
Софийски университет “Св. Климент Охридски“

Стопански факултет

---

# Иконометричен анализ на големи данни

Автореферат

на

дисертационен труд

*за присъждане на научна степен „Доктор“ по професионално направление*

*3.8 Икономика Научна специалност: Аналитични изследвания върху данни /Data Science/*

*Автор: Бориславa Петрова Толева*

*Научен ръководител: проф. Иван Ганчев Иванов*

*София 2021*

Дисертацията съдържа 161 стр., от които 32 таблици.

Използваната литература включва 112 литературни източника на английски и български език.

По темата на дисертационния труд са направени 4 публикации.

Дисертационния труд е обсъден и насочен за защита от разширен съвет на катедра „Статистика и иконометрия“ при Стопански факултет на Софийски университет „Св. Климент Охридски“.

Защитата на дисертационния труд ще се състои на от в …..зала на …..

Софийския университет на открито заседание на научно жури.

Автор: Бориславa Петрова Вригазова

Заглавие: Иконометричен анализ на големи данни

# I. Увод

Machine learning или алгоритмично учене от данни, намира широко приложение в ежедневието ни и икономиката. Събира се информация за посещаемостта на сайтове, обекти и магазини в различни местни и международни локации с цел дадена компания да подскаже на клиентите си къде се намира най-близкият техен магазин. Стартираха онлайн магазини, които позволяват намирането на определен модел стока само чрез снимката ѝ. Това спестява време на съвременния натоварен потребител и премахва физическото ограничение на потребителя, готов да направи покупка. Онлайн магазините се оказаха особено важни по време на ковид пандемията. Те бяха средството, чрез което много малки бизнеси оцеляха в пандемията.

Моделите от machine learning представляват широк спектър от статистически алгоритми, приложими при много големи множества данни с цел извличане на присъщата структура на данните и изграждане на прогнозни модели (Кабаиванов, 2020), (Иванов, 2018), (Mitchell, 1997), (James et. al., 2013). Алгоритъмът се учи от част от входните наблюдения сам, без изрично да му е зададена предварително структурата, присъща на данните. Те сами откриват структурата в данните и взимат решения на база на нея (Иванов, 2018), (Кабаиванов, 2020). Затова чрез опознаването на скритите характеристики на данните, те се „учат“ сами (machine learning) как да подобрят крайния резултат от тях (James et. al., 2013). Тъй като подобни алгоритми се състоят от изпълняването на последователност от стъпки за изграждане на прогноза, класификация или клъстер (Mitchell, 1997), ние приемаме терминът „алгоритмично учене от данни“ като превод на термина machine learning (Иванов, 2018).

*Актуалност на проблема*

Дисертационният труд има за цел да разшири приложенията на алгоритмичното учене от данни до панелните иконометрични модели и да усъвършенства вече съществуващи модели от алгоритмичното учене от данни за големи множества от наблюдения. Първа глава предлага интердисциплинарен подход при моделирането на панелни данни, докато втора и трета подобряват методи от алгоритмичното учене от данни. Първа глава надгражда буутстрап

резултатите на Брайман (1992, 1995) във връзка с приложението му при панелни данни. Втора и трета разширяват (Efron et. al., 1997), като изследват буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество.

*Цели и задачи на дисертацията*

**В дисертацията поставяме три цели, като всяка се разглежда в отделна глава. Целта на първа глава е да се предложи интердисциплинарен подход за моделиране на панелни данни, който да бъде по-ефективен от традиционния иконометричен подход.** Този подход съчетава класическата иконометрична теория за панелни данни и алгоритмично учене от данни. Формулира се хипотезата, че съществува интердисциплинарен подход за моделиране на панелни данни, обединяващ иконометричния подход и алгоритмичното учене от данни. За постигане на първата цел се решават пет задачи:

1. Да се построят иконометрични панелни модели на индекса на имуществените права.
2. Да се валидират панелните модели.
3. Да се приложат модели за избор на променливи от областта на алгоритмичното учене от данни /ridge, lasso, adaptive lasso/.
4. Да се изследва моделът nonnegative garrote (Breiman, 1995) в панелни данни.
5. Да се приложи интердисциплинарен подход, съчетаващ иконометрия и алгоритмично учене от данни.

**Втора глава цели да разгледа буутстрап процедурата като метод за разделяне на наблюденията в тренировъчно и тестово подмножество на (големи) множества от наблюдения. Формулира се хипотезата, че буутстрап процедурата може да бъде алтернатива на десеткратната крос валидация. Проверката на хипотезата преминава през** следните задачи:

1. Да се направи сравнителна характеристика между реализацията на буутстрап процедурата за разделяне на наблюденията на тренировъчно и тестово подмножество и десеткратна крос валидация по отношение на прогнозна точност.

2. Да се направи сравнителна характеристика между алгоритъма с буутстрап процедурата и алгоритъма с десеткратната крос валидация по отношение на класификационните показатели.

3. Да се изследва доколко буутстрап процедурата в така предложената реализация води до съкращаване на времето на прилагане на модела на опорните вектори.

**Трета глава има за цел да се предложат нови модификации на модела на поддържащите вектори и в метода за решаването му, които да подобряват ефективността (отношението на постигнатия резултат спрямо поставената цел) на класическия модел с ANOVA.** Изследва се хипотезата, че буутстрап версията на ANOVA моделът на опорните вектори, е по-ефективна от класическата ANOVA версия. Формулират се следните задачи на трета глава:

1. Да се направи сравнение между буутстрап модификациите и тяхната прогнозна точност и съществуващи модификации на модела на опорните вектори от академичната литература (Maldonado, 2014), както и други класификационни модели (Naiudu et al., 2012), (Miao et al., 2016), (Khanna et al., 2015), (Kodati et al., 2018), (Latha & Jeeva, 2019), (Nandipati, 2020) и др.
2. Да се проследи как се променят AUC точките при буутстрап модификациите спрямо тези, получени от метода на целочисленото линейно смятане (mixed linear integer approach, комбиниран с модела на опорните вектори), приложен в (Maldonado, 2014).

За удобство в автореферата е използвана номерация на литературните източници, както е направено в дисертационния труд. Не е включен списък на литературните източници. Номерацията на формулите и уравненията не е последователна и съвпада с тези от дисертационния труд.

# II. Структура и съдържание на дисертационния труд

Дисертацията се състои от увод, три глави, заключение и списък на цитираната литература. Номерацията на таблиците и формулите е двойна, включваща номера на главата и поредния номер на таблицата или формулата в главата.

**Глава 1. Приносът на алгоритмичното учене от данни към иконометричните процедури за избор на статистически значими променливи**

Първа глава предлага интердисциплинарен подход, който да се използва при моделиране на панелни данни.

### 1.1 Преглед на литературата

Панелните данни са особен вид данни, при които има две измерения – време и крос секционна единица (Greene, 2003). Традиционните иконометрични модели преминават през редица стъпки, за да моделират такива данни. Най-важните от тях са да се открият видовете ефекти в панела и да се изберат подходящите статистически значими променливи за модела (Wooldridge, 2012). За идентифицирането на панелните ефекти съществуват бързи и ефективни тестове (Wooldridge, 2012), но процедурата за избор на статистически значими променливи е дълга и тромава, тъй като не съществува автоматизиран начин те да бъдат избрани (Wooldridge, 2012).

В същото време в алгоритмичното учене от данни съществуват модели за избор на променливи, които предлагат автоматичен начин да бъдат разпознати статистически значимите променливи, тъй като техните коефициенти не са нулеви (James et. al., 2013). Но тези алгоритми не са широко застъпени в иконометричното моделиране (Wooldridge, 2012). Брайман (1992, 1995) предлага модел от алгоритмичното учене от данни, който се нарича nonnegative garrote. Той заключава, че когато този модел се прилага с буутстрап (Breiman, 1992), то той е подходящ за моделиране на фиксирани регресори, които са независими (Breiman, 1995). Ако регресорите са случайни и некорелирани, то тогава е подходящ nonnegative garrote с крос валидация (Breiman, 1995). Тази формулировка напомня на фиксираните и случайните ефекти при панелните

данни. Въпреки това тя е останала неизследвана в контекста на панелни данни. Първата глава на дисертацията цели да награди резултатите на Брайман, като предложи интердисциплинарен подход, включващ модела nonnegative garrote, за моделиране на панелни данни с цел постигането на по-ефективно построяване на панелен модел.

Тя изследва приложението на nonnegative garrote с буутстрап процедурата към панелни данни с фиксирани ефекти. Хипотезата, че nonnegative garrote със случайни ефекти може да се използва при панели със случайни ефекти не е предмет на дисертационния труд. Затова тя е възможна посока за бъдещо надграждане на получените резултати в него.

Интердисциплинарният подход, който предлагаме в главата се състои от няколко стъпки:

1. Да се направят необходимите трансформации на данните в панела, следвайки класическия иконометричен подход (Wooldridge, 2012), (Greene, 2003).
2. Да се установят видовете ефекти в панела чрез прилагането на традиционните иконометрични тестове (Wooldridge, 2012), (Greene, 2003).
3. Нова стъпка: ако се установят фиксирани ефекти, да се премине към откроряване на статистически значимите променливи чрез nonnegative garrote, като стойността на параметъра $\lambda$ се намери чрез малката буутстрап процедура на Брайман (Breiman, 1995).
4. Да се построи панелен модел на база на резултатите от стъпка 3 и да се премине към прогнозиране на данните.

Предимствата на този подход, които установяваме в процеса на изследването, са:

1. Извършва се ефективен избор на променливи в панелни данни чрез **автоматизиран алгоритъм**, отчитащ ефектите в панела.
2. **Изборът на променливи в панела се извърша по-бързо**, тъй като се прескача ръчното добавяне и премахване на променливи и се

пропускат класическите иконометрични тестове за изпуснати и излишни променливи.

