Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

DIPLOMOVÁ PRÁCA

na tému

Vývoj komponentu expertného systému pre analýzu lekárskych údajov s využitím zhlukovania

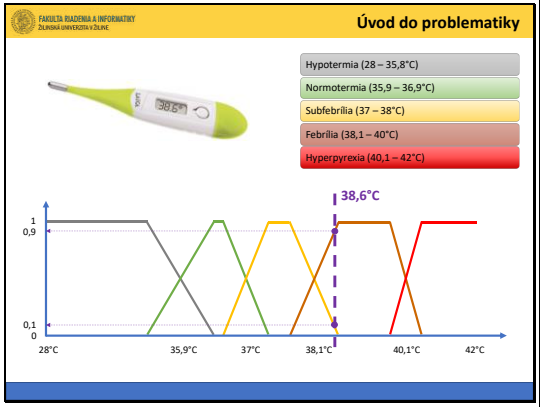
Bc. Volodymyr Ponomarenko

Vedúci: Prof., Dr. Paul Barach, MD, MPH
Tútor: prof. Ing. Vitaly Levashenko, PhD.

Cieľom práce je vývoj softvéru pre transformáciu numerických hodnôt na lingvistické, základom ktorej je zhlukovanie.

V analýze lekárskych údajov často získanie spojitej hodnoty nedáva dostatočnú informáciu pre správne určenie zdravotného stavu pacienta a predpísanie mu potrebného postupu liečby, lebo zdravotný stav sa popisuje kvalitatívnymi hodnotami. Teda je potrebným zabezpečiť kvalitatívny popis získanej numerickej hodnoty, inými slovami, transformovať numerickú hodnotu na zodpovedajúcu jej lingvistickú. Také lingvistické hodnoty môžu byť nasledovne relatívne jednoducho konvertované na neurčité dáta, zadané funkciou príslušnosti. Celý proces premenenia numerických hodnôt na fuzzy hodnoty má názov fuzzifikácia.

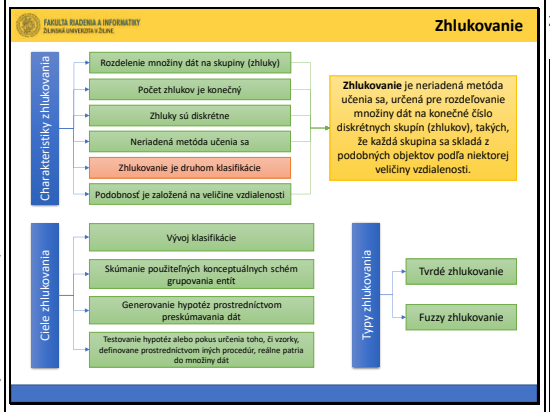
Na nasledovnej snímke ako príklad je uvedená telesná teplota človeka.



Nech u niektorého človeka bola nameraná teplota 38,6 stupňov. Síce pre lekára už samotná táto numerická hodnota nesie v sebe cennú informáciu o pacientovi, pre strojové učenie je viac dôležitá lingvistická hodnota, ktorá popisuje stav pacienta, lebo rozhodovanie sa uskutočňuje na základe toho, do ktorej skupiny respektíve do ktorých skupín patrí konkrétna hodnota telesnej teploty.

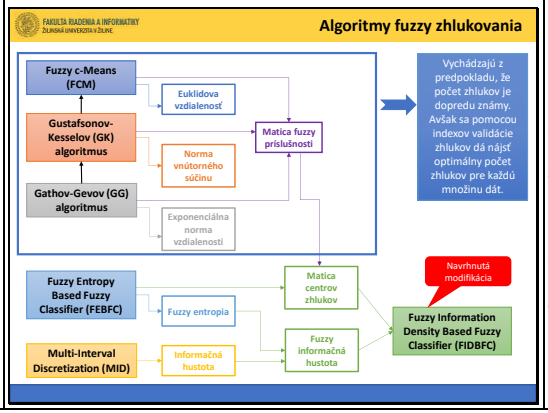
V diplomovej práci bola urobená analýza existujúcich definícií pojmu zhlukovania, výsledkom ktorej sú uvedené na ďalšej snímke najrozšírenejšie v literatúre charakteristiky zhlukovania.

Pod „zhlukovaním“ teda môžeme rozumieť neriadenú metódu učenia sa, určenú pre rozdeľovanie množiny dát na konečné číslo diskrétnych skupín (zhlukov), takých, že každá skupina sa skladá z podobných objektov podľa niektorej veličiny vzdialenosti.



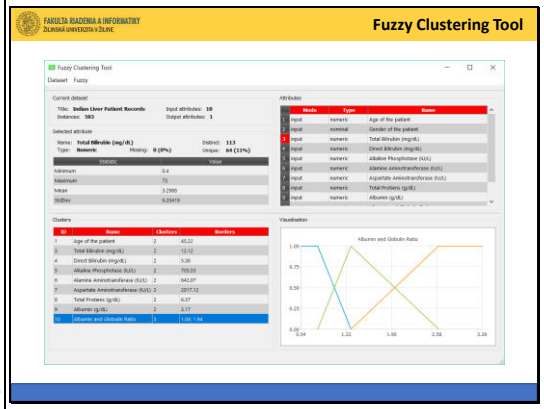
V diplomovej práci podrobne sa rozoberajú nasledovne algoritmy zhlukovania: Fuzzy c-Means (FCM), Gustafsonov-Kesselov (GK) algoritmus, Gathov-Gevov (GG) algoritmus, Multi-Interval Discretization (MID) a Fuzzy Entropy Based Fuzzy Classifier (FEBFC).

