XXIII CONIC
VII CONITI
IV ENIC

ALGORITMOS DE AGRUPAMENTO PARTICIONAIS SEMI-SUPERVISIONADOS NÃO-EXCLUSIVOS PARA DADOS QUANTITATIVOS

Roberto Costa Fernandes[1]; Francisco de Assis de Tenório de Carvalho[2]

[1]Estudante do curso de Engenharia da Computação – CIn – UFPE; E-mail: rcf6@cin.ufpe.br,
[2]Docente/pesquisador do Depto de Informática - CIn – UFPE; E-mail: fatc@cin.ufpe.br.

Sumário: Com a existência de muitas informações distintas e não-homogêneas, muitos dados ficam organizados de maneira difusa. Esse trabalho visa comparar dois algoritmos de agrupamentos, um que agrupa somete objetos e outro que agrupa simultaneamente objetos e variáveis. O agrupamento simultâneo também é conhecido como agrupamento em blocos, que tem como objetivo reorganizar a base de dados inicial em blocos homogêneos. Assim esses blocos podem ser definidos como subconjuntos da base de dados inicial. O repositório de dados da Universidade da Califórnia, em Irvine (UCI) é um dos repositórios com maior quantidade de dados e um dos mais utilizados, entre essas bases de dados, a mais conhecida é a Iris, que será utilizada nessa comparação de algoritmos.

Palavras-chave: clustering analysis; double k-means; euclidean distances

INTRODUÇÃO

Análise de agrupamento é o estudo de métodos e algoritmos de agrupamento, cujo objetivo é organizar objetos em grupos de acordo com suas características para formar grupos, em que objetos do mesmo grupo têm um alto grau de similaridade e objetos de grupos distintos apresentam um alto grau de dissimilaridade. Agrupamento não utiliza rótulos com identificação à priori, ou seja, rótulos de classe e por isso o agrupamento (aprendizagem não supervisionada) se difere da classificação ou análise discriminante (aprendizagem supervisionada).

A tarefa de agrupar geralmente agrupa somente os indivíduos ou somente as variáveis, mas um método que vem ganhando importância é o de agrupar simultaneamente indivíduos e variáveis. Esse método resulta em blocos de informação homogênea que sintetiza melhor as informações. Se agruparmos separadamente os indivíduos e as variáveis para depois juntar as informações iremos ter um maior custo computacional, quando comparado com o agrupamento simultâneo.

Os métodos de agrupamento de dados podem ser divididos em exclusivos e não-exclusivos: no método exclusivo o dado pertence somente a um grupo, já no método não-exclusivo o dado tem um grau de pertinência para cada grupo.

Esse trabalho tem como objetivo comparar a implementação e resultados dos algoritmos Fuzzy c-means (FCM), um método não exclusivo para agrupar somente os objetos, e o double c-means não-exclusivo, que agrupa simultaneamente objetos e variáveis.

MATERIAS E MÉTODOS

O FCM é um algoritmo que tem como entrada a descrição dos objetos que serão analisados, a quantidade máxima de iterações e a quantidade de grupos de dados desejados. Esse método utiliza apenas duas etapas em que se visa minimizar a função objetivo. Essa função utiliza o grau de pertinência de cada objeto em cada grupo, comparando a descrição de um objeto e a descrição do protótipo de cada grupo.

O algoritmo é inicializado selecionando aleatoriamente, entre os objetos, protótipos para cada grupo. Além disso é calculado o grau de pertinência de cada objeto em cada grupo, esse grau é calculado utilizando uma comparação entre a descrição de um objeto e a descrição do protótipo de cada grupo. Ao final desse cálculo esses dados obtidos são utilizados para a computação da função objetivo.

A etapa número um consiste no cálculo de novos protótipos para os grupos. Esse passo utiliza o grau de pertinência dos objetos e suas descrições para calcular novos protótipos. Os novos protótipos são calculados a partir da comparação entre a descrição de um objeto e a descrição do protótipo de um grupo. A segunda etapa é calcular os novos graus de pertinência de cada objeto. No final dessa etapa é calculado um novo valor para a função objetivo.

