СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ”

Факултет по математика и информатика

Катедра „Софтуерни технологии”

Сергей Миланов

Извличане на зависимости в потоци от данни

Автореферат

за присъждане на образователна и научна степен “Доктор”

професионално направление 4.6 „Информатика и компютърни науки“

докторска програма “Компютърни науки (Изкуствен Интелект)“

Научен ръководител: доц. д-р Олга Георгиева

София, 2017

Структура и обем на дисертационния труд

Дисертационният труд е с обем 152 страници и се състои от въведение, пет глави, заключение, общи изводи, предложения за използване на резултатите и виждания за насоките на по-нататъшната работа, а също и списък на цитираната литература - 107 заглавия. Дисертацията съдържа 32 фигури и 17 таблици. Номерата на включените в автореферата фигури и таблици съвпадат с тези в дисертационния труд.

Дисертационният труд е обсъден и насочен за защита от научно звено, включващо преподаватели от катедра „Софтуерни технологии” от Факултет по Математика и Информатика при Софийски университет „Св. Климент Охридски”, състояло се на 24 Януари 2017 година.

Материалите на докторанта са на разположение на заинтересованите във Факултет по Математика и Информатика, стая 209.

Автор: Сергей Миланов

Заглавие: Извличане на зависимости в потоци от данни

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

В днешно време сме заобиколени от данни – нарастващо количество от факти за различни проявления в реалния свят. Те се наблюдават и регистрират в най-различни сфери като икономика, здравеопазване, околната среда, телекомуникации и други, като обемът на данните нараства стремително през последните години.

Данни може да съществуват в изобилие без да се разбира напълно тяхното значение. Възможно е в тях да съществуват закономерности, които не са явни и могат да бъдат разкрити. *Извличането на зависимости от данни* е дисциплината, която се занимава с разкриване на скрити свойства и връзки в изследваните данни и тяхното структуриране в разбираем вид. Тя обхваща методи от изкуствения интелект, машинното самообучение, статистиката и други. Задачите биват разнообразни – разкриване на сходни групи и структури в данните, моделиране и обучение на базата на известни свойства на данните и прилагане на модела върху нови, непознати данни, намиране на асоциации или шаблони в данните, идентифициране на необичайно поведение, предвиждане, прогнозиране и други.

Една значителна част от натрупваните данни включват и времева характеристика, която отразява динамичната промяна и развитие на наблюдаваните обекти и явления във времето. Времевите данни със значителна продължителност и понякога без идея за техния край се наричат *потоци от данни*.

Настоящият труд изследва именно потоци от данни с цел решаване на една основна и особено актуална в момента задача, а именно отчитане на времевата характеристика, анализът и извличане на зависимости от времеви данни с голяма продължителност.

Цел на дисертационния труд

Да се разработи методология за автоматизирано извличане на съществуващи зависимости в потоци от данни, съставени от количествени динамични записи, натрупвани в дълъг период от време, която да се сравни и валидира чрез практически изследвания. Методологията да е приложима към разнообразни източници на количествени данни.

Задачи на дисертационния труд

За постигане на поставената цел се изследват следните задачи:

Задача 1: Обзор на темата за анализ и извличане на закономерности при потоците данни*;*

Задача 2: Разработване на метод за представяне на потоците от данни в нова структура, посредством извличане и анализ на техни характеристики;

Задача 3: Разработване на методология, представена като систематичен подход, за извличане на зависимости от потоци данни*;*

Задача 4: Приложения на предложената методология за анализ и извличане на зависимости в различни бази от динамични данни*:*

- ЕЕГ сигнали*;*
- Времеви редове от UCR колекция данни

Изследването включва разработване на нови методи, както и оригинално комбиниране на известни методи, притежаващи следните качества и отчитащи определени специфики и ограничения:

- Приложимост към данни от различни области и различно естество;
- Приложимост върху потоци от данни (потенциално неограничени);
- Разработените методи да зависят от възможно най-малък брой параметри и с ограничена нужда от специално нагласяване спрямо специфичната предметна област;
- Лесна приложимост за работа в реално време.

Публикации във връзка с дисертационния труд

Резултати от дисертационния труд са публикувани във водещо международно научно списание ***Neural Computing and Applications*** с **импакт фактор** (1.492 за 2015) и в международно научно списание ***International Journal of Reasoning-Based Intelligent*** Друга част от тях са представени като доклади и публикувани в сборниците на две IEEE конференции - ***IEEE Intelligent Systems 2016*** и ***IEEE INISTA 2013***, както и на докторантска конференция, организирана от ФМИ на СУ ”Св. Кл. Охридски“.

1) **Sergey Milanov**, Olga Georgieva, 2016, Pattern Frequency Representation for Time Series Classification, Proceedings of *IEEE 8th International Conference on Intelligent Systems*, 4-6 Sept., Sofia, Bulgaria, pp.478-483.

2) O. Georgieva, **S. Milanov**, P. Georgieva, I. M. Santos, A. T. Pereira and C. F. Silva, 2015, Learning to decode human emotions from event-related potential*, Neural Computing and Applications*, v. 26, issue 3, pp.573-580.

3) Olga Georgieva, **Sergey Milanov**, Petia Georgieva, 2014 Unsupervised EEG biosignal discrimination, *International Journal of Reasoning-based Intelligent Systems,* Vol.6, Nos3/4, pp.118-125.

4) **Milanov S**., O. Georgieva, P. Georgieva, 2013, Comparative Analysis of Brain Data Clustering, Proceedings of *Doctoral Conference in Mathematics, Informatics and Education* [MIE 2013], 19-29 Sept. Sofia, Bulgaria, pp. 94-101.

5) Georgieva O., **S. Milanov**, P. Georgieva, 2013, Cluster Analysis for EEG Biosignal Discrimination. *IEEE International Symposium on INnovations in Intelligent SysTems and Applications INISTA 2013*, 19-21 June, Albena, Bulgaria

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Глава 1. Теоретични основи на извличане на зависимости от потоци от данни

Първа глава изследва и анализира текущото състояние в областите *обработка и анализ на потоци от данни* и *извличане на зависимости от данни*, тясно свързани с целите на дисертационния труд и мотивира необходимостта от създаването на общ подход, обхващащ *извличане на зависимости от потоци от данни*.

1.1. Обработка и анализ на потоци от данни

Задачи на анализ на потоци данни

Класификацията и клъстерният анализ са базови техники в дисциплината извличане на зависимости от данни. Освен самостоятелно, към задачи за групиране и класифициране, тези методи могат да подпомогнат решения за предсказване, откриване на аномалии и тенденции. Затова дисертационният труд фокусира изследванията именно към тях. Прилагат се и се оценяват различни подходи за клъстеризация и класификация, така че да се открият и препоръчат алгоритмите, които са приложими за обработваните данни.

Подходи за обработка и анализ на времеви данни

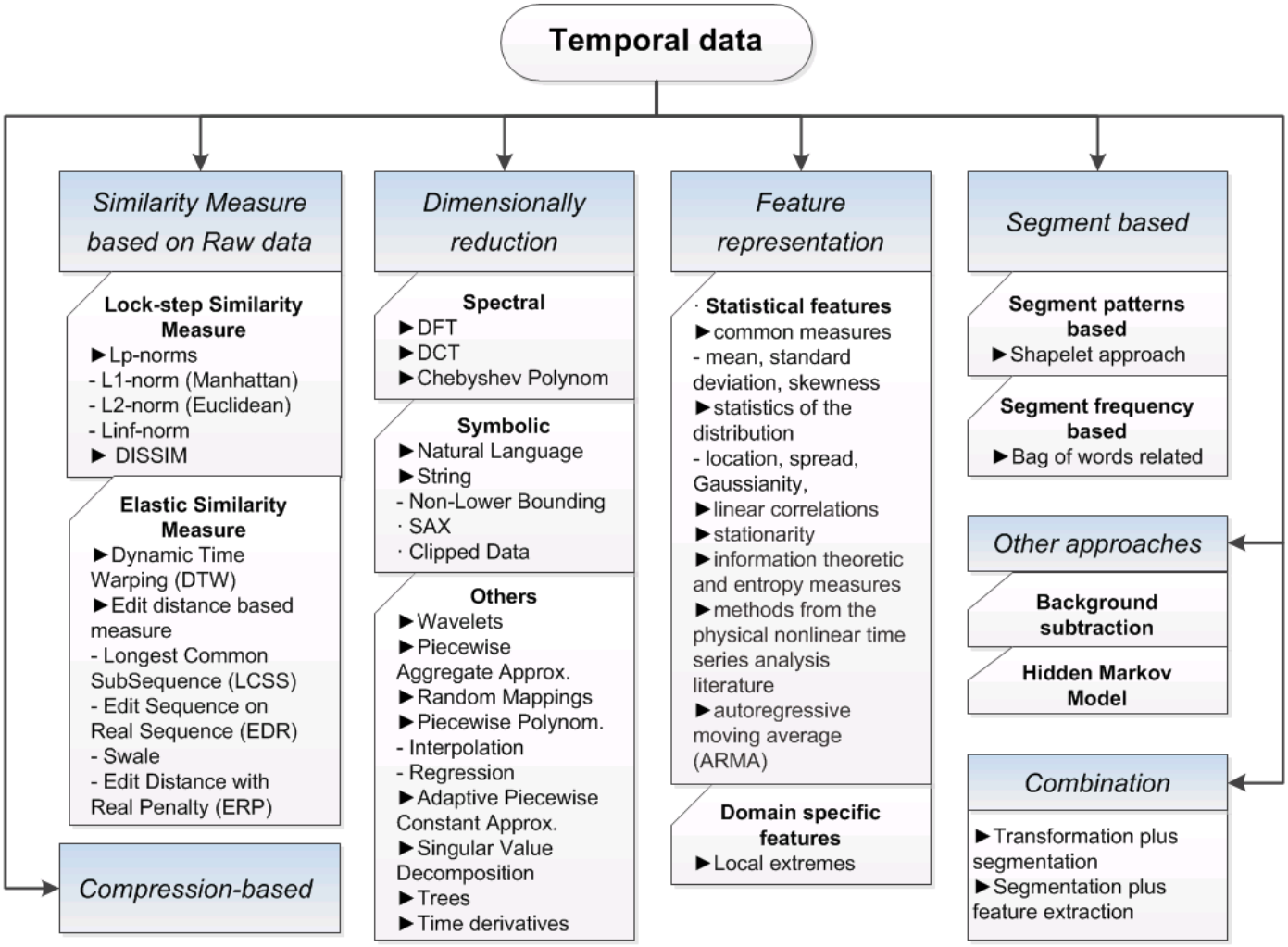
Разработени са различни методи за обработка и анализ на времеви редове, като се предложени разнообразни класификации на методите, групирани по специфични критерии. Съществуващите класификации са непълни, като при тях методите се разграничават в определена степен едностранчиво, по ограничен брой критерии.

