БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

Институт по Информационни и Комуникационни Технологии

асист. маг. Дорина Петрова Кабакчиева

Изследване на Data Mining модели за класификация

Автореферат

за придобиване на образователна и научна степен "Доктор"

Професионално направление 4.6. "Информатика и компютърни науки"

(по научна специалност 01.01.12. "Информатика")

Научен ръководител

акад. проф. д.н. Иван Попчев

София

Декември 2012г.

Дисертацията е обсъдена и допусната до защита на разширено заседание на секция „Интелигентни системи” в ИИКТ на БАН, състояло се на 10.12.2012г.

Дисертацията съдържа 193 стр., 25 формули, 75 фигури, 36 таблици, 6стр. Библиография, включваща 71 заглавия.

Защитата на дисертацията ще се състои на .....2013г. от ..... часа в Зала ..... на ..... на открито заседание на научно жури в състав:

1. Акад.проф.д.н.Иван Попчев - ИИКТ-БАН
2. Проф.д-р Румен Николов - УниБИТ
3. Проф.д.н.Благовест Шишков - ИМИ-БАН
4. Доц.д-р Силвия Илиева - ИИКТ-БАН
5. Доц.д-р Иван Койчев - ФМИ, СУ"Св. Кл. Охридски"

Резервни:

1. Доц.д-р Олга Георгиева - ФМИ, СУ"Св. Кл. Охридски"
2. Доц.д-р Данаил Дочев - ИИКТ-БАН

Материалите за защитата са на разположение на интересуващите се в ИИКТ на БАН, София, ул.”Акад. Г. Бончев”, бл.2.

Автор: Дорина Петрова Кабакчиева

Заглавие: Изследване на Data Mining модели за класификация

УВОД

Извличането на знания от големи обеми данни (Data Mining) е сравнително ново научно направление - възникнало през последните 20-25 години. В действителност, изследванията в тази област не са напълно нови и непознати за научната общност, тъй като 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 Извличането на знания от големи обеми данни (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.

Настоящият дисертационен труд е посветен на изследване на приложимостта и ефективността на различни Data Mining модели за класификация. Под *Data Mining модели за класификация* в дисертацията се разбира *обучени класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни*. Терминът *Data Mining* често се използва в текста на дисертацията вместо неговия български еквивалент - *извличане на знания от големи обеми данни*, с цел постигане на по-кратък и ясен изказ.

Изследванията са осъществени върху данни, произхождащи от две различни предметни области - данни за студенти и данни за открити морски цели, с което се цели да се потвърди още един път твърдението, че Data Mining методите и средствата могат да бъдат прилагани върху различни видове данни от различни научни области.

През последните години българските университети претърпяват различни промени и се изправят пред сериозни проблеми. Намалява броят на потенциалните кандидати за висшите учебни заведения и се засилва конкуренцията между образователните институции, което води до необходимостта от актуална информация, която да подпомага ръководствата при взимането на важни и своевременни управленски решения. Съвременните университети разполагат с големи обеми данни, които в повечето случаи не са ефективно анализирани и използвани. Това може да бъде постигнато чрез използване на Data Mining методи и средства.

„Educational Data Mining" е едно от най-новите направления в Data Mining, посветено на решаването на проблеми, свързани с привличането на подходящи студенти (targeted marketing), задържане на студенти и предотвратяване на тяхното отпадане от обучение (retention of students), повишаване качеството на образователния процес, управление на кандидатстудентските кампании, подобряване на организационната ефективност, управление на взаимоотношенията с бивши възпитаници (alumni management). Осъщественият анализ на публикуваните научни изследвания в областта на Educational Data Mining показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Успехът на студентите се предсказва най-често чрез решаване на Data Mining задача за предсказване, като в повечето случаи предсказваната величина е номинална променлива, т.е. решава се задача за класификация. Най-често използваните методи за генериране на модели за предсказване включват „Дърво на решенията", „Невронни мрежи", Байесови методи - наивен Байесов класификатор, Байесови мрежи, и логистична регресия. Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ модел за предсказване.

Проблемът, свързан с класификацията на открити радарни цели, занимава радарната научна общност от много време. Задачите за откриване и класификация на сигнали могат да се решават с методите за разпознаване на образи, защото радарните сигнали могат да се разглеждат като образи от данни. Тъй като при много методи за разпознаване на образи се използва статистическата теория на решенията, при априорна определеност и неопределеност, затова разпознаването или класификацията на образите се свежда всъщност до метода за многоалтернативно изпитване на хипотези. Съществува голямо разнообразие от методи за разпознаване на изображения, разработени в други научни области, например Data Mining, които биха могли да бъдат приложени за класификация на радарни цели, което е и работната хипотеза в настоящата дисертация.

С въпросите за класификация на наземни и морски цели при условията на forward scattering ефект научната общност се занимава само през последните няколко години. Един от водещите изследователски колективи в областта е екипът на проф. Черняков от Университета в Бирмингам, Великобритания. За момента не са известни изследвания, свързани с използването на Data Mining методи за класификация на радарни морски цели. Изследванията върху такъв тип данни в настоящия дисертационен труд имат за цел да проверят възможностите за ефективно прилагане на Data Mining алгоритми за класификация и по отношение на този нестандартен тип данни – forward scattering радиолокационни сигнали, и по-конкретно – за класификация на открити морски forward scattering радиолокационни цели. Особеностите тук са свързани с избор на информативни параметри на сигналите и необходимостта от предварителна обработка на записите на сигналите с цел оценка на параметрите им, за да се получат от тях данни, подходящи за аналитична обработка с Data Mining методи и средства. Друг проблем е невъзможността на този етап да бъдат получени голям обем от записи, което до голяма степен влияе върху качеството на създаваните класификационни модели.

На базата на извършените проучвания, описани по-горе, и благодарение на осигурения достъп до двете съвкупности от данни – данни за студенти, предоставени от ръководството на УНСС, и записи за морски цели, предоставени от изследователите от Бирмингамския университет, в рамките на действащи научноизследователски проекти, са дефинирани целта и основните задачи на дисертацията.

Целта на настоящия дисертационен труд е да се генерират и изследват Data Mining модели за класификация (обучени класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни), за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, и за класификация на морски обекти.

Дисертационният труд включва: съдържание, увод, три глави, заключение с научни и научно-приложни приноси, списък на фигурите, списък на таблиците и библиография. Първа глава е озаглавена "Обзор и актуално състояние на DM методите за класификация", втора глава - "Методика за провеждане на изследването. Подготовка на данните", а трета глава - "Резултати от изследването на Data Mining моделите за класификация". Основният текст е изложен на 197 страници. Включени са 25 формули, 75 фигури, 34 таблици, 6стр. Библиография, включваща 71 заглавия.

Получените резултати са публикувани в 1 статия в международно електронно списание (International Journal of Computer Science and Management Research, 2012), 3 статии, докладвани на реномирани международни конференции, 2 от които в чужбина (EDM 2011, IRS 2011) и 1 в България (S3T 2010), 1 статия, приета за публикуване в специализирано национално списание (International Journal of Cybernetics and Information Technologies), и 1 статия докладвана на международен семинар в България.

Публикации, свързани с дисертационния труд

1. Kabakchieva, D. (2012). *Student Performance Prediction by Using Data Mining Classification Algorithms.* International Journal of Computer Science and Management Research, Vol.1, Issue 4, November 2012, pp.686-690. Available at: www.ijcsmr.org
2. Kabakchiev, H., Kabakchieva, D., Cherniakov, M., Gashinova, M., Behar, V., Garvanov, I. (2011). *Maritime Target Detection, Estimation and Classification in Bistatic Ultra Wideband Forward Scattering Radar.* Conference Proceedings of the International Radar Symposium (IRS 2011), 7-9 September 2011, Leipzig, Germany, pp.79-84.
3. Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). *Analyzing University Data for Determining Student Profiles and Predicting Performance.* Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.
4. Kabakchieva, D., Stefanova, K., Kisimov, V., Nikolov, R. (2010). *Research Phases of University Data Mining Project Development.* Conference Proceedings of the Second International Conference on Software, Services and Semantic Technologies (S3T), 11-12 Sept. 2010, Varna, Bulgaria, pp.231-239.
5. Kabakchieva, D. (2012). *Predicting Student Performance by Using Data Mining Methods for Classification*. Accepted for Publication in the International Journal of Cybernetics and Information Technologies, Bulgarian Academy of Sciences.
6. Kabakchieva, D., Popchev, I. (2006). *Business Intelligence Initiatives – Valuable Opportunity and Great Challenge*. Conference Proceedings of the International Workshop "Distributed Computer and Communication Networks", 30 Oct–2 Nov 2006, Sofia, Bulgaria, organized by the Russian Academy of Sciences and the Bulgarian Academy of Sciences, pp.254-263.

Глава 1: Обзор и актуално състояние на областта Извличане на знания от големи обеми данни (Data Mining) и методите за класификация

В ***Първа глава*** е разгледана актуалността на темата на изследванията, направен е обзор на областта Извличане на знания от големи обеми данни (Data Mining), като е акцентирано върху задачата за класификация. Разгледани са четири Data Mining метода за генериране на модели за класификация, описани са различни мерки за оценка и сравнение на моделите. Направен е обзор на състоянието на научните изследвания относно използването на Data Mining методи и средства в избраните конкретни приложни области – обучението на студенти и класификацията на морски цели. Дефинирани са целта и задачите на изследването.

1.2.2. Същност на класификацията

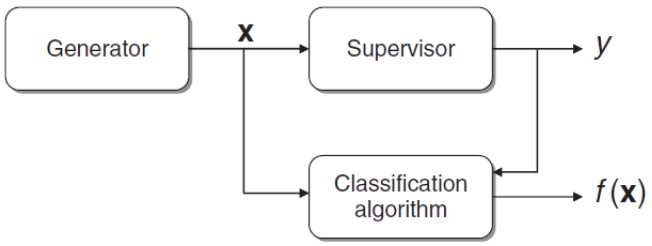
От направения обзор на Data Mining изследванията в областта на Educational Data Mining (т.1.3) следва, че най-често използваният подход е търсенето на класификационни модели за предсказване на "качеството" на студентите, като особено популярни са следните методи за класификация - „Дърво на решенията", „Невронни мрежи", „Генератор на правила", „К-най-близък съсед", Байесови класификатори и др. В обзора на областта, свързана с класификацията на морски цели (т.1.4), също е подчертана особената важност на задачата за класификация. Затова:

Основната Data Mining задача, решавана в настоящия дисертационен труд, е задачата за предсказване на променлива с дискретни стойности, т.е. задачата за извличане на знания от данни чрез обучение на класификатори.

Предсказването в този случай се разглежда като процес на изследване на реални данни, съдържащи вече класифицирани обекти, и създаване на модели, описващи закономерностите и взаимовръзките в изследваните данни, с цел да се определи (т.е. да се предскаже) класа, към който принадлежи даден некласифициран обект.

От математическа гледна точка, при решаването на задачата за класификация имаме m на брой примера, описани с двойката $(\boldsymbol{x}_i, y_i)$, $i \in M$, където $\boldsymbol{x}_i \in R^n$ е вектор от стойностите на n предсказващи променливи за i-тия пример, а с $y_i \in H=\{v_1, v_2, \ldots, v_H\}$ се обозначава съответния клас на целевата променлива. Всеки компонент x_{ij} на вектора $\boldsymbol{x}_i$ се разглежда като реализация на случайната променлива X_j, $j \in N$, представляваща атрибута a_j в съвкупността от данни D. При класификация с двоична предсказвана променлива $H=2$, т.е. имаме два класа, обозначавани например $H=\{0,1\}$ или $H=\{-1;1\}$, без да се загубва общия характер (without loss of generality).

Нека F е клас от функции $f(\boldsymbol{x}): R^n \rightarrow H$, наричани хипотези и представляващи хипотетичните взаимовръзки между y_i и $\boldsymbol{x}_i$. Класификационната задача се състои в дефинирането на подходящо хипотетично пространство F и на алгоритъм A_F, чрез който се открива функция $f^* \in F$, която да описва по оптимален начин взаимовръзката между предсказващите променливи и предсказваната променлива – класа. Вероятностното разпределение $\boldsymbol{P}_{x,y}(\boldsymbol{x},y)$ на примерите в съвкупността от данни D, дефинирано в пространството R^n xH, обикновено е неизвестно и повечето класификационни модели са непараметрични, тъй като не правят предварителни предположения за формата на разпределението $\boldsymbol{P}_{x,y}(\boldsymbol{x},y)$. Процесът е представен на Фиг.1.2 и включва три компонента – генератор на наблюдения, учител (supervisor) на класа и класификационен алгоритъм.



Фиг.1.2: Процес на обучение при класификация (Vercellis, 2008)

Задачата на генератора е да извлича случайни вектори $\boldsymbol{x}$, съгласно неизвестно вероятностно разпределение $\boldsymbol{P_x(x)}$. За всеки вектор $\boldsymbol{x}$ учителят (supervisor) връща класа, съгласно условна вероятност $\boldsymbol{P_{y|x}(y|x)}$, която също е неизвестна. Класификационен алгоритъм A_F, наричан още класификатор, избира функция $f^* \epsilon CF$ в хипотетичното пространство така, че да се минимизира подходящо избрана функция на загубите (loss function).

Част от примерите в съвкупността от данни D се използва за *обучаване*, т.е. за извличане на функционална взаимовръзка между целевата променлива и входните променливи, изразена посредством хипотезата $f^* \epsilon CF$. Останалите данни се използват за оценка точността на генерирания модел, както и за избор на най-добър модел сред генерираните с различни класификационни методи.

Следователно, създаването на класификационен модел включва три основни фази:

- *Обучение*

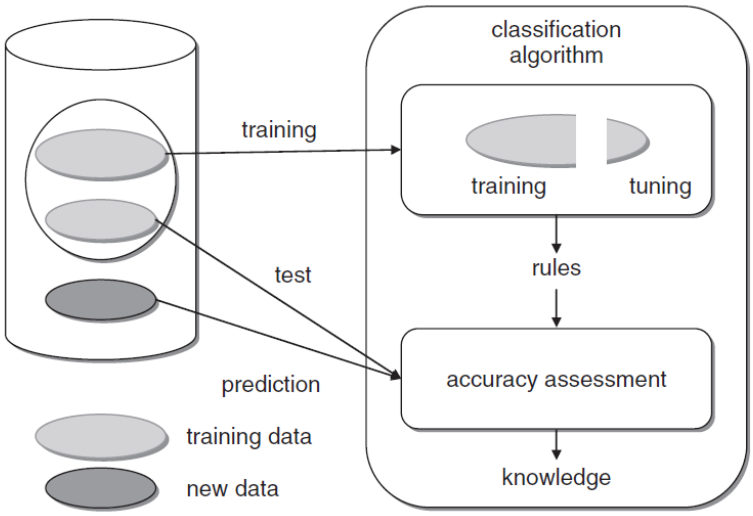
  Във фазата на обучение класификационният алгоритъм се прилага върху примерите, принадлежащи на подсъвкупност T на съвкупността от данни D, наричана обучаваща извадка, с цел извличане на класификационни правила, чрез които да се определи целевия клас y за всяко наблюдение $\boldsymbol{x}$.

- *Тестване*

  Във фазата на тестване правилата, генерирани по време на обучението, се използват за класифициране на примерите от съвкупността D, които не са били включени в обучаващата извадка и за които вече е известен класът (тестова извадка). За да се оцени точността на класификационния модел, действителният клас на всеки пример от тестовата извадка $V=D-T$ се сравнява с класа, предсказан от класификаторът. Обучаващата и тестовата извадки трябва да са напълно независими, т.е. примери от обучаващата извадка не трябва да се съдържат в тестовата извадка и обратно, за да се постигне реалистична оценка.

- *Предсказване*

  Фазата на предсказване представлява реалното използване на класификационния модел за определяне класа на нови примери, които ще бъдат регистрирани в бъдещ момент. Предсказване се получава чрез прилагане на правилата, генерирани по време на обучението, върху входните промеливи, описващи новите примери.



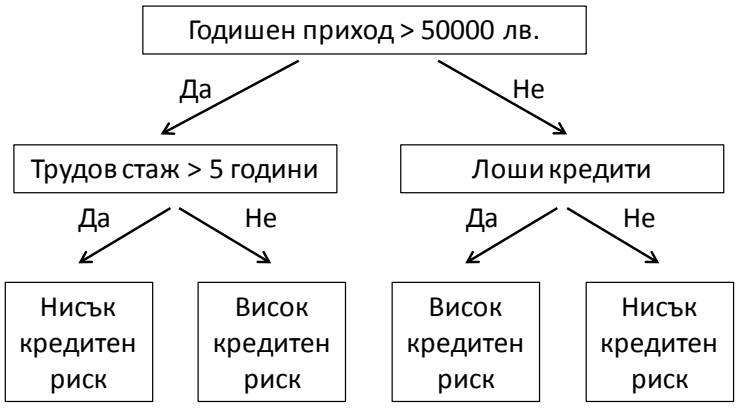
Фиг.1.3: Фази в процеса на обучение при класификационен алгоритъм (Vercellis, 2008)

1.2.3. Методи за извличане на знания от данни чрез обучение на класификатори, използвани в дисертацията

Съществува голямо разнообразие от методи за класификация. За целите на настоящия дисертационен труд са разгледани четири основни метода – „Дърво на решенията", „Генератор на правила", „К-най-близък съсед" и „Невронни мрежи".