3. **Изборът на променливи може да се извърши измежду голям набор от първоначални независими променливи,** което прави интердисциплинарния подход по-ефективен от класическия панелен подход.
4. **Достига се по-бързо до панелен модел, който да бъде валидиран чрез панелен подход и да бъде използван за по-нататъшно прогнозиране на данните.**

### 1.2 Казус

За да проверим формулираната хипотеза преминаваме през решаването на няколко задачи. Тяхното решаване се разглежда в контекста на панелен казус за моделиране на индекса на имуществени права. Той съдържа 32 страни, разгледани в периода 2000-2014г. Казусът цели да идентифицира статистически значимите променливи за моделиране на индекса на имуществените права и да провери дали неравенството на доходите, сивата икономика и неравенството на половете също повлияват индекса на имуществени права.

### 1.3 Панелни модели

Първата задача на главата е да се построят иконометрични панелни модели на индекса на имуществените права. Те целят да проверят статистическата значимост на основните независими променливи в модела. След като с помощта на иконометрични тестове (Wooldridge, 2012), са идентифицирани те, се построяват панелни модели с контролни променливи - сива икономика, полово неравенство и неравенството на доходите. Към основния модел се добавя по една контролна променлива и така се получават три модела с контролни променливи. Провеждат се иконометрични тестове, за да се провери статистическата значимост на всяка контролна променлива. Следващите уравнения представят базовия панелен модел и трите модела с контролни променливи:

**Модел 1.1: Базов модел – детерминанти на имуществените права**

$$lproperty_{it} = \beta f(dbirth_{it}, drate_{it}, mortality_{it}, unempl_{it}, urban_{it}, dmilitary_{it}) + c_{it} + \varepsilon_{it}$$

Където $c_{it}$ е фиксираният времеви ефект за всяка от 32-те страни в извадката, t= 2000:2014, $\varepsilon_{it}$ е смущението за панела.

**Модел 1.2: Контролна променлива – полово неравенство/ коефициент на участие на жените спрямо мъжете в работната сила/**

$$lproperty_{it} = \beta f(dbirth_{it}, drate_{it}, mortality_{it}, unempl_{it}, urban_{it}, dmilitary_{it} + dgender_{it}) + c_{it} + \varepsilon_{it}$$

Където $c_{it}$ е фиксираният времеви ефект за всяка от 32-те страни в извадката, t= 2000:2014, $\varepsilon_{it}$ е смущението за панела. Към базовия модел 1.1 прибавяме променливата приближение на половото неравенство, за да проверим ефекта ѝ върху имуществените права.

**Модел 1.3: Контролна променлива – неравенство на доходите / коефициент на Джини/**

$$lproperty_{it} = \beta f(dbirth_{it}, drate_{it}, mortality_{it}, unempl_{it}, urban_{it}, dmilitary_{it} + \lg ini_{it}) + c_{it} + \varepsilon_{it}$$

Където $c_{it}$ е фиксираният времеви ефект за всяка от 32-те страни в извадката, t= 2000:2014, $\varepsilon_{it}$ е смущението за панела. Към базовия модел 1.1 се прибавя неравенството на доходите, като се изважда от модела неравенството на половете.

**Модел 1.4: Контролна променлива – сива икономика**

$$lproperty_{it} = \beta f(dbirth_{it}, drate_{it}, mortality_{it}, unempl_{it}, urban_{it}, dmilitary_{it} + dshadow_{it}) + c_{it} + \varepsilon_{it}$$

Където $c_{it}$ е фиксираният времеви ефект за всяка от 32-те страни в извадката, t= 2000:2014, $\varepsilon_{it}$ е смущението за панела. Към базовия модел 1.1 се прибавя сивата икономика на мястото на неравенството в доходите.

### 1.4 Валидиране на панелните модели

Модели 1.1-1.4 са класически панелни модели. Иконометричната теория препоръча да се валидират получените резултати чрез робастни иконометрични методи (Wooldridge, 2012), за да сме сигурни, че оценките от модели 1.1—1.4 не са изместени. Втората задача на главата е да се валидират панелните модели както изисква класическата иконометрична теория (Wooldridge, 2012). Изпълняваме няколко стъпки за валидиране на иконометричния модел. Първо, използваме робастна ковариационна матрица или т.нар. sandwich estimator (Wooldridge, 2012). Тя изчиства ефектите от естествената хетерогенност в панела

и възможната корелация между наблюденията във времето. Ако оценките на коефициентите от робастната ковариационна матрица са близки от тези от модели 1.2-1.4, то тогава панелните модели с фиксирани времеви модели са правилно специфицирани.

Второ, прилагаме панелен генерализиран метод на моментите (GMM) (Wooldridge, 2012), (Ghysels, 2020), за да отчетем влиянието на хетерогенността и хетероскедастичността в панела и да се валидират оценките от панелната регресия. Този модел изисква дефинирането на инструментални променливи. Те трябва да бъдат корелирани със зависимата променлива и некорелирани със смущенията на независимите променливи (Hansen, 1982). За инструментални променливи избираме съответните независими променливи, участващи в модели 1.2-1.4 и индексът на икономическа свобода. По този начин се изпълнява условията рангът на инструменталните променливи да е по-голям от броя независими променливи (Wooldridge, 2012).

Трето, правят се тестове за излишни и пропуснати променливи (Wooldridge, 2012). Те се състоят в ръчно добавяне и махане на променливи от панелния модел, като се проследява как се променя коефициентът $R^2$ и статистическата значимост на променливите. Ако те се променят значително, променливата се добавя към модела и става част от него. Ако изпускането ѝ не води до съществени изменения в модели 1.2-1.4, то тя може да бъде изпусната. Тези тестове са времеемки, ако става въпрос за голям масив от данни. Въпреки това са често използвани в иконометричното моделиране (Wooldridge, 2012).

**По този начин изхождаме от класическите иконометрични панелни модели. Чрез решаването на първите две задачи построяваме валидиран панелен модел, резултатите от който използваме, за да анализираме резултатите от модели за избор на променливите от алгоритмичното учене на данни. Така стигаме до третата задача на главата – да решим казуса за имуществените права чрез методи за избор на променливите и да съпоставим резултатите с класическия панелен подход.**

### 1.5 Избор на променливи чрез алгоритмично учене от данни

Изпълняваме няколко метода за избор на променливи - ridge, lasso и adaptive lasso (James et. al., 2013). Трета задача е важна, тъй като тя анализира поведението на класическите модели за избор на променливите, при които панелните ефекти не са отчетени. Тук имаме две възможни посоки за задача четири в зависимост от резултатите от задача 3. Ако класическите модели за избор на променливи даваха подобни резултати на панелните модели, то тогава изводът щеше да бъде, че панелните ефекти не е необходимо да бъдат отчетени. Тъй като ние попадаме във втория случай, когато тези модели за избор на променливи не дават добри резултати, си задаваме въпроса: „Как фиксираните ефекти в панела да бъдат отчетени в моделите за избор на променливи?“

### 1.6 Nonnegative Garrote

**Тогава тестваме формулировката на Брайман (1992, 1995), че моделът nonnegative garrote с буутстрап процедурата може да се използва при фиксирани регресори, които са независими. Така стигаме до четвъртата задача на първа глава - да се изследва моделът nonnegative garrote (Breiman, 1995) в панелни данни.**

Резултатите ни показват, че той идентифицира същите статистически значими променливи като валидирания панелен модел.

### 1.7 Интердисциплинарен подход

В същото време този модел не може да бъде самостоятелно използван при панелни данни, защото трябва първо да бъдат идентифицирани видовете ефекти, а след това резултатите от nonnegative garrote да бъдат валидирани, най-добре чрез класическите панелни модели за целта (Wooldridge, 2012). По този начин смятаме, че се извършва ефективен подбор на променливи за панела, като се спестява време от ръчните панелни тестове. Чрез валидирането с панелни методи се избягва възможността nonnegative garrote да бъде подвеждащ. Този извод ни доведе до петата задача на главата - да се приложи интердисциплинарен подход (показан в началото на текста), съчетаващ иконометрия и алгоритмично учене от данни.

### 1.8 Резултати

**Панелни модели**

Таблица 1.1 представя резултатите от панелните модели 1.1-1.4, показвайки със звездички статистически значимите променливи, т.е. кои са променливите, които влияят на индекса на имуществени права.

Таблица 1.5: Резултати от панелни модели

| | **Model 1.1** | **Model 1.2** | **Model 1.3** | **Model 1.4** |
|---|---|---|---|---|
| dbirth | -0.06** | -0.06** | -0.06** | -0.05 |
| | (-28.66) | (-28.67) | (-27.47) | (-25.22) |
| drate | -0.10** | -0.10** | -0.10** | -0.11** |
| | (-26.17) | (-26.16) | (-25.81) | (-27.16) |
| mortality | -0.05*** | -0.05*** | -0.05*** | -0.05*** |
| | (-70.89) | (-72.08) | (-72.58) | (-70.78) |
| unempl | -0.02** | -0.02** | -0.02** | -0.02** |
| | (-27.96) | (-27.96) | (-25.91) | (-28.54) |
| urban | 0.01*** | 0.01*** | 0.01*** | 0.01*** |
| | (35.26) | (34.75) | (32.54) | (35.84) |
| dmilitary | 0.26* | 0.27* | 0.27* | 0.26* |
| | (24.76) | (24.78) | (25.06) | (24.08) |
| dgender | | -0.00 | | |
| | | (-0.05) | | |
| lgini | | | -0.09 | |
| | | | (-0.51) | |
| dshadow | | | | 0.07*** |
| | | | | (40.37) |
| $R^2$ | 0.67 | 0.67 | 0.67 | 0.68 |
| F-stat | 135.59*** | 115.93*** | 116.55*** | 118.45*** |
| Data set | 2000 -2014 | 2000 -2014 | 2000 -2014 | 2000 -2014 |
| | Annual | Annual | Annual | Annual |

Източник: изчисления на автора, Нивата на статистическа значимост са както следва: „***“ – 0,001; „**“ – 0,01; „*“ – 0,05; „ . “ – 0,1.

