Eduardo Cintra Simões

Clusterização Difusa com Múltiplos Representantes Baseada em Entropia para Dados Relacionais

Recife-PE

20 de Junho de 2017

Eduardo Cintra Simões

Clusterização Difusa com Múltiplos Representantes Baseada em Entropia para Dados Relacionais

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Universidade Federal de Pernambuco

Centro de Informática

Orientador: Francisco de Assis Tenório de Carvalho

Recife-PE

20 de Junho de 2017

Agradecimentos

Agradeço ao meu orientador, o professor Dr. Francisco de Assis Tenório de Carvalho, pela ajuda e pelas oportunidades que me proporcionou, tando durante a Iniciação Científica quanto agora.

Agradeço ao Dr. Diogo Philippini Pontual Branco pela implementação dos Algoritmos Nerf e Fuzzy.

Agradeço ao centro de informática da UFPE, e as pessoas que fazem parte do mesmo, pelos anos de aprendizagem e pela oportunidade.

Agradeço à minha família e amigos pelo apoio.

Resumo

Classificadores podem ser utilizados em diversas áreas do conhecimento. Para tal é necessário treiná-los com dados do domínio em que precisam atuar, de preferência com os resultados esperados. Contudo, nem sempre essas classes são conhecidas, por tanto é necessário métodos de classificação não-supervisionados (que não utilizam as classes no treinamento). Os classificadores também podem ser difusos, podendo classificar os elementos em mais de uma classe, com certos graus de pertinência. Classificadores Difusos são bons para encontrar a relação entre classes diferentes, além de melhor representarem elementos próximos do limiar entre duas classes.

Esse trabalho tem como objetivo a apresentação de um modelo de classificação difusa não-supervisionado inspirado no modelo FMMdd (MEI; CHEN, 2011) com resultados compatíveis aos modelos da literatura Nerf (HATHAWAY; BEZDEK, 1994) e Fanny (KAUFMAN; ROUSSEEUW, 1990a), contudo sendo consideravelmente mais rápido do que esses modelos.

Palavras-chaves: classificação, clusterização, difuso, dados relacionais, não-supervisionado

Lista de tabelas

Tabela 1 – Bases testadas 19
Tabela 2 – Parâmetros do algoritmo proposto 20
Tabela 3 – Resultados de cada algoritmo para as bases testadas 21
Tabela 4 – Porcentagem em que cada algoritmo obteve o maior ou menor valor de cada índice 26
Tabela 5 – Melhores Resultados de cada algoritmo para as bases testadas 27
Tabela 6 – Porcentagem em que cada algoritmo obteve o maior ou menor valor de cada índice para a melhor solução 29

Lista de abreviaturas e siglas

J	Energia do sistema (Entropia, Função-Objetivo)
N	Número de elementos
K	Número de Grupos
x_i	i-ésimo elemento
D	Matriz de Dissimilaridade
D(i,j)	Dissimilaridade entre x_i e x_j
U	Matriz de pertinências
u_{ik}	Pertinência do i-ésimo elemento no k-ésimo grupo
V	Matriz de pesos de representação
v_{kj}	Peso do j-ésimo elemento no k-ésimo grupo
T_u	Parâmetro que controla o cálculo das pertinências
T_v	Parâmetro que controla o cálculo dos pesos
t	passo da iteração
T_{MAX}	Limite de iterações
NERF	non-Euclidean relational fuzzy (algoritmo da literatura)
FANNY	Fuzzy Analysis (algoritmo da literatura)
FMMdd	Fuzzy clustering with multi-medoids (algoritmo de referência)
FMMdd-ENT	Fuzzy clustering with multi-medoids based on Entropy (algoritmo proposto)

Lista de símbolos

ϵ	letra grega épsilon, limiar mínimo da variação de energia entre iterações
ϕ	letra grega fi.
$\sum$	somatório
$\forall$	para todo
$\in$	pertence
$\vee$	ou lógico
$\leq$	menor ou igual

Sumário

Introdução **9**
1.1 Objetivo 9

Fundamentos **10**
2.1 Dados Relacionais 10
2.2 Entropia 10

Metodologia **11**
3.1 Entradas 11
3.2 Saídas 11
3.3 Algoritmo 12
3.3.1 Otimização do calculo das pertinências 12
3.3.2 Otimização do calculo dos pesos 13
3.4 Escolhendo os parâmetros 13

Experimentos **15**
4.1 Índices 15
4.1.1 Índices para partições exclusivas 15
4.1.1.1 F-Measure 15
4.1.1.2 Erro Geral 16
4.1.1.3 Rand Ajustado 16
4.1.1.4 Informação Mútua Normalizada (NMI) 16
4.1.1.5 Coeficiente de Silhueta 16
4.1.2 Índices para partições difusas 17
4.1.2.1 Rand Frigui 17
4.1.2.2 Rand Hüllermeier 17
4.1.2.3 Coeficiente de Partição (Vpc) 18
4.1.2.4 Coeficiente de Partição Modificado (Vmpc) 18
4.1.2.5 Entropia da Partição (Vpe) 18
4.1.2.6 Coeficiente de Silhueta Difusa 18
4.2 Bases 19
4.3 Parâmetros 19

Resultados **21**

Conclusão **30**

Referências **31**

Introdução

Classificar elementos em grupos de acordo com suas similaridades é a base para diversas áreas do conhecimento (Biólogos precisam classificar espécies em famílias, bancos precisam identificar se uma pessoa é boa pagadora apenas pelo seu perfil, entre vários outros exemplos). Os classificadores são treinados com elementos de exemplo do domínio, podendo esse treinamento ser supervisionado (caso conheça e utilize as classes desses elementos no treinamento) ou não-supervisionado (caso contrário).

Os Classificadores também podem ser divididos entre exclusivos (caso só se possa mapear cada elementos para um grupo) e difusos (caso os elementos possam participar de mais de um grupo, com um certo grau de pertinência)

É importante ressaltar que não existe um método de classificação ideal para qualquer situação. Por causa disso, é necessário o estudo de métodos novos, para se aumentar a quantidade de opções de classificação.

1.1 Objetivo

O Objetivo desse Trabalho de Graduação é apresentar um método de classificação difusa não-supervisionada para dados relacionais com uma única matrizes de dissimilaridade. Também será apresentado um método para se encontrar bons parâmetros para o método apresentado.

Fundamentos

2.1 Dados Relacionais

Dados Relacionais são aqueles representados não pelos valores dos elementos de forma direta, mas pela relação entre os elementos (como a distância entre os elementos). Para esse trabalho será considerado os dados relacionais representados por uma matriz de dissimilaridade.

Entende-se por matriz de dissimilaridade como uma matriz D na qual a célula D(i,j) representa a distância entre o i-ésimo elemento com o j-ésimo elemento.