1.2.3.1. Метод „Дърво на решенията"

Дървото на решенията представлява диаграма с дървовидна структура, при която всеки вътрешен възел (node) представлява тестова процедура за даден атрибут, всеки клон (branch) представлява резултат от теста, а крайните възли – листа на дървото (leaf nodes), представляват търсения клас. Началният възел се нарича корен на дървото (root node). Пример за дърво на решенията е показан на Фиг.1.4.



Фиг.1.4: Дърво на решенията – пример

Основните параметри на алгоритмите, чрез които се генерира дърво на решенията, са свързани с начините за избиране на най-добър разделящ тест (правило за разделяне), както и правила за спиране нарастването на дървото.

Правило за разделяне. За всеки възел на дървото се избира оптимално правило за разделяне на обектите и създаването на производни възли, включващо избор на атрибут за разделяне и стойности атрибута, по които да стане разделянето. Използват се различни критерии за избор на подходящо правило за разделяне, които се различават по броя на производните възли, броя на атрибутите и мерките за оценка на „чистотата" на създадените производни възли.

Критерии за спиране нарастването на дървото. Във всеки възел на дървото се прилагат различни критерии за спиране, за да се прецени дали нарастването на дървото трябва да продължи или възелът трябва да се счита за листо. В зависимост от използваните критерии за спиране на нарастването, при равни други условия, се получават класификационни дървета с различна топология.

Критерии за подрязване на дървото. Накрая е подходящо да се приложи някакъв критерий за подрязване на дървото, с цел да се избегне плътното прилягане на модела към обучаващите данни (overfitting). Подрязването на дървото може да се извърши по време на фазата на обучение (предварително подрязване), или след като е достигнат максималния размер на дървото (последващо подрязване). Предварителното подрязване има предимства от изчислителната гледна точка в сравнение със завършващото, но то може да доведе до прекалено рано спиране на ръста на дървото, което води до получаване на неоптимални решения. По тази причина най-често използвания метод за избягване на преспецификацията е завършващото подрязване.

1.2.3.2. Метод „Генератор на покриващи правила"

Извличането на правила се осъществява чрез генерирането на т.нар. „покриващи" правила (covering rules) – правила, които са валидни за определена част от обектите. Повечето алгоритми за извличане на правила започват работата си с откриването на „най-покриващото" правило, т.е. правилото, което е валидно за най-много обекти. След това тези обекти се премахват от изследваните данни и алгоритъмът се прилага отново с цел да се открие правилото, „покриващо" най-много обекти, и т.н. Генерираните правила на практика също разделят обектното пространство на области, подобно на дървовидните модели за предсказване. Отличават се по начина на разделяне на обектите и по графичното представяне. Извлечените правила обикновено са от типа „Ако (условие/условия), то (резултат)" (If – Then).

Един от често използваните алгоритми за извличане на правила за класификация е алгоритъмът 1R (1-rule), чрез който се генерира дърво на решенията на едно ниво, представено като съвкупност от правила, всяко от които съдържа тест на един единствен атрибут (входна променлива).

Идеята на алгоритъма 1R е следната: генерират се правила, с които се тества една единствена входна променлива и се получават съответни разклонения. Всеки „клон" съответства на различна стойност на входната променлива. За всеки „клон" се определя класа, към който принадлежат най-голям брой обекти, попаднали в това разклонение. Определянето на грешката се извършва като за всяко разклонение се преброяват обектите, които в действителност принадлежат към различен клас от определения за този „клон".

За всяка входна променлива се генерира различна съвкупност от правила, по едно правило за всяка стойност на променливата. Накрая се определя стойността на грешката за всяка

съвкупност от правила. Резултатът на класификационния алгоритъм е съвкупността от правила за тази единствена входна променлива, за която се получава най-малка грешка при класификацията.

1.2.3.3. Метод „К-най-близък съсед”

Методът „К-най-близък съсед” е известен още като Memory Based Reasoning метод, тъй като заключенията по отношение на нови обекти, които трябва да бъдат класифицирани, се правят на базата на минал опит. При този метод се използва фактът, че всеки обект може да бъде представен като точка в многомерното пространство, чрез неговия вектор - съвкупността от стойности, заемани от описващите го променливи. Алгоритъмът приема, че близостта на обектите в многомерното пространство може да се използва за дефиниране на подобността между техните вектори.

Методът „К-най-близък съсед” е data mining метод за предсказване. Предсказването се заключава в определяне стойностите на някоя от променливите (предсказвана променлива), описващи обектите от изследваната съвкупност от данни, като се използват известните стойности на останалите (предсказващи) променливи.

Прилагане на метода:

- Определя се броят К на най-близките обекти, които ще бъдат използвани при класификацията;
- Намират се най-близките К на брой съседни обекти до обекта с неизвестен клас на принадлежност, като се използва подходяща функция за измерване на разстоянието между два обекта;
- Определя се класът на новия обект на базата на заключенията, направени по отношение на намерените К най-близки съседи.

Разстоянието е начин за определяне на подобност при методите “най-близък съсед”. В случая трябва да се определя разстоянието между обекти в многомерното пространство, описвани чрез множество променливи от различен тип – цифрови и категорийни. Разстоянието между два обекта се определя по следния начин. Първо се намира разстоянието между двата обекта по отношение на всяка от променливите чрез използване на подходяща функция за разстояние. След това намерените разстояния се комбинират в една обобщена функция за разстояние, която представлява търсеното разстояние между двата обекта.

Четирите най-често използвани функции за разстояние при числови променливи са:

- Абсолютна стойност на разликата: $|A-B|$
- Квадрат на разликата: $(A-B)^2$
- Нормализирана абсолютна стойност:

$$|A-B|/(\text{максималната разлика между } A \text{ и } B) \qquad (1.5)$$

- Абсолютната стойност на разликата между стандартизираните стойности:

$$(A-\text{mean})/(\text{standard deviation})-(B-\text{mean})/(\text{standard deviation}) \qquad (1.6$$

което е еквивалентно на

$$(A-B)/(\text{standard deviation}) \qquad (1.7)$$

По-трудно е намирането на подходяща функция за разстояние при променливи от категориен тип (напр. какво е разстоянието между стойностите „мъжки” и „женски” на променливата „Пол”, или какво е разстоянието между стойностите „червен” и „син” на променливата „цвят”). Най-простата функция за разстояние при категорийни променливи е „идентично на”, която приема стойност 1, когато стойностите на изследваната променлива са еднакви за двата обекта, а стойност 0, когато стойностите на изследваната променлива са различни.

След като бъдат намерени разстоянията между стойностите на отделните променливи, те се обединяват в обща функция за разстоянието между два обекта. Най-често се използват три основни начина за сумиране:

- Сумиране (разстояние Манхатън):

$$d_{sum}(A,B) = d_{Var1}(A,B) + d_{Var2}(A,B) + ... + d_{VarN}(A,B) \quad (1.8)$$

- Нормализирано сумиране:

$$d_{norm}(A,B) = d_{sum}(A,B) / \max(d_{sum}) \quad (1.9)$$

- Евклидово разстояние:

$$d_{Euclid}(A,B) = sqrt\left(d_{Var1}(A,B)^2 + d_{Var2}(A,B)^2 + ... + d_{Var3}(A,B)^2\right) \quad (1.10)$$

След като са намерени търсените К на брой най-близки обекти, определянето на търсения клас за нов обект най-често се основава на т.нар. “демократичен принцип”. При класификация всеки от съседите гласува за своя клас. Пропорцията гласове за всеки клас съответства на вероятността класифицираният обект да принадлежи към този клас. Когато целта на задачата е да се определи точно един клас за новия обект, се избира класът, получил най-много гласове.

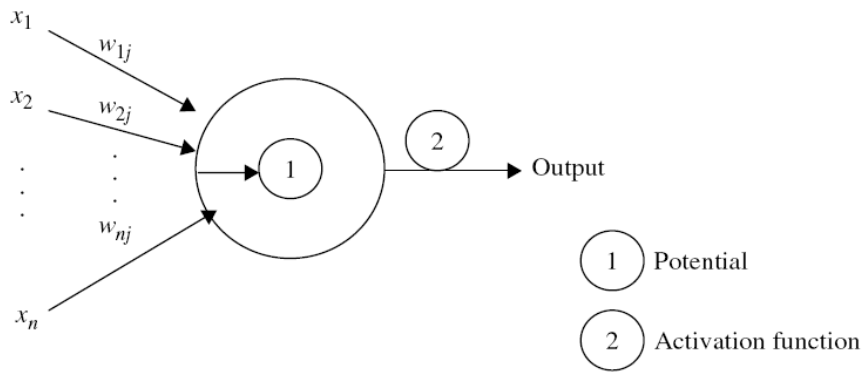
1.2.3.4. Метод „Невронни мрежи”

Невронните мрежи представляват особен интерес, защото предлагат средства за ефективно моделиране на големи и сложни проблеми, при които може да участват голям брой предсказващи променливи да съществуват множество взаимодействия. Невронните мрежи могат да бъдат използвани както за описание на данните, така и за решаване на задачи за предсказване. Първоначално са възникнали в областта на машинното обучение с цел да имитират нервно физиологичната дейност на човешкия мозък чрез комбинация от елементарни изчислителни елементи (неврони), работещи в една силно свързана система.

Невронната мрежа е съставена от множество елементарни изчислителни елементи, наречени неврони, свързани помежду си чрез претеглени връзки и организирани в слоеве. Формално процесите се описват по следния начин. Невронът j с прагова стойност θ_j, получава входни сигнали $x=[x_1,x_2,...,x_n]$ от елементите от предишния слой, с които е свързан. Всеки сигнал е свързан с тегло $w_j=[w_{1j},w_{2j},...,w_{nj}]$, определящо неговата важност.

Невронът обработва едновременно входните сигнали, техните тегла и праговата стойност чрез т.нар. комбинативна функция (combination function). Комбинативната функция произвежда стойност, наречена потенциал (potential) или нетен входен сигнал (net input).

Активираща функция (activation function) трансформира потенциала в изходен сигнал. На Фиг.1.5 схематично е представена работата на един неврон.



Фиг.1.5: Представяне действието на неврон в невронна мрежа

Комбинативната функция обикновено е линейна, затова потенциалът е претеглена сума на входните сигнали, умножени по теглата на съответните връзки. Тази сума се сравнява с праговата стойност. Потенциалът на неврон j се определя чрез следната линейна комбинация:

$$P_j = \sum_{i=1}^{n} (x_i w_{ij} - \theta_j) \tag{1.11}$$

За да се опрости изборът за потенциала, праговата стойност може да се абсорбира чрез добавяне на още един входен сигнал с постоянна стойност $x_0=1$, свързан с неврона j чрез тегло $w_{0j} = -\theta_j$:

$$P_j = \sum_{i=0}^{n} (x_i w_{ij}) \tag{1.12}$$

Изходният сигнал y_j на неврон j, се получава чрез прилагане на активиращата функция към потенциала P_j:

$$y_j = f(x, w_j) = f(P_j) = f\left(\sum_{i=0}^{n} x_i w_{ij} \right) \tag{1.13}$$

Величините x и w във функцията $f(x,w_j)$ са векторни величини.

Един от елементите, който трябва да бъде определен при дефиниране модела на невронната мрежа, е активиращата функция. Обикновено се използват три типа активираща функция: линейна (linear), стъпаловидна (stepwise) и сигмоидална (sigmoidal).

Линейната активираща функция се дефинира по следния начин:

$$f(P_j) = \alpha + \beta P_j \tag{1.14}$$

където P_j е реално число, а α и β са реални константи; в частния случай, когато α=0 и β=1, се получава идентичната функция (identity function), която обикновено се използва, когато

моделът изисква изходният сигнал на неврона да е равен на неговото активиращо ниво (неговия потенциал).

Стъпаловидната активираща функция се дефинира като:

$$f(P_j) = \begin{cases} \alpha, P_j \geq \theta_j \\ \beta, P_j < \theta_j \end{cases} \quad (1.15)$$

В този случай активиращата функция може да заема само две стойности в зависимост от това дали потенциалът превишава или не праговата стойност θ_j. При $\alpha=1$, $\beta=0$ и $\theta_j=0$ получаваме т.нар. знакова активираща функция (sign activation function), която приема стойност 0, когато потенциалът е отрицателен, и стойност +1, когато потенциалът е положителен.

Сигмоидалната, или S-образна активираща функция, е може би най-често използвана, тъй като е нелинейна и лесно разбираема. При нея винаги се получава положителен изходен сигнал в интервала [0,1]. *Сигмоидалната активираща функция* се определя като:

$$f(P_j) = \frac{1}{1+e^{-\alpha P_j}} \quad (1.16)$$

където α е положителен параметър, който определя наклона на функцията.

Невроните на невронната мрежа са организирани в слоеве. Тези слоеве могат да бъдат три вида: входящ, изходящ или скрит. Входящият слой получава информация само от външната среда, като на всеки неврон обикновено съответства входна (предсказваща) променлива. Във входящия слой не се извършват никакви изчисления, той предава информация към следващия слой. Изходящият слой произвежда крайните резултати, които се подават от мрежата навън от системата. Всеки от неговите неврони съответства на стойности на предсказваната променлива. Между входящия и изходящия слоеве може да има един или повече междинни слоеве, наречени скрити, тъй като те не влизат в директен контакт с външната среда. Тези слоеве се използват изцяло за аналитични цели, тяхната основна функция е да открият съществуващите взаимовръзки между входните и изходната променливи.

Под "архитектура" на невронната мрежа се разбира нейната организация: брой слоеве, брой елементи (неврони) принадлежащи на всеки слой, както и начина, по който са свързани елементите.

1.2.4. Оценка и сравнение на Data Mining модели за класификация (обучени класификатори)

Оценката на генерираните модели за класификация (обучени класификатори) може да се извърши чрез използването на няколко различни метода:

- *Разделяне на общата съвкупност от данни на обучаваща и тестова извадка (Holdout method)*

При този метод на оценка, обектите/записите от общата съвкупност от данни се разделят на две части – едната част се използва за обучението на модела и се нарича *обучаваща извадка*, а другата – за тестване на модела и се нарича *тестова извадка*. Обикновено, обектите в обучаващата извадка се избират на случаен принцип, останалите обекти

попадат в тестовата извадка. При някои алгоритми се постига еднакво пропорционално разпределението на обектите от различните класове в обучаващата и тестовата извадки. Големината на обучаващата извадка може да варира в зависимост от големината на общата съвкупност от данни. Най-често обучаващата извадка включва от 1/2 до 2/3 от общия брой записи в изследваната съвкупност от данни.

Точността на оценявания класификационен алгоритъм в този случай зависи от избора на тестова извадка, затова може да се получи надценяване или подценяване спрямо действителната точност на предсказване, при промяна на тестовата извадка. За да се получи по-устойчива оценка, и съответно по-надеждно да се оцени точността на класификационния модел, се използват някои от стратегиите, описани по-долу.

- *Многократно случайно избиране на обектите в обучаващата и тестовата извадки* (*Repeated random sampling*)
  При този метод процедурата по разделянето на общата съвкупност от данни на две части – обучаваща и тестова извадки, се извършва многократно. При всяко повторение двете извадки се избират на случаен принцип и се определя точността на класификация на получения модел за съответната итерация. Накрая, точността на модела се определя като средно-ариметично на получените оценки за отделните итерации.
  Броят на повторенията може да бъде определен чрез предварително прилагане на различни статистически техники като се взима предвид големината на изследваната обща съвкупност от данни.
  Този метод в много случаи се предпочита пред Holdout метода, тъй като при него може да бъде получена по-надеждна оценка за точността на класификационните модели. Но, при него не се контролира броят на попаденията на даден обект от общата съвкупност от данни в обучаващата и тестовата извадки. Това може да окаже неблагоприятен ефект върху генерираните класификационни правила и оценката на точността, особено при наличието на доминиращ обект, за който една или повече променливи приемат необичайни стойности (изключения).
- *Крос-валидиране* (*Cross-Validation*)

  Методът на крос-валидирането е разновидност на Repeated Random Sampling метода, като при него се гарантира еднакъв брой участия на всеки обект от изследваната обща съвкупност от данни в обучаващата извадка и едно единствено участие – в тестовата извадка.

  В практиката най-често се използва т.нар. tenfold cross-validation, което означава, че общата съвкупност от данни се разделя на 10 части и алгоритъмът се изпълнява 10 пъти, като всеки път 9 от 10-те части се използват като обучаваща извадка, а останалата 1 част се използва за тестова извадка.

Мерки за оценка на генерираните модели за класификация

Най-често използваните мерки за оценка на генерираните модели за класификация включват Матрица на класификация (Confusion Matrix), ROC крива, Lift диаграма и др.

