Universidade Federal de Pernambuco
Centro de Informática
Graduação em Engenharia da Computação

Modelo de Clusterização Fuzzy Não-Supervisionado para Dados Relacionais Baseado em Entropia

Proposta de Trabalho de Graduação

Aluno: Eduardo Cintra Simões (ecs4@cin.ufpe.br)
Orientador: Francisco de Assis Tenório de Carvalho (fatc@cin.ufpe.br)
Área: Aprendizagem não-supervisionada

17 de Abril de 2017

Resumo

Aprendizagem de máquina é a área da computação que, ao invés de implementar as soluções para determinado problema, busca métodos para o computador descobrir tais soluções, podendo ser supervisionada, caso o sistema seja treinado utilizando as respostas do conjunto de treinamento, ou não-supervisionada, caso contrário. Uma das suas subáreas é a clusterização, cujo objetivo é a divisão de um conjunto de elementos similares em classes. essa clusterização pode ser Hard, caso cada elemento pertença a apenas uma classe, ou Fuzzy, caso os elementos possam pertencer a mais de uma classe. O foco deste trabalho será a implementação de um modelo de clusterização fuzzy não-supervisionado para dados relacionais baseado em entropia.

Introdução

Aprendizagem de máquina é a área da computação que, ao invés de implementar as soluções para determinado problema, busca métodos para o computador descobrir tais soluções. Essa aprendizagem pode ser dividida em duas categorias: Aprendizagem Supervisionada, que utiliza as respostas do conjunto de treinamento, ou algum conhecimento sobre elas, durante o treino; E a aprendizagem não-supervisionada, que não utiliza as respostas no treino, o que se demonstra mais difícil e lento do que a aprendizagem supervisionada, mas pode ser utilizada numa gama maior de problemas.

Uma classe de problemas dentro da aprendizagem de máquina é a da clusterização, na qual se busca uma divisão dos elementos em grupos de similaridade. Esses problemas podem ser solucionados com classificadores hard, que colocam cada elemento em apenas uma classe, ou com classificadores fuzzy, que dizer que os elementos pertencem a mais de um grupo, com uma certa probabilidade. Um classificador é considerado bom quando ele consegue separar elementos distintos enquanto junta elementos semelhantes, reduzindo a entropia, ou caos, do sistema, normalmente representada por uma função-objetivo, também chamada de energia do sistema.

Este trabalho de graduação é baseado no trabalho de Jian-Ping Mei e Lihui Chen [1] e apresentará um método de clusterização fuzzy não-supervisionada para dados relacionais, descritos como uma matriz de dissimilaridade entre os elementos. Diferentemente do trabalho descrito em [1], a função-objetivo é dada pela Fórmula 1, sendo Tu e Tv constantes dadas na entrada, N o número de elementos, K o número de classes, u_{ic} a pertinência do elemento i na classe c, respeitando a fórmula 2, v_{cj} a influência do elemento j no protótipo implícito da classe c e D a matriz de dissimilaridade entre os elementos.

$$J = \sum_{c=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ic}v_{cj}D(i,j) + Tu\sum_{c=1}^{K}\sum_{i=1}^{N} u_{ic}ln(u_{ic}) + Tv\sum_{c=1}^{K}\sum_{j=1}^{N} v_{cj}ln(v_{cj})$$

Fórmula 1 - Função-objetivo do sistema

$$\sum_{c=1}^{K} u_{ic} = 1, \ \forall i = 1, 2, ...N$$

Fórmula 2 - condição da pertinência fuzzy

Objetivos

- Desenvolver um método de clusterização Fuzzy não-supervisionada que busca reduzir a função-objetivo da clusterização descrita pela fórmula 1.
- Encontrar um método de escolher boas constantes Tu e Tv.
- Comparar com outros modelos da clusterização fuzzy não-supervisionados.
- Gerar um relatório para ser apresentado no final do período.

Metodologia

O modelo possui uma inicialização e uma fase iterativa que computa as partições fuzzy e os protótipos de forma alternada. Na inicialização os elementos são particionados de forma aleatória, respeitando a Fórmula 2, a matriz V dos protótipos é então calculada utilizando a Fórmula 3, finalmente a energia inicial é calculada.

$$v_{cj} = \frac{exp\left(-\frac{\sum_{i=1}^{N} u_{ic} D(i,j)}{Tv}\right)}{\sum_{h=1}^{N} exp\left(-\frac{\sum_{i=1}^{N} u_{ic} D(i,j)}{Tv}\right)}$$

Fórmula 3 - Otimização dos protótipos

Na fase iterativa, as pertinências dos elementos são recalculados utilizando a Fórmula 4, então é a vez dos protótipos reusando a fórmula 3 para, finalmente, re-calcular a energia e verificar a condição de parada (dada pelo número de iterações ou se a variação da energia não superou um certo limiar dado como entrada).

$$u_{ic} = \frac{exp\left(-\frac{\sum_{i=1}^{N} v_{cj} D(i,j)}{Tu}\right)}{\sum_{h=1}^{K} exp\left(-\frac{\sum_{i=1}^{N} v_{cj} D(i,j)}{Tu}\right)}$$

Fórmula 4 - Otimização das pertinências

As entradas do modelo são a matriz de dissimilaridade entre os elementos, o número de clusters que serão gerados, as constantes Tu e Tv, que influenciam no passo de partição e de prototipação, respectivamente, e a variação mínima de energia e o número máximo de iterações, que fazem parte do critério de parada.

Se as constantes Tu e Tv forem muito pequenas, o classificador acaba ficando com um grande distância entre os protótipos e se torna um classificador Hard, contudo, se forem muito grandes, essa distância fica muito muito pequena e os elementos acabam sendo distribuídos igualitariamente entre as classes. Uma forma de se evitar isso é escolher Tu e Tv de forma à distância entre os protótipos seja 0.1 [2]. Para tal, será utilizada uma busca binária para se obter uma distância entre 0.09 e 0.12,

Cronograma

	Período															
Atividade	Março			Abril				Maio				Junho				Julho	
Revisão bibliográfica	X	X	X	X	X											
Estudo do algoritmo			X	X	X	X										
Implementação				X	X	X	X									
Experimentos							X	X	X							
Avaliação dos resultados									X	X	X	X				
Escrita do TG										X	X	X	X	X		
Preparação da apresentação													X	X		

Referências

[1] Jian-Ping Mei, Lihui Chen “Fuzzy relational clustering around medoids: A unified view” Fuzzy Sets and Systems; June 2011, Volume 18, Pages 44–56

[2] Veit Schwämmle and Ole Nørregaard Jensen “A simple and fast method to determine the parameters for fuzzy c–means cluster analysis” Bioinformatics; November 2010, Volume 26 Issue 22

Possíveis Avaliadores

Prof. Sergio Ricardo de Melo Queiroz

Prof. Ricardo Prudêncio

Assinaturas

Recife, 17 de Abril de 2017

_____

Eduardo Cintra Simões

(Aluno)

_____

Francisco de Assis Tenório de Carvalho

(Orientador)