# Приступни рад

|  |  |
| --- | --- |
| **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 података приликом решавања проблема предвиђања коришћењем регресије. У првом делу рада, описана је линеарна регресија као и методе импутације које ће се користити у експерименту. Конкретно, за импутацију података користиће се линеарна регресија, стохастичка линеарна регресија и шума стабала одлучивања. Током експеримента, скуп података са различитим процентима недостајућих вредности биће попуњен сваком од набројаних метода, и притом ће се упоређивати резултати саме импутације. Затим ће бити предложена нова метода импутације која се састоји из две фазе. Прва фаза је кластеровање података, док друга фаза представља импутацију података на нивоу кластера. Овом приликом, у првој фази ће се користити кластеровање К-средњих вредности (која подржава недостајуће врдности), док ће се у другој фази користити метода импутације која је показала најбоље резултате у првом експерименту. На крају рада ће бити дат упоредни приказ свих резултата добијених у раду; резултата из првог експеримента као и резултата добијених приликом импутације предложеном методом.* |

**УВОД**

Интернет као глобална појава омогућио је велику доступност података. Данас је веома лако приступити подацима о скоро свим појавама које нас окружују. Примери таквих података су подаци произведени од стране сензора, подаци који су коришћени у научним истраживањима, подаци који су постали доступни јавности одлуком држава (енг. open data) и многи други. (DeLorenzo & Kohun, 2009, стр. 4)

Са друге стране, таква доступност довела је до све веће популарности метода машинског учења. (Alsheikh et al. 2014, стр. 119) Велики подскуп тих метода се бави предвиђањем; односно коришћењем података одређене појаве креира се модел који је у могућности да предвиди дату појаву у будућности. На пример, на основу забележених података о временским приликама, креира се модел машинског учења који је у могућности да предвиди температуру за сутрашњи дан. (Lai et al. 2004, стр. 56)

Међутим, за успешно предвиђање није довољно имати само велику количину података, већ је често много битније имати добар ''квалитет'' података. ''Квалитет'' података се може огледати кроз различите параметре; колико различитих променљивих описује дату појаву (да ли се временске прилике описују само температуром или и подацима о падавинама, облачности...), какав је однос између променљивих (подаци о температури на две различите скале нису заиста две променљиве), или каква је заступљеност података (да ли скуп података садржи све вредности за сваку обсервацију). (Thomas, 1997) Због тога је једна од најдужих (и најбитнијих) фаза у креирању модела предвиђања, фаза припреме података. (Robert et al. 2013, стр 27)

Подаци прикупљени од стране сензора су често комплетни (све вредности су присутне). (Meng et al. 2013, стр. 91) Међутим то обично није случај са финансијским извештајима, временским серијама са интернета, интернет анкетама, мануелно скупљеним подацима итд. (Tak-chungFu, 2011, стр. 167) У набројаним примерима се неретко дешава да одређени проценат недостаје. У том случају, потребно је имати довољно ефикасну технику импутације (често се у литератури користе и термини попуњавање или уметање података) како би тачност модела предиђања била на високом нивоу. (Craig, 2010, стр. 33)

Овај рад се бави проблемом различитих метода импутације на скупу података коме недостају вредности. Најпре ће се описати најпопуларније методе импутације, (Alan et al. 2003, стр. 61) а затим ће бити предложена метода која се састоји од припреме скупа података за импутацију и саме импутације. Под припремом се подразумева кластеризација података, (Sarkar & Leong, 2001, стр. 591) а затим се примени метода импутације на сваки кластер појединачно. Очекује се вредности добијене предложеном методом буду веома ближе забележеним, ''очекиваним'', вредностима него вредности добијеним класичним методама импутације.

# 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 циљ овог истраживања упоредна анализа различитих метода импутације као и проналажење најбољег метода по различитим параметрима.

Рад ће се бавити проблемом одабира методе импутацијеа. Такође биће предложена нова метода у ком је потребно да се скуп података првобитно припреми пре него што се изврши сама импутација. 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 представља никакав проблем. Додатно, постоје пакети (алати) са готовим имплементацијама скоро свих алгоритама машинског учења, тако да данас често није потребно никакво кодирање како би се поменути алгоритми користили.

