РЕЦЕНЗИЯ

на дисертационен труд

„Извличане на зависимости в потоци от данни“

за присъждане на образователна и научна степен “Доктор”
в професионално направление 4.6. Информатика и компютърни науки,
докторска програма „Компютърни науки (Изкуствен интелект)“

Автор: Сергей Миланов

Докторант към Катедра „Софтуерни технологии”, Факултет по Математика и Информатика, СУ „Св. Климент Охридски“

Научен ръководител: доц. д-р Олга Георгиева , ФМИ- СУ

Рецензент: проф. д-р Анна Лекова, ИСИР-БАН

Настоящата рецензия се съставя въз основа на решение на Факултетния съвет на Факултета по математика и информатика (ФМИ) от 30.01.2017 г. за избор на Научно Жури, определено със Заповед № 38-132/21.02.2017 г. на Ректора на Софийския университет „Св. Климент Охридски" проф. дфн Анастас Герджиков, и съвкупността от критерии и показатели за придобиване на образователната и научна степен „доктор" в ЗРАСРБ, неговия Правилник и Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности на СУ и Факултета по математика и информатика на СУ „Св. Климент Охридски".

1. Основни сведения.

Представеният дисертационен труд е с обем 152 страници и се състои от въведение, пет глави, заключение, общи изводи и насоки за развитие на изследванията. Списъкът на цитираната литература се състои от достатъчно на брой заглавия - 107. Дисертацията съдържа 32 фигури и 17 таблици. Към документацията е приложен диск, който съдържа публикациите, автобиография, автореферат и дисертационния труд. Документацията и всички процедури по защитата са в съответствие с изискванията на закона и съответните правилници.

Авторът Сергей Миланов е задочен докторант към катедра „Софтуерни технологии”, ФМИ. Роден е 1974 г. Завършил е бакалавърска и магистърска програма на ФМИ – СУ през 1998г, а от 2000 г. до сега работи в „Кодикс България“ ЕАД като програмист и ръководител отдел „Java приложения”.

2. Съдържателен анализ на научните и научно-приложни приноси на дисертационния труд

Най-общо казано, дисертационният труд разглежда въпроси, свързани със стремителното нарастване на обема на данни с които сме заобиколени през последните години и

невъзможността да разберем напълно тяхното значение и наличие на закономерности, тъй като са неявни. Конкретно се разглеждат времеви данни със значителна продължителност (потоци данни). Много ясно са формулирани целта и задачите на дисертационен труд за *обработка и анализ на времеви потоци данни и извличане на зависимости от данни*, в резултат на което е разработен нов метод за подходящо представяне на времевите данни с цел тяхното структуриране в разбираем вид и е предложена, проектирана и разработена нова методология за разкриване на скрити свойства, връзки и зависимости в потоците данни. Разработено е софтуерно решение за извличане на зависимости от потоци данни и е приложено за различни бази от динамични данни: ЕЕГ сигнали и времеви редове от UCR базата данни.

Дисертационният труд започва с обзор на използваните понятия и теоретичната основа на главните задачи в процеса за извличане на зависимости от данни - избор и управление на характеристики, извличане на зависимости от данни с акцент класификация или клъстеризация, и оценка на резултата. Представено е задълбочено систематично проучване на съществуващите подходи за обработка и анализ на потоци от данни. Тук е предложена нова, цялостна класификация на методите за анализ на времеви данни. В четвърти раздел на първа глава е мотивирана необходимостта от създаване на нова методология за извличане на зависимости от потоците данни. В резултат на литературния обзор и забелязания недостиг на задълбочени изследвания в литературата за извличане на зависимости от потоци данни отчитайки времевите им специфики и свойства, докторантът смята за подходящо да се изследва и предложи нов систематичен подход, който съчетава различни методи за обработка и анализ на потоци от данни и подходящото им представяне във времето с цел тяхното редуциране и структуриране.