Essas duas etapas irão ficar se alternando até que a minimização da função objetivo não seja tão expressiva, ou seja, a diferença do valor da função objetivo atual com o valor da iteração anterior é menor que um número muito pequeno, como por exemplo 1e-10. Se essa diferença não for atingida essas duas etapas só serão executados pelo número máximo de iterações, que foi definido no início do programa.

Esse algoritmo é executado um número fixo de vezes para que a repetição com o menor valor da função objetivo seja utilizado como melhor resultado. No final da execução se terá o grau de pertinência de cada objeto em cada grupo e também se terá a descrição dos protótipos de cada grupo. Os grupos serão formados com os dados que têm maior grau de pertinência naquele grupo.

O double c-means não-exclusivo é um algoritmo em que três etapas ficam se alternando com o objetivo de tentar minimizar sua função objetivo, essa função objetivo se difere da função do FCM, pois também é utilizado o grau de pertinência de cada variável em cada grupo de variáveis. Esse método recebe com entrada a descrição dos objetos que serão analisados, a quantidade máxima de iterações, a quantidade de grupos de dados desejados e a quantidade de grupos de variáveis.

Esse algoritmo é inicializado escolhendo aleatoriamente o grau de pertinência de cada objeto em cada grupo de objeto e também escolhendo aleatoriamente o grau de pertinência de cada variável em um grupo de variável. Apesar dessas escolhas serem feitas aleatoriamente, a soma dos graus de pertinência de cada dado em todos os grupos tem que ser um, igualmente para as variáveis.

A primeira e a segunda etapa desse algoritmo são equivalentes a primeira e a segunda do FCM, a diferença é que na primeira etapa do double c-means é calculado o protótipo para cada bloco, a segunda continua sendo a do cálculo do grau de pertinência de cada indivíduo. A terceira etapa é o cálculo do grau de pertinência de cada variável.

Ao final dessas três etapas é calculado o novo valor da função objetivo, pois qualquer alteração nos protótipos e nos graus de pertinência gerará uma modificação no valor da função objetivo. Esse valor será comparado com o valor anterior da função objetivo e se a diferença entre esses valores for muito pequena ou se o número máximo de iterações for atingido o algoritmo acaba, caso contrário ele volta para a primeira etapa.

Após uma quantidade fixa de repetições o algoritmo utiliza a repetição com o menor valor da função objetivo como o melhor resultado. Ao final de todas as repetições irá se obter como saída os graus de pertinência dos dados e das variáveis, o protótipo de cada bloco e consequentemente também se terá os blocos com os dados homogêneos.

RESULTADOS

Para a execução dos algoritmos fuzzy c-means e double c-means não-exclusivo foi utilizado a base de dados Iris. Essa base de dados é uma das mais populares em testes de

algoritmos de agrupamento e ela está disponível no repositório de dados da Universidade da California, em Irvine (UCI - https://archive.ics.uci.edu/ml/).

Essa base de dados descreve flores do gênero Iris e contém 150 indivíduos e cada um com 5 informações, que são referentes à largura e comprimento em centímetro das pétalas e largura e comprimento em centímetro das sépalas. A última informação é a classe de cada flor só serve para comparar se as flores da mesma espécie ficaram no mesmo grupo.

Para a execução desses algoritmos foram utilizados os seguintes parâmetros:

- Quantidade de grupos de indivíduos e de variáveis: 3;
- Quantidade máxima de iterações e de repetições: 100;
- Parâmetro de erro: 1e-12;

Além desses parâmetros também foram utilizados parâmetros de correção que são definidos tetando para achar as melhores partições, esses parâmetros são m, alfa e beta, o primeiro é utilizado do FCM e os outros dois no double c-means. Nesses experimetros foram utilizados o valor de 1,1 para esses parametros. Na melhor repetição o FCM obteve o valor 60.5759555 para a função objetivo, enquanto o double c-means obteve 152.8339987 e a matriz de protótipos para os algoritmos ficou da seguinte forma o FCM:

	Variável_1	Variável_2	Variável_3	Variável_4
Grupo_1	6.775119	3.052431	5.646914	2.053608
Grupo_2	5.889200	2.761235	4.364255	1.397447
Grupo_3	5.003561	3.403036	1.485002	0.251541

