TECHNICKÁ UNVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY
KATEDRA KYBERNETIKY A UMELEJ INTELIGENCIE

Dokumentácia k zadaniu z predmetu Hybridná výpočtová inteligencia

Využitie HVI pri diagnostike rakoviny prsníkov

Nadiia Bezerova

Karina Cimborová

Stanislava-Saranda Halupkova

Diana Olejárová

2019

Obsah

1. Úloha

Obsahom zadania bolo navrhnúť a implementovať systém s využitím metód hybridnej výpočtovej umelej inteligencie (okrem ANFIS). Výstupom mala byť webová stránka s možnosťou daný systém odtestovať na vlastných dátach.

1.1 Dodefinovanie témy

Pre naše zadanie sme sa rozhodli pracovať s dátami o rakovine mliečnych žliaz a tým pádom navrhnúť neuro-fuzzy systém, ktorý by na základe uvedených dát predikoval výskyt danej rakoviny.

2. Analýza riešení predikcie rakoviny mliečnych žliaz

Množstvo dát, ktoré sú denne produkované už nie je možné spracovať klasickými analytickými metódami. Takto sa do popredia dostáva dolovanie dát a metódy strojového učenia. Využitie týchto prostriedkov ústí do nachádzania nových vzorov a vzťahov medzi dátami ako aj pri nachádzaní nových znalostí.

Pri väčšine ochorení je dôležité začať liečbu včas. Niekedy ale expert prehliadne súvislosti a zle vyhodnotí stav pacienta, čo môže pri vážnejších ochoreniach končiť zle. Tu prichádzajú metódy umelej inteligencie a strojového učenia, ktorých potenciál sa v tejto oblasti už plne využíva.

Rakovina mliečnych žliaz je najbežnejšia forma rakoviny medzi ženami a taktiež je aj hlavnou príčinou úmrtí žien [1]. Ak je rakovina detekovaná v počiatočnom štádiu, je 30% šanca, že jej liečba nebude veľmi komplikovaná a bude úspešná. Medzi tradičné metódy jej detekcie patrí mamografia, ultrasonografia či tenkoihlová aspiračná biopsia (FNA). Rozvoj technológií ale prináša možnosť hľadať iné spôsoby detekovania tejto choroby.

Prieskum štúdií nám ukázal, že datset, s ktorým sme sa rozhodli pracovať aj my je už dlhé roky využívaný v mnohých štúdiách. Prevažná väčšina výskumov sa zameriavala na porovnanie rôznych algoritmov strojového učenia medzi sebou a vyhodnotenie, ktorý dosahuje najväčšiu presnosť. Často išlo o Naivný Bayesov klasifikátor, J48 algoritmus a rôzne formy neurónových sietí . Pozoruhodné bolo, že Naivný Bayes väčšinou dosahovala najlepšie výsledky. Podarilo sa nám nájsť aj výskumy, ktoré navrhovali využitie hybridných systémov, hlavne fuzzy-(evolučné/neuro) systémy. Neuro-fuzzy systém, aký sme sa chystali využiť my, v zložení RBF siete a TSK regulátora sme nenašli, no nevylučujeme, že existujú štúdie, ktoré ho využívajú.

Taktiež sme videli, že niektoré štúdie zanedbávali niektoré atribúty datasetu.