2.2 Entropia

Entropia é uma medida do caos (aleatoriedade) de uma variável aleatória. sendo mínima caso um dos valores tenha 100% de chance de ocorrer, e máxima caso os valores sejam equiprováveis. A entropia para uma variável X com valores possíveis $x_i, i = 1, 2...N$ é dada pela equação 2.1

$$E(X) = \sum_{i=1}^{N} p(x_i) log\left(\frac{1}{p(x_i)}\right) \tag{2.1}$$

Metodologia

O método proposto é um modelo difuso de classificação não-supervisionado que busca resolver o problema de otimização que minimiza, de forma iterativa, a entropia da divisão entre os elementos. Ele é baseado no método proposto por Jian-Ping Mei, Lihui Chen (MEI; CHEN, 2011) que busca minimizar a equação 3.2. Sendo u_{ik} a pertinência do elemento x_i no grupo k, v_{kj} a influência do elemento x_j na representação do grupo k e D(i,j) a distância entre x_i e x_j.

$$J_{FMMdd} = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N}(u_{ik})^m(v_{kj})^n D(i,j) \tag{3.2}$$

No método proposto foram adicionados, à função-objetivo, termos referentes à entropia das pertinências u_{ik} e a dos pesos das representações v_{kj}, ao invez de utilizar as exponenciais do primeiro termo, como pode-se ser observado na equação 3.3. As matrizes U e V são calculadas iterativamente com equações obtidas pelo método de multiplicadores de Lagrange com o intuito de minimizar a função-objetivo.

$$J_{FMMdd-ENT} = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N}(u_{ik})(v_{kj})D(i,j) + T_u\sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ik})ln(u_{ik}) + T_v\sum_{k=1}^{K}\sum_{j=1}^{N}(v_{kj})ln(v_{kj}) \tag{3.3}$$

3.1 Entradas

- A matriz de dissimilaridade: D
- O número de grupos: K
- O número máximo de iterações: T_{MAX}
- O limiar mínimo da variação de energia: ϵ
- O parâmetro que controla o cálculo das pertinências : T_u
- O parâmetro que controla o cálculo dos pesos: T_v

3.2 Saídas

- Matriz de pertinências: U
- Matriz de pesos de representação: V

3.3 Algoritmo

Algoritmo
inicialização
1. Carrega os parâmetros e a matriz de dissimilaridade.
2. t = 0
3. Seleciona as pertinências aleatoriamente, respeitando a condição 3.4
4. Calcula os pesos de representação iniciais, com a equação 3.5
5. Calcula a entropia com a equação 3.3
Iteração
6. t = t + 1
7. Atualiza as pertinências utilizando a equação 3.6
8. Atualiza os pesos de representação utilizando a equação 3.5
9. Atualiza a entropia com a equação 3.3
10. Se a condição de parada 3.10 não foi atingida, volta para 6.

$$\sum_{k=1}^{K} u_{ik} = 1, \forall i = 1, 2, \ldots, N \tag{3.4}$$

$$v_{kj} = \frac{exp\left\{\frac{-\sum_{i=1}^{N}(u_{ik})D(i,j))}{T_v}\right\}}{\sum_{h=1}^{K} exp\left\{\frac{-\sum_{i=1}^{N}(u_{ik})D(i,h))}{T_v}\right\}} \tag{3.5}$$

$$u_{ik} = \frac{exp\left\{\frac{-\sum_{j=1}^{N}(v_{kj})D(i,j))}{T_u}\right\}}{\sum_{h=1}^{K} exp\left\{\frac{-\sum_{j=1}^{N}(v_{hj})D(i,j))}{T_u}\right\}} \tag{3.6}$$

3.3.1 Otimização do calculo das pertinências

Por questão de otimização, pode-se utilizar um vetor auxiliar de tamanho K que armazena os resultados parciais definidos pela equação 3.7, para então calcular as pertinências usando as equações 3.8 e 3.9.

$$aux_i[h] = exp\left\{\frac{-\sum_{j=1}^{N}(v_{hj})D(i,j))}{T_u}\right\} \tag{3.7}$$

$$norm_i = \sum_{h=1}^{K} aux_i[h] \tag{3.8}$$

$$u_{ik} = \frac{aux_i[k]}{norm_i} \tag{3.9}$$

$$\left\| J^{(t)}_{FMMdd-ENT} - J^{(t-1)}_{FMMdd-ENT} \right\| \leq \epsilon \vee t = T_{MAX} \tag{3.10}$$

3.3.2 Otimização do calculo dos pesos

Por questão de otimização, pode-se utilizar um vetor auxiliar de tamanho N que armazena os resultados parciais definidos pela equação 3.11, para então calcular os pesos usando as equações 3.12 e 3.13.

$$aux_h[j] = exp\left\{ \frac{-\sum_{i=1}^{N}(u_{ih})D(i,j))}{T_v} \right\} \tag{3.11}$$

$$norm_h = \sum_{j=1}^{N} aux_h[j] \tag{3.12}$$

$$v_{kj} = \frac{aux_k[j]}{norm_k} \tag{3.13}$$

3.4 Escolhendo os parâmetros

Os resultados são sensíveis aos parâmetros T_u e T_v, além de que os melhores valores para esses parâmetros dependem da base. Por tanto é necessário um método para escolher esses parâmetros, e de preferência de forma automática.

Caso os centroides dos grupos de um classificador difuso fiquem muito próximos, as pertinências ficarão em torno de 1/K, contudo, caso esses centroides fiquem muito distantes, o classificador basicamente se torna um classificador exclusivo (com cada elemento com uma das pertinências próxima de 1, e as outras de 0). Foi descoberto empiricamente que classificadores difusos possuem bons resultados quando a distância mínima entre os centroides se encontra em torno de 0,1 (SCHWäMMLE; JENSEN, 2010). Também é importante perceber que, na maioria dos casos, aumentar os valores de T_u e T_v diminui a distância entre os centroides, isso se dá pela natureza exponencial do uso desses parâmetros.

Os valores iniciais dos parâmetros foram calculados utilizando as equações 3.14 e 3.15 (obtidas dos passos de atualização dados pelas equações 3.5 e 3.6) logo após a inicialização. Vale ressaltar que esses valores não são bons. O que estamos interessados aqui é obter uma boa proporção entre T_u e T_v, além de um ponto inicial para busca que tenha alguma relação com a base. Acreditamos que essa proporção é boa pois ela se mostrou interessante empiricamente.

$$T_u = max_{1 \leq h \leq K} \left\{ \sum_{j=1}^{N} (v_{hj}) D(i,j)) \right\} \tag{3.14}$$

$$T_v = max_{1 \leq h \leq N} \left\{ \sum_{i=1}^{N} (u_{ih}) D(i,j)) \right\} \tag{3.15}$$

Agora uma busca é feita, modificando-se o parâmetro T_u e mantendo o parâmetro T_v proporcional, dessa forma o problema de duas variáveis se torna um de apenas uma. Essa busca foi feita como uma busca binária levando em consideração a média da distância mínima entre os centroides de 10 repetições e considerando como bom resultado caso a distância entre os centroides se encontrasse entre 0,09 e 0,12.

Contudo, nem sempre é possível utilizar essa estratégia, já que não é em todo caso que a relação dos parâmetros com a distância entre os centroides é uma função monótona. Para esses casos em que a busca binária não consegue encontrar bons parâmetros, utilizou-se uma busca em grade variando os parâmetros numa grade de possibilidades próxima do melhor resultado encontrado pela busca binária.

Experimentos

O algoritmo foi testado, junto com os algoritmos Nerf (HATHAWAY; BEZDEK, 1994) e Fanny (KAUFMAN; ROUSSEEUW, 1990a), para diversas bases, e a média e desvio padrão de alguns índices foram calculados sobre o resultado de 100 repetições de cada um dos algoritmos. As implementações foram feitas em C, e os experimentos foram feitos num computador com processador Core i7-4790 3.60 GHz e 12 GB de memória RAM, rodando Windows 10 64bits.

