Ministério da
Ciência e Tecnologia

INPE-16636-RPQ/841

CLUSTERIZAÇÃO DE DADOS BIOLÓGICOS ATRAVÉS DA METAHEURÍSTICA VNS

Dalila Ribeiro Serpa

Relatório final da disciplina Princípios e Aplicações de Mineração de Dados (CAP-359) do Programa de Pós-Graduação em Computação Aplicada, ministrada pelo professor Rafael Santos.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m19@80/2009/10.27.18.08>

INPE
São José dos Campos
2009

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6911/6923
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

AGRADECIMENTOS

Agradeço a colaboração de Antônio Augusto Chaves, Unesp de Guaratinguetá e Mariá C.V. Nascimento, USP de São Carlos, que me auxiliaram no desenvolvimento e teste do algoritmo.

RESUMO

O estudo de diferentes técnicas para clusterização de dados biológicos vem crescendo ultimamente. Muitos algoritmos já são utilizados para este fim, como por exemplo, K-Medias, K-Medianas, entre outros. Mas, o uso de metaheurísticas vem sendo estudado recentemente e já mostra resultados satisfatórios. Este trabalho apresenta uma abordagem com a metaheurística VNS para solucionar este problema de classificação de dados. Os resultados obtidos pelo GRASP[1] para o método euclidiano de cálculo de distâncias, serão utilizados com o intuito de comparar e verificar se há melhora nas soluções com o uso do VNS. Os dados usados nos testes são referentes a três bases de dados biológicos que trouxeram sua classificação original, facilitando assim a avaliação das soluções através do Corrected Rand Index. Ao final do trabalho, encontra-se uma comparação entre os resultados do VNS e do GRASP[1].

SUMÁRIO

Pág.

LISTA DE TABELAS
1 INTRODUÇÃO 1
2 DADOS BIOLÓGICOS 1
3 MODELO MATEMÁTICO 2
4 METAHEURÍSTICA VNS 4
5 CORRECTED RAND INDEX 7
6 EXPERIMENTO COMPUTACIONAL 7
7 CONCLUSÃO 8
REFERÊNCIAS BIBLIOGRÁFICAS 9

LISTA DE TABELAS

Pág.

6.1 – Resultados do VNS.....8
6.2 – Comparação dos resultados VNS x GRASP 8

1 INTRODUÇÃO

O estudo de técnicas para clusterização de dados biológicos vem crescendo ultimamente[1]. As proteínas, células cancerígenas e até mesmo dados sobre plantas são bastante utilizados nas buscas por melhores algoritmos para classificação de grande volume de dados, tarefa nada fácil para seres humanos.

Muitos algoritmos já conhecidos, são utilizados para este fim, como por exemplo K-Medias, K-Medianas, entre outros. Mas, o uso de metaheurísticas vem sendo estudado recentemente e já mostra resultados satisfatórios.
Inspirado na utilização do GRASP em [1], este trabalho, apresenta uma abordagem com a metaheurística VNS para solucionar este problema de classificação de dados.
Como os resultados obtidos em [1] foram satisfatórios, o VNS é aplicado com o intuito de comparar e, possivelmente, melhorar as soluções geradas pelo GRASP.

Os dados usados nos testes são referentes a três bases de dados biológicos que trouxeram sua classificação original, facilitando assim a avaliação das soluções através do Corrected Rand Index (índice que mede o quanto dois clusters são parecidos).
No decorrer do trabalho, encontram-se breves explicações sobre o VNS, a formulação matemática utilizada, o Corrected Rand Index e por fim a comparação dos resultados obtidos pelos dois algoritmos.

2 DADOS BIOLÓGICOS

Serão utilizadas três bases de dados biológicos distintas: Breast, Proteínas e Íris. Foram escolhidas porque apresentaram os melhores resultados em [1] se comparados aos outros métodos(K-Medias, etc).

A primeira trata-se de um conjunto de células cancerígenas. Possui 699 objetos, nove atributos e duas classes. A classificação desses objetos é dada pela caracterização de câncer maligno ou benigno que cada célula apresenta.

A segunda, Proteínas, tem 698 objetos com 125 atributos cada. Sua classificação se dá pelo tipo de enrolamento de cada proteína. Neste caso, serão quatro tipos (4 classes).

Já a terceira, Íris, é formada por 150 flores com 4 atributos cada uma. Sua classificação se dá pelo tamanho das pétalas e sépalas das flores. Dividi-se em 3 classes: Íris Versicolor, Íris Setosa e Íris Virgínica.

As três bases utilizadas são formadas apenas por dados numéricos, portanto, não foi necessário pré-processamento dos dados.

A base Íris e a Breast são encontradas em UCI Repository [2]. Já a base Proteínas é encontrada em http://ranger.uta.edu/~chqding/protein.

3 MODELO MATEMÁTICO

A partir da idéia de clusterização, pode-se obter um modelo matemático que resolva esse problema. Para este caso, a programação linear inteira mista é aplicável, conforme mostrado em [1].

