Primena genetskog programiranja u klasifikaciji spama

Milena Stojić

26. septembar 2021.

Table of contents

Problem klasifikacije spama

- Problem koji je privukao veliku pažnju

Problem klasifikacije spama

- Problem koji je privukao veliku pažnju
- Neželjeni (*spam*) mejlovi pune sanduče, zauzimaju prostor, nanose veliku finansijsku štetu kompanijama, a mogu sadržati i virus.

Problem klasifikacije spama

- Problem koji je privukao veliku pažnju
- Neželjeni (*spam*) mejlovi pune sanduče, zauzimaju prostor, nanose veliku finansijsku štetu kompanijama, a mogu sadržati i virus.
- Veliki se napori ulažu u razvoj dobrih klasifikatora. Dobar klasifikator spama bi trebalo da skoro uvek prepozna i obriše spam poruku, ali u isto vreme i da ne klasifikuje pogrešno poruke koje nisu spam.

- Većina poznatih algoritama klasifikacije poput neuronskih mreža, Bajesovskih klasifikatora i SVM je primenjeno i većina njih je dala odlične rezultate.

- Većina poznatih algoritama klasifikacije poput neuronskih mreža, Bajesovskih klasifikatora i SVM je primenjeno i većina njih je dala odlične rezultate.
- Takođe je razvijen i klasifikator koji se formira genetskim programiranjem.

- Klasifikatori bi bili algebarski izrazi koji bi sadržavali sve 4 algebarske operacije i fiksirane konstante. Vrednost izraza bi se računala nad transformisanim vrednostima atributa. Na osnovu dobijene vrednosti izraza bi se dodeljivala klasa.

- Klasifikatori bi bili algebarski izrazi koji bi sadržavali sve 4 algebarske operacije i fiksirane konstante. Vrednost izraza bi se računala nad transformisanim vrednostima atributa. Na osnovu dobijene vrednosti izraza bi se dodeljivala klasa.

- Sam algoritam je davao nešto lošije rezultate od rezultata dobijenih standardnim algoritmima klasifikacije, ali ansambl klasifikatora dobijenih genetskim programiranjem je davao bolje rezultate od svih prethodnih klasifikacionih algoritama.

- Genetsko programiranje (skraćeno GP) ima dosta primena u mašinskom učenju. Jedna od najčešćih primena je tokom preprocesiranja podataka. (odabir i ekstrakcija atirbuta)
- Mi smo u ovom radu pokušali da na drugačiji način primenimo GP u klasifikaciji.

Malo o genetskom programiranju

- Kao i genetski algoritam, vrsta evolutivnog izračunavanja.
- Dakle, imamo koncepte populacije, selekcije, ukrštanja i mutacije.
- Karakteriše ga drvolika struktura jedinki čija veličina nije fiksna. (za razliku od linearne strukture jedinki fiksne veličine u GA)
- Obično se radi sa velikim populacijama. (hiljade jedinki)

Opis predložene metode

- Jedinke su nam stabla odlučivanja za klasifikaciju.
- U neterminalnim čvorovima su atributi čije se vrednosti ispituju, a u terminalim dodeljene klase.

Opis predložene metode

- Jedinke su nam stabla odlučivanja za klasifikaciju.
- U neterminalnim čvorovima su atributi čije se vrednosti ispituju, a u terminalim dodeljene klase.
- Mera kvaliteta klasifikacije **tačnost** (eng. *accuracy*) nam predstavlja fitnes funkciju.
- Selekcija je ruletska (fitnes srazmerna).
- Ukrštanje se vrši razmenom podstabala. Mutacija se vrši zamenom sadržaja slučajno odabranog čvora odgovarajućom vrednošću.

Opis predložene metode

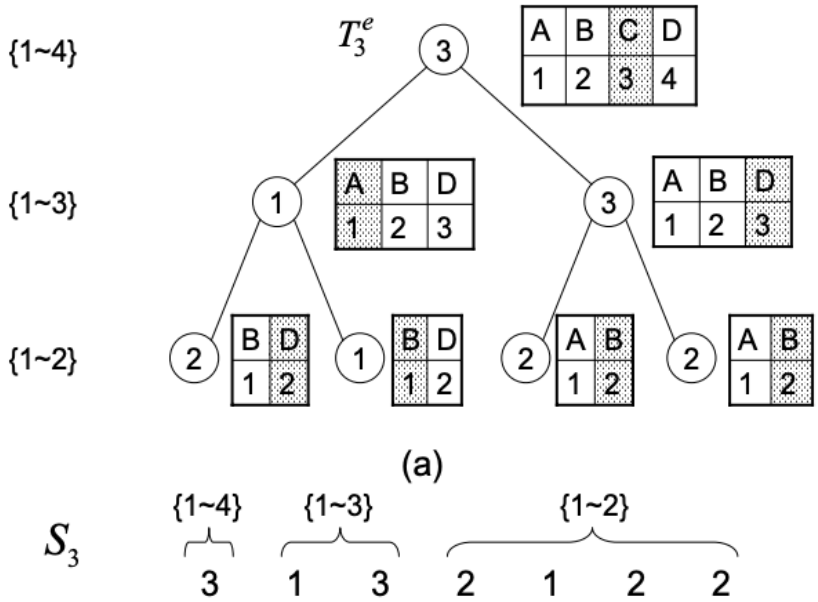
- Jedinke su nam stabla odlučivanja za klasifikaciju.
- U neterminalnim čvorovima su atributi čije se vrednosti ispituju, a u terminalim dodeljene klase.
- Mera kvaliteta klasifikacije **tačnost** (eng. *accuracy*) nam predstavlja fitnes funkciju.
- Selekcija je ruletska (fitnes srazmerna).
- Ukrštanje se vrši razmenom podstabala. Mutacija se vrši zamenom sadržaja slučajno odabranog čvora odgovarajućom vrednošću.
- Primenjen je i koncept elitizma.

