Проект по "Размити множества и приложения"

Автор: Любослав Карев

Въведение в решавания проблем и цел на проекта

Целта на проекта бе да се създаде програма, която използва размити клъстеризиращи алгоритми върху даден набор от данни, да се направи визуализация на тези данни, както и да се дефинират размити лингвистични правила, на база на резултатите от клъстеризацията.

Теоретична постановка и използван алгоритъм: Описание на приложения алгоритъм: дефиниции и извеждания, необходими за реализацията на поставената цел.

Алгоритъмът използван за клъстеризация е C-Means. Алгоритъмът се реализира в няколко стъпи:

1. Избор на брой клъстери
2. На случаен принцип се поставя степен на принадлежност на всеки един вектор от данните, към всеки един от клъстерите
3. Изчислява се центъра на всеки един от клъстерите
4. За всеки вектор от данните, се преизчислява неговата степен на принадлежност към всеки от клъстерите
5. Ако разликата между старите и новоизчислените коефициенти е по-голяма от предварително зададена константа, алгоритъма се връща на стъпка 3

   Център на клъстер се изчислява по следния начин:

$$c_k = \frac{\sum_{x \in X} w_k(x)^m x}{\sum_{x \in X} w_k(x)^m}$$

където k е клъстера, за който се изчислява центъра, X са данните ни, а m е параметър, показващ "размитостта" на множеството.

Степен на принадлежност на елемент i към клъстер j изчисляваме по следния начин:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}}$$

Лингвистичните правила са представени с функцията на принадлежност на Бел, като в общия случай:

$$bell(x; a, b, c) = \frac{1}{1 + \left|\frac{u-c}{a}\right|^{2b}}$$

В конкретната задача, за параметър c ще приемаме центъра на дадения клъстер c_i. Наклонът b ще бъде оставен като параметър равен на 2 (с възможност за промяна). Широчината a пък ще е разстоянието между конкретния клъстер c_i и най-близкия до него ($|c_i - c_j|$)

Описание на данните, предпроцесна обработка

Използваните данни са Iris данните - представящи три представителя на семейството цветя Iris - Iris-virginica, Iris-versicolor и Iris-setosa. Всеки от индивидите е представен със следните пет характеристики:

- височина на чашелистче (`sepal_height`)
- широчина на чашелистче (`sepal_width`)
- височина на венчелистче (`petal_height`)
- широчина на венчелистче (`petal_width`)
- клас към който принадлежи екземпляра (`class`)

	sepal_length	sepal_width	petal_length	petal_width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa

В базата от данни разполагаме с 150 екземпляра, разпределени равномерно между трите класа. Като част от предварителната обработка, ще премахнем информацията за класа от данните.

Експериментални/симулационни резултати: Представят се получени резултати и се визуализират - графично и/или таблично. Анализ на резултатите.

Основни изводи (идеи за бъдеща работа; възможни приложения на използваните подходи и реализиран алгоритъм).

Списък на използваната литература

Приложение: Код на програмната реализация