РЕЦЕНЗИЯ

ПО ДИСЕРТАЦИОНЕН ТРУД ЗА ПРИСЪЖДАНЕ НА ОБРАЗОВАТЕЛНА И НАУЧНА СТЕПЕН "ДОКТОР"

Професионално направление 4.6. Информатика и компютърни науки
(по научна специалност 01.01.12. "Информатика")

ТЕМА: **"Изследване на Data Mining модели за класификация"**

АВТОР НА ДИСЕРТАЦИОННИЯ ТРУД: **ас. Дорина Петрова Кабакчиева**

ИЗГОТВИЛ РЕЦЕНЗИЯТА **:Проф.д.н. Благовест Шишков – асоц. чл. на ИМИ-БАН**

1. Актуалност на проблема

В настоящия свят на ICT технологиите, мобилните технологии и използването на Internet във всяка точка на света, родиха нови научни направления като Data Mining, които се базират на постиженията на редица други познати и добре развити научни дисциплини като статистика, машинно обучение, изкуствен интелект, бази данни и др. Това, което отличава Data Mining от останалите дисциплини, е интердисциплинарният характер, който се заключава в интегрирането на знанията за бази данни, аналитични методи и средства, и бизнес познанията.

„Educational Data Mining" е едно от най-новите направления в Data Mining, разглеждащо въпроси, свързани с идентифицирането на подходящи студенти (targeted marketing), задържане на студенти и предотвратяване на тяхното отпадане от обучение (retention of students), повишаване качеството на образованието, с оценка на реализацията им чрез връзката с бивши възпитаници (alumni management). Предсказването успеха на студентите е много важен и актуален проблем за университетите и най-често се извършва чрез решаване на Data Mining задача за предсказване. В случаите, когато предсказваната величина е номинална променлива, задачата се трансформира в задача за класификация. Често използвани методи за генериране на модели за предсказване са „Дърво на решенията", „Невронни мрежи", Бейсови методи - наивен Бейсов класификатор, Бейсови мрежи, логистична регресия и др. Тази съвременна технология е приложима и за България, понеже в годините на прехода и българските университети претърпяват различни промени и се изправят пред сериозни проблеми, като: намаляване броя на потенциалните студенти; конкуренцията между университетите и др.

От друга страна е известно, че тази технология за класификация чрез обучение на класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни, може да се приложи и към различни по вид и съдържание масиви от данни. Прилагането на Data Mining подхода към задачата за класификация на радарни морски цели е особено актуален и оригинален, поради липса на подобни изследвания. Обичайно тези задачи за откриване и класификация на сигнали се решават с

методите на статистическата теория на решенията, при априорна определеност и неопределеност, или теорията за разпознаване на образи - ако радарните сигнали се разглеждат като двумерни изображения, което позволява да се използват и подходите на компютърните науки, като невронните мрежи, дърветата на решения и др.

Именно на решаването на някои частни задачи от тази актуална научна област е посветен настоящият дисертационен труд. По-конкретно, на изследване на приложимостта и ефективността на различни Data Mining методи за класификация, в две различни предметни области - данни за студенти и данни за открити морски цели.

2. Анализ на научните постижения на кандидата по съдържанието на дисертационния труд

Дисертацията се състои от увод 3 глави и заключение. В ***първа глава*** е направен много обширен обзор на Data Mining областта, разгледани са четири Data Mining метода за генериране на модели за класификация, описани са различни мерки за оценка и сравнение на моделите, както и обзор на състоянието на научните изследвания – обучението на студенти и класификацията на морски цели, на базата на съвременни публикации и монографии, както е прието в международната публикационна дейност. Прави добро впечатление на рецензента, че такива сериозни обзори са рядкост за докторски дисертации. На основата на направените изводи по обзора, са формулирани целта и основните задачи на дисертационния труд.

