

Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

DIPLOMOVÁ PRÁCA

na tému
**Vývoj komponentu expertného systému pre analýzu
lekárskych údajov s využitím zhlukovania**

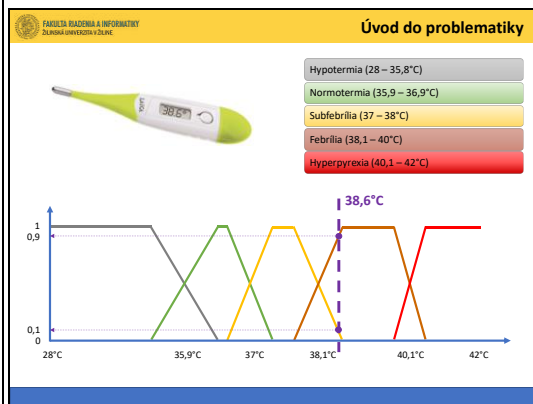
Bc. Volodymyr Ponomarenko

Veďúci: Prof. Dr. Paul Barach, MD, MPH
Tutor: prof. Ing. Vitally Levashenko, PhD.

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

V analýze lekárskych údajov často získanie spojenej 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 numerickej 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.

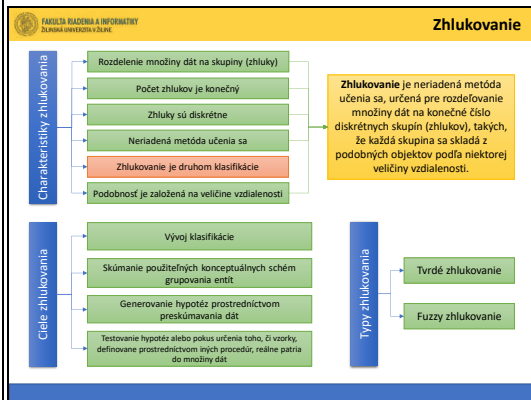
Na nasledovnej snímke ako príklad je uvedená telesná teplota človeka.



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.

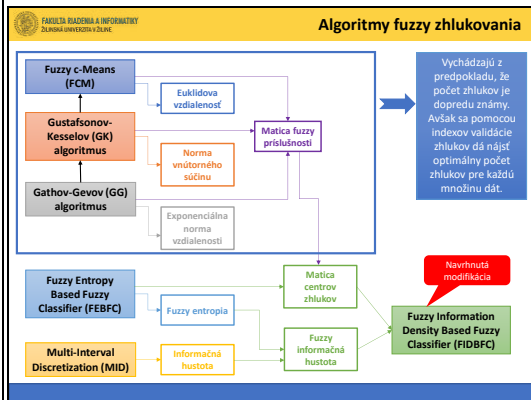
V diplomovej práci bola urobená analýza existujúcich definícií pojmu zhlukovania, výsledkom ktorej sú uvedené na ďalšej snímke najrozšírenejšie v literatúre charakteristiky zhlukovania.

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



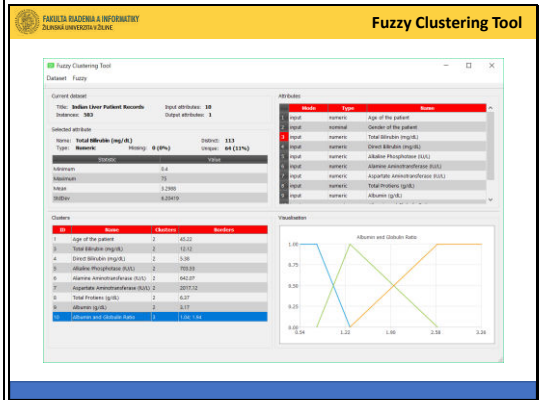
V diplomovej práci podrobne sa rozoberajú nasledovne algoritmy zhlukovania: Fuzzy c-Means (FCM), Gustafsonov-Kesselov (GK) algoritmus, Gathov-Gevov (GG) algoritmus, Multi-Interval Discretization (MID) a Fuzzy Entropy Based Fuzzy Classifier (FEBFC).