В дисертационния труд е представен нов начин за класификация на методите за анализ на времеви данни. Целта е създаване на цялостна класификация, която да разграничава съществуващите подходи задълбочено, с по-голяма степен на детайлност и с описание на йерархични връзки в съществуващите методи:

Структурата на предложеното ново разделяне на методите за обработка и анализ на времеви данни е дискутирана задълбочено в докторантурата, обобщена по-долу и представена на *Фигура 3*.

1. Определяне на сходство между времеви серии чрез *директно сравняване на суровите стойности* на времеви серии по различни алгоритми;
2. Определяне на сходство на времеви данни на базата на *компресия*;
3. *Трансформация* на времевите данни с цел намаляване на размерността на времевите серии с определена степен с цел интензифициране и подобряване на точността на сравнението;
4. Добиване на *глобални характеристики*, чрез които времевите серии да бъдат сравнявани и класифицирани;
5. Изследване на *локални сегменти* и откриване на фрагменти, които най-добре определят времевите данни;
6. *Комбинирани* методи;

7. *Други* методи.



Фигура 3 Методи за анализ на времевите данни (методите са дадени с оригинално въведените английски наименования)

Приложимост на методите за анализ на времеви данни към потоци данни

За разглежданите методи за обработка и анализ на времеви данни са направени следните изводи:

- Директната работа със сурови, необработени времеви данни е неефективна и резултатите се понижават в случаите на дълги потоци данни;

- Подходите за трансформация на времевите данни не са подходящи за много дълги или безкрайни потоци данни, защото дори и намалени, обемите на данните могат да са все още доста големи;

- Методите, базирани на изследване на части от потоците данни, се основават на търсене на характеристики в сегменти от данните. Отчита се локалното поведение на

времевите данни, което може да съдържа съществена информация за последващите изследвания за извличане на зависимости;

- Комбинацията от глобални и локални характеристики за представяне на времевия ред изглежда обещаващ подход за извличане на закономерности в потоците от данни;

- Използване на времеви производни вместо суровите стойности е вероятно да подобри резултатите при динамичните данни;

- Методи, при които се извличат свойства и потоците данни се представят чрез краен набор от съществени характеристики, предоставят добър вид на тяхната глобална структура. Резултатното представяне трябва да запазва важните свойства на времевите данни и по възможност да разкрива нови скрити качества.

Въз основа направените изводи докторантът счита за най-подходящ подходът за представяне на потоците данни във нова форма, подходяща за последващ анализ с цел добиване на зависимости. Необходимо е да се да извлекат фундаментални свойства както на глобално, така и на локално равнище.

Докато глобалните характеристики на времевите данни са изследвани подробно, то подходът за анализ на локални свойства на потоците данни и тяхното комбиниране за създаване на краен брой съществени характеристики, описващи изследваните потоци данни, не е достатъчно проучен.

Такъв подход според извършения анализ е обещаващ, но слабо изследван и има възможност да се развие и задълбочи. Това е и една основна задача на докторантурата - разработване на метод за представяне на потоците от данни в нова структура, посредством извличане на характеристики, които разкриват техни свойства.

1.2. Извличане на зависимости от данни

Интелигентният анализ на данните (*Data mining*) се дефинира като разкриване на неизвестна и потенциално полезна информация, съдържаща се в данните – скрити модели, структури, закономерности, връзки. То се разглежда като етап от цялостния процес на анализиране и откриването на знания в данни (*Knowledge Discovery from Data, KDD*) (***Фигура 6***), който се определя със следните етапи:

1. *Подбор на данни.* Само данните, имащи отношение към задачата за анализ са обект на изследване;

2. *Предварителна обработка.* Филтриране на данните - отстраняване на шума и противоречивите данни;

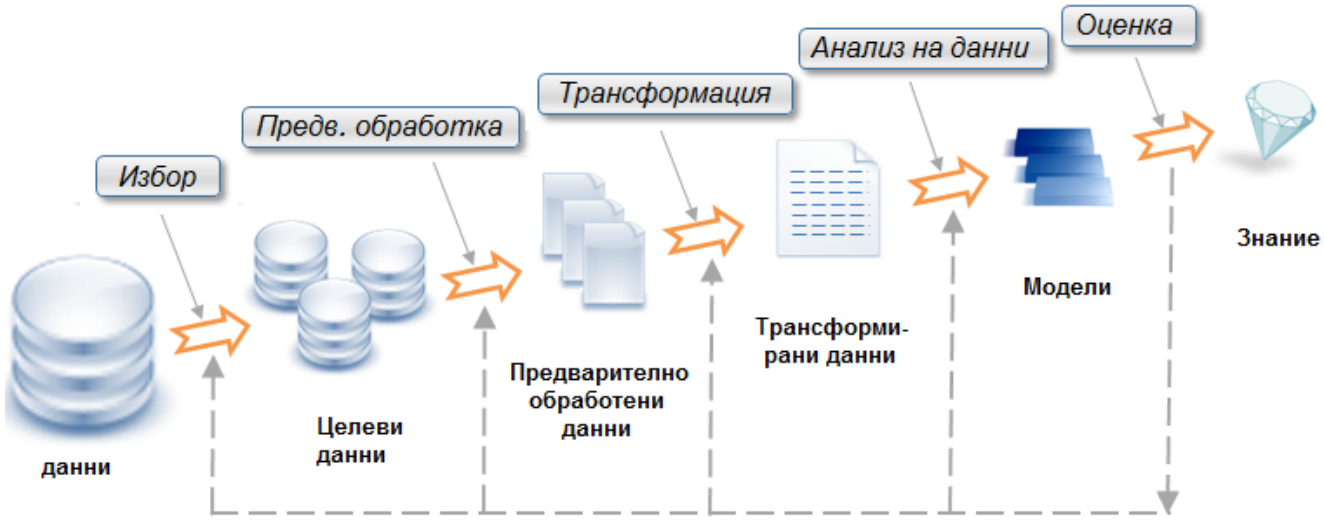
3. *Трансформация на данни*. Данните се трансформират и консолидират във форми, подходящи за извличане на закономерности;

4. *Интелигентен анализ на данните.* Същината на процеса, при който се прилагат специфични алгоритми за извличане на модели и зависимости в данните;

5. *Оценка на резултатите*. Тълкуване на резултатите, идентифициране на наличие на интересни модели, закономерности, знание;

6. *Представяне на придобитото знание.* Използват се техники за визуализация и представяне на знания;

7. *Оптимизация*. Търсене в пространството от решенията. Параметрите на моделите се калибрират, стремейки се към оптимални стойности.



Фигура 6. KDD процес на разкриване на знания

KDD процеса задава обща рамка, като съществуват разнообразни решения в зависимост от избора и прилагането на конкретни процедури и алгоритми в отделните етапи.

Най-важният етап в процеса е *Интелигентен анализ на данните* - аналитична стъпка в цялостния *KDD* процес. Правилната и основателна *оценка на постигнатите резултатите* е неизменна етап, необходим за определяне на стойността и успеха на извършените задачи. Другите стъпки, като подготовка на данни, избор на данни, почистване на данни, вграждане на подходящи предварителни знания, визуализация и тълкуване на получените резултати, макар и допълващи в *KDD* процеса, са от важно значение, за да се гарантира, че полезно знание се извлича от данните.

В докторантурата задълбочено и систематично са проучени най-съществените етапи в *KDD* процеса, а именно *интелигентен анализ на данните* (конкретно за задачите за клъстерен анализ и класификация) и *оценка на резултатите*. Етап *трансформация на данни* в контекста на анализа на времевите данни включва представяне им в подходящ вид, което подробно е изследвано в дисертацията. Предвид възможността за представяне на времевите данни чрез голям брой характеристики е разгледана и задачата за *избор и управление на характеристики* – смисъл и ползи, оценка и търсене на подмножество най-значими характеристики.