Матрица на класификация (Confusion Matrix)

Предсказването на бинарна променлива (променлива, приемаща само *2* възможни стойности), е най-простия и най-често срещан случай при решаването на Data Mining

задачи за класификация. Затова именно за този случай се разглежда матрицата на класификация по-долу (матрица с размерност *2x2*). Когато предсказваната величина не е бинарна променлива (т.е. може да приема *N* различни стойности), матрицата на класификация се получава по аналогичен начин и е с размерност *NxN*.

При наличието на бинарна предсказвана променлива класификационният модел разпределя всеки обект в един от два възможни класа, например Positive и Negative, съответно вярно (True) или невярно (False). Това води до четири възможни варианта за класификация на всеки обект: Тези възможности могат да бъдат представени във вид на матрица (Фиг.1.6), наричана *Confusion Matrix* или *Contingency Table*.

		Предсказан клас	
		Positive	**Negative**
Действителен клас	**Positive**	True Positive **TP**	False Negative **FN**
	Negative	False Positive **FP**	True Negative **TN**

Фиг.1.6: Матрица на класификация (Confusion Matrix)

В елементите от главния диагонал на матрицата се намират правилно предсказаните обекти (TP и TN), а в останалите елементи на матрицата – грешно предсказаните обекти (FP и FN).

На базата на матрицата на класификация се определят множество различни мерки, които се използват за оценка на генерираните модели:

- *Точност* (*Classification Accuracy*) и *грешка* (*Classification Error*) на класификация

  При решаването на Data Mining задача за класификация е естествено да се измерва работата на класификатора чрез стойността на грешката или чрез точността на предсказване. Тези две мерки са взаимно свързани.

  Оценката се извършва чрез сравняване на предсказания клас за всеки обект от тестовата извадка с неговата действителна стойност. Ако тези две стойности съвпадат – се отчита правилно предсказване, ако не съвпадат – се отчита грешка.

  Точността на класификация се определя като отношение между броя обекти, чийто клас е правилно предсказан от класификатора, и общия брой обекти в тестовата извадка. Често се представя в %.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1.17}$$

Класификационната грешка се определя като съотношение между броя обекти, чийто клас е грешно предсказан от класификатора, и общия брой обекти в тестовата извадка. Често се представя в %.

$$Error\ Rate = \frac{FP+FN}{TP+FP+TN+FN} \tag{1.18}$$

- Точност на предсказване на всеки отделен клас

  TP Rate показва каква част са обектите, правилно класифицирани в клас Positive, от общия брой обекти, които действително принадлежат към клас Positive, т.е. показва каква част от обектите в клас Positive са разпознати от класификационния модел. Тази мярка е еквивалентна на мярката Recall (представена по-долу):

$$TP\ Rate = \frac{TP}{TP+FN} \tag{1.19}$$

  FP Rate показва каква част са обектите, които са неправилно класифицирани в клас Positive, от общия брой обекти, които в действителност принадлежат към клас Negative:

$$FP\ Rate = \frac{FP}{FP+TN} \tag{1.20}$$

  Аналогично се определят *TN Rate* и *FN Rate*:

$$TN\ Rate = \frac{TN}{TN+FP} \tag{1.21}$$

$$FN\ Rate = \frac{FN}{FN+TP} \tag{1.22}$$

- *Precision, Recall* и *F-Measure*

  Мярката *Precision* показва каква част представляват обектите, които действително принадлежат към клас Positive, от общия брой обекти, които са класифицирани от модела в клас Positive:

$$Precision = \frac{TP}{TP+FP} \tag{1.23}$$

  Мярката *Recall* е еквивалентна на мярката *TP Rate*:

$$Recall = \frac{TP}{TP+FN} = TP\ Rate \tag{1.24}$$

  Мярката *F-Measure* е комбинирана мярка, обединяваща *Precision* и *Recall*:

$$F - Measure = \frac{2}{1/Precision+1/Recall} \tag{1.25}$$

Kappa Statistic

Друга мярка, която често се използва за оценка на класификационните модели, е *Kappa Statistic*, и измерва степента на съгласуваност между предсказвания и действителния клас. Стойността на Kappa Statistic е в интервала [0;1]. Когато стойността на Kappa Statistic е 1.0, това означава, че има пълно съгласуване между предсказвания и действителния клас.

ROC крива

Receiver Operating Characteristic (ROC) кривите са визуални средства, произхождащи от теорията на комуникациите, и използвани за определяне на оптималните параметри на

филтрите, отделящи полезните сигнали от смущенията. В областта на Data Mining ROC кривите се използват за визуална оценка на точността на класификаторите, както и за сравняване на различни класификационни модели.

ROC кривата е двумерна диаграма, която се получава чрез изобразяване на True Positive Rate (TP Rate) по вертикалната ос и False Positive Rate (FP Rate) по хоризонталната ос.

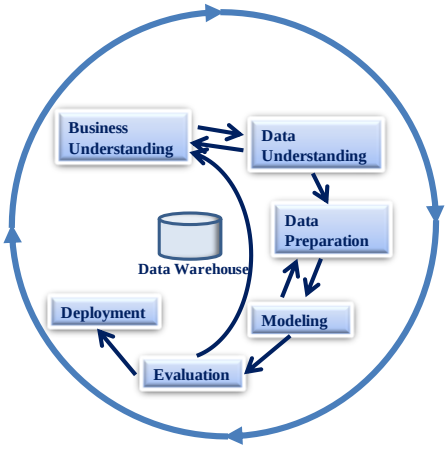
Чрез ROC кривата визуално се изобразява информацията от няколко последователно получени матрици на класификация за един и същи класификационен модел. За да се получи траекторията на ROC кривата за даден класификационен модел, трябва да се използват двойки стойности (FP Rate, TP Rate), получени за различни стойности на параметрите на модела.

Често като мярка за оценка работата на класификационните модели се използва площта над ROC кривата - *ROC Area*, приемаща стойности между 0 и 1. Колкото по-голяма е площта над ROC кривата, толкова по-добър е полученият класификационен модел. Ако ROC Area<0.5, то оценяваният модел работи по-лошо от случайния класификатор (с ROC Area=0.5).

1.2.5. Съществуващи подходи за реализация на Data Mining проекти

Извличането на знания от големи обеми данни е сложен процес, чиято успешна реализация изисква прилагането на системен подход, който да дефинира последователността от необходимите етапи и дейности, които трябва да бъдат осъществени от изследователите. Съществуващи подходи:

- ***Подходът 5A's – Assess, Access, Analyze, Act and Automate***
- **Подходът SEMMA: Sample, Explore, Modify, Model, Assess**
- ***Подходът CRISP-DM – Cross-Industry Standard Process for Data Mining*** – създаден през 1999-2000г. от консорциум от разработчици, консултанти и потребители на Data Mining софтуер – включва шест етапа и представлява цикличен процес с множество обратни връзки (Фиг.1.8).



Фиг.1.8: Подходът CRISP-DM – Cross-Industry Standard Process for Data Mining

Съгласно този модел, един цикъл на изследвания в областта на Data Mining може да се представи като процес, състоящ се от шест основни етапа:

- ***Разбирането на проблемната област*** (Business Understanding) - разбиране на целите на изследванията и формулиране на изисквания от гледната точка на потребителя.
- ***Разбирането на данните*** (Data Understanding) - събиране на данни и дейностите, целящи задълбочаване на знанията на изследователя за естеството на данните – идентифициране на проблеми, свързани с качеството на данни, получаване на първоначално мнение за характера на данните и др.
- ***Подготовката на данните*** (Data Preprocessing) включва всички дейности по създаване от първоначални "сурови" данни на крайното множество от данни (т.е. данни, които ще бъдат използвани от моделиращите средства). Включва изчистване на данните (data cleaning), интегриране на данните (data integration), трансформиране на данните (data transformation) и намаляване обема на данните (data reduction).
- ***Етапът на моделиране*** (Modeling) се състои в избора и прилагането на различни методи за моделиране, целящи извличане на закономерности от данните.
- ***Оценка на моделите*** (Evaluation) се прави с цел по-задълбочено разбиране на създадените модели от гледната точка не само на изследователя, но и на потребителя.
- ***Прилагане на моделите*** (Deployment). Готовите модели могат да се използват по два основни начина. Анализаторът може да препоръча предприемането на конкретни действия на базата на заключения от изградения модел и получените резултати, или моделът може да се приложи към други масиви от данни.

1.3. Обзор на Извличане на знания от данни в сферата на образованието (Educational Data Mining)

- Осъщественият анализ на публикуваните научни изследвания в областта на Educational Data Mining показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Резултатите от провежданите изследвания се използват за различни цели:
    - *Откриват се изоставащи студенти*, които евентуално ще отпаднат от обучение поради слабото си представяне, и на тези студенти се предоставя допълнителна помощ, за да бъдат задържани в университета.
    - *Откриват се добрите студенти* – това са студентите, които се справят най-успешно с обучението и които са най-желани за университета, и именно към такъв тип студенти се насочват бъдещите маркетингови кампании.
    - *Откриват се факторите, които в най-голяма степен влияят върху успеха* или слабото представяне на студентите, и това се превръща в ценна информация, на базата на която се взимат управленски решения за извършването на промени и за подобряване ефективността и качеството на предлаганото обучение.
- *Успехът на студентите се предсказва най-често чрез решаване на задача за предсказване, като*:
    - в повечето случаи *предсказваната величина е номинална променлива*, т.е. решава се задача за класификация
    - в по-редки случаи предсказваната променлива е цифрова, т.е. решава се задача за регресия.

- Най-често използваните методи за обучение на класификатори включват:
  - „Дърво на решенията”
  - „Невронни мрежи”
  - Байесови методи - Наивен Байесов класификатор, Байесови мрежи
  - Логистична регресия.
- Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ обучен класификатор, а не се използва един единствен алгоритъм.
- Впечатление прави и внимателното изучаване и подготовка на данните, както и изборът и форматирането на предсказваната величина.
- Често задачата за извличане на знания от данни чрез обучение на класификатори се решава за различни варианти на предсказваната променлива – дискретна величина, приемаща различен брой стойности (напр. 2,3,5,9).
- Представените резултати от изследванията обикновено показват, че при по-малък брой възможни стойности на предсказваната променлива често се получава и по-висока точност на предсказване.

1.4. Обзор на класификацията на радарни цели

- При обзора на състоянието на използваните методи за класификация на FSR цели бяха открити публикувани резултати само на представители на изследователския екип от Бирмингамския университет. В тези публикации се използват основно алгоритмите "K-най-близък съсед" (kNN) и "Невронни мрежи" (Multi-Layer Perceptron (MLP), след предварителна обработка на спектрите на сигналите от целите.
- Спектрите на сигналите от целите се обработват предварително с цел да се получат обучаващи сигнали по следния начин: първоначално спектърът на сигнала се нормира по носеща честота, по мощност и по скорост, а след това се "орязва", за да се използва най-информативната му част в ниските доплерови честоти.
- Използва се и методът Principle Component Analysis (PCA) **за** намаляване на размерността на параметричното пространство на спектрите на получените сигнали.

1.5. Цел и задачи на дисертационния труд

Настоящият дисертационен труд е посветен на изследване на приложимостта и ефективността на различни data mining методи за класификация. Изследванията са осъществени върху данни, произхождащи от две различни предметни области - данни за студенти и данни за открити морски цели.

Основните проблеми, свързани с извличане на полезна информация за обучаемите в университетите в България и разгледани в конкретния пример на УНСС, които представляват интерес за ръководството, могат да бъдат формулират както следва:

- Прогнозиране за успешно/слабо представящи се студенти в университета, в зависимост от техни персонални особености преди и след приемане във висшето учебно заведение.
- Определяне на факторите, от които най-силно зависи успеваемостта на студентите в университета.
- Подобряване на събирането и организирането на данните за студентите, с цел по-добро управление на качеството на обучение.

Решаването на тези проблеми може да бъде подпомогнато чрез използване на подходящи Data Mining методи и средства за анализ на наличните данни за студентите, и по-конкретно чрез решаване на Data Mining задача за класификация.

Важен проблем, който възниква при осъществяване на защита на морски граници, е определянето на вида на нарушителя (класа на плавателен съд). За решаването на този проблем също се предполага, че могат да бъдат използвани Data Mining методи за класификация.

Целта на настоящия дисертационен труд е да се генерират и изследват Data Mining модели за класификация, за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, и за класификация на морски обекти.

За постигане на посочената цел се поставят ***следните основни задачи***:

- Разработване на методика за решаване на задача за извличане на знания от данни (Data Mining) чрез обучение на класификатори.
- Избор на подходящи софтуерни средства за реализация на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори.
- Генериране на Data Mining модели за класификация (обучени класификатори) чрез избрани алгоритми.
- Оценяване на генерираните Data Mining модели за класификация (обучени класификатори) чрез избрани мерки за оценка.
- Сравняване на изследваните Data Mining модели за класификация (обучени класификатори).

1.6. Изводи по Първа Глава

В настоящата глава е постигнато следното:

- Направен е обзор на областта Извличане на знания от големи обеми данни (Data Mining), като е акцентирано върху задачата за класификация - основна задача, решавана в дисертационния труд чрез Data Mining методи и средства.
- Представени са четири метода - „Дърво на решенията", „Генератор на правила", „К-най-близък съсед" и „Невронни мрежи", които се използват за генериране на Data Mining модели за класификация (обучени класификатори) в дисертацията.
- Предложени са различни мерки за оценка и сравнение на генерираните модели за класификация (обучените класификатори).
- Разгледани са различни подходи за реализация на Data Mining проекти.
- Направен е обзор на състоянието на Извличането на знания от големи обеми данни в областта на образованието (Educational Data Mining - EDM).
- Направен е обзор относно използваните методи за класификация на морски цели.
- Формулирани са целта и задачите на дисертацията.

Глава 2: Методика за провеждане на изследването. Подготовка на данните

Във ***Втора глава*** е представена методиката за провеждане на изследването, включваща избор на подход за реализация на Data Mining проекта, избор на подходящи софтуерни средства за осъществяване на поставените задачи, и описание на конкретните дейности, които ще бъдат осъществени в рамките на изследването. Описани са и извършените дейности, свързани с предварителна подготовка и изследване на двете съвкупности от данни – за студенти и за морски цели, които се използват при генерирането на Data Mining моделите за класификация (обучени класификатори).

2.1. Методика за провеждане на изследването

2.1.1. Избор на подход за реализация на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори

Решаваната научна задача в този дисертационен труд е задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори. Затова, за целите на проведеното научно изследване в настоящия дисертационен труд е избран подходът CRISP-DM (Cross-Industry Standard Process for Data Mining), по-подробно описан в т.1.2.5, който се приема за стандартен подход за реализация на Data Mining проекти от специалистите в тази научна област, неутрален е по отношение на областите на приложение и е най-често използвания подход за провеждане на Data Mining анализи през последните години.

Извличането на знания от двете съвкупности от данни (университетски данни и FS радиолокационни данни за движещи се морски обекти), избрани за провеждане на научните изследвания, е извършено чрез последователно преминаване през петте основни етапа на CRISP-DM процеса - Business Understanding, Data Understanding, Data Preprocessing, Modeling and Evaluation, не са изпълнени само дейностите предвидени в последния етап – Deployment, тъй като те са свързани с конкретното практическо прилагане на получените аналитични резултати в производствена среда (такава не е използвана за целите на настоящия дисертационен труд).

2.1.2. Избор на софтуерни средства за провеждане на изследването

За осъществяването на научните изследвания, представени в настоящия дисертационен труд, се използват различни софтуерни решения на различни етапи:

- ***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.
- ***WEKA*** (v3.6) - за моделиране и оценка на получените резултати.

2.1.3. Описание на методиката на изследването

Научните изследвания, представени в настоящия дисертационен труд, са осъществени в следната последователност:

- ***Проучване на литературни източници*** в ***областта Data Mining*** (книги, научни публикации, описания на софтуерни средства), ***с цел избор и формулировка на решавните задачи, методи, класификационни модели, програмни продукти.***
- ***Подбор на данни от избраните приложни области и изучаване на тяхната същност, произход, начин на събиране, организиране и съхранение***.
- ***Изучаване и предварителна подготовка на наличните съвкупности от данни***
  - ***Визуално изследване на данните*** чрез софтуерните средства Excel, QlikView и WEKA.
  - ***Изследване индивидуалното влияние на всяка от входните променливи върху изходната/предсказваната променлива (класа)***.
- ***Разработване на модели за класификация чрез прилагане на избрани Data Mining алгоритми върху данните***.
- ***Оценка на разработените модели за класификация (обучени класификатори)***

  Получените модели за класификация на изследваните съвкупности от данни се оценяват на базата на резултатите, получени от работата на всеки от избраните алгоритми, по критерии и метрики за оценка на качеството на генерираните класификационни правила или модели (по-подробно описани в т.1.2.4).

- ***Формулиране на изводи и препоръки към крайния потребител на резултатите от Data Mining анализите***.

2.2. Подготовка на данните

Научните изследвания, представени в настоящия дисертационен труд, са осъществени за две различни съвкупности от данни – данни за студенти и данни за сигнали от радарни морски цели. Научните резултати за двата вида данни са получени чрез използване на една и съща методика, описана в т.2.1.3.