Както се вижда, статистически значимите променливи във всички модели са коефициентът на смъртност /drate/, коефициентът на детска смъртност /mortality/, коефициентът на безработица /unempl/, процентът градско население /urban/, както и военните разходи /dmilitary/. Моделите с контролни променливи 1.2-1.4 показват, че само темпът на нарастване на сивата икономика оказва влияние на индекса на имуществени права. Така се решава първата задача на главата.

### Валидиране на панелните модели

Таблица 1.2: Резултати от робастна ковариационна матрица при модели 1.1-1.4

| | **Model 1.5** | **Model 1.6** | **Model 1.7** | **Model 1.8** |
|---|---|---|---|---|
| dbirth | -0.06** | -0.06** | -0.06** | -0.05* |
| | (-28.66) | (-28.67) | (-27.47) | (-25.22) |
| drate | -0.10** | -0.10** | -0.10** | -0.11** |
| | (-26.17) | (-26.16) | (-25.81) | (-27.16) |
| mortality | -0.05*** | -0.05*** | -0.05*** | -0.05*** |
| | (-70.89) | (-72.08) | (-72.58) | (-70.78) |
| unempl | -0.02** | -0.02** | -0.02** | -0.02** |
| | (-27.96) | (-27.96) | (-25.91) | (-28.54) |
| urban | 0.01*** | 0.01*** | 0.01*** | 0.01*** |
| | (35.26) | (34.75) | (32.54) | (35.84) |
| dmilitary | 0.26* | 0.27* | 0.27* | 0.26* |
| | (24.76) | (24.78) | (25.06) | (24.08) |
| dgender | | -0.00 | | |
| | | (-0.05) | | |
| lgini | | | -0.09 | |
| | | | (-0.51) | |
| dshadow | | | | 0.07*** |
| | | | | (40.37) |
| $R^2$ | 0.67 | 0.67 | 0.67 | 0.68 |
| F-stat | 135.59*** | 115.93*** | 116.55*** | 118.45*** |

Източник: авторски изчисления

От таблицата се вижда, че робастните оценки и некоригираните оценки (таблица 1.1) не са различават. Всички коефициенти, които са значими при панелните модели, са значими и след прилагането на робастната ковариационна матрица. Колебания се забелязват в коефициента на раждаемост, който в модел 1.8 при некоригирани оценки е статистически незначим, а според робастния метод е значим на 0,05% ниво на статистическа значимост.Таблица 1.3 показва резултатите от панелния генерализиран метод на моментите (GMM), приложен към модели 1.1-1.4.

Таблица 1.3: Резултати от панелен GMM, приложен към панелните модели с фиксирани ефекти

| | **Model 1.5** | **Model 1.6** | **Model 1.7** | **Model 1.8** |
|---|---|---|---|---|
| DBIRTH | -0.06** | -0.08*** | -0.06** | -0.05* |
| | (-1.93) | (-2.57) | (-1.92) | (-1.60) |
| DRATE | -0.11*** | -0.10*** | -0.10*** | -0.11*** |
| | (-2.35) | (-2.53) | (-2.35) | (-2.41) |
| MORTALITY | -0.05*** | -0.05*** | -0.05*** | -0.05*** |
| | (-15.93) | (-14.30) | (-15.33) | (-15.78) |
| UNEMPL | -0.02*** | -0.02*** | -0.02*** | -0.02*** |
| | (-7.08) | (-7.56) | (-6.57) | (-7.07) |
| URBAN | 0.01*** | 0.01*** | 0.01*** | 0.01*** |
| | (10.21) | (9.95) | (10.01) | (10.32) |
| DMILITARY | 0.27*** | 0.26*** | 0.27*** | 0.26*** |
| | (3.96) | (3.83) | (4.01) | (3.86) |
| DGENDER | | 0.01 | | |
| | | (0.81) | | |
| LGINI | | | -0.09 | |
| | | | (-1.19) | |
| DSHADOW | | | | 0.08*** |
| | | | | (2.39) |
| $R^2$ | 0.67 | 0.63 | 0.67 | 0.68 |
| Instrument rank | 22 | 9 | 23 | 23 |
| J-stat | 139.34*** | 105.40*** | 149.22*** | 137.48*** |

Източник: изчисления на автора

J – статистиката (Wooldridge, 2012) потвърждава, че индексът на икономическа свобода е добър избор на инструментална променлива. Причината е статистическата значимост на 0,001% ниво на значимост. Ако тази статистика не беше значима на нито едно ниво на статистическа значимост, друга променлива щеше да бъде по-подходяща инструментална променлива.

От таблицата се вижда, че всички променливи запазват своята статистическа значимост, сравнени с резултатите от предишните модели. Както при робастните оценки, така и при панелния GMM с фиксирани времеви ефекти коефициентът на раждаемост е значима променлива. Забелязват се леки промени при нивото на статистическа значимост, както и стойността на оценките за някои променливи. Тъй като тези колебания са в интервала (0,001 – 0,2) може да се заключи, че дори при наличие на хетерогенност и хетероскедастичност в данните, детерминантите на имуществените права не се променят. Минималните

разлики между оценките от робастните модели и панелните GMM модели се дължи на разликите в начина на изчисляване на оценките.

Потвърждаваме резултатите от таблица 1.1 и чрез тестове за изпуснати и излишни променливи. Валидираме панелните модели. Резултатите от таблица 1.3 използваме, за да анализираме методите за избор на променливи върху същия казус.

**Избор на променливи чрез алгоритмично учене от данни**

Таблица 1.4 сравнява идентифицираните статистически значими променливи от ridge, lasso, adaptive lasso и валидираните резултати от панелния модел.

Таблица 1.4: Сравнение на статистически значимите променливи от панелните модели и моделите за избор на променливи от алгоритмичното учене от данни

| | LASSO | Тихонов /ridge/ | Адаптиран LASSO | Nonnegative garrote | Статистическа значимост панелни модели |
|---|---|---|---|---|---|
| Dbirth | -0,05 | -0,05 | 0 | 0,00 | **/* |
| Drate | -0,11 | -0,13 | -0,03 | -0,05 | ** |
| mortality | -0,04 | -0,04 | -0,04 | -0,48 | *** |
| unempl | -0,02 | -0,02 | -0,02 | -0,22 | ** |
| dshadow | 0,04 | 0,04 | 0,01 | 0,01 | *** |
| dgender | 0 | 0,01 | 0 | 0,00 | |
| Infl | 0 | 0 | 0 | 0,00 | |
| loginternet | -0,02 | -0,01 | 0 | 0,00 | |
| dexpect | -0,05 | -0,07 | 0 | 0,00 | |
| emissions | 0,01 | 0,01 | 0,01 | 0,00 | |
| Dhealth | 0 | 0 | 0 | 0,00 | |
| Urban | 0,01 | 0,01 | 0,01 | 0,33 | *** |
| Lgini | -0,07 | -0,11 | 0 | 0,00 | |
| dmilitary | 0,2 | 0 | 0,13 | 0,08 | * |

Източник: авторски изчисления

Вижда се, че всеки един от трите модела идентифицира някои от променливите като статистически значими както е при панелните модели. Но моделите ridge, lasso и adaptive lasso идентифицират и други променливи, които в панелите не са значими. Също така считат едни променливи за незначими, но в панела те са всъщност статистически значими. Трите модела не работят добре, тъй като не отчитат фиксираните ефекти. Проверяваме как nonnegative garrote ще се представи. Така изпълняваме третата задача на главата и преминаваме към четвъртата – тестване на nonnegative garrote.

**Nonnegative garrote**

Таблица 1.4 показва и резултатите от модела nonnegative garrote. Резултатите от него са много близки до **адаптирания** LASSO и съвпадат с панелните модели. Разликата между гарота и адаптирания LASSO е в променливата емисии на въглероден диоксид (emissions). Моделът nonnegative garrote показва тази променлива като статистически незначима, докато адаптираният LASSO – като значима променлива. Този резултат съвпада с панелните модели. Единствената разлика между построения модел и панелните регресии е в променливата коефициент на раждаемост. Моделът nonnegative garrote я идентифицира като незначима, но причината е същата както при адаптирания LASSO – коефициентът в панелната регресия е малък.

Нашите резултати показват, че буутстрап процедурата и моделът nonnegative garrote могат успешно да се приложат като метод за избор на променливата при панелни данни с фиксирани ефекти дори множеството данни да е малко. Причината за това е, че така се отчитат наличните ефекти в панела докато се извършва избор на променливи за моделиране на индекса на имуществените права.

**Интердисциплинарен подход**

Въпреки обещаващите резултати от модела nonnegative garrote, класическият иконометричен подход не може да бъде зачеркнат от моделирането на панелни данни. Той е бавен и тромав, затова алгоритмичното учене от данни може да го автоматизира и по този начин да го направи по-ефективен. На база на изпълнените задачи до момента предлагаме интердисциплинарен подход за моделиране на панелни данни. При него на първата стъпка се идентифицират ефектите в панела. На втората се прилага nonnegative garrote с буутстрап процедурата, ако панелните ефекти са фиксирани. По този начин се извършва ефективен избор на променливи без стандартните панелни тестове. Трета стъпка включва валидирането на получените резултати от nonnegative garrote с буутстрап чрез панелни робастни модели. Накрая получените резултати могат да се използват за прогнози или други цели.