1.3. Обобщение и изводи

В настоящата глава са изследвани и анализирани проблемните области, тясно свързани с целта и задачите на настоящия дисертационен труд – *обработка и анализ на времеви потоци данни* и *извличане на зависимости от данни*. Въз основа на това могат да бъдат направени някои важни заключения.

Анализ на времеви потоци данни

По отношение на **обработката и анализа на времеви потоци** са постигнати следните резултати:

- Разгледани са спецификите на времевите данни и типичните задачи, приложими над тях;
- Описани са различни подходи за обработка и анализ на времеви данни с разглеждане на предимствата и недостатъците им;
- Разгледана е целесъобразността от прилагане на различните методи за изследване на времеви данни над потоци от данни.

Като **принос** на дисертационния труд е предложена нова, цялостна класификация на методите за анализ на времеви данни. Тази класификация е по-задълбочена и с по-голяма степен на детайлност от съществуващите, основава се на повече критерии и включва и йерархични връзки в предложените групи.

На база на извършените анализи, аргументирано е предложено допълнително изследване за създаване на нов метод, основан на представяне на потоците данни посредством краен брой характеристики, получени на основата на локални свойства на потоците данни.

Извличане на зависимости от данни

Разгледан е систематичен подход за извличане на зависимости от данни (*Knowledge Discovery from Data)*. Той представлява структуриран процес от отделните етапи за извличане на зависимости от данни.

Разяснени са отделни стъпки от *KDD* процеса на анализиране и откриване на знания в данни. Задълбочено са разгледани основните, необходими етапи в *KDD* процеса:

- *Избор и управление на характеристики* – предназначение и ползи от избор на съществени характеристики и разяснение на основните компоненти - *оценка на подмножества от характеристики* и *стратегии за тяхното търсене*;

- *Методи за интелигентен анализ на данните* - подходи и методи за клъстерен анализ и класификация;

- *Оценка на резултата* - критерии за определяне на стойността от извършените задачи за класификация; оценка на валидността на клъстеризация при използване на външни и вътрешни критерии.

Интегриране на извличане на зависимости от данни с времеви потоци данни

По отношение на извличане на зависимости в потоци от данни са направени следните изводи:

- В областта на *обработка и анализ на времевите данни*, като самостоятелна тема, съществуват значителен брой изследвания в литературата. Недостатък е, че

времевите специфики на данните се проучват изолирано, а не като част от цялостен процес;

- *Извличане на зависимости от данни* също е широко изследвана тема. Налице са и значителен брой проучвания в областта на *извличане на зависимости от данни* и тяхното структуриране в цялостен процес. Тези изследвания се отнасят предимно към статични данни, като в голяма степен се игнорира времевата характеристика на данните, докато повечето реални данни се променят с времето;

- Съществуващите методи за анализ и извличане на зависимости от статични данни не винаги може да бъдат директно приложени върху времевите данни - времевата структура и значителното количеството при потоците данни трябва да се вземат предвид;

- Забелязва се недостиг на задълбочени изследвания в литературата за *извличане на зависимости от данни*, приложени върху *потоци данни* и отчитайки времевите им специфики и свойства.

Като обобщение, резултатите от литературния обзор потвърждават актуалността и значимостта на разглежданите проблемни области и подчертават необходимостта от създаването на систематичен подход за извличане на зависимости от потоци от данни.

Въз основа на направените изводи, докторантът смята за подходящо да се изследва и да се предложи нов систематичен подход, който съчетава подходящи методи за *обработка и анализ на потоци данни* и тяхното използване като стъпка от цялостния процес на *извличане на зависимости от данни*.

Глава 2. Извличане на зависимости от потоци данни

Настоящата глава предлага нова методология, разглеждана като цялостна систематизирана процедура, за изпълнение на процеса на *извличане на зависимости от потоци данни*. Подробно е разгледан първият етап от методологията, свързан с анализ и обработка на времевите свойства на потоците от данни, като е разработен нов метод за *представяне на потоците данни в нова структура*, основан на локални свойства на времевите данни.

2.1. Представяне на потоците данни в нова структура, чрез извличане на характеристики, разкриващи техни свойства

Разработен е нов подход за представяне на потоците от данни в структура от по-високо ниво посредством краен брой характеристики, описващи важни техни свойства. За целта се изследват локални свойства на потоците данни с прилагане на методи за добиване на техни свойства. Локалните свойства се комбинират и в резултат от този етап, времевите редове се представят чрез краен брой характеристики. Разглежда се и важната задача за подходящо измерване на сходство между новите описания на потоците данни.

Методите за извличане на закономерности се прилагат върху потоците данни представени в новата структура, чрез характеристичния им вектор.

Метод за представяне на времеви данни чрез честота на срещане на прототипи (ЧСП)

Разработен е нов метод представя времевите редове посредством *честота на срещане на прототипи* (*ЧСП*). За целта времевите редове се разделят на припокриващи се локални сегменти. Намирането на ограничен брой прототипи на сходни сегменти е основна стъпка в процеса. Всеки сегмент се представлява от своя прототип, изчисляват се срещанията на различните прототипи и накрая времевите редове или потоци данни се представят чрез честотата на появяване на отделните прототипи.

Съществена черта на метода е, че при изчисляването на сходство между сегменти не се работи директно с оригиналните (сурови) стойностите на сегментите, а се използват различни свойства на сегментите.

Изборът на такъв модел има предимствата да позволява:

- *Анализ на вида на прототипите*. Намерените прототипи описват времевата серия и логично може да се използват при последващия анализ на времевите серии и извличане на зависимости от тях;

- *Отчитане на честотата на срещане на прототипите*. Срещане на даден прототип рядко в дадена времева серия и често в друга е показател за разлика между самите времеви серии и процесите, които ги пораждат. И аналогично – времеви серии,

които съдържат подобни прототипи със сходна честота е по-вероятно да са от подобен вид и генерирани от подобни процеси.

Методът позволява прилагане в различни вариации, основно касаейки начина за сравнение на сегменти и подходът за определяне на прототипи.

Стъпки на метода за представяне на потоци данни чрез ЧСП

Четирите стъпки на представяне на времевите редове чрез честота на сещане на прототипи са изброени по-долу, обобщени в блок схема на *Фигура 11* и илюстрирани на *Фигура 12*:

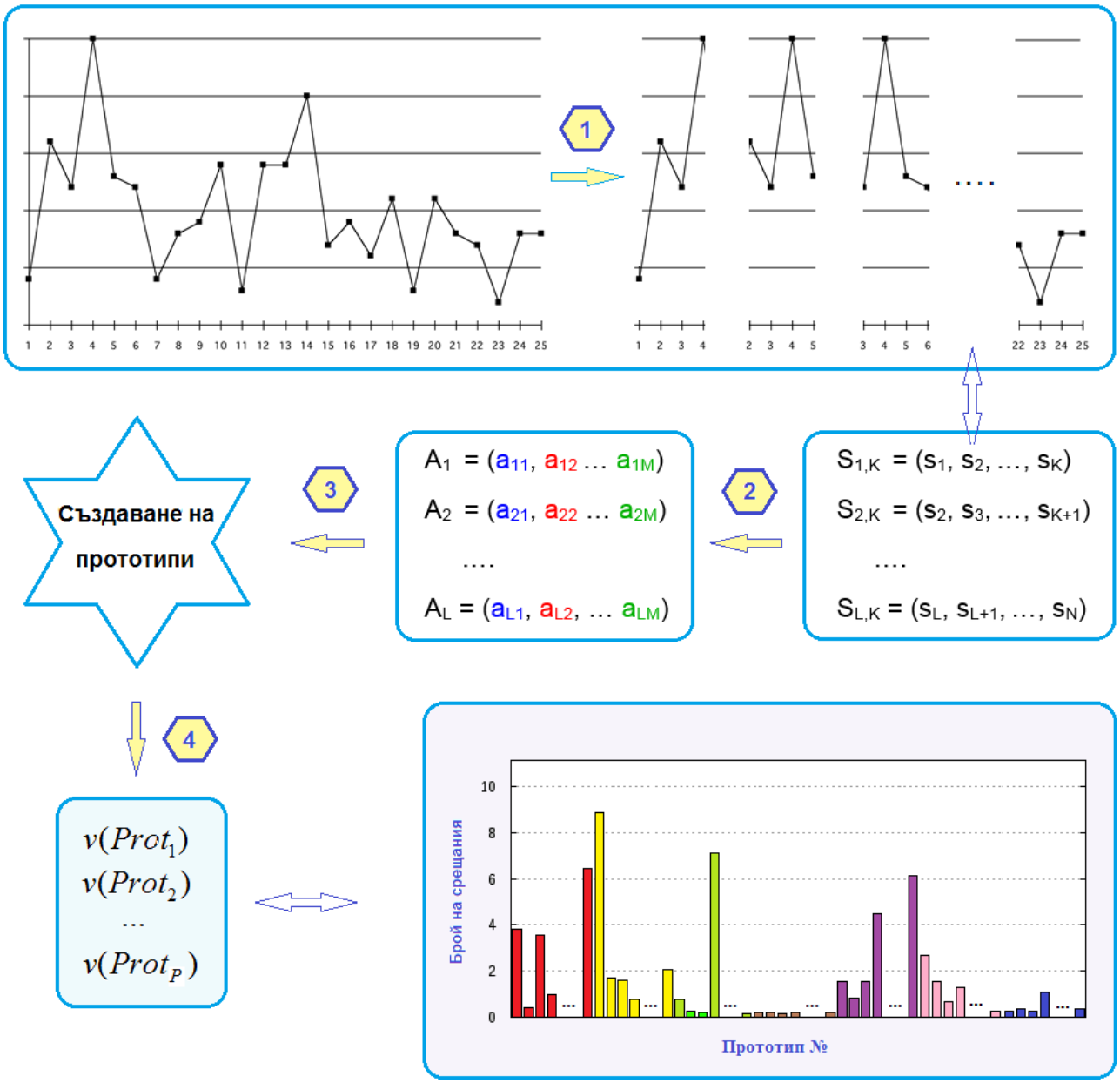
Стъпка 1. *Сегментация на потоци данни*;