Међутим, одређени алгоритам ће произвести различите моделе предвиђања за посматрану појаву, и ти модели ће имати различиту тачност, одностно могућност, предвиђања. Узрок за ту разноликост резултујућих модела лежи у квалитету података. Због тога, вероватно једна од тема која ће у будућности бити заступљенија у свету машинског учења је побољшање квалитета података за тренинг.

До сада, пакети (алати) 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) већ имплементирани за већину тих пакета тако да је потребно веома мало труда да се прошири и побољша сама моћ (функционалност) пакета.

# ОРГАНИЗАЦИЈА И МЕТОДЕ ИСТРАЖИВАЊА

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

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

Разлика између ове две групе се може описати 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 импутације које ће учествовати у експериментима су:

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

За сваку технику импутације ће се мерити показатељи успешности импутације:

1. Средња квадратна грешка – како ће се експеримент вршити над некомплетним подацима добијеним вештачким путем (одређен проценат вредности биће обрисан), биће могуће упоредити уметнуте вредности са оригиналним. Средња квадратна грешка ће бити израчуната користећи разлику између оригиналне и уметнуте вредности.
2. Корен средње квадратне грешке – слично као под 1), корен средње квадратне грешке се односи на разлику између оригиналне и уметнуте вредности
3. Просечна релативна грешка – грешка која се односи на исту разлику као под 1) и 2). Очекује се да просечна релативна грешка буде најбољи показатељ јер поред саме вредности грешке узима у обзир и распон вредности променљиве.
4. Корен средње квадратне грешке модела линеарне регресије – једини параметар који не посматра разлику између оригиналне и уметнуте вредности. Како ће се користити скуп података погодан за регресионе проблеме (зависна променљива је нумеричког типа), алгоритам машинског учења који ће се користити за предвиђање је линеарна регресија. Након импутације провериће се тачност (грешка) таквог предвиђања. Таква грешка је ће постати параметар за оцењивање импутације.

Додатно ће бити покушано да се подаци групишу у кластере (групе) као корак пред саму импутацију. Затим ће се подаци попунити методом која је претходно показала најбоље резултате. На крају ће бити дата упоредна анализа 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 биће 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 променљивама (колонама) које ће служити за одређивање уметнуте вредности. Уместо коришћења свих колона као улазни параметар у неку од техника, шредложиће се коришћење само добијених компоненти.

У наставку је приказан оквирни садржај рада:

1. Увод

2. Регресиона анализа

2.1. Линеарна регресија

2.2. Примена

3. Недостајуће вредности

3.1. Механизми недостајућих вредности

3.2. Технике уметања података

3.2.1. Уметање средње вредности

3.2.2. Преношење задњег запажања

3.2.3. Уметање података коришћењем линеарне регресије

3.2.4 Уметање података стохастичком регресијом

3.2.5 Уметање коришћењем (шума) стабала одлучивања

4. Експериментални подаци

4.1. Опис скупа података

4.2. Корелациона матрица

5. Експеримент

5.1. Конструисање тренинг скупа

5.2. Интерпретација резултата импутације података

5.2.1. Средња квадратна грешка импутације

5.2.2. Корен средње квадратне грешке

5.2.3. Просечна релативна грешка

5.2.4. Корен средње квадратне грешке линеарне регресије

5.3. Импутација података – експеримент

5.3.1. Импутација линеарном регресијом

5.3.2. Импутација стохастичком линеарном регресијом

5.3.3. Импутација шумом стабала одлучивања

5.3.4. Анализа резултата импутације

6. Предложени метод – импутација по групама

6.1. Кластеризација са непостојећим вредностима

6.2. Имплементација

6.3. Лакат метода

7. Експеримент над кластерованим подацима

7.1. Импутација

7.2. Анализа резултата

8. Даље истраживање

# 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. Разлог за коришћење два пакета је њихова непотпуност. Одређени алгоритми постоје у једном пакету, док у другом још нису имплементирани.

