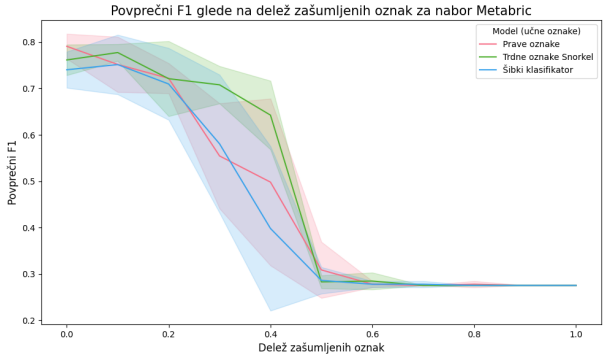
podatkov dobro izkazal, saj se performanse končnih modelov, naučenih na njegovih oznakah, ne razlikujejo bistveno od tistih, dobljenih na pravih oznakah. V nekaterih primerih so rezultati celo boljši, kar nakazuje na to, da so že v originalnih oznakah prisotne napake oz. šum.

### 4.3 Dodajanje šuma v oznake

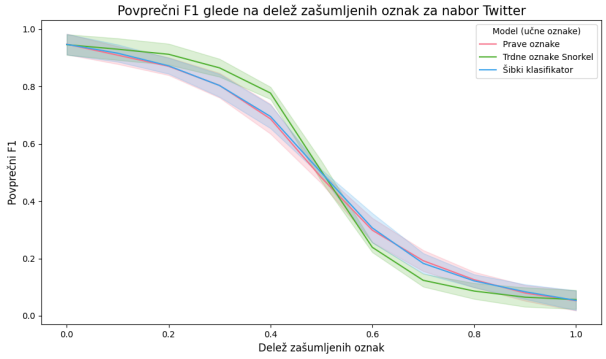
Ker z uporabo označevalnih modelov ne uporabljamo pravih oznak, želimo preveriti, ali je Snorkel učnikovito orodje za spopadanje s šumom v podatkovnih množicah. V množicah A1 in B1 smo uvedli šum v oznakah, postopek večkrat ponovili za različne stopnje šuma in primerjali uspešnost končnih modelov, naučenih z zašumljenimi oznakami ali z oznakami označevalnega modela Snorkel. Šum je bil dodan naključno, tako da smo pri danem deležu naključno izbranih primerov spremenili oznako.

Na slikah 4, 5 in 6 je predstavljen vpliv šuma na končne modele za različne nabore, ko sta zašumljeni množici A1 in B1. Na slikah se nahaja povprečje in standardni odklon čez vse končne modele. Vidimo, da se Snorkel s šumom učinkovito spopade. Uspešnost končnih modelov, naučenih z oznakami Snorkel, ovrednotena na B2, občutno pade šele, ko zašumimo več kot 40 % oznak, medtem ko uspešnost pri učenju iz zašumljenih oznak občutno pade že pri 20% šuma.

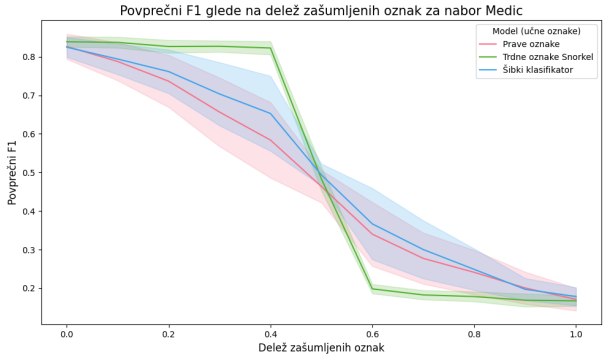
Uspešnosti spopadanja s šumom v učnih množicah je presenetljiva in kaže, da modeli, naučeni z oznakami orodja Snorkel, učinkovito obravnavajo šum in bi lahko bili primerni tudi za popravljanje oznak v množicah, kjer vemo, da imajo oznake v precejšnji meri zašumljene.




Slika 4: Vpliv dodanega šuma v množicah A1 in B1 na povprečni F1, ovrednoten na množici B2, za končne modele, naučene na pravih oznakah (z napakami), z oznakami generativnega modela Snorkel in na modelih, ki smo jih uporabili kot šibke klasifikatorje za nabor Metabric.




Slika 5: Vpliv dodanega šuma v množicah A1 in B1 na povprečni F1, ovrednoten na množici B2, za končne modele, naučene na pravih oznakah (z napakami), z oznakami generativnega modela Snorkel in na modelih, ki smo jih uporabili kot šibke klasifikatorje za nabor Twitter.




Slika 6: Vpliv dodanega šuma v množicah A1 in B1 na povprečni F1, ovrednoten na množici B2, za končne modele, naučene na pravih oznakah (z napakami), z oznakami generativnega modela Snorkel in na modelih, ki smo jih uporabili kot šibke klasifikatorje za nabor Medic.

## 5 Zaključek

Generativni označevalni model orodja Snorkel se je izkazal kot učinkovit in fleksibilen označevalni model na vseh preizkušenih naborih podatkov, saj poleg šibkih klasifikatorjev omogoča tudi enostavno vključevanje domenskega znanja s pomočjo označevalnih funkcij.

Končni modeli, ki smo jih naučili na oznakah označevalnega modela Snorkel, so pri vseh naborih po uspešnosti primerljivi z modeli, naučenimi na pravih oznakah, v nekaterih primerih celo malenkost boljši.

Kot presenetljiv rezultat lahko navedemo robustnost na vpliv dodanega šuma v oznakah v učni množici šibkih klasifikatorjev, ki smo jih nato uporabili kot označevalne funkcije v generativnem modelu. Snorkel učinkovito obravnava šumne označevalne funkcije in predstavlja zanimivo alternativo v obravnavanju šumnih podatkovnih naborov.

V nadaljnjem delu bi se bilo vredno posvetiti raziskavam na večrazrednimih problemih in na večjem številu podatkovnih zbirk, ter testirati scenarije z bistveno večjim deležem neoznačenih podatkov.

## Literatura

[1] Ratner, A. J., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases (Vol. 11, p. 269). NIH Public Access.

[2] Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., Ré, C. (2016). Data programming: Creating large training sets, quickly. Advances in neural information processing systems, 29, 3567-3575.

[3] Dunnmon, J. A., Ratner, A. J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., Rubin, D. L. (2020). Cross-modal data programming enables rapid medical machine learning. Patterns, 1(2), 100019.

[4] Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. The VLDB Journal, 29(2), 709-730.

[5] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature 486 (7403) (2012) 346–352.

[6] R. Agarwal, Twitter hate speech, (2018). URL: https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech

[7] B. Bračko (2023). Šibko nadzorovano programsko označevanje učnih primerov z orodjem Snorkel, magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.