Стъпка 2. *Изразяване на сегментите чрез атрибути и определяне на сходство между тях*;

Стъпка 3. *Определяне на прототипи от подобни сегменти*;

Стъпка 4. *Представяне на потоците данни чрез честотите на срещане на прототипите*.

Фигура 11. Блок-схема на представяне на времевите редове чрез ЧСП



Фигура 12. Стъпки от представянето на времевите редове чрез ЧСП

2.2. Методология за извличане на зависимости от потоци данни

В дисертационния труд се поставя задачата за разработване на подходяща, нова методология за извличане на зависимости от потоци данни. Методологията се основава на отделни етапи, обобщени в процеса на откриване на знания в данни (*Knowledge Discovery from Data* - Фигура), при който изборът и прилагането на конкретни процедури и алгоритми в отделните етапи на *KDD* процеса създава разнообразни решения.

В докторантурата се конкретизират стъпките от този процес и оригинално се специфицират за целите на анализ на потоци от данни. В някои от стъпките - на анализ и трансформация на данни, се предлагат нови решения и алгоритми.

Етапи в методологията за извличане на зависимости от потоци данни

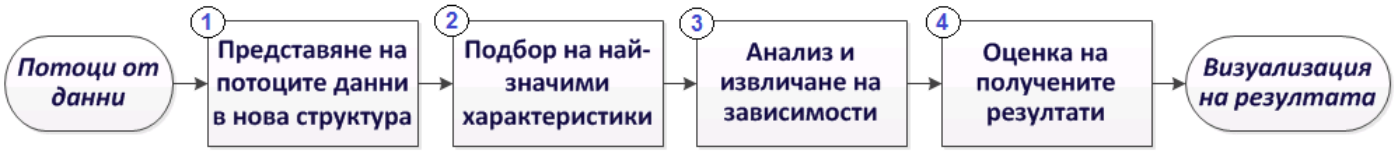
Предложената методология за извличане на зависимости от потоци данни е съставена от четири основни етапа, обобщени в блок схема на *Фигура 17*, визуализирани на *Фигура 18* и описани систематизирано както следва:

Етап 1. *Представяне на потоците данни в нова структура*. На базата на направените проучвания е предложен метод за трансформиране на потоците данни в нова форма, която запазва и подчертава основни свойства на наблюдаваните динамични обекти. Новият компактен вид на времевите редове подпомага последващите задачи за извличане на зависимости от тях като подобрява качеството и ефективността на използваните процедури.

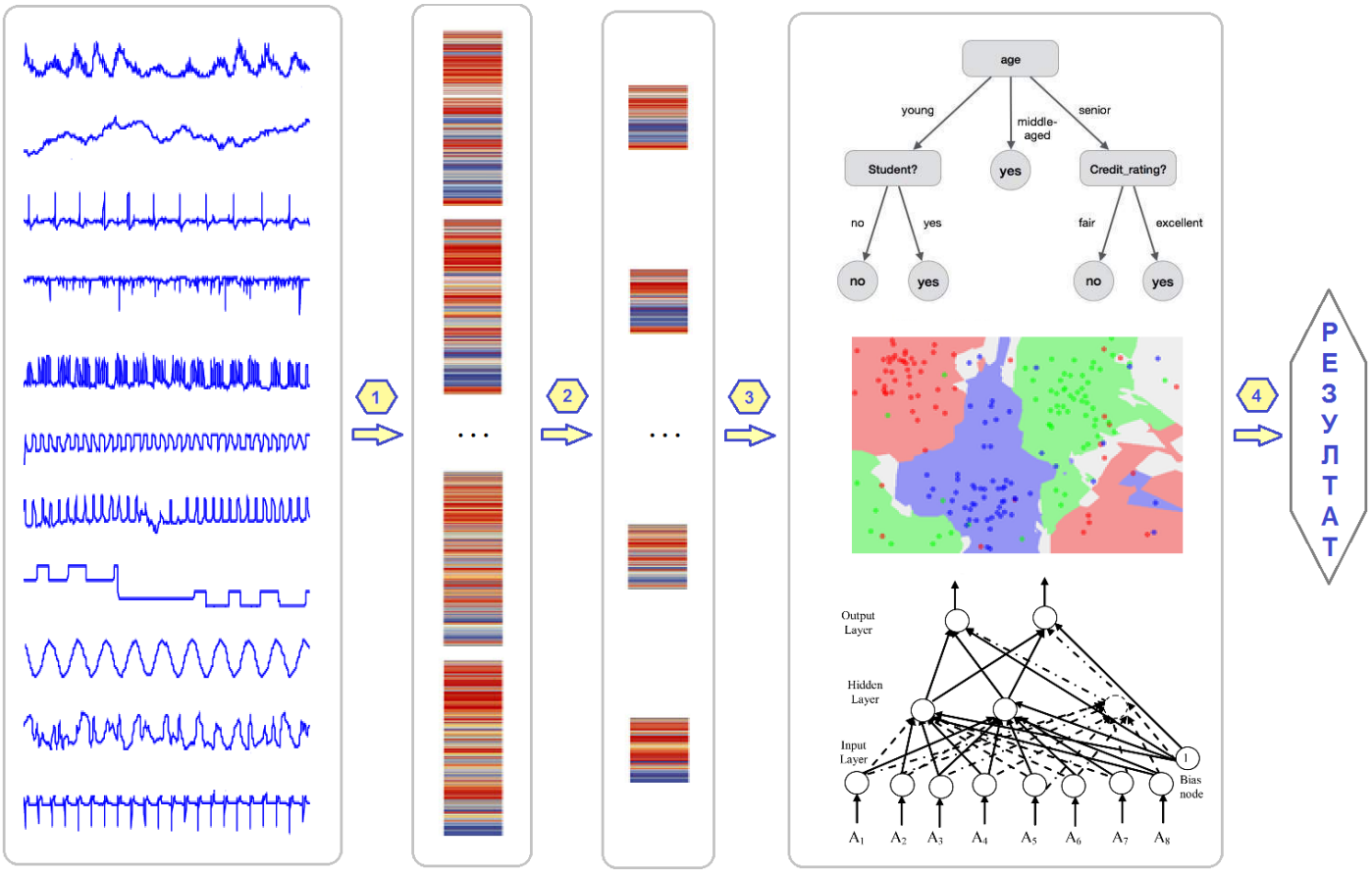
Етап 2. *Подбор на най-значими характеристики*. Когато генерираните в предишната стъпка характеристики са значителен брой е препоръчително да се включат процедури за избор на най-значимите от тях.

Етап 3. *Прилагане на методи за анализ и извличане на зависимости*. На базата на избраните характеристики се дефинира пространство от данни, над което се изпълняват различни алгоритми за клъстерен анализ и класификация.

Етап 4. *Оценка на изпълнението и получените резултати*. Резултатите от приложените методи за класификация и клъстерен анализ на потоци от данни се оценяват, използвайки вътрешни и външни валидационни измерители.



Фигура 17. Блок схема на методология за извличане на зависимости от ПД



Фигура 18. Методология за извличане на зависимости от потоци данни

Предложената методология представя решение за анализ на потоци от динамични данни и притежава следните качества, които я правят широко и лесно приложима и гарантират нейното автоматизирано изпълнение:

- *Приложимост над данни от различни области, от различно естество.* Предложената процедура не изисква и не използва знания за контекста и обкръжението, за които се отнасят времевите редове и потоци данни.

  Методологията не зависи от предметната област и така е приложима над данни от различни области на интерес

- С *възможно най-малко параметри*. Стремежът е използване на възможно най-малко параметри, което е постигнато чрез намаляване на използваните свободни променливи. По този начин се избягва необходимостта от специалното настройване на параметрите с подходящи стойности, множество сравнения за различни комбинации и свързаните усложнения и неточности.

Методологията зависи от малко параметри, което опростява инициализацията и настройката на цялостната процедура

- *Приложими за работа в реално време*. Използване на методи, които имат ограничени изисквания откъм изчислителни ресурси. Предвид обработката на потоци, които може да са значителни като обем, е възприето представяне в компресиран вид, което помага за ограничаване на капацитета при съхранение. Бързодействието и ефективността са важни характеристики и приложими за онлайн работа.

