III Pasaules latviešu zinātnieku kongress

Datorzinātnes un informācijas tehnoloģiju apakšsekcija

Zināšanu iegūšanas iespējas ar klasterizācijas tehnoloģiju palīdzību

Pēteris Grabusts
Rēzeknes Augstskola

25. oktobris, 2011

ZINĀŠANU IEGUVES METODES

Intelektuālo datu analīzes sistēmās būtisks ir jautājums – kas ir zināšanas un ar ko tās atšķiras no datiem.

Dati – tie ir atsevišķi fakti, kas raksturo objektus vai procesus kādā konkrētā pielietojuma jomā, kā arī to īpašības.

Zināšanas – dažādas definīcijas:

•„zināšanas ir informācija – saistīta ar likumiem, kas ļauj automātiski izdarīt slēdzienus, lai informācija varētu tikt pielietota lietderīgiem mērķiem";

•„zināšanas var tikt attēlotas kā likumi un fakti. Likums ir apgalvojums, kas ir vai nu patiess, vai aplams un satur IF daļu un THEN daļu. Fakts ir beznosacījuma apgalvojums, kas pieņemts kā patiess laikā, kad tas tiek lietots";

•„zināšanas – cilvēka pielietojuma jomas likumsakarības (principi, saites, likumi), kas iegūti cilvēka praktiskās darbības un profesionālās pieredzes rezultātā".

MOTIVĀCIJA

- analizējamo daudzdimensiju datu apjoms kļūst pārāk liels klasiskās statistiskās analīzes iespējām;

- populārās neironu tīklu metodes darbojas pēc „melnās kastes” principa, kas lietotājam apgrūtina interpretēt iegūtos rezultātus;

- iepriekš nezināmu likumsakarību atrašana datos;

- iespēja izteikt atrastās likumsakarības lietotājam uztveramā un saprotamā veidā.

LIKUMU IEGŪŠANAS METODES

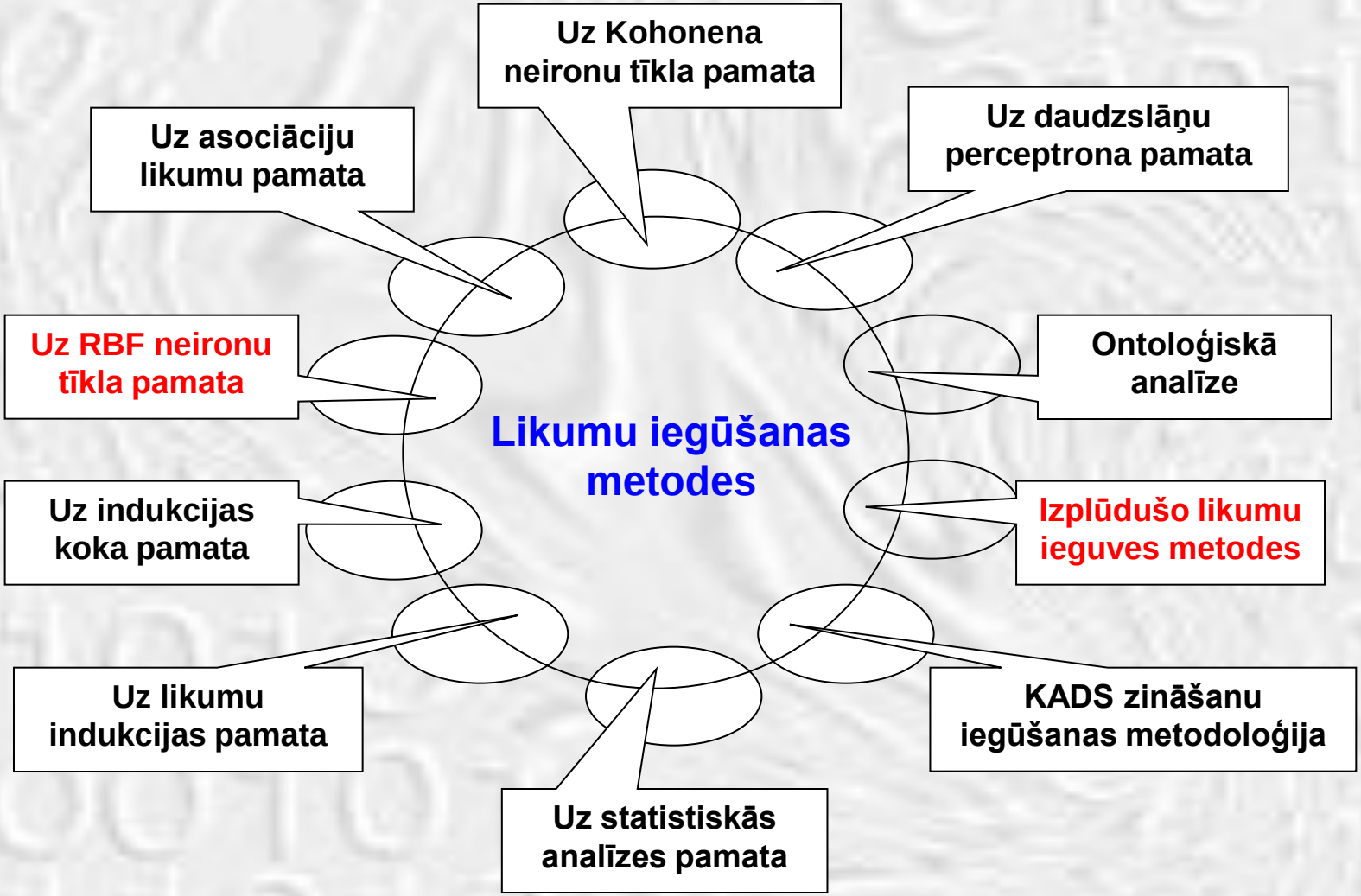
Galvenā prasība, kas tiek izvirzīta datu analīzes rezultātiem, ir tāda, ka vienmēr nepieciešams pēc iespējas korektāk interpretēt rezultātus. Likumi, kas izsaka atrastās likumsakarības, jāformulē vienkāršos un saprotamos loģisko izteikumu veidos, t.i., tiem jābūt šāda veida loģiskiem likumiem:

