# Предлог структуре приступног рада

|  |  |
| --- | --- |
| **ПРЕЗИМЕ И ИМЕ** | **Ступар Михаило** |
| **БРОЈ ИНДЕКСА** | **2015/3503** |
| **СТУДИЈСКИ ПРОГРАМ** | **Електронско пословање и управљање системима** |
| **ПРЕЗИМЕ И ИМЕ МЕНТОРА** | *Др Братислав Петровић* |
| **НАСЛОВ РАДА** | *Утицај метода импутације података на тачност*  *Предвиђања* |
| **АПСТРАКТ РАДА** | *Рад се бави утицајем различитих метода импутације на предбиђање коришћењем линеарне регресије. У првом делу рада је опсиана линеарна регресија као и методе импутације које ће се користити у експерименту (линеарна регресија, стохастичка регресија и стабла шума одлучивања). Након експеримента и приказа резултата објашњена је метода К-средњих вредности кластеризације са недостајућим вредностима. На крају је поновљен експеримент над кластерованим подацима коришћењем технике која је показала најбоље резултате при импутацији.* |

**УВОД**

Интернет као глобална појава омогућио је велику доступност података у данашње време. Са друге стране, таква доступност довела је до све веће популарности статистичких метода, као и метода машинског учења. Често се набројане методе користе за предвиђање, или другим речима користе се историсјки подаци да се направи одређени модел предвиђања и затим се такав модел користи да се предвиде до сад невиђене појаве.

За успешно предвиђање није довољно имати само велику количину података, већ је често много битније имати добар квалитет података. Због тога је једна од најдужих (и нјбитнијих) фаза у креирању модела предвиђања, фаза припреме података. Подаци који су прикупљени од стране људи често буду непотпуни, односно недостају им одређене врдности. То није толико чест случај са подацима прикупљених од стране уређаја повезаних на интернет (internet of things уређаји).

У случају недостајања одређених вредности, потребно је имати довољно ефикасну технику како би квалитет модела предиђања био а високом нивоу. Идеално би било да модел предвиђа на исти начин уколико је трениран са комплетним скупом података као и са скупом у ком недостају вредности.

Овај рад се бави проблемом разиличитих техника за попуњавање скупа података коме недостају вредности. Рад је концентрисан на скупове података у коме подаци недостају услед немарности испитаника. Најпре ће се описати до сада најкоришћеније технике за уметање података, а затим ће бити предложена комбинација до сада две познате методе. Најпре се кластеризују подаци и онда се примени метода уметања на сваки кластер појединачно. Очекује да овакав приступ произведе боље резултате, или другим речима да уметнуте вредности буду веома сличне заиста забележеним, ''очекиваним'', вредностима.

# ФОРМУЛАЦИЈА ПРОБЛЕМА

Квалитет података који се користе за тренирање модела предвиђања има важну улогу у квалитету коначно истренираног алгоритма. Уколико подаци који представљају улаз за тренинг не описују у поптуности одређену појаву, истрениран алгоритам неће бити у могућности да препозна све ситуације дате појаве.

Један од главних фактора квалитета података је комплетност података. Некомплетни подаци у већини случајева воде неквалитетном моделу. Због тога проблем некомплетног скупа података не може бити игнорисан, па је резултат овог истраживања упоредна анализа различитих метода за попуњавање (импутацију) података као и проналажење најбољег метода по различитим параметрима.

Рад ће се бавити проблемом одабира технике за попуњавање података. Такође биће предложена идеја да се скуп података првобитно припреми пре него што се изврши уметање (импутација) података. Биће испитано да ли таква припрема има утицај на сам квалитет импутације.

# МОТИВАЦИЈА: ИНТЕЛЕКТУАЛНА И ЛИЧНА МОТИВИСАНОСТ

Са повеђањем брзине рачунара повећала се и брзина тренирања алгоритама машинског учења. Некад је било потребно и по неколико дана да се истренирају дубоке неуронске мреже (deep neural networks), а данас је то могуће за пар сати. Такође, са масовном употребом интернета, савладавање разних техника машинског учења не предтавља никакав проблем. Додатно, постоје пакети са готовим имплементацијама скоро свих алгоритама машинског учења, тако да данас није потребно никакво кодирање како би се користило машинско учење (МУ).

Иако су алати МУ свима доступни, они производе различите моделе МУ (различитог квалитета предвиђања). Узрок за ту разноликост резултујућих модела лежи у квалитету података. Вероватно једна од тема која ђе у будућности бити заступљенија у свету статистике и машинског учења је побољшање квалитета података за трнинг.