Методологията отчита значителния обем и продължителност на данните

- *Единствено преминаване на изпълнение (one pass)*. Предвид възможната голяма продължителност при потоците данни, стремежът е за единична обработка на постъпващите данни. Потоците данни не се запазват в суров вид, а се обработват и представят в ограничен брой характеристики.

Предимство на подход, базиран на изложените характеристики е, че той е автоматизиран. Пълна автоматизация е предизвикателство задача, но стремежът за намаляване на човешкия фактор е постижима крачка в тази посока. Това подчертава значението на алгоритми, които са независими от типа на задачи и контекста, приложими към различни данни, в голяма степен автономни, изчислително ефективни и са в състояние да откриват и закономерности в данните.

2.3. Обобщение и изводи

Разработен е нов метод за *представяне на потоците данни в нова структура, посредством честота на срещане на прототипи (ЧСП)*, удобна за целите на последващо извличане на зависимости. Методът е основан на изследване на локални свойства на времевите данни и е съставен от 4 стъпки – *Сегментация на времевите редове чрез способа на плъзгащия прозорец*; *Изразяване на сегмент чрез множество атрибути с цел тяхното сравняване*; *Определяне на ограничен брой прототипи от сегменти* и накрая - *Представяне на потоците данни чрез честотите на срещане на прототипите*.

Създаденият метод за представяне на потоците данни в нова структура притежава следните качества:

- Приложим е върху потоци от данни (потенциално неограничени), чрез представянето им чрез краен брой съществени характеристики;
- Използва се по унифициран начин към бази времеви данни от различни области на приложение и не се нуждае от специфични настройки;
- Има ограничен брой параметри, което спомага и за лесно вграждане в цялостната методология за извличане на зависимости от потоци данни;
- Приложим е за работа в реално време.

По отношение на разработения метод за представяне на потоци данни чрез ЧСП съществуват различни реализации, в зависимост от начина на изразяване на сегмент чрез множество атрибути и процедурата за определяне на прототипите на сегментите. Един вариант на метода е разработен в четвърта глава и приложен над избрани бази от времеви данни.

Предложена е цялостна методология като *систематичен подход за извличане на зависимости от потоци данни*, оригинално специфицирана за целите на анализ на потоци от данни. Подробно са разгледани отделните етапи, следващи представянето на потоците данни в обобщен вид, а именно - *избор на характеристики*, *клъстеризационни и класификационни методи за анализ и извличане на зависимости от данните* и *оценка на резултатите*. Описани са ефективни алгоритми за всеки етап, които са използвани в трета и четвърта глава при изследване на приложения на предложената методология.

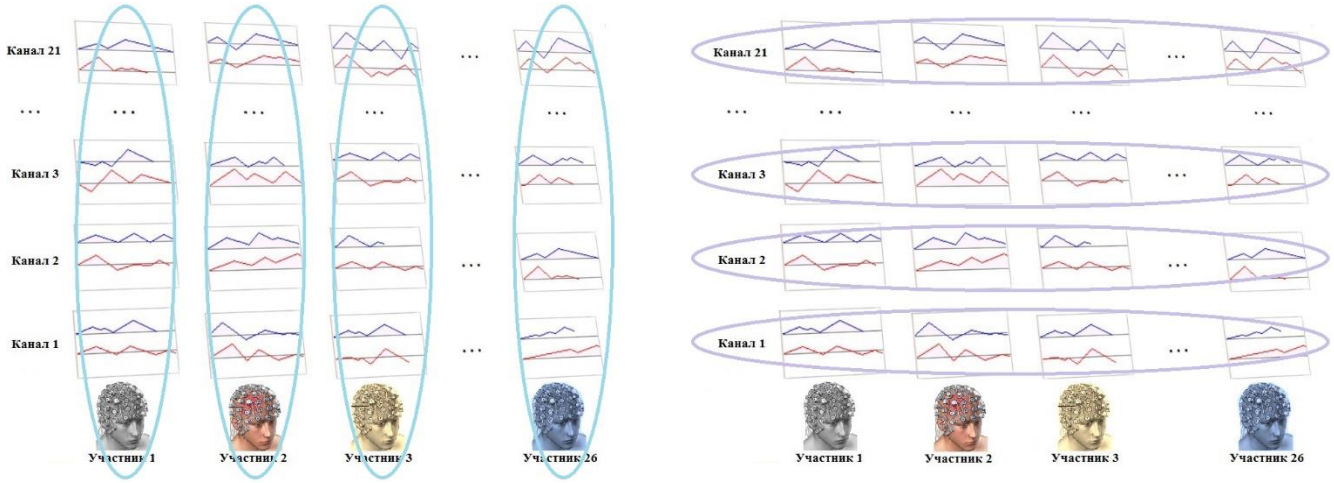
Практическото приложение на методологията е способността ѝ за решаване на особено актуалната в момента задача за отчитане на времевата характеристика и анализът на потоците данни с голяма продължителност.

Глава 3. Извличане на зависимости от ЕЕГ потоци данни

Изследвани са визуално подбудените био-сигнали, записани при наблюдение на картини с високо положително и отрицателно емоционално въздействие. Данните са регистрирани при показване на две картини (по една от всеки вид) на 26 доброволци. Получени са общо 52 потока от данни. От потока данни, регистриран за всеки ЕЕК канал, са определени 12 характеристики - амплитудите и времето на настъпване на първите шест екстремума.

Структуриране на ЕЕГ потоци от данни

Всеки опит се състои от записите на 21 ЕЕГ потоци данни, съответстващи на наблюдаваните канали. Всеки ЕЕГ поток данни от своя страна е представен посредством 12 характеристики, извлечени по описания алгоритъм на поредните екстремуми. Като резултат се получава масив от данни, който може да се разглежда в четири размерности: 26 Доброволци × 2 картини × 21 Канала × 12 Характеристики. Това дава възможност данните от ЕЕГ записите да бъдат комбинирани по различен начин, с цел в тях да бъдат търсени различни зависимости (*Фигура 25*)



Фигура 15. Групиране (а) по участници (Intra-subject); (б) по канали (Inter-subject)

Обща база данни. Една единствена база данни съдържа всички данни, групирани по отделно наблюдение. Тя се състои от 52 отделни случая - 26 участници x 2 наблюдения. Всеки случай агрегира показателите за всичките 21 канала за съответното наблюдение. *Задача*: Анализът на тази база данни може да даде информация за цялостния отговор на мозъка при визуално подбудена реакция на различно емоционално натоварени картини.

Формиране на бази данни по участници (*Intra-subject*). Наблюденията за всеки отделен участник образуват независима база данни. Това разделя данните в 26 бази данни - по един за данните за всеки от участниците (*Фигура 25 (а)*). Всяка база данни е образувана от 42 отделни случая с данни за отделен индивид (21 канала X 2 наблюдения). *Задача*: Чрез така формираните бази от данни могат да се оценяват реакциите на отделните индивиди на база на сравнителен анализ между тях при наличие на визуална задача от описания вид.

Формиране на бази данни по канали (*Inter-subject*). Данните се групират по всеки канал, при което се образуват 21 отделни бази данни - по един за всеки канал (*Фигура 25 (б)*). Всяка база данни съдържа 52 отделни случая – 26 участника по 2 наблюдения (положителен и отрицателен стимул) за съответния канал. *Задача*: Тази база от данни позволява да бъдат разкрити зависимости в отделните канали на мозъка като се направят изводи кои от тях са отговорни при наличие на визуална задача от описания вид.

3.1. Приложение на методологията към ЕЕГ данни

Към ЕЕГ био-сигналите е приложена методологията за извличане на зависимости от потоци данни, като отделните етапи са обяснени в следващите подраздели. Потоците данни са представени чрез 12 характеристики, което редуцира първа стъпка *Представяне на потоците данни в нова структура* от методологията.

За целите на това изследване са приложени алгоритми за извличане на зависимости, получени както чрез клъстеризация, така и класификация. Прилагат се различни техники за извличане на данни с цел идентифициране на най-подходящите методи и алгоритми.

- Клъстерният анализ е мощен метод за изследване на данни със сложна и неясна структура каквито са ЕЕГ сигналите. Целта при изследването е да се проучат възможностите за разпознаване, дискриминация и групиране на визуално подбудени ЕЕГ био-сигнали.

- Потенциалът на процедурите за класификация се изследва като приложение на ЕЕГ потоци от данни.

Търсят се секции от мозъка с най-добра дискриминативност на валентността (знака) на визуално предизвиканата емоция.

Представяне на потоците данни чрез краен брой характеристики

Методът за извличане на характеристики от потоците данни следва изложения подход, приложен в контекста на био-сигнали, а именно - ЕЕГ сигнала по всеки канал се представя посредством 12 характеристики, извлечени по алгоритъма на поредните екстремуми. Характеристики с номера от 1 до 6 съответстват на амплитуди (A_{min1}, A_{max1},

A_{min2}, A_{max2}, A_{min3}, A_{max3}), а номера от 7 до 12 – на латентността на поредните екстремуми, наблюдавани при био-сигналите (L_{min1}, L_{max1}, L_{min2}, L_{max2}, L_{min3}, L_{max3}).