JA {(1. notikums) un (2. notikums) un ...(N. notikums)} TAD ...

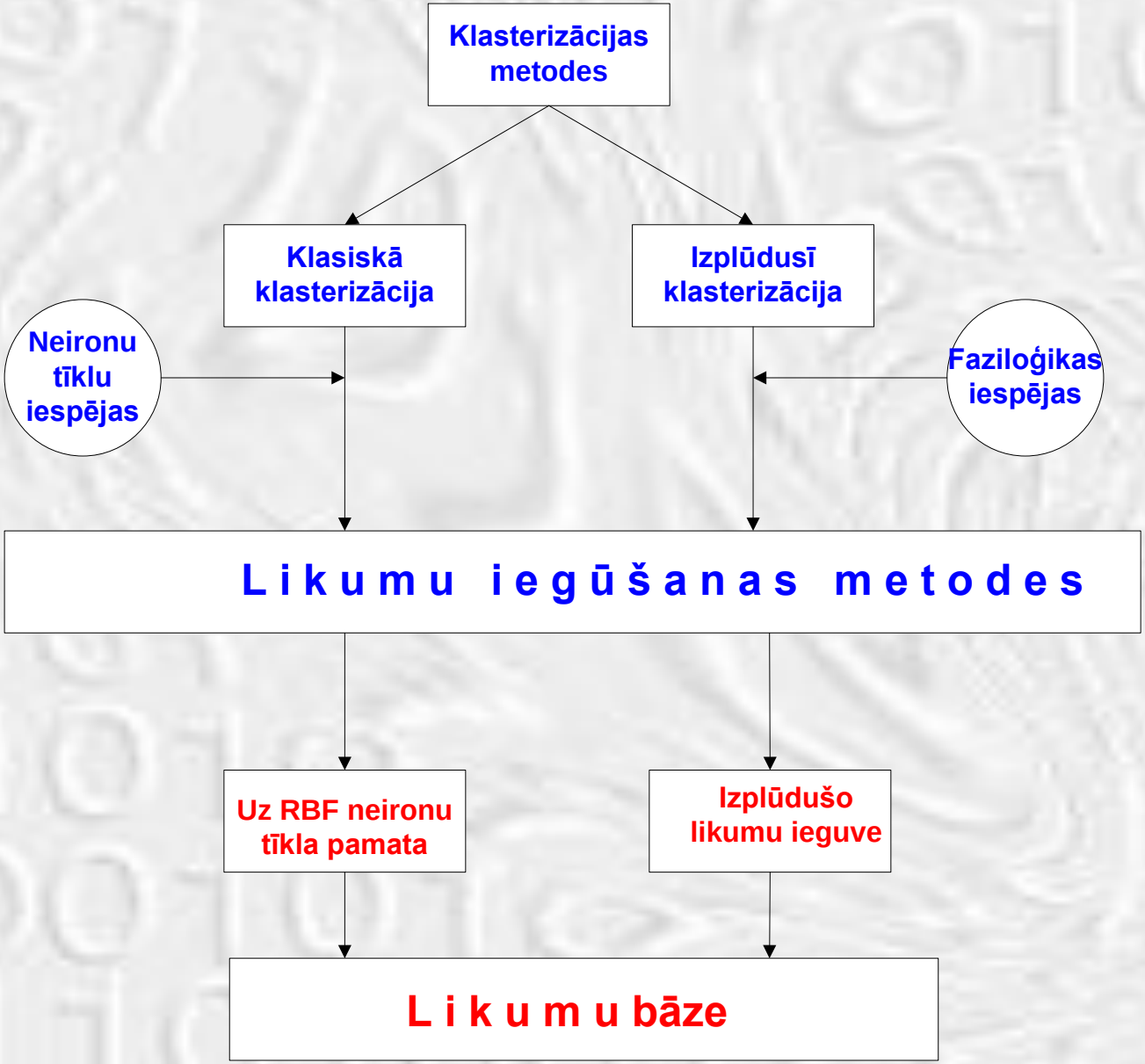
Vispārīgā veidā tie ir loģiskie nosacītie likumi (produkciju likumi):