As matrizes de dissimilaridade das bases foram calculadas com a distância euclidiana entre os elementos e então normalizada seguinda a equação 4.16, sendo D(i,j) a distância entre os elementos x_i e x_j e D'(i,j) a distância normalizada entre esses elementos.

$$D'(i,j) = \frac{D(i,j)}{\sum_{k=1}^{N} D(i,k)} \tag{4.16}$$

4.1 Índices

Índices são calculados para se verificar a qualidade da classificação e poder se comparar métodos diferentes. Os primeiros índices foram pensados para partições exclusivas, mas, com o surgimento de classificadores difusos, outros índices foram criados para verificar a qualidade das partições difusas.

4.1.1 Índices para partições exclusivas

Para classificadores difusos, a partição exclusiva pode ser obtida selecionando, para cada elemento, a partição difusa com maior pertinência.

4.1.1.1 F-Measure

Índice que leva em consideração a precisão e revocação. Porém, como o problema pode possuir mais de duas classes, é necessário utilizar uma variação dessa medida. Essa variação está representada pela equação 4.17, sendo K_c o grupo predominante da classe C. Ele varia entre 0 e 1, sendo 1 o melhor valor. (RIJISBERGEN, 1979)

$$F - measure = \sum_{c \in C} p(c) \frac{2p(c,K_c)}{p(c)p(K_c)} \tag{4.17}$$

4.1.1.2 Erro Geral

Proporção de elementos mapeados erroneamente. Como a quantidade de classes e de grupos não precisa ser a mesma, foi necessário utilizar uma variação do erro. (BREIMAN et al., 1984)

Será considerado que um elemento foi mapeado erroneamente se a sua classe não corresponde a classe predominante do grupo em que foi mapeado. Os valores variam entre 0 e 1, sendo 0 o melhor valor.

4.1.1.3 Rand Ajustado

O índice Rand é a razão de pares de elementos que concordam no mapeamento (mesmo grupo se, e somente se, forem da mesma classe) em relação ao total de pares. A versão ajustada está descrita pela equação 4.18. Ele varia entre 0 e 1, sendo 1 o melhor valor. (HUBERT; ARABIE, 1985)

$$AR = \frac{a+d-Adj}{a+b+c+d-Adj} \begin{cases} a = pares\ com\ mesma\ classe\ e\ mesmo\ grupo \\ b = pares\ com\ mesma\ classe,\ mas\ grupos\ diferentes \\ c = pares\ com\ classes\ diferentes,\ mas\ mesmo\ grupo \\ d = pares\ com\ classes\ e\ grupos\ diferentes \\ Adj = \frac{(a\ +\ b)(a\ +\ c)\ +\ (d\ +\ b)(d\ +\ c)}{a\ +\ b\ +\ c\ +\ d} \end{cases} \tag{4.18}$$

4.1.1.4 Informação Mútua Normalizada (NMI)

Considerando Classe e Grupo como duas Variáveis Aleatórias, temos que a Informação Mútua Normalizada é a razão da Informação Mútua entre essas variáveis e a soma das Entropias, como descrito pela equação 4.19. Ele varia entre 0 e 1, sendo 1 o melhor valor. (STREHL; GHOSH, 2003)

$$NMI = \frac{2I(C,G)}{H(C)+H(G)} = \frac{2\sum_{c\in C}\sum_{g\in G} p(c,g)log_2\left(\frac{p(c,g)}{p(c)p(g)}\right)}{\left[\sum_{c\in C} p(c)log_2\left(\frac{1}{p(c)}\right)\right]+\left[\sum_{g\in G} p(g)log_2\left(\frac{1}{p(g)}\right)\right]} \tag{4.19}$$

4.1.1.5 Coeficiente de Silhueta

Esse índice (ver equações 4.20 e 4.21) leva em consideração a distância de cada elemento para os outros elementos do seu grupo (equação 4.22), e a distância para para os elementos do grupo diferente mais próximo (equação 4.23), sendo G_i o grupo principal de x_i. Ele varia entre -1 e 1, sendo positivo caso os elementos estejam mais próximos dos elementos do próprio grupo do que dos outros grupos (o que é melhor). (KAUFMAN; ROUSSEEUW, 1990b)

$$S = \frac{\sum_{i=1}^{N} Si}{N} \tag{4.20}$$

$$Si = \frac{O(x_i) - I(x_i)}{max(O(x_i), I(x_i))} \tag{4.21}$$

$$I(x_i) = \frac{\sum_{e \in G_i, e \neq x_i} d(e, x_i)}{|G_i|} \tag{4.22}$$

$$O(x_i) = min_{G_j \neq G_i} \left\{ \frac{\sum_{e \in G_j} d(e, x_i)}{|G_j|} \right\} \tag{4.23}$$

4.1.2 Índices para partições difusas

Esses índices são únicos para classificadores difusos.

4.1.2.1 Rand Frigui

Esse índice é uma variação do Rand para classificações difusas. como descrito pela equação 4.24, considerando as equações 4.25 e 4.26. Ele varia entre 0 e 1, sendo 1 o melhor valor. (FRIGUI; HWANGA; RHEE, 2007)

$$RF = \frac{Nss + Ndd}{Nss + Nsd + Nds + Ndd} \begin{cases} Nss = \sum_{i \neq j} (\phi 1_{ij})(\phi 2_{ij}) \\ Nsd = \sum_{i \neq j} (\phi 1_{ij})(1 - \phi 2_{ij}) \\ Nds = \sum_{i \neq j} (1 - \phi 1_{ij})(\phi 2_{ij}) \\ Ndd = \sum_{i \neq j} (1 - \phi 1_{ij})(1 - \phi 2_{ij}) \end{cases} \tag{4.24}$$

$$\phi 1_{ij} = \sum_{k=1}^{K} (u_{ik})(u_{jk}) \tag{4.25}$$

$$\phi 2_{ij} = \begin{cases} 1, \ se \ x_i \ e \ x_j \ forem \ da \ mesma \ classe \\ 0, \ caso \ contrário \end{cases} \tag{4.26}$$

4.1.2.2 Rand Hüllermeier

Esse índice é uma variação do Rand sugerida por Hüllermeier 4.27. Ele utiliza o módulo da distância entre os vetores de pertinência (ver equação 4.28) ao invés do produto interno, como no Rand Frigui. Ele varia entre 0 e 1, sendo 1 o melhor valor. (HULLERMEIER; RIFQI, 2009)

$$RH = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} \| Ep(x_i, x_j) - Eq(x_i, x_j) \|}{N(N-1)/2} \tag{4.27}$$

$$Ep(x_i, x_j) = 1 - \frac{\sum_{k=1}^{K} \|u_{ik} - u_{jk}\|}{2} \tag{4.28}$$

$$Eq(x_i, x_j) = \begin{cases} 1, \ se\ x_i\ e\ x_j\ forem\ da\ mesma\ classe \\ 0, \ caso\ contrário \end{cases} \tag{4.29}$$