Додатно, делови кода написани за потребе овог рада могу употпунити библиотеке статистичког пакета Octave или R.

# РЕФЕРЕНТНА ЛИТЕРАТУРА, КОНЦЕПТИ И ТЕОРИЈСКИ ФОКУС

У раду је представљена упоредна анализа најпопуларнијих техника импутације.(Rubin et al. 1987, стр. 31-41) Oјашњене су технике линеарне регресије као и стохастичке линеарне регресије. У оба примера је објашњена важност корелације за примену набројаних техника као и предноси и мане.

Трећа техника која ће се користити у експериментима, шума стабала одлучивања, је уведена због немогућности регресионих алгоритама да превазиђу нелинеарне проблеме. Иако су стабла одлучивања одлична за проблеме класификације, могуће их је кориситти и код регресионих проблема. (Criminisi et al. 2011, стр. 44) Такође објашњен је и значај шуме стабала и шта доприноси случајна променљива приликом креирања саме шуме.

Подаци који се користе у свим експериментима су до сада већ виђени подаци у разним студијама, (Bradley et al. 2004, стр 112). Подаци који ће се користити у раду састоје се од десет независних и једне зависне променљиве. До сада је показано да се могу добити значајни резултати предвиђаза уколико се користе Ласо регресија или постепена (енг. Stagewise) регресија. Међутим, како тема мастер истраживања није предиђање него импутација података, у раду ће се користити једноставна линеарна регресија као један од параметара ефикасности импутације.

За анлализу метода импутације ће се користити софтверски пакет R, који има одличну подршку како за импутацију регресионим методама, тако и за импутацију стаблима одлучивања. (Daniel, 2011, стр 2-8).

СВРХА И ЦИЉЕВИ ИСТРАЖИВАЊА

Проблем недостајућих вредности је веома добро заступљен у литератури. Такође, решења описана у литератури су све више део стандардних софтверских алата. Циљ овог истраживања је да покаже могућности комбиновања већ постојећих решења како би се добили бољи резултати импутације. Такође, уколико се покаже да комбинација описана у овом раду (кластеризација па импутација) даје одличне резултате, могуће је тако нешто имплементирати у већ постојеће пакете.

РЕФЕРЕНЦЕ

Alan Olinskya, Shaw Chenb, Lisa Harlowc. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. European Journal of Operational Research, Volume 151, Issue 1, 16 November 2003, Pages 53-79

A. Criminisi, J. Shotton and E. Konukoglu. (2011). Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Microsoft Research technical report

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirant. (2004). Least Angle Regression. The Annals of Statistics

Craig K. Enders. (2010). Applied Missing Data Analysis. The Guilford Press, New York London

Daniel J. Stekhoven. (2011). Using the missForest Package

Gary J. DeLorenzo, Frederick G. Kohun, A. (2009). Data Driven Conceptual Analysis of Globalization. Issues in Informing Science and Information Technology Volume 6, 2009

L. L. Lai, H. Braun, Q. P. Zhang. (2004). Intelligent weather forecast. Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)

M. Sarkar, T. Y. Leong. (2001). Fuzzy K-means clustering with missing values. AMIA Annual Symposium Proceedings Archive

Meng Ma, Ping Wang, Chao-Hsien Chu. (2013). Data Management for Internet of Things: Challenges, Approaches and Opportunities. Green Computing and Communications (GreenCom)

Mohammad Abu Alsheikh, Shaowei Lin, Dusit Niyato. (2014). Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. IEEE Communications Surveys & Tutorials ( Volume: 16, Issue: 4, Fourthquarter 2014 )

Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. (2013). Data Preparation for Mining World Wide Web Browsing Patterns. Springer-Verlag Singapore Pte. Ltd. 1999

Rubin, Little, R.J.A. (1987). Statistical Analysis with Missing *Data*, John Wiley & Sons, New York

Tak-chungFu. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence Volume 24, Issue 1

Thomas C. Redman. (1997). Data Quality for the Information Age. Artech House, Inc. Norwood, MA, USA ©1997