Рецензия

на

Дисертационен труд за присъждане на научна степен „Доктор“ по професионално направление 3.8 Икономика, научна специалност Аналитични изследвания върху данни /Data Science, озаглавен:
„Иконометричен анализ на големи данни”

Докторант: Бориславa Петрова Толева
Научен ръководител: проф. д.м.н. Иван Ганчев Иванов
Рецензент: доц. д-р Боян Михайлов Ломев

Дисертацията е с обем от 161 страници и съдържа увод, три глави, заключение и списък с използваната литература.

В увода се прави обстоен критичен анализ на съществуващата литература. Обосновава се необходимостта от използване на методи, базирани на Алгоритмичното учене от данни (Machine Learning), като се цитира ясна дефиниция за тази нова научна област.

Определят се следните изследователски цели и съответстващите им хипотези, като всяка от тях е предмет на глава от дисертационния труд:

1. **Цел:** Да се съпостави представянето на моделите за алгоритмично учене от данни спрямо иконометричните

модели в контекста на панелни икономически данни. **Хипотеза:** Съществува интердисциплинарна методология, обединяваща иконометричен анализ и методи от алгоритмичното учене от данни, провеждаща ефективно определяне на статистическите променливи.

2. **Цел:** Да се изследва приложението на бутстрапа като метод за разделяне на данните на тренировъчно и тестово подмножество без избор на променливи. **Хипотеза:** Бутстрап процедурата е ефективна алтернатива на крос валидацията.
3. Да се предложат модификации на метода на опорните вектори, на база на бутстрап процедурата, които подобряват ефективността на класическия модел на поддържащите вектори.

   Хипотеза: Предложените модификации подобряват ефективността на класическия модел на поддържащите вектори.

Първа глава е с обем от 45 стр. и съдържа следните основни раздели:

- Литературен обзор;
- Казус;
- Панелни модели;
- Валидиране на панелните модели;
- Избор на променливи чрез алгоритмично учене от данни;
- Nonnegative garrote;
- Интердисциплинарен подход;
- Числени експерименти;
- Заключение;

- Дискусия.

По нея са публикувани две статии в международни научни списания с редакционна колегия.

Поставените изследователски задачи са:

1. Да се направи сравнителна характеристика между реализацията на бутстрап процедурата за разделяне на данните на тренировъчно и тестово подмножество и десеткратна крос валидация по отношение на прогнозна точност.
2. Да се направи сравнителна характеристика между алгоритъма с бутстрап процедурата и алгоритъма с десеткратната крос валидация по отношение на класификационните показатели.
3. Да се изследва доколко бутстрап процедурата в така предложената реализация води до съкращаване на времето на прилагане на модела на опорните вектори.

За числените експерименти се използват панелните данни за имуществените права за 32 държави / ЕС + Австралия, САЩ, Канада и Норвегия/, разгледани в периода 1999 - 2014 г. и съдържащи 25 променливи. Тествани са четири основни подхода за анализ и прогнозиране на панелни данни:

- Класически иконометрични модели, като за валидиране се използват робастна ковариационна матрица и панелен генерализиран метод на моментите.
- Избор на променливи чрез алгоритмично учене от данни - Регресия на Тихонов, LASSO регресия и Адаптирана LASSO регресия.
- Nonnegative garrote с бутстрап.
- Интердисциплинарен подход (предложен от автора), съчетаващ класическите панелни модели и моделите за избор на променливи, при който след необходимите трансформации на

данните и установяване на видовете ефекти, статистически значимите променливи се определят посредством метода nonnegative garrote.

Основните изводи за първа глава са, че моделите за алгоритмичното учене от данни могат да бъдат използвани успешно при панелни данни и че предложения интердисциплинарен подход е по-ефективен от останалите по отношение на използвано машинно време и автоматизация.

Във втора глава се разглежда приложение на бутстрап процедурата при модела на опорните вектори, а обемът е 45 стр.

Структурата на изложение е:

- Литературен обзор;
- Методология;
- Данни;
- Числени експерименти;
- Заключение;
- Дискусия.

По тази глава има една публикация в списание, индексирано в SCOPUS.

Конкретизираната изследователска цел е използването на бутстрап процедурата като метод за разделяне на наблюденията в тренировъчно и тестово подмножество при големи множества от данни. Формулираната основна изследователска задача за постигане на тази цел е да се модифицира бутстрап процедурата като метод за разделяне на данните на тренировъчно и тестово подмножество.

За подобряване на бутстрап метода се използва паралелно смятане като функционалност на Python 3.6. Предложената модификация се сравнява с четири други метода: десеткратна крос валидация, leave-one-out

крос валидация, немодифициран бутстрап и случайно разбъркване на данните с повторения (repeated train/test split).

За числените експерименти са използвани девет множества от данни от сайта www.kaggle.com с различна размерност. Всички изчисления са извършени с Python 3.6.

Общото време за трениране и тестване на модела на модифицирания от автора метод е значително по-кратко от това на останалите подходи, като постигнатата прогнозна точност е сходна с тази на другите методи.

Трета глава е озаглавена „Модификации на модела на поддържащите вектори (Support vector machines) на основа на бутстрап процедурата“ и е с обем 39 стр. По нея е представена една публикация в международно научно списание с редакционна колегия.

Структурата е аналогична на използваната за представяне на материала в глави първа и втора.

Поставени са следните изследователски задачи:

1. 1. Да се направи сравнение между бутстрап модификации с линейно и rbf ядро, предложени от автора на основа на резултатите от втора глава и съществуващи варианти на модела на опорните вектори от академичната литература.
2. 2. Да се проследи как се променят AUC точките при бутстрап модификациите спрямо тези, получени от метода на целочисленото линейно смятане (mixed linear integer approach, комбиниран с модела на опорните вектори).

За провеждане на числените експерименти са използвани девет множества от медицински данни с различна размерност, върху които са правени експерименти и от други автори. Източникът е : https://archive.ics.uci.edu/ml/datasets/.

Получените резултати демонстрират, че бутстрап модификациите запазват или подобряват прогнозната точност на модела на опорните вектори спрямо други негови версии и класификационни модели от други изследвания. При сравняване на AUC точките от бутстрап модификациите и линейния целочислен подход също се отчита подобряване на точността.

Списъкът с използваните източници съдържа 113 заглавия, голяма част от които публикувани през последните пет години.

Проверката за плагиатство на дисертационния труд показва много малък процент на съвпадение със съществуващи текстове.

Темата на предложената работа е актуална и с голяма научно-приложна значимост. Авторът е демонстрирал задълбочено познаване на изследваната област. Представените резултати съответстват на поставените изследователски цели и задачи и аргументирано обосновават претенциите за авторски приноси.

Бях рецензент на вътрешната защита на дисертационния труд. Основната ми забележка беше:

„В академичните среди е прието софтуер да не се признава за научен принос. В тази връзка смятам за удачно основния новаторски резултат – модификацията на bootstrapping процедурата в втора глава да не се представя като замяна на код на Python с друг, а като изменение на метода за изчисляване, като се разясни в текста какво точно се състои тази модификация.“

Считам, че в последната версия на дисертацията този коментар е взет предвид и нямам съществени забележки.

В заключение бих предложил на уважаемите членове на научното жури, сформирано съгласно заповед РД-38-303/08.07.2021 да присъдят образователната и научна степен „Доктор“ на Бориславa Петрова Толева.

рецензент: Boyan Lomev

Digitally signed by Boyan Lomev
Date: 2021.08.26 10:13:13 +03'00'