4.1.2.3 Coeficiente de Partição (Vpc)

Esse índice é útil para determinar o quanto as partições estão difusas. ele varia entre 1/K (caso as partições sejam equiprováveis) e 1 (caso as partições sejam exclusivas). O ideal é que esse índice não seja nem muito baixo (pois significa que os elementos foram mapeados para todos os grupos) e nem muito alto (pois significa que as partições se tornaram exclusivas). (BEZDEK, 1981)

$$Vpc = \sum_{i=1}^{N} \sum_{k=1}^{K} (u_{ik})^2 \tag{4.30}$$

4.1.2.4 Coeficiente de Partição Modificado (Vmpc)

Esse índice normaliza o coeficiente de Partição para os valores entre 0 e 1. (DAVE, 1996)

$$Vmpc = \frac{Vpc \cdot K - 1}{K - 1} \tag{4.31}$$

4.1.2.5 Entropia da Partição (Vpe)

Essa é a média das entropia das partições de cada elemento. Ele varia entre 0 e log(K). (BEZDEK, 1974)

$$Vpe = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{k=1}^{K} (u_{ik}) log(\frac{1}{u_{ik}}) \right] \tag{4.32}$$

4.1.2.6 Coeficiente de Silhueta Difusa

Versão difusa do Coeficiente de Silhueta (ver equação 4.33) e utiliza o cálculo de Si do coeficiente de silhueta (rever equação 4.21). Assim como o coeficiente de silhueta exclusivo, esse varia entre entre -1 e 1. (CAMPELLO; HRUSCHKA, 2006)

$$S = \frac{\sum_{i=1}^{N} (u_{iG_1} - u_{iG_2}) Si}{\sum_{i=1}^{N} (u_{iG_1} - u_{iG_2})} \begin{cases} u_{iG_1}, \ a\ maior\ pertinência\ de\ x_i \\ u_{iG_2}, \ a\ segunda\ maior\ pertinência\ de\ x_i \end{cases} \tag{4.33}$$

4.2 Bases

Foram utilizadas 10 bases de classificação obtidas no repositório UCI (https://archive.ics.uci.edu/ml/datasets.html), que é um repositório gratuito de bases utilizadas em aprendizagem de máquina. Como os elementos estão representados nessas bases por parâmetros, foi necessário calcular a matriz de dissimilaridade calculando as distâncias euclidianas entre os elementos.

Tabela 1: Bases testadas

Base	Nº de elementos	Nº de Classes
Breast Cancer Wisconsin	683	2
Glass Identification	214	6
Image Segmentation (train)	210	7
Ionosphere	351	2
Iris	150	3
Libras Movement	360	15
Seeds	210	3
Sonar, Mines vs. Rocks	208	2
Thyroid Disease (new-thyroid)	215	3
Wine	178	3

4.3 Parâmetros

Para todos os algoritmos, o número de grupos escolhidos para cada base é o mesmo do número de classes das bases e foi usado $\epsilon = 1e{-}9$. Para o algoritmo de referência FMMdd, foi usado $m = 1.9$ e $n = 1.5$. Para o algoritmo Fanny, foi usado $m = 2.0$ e $n = 2.0$. Para o algoritmo Nerf, foi usado $m = 2.0$ e $q = 2.0$.

Para o algoritmo proposto, os valores dos parâmetros Tu e Tv variaram para cada base e estão descritos na Tabela 2. Infelizmente não houve tempo para fazer uma busca nos parâmetros do Fanny, Nerf e FMMdd. por tanto foram usados valores normalmente utilizados ou apresentados no artigo de origem.

Tabela 2: Parâmetros do algoritmo proposto

Base	Tu	Tv
Breast Cancer Wisconsin	0.162522	24.5811
Glass Identification	0.039524	4.62473
Image Segmentation (train)	0.935430	296.990
Ionosphere	0.245895	88.4972
Iris	0.327030	19.9434
Libras Movement	0.223499	6.93447
Seeds	0.534109	78.2366
Sonar, Mines vs. Rocks	0.162522	24.5811
Thyroid Disease (new-thyroid)	0.545362	190.092
Wine	43.75360	5718.63

Resultados

Os resultados obtidos estão descritos na tabela 3 e, para facilitar a visualização, foi calculado a quantidade de vezes que cada algoritmo obteve os maior e menor valor de índice (como demostrado na tabela 6). Foram mantidos os resultados para as execuções com menor função-objetivo (ver tabelas 5 e 4).

Também é importante notar que o algoritmo proposto foi em torno de 39% mais rápido do que o modelo FMMdd, 1349% mais rápido do que o modelo Nerf, e 16706% mais rápido do que o modelo Fanny.

Tabela 3: Resultados de cada algoritmo para as bases testadas

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Breast Cancer Wisconsin				
F-Measure	0.9602 ± 0.0	0.9709 ± 0.195e-7	0.9678 ± 0.0	0.9693 ± 0.0
Erro Geral	0.0395 ± 0.124e-8	0.0293 ± 0.0	0.0322 ± 0.160e-8	0.0307 ± 0.0
Rand Ajustado	0.8464 ± 0.0	0.8855 ± 0.347e-8	0.8742 ± 0.0	0.8797 ± 0.0
NMI	0.7492 ± 0.992e-8	0.8112 ± 0.459e-7	0.7830 ± 0.508e-7	0.7910 ± 0.0
Silhueta	0.5983 ± 0.0	0.5817 ± 0.0	0.4861 ± 0.0	0.4858 ± 0.0
Rand Frigui	0.9224 ± 0.117e-7	0.5000 ± 0.105e-7	0.7872 ± 0.136e-7	0.6618 ± 0.0
Rand Hullermeier	0.9224 ± 0.129e-7	0.5444 ± 0.206e-7	0.8230 ± 0.0	0.7364 ± 0.0
Vpc	0.9985 ± 0.299e-7	0.5000 ± 0.0	0.8257 ± 0.166e-7	0.6899 ± 0.128e-7
Vmpc	0.9970 ± 0.118e-7	0.264e-11 ± 0.222e-12	0.6515 ± 0.182e-7	0.3799 ± 0.215e-7
Vpe	0.0024 ± 0.119e-9	0.6931 ± 0.130e-7	0.2966 ± 0.131e-7	0.4820 ± 0.589e-8
Silhueta Difusa	0.5995 ± 0.0	0.7015 ± 0.838e-8	0.5821 ± 0.162e-7	0.6148 ± 0.165e-7
Tempo de execusão (s)	0.1792 ± 0.0173	0.5547 ± 0.0838	2.2693 ± 0.1630	43.6666 ± 3.5244