**1.9 Заключение**

Чрез решаването на задачите, поставени в началото на първа глава, изпълняваме целта на главата – да се предложи интердисциплинарен подход за ефективно моделиране на панелни данни. Това е иновация при моделирането на панелните данни. Показваме ново приложения на модела nonnegative garrote на Брайман (1995) в областта на панелните данни, както и неговата неизследвана способност досега да отчита фиксирани ефекти в панели. Този извод е основополагащ за следващите две глави, новото му приложение от първа глава ни стимулира да потърсим и други негови недобре изследвани приложения. Такова е приложението му като метод за разделяне на наблюденията на тренировъчно и тестово подмножество. Така буутстрап процедурата става предмет на втора и трета глава.

## 1.10 Дискусия

В тази глава показваме, че моделите за избор на променливи, в частност nonnegative garrote (Breiman, 1995) могат да бъдат успешно прилагани за избор на статистически значими променливи при малки множества и панелни данни. Използването на такъв вид модели при панелните данни е нов подход в моделирането им. Подобни модели позволяват да бъдат спестени редица стъпки от класическия панелен подход за избор на статистически значими променливи в панелни модели. Най-голямото предимство на алгоритмичното учене от данни в панелни множества е, че предлага автоматизиран алгоритъм, който идентифицира променливите, които трябва да присъстват в модела, без да се налага ръчното добавяне и премахване на независими променливи и тяхното ръчно тестване. Моделите за избор на променливи могат успешно да извършат подбор на независими променливи дори при малки множества, каквито обикновено са панелните множества.

Най-важният извод, който достигаме, е, че въпреки предимствата на nonnegative garrote с буутстрап процедурата, при панелни модели класическият иконометричен подход не може да бъдат зачеркнат от моделирането на панелни данни. Най-добрият начин бързо и ефективно да се открояят статистически значимите променливи в панел с фиксирани ефекти е комбиниран подход между класическа иконометрична теория и алгоритмично учене от данни. Класическият иконометричен подход е необходим, за да разпознае видовете ефекти в панела, и

да валидира резултатите от подбора на променливи чрез алгоритмично учене от данни. Моделите за избор на променливи са необходими, за да се предложи автоматизирана и спестяваща време процедура за подбор на променливи. Така комбинираният подход може да се използва както за предварително опознаване на панелните данни, така и за прогнозиране и моделиране.

Така допринасяме и за разширяване на възможностите за комбиниран научен подход, което помага за сближаването на двете области за моделиране на данни, и показва неразкрити досега предимства на този подход.

Бутстрап процедурата има редица важни приложения в академичната литература, но предимствата ѝ като метод за разделяне на наблюденията на тренировъчно и тестово подмножество досега не са били изследвани в дълбочина. Резултатите, които получаваме за панелните модели, ни провокират да изследваме буутстрап процедурата по-задълбочено в контекста на други модели от алгоритмичното учене от данни. Тази тема е предмет на следващите две глави.

***Публикации по първа глава:***

- Vrigazova B., 2017, Property rights: Factors Contributing to the Generation of Income from Ownership of Property, Innovativity in Modeling and Analytics Journal of Research, vol. 2, pp.22-39 – научна статия в международно научно списание с редакционна колегия
- Vrigazova, B. 2018, Nonnegative Garrote as a Variable Selection Method in Panel Data, International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 1, January 2018 – научна статия в международно научно списание с редакционна колег

## Глава 2. Приложение на буутстрап процедурата при модела на опорните вектори

Втора глава има за предмет буутстрап процедурата като метод за получаване на тренировъчни и тестови данни, а обект е моделът на опорните вектори без избор на променливи.

### 2.1 Преглед на литературата

Моделът на опорните вектори е един от най-често използваните класификационни модели в практиката (Vapnik, 2012). Това е така, тъй като той се отличава с няколко предимства пред останалите модели в алгоритмичното учене от данни. На първо място той е удобен за работа както при количествени, така и при качествени регресори (Vapnik, 2012). Когато се използва за класификация, той е изключително гъвкав (James et.al., 2013), (Vieira, 2020), (Yeturu, 2020). От една страна, той притежава няколко параметъра, които могат да бъдат адаптирани към особеностите на данните. Например, параметърът C за определяне на границите между хипер равнините при C-версията на модела на опорните вектори (Cortes, 1995), (Vapnik, 2012). Тази версия на модела на опорните вектори се нарича още модел , съдържащ т.нар. soft margin (James et.al., 2013). Съществува версия на модела на поддържащите вектори с т.нар. hard margin (James et.al., 2013). Кой от двата модела е по-подходящ зависи от това по какъв начин трябва да бъдат определени границите между хипер равнините, съдържащи наблюденията от различните класове (Vapnik, 2012), (James et.al., 2013), (Vieira, 2020), (Yeturu, 2020). Степента на модела също може да бъде контролирана чрез вида ядро – линейно, нелинейно (Vapnik, 2012), (James et.al., 2013, (Vieira, 2020), (Yeturu, 2020), (Cortes, 1995).

От друга страна, моделът на опорните вектори работи добре при данни, които са с нелинейна независимост дори когато е трудно да бъде определена каква точно е тя (Cortes, 1995), (Maldonado et. al., 2014), (James et. al., 2013). Той работи със стандартизирани данни, което позволява единна интерпретация на мерните единици на данните (James et.al., 2013), (Vieira, 2020), (Yeturu, 2020). Моделът е лесно приложим към данни с голяма размерност, като позволява по-бързи изчисления в

сравнение с някои други класификационни модели като логистичната регресия (James et.al., 2013).

Поради широкообхватните предимства на модела на опорните вектори, той е обект на втора глава. Предимствата на тази процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество са загатнати от (Efron et. al., 1997) при метода на най-близките съседи, логистична регресия и дървета на решенията. След тях това приложение на буутстрап процедурата не е изследвано по-нататък и остава загатнато в литературата (James et. al., 2013), (Yeturu, 2020).

Целта на втора глава е да *изследваме буутстрап процедурата като метод за разделяне разделяне на наблюденията в тренировъчно и тестово подмножество на (големи) множества от наблюдения. Формулираме хипотезата, тя може да бъде алтернатива на десеткратната крос валидация.* За да изпълним целта предлагаме нова нейна реализация за Python 3.6. Предложената от нас модификация наричаме Алгоритъм 4. Формулираме следните задачи на главата:

1. Да се направи сравнителна характеристика между реализацията на буутстрап процедурата за разделяне на наблюденията на тренировъчно и тестово подмножество и десеткратна крос валидация по отношение на прогнозна точност.
2. Да се сравни алгоритъмът с буутстрап процедурата и алгоритъмът с десеткратната крос валидация по отношение на класификационните показатели.
3. Да се изследва въпросът дали така предложената модификация от нас ANOVA-Bootstrap-SVM води до съкращаване на времето на прилагане на модела на опорните вектори.

**2.2 Методология**

Всеки от изложените алгоритми в тази секция има за цел да реши C-оптимизационната задача на модела на опорните вектори (Cortes, 1995). Ние предлагаме Алгоритъм 4 като нова модификация на модела на опорните вектори. Алгоритми 1,2 и 3 се наричат класически алгоритми, които се дават от учебниците

(James et. al., 2013) като възможни за разделяне на наблюденията на тренировъчно и тестово подмножество.

### 2.2.1 Алгоритъм 1: Train/test split и десеткратна крос валидация

Стъпка 1: Данните се зареждат в Python 3.6. Определят се независимите променливи и целевата променлива (target variable) за всяко множество данни. Не се прилагат трансформации към данните.

Стъпка 2: Настройва се моделът на опорните вектори. Използва се функцията SVC от модула sklearn.SVC.svm в Python Запазват се настройките на параметрите, които са зададени по подразбиране, като се задава ядро, базирано на радиална базисна функция (radial basis function/rbf) kernel='rbf'. Фиксираме стойността на параметъра C да бъде равен на 1 (Pedregosa et. al., 2011).

Стъпка 3: Използва се пропорция 70/30 за разделяне на наблюденията на тренировъчно и тестово подмножество. Такава пропорция е често използвана в академичната литература (James et. al., 2013). Това става чрез функцията train _test_split и kfold = model_selection.KFold(n_splits=10, random_state=7).

Стъпка 4: В стъпка 4 се настройват параметрите за извършване на десеткратната крос валидация, чрез която данните се разделят на тренировъчно и тестово подмножество за прогнозиране (James et. al., 2013) на база на входните данни X и y. Това става чрез функцията KFold(): kfold = model_selection.KFold(n_splits=10, random_state=7) от модула model_selection в Python 3.6. Всички получени тренировъчни и тестови подмножества съдържат наблюдения без повторение (James et. al., 2013), (Vieira, 2020). Параметърът n_splits във функцията KFold() се използва като се зададе стойност 10. По този начин се задава десеткратната крос валидация в софтуера. Параметърът random_state се използва, за да е възможно възпроизвеждане на резултатите.

Стъпка 5: Използва се функцията cross_val_score, за да се построи модел на опорните вектори, който се изпълнява чрез данните от тренировъчното множество, получени чрез train _test_split и десеткратна крос валидация,т.е. чрез реда results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)

модула model selection в Python 3.6. По този начин се извършва десеткратна крос валидация, като са създадени допълнителни тренировъчно и тестово подмножество чрез train_test_split. Тази стъпка се нарича fitting и в таблиците с резултати в следваща секция е представено времето за изпълнението ѝ чрез термина time for fitting SVM.