$$IF \underbrace{(1.nosacījums)\,un\,(2.nosacījums)\,un\,...\,(N.nosacījums)}_{A}\,THEN\,\underbrace{(slēdziens)}_{B}$$

LIKUMU IEGŪŠANAS METODES



LIKUMU IEGŪŠANAS METODES



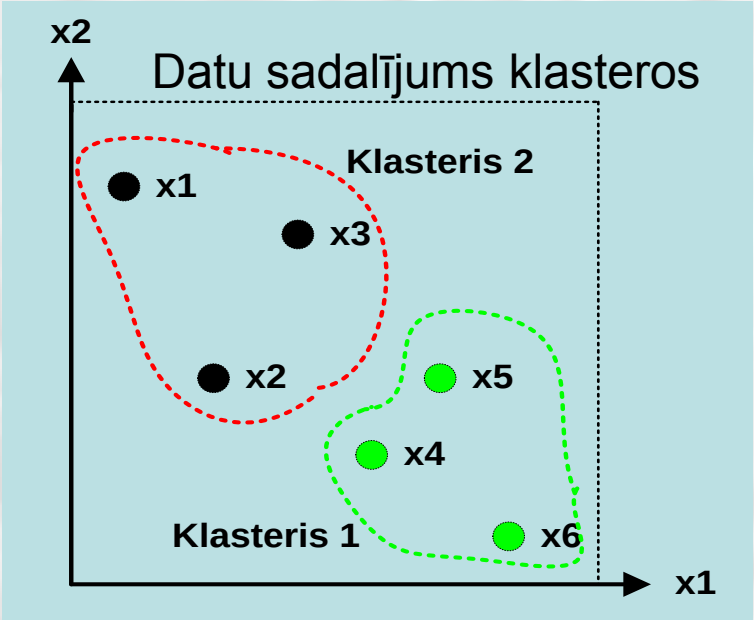
LIKUMU IEGŪŠANAS METODES

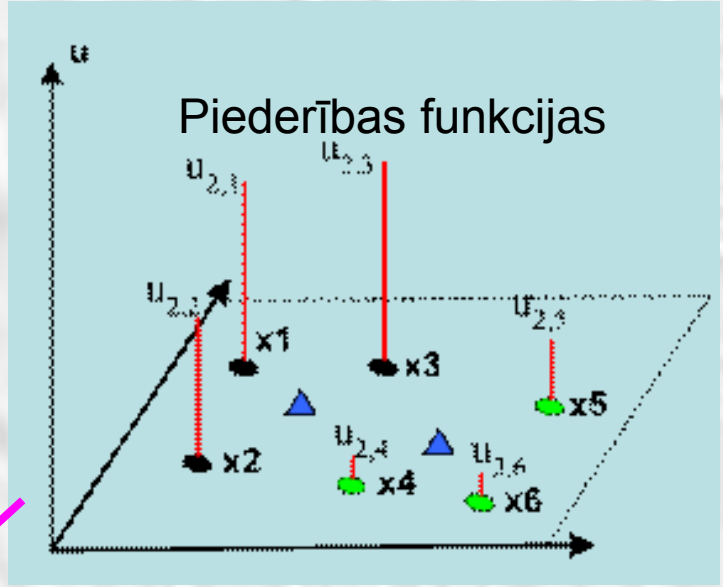
- **IZPLŪDUŠO LIKUMU BĀZES PROJEKTĒŠANA**
- **LIKUMU IEGŪŠANA NO APMĀCĪTIEM NEIRONU TĪKLIEM**

(Klasterizācijas pielietošana !)