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Glass Identification				
F-Measure	0.4916 ± 0.0335	0.4914 ± 0.0	0.4879 ± 0.139e-7	0.5381 ± 0.0015
Erro Geral	0.4286 ± 0.0220	0.5888 ± 0.0	0.4439 ± 0.0	0.4579 ± 0.215e-7
Rand Ajustado	0.1731 ± 0.0274	0.1646 ± 0.661e-8	0.2029 ± 0.211e-8	0.2416 ± 0.0019
NMI	0.2721 ± 0.0349	0.2347 ± 0.0	0.3394 ± 0.0	0.3787 ± 0.0015
Silhueta	-0.9507 ± 0.2417	-1.0000 ± 0.0	-0.7617 ± 0.245e-7	-0.7102 ± 0.0005
Rand Frigui	0.6882 ± 0.0141	0.6601 ± 0.0	0.6815 ± 0.0	0.6671 ± 0.148e-7
Rand Hullermeier	0.6868 ± 0.0140	0.2598 ± 0.852e-7	0.5275 ± 0.0	0.4047 ± 0.622e-6
Vpc	0.9381 ± 0.0087	0.1667 ± 0.552e-8	0.3288 ± 0.0	0.2079 ± 0.349e-6
Vmpc	0.9257 ± 0.0104	0.601e-12 ± 0.221e-12	0.1945 ± 0.0	0.0495 ± 0.419e-6
Vpe	0.1072 ± 0.0161	1.7918 ± 0.0	1.3343 ± 0.232e-7	1.6597 ± 0.121e-5
Silhueta Difusa	-0.9497 ± 0.2465	-1.0000 ± 0.0	-0.3819 ± 0.0274	-0.0510 ± 0.0409
Tempo de execusão (s)	0.1628 ± 0.0591	0.0454 ± 0.0066	5.4494 ± 0.3983	23.7977 ± 5.0227
Image Segmentation (train)				
F-Measure	0.5904 ± 0.0381	0.3525 ± 0.0137	0.5869 ± 0.0082	0.3511 ± 0.0
Erro Geral	0.4466 ± 0.0450	0.7293 ± 0.0153	0.4243 ± 0.0081	0.7286 ± 0.294e-7
Rand Ajustado	0.3569 ± 0.0409	0.1083 ± 0.0148	0.3505 ± 0.0065	0.1103 ± 0.0
NMI	0.5158 ± 0.0410	0.1901 ± 0.0232	0.4661 ± 0.0046	0.1903 ± 0.0
Silhueta	-0.8144 ± 0.4574	-0.9996 ± 0.0015	-0.7098 ± 0.0049	-1.0000 ± 0.0
Rand Frigui	0.8156 ± 0.0218	0.7580 ± 0.479e-7	0.7817 ± 0.447e-4	0.7580 ± 0.0
Rand Hullermeier	0.8156 ± 0.0218	0.1388 ± 0.119e-7	0.5032 ± 0.0009	0.1388 ± 0.188e-6
Vpc	0.9919 ± 0.0017	0.1429 ± 0.177e-8	0.2937 ± 0.0003	0.1429 ± 0.438e-8
Vmpc	0.9906 ± 0.0020	0.117e-14 ± 0.878e-15	0.1759 ± 0.0003	0.413e-11 ± 0.271e-12
Vpe	0.0160 ± 0.0028	1.9459 ± 0.0	1.5432 ± 0.0008	1.9459 ± 0.559e-8
Silhueta Difusa	-0.8151 ± 0.4582	-1.0000 ± 0.100e-4	-0.5979 ± 0.0430	-1.0000 ± 0.0
Tempo de execusão (s)	0.3430 ± 0.3231	0.0413 ± 0.0088	8.5064 ± 0.0153	30.5286 ± 1.5735

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
		Ionosphere		
F-Measure	0.6914 ± 0.344e-7	0.7181 ± 0.400e-7	0.7123 ± 0.0	0.7040 ± 0.290e-7
Erro Geral	0.3134 ± 0.0	0.2877 ± 0.0	0.2934 ± 0.0	0.3020 ± 0.0
Rand Ajustado	0.1356 ± 0.382e-8	0.1780 ± 0.0	0.1682 ± 0.502e-8	0.1545 ± 0.106e-7
NMI	0.0925 ± 0.547e-8	0.1467 ± 0.0	0.1314 ± 0.0	0.1231 ± 0.0
Silhueta	0.3026 ± 0.209e-7	0.2889 ± 0.0	0.2539 ± 0.137e-7	0.2530 ± 0.0
Rand Frigui	0.5520 ± 0.325e-7	0.5000 ± 0.167e-7	0.5181 ± 0.175e-7	0.5000 ± 0.0
Rand Hullermeier	0.5560 ± 0.169e-7	0.5385 ± 0.0	0.5395 ± 0.701e-8	0.5385 ± 0.0
Vpc	0.8990 ± 0.380e-7	0.5000 ± 0.0	0.5910 ± 0.107e-6	0.5000 ± 0.0
Vmpc	0.7980 ± 0.687e-7	0.297e-12 ± 0.156e-12	0.1821 ± 0.214e-6	0.205e-8 ± 0.202e-9
Vpe	0.1695 ± 0.438e-7	0.6931 ± 0.0	0.5955 ± 0.123e-6	0.6931 ± 0.0
Silhueta Difusa	0.3415 ± 0.525e-7	0.4105 ± 0.163e-7	0.2949 ± 0.822e-8	0.2906 ± 0.490e-5
Tempo de execusão (s)	0.0724 ± 0.0071	0.0290 ± 0.0081	1.1856 ± 0.1319	68.8237 ± 12.7258
		Iris		
F-Measure	0.9267 ± 0.384e-7	0.9265 ± 0.263e-7	0.8995 ± 0.0	0.9131 ± 0.515e-7
Erro Geral	0.0733 ± 0.249e-8	0.0733 ± 0.249e-8	0.1000 ± 0.676e-8	0.0867 ± 0.0
Rand Ajustado	0.8015 ± 0.637e-7	0.8019 ± 0.0	0.7424 ± 0.0	0.7711 ± 0.0
NMI	0.7900 ± 0.554e-7	0.7959 ± 0.383e-7	0.7518 ± 0.0	0.7705 ± 0.122e-7
Silhueta	0.5371 ± 0.0	0.5316 ± 0.270e-7	-0.3267 ± 0.0	-0.3267 ± 0.0
Rand Frigui	0.8322 ± 0.123e-6	0.6419 ± 0.688e-8	0.8022 ± 0.902e-8	0.6994 ± 0.492e-7
Rand Hullermeier	0.8466 ± 0.902e-7	0.6001 ± 0.0	0.8115 ± 0.0	0.6962 ± 0.596e-7
Vpc	0.7922 ± 0.257e-6	0.4661 ± 0.0	0.7503 ± 0.186e-7	0.5677 ± 0.118e-6
Vmpc	0.6883 ± 0.387e-6	0.1992 ± 0.0	0.6254 ± 0.506e-7	0.3515 ± 0.176e-6
Vpe	0.3258 ± 0.342e-6	0.8869 ± 0.125e-7	0.4489 ± 0.388e-7	0.7464 ± 0.142e-6
Silhueta Difusa	0.6411 ± 0.923e-7	0.7953 ± 0.270e-5	-0.4466 ± 0.0076	-0.5276 ± 0.0064
Tempo de execusão (s)	0.0300 ± 0.0111	0.0191 ± 0.0107	0.2903 ± 0.0214	0.9879 ± 0.1505