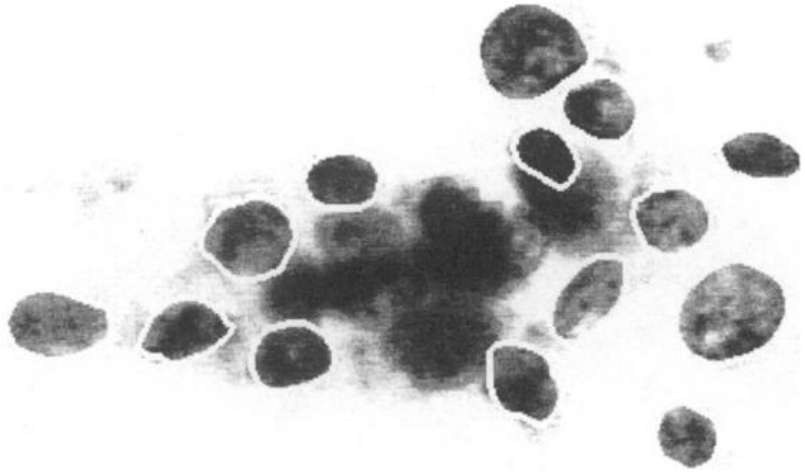

Figure 1 Zväčšený obrázok výsledku tenkoihlovej aspiračnej biopsie = zhubné

[2] popisuje porovnanie troch algoritmov pri predikcii rakoviny prsníkov. Konkrétne ide o algoritmus Naivného Bayesa, J48 a taktiež RBF sieť. Spomínajú, že výber RBF siete bol na základe toho, že konverguje rýchlejšie než MLP sieť a je aj spoľahlivejší. Výsledok tohto experimentu ukázal, že Naivný Bayes dosahoval lepšie výsledky než RBS a J48. RBF vyhrávala jedine z hľadiska výkonu. Z hľadiska presnosti teda Naivný Bayes dosiahol 97.36%, RBF sieť 96.77% a J48 93.41%.

[3] sa taktiež zameriava na porovnanie dvoch metód strojového učenia pri predikcii rakoviny mliečnych žliaz. Opäť išlo o Naivného Bayesa a J48. Práca kladie dôraz na predspracovanie dát. Výsledkom bolo opäť to, že Naviný Bayes mal vyššiu presnosť (97.80%), ale dáta bolo potrebné predspracovať, pretože algoritmus si neporadí s chýbajúcimi hodnotami a taktiež sa mu lepšie pracuje s diskretizovanými nominálnymi hodnotami. Naproti tomu J48 obstál horšie. Jeho presnosť bola 96.05%, čo ale nie je prekvapujúce vzhľadom na povahu tohto algoritmu.

V [4] išlo o hybridný fuzzy-genetický systém. Porovnávali to, ako rozdelenie na testovaciu a trénovaciu vzorku ovplyvní presnosť. Taktiež porovnávali to, ako počet pravidiel ovplyvní presnosť. Poukázali na to, že aj s menším počtom jednoduchých pravidiel dosiahli lepšiu

presnosť než iné, v tej dobe publikované výskumy. Presnosť, ktorú s týmto hybridným systémom dosiahli bola 97.80%.

[5] taktiež využíval hybridný systém, konkrétne Fuzzy-AIS-kNN systém, kde sa spájal fuzzy autoimunitný algoritmus s algoritmom kNN. Išlo o pokus navrhnúť nový hybridný systém. Fuzzy AIS mal za úlohu spracovať a redukovať dáta pre kNN algoritmus. Tým sa dosiahol menšší počet trénovacích dát pre kNN čo viedlo aj k zníženiu klasifikačného času kNN. Presnosť, s ktorou tento systém klasifikoval pacientov s rakovinou prsníkov bola 99.56%, takže sa dá zhodnotiť, že navrhnutá architektúra má potenciál.

[6] porovnáva viaceré typy neurónových sietí pri predikcii rakoviny mliečnych žliaz. Išlo o MLP (Multilayer Perceptron), RBF (Radial Base Function) siete, PNN (Probabilistic Neural Networks) a GRNN (Generalized Regression Neural Networks). Na trénovanie bolo využitých 50% dát z datasetu. Najlepšiu presnosť klasifikácie testovacích dát dosiahla GRNN – 98.8%. Druhá najlepšia bola PNN – 97.0%. RBF mala 96.18% a MLP 95.74%. Vo výsledku štúdia hodnotí, že využitie neurónových sietí na tento typ problému má zmysel a dokonca by mohli byť využívané na diagnostiku priamo onkológmi.