Стъпка 6: Изпълнява се моделът на опорните вектори за класификация чрез функцията cross_val_predict, оригиналните данни (променливите X и таргета y) и параметър cv=10 от модула cross_validation. По този начин се получава прогноза за принадлежността на всяко наблюдение от тестовите данни, т.е. прогнозна точност на база на тестовите данни.

### 2.2.2 Алгоритъм 2: Train/test split и Leave-one-out (looCV) крос валидация

Стъпка 1: Данните се зареждат в Python 3.6. Определят се независимите променливи и таргета (target variable) за всяко множество данни. Не се прилагат трансформации към данните.

Стъпка 2: Настройва се моделът на опорните вектори. Използва се функцията SVC от модула sklearn.SVC.svm в Python Запазват се настройките на параметрите, които са зададени по подразбиране, като се задава ядро, базирано на радиална базисна функция (radial basis function/rbf) kernel='rbf'. Фиксираме стойността на параметъра C да бъде равен на 1 (Pedregosa et. al., 2011).

Стъпка 3: Използва се пропорция 70/30 за разделяне на наблюденията на тренировъчно и тестово подмножество. Това става подобно на предния алгоритъм чрез функцията train _test_split, но разликата е, че вместо KFold() функцията се използва функцията LeaveOneOut(p=N). По този начин се изпълнява leave-one-out крос валидация (James et. al., 2013), (Pedregosa et. al., 2011). Тя се различава от десеткратната крос валидация по това, че K=N. Това означава, че броят итерации и получените тренировъчни и тестови подмножества не са десет а са равни на броя наблюдения в множеството данни. Това прави този метод за получаване на тренировъчно и тестово подмножество значително по-бавен от десеткратната крос валидация (James et. al., 2013), (Yeturu, 2020), (Vieira, 2020). Тъй като използваме

девет множества с различна размерност за провеждане на експериментите, K при този алгоритъм е различно за всяко множество от данни, тъй като всяко съдържа различен брой наблюдения.

Стъпка 4: Използва се функцията cross_val_score, за да се построи модел на опорните вектори, който се изпълнява чрез данните от тренировъчното множество, получени чрез train _test_split и leave out cross validation в Python 3.6. По този начин се извършената leave-one-out cross validation също съдържа допълнително тренировъчно и тестово подмножество, получено чрез функцията train_test_split. Тази стъпка се нарича fitting и в таблиците с резултати в следваща секция е представено времето за изпълнението й чрез термина time for fitting SVM.

Стъпка 5: Изпълнява се моделът на опорните вектори за класификация чрез функцията cross_val_predict, оригиналните данни (променливите X и таргета y) и leave-one-out cross validation. Записано е времето да извършването на тази стъпка в таблиците с резултатите под името “time for prediction”.

### 2.2.3 Алгоритъм 3: Repeated random train/test split

Стъпка 1: Данните се зареждат в Python 3.6. Определят се независимите променливи и таргета (target variable) за всяко множество данни. Не се прилагат трансформации към данните.

Стъпка 2: Настройва се моделът на опорните вектори. Използва се функцията SVC от модула sklearn.SVC.svm в Python Запазват се настройките на параметрите, които са зададени по подразбиране, като се задава ядро, базирано на радиална базисна функция (radial basis function/rbf) kernel=’rbf’. Фиксираме стойността на параметъра C да бъде равен на 1 (Pedregosa et. al., 2011).

Стъпка 3: Използва се пропорция 70/30 за разделяне на наблюденията на тренировъчно и тестово подмножество. За да се приложи методът за разделяне на наблюденията на тренировъчно и тестово подмножество, наречен repeated random train/test split (случайно разбъркване с повторение) се използва функцията ShuffleSplit от model_selection модула.

Стъпка 4: Използва се функцията cross_val_score, за да се построи модел на опорните вектори, който се изпълнява чрез данните от тренировъчното подмножество, получени чрез функцията ShuffleSplit в Python 3.6 Тази стъпка се нарича fitting и в таблиците с резултати в следваща секция е представено времето за изпълнението й чрез термина time for fitting SVM.

Стъпка 5: Изпълнява се моделът на опорните вектори за класификация чрез функцията cross_val_predict и функцията ShuffleSplit. Записано е времето да извършването на тази стъпка в таблиците с резултатите под името “time for prediction”.

За всеки от трите класически алгоритми се записва прогнозната точност, получена от тестовите данни и се изчисляват класификационните показатели precision, recall, $f_1$ -score.

### 2.2.4 Алгоритъм 4: Буутстрап алгоритъм

Стъпка 1: Дефинираме независимите променливи и целевата променлива, без да има трансформации на данните.

Стъпка 2: Определяме пропорция за получаване на тренировъчни и тестови данни 30:70. Тази пропорция се различава от алгоритми 1-3. На тази стъпка може да бъде избрана и друга пропорция по преценка на автора. В случая се използва пропорция 30:70, за да се провери как това ще се отрази на прогнозната точност на модела на опорните вектори в комбинация с буутстрап процедурата.

Стъпка 3: Използваме същите настройки на функцията SVC в Python 3.6 както при алгоритми 1-3. Това са rbf ядро и параметър C фиксиран на 1.

Стъпка 4: Изпълняваме буутстрап процедурата (Efron, 1979), (James et. al., 2013), като начин да разделим множеството на тренировъчно и тестово вместо крос валидацията в алгоритми 1-2 и случайното разделяне с повторение в алгоритъм 3. Редица приложения на буутстрап процедурата изискват да бъдат направени между 100 и 1000 итерации, за да е надежден резултатът (Efron, 1979). Ние предлагаме

използването на 10 итерации на буутстрап процедурата, за да се види дали той може да бъде аналог на десеткратната крос валидация.

Стъпка 5: Получаваме тренировъчните данни, върху които се построява моделът на опорните вектори. Времевият резултат от този процес е описан в резултатите като time for fitting.

Стъпка 6: Тестовите данни се използват, за да се направи прогноза за класа, към който принадлежат. Времевият резултат по-нататък от тази процедура е описана като time for prediction.

### 2.3 Данни

Таблица 2.1 описва множествата данни и таргетите:

Таблица 2.1: Описание на използваните в експериментите множества

| Множество | Брой наблюдения (N) | Брой променливи (p) | Таргет (y) |
|---|---|---|---|
| glass | 175 | 9 | Type |
| leaf | 286 | 7 | arch |
| wells | 3020 | 4 | association |
| fraud | 3255 | 4 | IsFraud |
| abalone | 4177 | 8 | Rings |
| ed | 5785 | 5 | difference |
| monica | 6367 | 11 | outcome |
| food | 23971 | 5 | sex |
| adult | 45222 | 13 | income |

Източник: www.kaggle.com

## 2.4 Числени експерименти

### 2.4.1 Прогнозна точност

Таблица 2.2

| множество | n | p | 10-кратна cv - общо време | btsp - общо време | прогнозна точност при 10-кратната cv | прогнозна точност при btsp |
|---|---|---|---|---|---|---|
| glass | 175 | 9 | 0.03 | 0.03 | 0.69681 | 0.65297 |
| leaf | 286 | 7 | 0.08 | 0.00 | 0.58042 | 0.69212 |
| wells | 3020 | 4 | 4.49 | 0.84 | 0.53510 | 0.53899 |
| fraud | 3255 | 4 | 9.05 | 2.11 | 0.65131 | 0.65200 |
| abalone | 4177 | 8 | 7.57 | 2.23 | 0.52909 | 0.52899 |
| ed | 5785 | 5 | 27.62 | 5.62 | 0.86845 | 0.86717 |
| monica | 6367 | 11 | 15.62 | 2.03 | 0.87655 | 0.87303 |
| food | 23971 | 5 | 2294.93 | 93.09 | 0.86037 | 0.86032 |
| adult | 45222 | 13 | 7219.93 | 830.44 | 0.75227 | 0.75321 |

Източник: авторски изчисления

Както се вижда от таблица 2.2 в повечето случаи буутстрап процедурата запазва прогнозната точност, получена чрез крос валидация или я повишава. Затова десеткратният буутстрап, изпълнен по начина, предложен в секция „Методология“, може да използва като алтернатива на десеткратната крос валидация при големи данни. Буутстрап процедурата съкращава изчислителното време на модела на опорните вектори във всички случаи – от малки до големи по размер множества. По този начин данните могат да бъдат опознати по-бързо и да се прецени дали даденият модел е подходящ за тях. Така се решава първата задача на втора глава – прави се сравнителна характеристика между реализацията на буутстрап процедурата за разделяне на наблюденията на тренировъчно и тестово подмножество и десеткратна крос валидация по отношение на прогнозната точност.

### 2.4.2 Recall, precision, f1-score

В дисертацията представяме резултатите от класификационните показатели recall, precision и f1-score в отделни таблици за всяко множество. Тъй като тези резултати са обемни, таблиците не са показани тук. Но чрез проведения сравнителен анализ се решава втората задача на провеждания анализ - да се направи сравнителна

характеристика между алгоритъма с буутстрап процедурата и алгоритъма с десеткратната крос валидация по отношение на класификационните показатели.

Накратко, изводите, които можем да направим, се свеждат до обобщението, че буутстрап процедурата не води до загуба на прогнозна точност. В някои случаи води до подобряването ѝ, както и на класификационните метрики. В повечето случаи моделът на опорните вектори с буутстрап процедурата води до сходна прогнозна точност както при десеткратната крос валидация. Класификационните показатели precision, recall и $f_1$ – score, резултат от разделянето на данните на тренировъчно и тестово подмножество чрез буутстрап процедурата, са сходни на получените от десеткратната крос валидация.