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Libras Movement				
F-Measure	0.2564 ± 0.0053	0.2539 ± 0.0220	0.2407 ± 0.0173	0.2554 ± 0.0215
Erro Geral	0.8007 ± 0.0060	0.7870 ± 0.0230	0.8047 ± 0.0181	0.7900 ± 0.0230
Rand Ajustado	0.0848 ± 0.0043	0.0862 ± 0.0176	0.0797 ± 0.0136	0.0910 ± 0.0152
NMI	0.2470 ± 0.0071	0.2627 ± 0.0313	0.2452 ± 0.0257	0.2652 ± 0.0287
Silhueta	-1.0000 ± 0.0	-0.9980 ± 0.0027	-0.9934 ± 0.0123	-0.9886 ± 0.0258
Rand Frigui	0.8356 ± 0.0015	0.8778 ± 0.0	0.8778 ± 0.0	0.8778 ± 0.291e-7
Rand Hullermeier	0.6048 ± 0.0100	0.0641 ± 0.197e-7	0.0641 ± 0.392e-6	0.0641 ± 0.750e-7
Vpc	0.4065 ± 0.0083	0.0667 ± 0.0	0.0667 ± 0.0	0.0667 ± 0.152e-8
Vmpc	0.3641 ± 0.0088	0.598e-14 ± 0.174e-14	0.325e-11 ± 0.737e-12	0.841e-13 ± 0.250e-13
Vpe	1.6640 ± 0.0245	2.7081 ± 0.109e-6	2.7081 ± 0.0	2.7081 ± 0.383e-7
Silhueta Difusa	-1.0000 ± 0.0	-0.9998 ± 0.0005	-0.9995 ± 0.0017	-0.9984 ± 0.0065
Tempo de execusão (s)	2.7545 ± 0.5404	0.1158 ± 0.0089	5.8616 ± 0.1380	36.3513 ± 0.1903
Seeds				
F-Measure	0.8830 ± 0.0	0.6880 ± 0.238e-7	0.8954 ± 0.498e-7	0.8904 ± 0.0
Erro Geral	0.1143 ± 0.731e-8	0.3333 ± 0.657e-8	0.1048 ± 0.718e-8	0.1095 ± 0.749e-8
Rand Ajustado	0.7006 ± 0.130e-7	0.4462 ± 0.0	0.7166 ± 0.0	0.7056 ± 0.521e-7
NMI	0.6900 ± 0.0	0.5215 ± 0.243e-7	0.6949 ± 0.0	0.6793 ± 0.210e-8
Silhueta	0.4592 ± 0.0	-1.0000 ± 0.0	-0.4190 ± 0.287e-7	0.0561 ± 0.197e-8
Rand Frigui	0.8184 ± 0.130e-6	0.5566 ± 0.0	0.7456 ± 0.123e-6	0.6461 ± 0.357e-7
Rand Hullermeier	0.8237 ± 0.111e-6	0.3301 ± 0.591e-7	0.7401 ± 0.129e-6	0.6102 ± 0.531e-7
Vpc	0.8361 ± 0.673e-7	0.3333 ± 0.869e-8	0.6920 ± 0.348e-7	0.5039 ± 0.324e-7
Vmpc	0.7541 ± 0.108e-6	0.566e-11 ± 0.310e-12	0.5380 ± 0.530e-7	0.2559 ± 0.441e-7
Vpe	0.2832 ± 0.109e-6	1.0986 ± 0.333e-7	0.5590 ± 0.527e-7	0.8502 ± 0.426e-7
Silhueta Difusa	0.5099 ± 0.970e-7	-1.0000 ± 0.0	-0.3372 ± 0.104e-5	0.1781 ± 0.887e-7
Tempo de execusão (s)	0.0547 ± 0.0044	0.1484 ± 0.0146	0.5201 ± 0.0558	2.6520 ± 0.2591

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Sonar, Mines vs. Rocks				
F-Measure	0.5514 ± 0.0	0.5508 ± 0.0132	0.5531 ± 0.0013	0.5597 ± 0.0091
Erro Geral	0.4663 ± 0.257e-7	0.4479 ± 0.0108	0.4470 ± 0.0014	0.4401 ± 0.0090
Rand Ajustado	-0.0025 ± 0.0	0.0063 ± 0.0057	0.0065 ± 0.0006	0.0099 ± 0.0059
NMI	0.0032 ± 0.0	0.0088 ± 0.0040	0.0088 ± 0.0005	0.0116 ± 0.0041
Silhueta	0.2079 ± 0.843e-8	0.1995 ± 0.0143	0.1778 ± 0.0053	0.1644 ± 0.0131
Rand Frigui	0.5011 ± 0.290e-7	0.5000 ± 0.745e-8	0.5000 ± 0.146e-7	0.5000 ± 0.0
Rand Hullermeier	0.5018 ± 0.306e-7	0.4999 ± 0.0	0.4999 ± 0.0	0.4999 ± 0.105e-7
Vpc	0.7361 ± 0.254e-6	0.5000 ± 0.149e-7	0.5000 ± 0.0	0.5000 ± 0.0
Vmpc	0.4722 ± 0.508e-6	0.132e-12 ± 0.151e-12	0.154e-9 ± 0.207e-10	0.234e-11 ± 0.118e-11
Vpe	0.4184 ± 0.347e-6	0.6931 ± 0.877e-8	0.6931 ± 0.812e-8	0.6931 ± 0.0
Silhueta Difusa	0.2496 ± 0.157e-6	0.2580 ± 0.0085	0.2237 ± 0.0027	0.2196 ± 0.0157
Tempo de execusão (s)	0.0419 ± 0.0086	0.0074 ± 0.0020	0.4190 ± 0.0731	1.1541 ± 0.1052
Thyroid Disease (new-thyroid)				
F-Measure	0.5763 ± 0.0320	0.5458 ± 0.0030	0.6156 ± 0.401e-7	0.4742 ± 0.0
Erro Geral	0.2641 ± 0.0393	0.3019 ± 0.0042	0.3023 ± 0.160e-7	0.3023 ± 0.160e-7
Rand Ajustado	0.0950 ± 0.0586	0.0387 ± 0.0060	0.1355 ± 0.565e-8	0.0323 ± 0.0
NMI	0.1645 ± 0.0691	0.0982 ± 0.0074	0.1973 ± 0.826e-8	0.1475 ± 0.0
Silhueta	-0.3489 ± 0.6643	-0.9781 ± 0.1588	0.1432 ± 0.0	0.3637 ± 0.125e-7
Rand Frigui	0.5479 ± 0.0238	0.4898 ± 0.123e-7	0.5143 ± 0.298e-6	0.4908 ± 0.126e-7
Rand Hullermeier	0.5483 ± 0.0238	0.5305 ± 0.165e-7	0.5199 ± 0.374e-6	0.5107 ± 0.109e-6
Vpc	0.9559 ± 0.0012	0.3333 ± 0.755e-8	0.5289 ± 0.334e-6	0.3699 ± 0.117e-6
Vmpc	0.9339 ± 0.0018	0.804e-13 ± 0.374e-13	0.2934 ± 0.500e-6	0.0549 ± 0.177e-6
Vpe	0.0726 ± 0.0029	1.0986 ± 0.403e-7	0.8047 ± 0.473e-6	1.0432 ± 0.183e-6
Silhueta Difusa	-0.3397 ± 0.6737	-0.9766 ± 0.1677	0.1701 ± 0.212e-5	0.4870 ± 0.180e-5
Tempo de execusão (s)	0.1324 ± 0.0556	0.0224 ± 0.0080	2.0978 ± 0.2508	11.2677 ± 2.0174