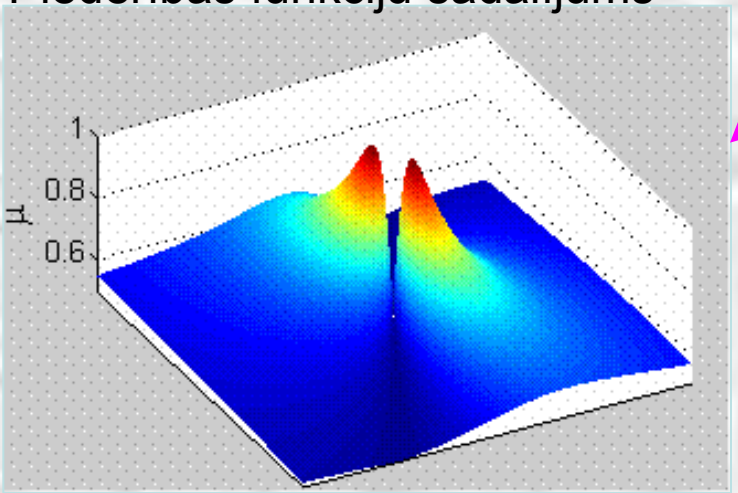
I. IZPLŪDUŠO LIKUMU BĀZES PROJEKTĒŠANA

IZPLŪDUSĪ KLASTERIZĀCIJA





Piederības funkciju sadalījums



FCM

Izplūdušais C-Mean algoritms (FCM) atšķiras no klasiskā C-Mean algoritma ar to, ka tas izmanto izplūdušu kopu elementus - klasteru punkti tiek nodalīti ar piederības pakāpi.

- Punkti starp diviem klasteru centriem nosaka piederības attiecības starp šiem klasteriem

- Piederības matrica - [0,1] :

$$m_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(q-1)}}$$

- **Mērķa funkcijas minimizācija :**

$$J(M, c_1, c_2, \ldots . c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k=1}^{K} m_{ik}^q d_{ik}^2$$

- Optimālā centra noteikšana:

$$c_i = \frac{\sum_{k=1}^{K} m_{ik}^q u_k}{\sum_{k=1}^{K} m_{ik}^q}$$

FCM ALGORITMS

FCM algoritms

(1)	Inicializē piederības matricu M ar gadījumvērtībām starp 0 un 1.
(2)	Izskaitļo c klasteru centrus c_i (i=1,2….c)
(3)	Izskaitļo mērķa funkciju. Ja sasniegts apstāšanās nosacījums – beigt darbu, citādi turpināt.
(4)	Izskaitļo jauno piederības matricu
(5)	Pāriet uz soli (2).

Tiks izmantots likumu bāzes konstruēšanā piederības funkciju iegūšanai

IZPLŪDUŠU LIKUMU BĀZES IEGŪŠANA

Izplūdušā klasifikatora būtība:

- bāzējas uz galīgu likumu R kopu, kam izpildās:

R: IF x_1 ir $\mu^{(1)}_R$ un ….. un x_p ir $\mu^{(p)}_R$ THEN klase ir C_R

kur $C_R \in C$. $\mu^{(i)}_R$ ir X_i izplūdusī kopa t.i. . $\mu^{(i)}_R : X_i \rightarrow [0,1]$

Likumu sadalījums:

$$R(x_1,....x_p) = \begin{cases} C, & ja\ \mu_C^{(R)}(x_1,...,x_p) > \mu_D^{(R)}(x_1,...,x_p)\ \ visiem\ D \in C,\ D \neq C \\ \notin C, & citādi \end{cases}$$

Divdimensiju gadījumā:

IF x ir μ1 un y ir v1 THEN klase ir A

IF x ir μ2 un y ir v2 THEN klase ir B

5 Etapi

IZPLŪDUŠU LIKUMU BĀZES IEGŪŠANA

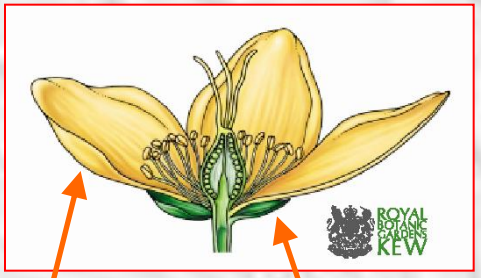
5 Etapi

1. Ieejas un izejas datu nodalīšana -Intervālu noteikšana: (2N+1)
2. Likumu konstruēšana uz apmācošās kopas datiem:
   - Katram apmācošo datu pārim var uzdot vienu likumu
3. Ticamības pakāpes noteikšana katram likumam.
4. Likumu bāze tabulas formā.
5. Defazifikācija

Iegūta likumu bāze

IRIS PIEMĒRS

IRIS datu bāze: satur trīs ziedu klases:
pa 50 elementiem katrā:
setosa, versicolor un virginica.