2.2.1. Анализ на данните за студенти

2.2.1.1. Запознаване с университетските бази данни за студентите

На базата на извършените на този етап дейности са дефинирани целите, които трябва да бъдат постигнати от гледна точка на потребителите – в случая ръководителите на университета. Тези цели впоследствие са трансформирани в научни задачи, които се решават чрез прилагане на избрани data mining методи и средства. Целите и задачите на научното изследване са представени в Глава 1, т.1.5.

2.2.1.2. Избор на данни за изследването

В резултат от извършените дейности се стига до следните изводи – университетските данни се съхраняват в две отделни релационни бази данни – в едната се съдържат данните, свързани с провеждането на кандидат-студентските кампании, а в другата – данните за обучението на приетите студенти.

Резултатът от извършените на този етап дейности е изборът на съвкупност от данни, които да бъдат използвани за целите на провежданите научни изследвания. Тъй като избраните атрибути на данните се съдържат в двете отделни бази данни, крайната съвкупност от

данни е получена чрез екстракция на необходимите параметри от двете бази и интегрирането им в един общ файл в Excel формат.

2.2.1.3. Предварителна подготовка на данните

Предоставената от университета съвкупност от данни съдържа 10330 записа за студентите в УНСС, приети по време на кандидат-студентските кампании (КСК) през 2007, 2008 и 2009 години. Информацията за всеки студент в оригиналните данни съдържа следните 20 променливи (атрибути).

Извършени са различни преобразувания на оригиналните данни с цел, от една страна, да се повишат възможностите за извършване на по-задълбочени анализи, които да подпомагат реалните потребители при взимане на решения (в случая, ръководството на университета), и от друга страна, да се използват по-пълноценно аналитичните възможности на използваните BI и Data Mining методи и средства.

А. Премахване на променливи или стойности на променливи

Част от променливите, съдържащи се в оригиналните данни, не представляват интерес за целите на изследването в настоящата дисертация. Например, променливите „Държава" и „Вид образование, получено до момента на кандидатстване" приемат по една единствена стойност за всички записи (съответно, „България" и „Средно"), тъй като данните са за българските студенти в УНСС и всички те са завършили средно образование до момента на кандидатстване.

Б. Преобразуване на променливи от тип „свободен текст"

Една от основните стъпки при предварителната подготовка на предоставените данни е обработката на променливите от типа „свободен текст" (т.е. стойностите, които приемат тези променливи, не са предварително стандартизирани и представляват свободен текст, въвеждан от служителите при приемане на документите на кандидат-студентите), които се считат за особено важни за осъществяване на предвижданите анализи. Такъв тип данни трудно могат да бъдат анализирани така, че да се получат смислени резултати от анализите. Поради това е извършена предварителна обработка на стойностите на тези параметри – свободният текст е заменен с определен брой стойности, извлечени по смисъл от наличния текст. Такъв вид преобразуване е извършено за променливите „Профил на училището" и „Селище, в което се намира учебното заведение, където е получено образованието".

Променливата „Профил на училището", величина от типа „свободен текст" в оригиналните данни, е трансформирана в номинална променлива, приемаща краен брой стойности, съответстващи на различните профили на средните училища в България. При преобразуването е взето предвид профилирането на училищата, в които получават средното си образование учениците в България (непрофилирани СОУ и профилирани училища, в които се извършва прием след завършване на 7 и 8 клас), като информация е взета основно от интернет страницата на Министерството на образованието, младежта и науката (www.minedu.government.bg).

Променливата „Селище на образование", съдържаща наименования на различни селища в България, е преобразувана във величина с краен брой стойности. Първоначално са запазени стойностите, съответстващи на 12-те най-големи града в България (София,

Пловдив, Варна, Бургас, Русе, Велико Търново, Сливен, Стара Загора, Шумен, Добрич, Благоевград, Плевен), а останалите са заменени с нови 7 стойности, съответстващи на 7 региона в България - София-област, северозападен (СЗ), северен централен (СЦ), североизточен (СИ), югозападен (ЮЗ), южен централен (ЮЦ) и югоизточен (ЮИ), на базата на принадлежността на селищата към съответните региони на страната (използвана е основно информацията, публикувана на интернет страницата http://bg.guide-bulgaria.com/). В последствие, след като при анализирането на данните се установява, че големият брой стойности, които заема тази променлива, не водят до съществени изводи, техният брой се намалява до 7 – София и шестте региона в страната (ЮЗ, ЮЦ, ЮИ, СЗ, СЦ, СИ).

Преобразувания се извършват и на променливата „Наименование на Направление/Поднаправление/Специалност". Оригиналните данни са взети в началото на 2011г., което означава, че студентите, приети през периода 2007-2009г., в този момент са в различни семестри на обучение (1-10 семестър, в зависимост от направление, прекъсване на обучението и др.). Специфичното в случая е, че студентите в УНСС първоначално (1-4 семестър) се приемат в съответното направление или поднаправление, а в последствие се разпределят по специалности. Поради това, в данните се съдържат наименования както на направления и поднаправления, така и на конкретни специалности, което е причина за наличието на голям брой различни стойности на променливата „Наименование на Направление/Поднаправление/Специалност", което силно затруднява анализите. За да се получат по-смислени анализи, в крайната съвкупност от данни стойностите на тази променлива са редуцирани. Запазени са стойностите на 5 от 6-те професионални направления - „Социология, антропология и наука за културата", „Политически науки", „Обществени комуникации и информационни науки", „Право" и „Администрация и управление", а шестото направление „Икономика" (2/3 от общата съвкупност от данни се отнася за студенти в това направление) е представено чрез своите 5 поднаправления - „Икономика и бизнес", „Приложна информатика, комуникации и икометрия", „Икономика с чуждоезиково обучение", „Икономика с преподаване на английски език", „Финанси, счетоводство и контрол".

В. Създаване на нови променливи

При изследването и подготовката на данните често се стига до необходимостта от създаването на нови променливи, които се считат важни за извършването на предвидените анализи. В случая, за целите на анализа в настоящата дисертация е създадена новата променлива „Възраст на кандидат-студентите", която е формирана като разлика между величините „Година на приемане в УНСС" и „Година на раждане".

Г. Преобразуване на количествени в качествени променливите

При решаването на дефинираните data mining задачи и при прилагането на избраните data mining методи и средства се извършват и някои допълнителни преобразувания на променливите. Например, част от променливите заемат числови стойности, но на практика не носят числов смисъл (т.е. не е логично да участват в различни математически изчисления), затова те се трансформират от числови в номинални променливи (напр. „Година на приемане в УНСС", „Семестър", „Възраст"), тъй като са много по-информативни при извършването на анализите.

Една от целите в настоящата дисертация е да се генерират и изследват Data Mining модели за класификация, за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета. Постигането на тази цел се осъществява чрез решаването на data mining задача за класификация, а при класификацията изходната (целевата) променлива, т.е. променливата, която ще бъде предсказване, е дискретна величина. Затова, оригиналната променлива „Среден успех от I курс с 2-ки", която е непрекъсната числова величина, е преобразувана в променлива с краен брой дискретни стойности. Преобразуването е извършено по два начина – в единия случай дискретната величина заема 5 различни стойности, а във втория случай – 2 различни стойности, като е изследвано поведението на приложените data mining алгоритми при тези различни условия.

Преобразуването на непрекъсната числова променлива „Среден успех от I курс с 2-ки" в дискретна променлива с 5 стойности е извършено въз основа на възприетата в българското образование шестобална система за оценяване, както е показано в Табл.2.1:

Таблица 2.1: Преобразуване на променливата „Среден успех от I курс с 2-ки" в дискретна променлива с 5 стойности

Оригинални стойности на променливата „Среден успех от I курс с 2-ки"	**Нови стойности на променливата „Среден успех от I курс с 2-ки"**
[0.00;2.49]	Слаб (Bad)
[2.50;3.49]	Среден (Average)
[3.50;4.49]	Добър (Good)
[4.50;5.49]	Много добър (Very Good)
[5.50;6.00]	Отличен (Excellent)

Преобразуването на непрекъсната числова променлива „Среден успех от I курс с 2-ки" в дискретна променлива с 2 стойности се извършва с цел да се провери поведението на избраните data mining алгоритми при наличието на двоична изходна променлива. В този случай изходната променлива представлява величина с две отчетливи стойности, т.е. по този параметър студентите се разделят на две групи (два класа) – „силни" и „слаби". Праговата стойност за формиране на тези два класа е 4.50, както е показано в Табл.2.2:

Таблица 2.2: Преобразуване на променливата „Среден успех от I курс с 2-ки" в дискретна променлива с 2 стойности

Оригинални стойности на променливата „Среден успех от I курс с 2-ки"	**Нови стойности на променливата „Среден успех от I курс с 2-ки"**
[0.00;4.49)	Слаби (Weak)
[4.50;6.00]	Силни (Strong)

Д. Други преобразувания на променливите

При изследването на данните, проведено в дисертацията, са открити някои грешки (наличие на стойности, които са извън допустимия интервал за съответната променлива), които са коригирани след консултиране с представителите на университетската администрация, отговорни за данните.

При прилагането на data mining алгоритмите са извършвани и други преобразувания на основната съвкупност от данни, използвани при проведените изследвания, които са описани при представянето на резултатите (Глава 3).

Друго преобразуване на данните, което е извършено поради спецификата на данните и използваните софтуерни средства, е замяната на оригиналните текстови стойности на променливите, на български език, с техните еквиваленти - на английски език, за да е възможно разчитането на получените резултати, а и за да бъдат използвани резултатите при разработените научни публикации.

2.2.1.4. Формиране на съвкупност от данни за изследванията

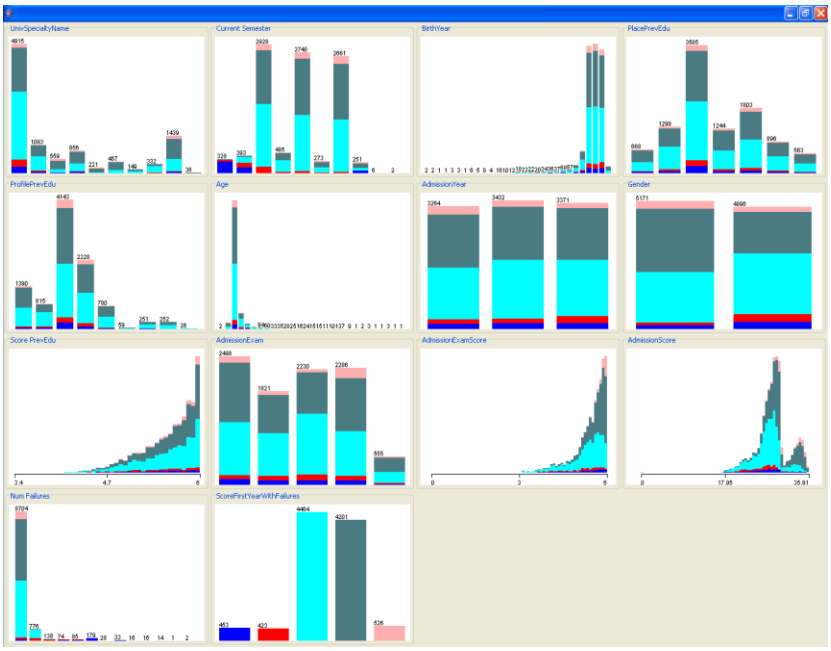
Като резултат от описаните по-горе преобразувания на оригиналните данни, се достига до крайната съвкупност от данни, използвана при проведените изследвания, представени в настоящия дисертационен труд.

Съвкупността от данни, използвана при анализите, съдържа 10067 записа (т.е. данни за 10067 студента, приети в УНСС през периода 2007-2009г.) и 14 атрибута (променливи, които характеризират всеки от приетите студенти), и е представена в обобщен вид в Табл.2.3 и на Фиг.2.3.

Таблица 2.3: Обобщено описание на данните, използвани при анализите

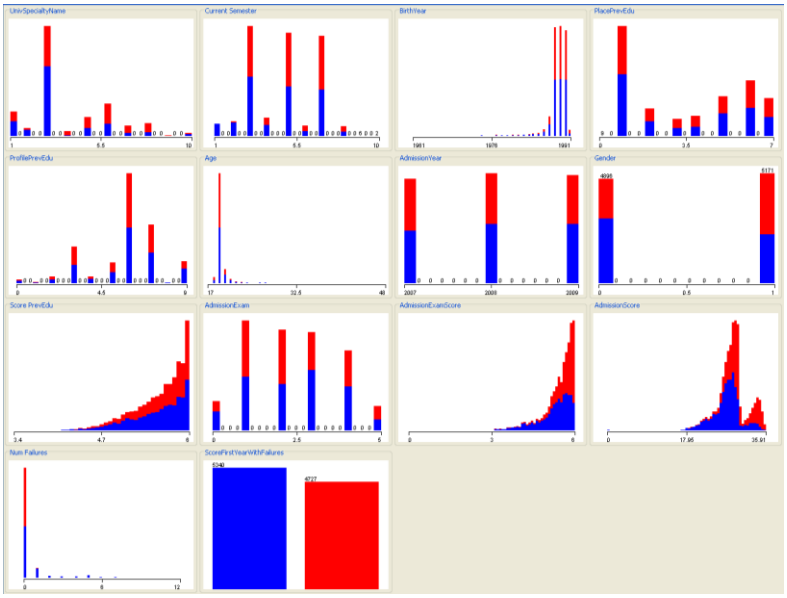
No.	Име на Атрибут	Тип променлива	Разпределение по стойности	Липсващи стойности
1	Наименование на направление/поднаправление	Номинална	10 различни стойности – съответстващи на 5 професионални направления и 5 поднаправления в УНСС	0 (0%)
2	Семестър	Номинална	10 различни стойности – 1,2,3,4,5,6,7,8,9,10	0 (0%)
3	Година на раждане	Номинална	29 различни стойности в интервала 1961-1991	0 (0%)
4	Местоположение (средно образование)	Номинална	7 различни стойности	9 (0%)
5	Профил (средно образование)	Номинална	9 различни стойности	123 (1%)
6	Възраст	Номинална	29 различни стойности в интервала 17-48	0 (0%)
7	Година на приемане в УНСС	Номинална	3 различни стойности – 2007, 2008, 2009	0 (0%)
8	Пол	Номинална	2 различни стойности – м/ж	0 (0%)
9	Среден успех (средно образование)	Числова	Min=3.4, Max=6, Mean=5.559, StdDev=0.391	219 (2%)
10	Изпит на приемане при КСК	Номинална	5 различни стойности	677 (7%)
11	Оценка от изпита на приемане	Числова	Min=0, Max=6, Mean=5.431, StdDev=0.55	677 (7%)
12	Общ бал при приемане	Числова	Min=0, Max=35.91, Mean=28.282, StdDev=3.41	682 (7%)
13	Брой невзети изпити	Номинална	13 различни стойности – 0-12	1 (0%)
14	Среден успех от I курс в УНСС	Номинална	1 Вариант: 5 различни стойности – слаб, среден, добър, много добър, отличен 2 Вариант: 2 различни стойности – слаби и силни	0 (0%)

На Фиг.2.3 е представено разпределението на 14-те атрибута, които се съдържат в данните, по отношение на петте стойности на изходната променлива (Вариант 1) - „Среден успех от I курс” (слаб, среден, добър, много добър, отличен – изобразени с различни цветове).



Фиг.2.3: Разпределение на атрибутите в данните относно 5-те стойности на предсказваната променлива в WEKA

На Фиг.2.5 е показано разпределението на атрибутите в данните относно 2-те стойности на предсказваната променлива.



Фиг.2.5: Разпределение на атрибутите в данните относно 2-те стойности на предсказваната променлива в WEKA

Изводи:

- Получена е крайна съвкупност от данни за студентите, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.

- Съвкупността от данни съдържа 10067 записа за студенти, описани чрез 14 атрибута (параметъра), 11 качествени променливи (nominal variables)и 3 числови непрекъснати променливи (numeric variables).

- Много малко са липсващите стойности (за 3 от атрибутите - около 7%, за други 2 атрибута - съответно 2% и 1%), което означава, че това няма да се отрази съществено върху резултатите от изследванията и затова не е необходимо да се предприемат действия по отношение попълването на тези липсващи стойности.

- Изходната/предсказваната променлива „Среден успех от I курс" е преобразувана в номинална променлива по два начина (Вариант 1 - номинална променлива с пет стойности - слаб, среден, добър, много добър, отличен; Вариант 2 - номинална променлива с 2 стойности - слаби и силни).

2.2.2. Анализ на данните за класификация на морски цели

Изходните данни са събрани от изследователския екип на Бирмингамския университет в периода Февруари-Март 2010г. Параметрите, които характеризират всеки запис на радарна морска цел, включват както цифрови, така и номинални променливи, и описват различни аспекти - разстояние между радарите в използваната тестова радарна система, параметри на антената, метеорологични характеристики като скорост и посока на вятъра, записи на Доплеровия сигнал получен при преминаването на лодката през базовата линия между приемника и предавателя на радара.