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
		Wine		
F-Measure	0.7217 ± 0.0	0.7249 ± 0.113e-7	0.7196 ± 0.329e-7	0.7196 ± 0.329e-7
Erro Geral	0.2865 ± 0.0	0.2809 ± 0.176e-7	0.2921 ± 0.0	0.2921 ± 0.0
Rand Ajustado	0.3863 ± 0.165e-7	0.4019 ± 0.197e-7	0.3676 ± 0.0	0.3676 ± 0.0
NMI	0.4224 ± 0.0	0.4173 ± 0.200e-7	0.4179 ± 0.277e-7	0.4179 ± 0.277e-7
Silhueta	0.5683 ± 0.0	0.5562 ± 0.455e-7	-0.2584 ± 0.0	-0.2584 ± 0.0
Rand Frigui	0.7218 ± 0.0	0.5874 ± 0.0	0.6866 ± 0.109e-7	0.6361 ± 0.162e-7
Rand Hullermeier	0.7167 ± 0.130e-7	0.5043 ± 0.0	0.6755 ± 0.0	0.6007 ± 0.0
Vpc	0.8239 ± 0.309e-7	0.4205 ± 0.0	0.7925 ± 0.0	0.6043 ± 0.119e-7
Vmpc	0.7358 ± 0.575e-7	0.1307 ± 0.329e-8	0.6888 ± 0.0	0.4065 ± 0.913e-8
Vpe	0.2942 ± 0.592e-7	0.9545 ± 0.149e-7	0.3924 ± 0.0	0.7004 ± 0.129e-7
Silhueta Difusa	0.6108 ± 0.224e-7	0.6034 ± 0.0	-0.2248 ± 0.206e-6	-0.2179 ± 0.121e-6
Tempo de execusão (s)	0.0503 ± 0.0078	0.3148 ± 0.0495	0.5251 ± 0.0502	2.1902 ± 0.1889

Tabela 4: Porcentagem em que cada algoritmo obteve o maior ou menor valor de cada índice

Índice	FMMdd-ENT		FMMdd		Nerf		Fanny	
	menor	maior	menor	maior	menor	maior	menor	maior
F-Measure	20.0	30.0	20.0	30.0	40.0	20.0	30.0	20.0
Erro Geral	30.0	30.0	50.0	30.0	20.0	40.0	10.0	20.0
Rand Ajustado	30.0	10.0	30.0	40.0	30.0	20.0	20.0	30.0
NMI	30.0	20.0	50.0	30.0	20.0	20.0	0.0	30.0
Silhueta	10.0	60.0	30.0	0.0	20.0	10.0	60.0	30.0
Rand Frigui	10.0	90.0	90.0	10.0	10.0	10.0	30.0	10.0
Rand Hullermeier	0.0	100.0	80.0	0.0	10.0	0.0	30.0	0.0
Vpc	0.0	100.0	100.0	0.0	20.0	0.0	40.0	0.0
Vmpc	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
Vpe	100.0	0.0	0.0	100.0	0.0	20.0	0.0	40.0
Silhueta Difusa	10.0	20.0	30.0	40.0	20.0	10.0	40.0	30.0

Tabela 5: Melhores Resultados de cada algoritmo para as bases testadas

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Breast Cancer Wisconsin				
F-Measure	0.9602	0.9709	0.9678	0.9693
Erro Geral	0.0395	0.0293	0.0322	0.0307
Rand Ajustado	0.8464	0.8855	0.8742	0.8797
NMI	0.7492	0.8112	0.7830	0.7910
Silhueta	0.5983	0.5817	0.4861	0.4858
Rand Frigui	0.9224	0.5000	0.7872	0.6618
Rand Hullermeier	0.9224	0.5444	0.8230	0.7364
Vpc	0.9985	0.5000	0.8257	0.6899
Vmpc	0.9970	0.309e-11	0.6515	0.3799
Vpe	0.0024	0.6931	0.2966	0.4820
Silhueta Difusa	0.5995	0.7015	0.5821	0.6148
Glass Identification				
F-Measure	0.5383	0.4914	0.4879	0.5382
Erro Geral	0.3878	0.5888	0.4439	0.4579
Rand Ajustado	0.2126	0.1646	0.2029	0.2418
NMI	0.3392	0.2347	0.3394	0.3788
Silhueta	-1.0000	-1.0000	-0.7617	-0.7103
Rand Frigui	0.7034	0.6601	0.6815	0.6671
Rand Hullermeier	0.7023	0.2598	0.5275	0.4047
Vpc	0.9428	0.1667	0.3288	0.2079
Vmpc	0.9313	0.112e-11	0.1945	0.0495
Vpe	0.0963	1.7918	1.3343	1.6597
Silhueta Difusa	-1.0000	-1.0000	-0.3766	-0.0466
Image Segmentation (train)				
F-Measure	0.6625	0.3506	0.5951	0.3511
Erro Geral	0.3429	0.7333	0.4143	0.7286
Rand Ajustado	0.4480	0.1056	0.3570	0.1103
NMI	0.5734	0.1837	0.4701	0.1903
Silhueta	0.3142	-1.0000	-0.7143	-1.0000
Rand Frigui	0.8657	0.7580	0.7818	0.7580
Rand Hullermeier	0.8657	0.1388	0.5036	0.1388
Vpc	0.9905	0.1429	0.2938	0.1429
Vmpc	0.9889	0.122e-14	0.1761	0.361e-11
Vpe	0.0192	1.9459	1.5428	1.9459
Silhueta Difusa	0.3172	-1.0000	-0.6249	-1.0000
Ionosphere				
F-Measure	0.6914	0.7181	0.7123	0.7040
Erro Geral	0.3134	0.2877	0.2934	0.3020
Rand Ajustado	0.1356	0.1780	0.1682	0.1545
NMI	0.0925	0.1467	0.1314	0.1231
Silhueta	0.3026	0.2889	0.2539	0.2530
Rand Frigui	0.5520	0.5000	0.5181	0.5000
Rand Hullermeier	0.5560	0.5385	0.5395	0.5385
Vpc	0.8990	0.5000	0.5910	0.5000
Vmpc	0.7980	0.657e-12	0.1821	0.113e-9
Vpe	0.1695	0.6931	0.5955	0.6931
Silhueta Difusa	0.3415	0.4105	0.2949	0.2907

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
		Iris		
F-Measure	0.9267	0.9265	0.8995	0.9131
Erro Geral	0.0733	0.0733	0.1000	0.0867
Rand Ajustado	0.8015	0.8019	0.7424	0.7711
NMI	0.7900	0.7959	0.7518	0.7705
Silhueta	0.5371	0.5316	-0.3267	-0.3267
Rand Frigui	0.8322	0.6419	0.8022	0.6994
Rand Hullermeier	0.8466	0.6001	0.8115	0.6962
Vpc	0.7922	0.4661	0.7503	0.5677
Vmpc	0.6883	0.1992	0.6254	0.3515
Vpe	0.3258	0.8869	0.4489	0.7464
Silhueta Difusa	0.6411	0.7953	-0.4517	-0.5262
		Libras Movement		
F-Measure	0.2765	0.2183	0.2748	0.2275
Erro Geral	0.7889	0.8361	0.7722	0.8278
Rand Ajustado	0.0992	0.0625	0.1011	0.0462
NMI	0.2754	0.2011	0.2810	0.2019
Silhueta	-1.0000	-1.0000	-0.9861	-0.9833
Rand Frigui	0.8306	0.8778	0.8778	0.8778
Rand Hullermeier	0.6467	0.0641	0.0641	0.0641
Vpc	0.4636	0.0667	0.0667	0.0667
Vmpc	0.4253	0.910e-14	0.202e-11	0.397e-13
Vpe	1.4670	2.7081	2.7081	2.7081
Silhueta Difusa	-1.0000	-1.0000	-0.9959	-0.9995
		Seeds		
F-Measure	0.8830	0.6880	0.8954	0.8904
Erro Geral	0.1143	0.3333	0.1048	0.1095
Rand Ajustado	0.7006	0.4462	0.7166	0.7056
NMI	0.6900	0.5215	0.6949	0.6793
Silhueta	0.4592	-1.0000	-0.4190	0.0561
Rand Frigui	0.8184	0.5566	0.7456	0.6461
Rand Hullermeier	0.8237	0.3301	0.7401	0.6102
Vpc	0.8361	0.3333	0.6920	0.5039
Vmpc	0.7541	0.623e-11	0.5380	0.2559
Vpe	0.2832	1.0986	0.5590	0.8502
Silhueta Difusa	0.5099	-1.0000	-0.3372	0.1781
		Sonar, Mines vs. Rocks		
F-Measure	0.5514	0.5577	0.5581	0.5437
Erro Geral	0.4663	0.4423	0.4423	0.4567
Rand Ajustado	-0.0025	0.0085	0.0085	0.0027
NMI	0.0032	0.0105	0.0094	0.0053
Silhueta	0.2079	0.2063	0.1557	0.0727
Rand Frigui	0.5011	0.5000	0.5000	0.5000
Rand Hullermeier	0.5018	0.4999	0.4999	0.4999
Vpc	0.7361	0.5000	0.5000	0.5000
Vmpc	0.4722	0.635e-12	0.614e-10	0.290e-12
Vpe	0.4184	0.6931	0.6931	0.6931
Silhueta Difusa	0.2496	0.2621	0.2214	0.1377