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 vzdialenosti, 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 sčítavaní bol urobený pokus o zlepšenie presnosti a tak bol získaný algoritmus Fuzzy Information Density Based Fuzzy Classifier.

Pre dosiahnutie cieľov diplomovej práce bol vyvinutý softvér Fuzzy Clustering Tool, hlavne okno ktorého je zobrazené 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 príslušnosti; grafickú vizualizáciu výsledkov zhlukovania; zapisovanie výsledkov fuzzifikácie do súboru.

Hlavné okno nástroja sa skladá z nasledovných častí:

- Ovládací panel
- Krátka informácia o načítanej množine dát
- Zoznam 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 lekárske 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 s inými algoritmi. Celkovo pre množinu dát Breast Cancer Wisconsin navrhnutá v práci modifikácia algoritmu fuzzy zhlukovania bola najpresnejšia v porovnaní s ostatnými algoritmi.

Experimentálny výskum

Breast Cancer Wisconsin (Diagnostic) Data Set

Index	FCM	GK	GG	MID	FEBFC	FIDBFC
Partition Coefficient Index	0.828692	0.216076	0.783207	0.832107	0.780111	0.780111
Partition Entropy Index	0.415706	0.613413	0.372723	0.364179	0.479214	0.479214
Fukayama-Sugeno Index	0.676791	0.511188	0.522117	0.032811	0.600726	0.650277
Xie-Beni Index	0.176355	136.80074	30.854601	0.25205	0.107611	0.107611
Purity Index	0.781407	0.198229	0.76514	0.762009	0.762009	0.762009
Normalized Mutual Information Index	0.208633	0.198229	0.208203	0.199264	0.199264	0.199264

Normalizované dáta

Index	FCM	GK	GG	MID	FEBFC	FIDBFC
Partition Coefficient Index	0.491035	0.079354	0.000000	0.000897	0.380657	0.380657
Partition Entropy Index	0.505351	0.921676	0.000000	0.394325	0.631379	0.631379
Fukayama-Sugeno Index	0.000000	0.788214	0.718684	0.070648	0.521131	0.521131
Xie-Beni Index	0.000279	0.563363	0.126419	0.000000	0.000047	0.000047
Purity Index	0.918849	0.805203	0.805203	0.250748	0.250748	0.250748
Normalized Mutual Information Index	0.918849	0.316417	0.911063	0.344032	0.344032	0.344032

Vyhodnotenie podľa indexu

Index	FCM	GK	GG	MID	FEBFC	FIDBFC
Partition Coefficient Index	3	5	5	2	4	4
Partition Entropy Index	3	5	4	2	4	4
Fukayama-Sugeno Index	3	5	4	2	4	4
Xie-Beni Index	2	3	3	5	4	4
Purity Index	2	5	3	4	4	4
Normalized Mutual Information Index	2	3	3	4	4	4
Celkové vyhodnotenie	3,33	3,83	3,50	3,17	3,17	3,17

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:

Experimentálny výskum

Porovnanie výsledkov fuzzifikácie na rôznych množinách lekárskych dát

Množina lekárskych dát	FCM	GK	GG	MID	FEBFC	FIDBFC
Breast Cancer: Wisconsin (Diagnostic) Data Set	3,33	3,83	3,50	3,17	3,17	3,17
Heart Disease Database	3,67	3,67	3,67	2,83	3,33	3,33
Chronic Kidney Disease	3,67	3,67	3,67	2,83	3,33	3,33
Indian Liver Patient Records	3,33	3,33	3,33	3,33	3,67	3,67
Pima Indians Diabetes Database	2,83	2,83	3,50	3,00	3,00	3,00
Celkové priemerne vyhodnotenie	3,33	2,93	3,07	3,07	2,97	2,97

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 porovnané množiny dát presnejší vychádza pôvodný algoritmus, môžeme však tvrdiť, že na niektorých lekárske množinách dát navrhnutý FIDBFC algoritmus je vhodnejší pre transformáciu numerických hodnôt na lingvistické a fuzzy hodnoty.