Суровите данни, или записите на Доплеровия сигнал за засечените морски цели, събирани по време на реалните изпитания с разработената за целта FS радарна система, са обработени допълнително така, че от тях да се получи съвкупност от данни, подходяща за Data Mining анализи. Смисълът на тази предварителна обработка на сигнали от движещи се морски цели се състои в това, че по наличните записи на FS доплерови сигнали, получени при пресичането на базовата линия между радарите, се извършва предварително задачата на автоматично откриване на целта и последваща оценка на информативни параметри на сигнала или целта, необходими за последващото й класифициране с методите на Data Mining. Тези параметри на сигналите са: дължина на целта и енергия на отразения Доплеров сигнал - косвено определяща общата големина на морския обект. Времетраенето на отразения от лодките сигнал се изчислява като произведение на времевия интервал между импулсите, облъчващи целта, и техния брой.

Параметрите на сигнала са измерени на изхода на оригинална структура на MTI CA CFAR K/M-L процесор, работещ във времевата област, разработена от екипа на българските изследователи.

Така извлечените данни са организирани в плосък файл – в случая Excel файл, тъй като този формат е подходящ за работа с избрания Data Mining софтуер WEKA, а и данните са ограничени. В случай, че радарната система работи в реални условия, количеството

събирани данни ще бъде доста голямо, което ще доведе до необходимост от организирането им в база от данни.

Съвкупността от данни, с която са проведени Data Mining анализите, съдържа 80 записа на радарни морски цели, описани чрез 16 параметъра (включващи и предсказваната величина), представени в Табл.2.4.

Таблица 2.4: Обобщено описание на данните, използвани при класифицирането на радарни морски цели

Variable Name	VarType	Values	Missing
Trial Date	Nom	17/02/10 (43), 18/02/10 (10), 21/03/10 (14), 23/03/10 (13)	0 (0%)
Distance Between Radars	Num	Min=300m, Max=500m, Mean=330.6m, StdDev=57.22 (300m, 316m, 500m)	0 (0%)
Antenna	Nom	A1/2/V/A2/1/V; A1/3/H/A2/1/H	0 (0%)
Weather	Nom	Sunny (56), Gloomy (11), Raining (13)	0 (0%)
Wind Speed	Num	1 - 5.1 m/s	2 (3%)
Wind Direction	Nom	SE (42), S (22), NW (1), SW (10), W (3)	2 (3%)
Boat Direction	Nom	South (11), North (12)	57 (71%)
S/N Ratio Before PC	Num	0 – 93.04, Mean=37.246, StdDev=26.207	12 (15%)
S/N Ratio After PC	Num	0 – 65.59, Mean=17.937, StdDev=22.605	14 (18%)
Number of Pulses Before PC	Num	0 – 2557, Mean=1148.892, StdDev=720.049	15 (19%)
Number of Pulses After PC	Num	0 – 4361, Mean=863.424, StdDev=1113.801	14 (18%)
Correlation Before PC	Num	0.62 – 1, Mean=0.954, StdDev=0.099	17 (21%)
Correlation After PC	Num	0.008 – 0.982, Mean=0.676, StdDev=0.322	18 (23%)
Energy Before PC	Num	0 – 2.939, Mean=0.599, StdDev=0.712	14 (18%)
Energy After PC	Num	0 – 0.499, Mean=0.024, StdDev=0.067	13 (16%)
Target	Nom	BigBoat (11), MISL_Boat (62), AverageBoat (6)	1 (1%)

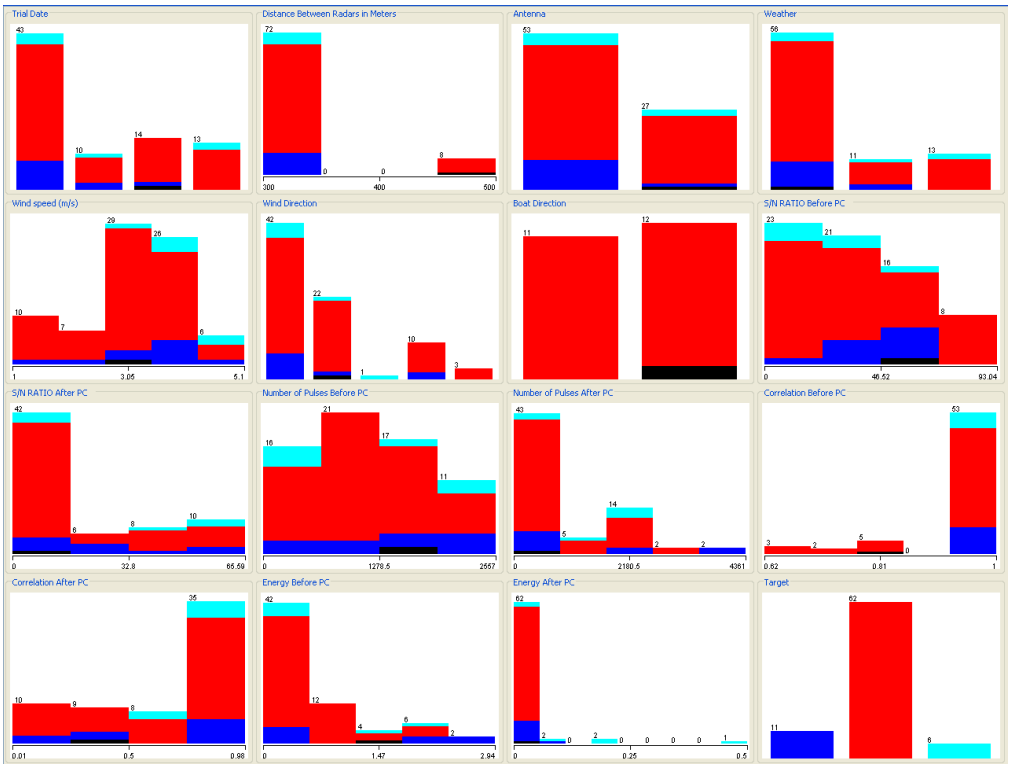
Както се вижда от Табл.2.4, за голяма част от параметрите има много липсващи стойности. Това се дължи както на липсваща информация в самите сурови данни от изпитанията, както и на затруднения при измерването на някои от параметрите.

Предсказваната целева променлива е *вида на откритата радарна цел*, който трябва да бъде определен чрез решаването на задачата за класификация. Оригиналните данни от изпитанията съдържат записи за 14 различни морски цели. Тъй като наличният обем от данни е много ограничен (80 записа), е счетено за целесъобразно тези 14 морски цели да бъдат обединени в три класа и така е получена предсказвана променлива от категориен тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*.

В класа *MISL Boat* попадат записите на малка гумена моторна лодка, която изследователският екип от Бирмингамския университет използва за реалните изпитания.

В класа *Big Boat* попадат записите за по-големи лодки, които са били регистрирани по време на изпитанията и които в оригиналните данни фигурират като *Big boat*, *Big fishing boat*, *Big white boat*, *Huge boat*.

В класа *Average Boat* попадат записите за средно големи лодки, фигуриращи в оригиналните данни като *Life boat*, *Police boat*, *Speed boat*, *Medium yacht*, *Fast red boat*. Разпределянето на записите в избраните три класа е извършено чрез консултации с членовете на екипа, участващи в реалните изпитания.



Фиг.2.33: Разпределение на атрибутите в данните относно стойностите на предсказваната променлива в WEKA

На Фиг.2.33 е представено разпределението на 16-те атрибута (променливи), които се съдържат в данните, по отношение на трите стойности на изходната предсказвана променлива (Target – вида на откритата морска цел) – *MISL Boat* (с червен цвят), *Big Boat* (тъмно син цвят), *Average Boat* (светло син цвят).

Броят на записите в трите класа е различен. Най-много данни са налични за клас *MISL Boat* (62 записа), а доста по-ограничен е броят са записите за клас *Big Boat* (11) и клас *Average Boat* (6). За целевата променлива има само една липсваща стойност (Target Variable - 1% Missing Values - виж Табл.2.4), затова общият брой на записите в изследваната съвкупност от данни е 79.

Изводи:

- Получена е крайна съвкупност от данни за морски цели, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.
- Съвкупността от данни съдържа 80 записа за радарни морски цели, описани чрез 16 параметъра (включващи и предсказваната променлива).

- За голяма част от параметрите има много липсващи стойности, което се дължи както на липсваща информация в самите сурови данни от изпитанията, така и на затруднения при измерването на някои от параметрите.
- Предсказваната целева променлива е вида на откритата радарна цел - променлива от номинален тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*. Броят на записите в трите класа е различен. Най-много данни са налични за клас *MISL Boat* (62 записа), а доста по-ограничен е броят са записите за клас *Big Boat* (11) и клас *Average Boat* (6).

2.3. Изводи по Втора Глава

В настоящата глава е постигнато следното:

- Избран е CRISP-DM подход за решаването на избраната Data Mining задача.
- Разгледани са различни софтуерни средства, които биха били подходящи за осъществяване на формулираните задачи. Избрани са *WEKA*, *QlikView* и *Excel*, които се използват за предварителна подготовка на данните и генериране на Data Mining модели за класификация (обучение на класификатори).
- Разработена е методика за решаване на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори. Избрани са мерки за оценка и сравнение на генерираните Data Mining модели за класификация (обучените класификатори).
- Съгласно по-горе посочените подход и методика, е осъществена задачата за предварителна подготовка и изследване на данните за студентите,
- Получена е крайна съвкупност от данни за студентите, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.
- Получена е крайна съвкупност от данни за морски цели, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.

Глава 3: Резултати от изследването на Data Mining моделите за класификация

В ***Трета глава*** са представени резултатите от изследването, осъществено чрез прилагане на избраните Data Mining алгоритми за класификация (обучените класификатори) 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.

3.1. Генериране и оценка на Data Mining класификационни модели (обучени класификатори) за предсказване успеваемостта на студентите

3.1.1. Класификационни алгоритми, реализирани в Data Mining софтуер WEKA

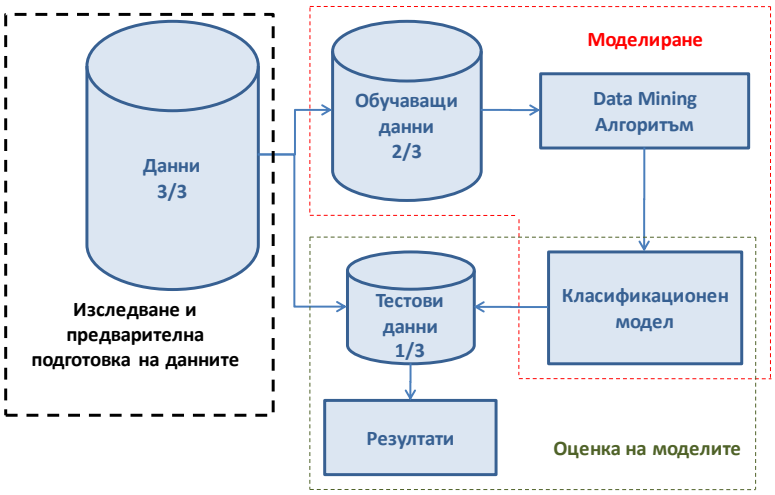
За решаването на Data mining задачата за класификация са използвани различни алгоритми – генератор на правила (Rule Learner), дърво на решението (Decision Tree), невронна мрежа (Neural Network) и К-най-близък съсед (K-Nearest Neighbour).

Тези алгоритми са избрани, тъй като всеки от тях работи на различен принцип, т.е. по различен начин генерира модела за предсказване на целевата променлива. Затова, прилагането на тези алгоритми върху едни и същи данни може да доведе до получаването на различни резултати – в някои случаи определени алгоритми дават по-добри резултати, а в други случаи – други алгоритми се справят по-добре, по отношение предсказването на класа (това действително се вижда от получените резултати в представеното изследване). При решаването на data mining задачи винаги се препоръчва да не се разчита на прилагането на един единствен метод, дори когато той дава добри резултати, а да се сравни работата на няколко различни метода и да се избере оптималното решение за последващо прилагане върху нови данни. Друга причина за избора на тези алгоритми е тяхното често използване при решаването на различни Data Mining задачи в сферата на образованието (виж т.1.3).

Работата на избраните data mining алгоритми за класификация се изследва за два варианта на целевата променлива (с 5 и с 2 стойности), тъй като много често точността на алгоритмите зависи от броя на стойностите, които заема целевата променлива, което е пряко свързано и с представителността на всеки от предсказваните класове в общата съвкупност от данни.

Data mining анализите в дисертацията се реализират със софтуерния пакет WEKA, предлагащо богато разнообразие от Data Mining алгоритми. Избраните data mining алгоритми за класификация са приложени върху наличните данни при различни условия (test options): “percentage split” и „stratified 10-fold cross validation”.

При “percentage split test option” (holdout method - разделяне на общата съвкупност от данни на обучаваща и тестова извадка), 2/3 (66%) от данните се използват за обучаваща извадка, а останалата 1/3 (34%) – за оценка на генерирания модел. Тестовата схема “percentage split” (66%/34%) е представена на Фиг.3.1.



Фиг.3.1: Тестова схема "Percentage Split"

При "stratified 10-fold cross validation" общата съвкупност от данни се разделя на 10 части и алгоритъмът се прилага 10 пъти, като всеки път 9 от 10-те части се използват като обучаващи данни (training data), а останалата 1 част се използва за оценка на модела. Това е стандартен подход, който се прилага за определяне грешката на предсказване на класификационните алгоритми при наличието на една постоянна съвкупност от данни. Стратификацията на данните означава, че съвкупността от данни ще бъде разделена произволно на 10 части, но във всяка част всеки клас ще е представен приблизително в същите пропорции, както в цялата съвкупност от данни. Крайната грешка на алгоритъма се определя като средно аритметично от грешките, получени при 10-те изпълнения на алгоритъма.

За целите на изследването в дисертацията, за целева променлива (променлива, която ще бъде предсказвана) е избрана величината „Среден успех от I курс". В оригиналната версия на наличните университетски данни тази променлива е от числов тип, затова е извършено нейното предварително преобразуване във величина от номинален тип (дискретна качествена променлива). Това преобразуване е осъществено в два варианта (описано по-подробно в т.2.2.1.3.Г):

- ***Вариант 1***: целевата променлива „Среден успех от I курс" заема 5 различни стойности (слаб, среден, добър, много добър, отличен)
- ***Вариант 2***: целевата променлива „Среден успех от I курс" заема 2 различни стойности (слаби, силни)

Останалите променливи, които се съдържат в данните, се разглеждат като входни или предсказващи променливи (описание на променливите е дадено в т.2.2.1.4).

Работата на така избраните data mining алгоритми за класификация се изследва в дисертацията за двата варианта на целевата променлива (с 5 и с 2 стойности), тъй като понякога точността на алгоритмите се влияе от броя на стойностите, които заема целевата променлива, което е пряко свързано и с представителността на всеки от предсказваните класове в общата съвкупност от данни.

Реализацията на класификационната задача в двата случая, за двата варианта на предсказваната променлива, позволява да се изследва устойчивостта на предсказващия

модел към промяна на данните, и/или да се избере по ефективно класификационно правило.

3.1.2. Резултати за предсказвана променлива с пет стойности

Предсказвана променлива е от *категориен тип* и заема *5 различни стойности* – „Слаб" (Bad), "Среден" (Average), „Добър" (Good), „Много добър" (Very Good) и „Отличен" (Excellent), създадени на базата на правилата в използваната шестобална система за оценяване в България (подробно обяснение е дадено в т.2.2.1.3.Г).

Използвани са четири Data Mining метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk).

Генерираните модели за класификация (обучени класификатори) са оценени и сравнени чрез получените стойности на избрани параметри за оценка (по-подробно описани в т.1.2.4 и т.2.1.3). Оценено е и подобрението по отношение точността на предсказване, което се получава чрез генерираните класификационни модели с избраните алгоритми, спрямо наивната класификация (Naïve Classification) - класификация на случаен принцип без използване на модел за предсказване, изготвен на базата на обучаваща извадка, чрез параметъра *Model/Naïve Correct/Incorrect Ratio*. Това се прави с цел да се получи реална оценка за качеството на получените модели за класификация (обучени класификатори), т.е. за да се види реалното подобряване на класификацията чрез генерираните модели.

Сравнение на получените модели за класификация

В Табл.3.4 са представени резултатите, получени при прилагането на избраните четири различни класификационни алгоритми върху изследваната съвкупност от данни (изходната променлива е номинална от тип категория - с 5 различни стойности) – генератор на правила 1R (WEKA OneR), дърво на решенията (WEKA J48), невронна мрежа (WEKA MultilayerPerceptron) и К-най-близък съсед (kNN – WEKA IBk).

Таблица 3.4: Резултати, получени в WEKA с Data Mining алгоритми *OneR, J48, MultiLayer Perceptron* и *IBk* при предсказвана променлива с 5 стойности

	1R Classifier	**Decision Tree**	**Neural Network**	**kNN Algorithm**
Correctly classified instances	55.77%	63.48%	57.026%	59.71%
Kappa Statistic	0.2241	0.3884	0.2816	0.2989
ROC Area	0.612	0.76	0.71	0.742
Model Correct/Incorrect Ratio	1.26	1.74	1.33	1.48
Model/Naïve Correct/Incorrect Ratio	2.07	2.85	2.18	2.43

От сравнението на моделите за класификация, генерирани с избраните Data Mining алгоритми за предсказвана номинална променлива с 5 стойности, могат да се направят следните изводи.