Petal Sepal

Katram ziedam ir 4 parametri:

SL – kauslapas garums *(sepal length)*;
SW – kauslapas platums *(sepal width)*;
PL - ziedlapas garums *(petal length)*;
PW – ziedlapas platums *(petal width)*.

Setosa			
SL	SW	PL	PW
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
.....	.....	.....	.....

Versicolor			
SL	SW	PL	PW
7.0	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.9	3.1	4.9	1.5
5.5	2.3	4.0	1.3
6.5	2.8	4.6	1.5
.....	.....	.....	.....

Virginica			
SL	SW	PL	PW
6.3	3.3	6.0	2.5
5.8	2.7	5.1	1.9
7.1	3.0	5.9	2.1
6.3	2.9	5.6	1.8
6.5	3.0	5.8	2.2
.....	.....	.....	.....

Setosa-lineāri atdalāma.
Versicolor un Virginica – nē.

PIEMĒRS

Tika izskaitļotas 3 piederības funkcijas 3 klasēm –
- iegūti 4 likumi I klasei, 3 likumi II klasei un 11 likumi III klasei (kopā 18)

Iegūto likumu fragments:

Likums 1: **IF X1 ir MF1 ar piederibu 1 AND X2 ir MF2 ar piederibu 0.83333 AND X3 ir MF1 ar piederibu 0.9661 AND X4 ir MF1 ar piederibu 1 THEN KLASE ir 1 ar piederibu 0.76483**

Likums 2: **IF X1 ir MF2 ar piederibu 0.77778 AND X2 ir MF3 ar piederibu 1 AND X3 ir MF1 ar piederibu 0.83051 AND X4 ir MF1 ar piederibu 0.75 THEN KLASE ir 1 ar piederibu 0.46024**

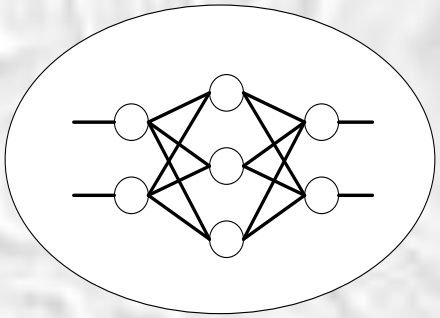
Likums 3: **IF X1 ir MF2 ar piederibu 0.66667 AND X2 ir MF2 ar piederibu 0.75 AND X3 ir MF1 ar piederibu 0.89831 AND X4 ir MF1 ar piederibu 0.91667 THEN KLASE ir 1 ar piederibu 0.39114**

Likums 4: **IF X1 ir MF1 ar piederibu 0.88889 AND X2 ir MF1 ar piederibu 0.75 AND X3 ir MF1 ar piederibu 0.89831 AND X4 ir MF1 ar piederibu 0.83333 THEN KLASE ir 1 ar piederibu 0.47411**