3. Riešenie

Z metód výpočtovej inteligencie sme sa rozhodli využili RBF sieť a TSK regulátor, ktoré sme programovali v programovacom jazyku Python. Webstránka bola deploynutá na Azure Portal.

3.1 Dataset

Dáta, s ktorými sme pracovali boli z *Breast Cancer Wisconsin Data* datasetu. Jednotlivé príznaky sú vypočítavané z digitalizovaného obrazu prsníka a popisujú charakteristiky bunkových jadier nachádzajúcich sa na obraze. Tieto príznaky sú polomer, textúra, obvod, plocha, hladkosť, kompaktnosť, konkávnosť, konkávne body, symetria a fraktálny rozmer, ktoré sú vyjadrené reálnymi číslami. Z nich boli vypočítané priemery, štandardné odchýlky a najväčšie priemerné hodnoty, čo viedlo k celkovému počtu 30 príznakov.

Cieľovým atribútom je *diagnosis*, ktorý môže nadobudnúť len 2 hodnoty – M (malignant – zhubný nádor) alebo B (benign – nezhubný nádor).

Rozdelenie vzoriek do tried je: 357 zhubných (B), 212 nezhubných (M).

3.2 RBF sieť

Ide o doprednú neurónovú sieť s tromi vrstvami – vstupná, jedna skrytá a výstupná. Aktivačná funkcia skrytej vrstvy je niektorá z nelineárnych radiálnych funkcií (najčastejšie Gaussova, multikvadratická). Neuróny skrytej vrstvy predstavujú „prototyp“ – príklad z trénovacej množiny. Často sa označuje ako stred neurónu, pretože napr. pre Gaussovu aktivačnú funkciu predstavuje hodnotu zodpovedajúcu stredu zvonovej krivky. Výstupom z tejto siete je lineárna kombinácia radiálnej funkcie vstupov a parametrov neurónov.

Keď chceme klasifikovať nový vstup, každý neurón si vypočíta Eukidovskú vzdialenosť medzi svojim „prototypom“ a vstupom. Čím menšia vzdialenosť je medzi nimi, tým skôr patria do rovnakej triedy.

Pre naše zadanie sme ako aktivačnú funkciu využili Gausssovu funkciu $e^{-\frac{||x-c||^2}{2\sigma^2}}$.