Implementacija stabla odlučivanja

- Stablo će interno biti čuvano u nizu fiksne veličine. (ne moraju se svi delovi niza koristiti)
- Stablo će biti u okviru omotač klase *Jedinka* u okviru koje se takođe nalazi i metoda za računanje fitnesa.

Implementacija stabla odlučivanja

- Stablo će interno biti čuvano u nizu fiksne veličine. (ne moraju se svi delovi niza koristiti)
- Stablo će biti u okviru omotač klase *Jedinka* u okviru koje se takođe nalazi i metoda za računanje fitnesa.
- Atributi će se u stablu kodirati indeksima u *trenutnom nizu atributa.*



Eksperimentalni rezultati

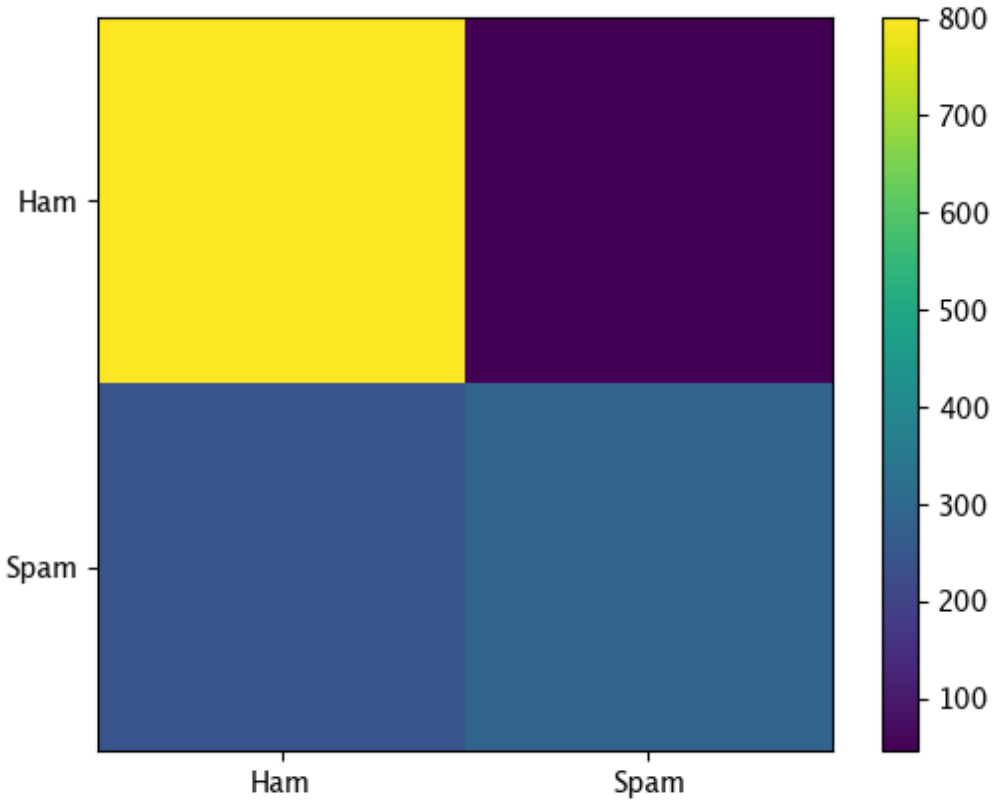
- Bolji rezultati su na većim populacijama.

Eksperimentalni rezultati

- Bolji rezultati su na većim populacijama.

Br. iteracija	Populacija	pm	Elitizam	Tačnost
1	5000	0.1	400	0.8298
50	100	0.05	400	0.6698
20	2000	0.05	400	0.7385

Matrica konfuzije



Poređenje sa postojećim rezultatima

Klasifikator	**Tančost**	**F-mera**
GP	0.848	0.848
Unapređeni GP (ansambl)	0.941	0.941
SVM	0.938	0.938
Nasumična šuma	0.939	0.939

Zaključak i dalji mogući pravci unapređivanja

- Ovde je bio fokus na samoj metodi i nismo se dublje bavili ostalim aspektima obrade podataka. (prvenstveno preprocesiranjem)
- Može da da dobre rezultate, ali bi trebalo još dosta raditi.

Zaključak i dalji mogući pravci unapređivanja

- Ovde je bio fokus na samoj metodi i nismo se dublje bavili ostalim aspektima obrade podataka. (prvenstveno preprocesiranjem)
- Može da da dobre rezultate, ali bi trebalo još dosta raditi.
- Trebalo bi se više raditi na odabiru i ekstrakciji atributa.
- Prilagoditi fitnes funkciju.

Hvala na pažnji :)