Във втора глава е представена нова систематизирана методология за извличане на зависимости от времеви редове. В първия раздел е разработен нов метод за представяне на времевите редове чрез честотата на срещане в тях на създадени прототипи. Във втори раздел е много добре изложена разработената методология за извличане на закономерности от времеви данни. Добре е представена общата теоретична рамка, чрез поредица от етапи, и разработените и използваните методи и алгоритми за всеки етап. Уместни са дадените примери.

В трета глава се прилага разработената методология за извличане на зависимости от ЕЕГ потоци данни, получени при наблюдение на картини с високо положително и отрицателно емоционално въздействие. Първи раздел илюстрира как ЕЕГ потоците от данни се представят в нова структура - чрез 12 на брой характеристики на базата на първите шест локални екстремума на био-сигнала и времето на тяхното настъпване. Данните от ЕЕГ записите се групират по различен начин, по ЕЕГ канала или по участници, което позволява в тях да бъдат търсени различни зависимости. Във втори раздел е показано как е приложена методология към ЕЕГ потоците данни, като са разгледани подробно отделните етапи и приложените алгоритми за извличане на зависимости, получени както чрез клъстеризация, така и класификация. В трети раздел са анализирани резултати от множество извършени експерименти по отношение на надеждността на алгоритмите правилно да разграничават двата класа, свързани с положителни и отрицателни емоции. Четвърти раздел резюмира резултатите от тази глава като много удачно е взета под внимание *невро-биологичната гледна точка*, която доказва на практика удачното

прилагане на развитата методология при избора на характеристики и извличане на зависимости от ЕЕГ потоци данни.

Четвърта глава подробно илюстрира извличане на зависимости от UCR колекция от времеви данни. Добре са подбрани използваните UCR времеви редове, от изключително разнообразни области на приложение и силно различаващи се по дължина, броя на класовете в тях и размерите на обучаващата и тестова извадка. Подробно са разгледани представянията на UCR времевите данни чрез честота на срещане на прототипи, описани са извършените експерименти и е направен анализ на получени резултати по отношение на класификационния алгоритъм и методите за категоризация. Резултатите от прилагането на новата методология са сравнени с най-добрите изследвания и постижения в областта и илюстрирани много добре в таблици 13 и 14. Представянето на времевите данни чрез ЧСП постига често 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.

Пета глава представя разработеното софтуерно решение Data Expert за извличане на зависимости от потоци данни като е описана архитектурата на изграждащите го модули за представяне в нова структура и разкриване на закономерности в потоци от данни, както и визуализация и анализ на получените резултати. Силно впечатление прави разработката на Data Expert - базиран на съвременни технологична база, спазва управляем работен процес за разработка и поддръжка на софтуера. Разработен като средство за провеждане на научни експерименти с гъвкавост и лесната конфигурация на различни научни опити и разбираемостта на програмния код, Data Expert може да се разшири с нови функционалности и методи, както и да се извършат подобрения по отношение на оптимизация на използваните алгоритми.

Основните заключения и изводи в дисертационния труд са много добре обобщени в завършващата шеста глава. Анализират се постигнатите резултати и приноси, обсъждат се направления за бъдещи изследвания и за развитие и усъвършенстване на разработените методи и подходи. Приносите на дисертационния труд са подходящо и много ясно формулирани в три научни, три научно-приложни и един приложен принос.

Съдържанието на *автореферата* съвпада с увода на дисертацията и правилно и изчерпателно отразява основните резултати, научните, научно-приложни и приложни приноси на дисертацията. Номерата на включените в автореферата фигури и таблици съвпадат с тези в дисертационния труд с изключение на Фигура 15, която всъщност трябва да е 25, както и Фигура 29, която всъщност трябва да е 31.

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“. Те отговарят като брой и качество на ЗРАСРБ и препоръчителните изисквания на ФМИ към СУ. Въз основа на авторските претенции в дисертацията, считам, че кандидатът има водещо или поне равностойно участие в съвместните публикации.

3. Критични бележки и препоръки.