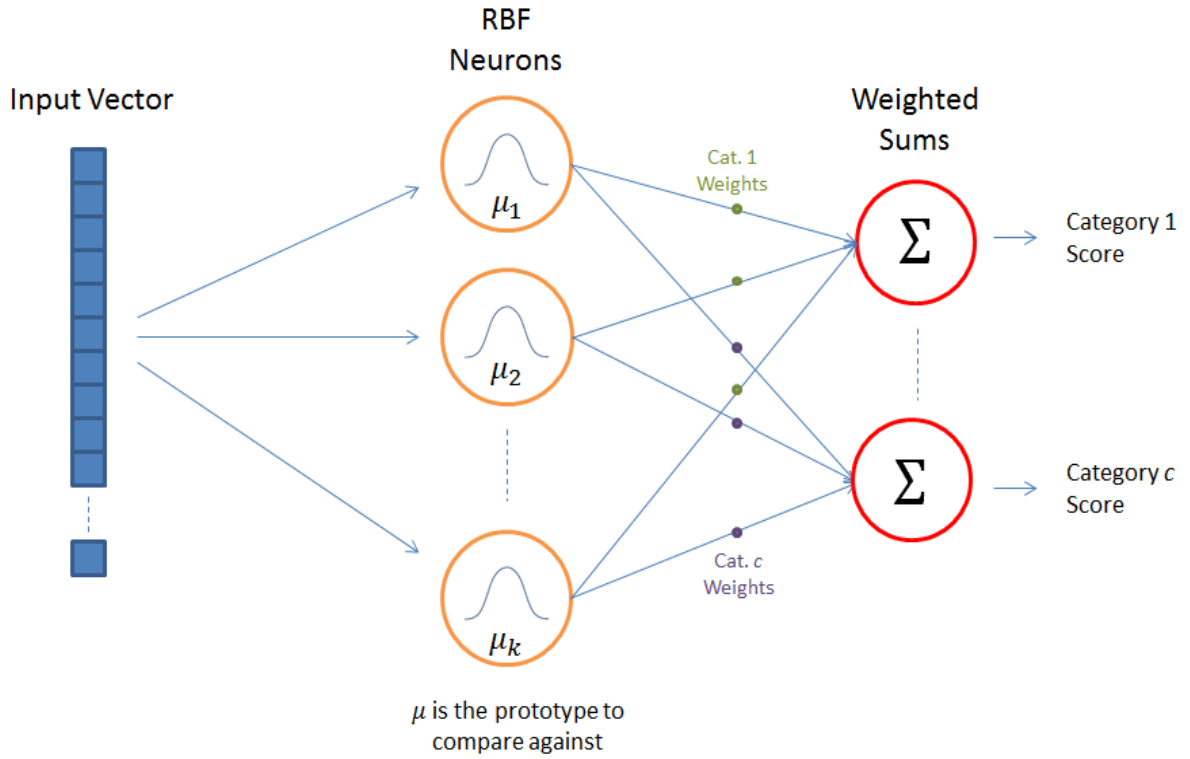

Figure 2 Schéma RBF siete

3.3 TSK regulátor a jeho prepojenie s RBF sieťou

TSK (Takagi-Sugeno-Kang), často označovaný ako Sugeno regulátor je jeden z dvoch základných typov fuzzy regulátorov. Jeho výhodou oproti Mamdaniho regulátoru je, že je výpočtovo efektívnejší a je lepšie uspôsobený na využitie optimalizačných a adaptívnych metód.

TSK regulátor je ekvivalentný s RBF sieťou pri splnení nasledujúcich podmienok:

1. Pravidlá bázy znalostí TSK regulátora v tvare

   (q:) Ak x_1 je LX_1^q& ... &$x_m$je LX_{mq} Potom $u^*_{1q}=b_{1q}$& ... & $u^*_{nq}=b_{nq}$
2. Nosiče funkcií príslušnosti pravidiel pre TSK regulátor musia byť nekonečne veľké.
3. Na lingvistických premenných sú fuzzy partície.
4. Hodnoty ϕ_q sú normalizované.
5. Výsledok agregácie LX_1^q x...x Lx_m^q = ϕ_q.
6. $\sum_{q=1}^{m} \phi_q$.

Ak je TSK ekvivalentný s RBF, vieme zo siete extrahovať znalosti slúžiace na návrh produkčných pravidiel regulátora. Extrakcia prebieha v dvoch krokoch:

1. Pre každý neurón skrytej vrstvy vykonáme jeho projekciu na lingivstickú premennú, čím získame predpokladovú časť pravidla.
2. Dôsledkovú časť pravidla získame tak, že vezmeme váhu zo spojenia daného neurónu, ktorý uvažujeme s výstupným neurónom.

Takto dostaneme pravidlo v tvare (q:) Ak x_1 je LX_1^q& ... &$x_m$je LX_{mq} Potom $u^*_{1q}=b_{1q}$& ... & $u^*_{nq}=b_{nq}$.

3.4 Postup riešenia

Nakoľko ide o neuro-fuzzy systém, prv musia dáta prejsť RBF sieťou. Po načítaní dát zo súboru bolo potrebné upraviť cieľové triedy aby malignant (zhubný) M=[1,0] a benign (nezhubný) B=[0,1]. Ďalšie spracovanie dát sa týkalo ich škálovania. Potom sa dáta rozdelili na trénovaciu a testovaciu vzorku v pomere 80:20. Z trénovacích dát sa vygenerovalo 5 prototypov – centier zhlukov pre obe triedy, pre ktoré sa potom počítala ich maximálna vzájomná vzdialenosť - sigma. Ďalší krok je trénovanie siete s využitím už spomínanej Gaussovej aktivačnej funkcie. Výstup z RBF siete je v tvare [x, y], pričom x, y $\in \langle 0,1 \rangle$ čo predstavuje príslušnosť ku triede M (zhubný) pre x a príslušnosť ku triede B (nezhubný) pre y, pričom 1 reprezentuje maximálnu príslušnosť a 0 minimálnu.