Разглеждам приносите в тази глава като научно - приложни, и по-точно в класа: ***приноси за внедряване,*** като: *методи, конструкции, технологии, схеми, реализация на по-ефективни средства за приложения.*

Във втора глава докторантката, въз основа на така формулираните научни задачи, е представила: методиката за провеждане на изследването. Избраният подход за реализация на Data Mining проекта; избора на подходящи софтуерни средства; предварителната подготовка и изследване на двете съвкупности от данни – за студенти и за морски цели се използват при генерирането на Data Mining моделите за класификация. По мнение на рецензента в тази глава са получени важни научно-приложни резултати, в частта подготовка на данни за обучение на класификатори и за двете изследвани области. Формирана е, на база емпиричните умения на докторанта, крайна съвкупност от данни за студентите, която се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA, като изходната/предсказваната променлива „Среден успех от I курс" е преобразувана в номинална променлива с пет стойности и номинална променлива с 2 стойности. В резултат от специфична сигнална обработка на записи от морски цели, включваща откриване/измерване на параметрите на сигналите от целите на фона на море, с използването на CFAR откривател и K/M-L параметричен откривател/измервател на дължината на семплите на сигнала, т.е. броя на откритите сигнални семпли, успешно са преведени в пространството на параметрите на сигналите като: дължина на открития сигнал, мощност, и съответно енергия, която е произведение на двете предходни величини. В този смисъл сигналните данни са трансформирани в подходящ формат за прилагане на Data Mining методи и средства, получени при съвместна работа на докторантката с изследователския международен колектив по проект с Националния Фонд "Научни Изследвания". Изследваната целева променлива е от номинален тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*. Разглеждам приносите в

тази глава, като научно-приложни, и по-точно в класа: ***приноси за внедряване,*** като: *методи, конструкции, технологии, схеми, реализация на по-ефективни средства за приложения.*

В трета глава са представени резултатите от изследването на избраните Data Mining алгоритми за класификация върху двете подготвени съвкупности от данни, за студенти и за морски цели, с помощта на Data Mining софтуер WEKA. Оценени и сравнени са генерираните Data Mining модели за класификация чрез избраните мерки за оценка. В тази глава са получени и представени основните оригинални научно-приложни приноси на дисертантката. Извършено е обучението са класификатори за извличане на знания от данни, чрез избраните класификационни алгоритми, за две съвкупности от данни, в Data Mining софтуер *WEKA*. Обучените класификатори са оценени по избраните мерки за оценка.

Другото, което заслужава да се отбележи е, че в дисертацията са избрани и изследвани обучени класификатори, на базата на една и съща методика, за две различни съвкупности от данни. Това е известно, но за случая на дисертацията - за изпитни оценки, и радарни сигнали, никак не е тривиално да се получи. Затова следват и бързите цитирания (в рамките на една година) от публикуването на материалите по дисертацията и в двете изследвани съвкупности.

Разглеждам приносите в тази глава като ***научно-приложни, по точно с характер на обогатяване на съществуващи знания*** - *получаване на нови факти, зависимости, изводи и заключения.*

Всичко това ми дава основание да твърдя, че докторантката е решила коректно поставената научна задача в дисертацията, като е получила необходимите научни резултати. В тази глава са получени редица важни за практиката резултати. По мнение на рецензента, получените резултати не противоречат на известни такива, допълват ги и ги подтвърждават. Предлагат се подходи, Data Mining алгоритми, методики и техники за обучени класификатори, на базата на една и съща методика, за две различни съвкупности от данни.

Като обобщение може да се каже, че дисертантката е получила добро образование, навлязла е в областта на Data Mining алгоритми за класификация, извършила е изследвания върху двете подготвени съвкупности от данни, за студенти и за морски цели, с помощта на Data Mining софтуер WEKA, усвоила е технологията за избор на модели, алгоритми и техники, с цел подготовка на данни за обучение Data Mining на класификатори, тяхното изследване с правилно избрани метрики, като е получила нови научни резултати, публикувани в две списания и четири доклада на международни конференции.

Относно литературната осведоменост на авторката може да се каже, че тя е съвременен изследовател, добре запознат с последните постижения в областта на изследванията. Практическите й умения в областта се демонстрират чрез проведените реални експерименти с помощта на персонален компютър върху програмни продукти: ***Microsoft Excel*** (Microsoft Office 2007) - за избор и интегриране на данни от различни източници, за изчистване и предварителна подготовка на данните, за подготовка на данните във формат, подходящ за използване в Data Mining софтуер; ***Microsoft Excel*** (Microsoft Office 2007); ***QlikView*** (v9.0) и ***WEKA*** (v3.6) - за изследване, опознаване и описание на данните; както и ***WEKA*** (v3.6) - за моделиране и оценка на получените резултати.