Como essas bases são formadas por um número muito grande de objetos, não é possível obter uma solução ótima em tempo computacional aceitável. Neste caso, usam-se técnicas que se aproximam da solução ótima, como as metaheurísticas, por exemplo. A partir da formulação matemática, é possível aplicar um desses algoritmos aproximativos. O problema passa a ser de otimização combinatória.

A modelagem matemática encontrada em [1] será a mesma utilizada neste trabalho. Ela se resume em:

$$\textit{Min} \quad \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} y_{ij}$$

$$\textit{Sujeito a} \quad \sum_{k=1}^{M} x_{ik} = 1, \qquad i = 1,..,N \qquad (1)$$

$$\sum_{i=1}^{N} x_{ik} \geq 1, \qquad k = 1,..,M \qquad (2)$$

$$y_{ik} \geq x_{ik} + x_{jk} - 1, \qquad i = 1,..,N,\ j = i+1,..,N,\ k = 1,..,M \qquad (3)$$

$$y_{ij} \geq 1, i = 1,..,N,\ j = i+1,..,N \qquad (4)$$

$$x_{ik} \in \{0,1\}, i = 1,..,N, k = 1,..M \qquad (5)$$

onde, a restrição (1) garante que o objeto esteja alocado a apenas um cluster e (2) garante que um cluster tenha no mínimo um objeto. As retrições (3) e (4) garantem que y_{ij} assuma o valor 1 se ambos os x_{ik} e x_{jk} forem 1. E por último, a restrição (5) garante que as variáveis x_{ik} sejam binárias.

A função objetivo trata-se da minimização da soma das distâncias entre objetos de mesmo cluster.

Para o cálculo das distâncias, o método utilizado foi euclidiano. A formula é a seguinte:

$$d_{ij} = \sqrt{\sum_{k=1}^{L} (a_{ik} - a_{jk})^2}$$

onde, d_{ij} será a distância calculada com os atributos a das bases.

Com essas formulações, o VNS foi modelado e testado. No próximo capítulo, segue a descrição da modelagem do VNS.

4 METAHEURÍSTICA VNS

O algoritmo VNS[3] é uma metaheurística que, diferentemente das outras, não foca apenas na busca local, mas sim na busca feita em vizinhos distantes. Com essa busca, estará sempre explorando soluções diferentes. Mas com isso, pode perder as melhores soluções e é por esse motivo que deve ser implementado utilizando uma busca local.

A modelagem do algoritmo é baseada na formulação matemática apresentada anteriormente. As primeiras rotinas da aplicação são as de leitura dos dados e cálculo da distância euclidiana, tendo as distâncias armazenadas numa matriz.

A estrutura que armazena as soluções a seguinte:

```
Estrutura Solucao
    inteiro classes[Numero total de Objetos];
    double fo; (função objetivo)
```

A seguir, tem-se a rotina de geração da solução inicial, conforme o pseudocódigo abaixo:

```
GeraSolucaoInicial()
Variáveis
iteracoes, maxiteracoes = 500: integer;
solucaoNova : estrutura solucao;
Inicio
        Inicializa solucaoFinal.fo;

Enquanto iteracoes < maxIteracoes faça
Inicio
                Para i de 0 até o total de objetos faça
                        solucaoNova.classe = Numero inteiro randômico entre 1 e o total de classes;

                solucaoNova.fo = Calcula a função objetivo para a nova solução;
```

```
            Se solucaoNova.fo < solucaoFinal.fo então
        solucaoFinal = solucaoNova;
        maxIteracoes = maxIteracoes – 1;
        Fim-enquanto;
Fim.
```

Dessa forma, são geradas várias soluções aleatórias, mas apenas a que tem menor função objetivo é armazenada.

Depois de encontrada a solução inicial, inicia-se o VNS. Logo no início, o algoritmo faz trocas entre clusters, ou seja, aleatoriamente escolhe duas posições do vetor de classes que sejam de clusters diferentes e faz a troca entre elas. Em seguida, inicia-se a busca local.

Abaixo, tem-se a rotina VNS da aplicação:

```
VNS()
Variáveis
iteracoes, k, iter: inteiro;
sLinha, sStar : Estrutura Solucao;
Inicio
k = 1;
Enquanto iteracoes < (maximo_iteracoes) faça
Inicio
sLinha = solucaoFinal;
Caso k
1: Troca 1 posicao de cluster;
2: Muda 1 posicao de cluster;
3: Troca 2 posicoes de cluster;
4: Muda 2 posicoes de cluster;
5: Troca 3 posicoes de cluster;
6: Muda 3 posicoes de cluster;

sLinha = Calculo a Função Objetivo para sLinha;
sStar = sLinha;
BuscaLocal(sLinha, sStar);
      Se sLinha.fo < solucaoFinal.fo então
         solucaoFinal = sLinha;
```

```
        k = 1;
      Senão
        k += 1;

Se k > 6
k = 1;
Fim-enquanto;
Fim.
```