Báza pravidiel TSK regulátora bola navrhnutá tak, ako je popísané v podkapitole 3.3, teda na základe váh smerujúcim k neurónom výstupnej vrstvy naučenej RBF siete a neurónov skrytej vrstvy. Avšak využitie takto navrhnutého regulátora neprinieslo zvýšenie presnosti, práve naopak. Presnosť nepresiahla ani 50%.

Webová stránka obsahuje 4 obrazovky – Home, About, Downloads, Test. Na Home (Domovská obrazovka) sa nachádza stručný popis nášeho projektu s možnosťou presmerovať sa na ostatné obrazovky. Taktiež sa tu nachádza odkaz na nami využívaný dataset Breast Cancer Wisconsin.

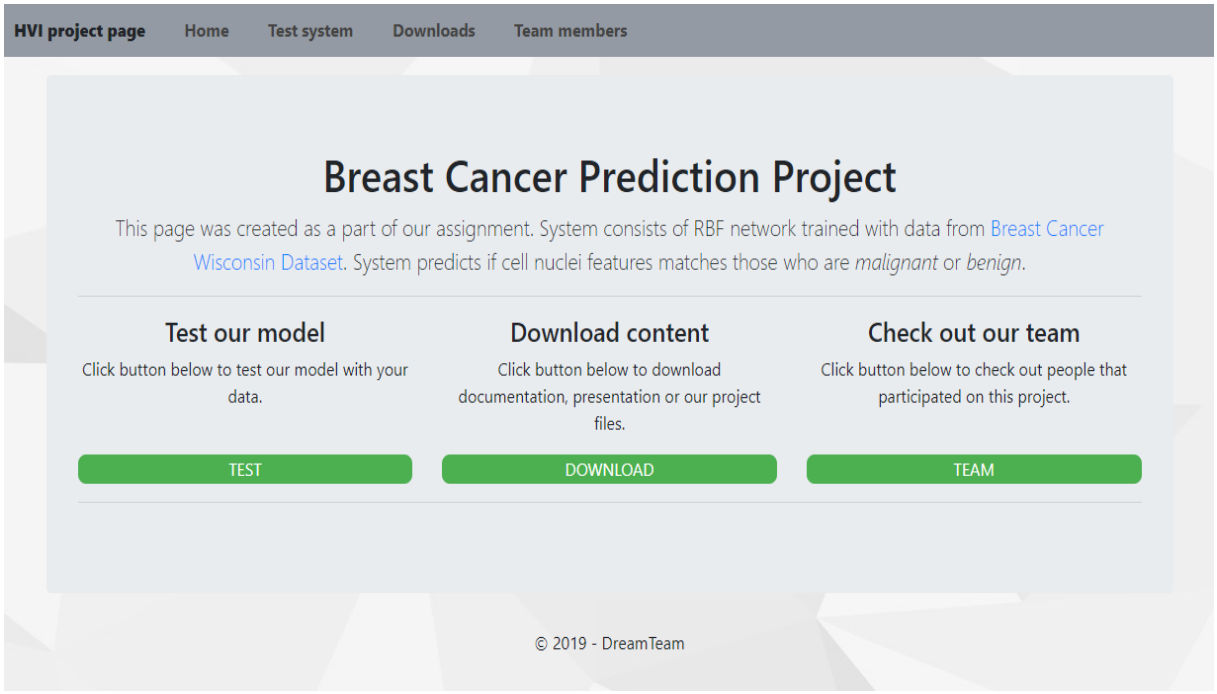

Figure 3 Domovská obrazovka

About (O nás) obsahuje fotografie a mená členov tímu. Obrazovka Downloads (Na stiahnutie) slúži na stiahnutie dokumentácie projektu rovnako ako prezentácie, ktoré sme mali pripravené. Taktiež sa tu nachádza odkaz na náš projekt na Git-e. Poslednou a z hľadiska projektu najdôležitejšou obrazovkou je Test (Testovanie), ktorá slúži na testovanie RBF siete. Po načítaní vlastného súboru dát sa vypíše výsledok klasifikácie.