Prvých tri algoritmy sa považujú za klasické algoritmy fuzzy zhlukovania. Sú založené na hľadaní optimálnej matice fuzzy príslušnosti. Pôvodný FCM algoritmus používa Euklidovu vzdialenosť vo výpočtoch. GK algoritmus je jeho modifikáciou a používa normu vnútorného súčinu namiesto Euklidovej vzdialenosť, a GG pre tie isté účely používa exponenciálnu normu vzdialenosti. Tieto algoritmy ale vychádzajú z predpokladu, že počet zhlukov je dopredu známy. Avšak sa pomocou indexov validácie zhlukov dá nájsť optimálny počet zhlukov pre každú množinu dát.



MID algoritmus je založený na informačnej hustote. FEBFC hľadá optimálnu maticu centrov zhlukov pomocou fuzzy entropie. Pri tom fuzzy entropie jednotlivých zhlukov sa jednoducho sčítavajú, čo môže viesť k nepresným výsledkom zhlukovania. Nahradením fuzzy entropie fuzzy informačnou hustotou s použitím váhových koeficientov pri sčítaní bol urobený pokus o zlepšenie presnosti a tak bol získaný algoritmus Fuzzy Information Density Based Fuzzy Classifier.

Pre dosiahnutie cieľov diplomovej prace bol vyvinutý softvér Fuzzy Clustering Tool, hlavne okno ktorého je zobrazene na snímke:



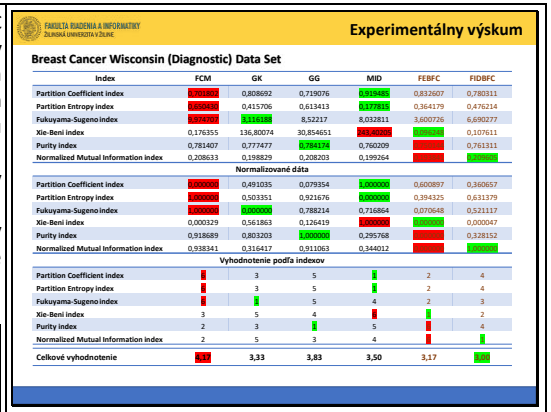
Softvér môže slúžiť ako komponent Expertného Systému. Vykonáva transformáciu numerických hodnôt na lingvistické.

Umožňuje načítavanie množiny dát zo súboru; základnú analýzu prvotnej množiny dát; grafickú vizualizáciu výsledkov základnej analýzy; importovanie výsledkov zhlukovania, získaných v iných nástrojoch; vypočítanie funkcií príslušností; grafickú vizualizáciu výsledkov zhlukovania; zapisovanie výsledkov fuzzifikácie do súboru.

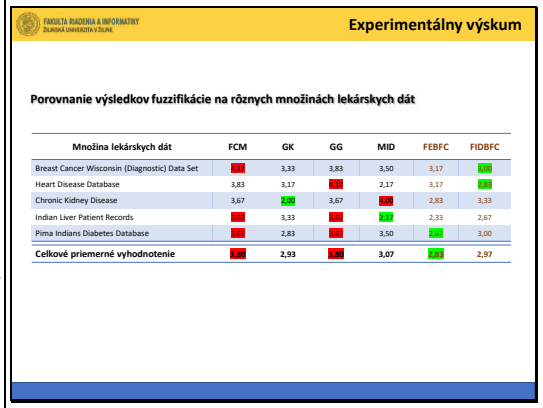
Hlavné okno nástroja sa skladá z nasledovných časti:

- Ovládací panel
- Krátka informácia o načítanej množine dát
- Zoznám atribútov množiny dát
- Podrobná informácia o vybranom atribúte
- Výsledok zhlukovania
- Vizualizácia fuzzy zhlukovania vybraného atribútu

Pre účely experimentálneho výskumu boli použité lekárske množiny dát z rôznych lekárskych odborov. Napríklad, Breast Cancer Wisconsin (Diagnostic) Data Set. Výsledky porovnania algoritmov fuzzy zhlukovania na množine Breast Cancer Wisconsin sú zobrazené na nasledovnej snímke. Vypočítané hodnoty indexov presnosti zhlukovania boli následne normalizované, na základe čoho každý z algoritmov bol ohodnotený od 1 do 6 podľa miery presnosti v porovnaní z inými algoritmami. Celkovo pre množinu dát Breast Cancer Wisconsin navrhnutá v práci modifikácia algoritmu fuzzy zhlukovania bola najpresnejšia v porovnaní z ostatnými algoritmami.



Podobný výskum bol urobený aj pre množiny dát Heart Disease Database, Chronic Kidney Disease, Indian Liver Patient Records, Pima Indians Diabetes Database. Tabuľka z výsledkami vyhodnotenia je predvedená na snímke:



Ako môžeme vidieť, na 2 z 5 množinách lekárskych dát navrhnutý algoritmus FIDBFC dáva najpresnejšie výsledky, čo je lepšie ako pôvodný algoritmus, ktorý je najpresnejší iba na jednej z množín.

Teda, síce priemerne pre všetky porovnávané množiny dát presnejším vychádza pôvodný algoritmus, môžeme však tvrdiť, že na niektorých lekárskych množinách dát navrhnutý FIDBFC algoritmus je vhodnejší pre transformáciu numerických hodnôt na lingvistické a fuzzy hodnoty.