II LIKUMU IEGŪŠANA NO APMĀCĪTIEM NEIRONU TĪKLIEM

LIKUMU IEGUVES PROCESS

Neironu tīkls



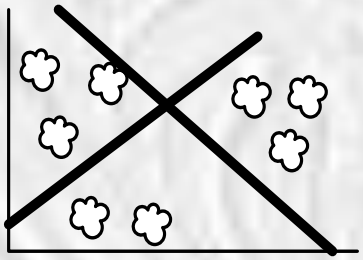
Tīkla parametri

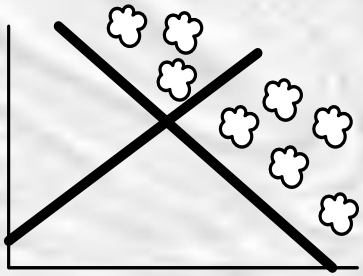
Svari

Sliekšņa vērtības

Analīzes algoritms

Klasifikācijas robežvērtības izteiktas kā likumi

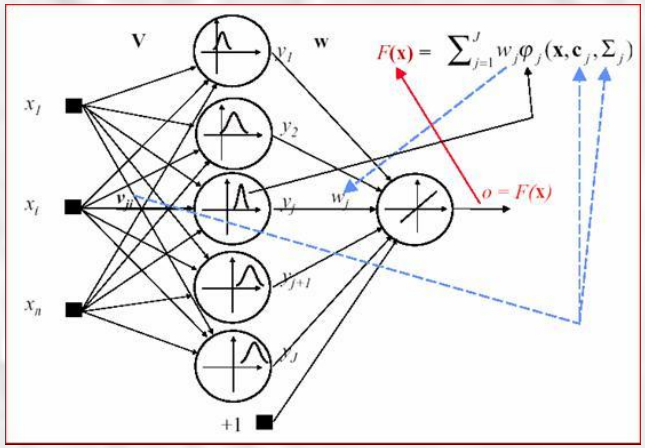




Konvertācija uz likumu bāzi

IF..
THEN
ELSE..

RBF NEIRONU TĪKLA PIEMĒROTĪBA



1. Slēptā slāņa neironi ar radiālo (Gausa) aktivācijas funkciju:

$$\Phi(x) = e^{-\frac{\|x_i - c_i\|}{\sigma^2}}$$

2. Katrs slēptais neirons producē vienu likumu formā:

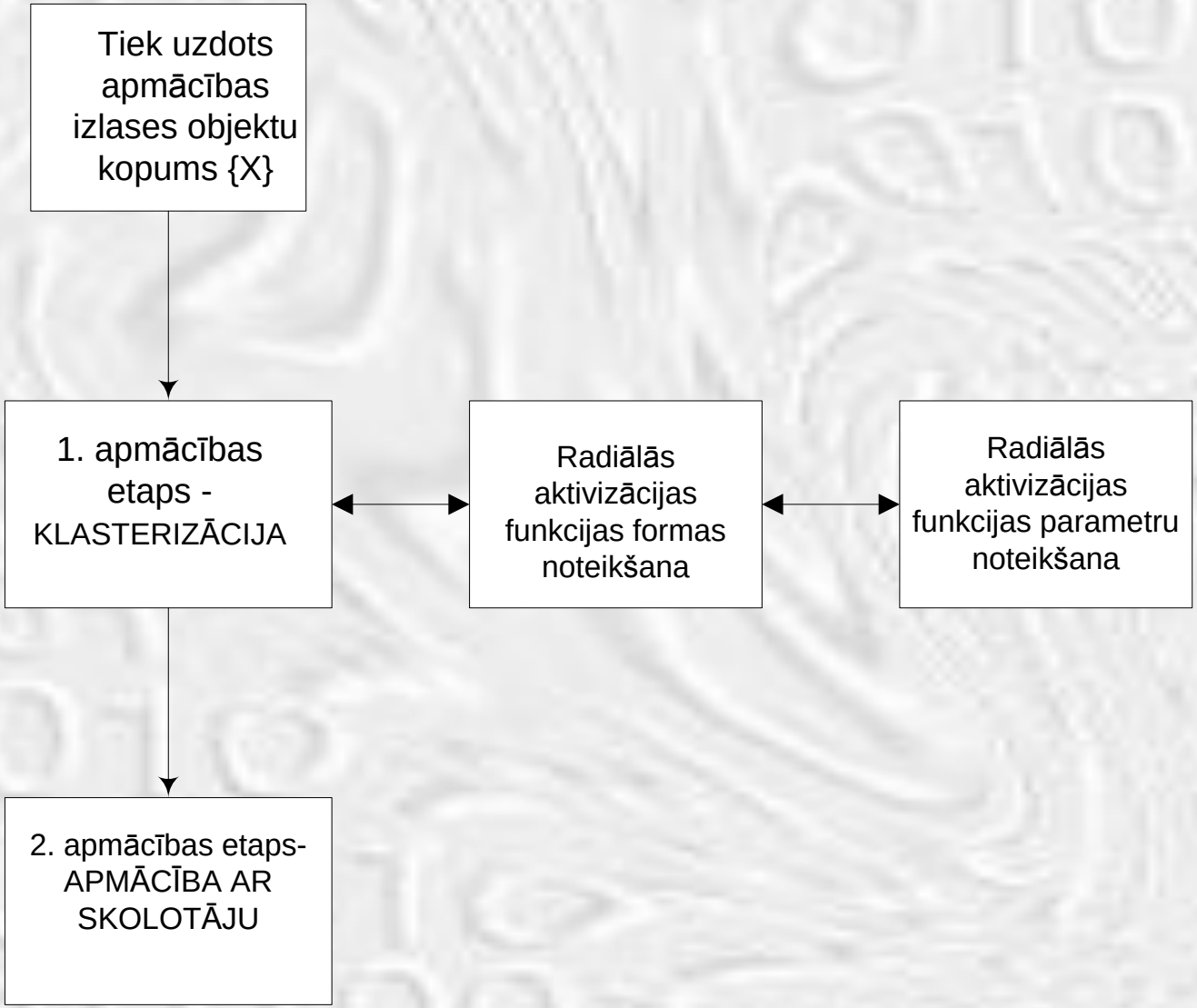
IF Feature$_1$ is TRUE AND
IF Feature$_2$ is TRUE AND
IF Feature$_n$ is TRUE
THEN Class$_x$