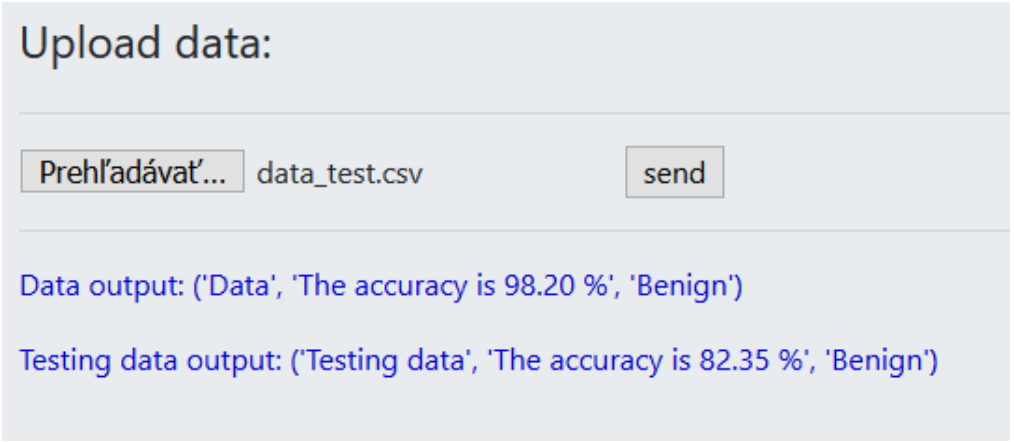

Figure 4 Výsledok načítania vlastných dát

Formát dát, ktoré sú zo stránky načítavané, je csv (comma-separated values), čiže riadok, kde sú jednotlivé hodnoty oddelené čiarkou.

Riadok dát, ktoré sú zadané v data_test.csv vo Figure 4:
924084, B, 12.77, 29.43, 81.35, 507.9, 0.08276, 0.04234, 0.01997, 0.01499, 0.1539, 0.05637, 0.2409, 1.367, 1.477, 18.76, 0.008835, 0.01233, 0.01328, 0.009305, 0.01897, 0.001726, 13.87, 36, 88.1, 594.7, 0.1234, 0.1064, 0.08653, 0.06498, 0.2407, 0.06484

Ako je vidieť, atribútov je veľa, čo by predstavovalo problém pre užívateľský vstup, preto sme sa rozhodli vkladať dáta na testovanie vo forme v akej sú v datasete. Ak by bola aplikácia mierená pre využitie v reálnom svete, bolo by potrebné vylepšiť interface, aby obsahoval textové polia, do ktorých by používateľ-expert zadával hodnoty vlastností jadier. Potenciálne by čítanie dát z .csv mohlo ostať ako je, ak by sa tento proces automatizoval.