Изводи:

- Най-добър е моделът, получен чрез алгоритъма *Дърво на решенията*, тъй като дава най-висока точност на предсказване ≈64%, както и най-високи стойности на параметрите Kappa Statistic=0.3884 и ROC Area=0.76;
- Моделите, получени чрез алгоритмите *kNN* и *Невронна мрежа,* работят по-слабо от полученото *Дърво на решенията.* Но, тъй като и за тези модели стойностите на параметъра ROC Area са >0.7 (за kNN 0.742, за Невронната мрежа 0.71), това означава, че и тези два модела могат да се използват успешно за предсказване.
- И четирите модела осигуряват поне *2 пъти по-точно предсказване* спрямо наивната класификация (класификация на случаен принцип) (стойностите на отношението Model/Naive Correct/Incorrect Ratio са >2 за всички използвани алгоритми).
- Най-добрият модел, генериран чрез алгоритъма *Дърво на решенията*, подобрява предсказването *2.85 пъти* спрямо наивната класификация.
- Моделите не предсказват с еднаква точност петте класа:
  - *Четирите модела* предсказват с по-висока точност класовете „Добър" и „Много Добър".
  - *Дървото на решенията* и *Невронната мрежа* предсказват с по-висока точност и клас „Слаб".
  - *Четирите модела* не предсказват с висока точност клас „Среден".
  - Най-ниска е точността на предсказване за клас „Отличен".
- Проведените изследвания показват, че точността на предсказване при всички методи за класификация не е достатъчно висока, за да бъде използван един от тях на практика;

3.1.3. Резултати за предсказвана променлива с две стойности

Проучените до момента научни изследвания, както и проведените в настоящата дисертация, показват, че точността на предсказване при голяма част от съществуващите методи за класификация се подобрява, когато предсказваната променлива заема по-малък брой стойности. Тъй като получените резултати от класификацията при предсказвана променлива с пет различни стойности са недостатъчно добри, е предприето трансформиране на предсказваната променлива в двоична величина и провеждане на същите Data Mining анализи при тези нови условия. В този случай, изходната предсказвана променлива е двоична величина от тип категория, приемаща две стойности „Слаби" (Weak) и „Силни" (Strong), съответстващи на студенти с общ успех постигнат в I курс от обучението си в университета, съответно <4.50 и ≥4.50 (виж т.2.2.1.3.Г, Фиг.2.4 и Фиг.2.5).

На етап „Изследване на данните", допълнително в този случай е извършено изследване на *индивидуалното влияние на всяка от входните променливи върху изходната/предсказваната променлива*. За целта е използван алгоритъм „Дърво на решението" (WEKA J48), който се прилага последователно само върху две от променливите, съдържащи се в данните – изследваната входна променлива и изходната/предсказвана променлива. Във всеки от тези случаи на прилагане на алгоритъма, се получава опростено дърво на решенията, чрез което данните се разделят на два класа „Слаби" и „Силни" (двете възможни стойности на изходната променлива) само

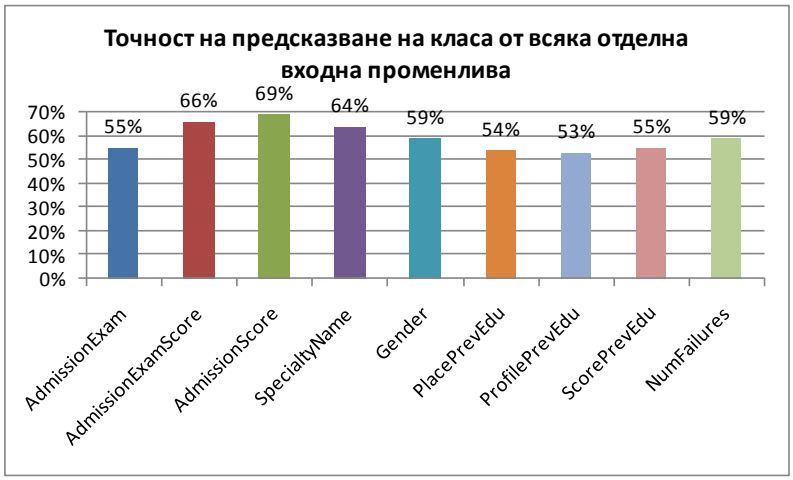
чрез стойностите на изследваната входна променлива (категориите, които заема, ако е дискретна променлива от категориен тип, или прагови стойности, ако е непрекъсната числова променлива).

На етап „Моделиране" са използвани същите четири метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk), използвани в случая на предсказвана променлива с 5 различни стойности. По аналогичен начин са оценени и сравнени генерираните модели за класификация чрез получените стойности на избраните параметри за оценка.

В този случай (при двоична предсказвана променлива) *допълнително е изследвана работата на алгоритмите при различни стойности на избрани техни параметри*, чрез които може да се подобряват генерираните модели за класификация.

3.1.3.1. Изследване влиянието на всяка входна променлива върху предсказването

В изследваната съвкупност от данни се съдържат общо 13 входни променливи и една изходна променлива – класа. На Фиг.3.18 са представени обобщените резултати от изследването на индивидуалното влияние на 9 от входните променливи върху класификацията, и по-точно резултатите по отношение точността на класификация на избрания класификационен алгоритъм (Дърво на решенията).



Фиг.3.18: Резултати от от изследването на индивидуалното влияние на входните променливи върху класификацията

Изводи:

- ***Най-висока точност на предсказване*** се получава в случая, когато като единствена входна променлива е използвана величината „***Бал на приемане"*** (AdmissionScore).
- ***Точността на предсказване чрез параметъра „Бал на приемане" е 69%***, което означава, че броят на правилно класифицираните записи е 2.23 пъти (69%/31%) по-голям от броя на неправилно класифицираните записи. Това е една доста висока точност на класификация като се има предвид, че само една единствена входна променлива (от общо 13) се използва за предсказване.

- Другите входни параметри, при които се получава сравнително висока точност на предсказване са *Оценка от приемен изпит* (AdmissionExamScore, 66%), *Направление/поднаправление* в което е приет кандидат-студента (SpecialtyName, 64%), *Пол* (Gender, 59%) и *Брой слаби оценки – 2-ки* (NumFailures, 59%).
- Получените резултати потвърждават предварителното предположение, че ***индивидуалното въздействие на всяка от входните променливи***, свързани с *Профила*, *Региона* и *Успеха от средното образование* на кандидат-студентите, и *Изпита*, с който са приети в университета, ***върху класификацията е по-слабо*** (получената точност на предсказване е 53-55%). Следователно, ***тези параметри не могат да бъдат самостоятелно използвани за качествено разделяне на студентите на „Слаби и „Силни".***

3.1.3.2. Генериране на модели за класификация с избраните Data Mining алгоритми в WEKA

Изследванията са осъществени върху цялата съвкупност от данни (с всички 14 променливи – 13 входни и 1 изходна). *Предсказвана променлива* е от *категориен тип* и заема *2 стойности* – „Слаби" (Weak) и „Силни" (Strong) (т.2.2.1.3.Г, Фиг.2.4 и Фиг.2.5).

Използвани са четири Data Mining метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk), както и в случая за предсказвана променлива с 5 стойности. Избраните алгоритми са приложени при различни условия – за различни стойности на някои от параметрите за настройка на самите класификационни алгоритми, с цел да се изследват допълнителни възможности за повишаване на точността на предсказване.

Генерираните модели за класификация са оценени и сравнени чрез получените стойности на избрани параметри за оценка (по-подробно описани в т.1.2.4 и т.2.1.3), както това бе направено в случая за предсказвана променлива с 5 стойности (виж. т.3.1.3).

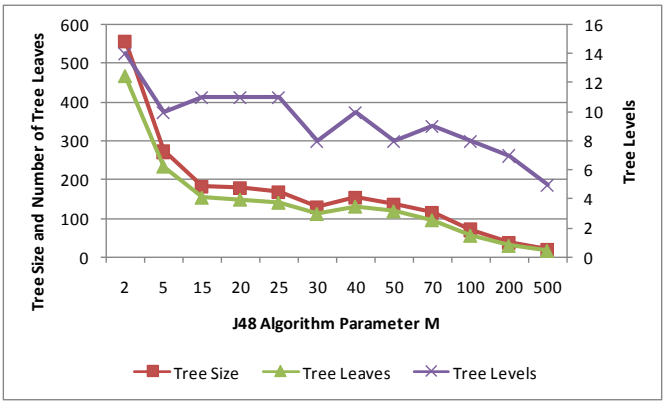
Алгоритъм за класификация 1R

- *Алгоритъмът 1R* извършва класификацията чрез входната променлива *AdmissionScore*, т.е. "*Общ бал*" при приемане на студентите в университета е атрибутът, който дава минимална грешка при класификацията.
- 1R класификаторът работи с малко по-голяма точност (около 1%) при % разделяне на съвкупността от данни на две части (66% за обучение и генериране на модела, и 34% за тестване), отколкото при крос-валидиране (десетократно изпълнение на алгоритъма като всеки път 9/10 от данните се използват за обучение и генериране на модела, а 1/10 - за тестване).
- Точността на класификация= 66%.
    - *Дървото на решението*, получено чрез единствената входна променлива "*Бал от приемането*" (*AdmissionScore*) (виж т.3.1.4.1), извършва класификацията с точност 69%, т.е. с 3% по-висока точност от 1R алгоритъма.
    - Kappa Statistic=0.3439
    - Model Correct/Incorrect Ratio=2.07

- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.05$

Алгоритъм за класификация „Дърво на решенията" - WEKA J48

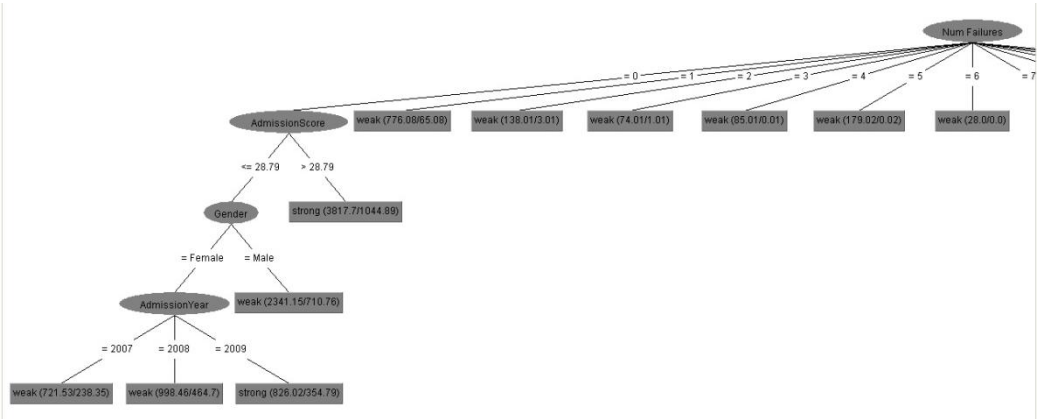
На Фиг.3.20 е представена зависимостта на големината на полученото дърво на решенията от параметъра M (минимален брой записи в едно листо на дървото) на избрания алгоритъм.



Фиг.3.20: Зависимост на големината на полученото *Дърво на решенията* от параметъра M (минимален брой записи в едно листо на дървото)

От данните в Табл.3.12 и на Фиг.3.20 се вижда, че с увеличаване на минималния брой записи в листата, намалява общият размер на дървото и броят на листата (от 12 случая има 1 изключение). Като цяло, има тенденция и за намаляване броя на нивата на дървото, но тя не е толкова ясно изразена (от 12 случая има 3 изключения).

При по-малки стойности на параметъра M на алгоритъма се получава дърво на решенията с голяма размерност. Такова дърво не може да бъде представено така, че лесно да бъде обхванато с поглед и анализирано. При стойност M=500 се получава дърво на решенията организирано на 5 нива, с общо 21 възела, 17 от които са листа на дървото. Именно това дърво (липсва малка част) е представено на Фиг.3.21.



Фиг.3.21: Полученото *Дърво на решенията* чрез алгоритъма *J48* в WEKA (при M=500)

Точността на предсказване при M=500 намалява до 70.17%, ROC Area=0.762, но това дърво е много по-лесно за интерпретиране. От него могат да бъдат извлечени следните правила за класифициране на студентите в двата класа „Слаби" и „Силни":

IF NumFailures=1,2,3,4,5,6,7,8,9,10,11,12, THEN Class=Weak

IF NumFailures=0 AND AdmissionScore>28.79, THEN Class=Strong

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Male, THEN Class=Weak

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Female AND AdmissionYear=2009, THEN Class=Strong

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Female AND AdmissionYear=2007,2008, THEN Class=Weak

Съгласно получените класификационни правила:

- В клас „Силни" се определят студентите, които:
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета и са приети с Бал >28.79;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.79, от женски пол са и са приети през 2009г.
- В клас „Слаби" се определят студенти, които:
  - Имат слаби оценки (2-ки), получени на изпитите в края на първата учебна година в университета;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.78 и са от мъжки пол;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.78, от женски пол са и са приети през 2007 или през 2008г.

По аналогичен начин може да бъде интерпретирано всяко дърво, което се получава като резултат от прилагането на избран алгоритъм „Дърво на решението" върху изследваните данни. Дървото на решения с по-голяма размерност и по-голям брой листа (и съответно по-малък брой записи в листата) характеризира по-детайлно съществуващите взаимовръзки между променливите в изследваната съвкупност от данни, но при него в по-голям риска от описване и на наличния шум в данните. Обратно, дърво на решения с по-малка размерност и по-малък брой листа (съответно по-голям брой записи в листата) описва по-общо съществуващите взаимовръзки между променливите в изследваната съвкупност от данни, но е по-лесно за интерпретация и е с по-малък риск относно характеризирането на наличния в данните шум.

<u>Изводи за модела на класификация, получен чрез алгоритъма "Дърво на решенията":</u>

- Най-добри параметри на модела за класификация чрез алгоритъма "Дърво на решенията" се получават при % разделяне на изследваната съвкупност от данни на 2 части (2/3 за генериране на модела и 1/3 за тестване.
- Постигната най-висока Точност на класификация =72.7%
- Най-висока стойност на параметъра Kappa Statistic=0.4524
- Най-висока стойност на параметъра ROC Area=0.784
- Model Correct/Incorrect Ratio=2.67
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.64$

- Входните променливи, които най-добре разделят студентите на два класа „Слаби” и „Силни”, са *Брой 2-ки (NumFailures), Бал на приемане* (*AdmissionScore), Пол* (*Gender)* и *Направление*/*Поднаправление* (*UnivSpecialtyName).*
- Промяната на параметъра M (минимален брой записи в едно листо на дървото) на алгоритъма „Дърво на решенията” води до промяна и на моделът за класификация, като увеличаването на M води до намаляване на общия размер на дървото, броя на листата и броя на нивата в дървото.
- Точността на предсказване намалява с намаляване размерността на дървото.
- Дърветата с по-малка размерност се интерпретират по-лесно – от тях се извличат разбираеми за потребителя класификационни правила.

Алгоритъм за класификация „Невронна мрежа” - WEKA MultilayerPerceptron

Една от основните особености при прилагането на класификационен алгоритъм от типа „Невронна мрежа” е необходимостта от преобразуване на променливите, тъй като невронните мрежи работят с числови величини и са особено чувствителни при наличието на колинеарност между входните променливи.

Избраният алгоритъм „Невронна мрежа” (WEKA MultilayerPerceptron) е приложен със стойностите по подразбиране (default parameters), като е променена само стойността на параметъра HiddenLayers (H), чрез който се задава броят на невроните в скрития слой на невронната мрежа, и в случая е зададена стойност H=1. Алгоритъмът е реализиран при тестова опция „% разделяне” като отново изследваната съвкупност от данни е разделена на две части – обучаваща извадка (66%, т.е.2/3 от общата съвкупност) и тестова извадка (34%, т.е. 1/3 от общата съвкупност).

Един от начините за подобряване точността на предсказване на невронните мрежи е да се експериментира с нейната топология. Затова, избраният алгоритъм, невронна мрежа MultilayerPerceptron, е приложен и при стойност по подразбиране за параметъра (H)HiddenLayers=a (a=(attributes+classes)/2).