- При общото групиране, всеки обект се представя чрез 256 характеристики – обединения на всички характеристики по всички канали

$$[(A^1_{\min 1}, A^1_{\max 1},...L^1_{\max 3}),(A^2_{\min 1}, A^2_{\max 1},...L^2_{\max 3}),.....(A^{21}_{\min 1}, A^{21}_{\max 1},...L^{21}_{\max 3})]$$

- При групиране по участници (*Intra-subject*) и групиране по канали (*Inter-subject*) всеки обект се представя чрез тези 12 характеристики и класова принадлежност.

$$(A_{min1}, A_{max1}, A_{min2}, A_{max2}, A_{min3}, A_{max3}, L_{min1}, L_{max1}, L_{min2}, L_{max2}, L_{min3}, L_{max3})$$

Подбор на най-значими характеристики

В това проучване са приложени 4 метода за избор на характеристики върху формираните бази данни - три филтърни (*Pearson product-moment correlation coefficient*, *Gain Ratio* и *Relief*) и един обхващащ метод (*Feature selection with Support Vector Machines*). В експерименталните изследвания също е включена и оценка на резултатите без извършване на стъпката за избор на характеристики с намерение да се разкрият ползите от включването на този етап в цялостната процедура.

По отношение на вида използваните алгоритми за избор на характеристики, филтърните методи са преобладаващи, тъй като те разчитат единствено на свойства на данните и не са зависими от последващите модели и алгоритми за извличане на зависимости от тях.

Анализ и намиране на зависимости в данните

Класове и клъстери. Всеки от разглежданите ЕЕГ потоци данни принадлежи на един от двата класа наблюдения, свързани с положителните и отрицателните емоционални въздействия, наричани съответно *Стимул_P* (положителна валентност) и *Стимул_N* (отрицателна валентност). Затова и в използваните алгоритми за клъстеризация и класификация се търси разделяне в два клъстера/класа, които да отговарят на двата вида стимули. Клъстерът, който има повече случаи от *Стимул_P се* означава като *Клъстер_P*. *Клъстер_N* съответно обхваща преобладаващо наблюдения от *Стимул_N*.

Оценка на модела

Различните методи на клъстеризация се оценяват по отношение на тяхната надеждност правилно да разграничават двата класа, съответно свързани с положителни и отрицателни емоции. Известната класова принадлежност на ЕЕГ биосигналите се използва като външен измерител.

Оценката на двата резултатни клъстера *Клъстер_P* и *Клъстер_N* се изчислява като процента на обектите от съответния свързан клас (*Стимул_P* и *Стимул_N*), които са определени правилно към клъстера.

3.2. Извършени експерименти и получени резултати

Приложимостта на методологията се оценява по отношение на различните формирани бази данни – обединени по индивиди (*intra-subject*), групирани по ЕЕГ канал (*inter-subject*) и обща база данни по експеримент. Изследванията включват различните етапи от методологията – избор на характеристики, извличане на закономерности, оценка на резултата, търсене на най-подходящи решения.

Извършени са множество експериментите в следните групи:

- Избор на характеристики
  - Бази данни обединени по участник (*intra-subject*)
  - Бази данни обединени по канал (*inter-subject*)
  - Обща база данни по експеримент
- Клъстеризация *Fuzzy-C-Means* (*FCM*)
  - *FCM* клъстеризация на бази данни, обединени по индивиди (intra-subject)
  - *FCM* клъстеризация по ЕЕГ канали (inter-subject)
- Клъстеризация (*HC*, *EM*, *KM*, *ZM* и *FF*) след избор на характеристики (*Pearson Correl. Coefficient* и *Gain Ratio*)
- Класификация на ЕЕГ сигнали

3.3. Обобщение и изводи

Изследвани са подходи за оценка на различни състояния на активността на мозъка посредством анализ на ЕЕГ данни, получени от група доброволци при наблюдение на картини с високо положително и отрицателно емоционално въздействие.

Първоначално в настоящата глава е разяснена същността на набора от данни от електроенцефалограми (ЕЕГ). Разяснен е методът за тяхното представяне чрез краен брой характеристики (амплитуда и латентност на първите екстремуми) и са описани подробно три начина на структуриране ЕЕГ данните – бази данни основани на *групиране по участници*, *групиране по канали* и *общо групиране* за целите на извличане на различни зависимости от тези данни.

Методологията за извличане на зависимости от потоци данни е приложена над различните формирани бази данни по отношение на задачите за клъстеризация и класификация. Извършени са множество разнообразни експерименти - *избор на характеристики* за всички видове групирания; *FCM клъстеризация* за *групиране по участници и канали*; прилагане на множество клъстеризационни методи в комбинация с избор на характеристики; *класификация* в комбинация с избор на характеристики при *общо групиране*.

За всеки експеримент е извършен анализ на резултатите от дискриминацията на положителен и отрицателен стимул. Въз основа на това са направени някои важни заключения по отношение на:

- *Локализиране на каналите, осигуряващи по-висока клъстеризационна точност*.

  Данните са много добре разграничими по канал 6 (*Fz*), което отговаря на неговата връзка с възприемането на емоции. Други канали с добра оценка са 10 (*C3*) и 11 (*Cz*), които се свързват с визуалното възприятие.

- *Анализира се значението на пространственото разположение на съответните зони*

  Централните канали (в частност *Fz*, *Cz* и *Pz*) осигуряват по-висока клъстеризационна точност, което е в съответствие с констатациите, че визуалните стимули са свързани с по-силна реакция в централната кортикална линия [28].

- *Оценка на представянето при положителни и отрицателни стимули*.

Методологията е в по-голяма степен подходяща за дискриминиране на състояние *P*

- *Анализ по отношение на алгоритмите за избор на характеристики, клъстеризация и класификация*.
  - Алгоритъм SVM превъзхожда останалите алгоритми за избор на характеристики при последваща класификация;
  - Клъстеризационните алгоритми, следващи подхода за разпределяща клъстеризация с оптимизация на целева функция, демонстрират най-добра оценка на групиране;
  - Алгоритми *Multilayer Perceptron* и *SMO* демонстрират най-добри резултати при задачите за класификация.

Оценката осигурява теоретична основа за автоматизиране на прогноза на човешкото състояние и съдържащите се експериментални доказателства. Изследванията може да се считат като първоначален опит за разкриване на общи представителни модели на емоционални състояния върху множество индивиди. Такива изградени модели могат да бъдат използвани като софтуерни сензори за откриване на емоционалните състояния на субекти извън участниците в експеримента. Това знание е от значение и може да бъде използвано при следните случаи:

- добиване на допълнителна информация за състоянието на изследвания човек, за да се предложат навременни решения или план за действие;
- за диагностициране на ментални процеси при пациенти с мозъчни увреждания;
- към задачата за категоризация на човешката личност (като например дискриминация между хора с високо и хора с ниско ниво на невротизъм).

Изследвания и приложения на методологията върху ЕЕГ потоци данни са представени теоретично, практически приложени и публикувани в статии с участие на докторанта [96]*,* [27]*,* [28] *и* [46]*.*

Глава 4. Извличане на зависимости от UCR колекция от времеви данни

Предложената в докторантурата методология за извличане на зависимости в потоци данни и разработеният нов метод за представяне на времевите данни чрез честота на срещане на прототипите (ЧСП) са резултат от анализ на съществуващите подходи, обосноваване на изисквания при потоците данни, разработване и оригинално комбиниране на различни методи. Те представят обещаващ подход, който е необходимо да се съпостави с други методи с цел експериментално валидиране на неговата приложимост. За пълноценно сравнение се изисква прилагане върху известни, публични бази времеви данни със значителен брой изследвания над тях.

За изследване е избрана колекция от бази времеви данни на Университетът в Калифорния (*The University of California, Riverside UCR*), която претендира да е най-голямата в света, включващи принадлежност към даден клас, като различните бази данни съдържат реални данни от изключително разнообразни области на приложение.

Многообразието на UCR данните ги прави подходящи за проверка и валидиране на целта на дисертационния труд за:

- Разработване на методология за извличане на зависимости в потоци от данни, наблюдавани и регистрирани от най-различни области на заобикалящата ни среда;

- Приложение на методологията към множество различни UCR бази данни по унифициран начин, без нагласяване към специфики на различните бази данни;

- Методът за представяне на потоците данни чрез ЧСП автоматично се нагласява към изследваните данни.

4.1. Представяне на UCR времевите данни чрез честота на срещане на прототипи (ЧСП)

Времевите редове от *UCR* базите данни достигат в суров вид като стойности на наблюдавана променлива във времето. Поради това методологията, представена в *Раздел 2.2* се прилага в цялостност, като последователност от всички описани етапи. Първият етап е представяне не потоците от данни посредством краен брой извлечени характеристики, подлежащи на последващ анализ.

UCR времевите данни са представени по новия метод, разработен в дисертацията, чрез честота на срещане на прототипи (*ЧСП*). Следва описание на реализацията на *ЧСП* метода, използвана в изследването на UCR колекция от времеви данни.

Сегментация на времевите редове

Използва се плъзгащ прозорец с дължина K=*4* като се дискутират основанията за този избор.