Следователно, като цяло научно–приложните приносите на авторката отговарят на изискванията, установени от „Закона за развитието на академичния състав в Република България“ и Правилника на ИИКТ, за получаване на исканата научна и образователна степен ”Доктор”, и са ***научно-приложни,*** в областите - ***обогатяване на съществуващи знания*** и по-точно*: получаване на*

нови факти, зависимости, изводи и заключения; и ***приноси за внедряване,*** като: *методи, конструкции, технологии, схеми, реализация на по-ефективни средства за приложения.*

3. Научно-приложни приноси, реализирани в дисертацията

По-конкретно, научно-приложните приноси на авторката са в избор и изследване на обучени класификатори, на базата на една и съща методика, за две различни съвкупности от данни, както и в предварителната подготовка на двете съвкупности от данни, за студенти и за морски цели, почиващи на добре обосновани методи, както и използването на подходящ за целта софтуер.

По-важните научно–приложни приноси, в областта на ***обогатяване на съществуващи знания и приноси за внедряване,*** които отбелязвам, са в областта на ***обогатяване на съществуващи знания,*** *получаване на нови факти, зависимости, изводи и заключения:*

3.1. Извършено е откриване на знания в данни чрез обучение на класификатори за две съвкупности от данни в различни предметни области: за предсказване успеваемостта на студентите, на базата на техни характеристики, и за предсказване на класа на морски обекти, и е оценено качеството на тези класификатори.

3.2. Качеството на предсказване на избраните модели на обучени класификатори на морски цели, работещи във времевата област, е сравнимо с резултатите, получени от други изследователи, използвали подобни алгоритми за класификация на движещи се наземни цели, работещи в честотната област.

3.3. Избраните и изследвани обучени класификатори извличат знания на базата на една и съща методика, за две различни съвкупности от данни, в две различни предметни области, което е едно добро подтвърждение на универсалността на Data Mining подхода.

3.4. Апробирана е методика на изследването, която може да се използва и в други случаи, за извличане на закономерности от данни за учениците и студентите в различни училища, университети, или данни за радарни и GPS сигнали от различни видове цели.

4. Относно публикациите по дисертационния труд

Публикациите на дисертантката са шест на брой, като пет са в съавторство с колеги и нейния ръководител. В тях авторката представя основните си резултати, получени в дисертацията си. В пет публикации дисертанката е на първо място. Всички са на латиница и са публикувани в две списания и четири доклада на международни конференции. Списанията са: международно електронно списание "International Journal of Computer Science and Management Research" и национално специализирано списание "International Journal of Cybernetics and Information Technologies" (приета за печат). Докладите на международните конференции са в Германия, Холандия, Варна, София, публикувани през 2006, 2011, и 2012г. На рецензента му е известно, трудовете на дисертантката са цитирани в статия в международно специализирано списание с висок импакт фактор и международна конференция в чужбина, което по себе си е висока международна оценка и за двете изследвани области и рядкост за докторска дисертация.

5. Критични бележки

Бих направил 2 забележки:

1. Да не се изключват асимптотическите методи при извличането на знания от големи обеми данни, както това е направно в дисертацията, защотото асимптотически оптималните алгоритми, съчетани с адаптация на класификаторите са най-мощният способ за преодоляване на априорната неопределеност в информационните системи.

2. Да не се премълчава ролята на руските учени в областта на теорията на вероятностите и информационните системи и особено в задачата за прогнозирането(forecasting), която за първи бе решена от А.Н. Колмогоров.

6. Препоръки

1. Би било добре дисертантката да се насочи в бъдещата си работа към автоматизирани процедури за подготовка на данни.
2. Препоръчвам на авторката, в бъдеще получените научни резултати да бъдат публикувани повече в наши и международни специализирани списания.

7. Заключение

На основание на горе казаното, и по моя преценка, ас. Дорина Петрова Кабакчиева е изграден научен работник, умеещ самостоятелно да формулира и решава научни задачи. Смятам, че представеният дисертационен труд отговаря на изискванията и има необходимите качества, установени от „Закона за развитието на академичния състав в Република България“, както и на правилника на ИИКТ-БАН за получаване на исканата образователна и научна степен "Доктор".

Поради това предлагам убедено на почитаемото научно жури по обявената процедура за защита на докторска дисертация в ИИКТ-БАН, по професионално направление 4.6. Информатика и компютърни науки (по научна специалност 01.01.12. "Информатика"), да присъди образователната и научна степен „Доктор" на ас.Дорина Петрова Кабакчиева.

.

15.02.2013
София