3. Elements Feature sastāv no:

- **izskaitļotajām RBF centru μ_n augšējām un apakšējām robežvērtībām**
- **RBF platuma jeb rādiusa σ**
- **Gausa funkcijas parametra S (*Steepness*)**

4. Likumu ieguves algoritms RULEX

RBF APMĀCĪBA

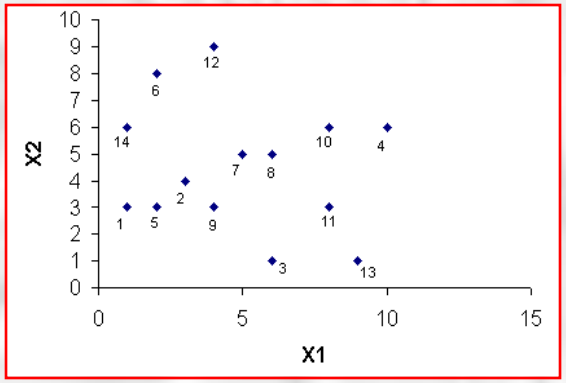


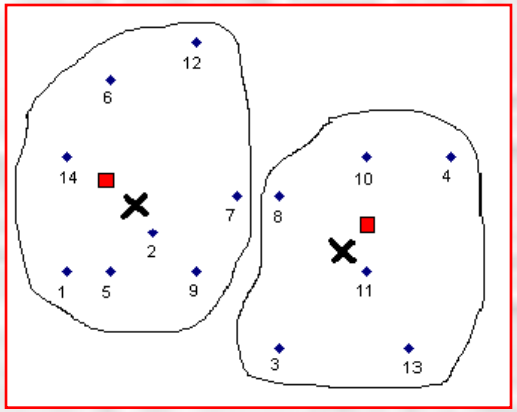
RULEX ALGORITMS

Ieeja:	RBF centra vērtības μ Gausa funkcijas platums σ Parametrs S
Izeja:	Viens likums katram slēptajam elementam
Procedūra:	RBF tīkla apmācības procedūra Katram slēptajam elementam: Katram μ_i $X_{lower} = \mu_i - \sigma_i + S$ $X_{upper} = \mu_i + \sigma_i - S$ Konstruējam likumu ar: antecedents=[X_{lower}, X_{upper}] Apvienojam antecedentus ar AND Pieliekam klases iezīmi Ierakstām likumu.

LIKUMU IEGŪŠANAS DEMONSTRĀCIJA

Dati piemēram:





Centri : μ_1=(-0.73; 0.26) un μ_2=(0.97;-0.35).

Rādiusi: σ_1^2 = 1.07 un σ_2^2 = 1.04

Cluster 1. X_{1_lower}= -0.73 - 1.03 + 0.6= - 1.16; X_{2_lower}=0.26 - 1.03 +0.6= -0.17;
X_{2_upper}= -0.73+ 1.03 - 0.6 = -0.3; X_{2_upper}=0.26 +1.03 - 0.6= 0.69.

Cluster 2. X_{1_lower}= 0.97 - 1.01 + 0.6= 0.56; X_{2_lower}= -0.35 - 1.01 +0.6= -0.76;
X_{2_upper}= 0.97+ 1.01 - 0.6 = 1.38; X_{2_upper}= -0.35 +1.01 - 0.6= 0.06.

Rezultāti pie S=0.6

IF ($x_1 \geq$ -1.16 AND $\leq$ -0.3) AND IF ($x_2 \geq$ - 0.17 AND $\leq$ 0.69) THEN CLASS 1
IF ($x_1 \geq$ 0.56 AND $\leq$ 1.38) AND IF ($x_2 \geq$ - 0.76 AND $\leq$ 0.06) THEN CLASS 2