### 2.4.3 Изчислително време

Таблица 2.3 Времена за изпълнение на тренировъчни и тестови модели

| | | | Time for fitting SVM (s)/Време за тренировъчен модел | | Time for prediction (s)/ Време за тестови модел | |
|---|---|---|---|---|---|---|
| множество | n | p | 10-кратна cv | btsp | 10-кратна cv | btsp |
| glass | 175 | 9 | 0.02 | 0.00 | 0.02 | 0.03 |
| leaf | 286 | 7 | 0.04 | 0.00 | 0.04 | 0.00 |
| wells | 3020 | 4 | 0.56 | 0.32 | 3.93 | 0.52 |
| fraud | 3255 | 4 | 3.01 | 0.59 | 6.04 | 1.52 |
| abalone | 4177 | 8 | 2.51 | 0.64 | 5.06 | 1.60 |
| ed | 5785 | 5 | 9.15 | 1.86 | 18.47 | 3.76 |
| monica | 6367 | 11 | 5.21 | 0.93 | 10.41 | 1.11 |
| food | 23971 | 5 | 763.03 | 30.33 | 1531.90 | 62.76 |
| adult | 45222 | 13 | 2434.45 | 528.61 | 4785.48 | 301.83 |

Източник: авторски изчисления

Таблица 2.3 показва, че при най-голямото множество – adult, буутстрап процедурата намалява времето за изграждане на тренировъчен модел от близо 40,5 часа на около 9 часа. Докато при построяването на прогнозен модел, той се справя за около 5 часа вместо 80 часа при десеткратната крос валидация.

Буутстрап процедурата съкращава изчислителното време за построяване на модела и прогнозиране на данните няколко пъти в сравнение с десеткратната крос валидация. Изчислителното предимство на буутстрап процедурата става по-ясно изразено с нарастването на размера на множествата. При най-малкото множество glass / 175 наблюдения/, буутстрап процедурата и крос валидацията имат сходно

време за построяване и прогнозиране на модела. Когато размерът на данните се увеличи, например се вземе средно по размер множество като monica, буутстрап процедурата построява тренировъчния модел пет пъти по-бързо и прави прогноза десет пъти по-бързо от десеткратната крос валидация. Подобни изводи могат да бъдат направени и за останалите множества от таблица 2.3.

Изводът от таблица 2.3, до който стигаме е, че предложената имплементация на буутстрап процедурата в секция „Методология“ води до значително подобряване на времената за получаване на тренировъчни и тестови модели, като това предимство става по-ясно изразени при големи множества от наблюдения. Този резултат е важен, тъй като показва как може да бъде съкратено изчислителното време при големи множества от наблюдения, където изчисленията могат да отнемат дни и седмици поради обема на данните.

По този начин се решава третата задача на главата– да се изследва доколко буутстрап процедурата в така предложената имплементация води до съкращаване на времето на прилагане  на модела на опорните вектори.

## 2.5 Заключение

Най-важните изводи, които са показани в секция „Числени експерименти“ могат да бъдат обобщени в няколко посоки:

1. Буутстрап процедурата може да бъде използвана като надеждна алтернатива на крос валидацията и нейни разновидности като десеткратната и leave-one-out в модела на опорните вектори. Това се дължи на първо, способността на буутстрап процедурата да получи прогнозна точност и класификационни показатели, близки или по-добри от тези, получени в резултат на крос валидацията. Второ, той извършва класификацията в пъти по-бързо от крос валидацията.
2. Това свойство на буутстрап процедурата се запазва с увеличаване на размерността на извадката. Така разликата в изчислителното време между буутстрап процедурата и десеткратната крос валидация при големи множества може да бъде съкратена от няколко дена до няколко часа.

3. Значителното изчислително предимство на буутстрап процедурата има важно приложение за големите множества данни, където може да се използва както като метод за разделяне на наблюденията на тренировъчно и тестово подмножество при модела на опорните вектори за построяването на самостоятелна класификация. Но той в комбинация с модела на опорните вектори може да се използва и като модел за първоначално опознаване на данните и преценяване дали моделът на опорните вектори или друг класификационен модел е най-подходящ. Това е възможно именно заради изчислителното предимство на буутстрап процедурата.
4. Друг важен извод е, че изчислителното предимство на буутстрап процедурата се дължи на десетте итерации. Те са достатъчни, за да се съкрати изчислителното време за модела, като се запази или подобри прогнозната точност спрямо десеткратната крос валидация. Приложението на буутстрап процедурата с десет итерации е нов подход, тъй като в практиката се е наложило правилото, че той трябва да се използва с голям брой итерации, например 100,1000 и т.н. (James et. al., 2013).
5. Разглеждането на буутстрап процедурата не само като теоретичен метод, но и като софтуерен алгоритъм също е важно за ускоряването на резултатите от него. Съществуващите версии на буутстрап процедурата в Python 3.6 (Brownlee, 2017), които могат да бъдат използвани като метод за разделяне на наблюденията на тренировъчно и тестово подмножество, не са достатъчно бързи. Затова е необходимо да се предложи оптимизирана версия на съществуващия алгоритъм, който променя само синтаксиса на програмния код, но не и теоретичната постановка зад метода.

## 2.6 Дискусия

Въпреки че буутстрап процедурата не е нов статистически метод, втора глава представи нови практически предимства на буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество. Използван като такъв, буутстрап процедурата позволява да се запази/подобри прогнозната точност на модела на опорните вектори, като се ускорява класификацията. Това

практическо предимство на буутстрап процедурата позволява да бъдат тествани значително по-бързо както различни версии на модела на опорните вектори, така и различни класификационни модели върху големи данни.

От една страна резултатите, които получаваме при различните класификационни модели, показват предимствата на новата имплементация на буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество. От друга страна, трябва да се има предвид, че буутстрап процедурата може да не работи добре при всеки класификационен модел, което прави крос валидацията универсален метод за разделяне на наблюденията на тренировъчно и тестово подмножество.

В същото време буутстрап процедурата и предложената нова негова имплементация дават много добри резултати при едни от най-често използваните класификационни модели като модела на опорните вектори. Неговата способност да запази прогнозната точност на модела, като съкрати значително изчислителното време има важно значение в контекста на големите данни.

Предложената от нас буутстрап процедура позволява както много по-бързо да бъдат откривани подходящите класификационни модели за големи множества от наблюдения, така и да бъде извършвана много по-бърза прогноза с тях. С това се разширяват известните досега предимства на буутстрап процедурата в литературата, като показваме, че той може да бъде надеждна алтернатива на крос валидацията.

***Списък с публикации по втора глава:***

- Vrigazova, Borislava & Ivanov, Ivan. (2020). TENFOLD BOOTSTRAP PROCEDURE FOR SUPPORT VECTOR MACHINES. Computer Science 21. 241-257. 10.7494/csci.2020.21.2.3634. – научна статия в международно списание, индексирана в Скопус

## Глава 3. Модификации на модела на поддържащите вектори (Support vector machines) на основа на буутстрап процедурата

Втора глава показва, че буутстрап процедурата може да бъде алтернатива на крос валидацията при модела на опорните вектори без избор на променливи. Буутстрап процедурата отново е предмет на трета глава, но обект е моделът на опорните вектор с избор на променливи чрез ANOVA.

### 3.1 Преглед на литературата

В статията „Pattern recognition using generalized portrait method' от 1963г., Вапник и Лърнър предлагат модел, който да открива общи характеристики в последователни изображения и да ги разделя в различни групи на база на установени граници между характеристиките на наблюденията. Те правят опит в статията си да формализират и изяснят по-добре концепции като „разпознаване на характеристики" (pattern recognition), но успяват да създадат нови понятия и да предложат нов алгоритъм, който да разграничава последователни изображения на база на праг (threshold) на подобие/различие. Те предлагат подход, който първо разпознава (recognize) характеристиките на изображенията, а след това ги разграничава в групи (distinguish).

Година по-късно Вапник (Vapnik, 1964) доразвива идеята за generalized portrait, т.е. за образуването на клъстъри от подобни наблюдения на база на праг за разделяне на наблюденията. През 1992г. Вапник и Бозер предлагат начин да се избере най-подходящият праг за разграничаване и класифициране на наблюденията, който те наричат margin (Boser, 1992). Съвременната версия на модела на поддържащите вектори е заложена в (Cortes, 1995) и имплементирана в библиотеките на различни софтуерни продукти, като в R и Python на база на библиотеката, написана от Chang (Chang, 2001).

Базовите версии на метода за решаване на модела на поддържащите вектори, които Вапник и Кортес (Cortes, 1995) предлагат, са широко използвани и до днес. Една версия, която предлагат, се нарича C-оптимизация. При нея класификацията зависи от стойността на оптимизационен параметър C. Именно тази версия на модела на опорните вектори е обект както на втора, така и на трета глава.

*Целта в трета глава е да се предложат нови модификации на модела на поддържащите вектори и в метода за решаването му, които да подобряват ефективността (отношението на постигнатия резултат спрямо поставената цел) на класическия модел с ANOVA.* Формулира се хипотезата, че съществуват модификации на модела на опорните вектори с буутстрап процедурата и ANOVA, които подобряват ефективността на модела.

Задачите, които се поставят в главата са следните:

1. Да се направи сравнение между буутстрап модификациите и тяхната прогнозна точност и съществуващи модификации на модела на опорните вектори от академичната литература (Maldonado et. al., 2014), както и други класификационни модели (Naidu et al., 2012), (Miao et al., 2016), (Khanna et al., 2015), (Kodati et al., 2018), (Latha & Jeeva, 2019), (Nandipati, 2020) и др.
2. Да се проследи как се променят AUC точките при буутстрап модификациите спрямо тези, получени от метода на целочисленото линейно смятане (mixed linear integer approach, комбиниран с модела на опорните вектори), приложен в (Maldonado et. al., 2014).