A rotina de busca local fará a troca de clusters a partir de uma posição aleatória, gerando uma nova classe aleatória para essa posição. Assim as soluções são melhoradas com muita rapidez. Segue abaixo o pseudocódigo:

```
BuscaLocal(Estrutura solucao: sLinha, sStar)
Variaveis
posTroca: inteiro;
        iteracoes: inteiro;
Inicio
Iteracoes = 50% do numero de objetos;
Enquanto iteracoes > 0 faça
Inicio
                posTroca = Numero inteiro aleatório entre 0 e o numero de objetos - 1;
sLinha.classe[posTroca] = Numero inteiro aleatório entre 1 e o numero de classes;
                Se sLinha.fo < sStar.fo então
                sStar = sLinha;
                Senão
                sLinha = sStar;
Iteracoes = iteracoes – 1;
Fim-enquanto;
Fim.
```

A quantidade de iterações do VNS é relativa à quantidade de objetos. Assim, quanto mais objetos na base de dados, maior o número de iterações.

Depois de encontrar as menores funções objetivo, as soluções foram avaliadas. No próximo capítulo, encontra-se uma descrição sobre essa avaliação.

5 CORRECTED RAND INDEX

O Corrected (Adjusted) Rand Index[4] é um índice que mede o quanto dois clusters são semelhantes. Com ele, as soluções geradas pelo VNS foram avaliadas.

Por ser um cálculo bastante complexo, o software R-Project[5] foi utilizado. É um software livre e de código aberto. Possui uma função que oferece esse cálculo. Basta entrar com os dois clusters e ele retorna o valor do índice. Quanto mais próximo de 1 for o índice, mais semelhantes são os clusters. Se o índice for 1, os clusters são iguais. Esse valor varia entre -1 e 1.

Neste trabalho, as bases utilizadas trouxeram suas classificações reais, assim o cálculo foi realizado entre a classificação real e a gerada pelo VNS.

6 EXPERIMENTO COMPUTACIONAL

Como visto anteriormente, o objetivo do trabalho é comparar os resultados obtidos pelo VNS com os resultados obtidos pelo GRASP em [1], utilizando o método euclidiano de cálculo de distâncias. Para isso, os testes foram feitos utilizando a mesma quantidade de classes que apresentaram os melhores resultados em [1]. Assim, para as bases Breast, Proteínas e Íris, foram usadas duas, quatro e três classes, respectivamente.

Para cada base, o algoritmo foi executado dez vezes. Dessas dez execuções, apenas a solução que apresentou o menor valor na função objetivo é que foi avaliada pelo Corrected Rand Index, chamado de CRandIndex.

Na Tabela 6.1, são apresentados os índices obtidos com as melhores soluções do VNS para cada base:

Tabela 6.1 – Resultados do VNS

Base	Classes	CRandIndex
Breast	2	0.877
Proteínas	4	0.322
Íris	3	0.756

Na Tabela 2, são apresentados os índices encontrados pelo VNS e pelo GRASP[1]:

Tabela 6.2 – Comparação dos resultados VNS x GRASP

Base	Classes	CRandIndex VNS	CRandIndex GRASP
Breast	2	0.877	0.877
Proteínas	4	0.322	0.322
Íris	3	0.756	0.756

Conforme a tabela 6.2, os resultados do VNS foram iguais aos do GRASP[1], por isso não foi necessário fazer testes no CPLEX, visto que em [1], os testes já haviam sido feitos e comprovaram o funcionamento do algoritmo.

7 CONCLUSÃO

Este trabalho apresentou uma nova técnica para clusterização de dados biológicos. Baseado numa formulação matemática, o VNS, gerou possíveis soluções que foram avaliadas através do Corrected Rand Index.

Ao comparar os resultados do VNS com os do GRASP[1], percebeu-se que eles foram iguais. Ou seja, os dois algoritmos são válidos para resolução deste problema quando usam o método euclidiano de cálculo de distâncias.

Essa técnica pode ser aplicada a outros tipos de dados, desde que sejam numéricos, pois há a necessidade de calcular a distancia entre os atributos.

Por ter apresentado resultados melhores que algoritmos bastante utilizados para esse fim, como K-Medias, por exemplo, pode ser tratado como uma nova técnica de mineração de dados.

Alterações no VNS ou aplicação de outras técnicas ficam como desafios futuros com o intuito de melhorar os resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] – Nascimento MCV, et al. Investigation of a new GRASP-based clustering algorithm applied to biological data. Computers and Operations Research (2009), doi: 10.1016/j.cor.2009.02.014

[2] – Site: http://archive.ics.uci.edu/ml/ Acesso em: 26/10/2009.

[3] – Hansen P., Mladenovic N. Variable Neighborhood search: Principles and applications. European Journal of Operational Research 130(2001) 449-467

[4] – Hubert L., Arabie P. Comparing partitions. Journal of Classification 1985;2;193-218.

[5] – Site: http://www.r-project.org/ Acesso em: 26/10/2009