Изразяване на сегменти чрез множество атрибути

В изследването за атрибути на сегментите се използват *времеви производни*, изчислени като разлика между последователните стойности. Подходът предполага сходство основано на формата на сегментите, тъй като сегменти с подобна форма произвеждат подобни характеризиращи вектори. Методът е и инвариантен спрямо отдалечеността на сегмента по Y координатата. Методът е интуитивен и разбираем, с ниски изчислителни разходи.

Метод за определяне на прототипи

Избран е ефективният решетъчен метод (*grid-based*) за създаване на прототипи, което би позволило приложение в реално време. Изгражда се многомерна мрежа от клетки над пространството на данните, като всички операции за клъстеризация се извършват над формираната решетъчна структура. Изчисляване на броя на информационните обекти, попадащи във всяка клетка, е измерител на плътност. Клъстеризация, базирана на решетка обикновено е независима от броя на обектите на данни, а зависи само от разделянето по всяка размерност, което води до предимство в краткото време за обработка. Всички сегменти от всички потоците данни с изследват обединение и се изследват общо с цел откриване на общи прототипи за всички от тях. Решетъчният метод използва *категоризация на характеристики*

Избор на брой атрибути и категории. Общият брой клетки или прототипи е $c_1 \times c_2 . \times \ldots \times c_M$, където c_i е *броят на категориите*, на които е разделен всеки атрибут. Този брой участва като множител в резултата. Всеки атрибут създава ново измерение в пространството на характеристиките и така *броят на атрибутите* участват като степен в общия брой прототипи.

Брой атрибути. Предвид стремежа броят на прототипите да е малък, то решетъчния подход е логично да се използва при малко на брой атрибути, всеки от тях разделен на ограничен брой категории. Именно затова, с цел ограничаване на изчислителната сложност са избрани 3 атрибута за представяне на сегмент.

Брой категории. Броят на категориите, на които се разделя всеки атрибут е друг параметър при метода на решетка (*grid*).

За избягване на нуждата от задаване на параметър *брой категории* е предложен нов алгоритъм, при който не се задава изрична стойност за брой на категориите. Това е основно предимство като се има предвид идеята за използване на минимален брой параметри в цялостната процедура по извличане на зависимости.

Изчисляване на честотата на срещане на прототипите

Следващият етап е построяване на честота на срещане на прототипите. Като краен резултат, времевите редове са представени посредством краен брой целочислени стойности – срещания на съответните прототипи, което важи и за потоци от данни от порядъка на хиляди и милиони стойности. Откъм гледна точка на ефективност, операциите в процеса на представяне на времевите данни в ЧСП са ефективни, бързи и ресурсо-икономични.

4.2. Експериментален анализ

Изследвано е приложение на предложената методология за класификация на двадесетте описани бази данни от времеви данни от колекцията на *UCR*. Всички времеви данни са представени посредством честота на срещане на прототипи (*ЧСП*).

Приложение на методологията към UCR бази данни

Всички времеви данни се представят в *ЧСП* формат по трите метода за категоризация (*равна ширина*, *честота* и *статистически измерител средната стойност на използваните атрибути*).

За всяка база времеви данни, представен чрез *ЧСП,* са обучени три класификатора по методите на първи най-близък съсед (*1NN*), Наивен Бейс (*Naive Bayes*) класификатор и

Support Vector Machines (SVM). За обучение на *SVM* се използва алгоритъм *Sequential Minimal Optimization algorithm for training a SVM classifier* [95].

Оценката за представянето за всеки класификатор е класификационната грешка (*error-rate / ER)* - броят на случаите с неправилна прогноза се разделя на броя на всички обекти, представен като процентна стойност. Оценка класификационна грешка е възприета в настоящето проучване с цел възможност за сравнение на предложената методология с другите изследвания над *UCR* бази времеви данни, където е използван този измерител ([98], [12], [18], [25]).

Анализира се цялостното представяне на различни класификационни техники и влиянието на прилагания метод за категоризация.

- Анализ по отношение на класификационния алгоритъм
- Анализ на методите за категоризация.

Сравнение с други изследвания

Извършено е сравнение на резултатите, получени при класификация на *UCR* колекция от времеви бази данни чрез представяне в *ЧСП* формат с други проучвания, приложени върху същите *UCR* бази данни и публикувани в престижни издания - от *DTW* [98], *TSBF* [12], *Bag of patterns (BoP)* приложен върху *SAX* трансформирани данни [18], *Feature-based* [25].

Предложеният в настоящия дисертационен труд подход на ЧСП постига най-малка грешка (*НМГ*) в **7** (4 + 3) от двадесетте разглеждани *UCR* бази времеви данни. Подходи *DTW* и *BoP* следват с **5** *НМГ* (1 + 4 за *DTW* и 2 + 3 за *BoP*). Подходът на *TSBF* постига *НМГ* при **4** (3 + 1) бази данни и *Feature based* има **2** *НМГ*..

Резултатите показват превъзходство в представянето на предложената в докторантурата методология и разработения нов метод за представяне на времевите данни чрез честота на срещане на прототипите (ЧСП), което може да служи като експериментално доказателство на аргументирания избор и качествата на подходите.

4.3. Обобщение и изводи

Изследвано е приложението на предложената методология над колекцията *UCR* от бази времеви данни с класова принадлежност.

Използван е разработения метод за представяне на времевите данни чрез честота на срещане на прототипи (ЧСП). При реализацията на метода се използват ***времеви производни*** за атрибути на сегментите и решетъчен метод за определяне на прототипите, като са изследвани 3 алгоритъма за категоризация.

Важно е да се отбележи, че ЧСП методът не се нуждае от специфични настройки и е приложен по унифициран начин към всички изследвани бази времеви данни. *UCR* базите данни са от изключително разнообразни области на приложение, което потвърждава широка приложимост на ЧСП метода над времеви данни от различни сфери.

Извършена е задачата за класификация на времеви данни от колекцията на *UCR* по предложената методология по три класификационни алгоритъма. Анализирано е представянето по отношение на класификационния алгоритъм и използваните методи за категоризация, като са направени изводи за методите на клъстеризация и категоризация, демонстриращи най-добри резултати (*SVM* клъстеризация, съчетано с категоризация с равна ширина и категоризация, базирана на осреднената стойност).

Резултатите от прилагането на предложената методология с представяне на времевите данни чрез ЧСП са сравнени с най-добрите изследвания и постижения в областта. Представянето чрез разработената методология постига най-добър резултат измежду всички подходи. Това потвърждава приложимостта и практическото значение на възприетия подход, както и значимостта на предложеното в дисертацията представяне на времевите данни чрез ЧСП за целите на извличане на зависимости в данни.

Глава 5. Софтуерно решение за извличане на зависимости от потоци данни

В настоящата глава се разглежда реализацията на софтуерен пакет, наречен *Data Expert*, за провеждане на научни експерименти по предложената методология за извличане на зависимости от времеви данни. Дефинирани са изискванията на софтуерното решение и

областта му на приложение, избора на технологии и инструменти, архитектура, организация на резултати.

5.1. Изисквания към софтуера

Предложената методология за извличане на зависимости от времеви данни се нуждае от съответстващ инструментариум за извършване на научни изследвания. Необходимо е създаване на софтуерно решение, което да реализира отделните етапи от методологията. Използването на *Data Expert* софтуера е в научно-експерименталната област, като трябва да позволява провеждане на разнообразни експерименти с използване и комбинация от различни методи и алгоритми, както и възможност за удобно представяне на получените резултати за последващ анализ. Следните изисквания, представени обобщено, са приложими и с практическо значение към разработваното решение:

Функционални изисквания

- Софтуерна реализация на създадения метод за представяне на времеви данни чрез честота на срещане на прототипи *ЧСП*; Разработване на предложената процедура на методологията за извличане на зависимости от времеви данни; Поддръжка на широк набор от алгоритми при изпълнение на отделните етапи на методологията;

- Възможност за изпълнение на различни комбинации от методи над разнообразни данни; Лесна конфигурация за стартиране на нови експерименти с определени настройки – изследвани данни, задача за изпълнение, метод за представяне на времевите данни, избор на характеристики, интелигентен анализ на данните, запазване на резултата; Способност за стартиране на множество експерименти;

Нефункционални изисквания

- Ограничени нужди от изчислителни ресурси с цел възможност за извършване на проучвания на локален компютър;

- Желателно софтуерът да притежава скалируемост като има потенциална възможност за разширение в размер и обем при прилагане в мощни хардуерни конфигурации;

- Лесен за поддържане програмен код, четим и разбираем, с достатъчно коментари и обяснения.