### 3.2 Методология

За разлика от алгоритмите в предишната глава, в тази глава правим предварителна трансформация на входните данни, която е описана чрез уравнения (3.1) и (3.2).

Независимите променливи може да бъдат нормализирани, ако са налични отрицателни числови стойности в тях (James et. al., 2013). Това означава данните да имат стойности между 0 и 1. Уравнение (3.1) показва нормализацията:

$$n_{ij} = \frac{z_{ij} - \min(z_{ij})}{\max(z_{ij}) - \min(z_{ij})}, \tag{3.1}$$

където $z_{ij}$ бележи стандартизираните данни.

Препоръчваното в литературата (James et. al., 2013), (Vapnik, 2012) стандартизиране на независимите променливи има вида:

$$z_{ij} = \frac{x_{ij-\mu}}{\sigma} \ , \qquad i = 1,..l \ , \ j = 1,..p \ , \tag{3.2}$$

където μ и σ обозначават средната стойност и стандартното отклонение на дадената променлива.

Така класическата версия на модела на опорните вектори представяме чрез следния алгоритъм, който се различава от представения класически алгоритъм в предишната глава.

**Алгоритъм ANOVA-CV-SVM**. Класически подход за решаване на модела на опорните вектори

1. Методът ANOVA (Smalheiser, 2017) се прилага към всеки процентил[1] от броя на независимите променливи. Той се изпълнява в комбинация от следните стъпки:

1.1. Нормализиране на данните в интервала 0-1 чрез уравнение (3.1). Това се прави с цел да се избегне наличието на отрицателни стойности в независимите променливи.

1.2. Независимите променливи се стандартизират чрез уравнение (3.2)

1.3.Стойността на параметъра C се фиксира 1 (Pedregosa et. al., 2011).

1.4.Прилагат се два вида ядро – линейно или rbf ядро.

1.5.Данните се разделят на тренировъчно и тестово подмножество (крос валидация. Въпреки че се препоръчва тя да бъде десеткратна или к-кратна (k-fold) (James et. al., 2013), ние прилагаме т.нар. randomized CV splitters, като избираме разновидността, наречена ShuffleSplit от sklearn.model_selection модула в Python 3.6 (Pedregosa et. al., 2011).

1.6. Решаваме оптимизационната задача (2.1)-(2.2) за тренировъчното подмножество, като вече са удовлетворени и ограничения (3.1) и (3.2).

[1] Процентилите са десети, двадесети, тридесети, четиридесети, петдесети, шестдесети, седемдесети, осемдесети, деветдесети и стотния.

1.7.Отчитат се прогнозната точност за всеки процентил от броя променливи в данните.

2. Избира се процентила, осигуряващ най-висока прогнозна точност.

3. Отчитат се точността, броят променливи и AUC точките (Yeturu, 2020) и получените резултати върху шестте множества се сравняват с тези на Малдонадо (2014).

В следващата подсекция предлагаме две модификации на алгоритъм **ANOVA-CV-SVM**. Когато алгоритъмът **ANOVA-CV-SVM** се изпълнява с линейно ядро, той се нарича ANOVA-CV-L-SVM, а с rbf ядро – ANOVA-CV-RBF-SVM. Модификациите, които представяме се базират на модела ANOVA-BOOTSTRAP-SVM, който описваме в главата.

**Модификация: ANOVA-BOOTSTRAP-SVM**

Модификацията, която в тази глава предлагаме, се базира на класическата ANOVA-CV-SVM, но за получаване на тренировъчното и тестовото множество се използва буутстрап процедурата (Efron, 1979). Модификацията се състои от няколко стъпки:

1.За всеки процентил[2] от броя на независимите променливи се изпълнява метода ANOVA (Smalheiser, 2017), като той се комбинира със следните стъпки:

1.1.Независимите променливи се стандартизират чрез уравнение (3.2).

1.2.Нормализиране на данните поради наличието на отрицателни стойности в някои от тях чрез уравнение (3.1).

1.3.Стойността на параметъра C се фиксира на 1 (Pedregosa et. al., 2011).

1.4.Прилагат се два вида ядро – линейно или rbf ядро.

[2] Процентилите са десети, двадесети, тридесети, четиридесети, петдесети, шестдесети, седемдесети, осемдесети, деветдесети и стотния.

1.5.Данните се разделят на тренировъчно и тестово подмножество чрез **буутстрап процедурата** (Efron, 1979), (James et. al., 2013). Това е стъпката, която отличава предложената модификация от класическия ANOVA-CV-SVM.

1.6.Решаваме оптимизационната задача (2.1)-(2.2) за тренировъчното подмножество, като вече са удовлетворени ограничения (3.1) и (3.2).

1.7.Отчита се прогнозната точност за текущия процентил от променливи.

2.Избира се процентила, осигуряващ най-висока прогнозна точност.

3. Записват се точността, броят променливи и AUC точките (Yeturu, 2020), (Veiera, 2020) за множествата от (Maldonado et. al., 2014) и прогнозната точност за останалите множества.

Версията с линейно ядро се нарича ANOVA-Bootstrap-L-SVM, а с rbf ядро – ANOVA-Bootstrap-RBF-SVM. Това са двете модификации на модела на опорните вектори, които предлагаме.

### 3.3 Данни

Таблица 3.1: Използвани данни

| Множество данни | Брой наблюдения (l) | Брой променливи (p) |
|---|---|---|
| Wisconsin breast cancer (WBC) | 569 | 30 |
| Australian credit approval | 690 | 14 |
| PIMA diabetis dataset | 768 | 8 |
| ionosphere dataset | 351 | 35 |
| splice dataset | 3190 | 61 |
| Colorectal cancer | 63 | 2000 |
| Liver dataset | 345 | 6 |
| Cleveland heart disease | 303 | 14 |
| HCV | 1385 | 29 |

Източник: https://archive.ics.uci.edu/ml/datasets/

### 3.4 Числени експерименти

Следващите подсекции показват получените резултати от описаните модели в секцията „Методология“. Те се съпоставят както с изпълнената класическа версия

на модела на опорните вектори с ANOVA, така и с изследванията на други автори. В (Maldonado et. al., 2014) са проведени експерименти с първите шест множества, показани в таблица 3.1, но чрез целочислен линеен метод (mixed linear integer approach), комбиниран с модела на опорните вектори. Първите шест подсекции съдържат таблици, които сравняват нашите резултати с тези, получени от (Maldonado et. al., 2014). Останалите три множества използваме, за да сравним буутстрап модификациите с други класификационни модели, използвани в други изследвания (Naidu et al., 2012), (Miao et al., 2016), (Khanna et al., 2015), (Kodati et al., 2018), (Latha & Jeeva, 2019), (Nandipati, 2020) и др.

За всяко множество в дисертацията са представени отделни таблици с резултати, но тъй като са обемни, тук ще обобщим основните изводи.

В хода на експериментите е необходимо за някои множества да се определи подходящата стойност на параметъра C, за да се повиши прогнозната точност на модела. Също така при някои данни е по-подходящо линейно ядро, при други – rbf ядро. Така ограничения на новия модел ANOVA-Bootstrap-SVM, който предлагаме, се явяват параметъра C и ядрото.

Въпреки това емпиричните резултати от проведените експерименти за тези множества показват, че предложените буутстрап модификации могат да се използват, за да се получи по-висока прогнозна точност и да се подобрят AUC точките в сравнение с модела на опорните вектори, комбиниран с метода на линейното целочислено смятане (Maldonado et. al., 2014). Нашите модификации запазват своите предимства дори когато е нарушено правилото броят променливи да е по-малък от броя наблюдения. По този начин се предлага начин да се получат надеждни резултати при случай, който академичната литература разглежда като изключение, за което са необходими други модели (James et. al., 2013).

При останалите три множества буутстрап процедурата доведе до повишаване на пронозната точност не само спрямо класическата версия на модела на опорните вектори, но и в спрямо други класификационни модели като дървото на решенията, модела на най-близките съседи и други. Резултати, постигнати от други автори върху тези множества, са подобрени чрез от нас чрез буутстрап процедурата. Таблиците,

разглеждащи тези научни постижения за всяко множество, са представени в дисертацията.

Така чрез проведените експерименти се решават поставените две задачи в началото на главата. Чрез тях се изпълнява целта на тази глава, а именно да се предложат нови модификации на модела на поддържащите вектори и в метода за решаването му, които да подобряват ефективността (отношението на постигнатия резултат спрямо поставената цел) на класическия модел.

### 3.5 Заключение

В тази глава се предлагат модификации на модела на опорните вектори, които целят да се повиши прогнозната точност на класификацията спрямо съществуващи модели в литературата. Резултатите от нашите експерименти водят до няколко извода:

1. Буутстрап модификациите могат да повишат както прогнозната точност, така и AUC точките, получени от модификации на модела на опорните вектори, комбиниран с алгоритъма на целочисленото линейно смятане, и в частност модификациите, предложени от Малдонадо (Maldonado et. al., 2014).
2. Буутстрап модификациите могат да подобрят прогнозната точност, резултат както от други версии на модела на опорните вектори (Nandipati, 2020), (Li et. al., 2011), (Gestel et. al., 2002), (Lee and Mangasarian, 2001), така и от други класификационни модели като модела на най-близките съседи, random forest, невронни мрежи (Nandipati, 2020), хибридни модели (Ozsen and Gunes, 2008) и др.
3. При някои множества C=1 е достатъчно, за да бъде подобрена прогнозната точност на модела, докато при други тя трябва да бъде правило избрана, като изследователят преценява как да я избере.
4. Буутстрап модификациите успяват да подобрят прогнозната точност и при множество, където броят на променливите е по-голям от броя на наблюденията. С това се разширяват възможностите в литературата за

решаване на класификационни проблеми в подобен тип данни и се допълват практическите предимства на бустрапа .