4. Výsledky a zhodnotenie

Snaha navrhnúť bázu znalostí TSK regulátora na základe výstupu RBF siete sa ukázala vo výsledku neúspešná, presnosť bola nižšia než po klasifikácii len RBF neurónovou sieťou. Ako je vidieť na obrázkoch nižšie, presnosť RBF siete bola 95.50%, zatiaľ čo presnosť TSK regulátora po návrhu bázy znalostí z výstupu RBF siete bola pod 50%. To mohlo byť spôsobené veľkým počtom príznakov, s ktorými sme pracovali (30). Nakoľko v praxi je dôležitejšia presnosť, rozhodli sme sa, že webová stránka bude obsahovať len natrénovanú RBF sieť a teda výstup, ktorý sa na obrazovke po načítaní vlastných dát zobrazí je výstupom tejto siete. Na Figure 5 vidieť výstup z RBF siete. Tento výpis nezobrazujeme na stránke, slúžil nám na kontrolu. Prvý riadok každého zápisu predstavuje [x y], ktoré vyjadrujú príslušnosť k jednotlivým triedam zhubný/nezhubný. Druhý riadok hovorí o triede, ktorej sú dáta v datasete priradené. Tretí riadok popisuje triedu, ku ktorej tieto dáta priradil náš algoritmus.

```
---------------------------------- ----------------------------------
[0.0622289  0.94025717]            [0.17675475 0.82937399]
Class is  B                        Class is  B
Given Class  M                     Given Class  B
---------------------------------- ----------------------------------
[-0.11626759  1.13386133]          [0.08744454 0.97075405]
Class is  B                        Class is  B
Given Class  B                     Given Class  B
---------------------------------- ----------------------------------
[0.72225019 0.29531284]            [0.417555  0.5752137]
Class is  M                        Class is  B
Given Class  M                     Given Class  B
---------------------------------- ----------------------------------
[0.1834392  0.82584882]            [0.02776146 1.02694063]
Class is  B                        Class is  B
Given Class  B                     Given Class  B
---------------------------------- ----------------------------------
```

Figure 5 Príklad klasifikácie testovacích dát RBF sieťou

Na Figure 6 môžete vidieť kontigenčnú tabuľku, ktorá je výsledkom nad testovacou množinou, ktorá predstavovala 20% z datasetu. Trieda 1 predstavuje zhubné jadro a 2 predstavuje

nezhubné. Je vidieť, že algoritmus sa viac mýlil v zaraďovaní nezhubných jadier, čo bolo zrejme spôsobené tým, že dataset obsahoval o viac než 100 príkladov tejto triedy menej. Dosiahnutá presnosť klasifikácie je 95.50%.

Predicted	1	2
Actual		
1	37	1
2	4	69

The accuracy is 95.50 %

Figure 6 Kontigenčná tabuľka pre RBF

5. Potenciál využitia v praxi

Ako bolo spomínané v kapitole 2, analýza dát a hlavne tých medicínskych je často otázkou života a smrti. Vieme si predstaviť, že s vylepšeniami frontendu, aby bol viac prispôsobený potrebám potenciálneho používateľa, by aj náš natrénovaný model mohol slúžiť ako pomôcka pri identifikácii potenciálneho výskytu rakoviny mliečnych žliaz. Rozhodne netvrdíme, že by mohol nejakým spôsobom nahradiť alebo prevážiť slovo odborníka v danej oblasti.

Natrénovaním RBF siete na datasete súvisiacom s iným ochorením dosiahneme predikčný model iného ochorenia. Otázne ostáva, kde je hranica presnosti, s ktorou sme ochotní systémy pustiť do prevádzky.

Sme toho názoru, že systémy využívajúce rôzne metódy strojového učenia a umelej inteligencie sa postupne budú dostávať do bežného života.

Záver

Výsledkom zadania mal byť neuro-fuzzy systém pre predikciu rakoviny mliečnych žliaz. Išlo o klasifikačný problém s dvomi triedami – zhubný/nezhubný. Systém bol naprogramovaný v programovacom jazyku Python, TSK regulátor sa nepodarilo sfunkčniť tak, aby dosiahol dostatočnú prensosť, avšak už samotná RBF sieť klasifikovala s vysokou presnosťou.

Zdroje

[1]. IARC. World cancer report: International agency for research on cancer. 2008

[2]. Chaurasia, Vikas et al.: Prediction of benign and malignantbreast cancer using data miningtechniques. 2018

[3]. Borges, Lucas: Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. 2015

[4]. Peña-Reyes, C. A., Sipper, M.: A fuzzy-genetic approach to breast cancer diagnosis. 1999

[5]. Şahan, S. et al.: A new hybrid method based on fuzzy-artificial immune system and -nn algorithm for breast cancer diagnosis. 2007

[6]. Kiyan, Tuba: Breast Cancer Diagnosis Using Statistical Neural Networks. 2004