Índice	FMMdd-ENT	FMMdd	Nerf	Fanny
Thyroid Disease (new-thyroid)				
F-Measure	0.5455	0.5455	0.6156	0.4742
Erro Geral	0.3023	0.3023	0.3023	0.3023
Rand Ajustado	0.0382	0.0382	0.1355	0.0323
NMI	0.0974	0.0974	0.1973	0.1475
Silhueta	-1.0000	-1.0000	0.1432	0.3637
Rand Frigui	0.5248	0.4898	0.5143	0.4908
Rand Hullermeier	0.5252	0.5305	0.5199	0.5107
Vpc	0.9571	0.3333	0.5289	0.3699
Vmpc	0.9357	0.168e-12	0.2934	0.0549
Vpe	0.0699	1.0986	0.8047	1.0432
Silhueta Difusa	-1.0000	-1.0000	0.1701	0.4870
Wine				
F-Measure	0.7217	0.7249	0.7196	0.7196
Erro Geral	0.2865	0.2809	0.2921	0.2921
Rand Ajustado	0.3863	0.4019	0.3676	0.3676
NMI	0.4224	0.4173	0.4179	0.4179
Silhueta	0.5683	0.5562	-0.2584	-0.2584
Rand Frigui	0.7218	0.5874	0.6866	0.6361
Rand Hullermeier	0.7167	0.5043	0.6755	0.6007
Vpc	0.8239	0.4205	0.7925	0.6043
Vmpc	0.7358	0.1307	0.6888	0.4065
Vpe	0.2942	0.9545	0.3924	0.7004
Silhueta Difusa	0.6108	0.6034	-0.2248	-0.2179

Tabela 6: Porcentagem em que cada algoritmo obteve o maior ou menor valor de cada índice para a melhor solução

Índice	FMMdd-ENT		FMMdd		Nerf		Fanny	
	menor	maior	menor	maior	menor	maior	menor	maior
F-Measure	20.0	40.0	30.0	30.0	30.0	30.0	30.0	0.0
Erro Geral	30.0	30.0	50.0	40.0	30.0	20.0	0.0	10.0
Rand Ajustado	30.0	10.0	30.0	50.0	20.0	30.0	30.0	10.0
NMI	40.0	20.0	60.0	40.0	10.0	30.0	0.0	10.0
Silhueta	30.0	70.0	50.0	0.0	20.0	0.0	60.0	30.0
Rand Frigui	10.0	90.0	90.0	10.0	10.0	10.0	30.0	10.0
Rand Hullermeier	0.0	90.0	90.0	10.0	10.0	0.0	40.0	0.0
Vpc	0.0	100.0	100.0	0.0	20.0	0.0	40.0	0.0
Vmpc	0.0	100.0	90.0	0.0	0.0	0.0	10.0	0.0
Vpe	100.0	0.0	0.0	100.0	0.0	20.0	0.0	40.0
Silhueta Difusa	30.0	30.0	50.0	40.0	20.0	10.0	40.0	20.0

Conclusão

Nesse Trabalho apresentamos um modelo de classificação difusa não-supervisionado para dados relacionais compatível com modelos da literatura como o Fanny e Nerf, além do modelo de referência FMMdd. Os modelos obtiveram F-Measure e Erro Geral semelhantes e o modelo proposto obteve um Rand Ajustado menor na maioria das bases. Contudo ele obteve maiores Rand Frigui, Rand Hulçermeier, Coeficiente de Partição, Entropia da partição, Coeficiente de Silhueta e Coeficiente de Silhueta Difusa, além de ser consideravelmente mais rápido o que o torna interessante.

Trabalhos futuros serão fazer versões para múltiplas matrizes, podendo ser usadas formas de ponderação diferentes para a influência das matrizes. Também pode-se buscar variações da função-objetivo e formas mais eficientes de selecionar os parâmetros, ou de ajustá-los durante a execução.

Referências

BEZDEK, J. C. Cluster validity with fuzzy sets. *J. Cybernetics*, v. 3, p. 58–72, 1974. Citado na página 18.

BEZDEK, J. C. Pattern recognition with fuzzy objective function algorithms. *Plenum Press*, New York, 1981. Citado na página 18.

BREIMAN, L. et al. Classification and regression trees. *Chapman and Hall/CRC*, Boca Raton, 1984. Citado na página 16.

CAMPELLO, R. J.; HRUSCHKA, E. R. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, v. 157, n. 21, p. 2858–2875, 2006. Citado na página 18.

DAVE, R. N. Validating fuzzy partition obtained through c-shells clustering. *Pattern Recognition*, v. 17, p. 613– 623, 1996. Citado na página 18.

FRIGUI, H.; HWANGA, C.; RHEE, F. C. H. Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition*, v. 40, n. 11, p. 3053–3068, 2007. Citado na página 17.

HATHAWAY, R. J.; BEZDEK, J. C. Nerf c-means: non-euclidean relational fuzzy clustering. *Pattern Recognition*, v. 27, n. 3, p. 429–437, 1994. Citado 2 vezes nas páginas 3 e 15.

HUBERT, L.; ARABIE, P. Comparing partitions. *J. Classification*, v. 2, p. 193–218, 1985. Citado na página 16.

HULLERMEIER, E.; RIFQI, M. A fuzzy variant of the rand index for comparing clustering structures. *Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA-EUSFLAT*, 2009. Citado na página 17.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: An introduction to cluster analysis. In: . [S.l.]: John Wiley & Sons, Inc, 1990. cap. 4. Citado 2 vezes nas páginas 3 e 15.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: An introduction to cluster analysis. In: . [S.l.]: John Wiley & Sons, Inc, 1990. Citado na página 16.

MEI, J.-P.; CHEN, L. Fuzzy relational clustering around medoids: A unified view. *Fuzzy Sets and Systems*, v. 18, p. 44–56, 2011. Citado 2 vezes nas páginas 3 e 11.

RIJISBERGEN, C. J. van. Information retrieval. *Butterworth-Heinemann*, London, 1979. Citado na página 15.

SCHWäMMLE, V.; JENSEN, O. N. A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. *Bioinformatics*, v. 26, n. 22, p. 2841–2848, 2010. Citado na página 13.

STREHL, A.; GHOSH, J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, p. 583–617, 2003. Citado na página 16.