### 3.6 Дискусия

Получените от нас резултати за модела ANOVA-Bootstrap- SVM надминават редица резултати от други научни изследвания, което показва практическия принос на нашите резултати в академичната литература. Въпреки това нашите изводи могат да бъдат задълбочени и по посока сравнение на изчислителното време между предложения модел от нас ANOVA-Bootstrap-SVM и други съществуващи модификации на модела на опорните вектори. В хода на експериментите, които провеждаме, правим сравнение за някои данни по показателя „време“ между ANOVA-Bootstrap-SVM и наложилият се в литературата ANOVA-CV-SVM (James et. al., 2013). Изводите, които могат да се направят, са, че моделът ANOVA-Bootstrap-SVM се справя по-бързо от модела ANOVA-CV-SVM дори при големи данни. Въпреки това е необходимо да се проведат повече изследвания, за да разбере за кои видове класификационни модели е подходящ буутстрап с ANOVA, както и спрямо кои други модели той има изчислително предимство и дали винаги е налице то.

Въпреки че нашите изследвания могат да бъдат задълбочени в различни посоки, те представят първите крачки към по-задълбочен анализ на буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество, тъй като той има важни практически приложения, част от които до момента не са били достатъчно изследвани и широко прилагани.

Загатнатите, но неизследвани приложения на буутстрап процедурата в панели с фиксирани ефекти (Breiman, 1995), както и неговите възможни приложения като алтернатива на крос валидацията (Efron et. al., 1979) налагат да проведем първото най-мащабно изследване на буутстрап процедурата като метод от алгоритмичното учене от данни.

В дисертацията представяме три глави, всяка от които постига основна цел. Целта на първа глава е да предложим интердисциплинарен подход при моделиране на панелни данни. Въпреки че първата цел изглежда общо формулирана, нейното

изпълнение минава през редица конкретни модели от класическата иконометрична теория и алгоритмичното учене от данни, за да се моделира индексът на имуществените права. Основните изводи, които можем да направим, след постигането ѝ, могат да изкажат първата цел по различен начин.

Изводът, че методите за избор на променливи трябва да отчитат панелните ефекти, за да са конкурентни на иконометричните модели, доведе до откритието, че буутстрап процедурата може да отчете фиксирани ефекти и да се използва с метода за избор на променливи nonnegative garrote. Това откритие, от друга страна, позволява първата цел на дисертацията, съответно на първа глава, да бъде преформулирана като „Да покажем, че буутстрап процедурата може да отчита фиксирани панелни ефекти и да се използва с nonnegative garrote за избор на променливи при панелни данни“. Именно тази формулировка е ключова за цялата дисертация, тъй като тя превръща буутстрап процедурата предмет на дисертацията. Така постигайки целта на първа глава, ние показваме иновативни приложения на буутстрап процедурата при панелни данни. Именно съпоставката на двата подхода и търсенето на интердисциплинарен подход ни доведе до това научно откритие и до целта на втора глава.

Целта на втора глава  е да изследваме буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество при класификационни проблеми без избор на променливи. Тя е по-конкретно формулирана, тъй като вече знаем предимството на буутстрап процедурата при панели и логично бе да изследваме и въпроса при методите от алгоритмичното учене от данни. Тъй като буутстрап процедурата е подобен на крос валидацията, а алгоритмичното учене от данни изисква разделяне на тренировъчно и тестово подмножество, ние задълбаваме в приложението на буутстрап процедурата като такъв метод.

Тъй като тези приложения не са добре проучени до момента, започваме без избор на променливи. В зависимост от получените резултати имаме две възможности. Първо, буутстрап процедурата да не постига подобни класификационни резултати като крос валидацията. Това би означавало, че

изпълнението на втората цел не води до смислен научен резултат. Втората възможност е буутстрап процедурата да дава сходни или по-добри резултати от крос валидацията, т.е. целта на втора глава да се изпълни чрез научен принос. Експериментите ни показват, че попадаме във втората възможност - буутстрап процедурата може да се използва като алтернатива на десеткратната крос валидация. Така разширяваме практическите му приложения.

Този извод може да преобразува втората цел на: „Да покажем, че буутстрап процедурата може да се използва като алтернатива на десеткратната крос валидация при модела на опорните вектори без избор на променливи“. По този начин става ясно защо втората цел има ключова роля за третата цел.

Целта на трета глава - да изследваме буутстрап процедурата като метод за разделяне на наблюденията на тренировъчно и тестово подмножество при класификационни проблеми с избор на променливи. Третата цел следва от втората. След като показваме, че буутстрап процедурата може да получи тренировъчно и тестово подмножество при модела на опорните вектори без избор на променливи, е необходимо да проверим хипотезата, че това е така и когато има избор на променливи. **Т**ъй като резултатите в трета глава потвърждават тази хипотеза, то целта на трета глава може да бъде преформулирана по следния начин: „да се покаже, че буутстрап процедурата като метод за получаване на тренировъчно и тестово множество е приложим при модела на опорните вектори и с избор на променливи чрез ANOVA“.

Досега буутстрап процедурата се е прилагала основно в статистиката при Монте Карло симулации, за апроксимиране на нормалното разпределение и др. Но приложенията му в алгоритмичното учене от данни като метод за разделяне на наблюденията на тренировъчно и тестово подмножество и приложенията му в панелни данни не са изследвани. Затова дисертационният труд, който представяме, може да се счете за най-мащабното проучване на буутстрап процедурата като метод от алгоритмичното учене от данни.

***Списък с публикации по трета глава:***

- Vrigazova, Borislava & Ivanov, Ivan. (2019). Optimization of the ANOVA Procedure for Support Vector Machines. International Journal of Recent Technology and Engineering. 8. 5160. 10.35940/ijrte.D7375.118419. – статия в международно научно списание с редакционна колегия

## III. Научен принос на проведеното авторско изследване

Научният принос на проведеното изследване в дисертацията може да бъде обобщен по глави в няколко посоки.

**Първа глава**

Първа глава изследва приложението на буутстрап процедурата на Брайман (Breiman, 1979) в контекста на панелни иконометрични данни. Тя изследва хипотезата на Брайман, че моделът nonnegative garrote може да бъде използван с буутстрап процедурата в данни с фиксирани регресори. Тази хипотеза проверяваме в контекста на панелни данни с фиксирани ефекти. Формират се следните научни приноси:

1. Разширява се приложението на буутстрап процедурата на Брайман до панелни данни с фиксирани ефекти.
2. Предлага се интердисциплинарна методология за моделиране на панелни данни, при която се обединява класическа иконометрична теория и алгоритмично учене от данни, с цел да се открият ключови променливи за панелния модел.

**Втора глава**

Втора глава изследва приложението на буутстрап процедурата на Ефрон (1979) като метод за разделяне на наблюденията на тренировъчно и тестово подмножество при модела на опорните вектори. Главата допълва съществуващата академична литература, като:

1. Предлагаме буутстрап процедурата на Ефрон като алтернатива на десеткратната крос валидация при модела на опорните вектори за разделяне на наблюденията на тренировъчно и тестово подмножество. По този начин

разширяваме сферите на приложение на буутстрап процедурата в академичната литература.

2. Показваме, че десет итерации на буутстрап процедурата на Ефрон са достатъчни, за да може той да бъде използван като алтернатива на крос валидацията. Това приложение е ново, тъй като Ефрон (1979) препоръча използването на буутстрап процедурата с голям на брой итерации – 100, 1000 и т.н.
3. Предлагаме модификация на буутстрап процедурата, която може да съкрати значително времето за изчисляване, като това разрешава проблемът с бавното смятане при големи множества от наблюдения.

**Трета глава**

В трета глава допълваме приложенията на буутстрап процедурата на Ефрон, като:

1. Предлагаме той да бъде използван като алтернатива на десеткратната крос валидация и при ANOVA модел на опорните вектори.
2. Предлагаме модификация на ANOVA, която наричаме, ANOVA-Bootstrap-SVM, която може да бъде по-ефективна от класическия ANOVA модел.
3. С помощта на предложената от нас модификация успяваме да подобрим получените резултати върху вече изследвани множества от наблюдения. Така разширяваме набора от модели, с които данните могат да бъдат опознати и моделирани по-ефективно.

# IV. Благодарности

Благодаря на моя научен ръководител проф. Иван Ганчев Иванов за оказаната подкрепа по време на всеки етап от разработването и подготвянето на дисертационния труд и публикуваните изследвания по него. Благодарности изказвам и към рецензентите за техните бележки и коментари.

Благодарности бих искала да изкажа и на Софийския университет за получената финансова подкрепа при реализирането и представянето на получените резултати от дисертационния труд пред научната общност.

## V. Списък с публикации по дисертацията

- Vrigazova B., 2017, Property rights: Factors Contributing to the Generation of Income from Ownership of Property, Innovativity in Modeling and Analytics Journal of Research, vol. 2, pp.22-39 – научна статия в международно научно списание с редакционна колегия
- Vrigazova, B. 2018, Nonnegative Garrote as a Variable Selection Method in Panel Data, International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 1, January 2018 – научна статия в международно научно списание с редакционна колегия
- Vrigazova, Borislava & Ivanov, Ivan. (2019). Optimization of the ANOVA Procedure for Support Vector Machines. International Journal of Recent Technology and Engineering. 8. 5160. 10.35940/ijrte.D7375.118419. – статия в международно научно списание с редакционна колегия
- Vrigazova, Borislava & Ivanov, Ivan. (2020). TENFOLD BOOTSTRAP PROCEDURE FOR SUPPORT VECTOR MACHINES. Computer Science 21. 241-257. 10.7494/csci.2020.21.2.3634. – научна статия в международно списание, индексирана в Скопус