5.2. Избор на технологии и инструменти за разработка на софтуерното решение

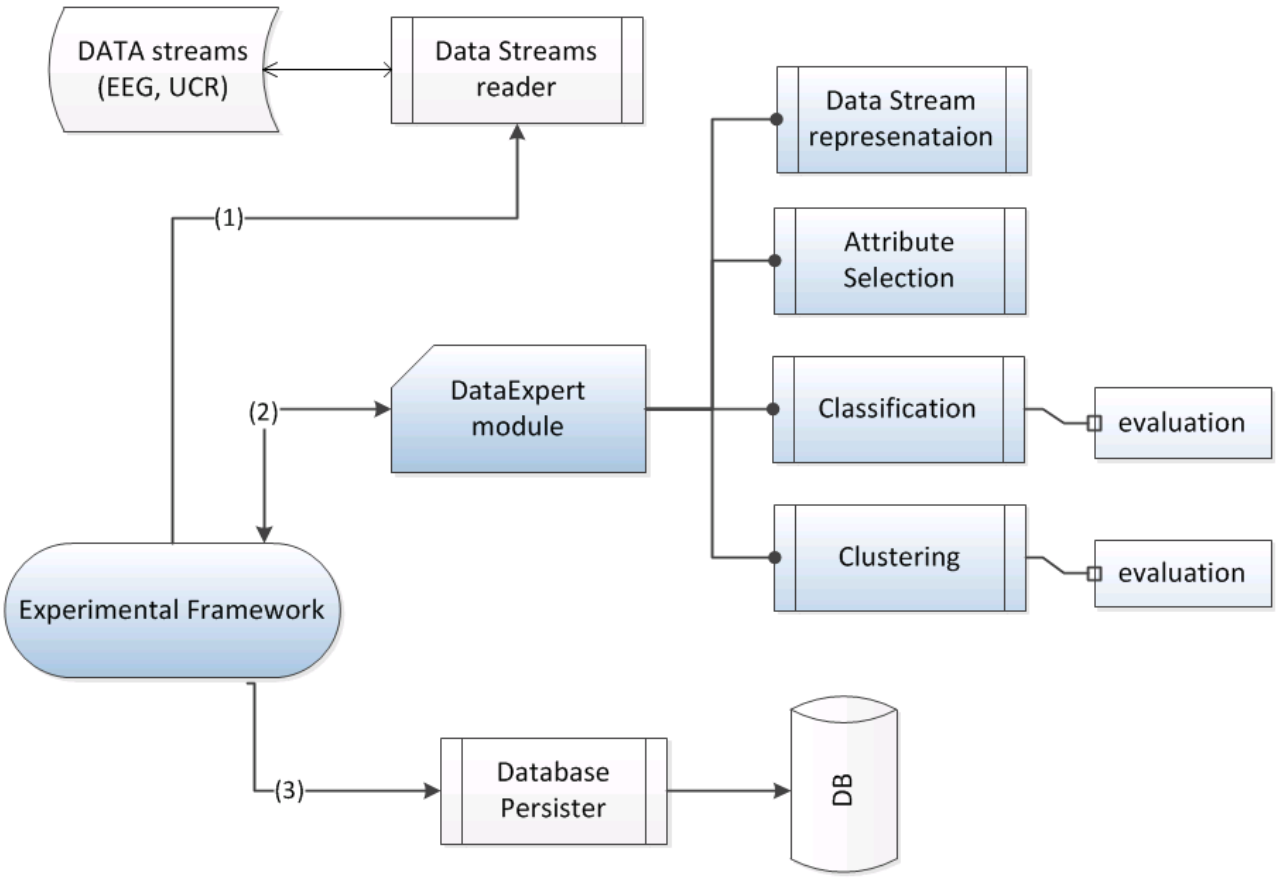
На база на описаните функционални изисквания към софтуера са избрани подходящи технологии и инструменти за неговата реализация, както е проектирана неговата архитектура. Разгледани са следните аспекти:

- Избор на тип приложение
- Избор на програмни средства за разработка на софтуера
- Използване на механизма за отражение (*Reflection*)
- Използване на система за контрол на версиите

5.3. Основни модули

Приложението *Data Expert* е проектирано, използвайки модулна структура. Това позволява запазване на общата архитектурна функционалност и независимо разширяване с допълнителни нови методи във всеки етап от методологията.

Функционирането на системата е представено в обобщен вид на следващата диаграма на основните модули (*Фигура 29*).



Фигура 29. Основни модули на софтуерната система

5.4. Визуализация и анализ на получените резултати

Визуализацията на получените резултати се възлага (делегира) на външни програми. Различни изгледи на резултатите се извличат чрез езика *SQL*. представени са няколко примера:

- Представяне на резултатите от клъстеризация на ЕЕГ потоци данни
- Представяне на резултатите от класификация на UCR потоци данни

5.5. Обобщение

Представеният софтуерен продукт е разработван от автора в периода 2013-2017 г. паралелно с настоящия дисертационен труд. По лично мнение на автора, реализацията на проекта отговаря на съвременните функционални и технологични изисквания, осъществявайки една успешна система за извличане на зависимости от потоци данни.

Разработката на софтуерния пакет *Data Expert* е базирана на съвременни технологична база, спазва управляем работен процес за разработка и поддръжка на софтуера, прилага системи за управление на програмния код.

Софтуерното решение следва съвременен работен процес, използва системи за управление на програмния код, достатъчно документация в кода и между различните версии. Приложението позволява разширяване с допълнителни нови методи във отделните съставни части, благодарение на модулната си архитектура, както и лесна настройка за изследване на нови задачи.

ЗАКЛЮЧЕНИЕ

В дисертационния труд е разработена методология като систематизиран подход за анализ на времеви данни, която използва и комбинира множество методи и алгоритми. Тяхното изпълнение е обвързано в единна процедура, позволяваща да бъде обогатявана с нови методи/алгоритми в отделните етапи. Методологията може да се използва в приложения от разнообразни области, притежава научни нововъведения и оригиналност. Резултатите от прилагането ѝ са сравними с най-добрите изследвания и постижения в областта и за определени случаи ги превъзхождат.

Приносите на дисертационния труд могат да се обобщят както следва:

Научни приноси

- Предложена е нова класификация на методите за анализ на потоци от данни, с отчитане на спецификите на времевите данни и типичните задачи, приложими над тях и съществуващите подходи за обработка и анализ на потоци от данни;

- Разработен е нов метод за представяне на потоците данни в структура, посредством честота на срещане на прототипи (ЧСП). Методът е основан на изследване на локални свойства на времевите данни и е съставен от 4 стъпки - сегментация на времевите редове чрез способа на плъзгащия прозорец; представяне на сегмент чрез множество атрибути с цел сравняване на сегменти; определяне на ограничен брой прототипи от сегменти и накрая - представяне на потоците данни чрез честотите на срещане на прототипите;

- Предложена е цялостна методология, формулирана като систематичен подход, за извличане на зависимости от потоци данни. Подробно са разгледани отделните етапи, следващи анализа и представянето на потоците от данни в обобщен вид, а именно - избор на характеристики, клъстеризационни и класификационни методи за анализ и извличане на зависимости от данните и интерпретация на получените резултати.

Научно-приложен принос

- Прилагайки методологията за извличане на зависимости над ЕЕГ данни на активността на мозъка при наблюдение на картини с високо положително и отрицателно емоционално въздействие, са направени изводи във връзка с избора на значими характеристики на ЕЕГ данните, както и възможността за използване на различни клъстеризационни и класификационните алгоритми. Локализирани са ЕЕГ каналите, осигуряващи по-висока клъстеризационна точност с цел дефиниране на пространственото разположение на зоните в мозъка, отговорни за решението на тази визуална задача; представени са важни изводи за възможностите за дискриминация по канали и по участници в експеримента. Изследванията осигуряват основа за автоматизиране на прогнозата на човешкото състояние, представлявайки първоначален опит за разкриване на общи представителни модели на емоционалното състояние на индивидите.

- Изследвани са резултатите от приложението на предложената методология при задачата за класификация над голяма и разнообразна колекция *UCR* бази от времеви данни с класова принадлежност като е използван разработения нов метод за представяне на времевите данни чрез честота на срещане на прототипи. Сравнителният анализ с други изследвания върху тези данни показва значим резултат и потвърждава приложимостта и практическото значение на създадения метод за представяне чрез ЧСП и последващ анализ на времеви данни.

Приложен принос

Разработен е софтуерен пакет, който реализира предложената методология за извличане на закономерности от данни.

Бъдещи направления за развитие и усъвършенстване на изследванията

Набелязани са няколко основни направления за бъдещо усъвършенстване и развитие на предложените в дисертационния труд подходи.

- По отношение на разработения метод за *представяне на потоци от данни чрез честота на срещане на прототипи* е интересно да се изследват различни вариации при отделните стъпки:
  - чрез алтернативен метод за изчисляване на атрибутите на сегментите;
  - посредством друг метод за определяне на прототипите.
- По отношение на предложената в дисертационния труд *методология за извличане на зависимости от потоци данни* би било полезно разширяване на приложението ѝ в проучените области, както и прилагане в нови области:
  - Използване на подхода над различни ЕЕГ потоци данни - с повече участници, повече класове или при друг вид възбуда (например звукова);
  - Приложение върху потоци данни от други области - финансови пазари, медицински и биологични записи, метеорологични данни, стойности, получени от различни сензори.
- *Софтуерното решение за извличане на зависимости от потоци данни* е разработено като средство за провеждане на научни експерименти. Подчертани са гъвкавостта и лесната конфигурация на различни научни опити, както и на разбираемостта на програмния код. Софтуерният пакет може да се разшири с нови функционалности и методи, както и да се извършат подобрения по отношение на оптимизация на използваните алгоритми. Добре е да се обмисли евентуалното му предлагане като услуга (включително облачна услуга) или представяне като отворен код.