До сада, алати (пакети) за машинско учење немају добру подршку за импутацију података. Често се прибегава техници игнорисања обсервација уколико неки податак недостаје. Занимљиво је да су сви алгоритми (који се у овом раду користе) стандард за већину тих алата (пакета) тако да је потребно веома мало труда да се прошири и побољша сама њихова фукционалност.

# ОРГАНИЗАЦИЈА И МЕТОДЕ ИСТРАЖИВАЊА

Методе за ипутацију вредности у некомплетан скуп се грубо могу поделити у две групе:

1. Импутација коришћењем једне променљиве
2. Импутације коришћењем више променљивих

Ралика између ове две групе ће бити показана на примеру скупа података са три независне променљиве (колоне) и једном зависном, где свака од независних променљивих садржи бар једну недостајућу вредност.

Уколико се импутација врши коришћењем једне променљиве, онда би се вредности из прве колоне попуниле вредностима искључиво из прве колоне. У другом случају, вредности из прве колоне се попуњавају коришћењем вредности из осталих колона.

Показано је да за већину проблема прва група не даје добре резултате, па ће се овај рад бавити искључиво техникама које припадају другој групи.

Технике које ће учествовати у експериментима су:

1. Линеарна регресија
2. Стохастичка линеарна регресија
3. Шума стабала одлучивања

За сваку тегнику импутације ће се мерити показатељи успешности импутације:

1. Средња квадратна грешка
2. Корен средње квадратне грешке
3. Просечна релативна грешка
4. Корен средње квадратне грешке модела линеарне регресије

Битно је напоменути да све набројане технике захтевају (подразумевају) одређени ниво корелације између променљивих. Додатно ће бити покушано да се подаци групишу у кластере (групе) као корак пред саму импутацију. Затим ће се подаци попунити техником која је претходно показала најбоље резултате. На крају ће бити дата упоредна анаиза свих техника импутације.

Идеја да се подаци групишу у кластере као корак припреме је настала због једноставности проблема. Очекује се да распон вредности у оквиру једне променљиве унутар једне групе (кластера) буде знатно мањи него на целом скупу података. Самим тим, параметри регресије неће бити подешени да подрже широк распон података. Такође, стабла одлучивања би требало да буду мање дубине него када се обрађују све обсервације одједном.

На крају рада ће бити дат преглед могућности за даље истраживање. Предлог ће свакако зависити од резултата добијених у току рада, међутим једна од ствари која се може одмах предложити је додавање још једног корака пред саму импутацију. Може се извршити анализа главних компоненти (PCA) над променљивима (колонама) које ће служити за одређивање уметнуте вредности. Уместа коришћења свих колона као улазни парамерар у неку техника, шредложиће се коришћење само добијених компоненти.

У наставку је приказан оквирни садржај рада:

# ОПИС ДЕЛА СВЕТА КОЈИ ЋЕ БИТИ ИЗУЧАВАН

Заједница отвореног кода (open source code) је данас све заступљенија међу програмерима. Људи из целог света пристају да дају свој допринос заједници и на тај начин се стварају веома моћни алати (пакети, библиотеке). Неки од њих су писани у програмском језику python (python са библиотеком scikit-learn), програмски језик R (пуно различитих библотека подржава машинско учење) или Octave који још нема добру подршку за машинско учење (Octave који је по синтакси веома сличан Matlab-u, само што је бесплатан за коришћење).

Готова решења неких од ових алата ће бити коришћени у раду. На пример, за технике импутације података и једноставну статистичку анализу биће коришћен програмски језик R. За писање алгоритма који врши кластеризацију са непостојећим вредностма биће коришћен Octave.

Додатно, уколико новонаписан код (алгоритам) за кластеризацију прође одређене тестове, може постати део стандардног статистичког пакета Octave.

# РЕФЕРЕНТНА ЛИТЕРАТУРА, КОНЦЕПТИ И ТЕОРИЈСКИ ФОКУС

* Наведите прелиминаран списак референтне литературе коју ћете користити при истраживању Вашег проблема (односно у мери којом сте већ овладали темом истраживања) са назнаком разлога увођења.
* Наведите кључне концепте које ћете користити током свога истраживања, заједно са кључним референцама које покривају теоријски фокус, и образложите потребу за њиховим увођењем.
* [Потребно је да на основу отворене литературе (часописи, конференције, поглавља у књигама) припремите преглед стања у областима која су релевантна за одговоре на истраживачка питања]

СВРХА И ЦИЉЕВИ ИСТРАЖИВАЊА

* Шта је сврха, а шта су циљеви, Вашег истраживања.
* Опишите конкретне резултате Вашег истраживања у терминима који се јављају у пракси (*на пример, прототип, производ, симулатор, односно нов или побољшан приступ, поступак, ,...*)

РЕФЕРЕНЦЕ

Потпуна листа документа реферисаних у овом извештају.