E da seguinte forma no double c-means não-exclusivo:

	Grupo de Variáveis_1	Grupo de Variáveis_2	Grupo de Variáveis_3
Grupo de Indivíduos_1	4.926861	0.610838	2.881154
Grupo de Indivíduos_2	6.075544	1.691657	3.437630
Grupo de Indivíduos_3	6.075526	1.691642	3.437620

O grau de pertinência de alguns indivíduos ficou da seguinte forma para o FCM:

	Grupo_1	Grupo_2	Grupo_3
Indivíduo_1	0.001163	0.002501	0.996336
Indivíduo_2	0.007156	0.015887	0.976957
Indivíduo_3	0.006261	0.013427	0.980312
...	...	...	...
Indivíduo_148	0.831250	0.156620	0.012130
Indivíduo_149	0.789262	0.189147	0.021591
Indivíduo_150	0.391020	0.582042	0.026938

E da seguinte forma para o double c-means:

	Grupo_1	Grupo_2	Grupo_3
Indivíduo_1	0.567009	0.216494	0.216497
Indivíduo_2	0.629814	0.185091	0.185095
Indivíduo_3	0.618419	0.190789	0.190792
...	...	...	...
Indivíduo_148	0.133617	0.433194	0.433188

Indivíduo_149	0.133804	0.433101	0.433095
Indivíduo_150	0.155759	0.422121	0.422120

Os graus de pertinência das variáveis no double c-means ficaram:

	Grupo_1	Grupo_2	Grupo_3
Variável_1	0. 918530	0. 020743	0. 060728
Variável _2	0. 039512	0. 077347	0. 883141
Variável _3	0. 252676	0. 196797	0. 550527
Variável _4	0. 016383	0. 912058	0. 071559

A partir da matriz com os graus de pertinência é possível obter uma partição exclusiva. Para isso olhamos os graus de pertinência de cada indivíduo e dizemos que ele participará do grupo que ele tem o maior grau de pertinência. Com essa nova tabela conseguirmos comparar as classes descritas na base de dados e o resultado obtido com o algoritmo, para assim obtermos o índice chamado de rand corrigido, que informa a compatibilidade entre as classes a priori e o agrupamento fornecido pela execução do algoritmo. Esse índice é um valor que pode ser no mínimo 0 e no máximo 1, quando mais peto de 1 maior é a compatibilidade entre os agrupamentos. Na execução do FCM o índice de rand corrigido ficou 0.729420, e na execução do double c-means ficou 0.533061.

DISCUSSÃO

Com a execução dos dois algoritmos, para essa base de dados e para alguns parâmetros de correção, pode-se perceber que ao agrupar apenas indivíduos obtêm-se um maior grau de compatibilidade com a partição a priori, quando comparado com o agrupamento simultâneo. Isso não significa que para todas as bases de dados e para todos parâmetros iniciais esse resultado irá se repetir, pois pode acontecer do double c-means obter uma melhor compatibilidade.

O double c-means tem a vantagem de sempre agrupar indivíduos e variáveis. Além disso esse método cria blocos de informação homogênea. O grau de similaridade desse bloco pode ser alterado de acordo com os parâmetros iniciais.

CONCLUSÕES

Com a grande quantidade de algoritmos de agrupamento esse trabalho comparou a execução de dois algoritmos na base de dados Íris do UCI. O primeiro algoritmo visa agrupar somente os indivíduos em partições não-exclusivas. O segundo tem como objetivo agrupar simultaneamente indivíduos e variáveis, em partições não exclusivas, para criar blocos de informação.

Dentro de alguns parâmetros o FCM obteve um melhor resultado na compatibilidade com a partição inicial. Que apesar de ser um dos algoritmos de agrupamento mais antigo ainda mostrar que em alguns casos pode ser melhor do que algoritmos que foram criados ao longo da existência dele.

AGRADECIMENTOS

Agradecemos ao PIBIC, à UFPE, e ao CNPq pelo apoio para a realização da pesquisa.

REFERÊNCIAS

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J.. Data Clustering: A Review. Acm Computing Surveys, Irati, 1999.

JAIN, Anil K. 2009. Data clustering: 50 years beyond K-means.