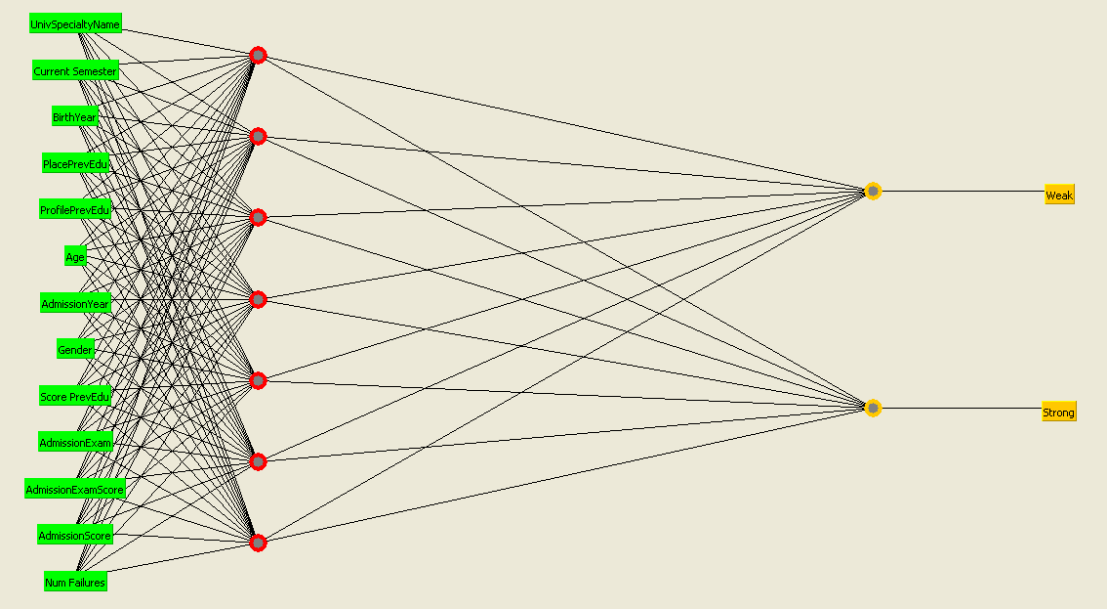
Върху точността на предсказване на невронната мрежа оказва влияние типа на променливите в изследваната съвкупност от данни. Затова е създаден нов файл с данни, в който категорийните променливи са трансформирани в цифрови.

Таблица 3.16: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=1 и H=a)

MultiLayer Perceptron	**Test Option - % Split (66/34)**					
	H=1			**H=a**		
	Weak	**Strong**	**Weighted Average**	**Weak**	**Strong**	**Weighted Average**
Correctly Classified Instances	71.3701%			73.5904%		
Incorrectly Classified Instances	28.6299%			26.4096%		
Kappa Statistic	0.4326			0.4730		
TP Rate	0.625	0.813	0.714	0.704	0.772	0.736
FP Rate	0.188	0.375	0.276	0.228	0.296	0.26
Precision	0.789	0.66	0.728	0.775	0.7	0.74
Recall	0.625	0.813	0.714	0.704	0.772	0.736
F-Measure	0.698	0.728	0.712	0.738	0.734	0.736
ROC Area	0.784	0.784	0.784	0.823	0.823	0.823

Резултатите от класификацията са представени в Табл.3.16, при две различни стойности на параметъра H (HiddenLayers) – H=1 и H=a=(attribute+class)/2.

Точността на класификация е по-висока за стойност H=a (73.59%), отколкото за стойност H=1 (71.31%). Топологията на мрежата при H=a е представена на Фиг.3.27 и съдържа входящ слой с 13 неврона, съответстващи на 13-те входни променливи, 1 скрит слой със 7 неврона (a=(13+1)/2=14/2=7) и изходящ слой с 2 неврона, съответстващи на двете стойности на изходната променлива – класа (Слаби и Силни).



Фиг.3.27: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=a)

В случая, когато категорийните променливи са предварително трансформирани в цифрови, се получава по-висока точност на класификация, отколкото ако трансформирането на категорийните променливи в двоични цифрови величини се извършва автоматично от невронната мрежа.

Изводи за модела на класификация, получен чрез алгоритъма "Невронна мрежа":

- При прилагането на алгоритъма „Невронна мрежа", по-добри модели за класификация се получават, когато се извърши подходящо предварително преобразуване на променливите от категориен тип в цифрови величини. Този извод напълно съответства на очакванията, тъй като самата природа на този алгоритъм изисква работа с цифрови величини, затова в този вид алгоритми е предвидена процедура за предварително преобразуване на качествените променливи.
- Премахването на две от променливите, които са взаимно свързани, от наличната съвкупност от данни, не води до получаването на модел за класификация с по-добри параметри. Обикновено очакванията в това отношение се различават от получените резултати, тъй като премахването на взаимосвързани променливи обикновено води до повишане на точността на предсказване.
- Върху точността на предсказване на моделите, генерирани чрез алгоритъм „Невронна мрежа", оказва влияние топологията на мрежата. В конкретните изследвания,

невронната мрежа с по-голям брой неврони в скрития слой работи с по-голяма точност от невронната мрежа с един неврон в скрития слой.

- Най-добрият модел за класификация чрез алгоритъма „Невронна мрежа" се получава за пълната съвкупност от данни (с всички 14 променливи), при предварително подходящо преобразуване на всички дискретни променливи в цифрови величини, и при наличието на 7 неврона в скрития слой (стойност на параметъра H=(attribute+class)/2=7).
- Постигната най-висока Точност на класификация =73.59%
- Най-висока стойност на параметъра Kappa Statistic=0.473
- Най-висока стойност на параметъра ROC Area=0.823
- Model Correct/Incorrect Ratio=2.79
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.76$
- Невронната мрежа е единственият модел за предсказване, генериран в рамките на изследванията, при който точността на предсказване на клас „Силни" е по-висока от точността на предсказване на клас „Слаби".

Алгоритъм за класификация „К-най-близък съсед" – WEKA IBk

Алгоритъмът К-най-близък съсед (WEKA IBk) е приложен за различни стойности на параметъра К – предварително зададеният брой обекти, на базата на които се определя класът на нов обект..

Получените резултати показват, че алгоритъмът работи с най-голяма точност на класификация при К=50. В този случай висока стойност имат и параметрите Kappa Statistic=0.4085 (неговата най-висока стойност е 0.4086 за К=60) и ROC Area=0.784 (неговите най-високи стойности са 0.786 за К=80 и 0.785 за К=60,70). Именно за случая К=50 са и следващите представени резултати. В Табл.3.19 са дадени резултатите от прилагането на алгоритъма за К=50.

Изводи за модела на класификация, получен чрез алгоритъма "К-Най-близък съсед":

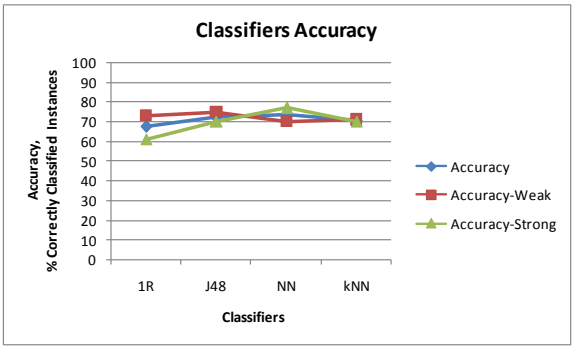
- Параметърът К – предварително зададеният брой обекти, на базата на които се определя класът на нов обект, е един от основните параметри на алгоритъма, чрез които може да се променя неговата точност на класификация. В проведеното изследване, модел за класификация с най-висока точност на предсказване е получен при К=50.
- Точност на предсказване = 70.47%
- Kappa Statistic = 0.4085
- ROC Area = 0.784
- Model Correct/Incorrect Ratio=2.39
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.37$

Сравнение на резултатите, получени за различните алгоритми на класификация

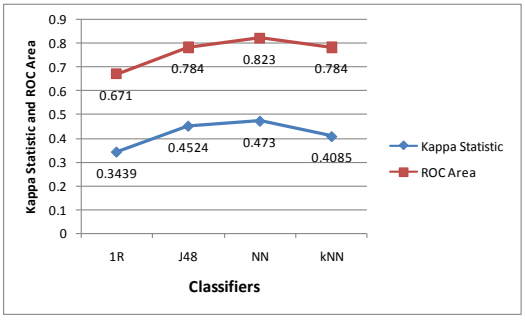
Таблица 3.21: Резултати, получени в WEKA с Data Mining алгоритми *OneR, J48, MultiLayer Perceptron* и *IBk* при предсказвана променлива с 2 стойности

	1R Classifier	Decision Tree	Neural Network	kNN Algorithm
Correctly classified instances	67.46%	72.74%	73.59%	70.47%
Correctly classified instances for Class Weak	73%	75%	70%	71%
Correctly Classified Instances for Class Strong	61%	70%	77%	70%
Kappa Statistic	0.3439	0.4524	0.4730	0.4085
ROC Area	0.671	0.784	0.823	0.784
Model Correct/Incorrect Ratio	2.07	2.67	2.79	2.39
Model/Naïve Correct/Incorrect Ratio	2.05	2.64	2.76	2.37

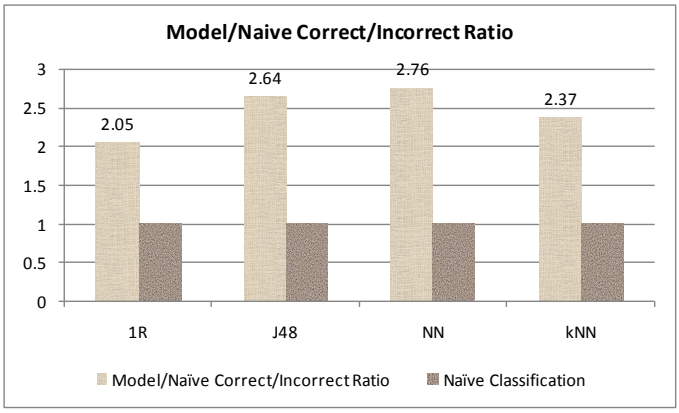
На Фиг.3.28 е представено сравнение на избраните класификационни алгоритми – генератор на правила (1R), дърво на решението (J48), невронна мрежа (NN) и К-най-близък съсед (kNN), по отношение на точността на предсказване.



Фиг.3.28: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA, по отношение на *точността на предсказване*



Фиг.3.29: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA



Фиг.3.30: Сравнение на моделите за класификация, получени чрез алгоритмите OneR, J48, Multilayer Perceptron и IBk в WEKA, по отношение на коефициентите на подобрение на предсказването спрямо случайната класификация

Изводи:

- Сред избраните класификационни алгоритми, *най-висока точност на предсказване* е получена при модела, генериран чрез алгоритъма „*Невронна мрежа*" *– 73.59%.*
- *Невронната мрежа* е единствения класификационен модел, при който *точността на предсказване на клас „Силни" (77%) е по-висока от точността на предсказване на клас „Слаби"*, и това е най-високата постигната точност за предсказване на някой от двата класа.
- Полученото *Дърво на решението* също е *модел с висока точност на класификация – 72.74%*, като този модел е *по-разбираем* и *лесен за интерпретация* за потребителите.
- Моделът, създаден чрез алгоритъма *К-най-близък съсед*, извършва класификацията с *точност 70.5%* и показва *приблизително еднакви точности на предсказване на двата класа „Слаби" и „Силни"*.
- *Най-ниска точност на предсказване* е получена за модела, създаден чрез *генератора на правила 1R.*
- *Най-висока стойност* на параметъра *Kappa Statistic=0.473* е получена за модела „*Невронна мрежа*".
- *Най-висока стойност* на параметъра *ROC Area=0.823* е получена за модела „*Невронна мрежа*".
- Стойностите на параметъра *ROC Area* за три от избраните алгоритми – *Невронна мрежа, Дърво на решенията* и *К-най-близък съсед*, са *>0.7*, което означава, че и трите получени модела за класификация са добри.
- *Най-висок коефициент за подобряване на предсказването спрямо наивна класификация* е получен за *Невронната мрежа – 2.76*, следва *Дърво на решенията – 2.64.*
- За *четирите използвани алгоритъма* получените *коефициенти са* >2, т.е. и с четирите модела предсказването ще се извършва поне 2 пъти по-точно в сравнение с наивната класификация.

3.1.4. Сравнение на получените Data Mining модели за класификация за двата варианта на предсказваната променлива

Въз основа на резултатите, получени при генерирането на Data Mining модели за класификация, върху съвкупността от данни за студентите, за двата варианта на предсказваната променлива - с пет и с две стойности, могат да бъдат направени следните изводи.

Изводи:

- Генерираните модели за класификация чрез четирите избрани алгоритми дават *по-голяма точност на предсказване*, когато *изходната променлива съдържа по-малък брой различни стойности*.
- За *двоична предсказвана променлива* (с две стойности - „Слаби" и „Силни"):
    - *Всички приложени алгоритми, с изключение на 1R класификатора*, работят с обща *точност на предсказване >70%* (weighted average accuracy), като се справят сравнително добре и с двата класа.
    - И за *трите алгоритъма (Невронна мрежа, Дърво на решението и К-Най-близък съсед*) се получават високи стойности на параметрите *Kappa Statistic (>0.40) и ROC Area (>0.78)*, което означава, че генерираните модели са подходящи за извършване на класификацията с висока точност.
    - *Невронната мрежа* е генерираният *модел за класификация с най-добри параметри*.
    - *Невронната мрежа* се справя най-добре при предсказването на *клас „Силни"*, което в конкретния бизнес контекст е особено важно.
    - Чрез прилагането на създадената *невронна мрежа* в реални условия, върху нови данни, *точността на предсказване може да се подобри 2.76 пъти в сравнение с наивната класификация*.
- За *предсказвана променлива с пет различни стоности* – „Слаб", „Среден", „Добър", „Много Добър" и „Отличен":
    - Получена е *по-ниска точност на предсказване* и при *четирите използвани алгоритъма (<70%)*
    - Наблюдават се *съществени разлики в точността на предсказване на различните класове*
    - *Най-ниски точности на предсказване* се получават по отношение на *клас „Отличен"*.
    - *Най-добър модел* се получава чрез алгоритъма „*Дърво на решенията*", който дава *2.85 пъти подобрение в сравнение с наивната класификация*.

3.2. Генериране и оценка на Data Mining класификационни модели за предсказване вида на открити морски цели

Data Mining задачата за класификация и в случая на радарни данни се осъществява на базата на CRISP-DM (Cross-Industry Standard Process for Data Mining) модела – стандартен подход за реализация на Data Mining проекти, и включва:

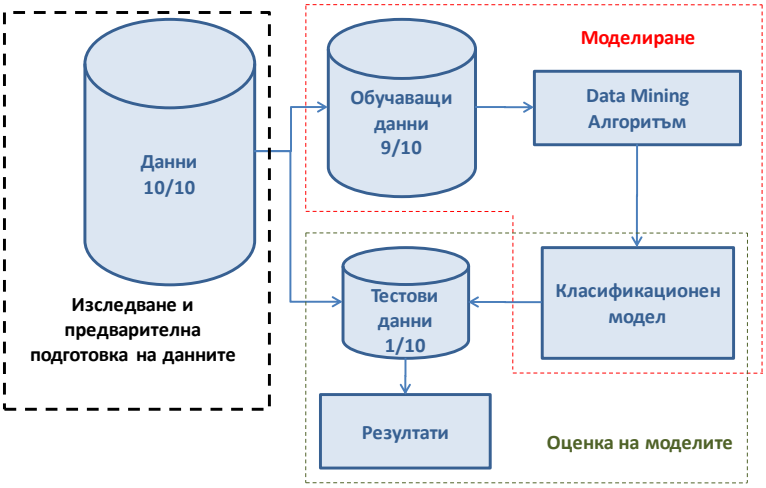
- Избор и извличане на данни за Data Mining анализи – описано в т.2.2.2 на дисертацията;

- Изследване и предварителна подготовка на данните - описано в т.2.2.2 на дисертацията;
- Избор на Data Mining методи и генериране на модели за предсказване вида на откритите радарни морски цели чрез софтуерен пакет WEKA;
- Сравнение и оценка на получените модели за предсказване.

Data Mining анализите са осъществени върху *съвкупността от данни*, съдържаща *80 записа* на радарни морски цели, описани чрез *16 параметъра* (описани в т.2.2.2 и Табл.2.4). *Предсказваната целева променлива* е *вида на откритата радарна цел - променлива от категориен тип* с *три различни стойности - MISL Boat, Big Boat* и *Average Boat*. Броят на записите в трите класа е различен. Най-много данни са налични за клас MISL Boat (62 записа), а доста по-ограничен е броят са записите за клас Big Boat (11) и клас Average Boat (6). За целевата променлива има само една липсваща стойност (Target Variable - 1% Missing Values - виж Табл.2.4), затова общият брой на записите в изследваната съвкупност от данни е 79.

За решаването на Data mining задачата за класификация са използвани различни алгоритми: *Генератори на правила* (Rule Learners) – WEKA *OneR* и *JRip*, *Дърво на решението* (Decision Tree) – WEKA *J48* (базиран на алгоритъма C4.5) и *RandomForest*, *Невронна мрежа* (Neural Network) – WEKA *Multilayer Perceptron*, *К-най-близък съсед* (K-Nearest Neighbour) – WEKA *IBk*, два Байесови класификатора (*NaiveBayes* и *BayesNet*) и Логистична регресия - *SimpleLogistic*. Тези алгоритми са избрани, тъй като всеки от тях работи на различен принцип, т.е. по различен начин генерира модела за предсказване на целевата променлива.