3.1. Дисертацията е написана ясно и разбрано. Има незначителен брой правописни грешки, както и някои технически. Например, при илюстриране на ЧСП метода, описанието на четирите сегмента на стр. 63 не съответства на графичното им представяне на фиг.14 на стр. 62 - по ординатата интервалът трябва да е от 0 до 20.

3.2. На 39 стр. алгоритмите за избор на характеристики се разделят на три основни категории по отношение на метода използван за оценка: обхващащи (wrappers), филтърни (filters) и вградени (embedded), но може да се спомене и четвърта категория - статистически алгоритми за обучение като Markov Blanket.

3.3. Типовете използвани алгоритми за избор на характеристики могат да бъдат допълнени с широко известните и често използвани алгоритми за редуциране броя на характеристики чрез извличане и конструиране на нови по-значими характеристики, такива като вероятностно клъстеризиране, Principal Component Analysis (PCA), ISO maps и др. Тъй като разработената методология използва и комбинира множество методи и алгоритми и позволява да бъде обогатявана с нови методи/алгоритми в отделните етапи, тези алгоритми би трябвало лесно да могат да бъдат добавени в софтуера. А особено ако Data Expert се предостави като отворен код, самите потребители ще могат да добавят необходимата им функционалност.

3.4. Въпреки, че Support Vector Machine classifier (SVM_SMO) е един от най-широко използваните алгоритми за класификация на био-сигналите, препоръчвам на автора с цел развитие и усъвършенстване на изследванията в бъдеще да анализира и другият масово използван метод за изчисляване на сходство между времеви редове съставени от био-сигнали - Dynamic Time Warping (DTW). В първа глава авторът го класифицира към подходите за анализ на времеви редове базирани на суровите стойности, но той може да бъде също така прилаган и към атрибути/характеристики (в сегментите) на потоците от данни. Неговото експериментиране в рамките на предложената методология върху ЕЕГ сигналите би обогатило получените резултати в сравнение с клъстеризационните и класификационни алгоритми, тъй като DTW се очаква да е по-гъвкав към промените във фазата на времевите данни, т.е. времето на появяване (латентността) не емоционалната реакция при различните участници. Самият автор казва в трета глава, че при оценка на базите данни, обединени по канал, характеристиките от амплитуден тип са еднакво важни с характеристиките от тип латентност. Затова препоръчвам DTW алгоритъма да се добави от библиотеки WEKA и JavaML, за да се анализира емоционалната реакция по канали без тя да зависи от скоростта на физиологичния сигнал при различните участници.

4. ЗАКЛЮЧЕНИЕ

Представеният дисертационен труд има всички качества на сериозно научно изследване в областта на извличане на зависимости в потоци от данни. Получените интересни резултати предоставят нов систематичен подход за извличане на зависимости от потоци данни и метод за представяне на потоците от данни в нова структура, посредством извличане и анализ на техни характеристики. Една част от приносите имат определено теоретичен характер, а друга част дават нови методи и алгоритми за представяне и категоризация на времеви данни. Добро впечатление остава работата с реални данни и получаването на крайни резултати, сравними и дори превъзхождащи най-добрите изследвания и постижения в областта. Всичко това показва, че авторът е навлязъл сериозно в научната тематика.

Получените резултати са докладвани в две международни списания, едното от които с импакт фактор, както и на две IEEE международни конференции. Всичко това ми дава пълно основание да считам, че са удовлетворени всички условия от ЗРАСРБ и неговия правилник, както и този на ФМИ - СУ, които се прилагат за получаване на научната степен „Доктор“ по съответната специалност. Ето защо категорично предлагам на почитаемото жури да оцени подобаващо високо представения дисертационен труд, а на неговия автор Сергей Миланов да бъде присъдена образователната и научна степен „ДОКТОР“ в професионално направление 4.6. Информатика и компютърни науки, докторска програма „Компютърни науки (Изкуствен интелект)“

Рецензент:

/професор д-р Анна Лекова/

24.04.2017