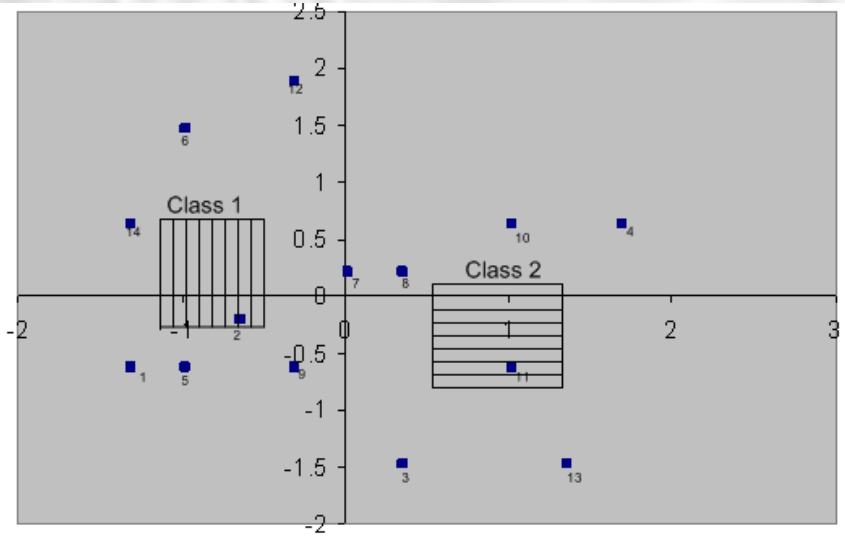
Rezultāti pie S=0

IF ($x_1 \geq$ -1.76 AND $\leq$ 0.3) AND IF ($x_2 \geq$ - 0.77 AND $\leq$ 1.29) THEN CLASS 1
IF ($x_1 \geq$ -0.04 AND $\leq$ 1.98) AND IF ($x_2 \geq$ - 1.36 AND $\leq$ 0.66) THEN CLASS 2

LIKUMU IEGŪŠANAS DEMONSTRĀCIJA

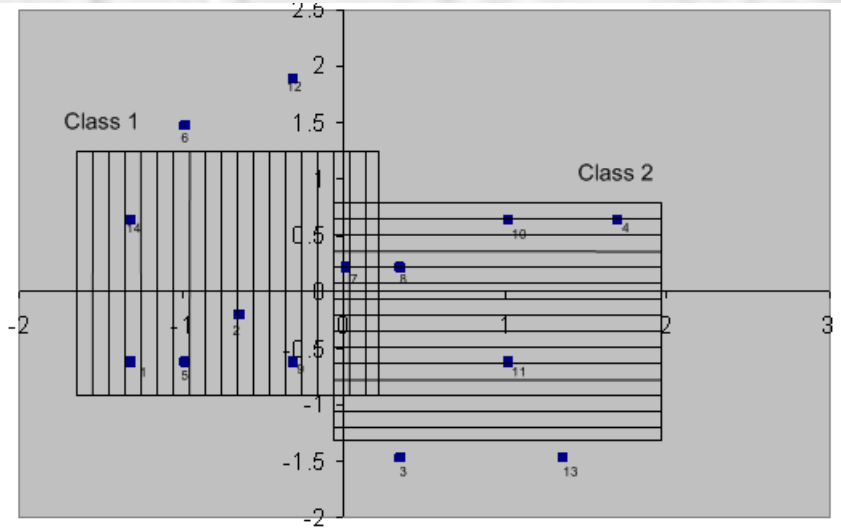
S=0.6

Kļūdaini - 12 ieejas vektori no 14 (86% kļūdu)



S=0

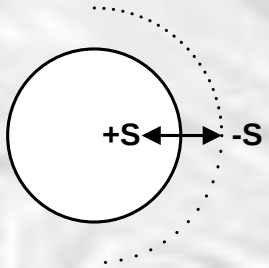
Kļūdaini - 4 ieejas vektori no 14 (28% kļūdu) - 3, 6, 12, 13 punkti



IRIS PIEMĒRS

Likumiem atbilstošo elementu kopskaita (%) atkarība no S

S	Parametra S vērtības												
	-0,8	-0,7	-0,6	-0,5	-0,4	-0,3	-0,2	-0,1	0	0,1	0,2	0,3	0,6
%	98,7	98,7	97,3	97,3	92,7	87,3	84	70,7	58,7	47,3	34,7	23,3	4



SECINĀJUMI

- Metodes uz RBF tīkla pamata un izplūdušo likumu metodes darbības rezultātus var salīdzināt tikai pēc iegūto likumu „kvalitātes” rādītāju pārbaudes, t.i., izskaitļojot likumus pareizi aprakstošo elementu skaitu.

- Klasterizāciju dažādās metodēs var izmantot likumu iegūšanā no daudzdimensiju datiem.

- Iegūtie likumi dod iespēju analizēt slēptās likumsakarības datos.

PALDIES PAR UZMANĪBU !