Избраните data mining алгоритми за класификация са приложени върху наличните данни като са използвани стойностите по подразбиране на техните параметри за настройка (WEKA default parameters), освен в случаите, когато това е допълнително указано. Моделите за предсказване вида на откритата радарна морска цел са получени като е използвана тестова схема „stratified 10-fold cross validation" (крос-валидиране), представена на Фиг.3.31.



Фиг.3.31: Тестова схема за класификация крос-валидиране („10-fold Cross Validation")

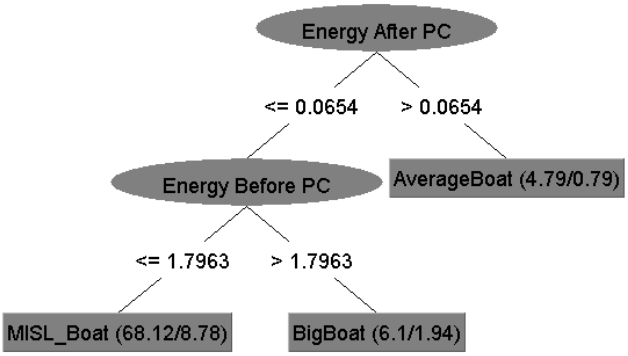
При "stratified 10-fold cross validation test option" общата съвкупност от данни се разделя на 10 части и алгоритъмът се прилага 10 пъти, като всеки път 9 от 10-те части се използват

като обучаващи данни (training data), а останалата 1 част се използва за оценка на модела. Това е стандартен подход, който се прилага за определяне грешката на предсказване на класификационните алгоритми при наличието на една постоянна съвкупност от данни. Стратификацията на данните означава, че съвкупността от данни ще бъде разделена произволно на 10 части, но във всяка част всеки клас ще е представен приблизително в същите пропорции, както в цялата съвкупност от данни. Крайната грешка на алгоритъма се определя като средно аритметично от грешките, получени при 10-те изпълнения на алгоритъма.

Всеки от моделите за класификация, получени чрез приложените различни Data Mining алгоритми, е оценен по критерии, по-подробно описани в т.1.2.4 на дисертацията.:

Алгоритъм „Дърво на решенията"

На Фиг.3.32 е представено полученото „Дърво на решенията" чрез алгоритъма J48:



Фиг.3.32: Полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

- *Общата точност на предсказване* на модела е висока – 81%, но има съществени разлики по отношение предсказването на трите класа.
- Най-висока е *точността на предсказване* на клас *MISL Boat*, което най-вероятно се дължи на факта, че именно този клас има най-много представители в изследваната съвкупност от данни (78%).
- *Точността на предсказване* на другите два класа, *Big Boat* и *Average Boat*, е значително по-ниска.
- *Kappa Statistic*=0.3328 (при максимална възможна стойност 1.0), следователно степента на съгласуваност между предсказания и реалния клас не е много висока, т.е. някои класовете се предсказват с много ниска точност.
- *ROC Area*=0.596 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.
- *Model Correct/Incorrect Ratio*=4.27
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.62$, следователно се постига 2.62 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).
- Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е*

разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.

Алгоритъм за класификация JRip

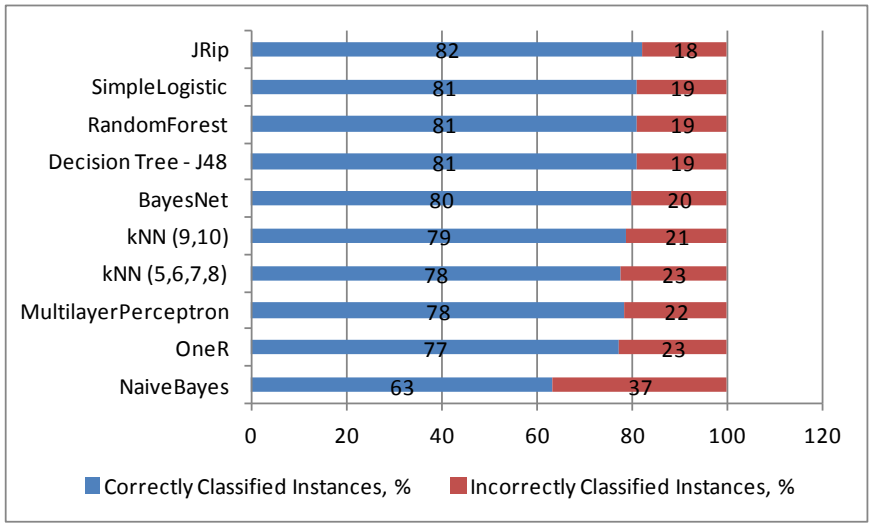
- *Общата точност на предсказване* на модела е висока – 82%, но има съществени разлики по отношение предсказването на трите класа.
- Най-висока е *точността на предсказване* на клас *MISL Boat*, което най-вероятно се дължи на факта, че именно този клас има най-много представители в изследваната съвкупност от данни (78%).
- *Точността на предсказване* на другите два класа, *Big Boat* и *Average Boat*, е значително по-ниска.
- *Kappa Statistic*=0.3333, следователно степента на съгласуваност между предсказания и реалния клас не е много висока, т.е. някои класовете се предсказват с много ниска точност.
- *ROC Area*=0.583 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.
- *Model Correct/Incorrect Ratio*=4.64
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.85$, следователно се постига 2.85 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).

Алгоритъм за класификация SimpleLogistic

- *Общата точност на предсказване* на модела е висока – 82%, но има съществени разлики по отношение предсказването на трите класа.
- Алгоритъмът предсказва без грешка клас *MISL Boat.*
- Алгоритъмът предсказва най-често клас *MISL Boat* за обектите от другите два класа - *Big Boat* и *Average Boat.*
- *Kappa Statistic*=0.1828 (<<1), следователно степента на съгласуваност между предсказания и реалния клас е много ниска, т.е. потвърждава се неспособността на алгоритъма да предсказва точно някои от класовете.
- *ROC Area*=0.659 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.
- *Model Correct/Incorrect Ratio*=4.27
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.62$, следователно се постига 2.62 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).

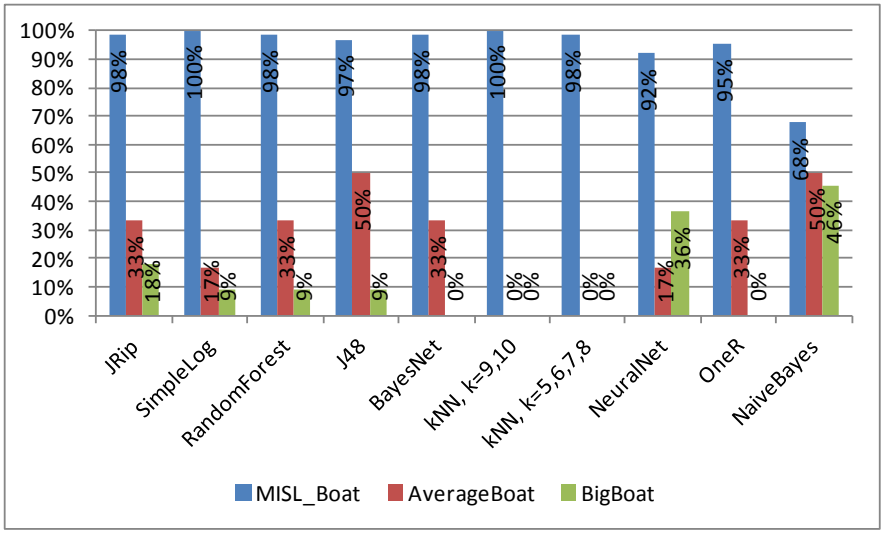
Сравнение на резултатите от класификацията, получени чрез прилагането на избраните Data Mining алгоритми

Получените резултати за Точността/Грешката, получена при предсказване на вида на откритите FS радарни морски цели (класа) чрез различните модели за класификация, генерирани с избраните Data Mining алгоритми, са представени на Фиг.3.33.



Фиг.3.33. Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността/грешката на предсказване

Получените резултати от сравнението на генерираните модели за класификация чрез различните Data Mining методи, относно способността им за предсказване на трите различни класа – *MISL Boat*, *Average Boat* и *Big Boat*, на базата на параметъра True Positive Rate (TP Rate), са представени на Фиг.3.34.



Фиг.3.34: Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността на класификация по класове

Изводи:

- Всички избрани класификационни алгоритми - Генератори на правила OneR и JRip; Дърво на решението - J48 и RandomForest; Невронна мрежа - Multilayer Perceptron; К-най-близък съсед IBk; Байесови класификатори - NaiveBayes и BayesNet, и Логистична регресия - SimpleLogistic, предсказват с висока точност (68% - 100%)) клас MISL Boat.
- Точността на предсказване е доста по-ниска за клас Average Boat и най-ниска за клас Big Boat, които са по-слабо представени в общата съвкупност от данни
- Алгоритмите BayesNet, OneR и kNN, са абсолютно неспособни да предсказват класовете, които са по-слабо представени в наличната съвкупност от данни.
- Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC*, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.
- Data Mining моделът за класификация, получен чрез алгоритъма Дърво на решенията (WEKA J48), е със сравнително висока точност на предсказване за два от класовете (97% за MISL Boat, 50% за Average Boat) и с ниска точност за третия клас (9% за Small Boat). Но, полученият модел е много разбираем, лесен за интерпретация от потребителите, и по-лесно може да се реализира за работа в реално време.
- Data Mining моделът за класификация, получен чрез Наивния Байесов класификатор (WEKA NaiveBayes), е с близки точности на предсказване за трите класа (68% за MISL Boat, 50% за Average Boat, 46% за Small Boat). Това е единственият модел, който дава сравнително приемлива точност на класификация и за трите вида морски цели.
- Получените резултати са съизмерими с резултатите, получени от изследователския колектив на Бирмингамския университет, отнасящи се за класификация на движещи се наземни цели (автомобили) чрез използване на Forward Scattering радарна система, представени в т.1.4.

3.3. Изводи по Трета Глава

В настоящата глава е постигнато следното:

- Обучени са класификатори за извличане на знания от данни (генерирани са Data Mining модели за класификация) чрез избрани класификационни алгоритми, за две съвкупности от данни, в Data Mining софтуер *WEKA*.
- Качеството на обучените класификатори е оценено по избраните мерки за оценка.

Заключение

В настоящия дисертационен труд е решена и изследвана задачата за откриване на знания в данни чрез обучение на класификатори, за две съвкупности от данни в различни предметни области - за предсказване успеваемостта на студентите в зависимост от техни характеристики преди и след приемане в университета, и за предсказване класа на морски обекти в зависимост от избрани параметри на отразените сигнали от тези обекти.

Научно-приложни приноси

Най-важните *научно-приложни приноси*, с характер *обогатяване на съществуващи знания и приложение на научните постижения в практиката,* са представени в следните групи - ***получаване и доказване на нови факти***, ***подтвърдителни факти, приноси за внедряване***, както следва:

А) Получаване на нови факти, зависимости, изводи и заключения

Решена и изследвана е задача за 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 преди и след приемане в университета, за два варианта на предсказваната променлива (с пет и с две стойности);
- за предсказване на класа на морски обекти, представен като номинална променлива с три стойности, в зависимост от избрани времеви параметри на отразени сигнали от целите.

Б) Получаване на подтвърдителни факти

Получени са подтвърдителни факти за двете изследвани съвкупности от данни:

- Разпределянето на студентите в избраните класове (2 или 5) в най-голяма степен зависи от предсказващите променливи *Бал на приемане, Оценка от приемен изпит, Направление в което са приети студентите, Брой двойки*.
- Точността на предсказване на получените модели за класификация на морски цели е сравнима с резултатите, получени от изследователи в Бирмингамския университет, използвали подобни честотни алгоритми за класификация на движещи се наземни цели.
- Потвърдена е възможността за успешно прилагане на една и съща методика за откриване на знания в данни чрез обучение на класификатори върху две съвкупности от данни в различни предметни области.

В) Приноси за внедряване: методи, конструкции, реализация на по-ефективни средства за приложения

- Апробирана е методика на изследването, включваща предварителна подготовка на данните, обучение на класификатори за откриване на знания в данни, и оценка и сравнение на получените модели, която може да се използва и в други случаи, за извличане на закономерности от:

- данни за учениците и студентите в различни училища или университети, за нуждите на подобряване на качеството на обучение в Университети и училища.
- данни за радарни и GPS сигнали от различни видове цели - наземни, въздушни и морски.

- Получени са множество резултати от откриване на знания в данни чрез обучение на класификатори за предсказване успеваемостта на студентите, описващи зависимости между успеха на студентите и техните индивидуални особености: пол, приемен изпит, профил и регион на училището, където е получено средно образование, професионално направление, в което се обучават, и др., в табличен и графичен формат, и под формата на изводи.

- Част от получените резултати се използват в проекти:
  - 2010-2013: Проект No.ДДВУ02/50/20.12.2010г. на СУ ”Св. Кл. Охридски” – НИС, финансиран от Фонд „Научни изследвания”, на тема „Разработка на програмна система за изследване и проектиране на радио мрежи, базирани на радиотехники за разпространение на сигнали „напред”, за лоциране на движещи се обекти на фона на море и електронни смущения с цел защита на морски зони и граници”.
  - 2009-2013: Проект No.ДТК 02/28/2009г. на ИИКТ-БАН, финансирани от Фонд „Научни изследвания”, на тема "Откриване, оценка на параметрите на слаби GPS сигнали и подобряване на ефективността на системата чрез подтискане на радиочестотни смущения и намаляване на навигационната грешка".
  - 2012-2013: Университетски проект № НИД НИ 2–1/2012г., финансиран от УНСС-НИД, на тема „Създаване на профил на студента в УНСС чрез средствата на Data Mining”.

Цитирания

Открито е цитиране на труд No.22 от Библиографията на дисертацията, в списание с импакт фактор около 1 (0,878 за 2011, виж www.bioxbio.com/.../IET-R):

- труд N22, Kabakchiev C., D. Kabakchieva, M. Cherniakov, M. Gasninova, V. Behar, I. Garvanov,”Maritime Target Detection, Estimation and Classification in Bistatic Ultra Wideband Forward Scattering Radar”, Intern. Radar Symp. IRS-2011,7-9 Sept., Leipzig, 2011, pp.79-84:

цитиран през 2012г. в списание "IET Radar, Sonar and Navigation"

- Ai, X. , Li, Y. , Wang, X. "Some results on characteristics of bistatic high-range resolution profiles for target classification"

Благодарности

Изследванията в дисертацията, свързани с въпросите за генериране на Data Mining модели за класификаци, за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, са осъществени с подкрепата и в резултат от дейностите, проведени по проект 2012-2013, Университетски проект № НИД НИ 2–1/2012г., финансиран от УНСС-НИД.

Изследванията в дисертацията, свързани с въпросите за генериране на Data Mining модели за класификация на морски цели по параметрите на FSR сигналите, получени от тях, са осъществени с подкрепата и в резултат от дейностите, проведени по проект 2010-2013: Проект No.ДДВУ02/50/20.12.2010г. на СУ "Св. Кл. Охридски" – НИС, и Проект ДТК 02/28/2009г. на ИИКТ-БАН, финансирани от Фонд „Научни изследвания".

Дисертантът изказва благодарност на неговия научен ръководител акад.проф.д.н.Иван Попчев, както и на секция "Интелигентни системи" с ръководител доц.д-р Любка Дуковска.

Библиография (избрани източници)

1. Baker, R., Yacef, K. (2009). *The State of Educational Data mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining, Vol.1, Issue 1, Oct. 2009, pp.3-17.
2. Cherniakov, M., Raja Abdullah, R.S.A., Jancovic, P., Salous, M. (2005). *Forward Scattering Micro Sensor for Vehicle Classification*. Proceedings of the IEEE International Radar Conference, Washington DC, US, pp. 184-189, 2005.
3. Guidici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. John Wiley & Sons Ltd.
4. Han, J., Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Simon Fraser University. Morgan Kaufmann Publishers.
5. Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.
6. Larose, D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc.
7. Larose, D. (2006). *Data Mining: Methods and Models*. John Wiley & Sons, Inc.
8. Rashid, N., Antoniou, M, Jancovic, P, Sizov, V, Abdullah, R, Cherniakov, M. (2008). *Automatic Target Classification in a Low Frequency FSR Network*. Proc. of EuRAD Conf. 2008, pp. 68-71.
9. Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers Inc.
10. Pyle, D. (2003). *Business Modeling and Data Mining*. Morgan Kaufmann Publishers Inc.
11. Tan, P., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley.
12. Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons Ltd, Chichester, West Sussex, UK.
13. Vialardi, C., Bravo, J., Shafti, L., Ortigosa, A. (2009). *Recommendation in Higher Education Using Data Mining Techniques*. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp. 190-199.
14. Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J., Ortigosa, A. (2011). *A data mining approach to guide students through the enrollment process based on academic performance*. User Model User-Adap Inter (2011) 21:217–248, Springer Science+Business Media B.V. 2011.
15. Witten, I., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Elsevier Inc. 2005.