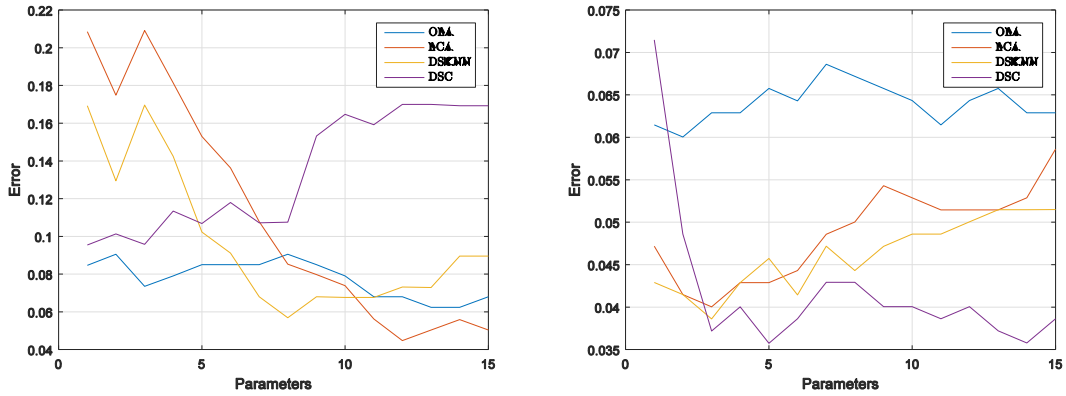
Combinação de Classificadores - Lista 4 - Seleção Dinâmica de Classificadores.

Nome: Pedro Diamel Marrero Fernández.

QUESTÃO 1

Os métodos escolhidos foram *OLA*, *LCA*, *DS-KNN*, *DS-Clusters*. Usou-se 2 bases de dados: *Breast Cancer Wisconsin (BC)*, *Wine (WN)*, da *UCI Machine Learning Repository* [1]. Empelaram-se k-fold (k=10) para separar os conjuntos de Treinamento, Teste e Validação (8,1,1). Valorou-se o desempenho dos quatro métodos frente à variação do parâmetro $K = 1, \ldots, 15$. Selecionou-se um tamanho do pool igual 100, N' e N'', 5 y 3 respectivamente, para o caso de os métodos *DS-KNN* e *DS-Clusters*. Os resultados se mostram em a Fig. 1. O melhor resultado é alcançando por *LCA* em *WN* dataset e *DS-Clusters* em *BC* dataset. A tabela 1 mostra os resultados obtidos para K=5.



a)WN b)BC

Figure 1. Resultados obtidos por os métodos OLA, LCA, DS-KNN e DS-Clusters.

Tabela 1. Resultados dos métodos para K=5.

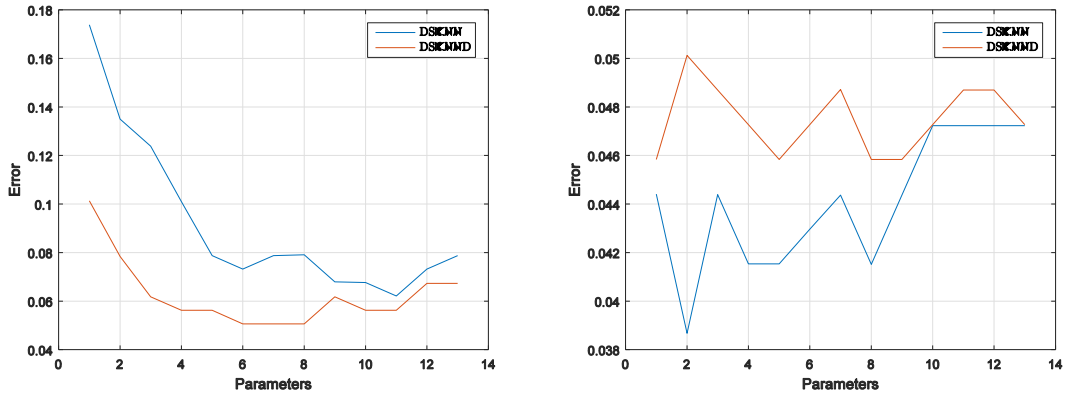
ÁRVORE DE DECISÃO	WINE (WIN)	CANCER (BC)
OLA	**0,084**	0,065
LCA	0,152	0,042
DS-KNN	0,102	0,045
DS-CLUSTER	0,106	**0,035**

Como se pode observar em a tabela 1, os melhores resultados foram obtidos por *OLA* em *WIN* e *DS-Clusters* do caso de *BC*, mas, estou resultados variam na medida que aumenta o parâmetro *K*, donde os resultados de *LCA* ficam melhores para K>11. O método DS-Clusters em WN dataset, não brinda bons resultados, estou poderia ser por que WN dataset conta com 178 objetos em seu corpus, o qual provoca que os grupos de validação

fiquem pequenos (18 objetos), por tanto os grupos formados por k-means sejam cada vez mais pequenos.

QUESTÃO 2

As técnicas de diversidade não-pareada avaliam a diversidade do pool. No caso de DS-KNN gera um ranking dos pares mais diversos do *N'* classificadores do pool selecionado a partir de LCA ranking. Para aplicar uma medida não-pareada de diversidade se avalio a diversidade do pool formado por os *N'* classificadores de maior ranking. Posteriormente se avalio os pools obtidos como resultado de ir sacando os classificadores de menor ranking hasta ficar com só duos. Finalmente o pool selecionado é o de maior diversidade. A medida de diversidade não pareada usada foi Generalized Diversity. A Fig. 2 mostra os resultados obtidos pelo método proposto vs. DS-KNN.



a)WN b)BC

Figure 2. Resultados obtidos para DS-KNN vs. Proposto.

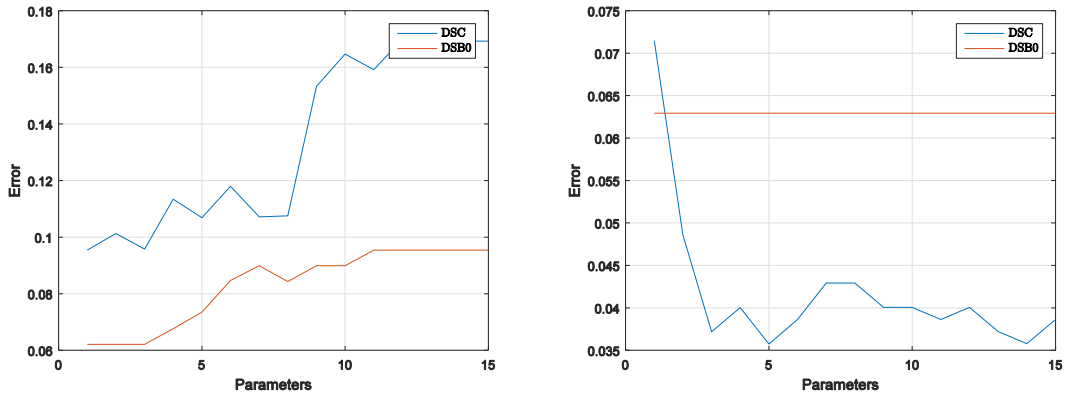
QUESTÃO 3

Para fazer um upgrade do DS-CLUSTER, selecionou-se o algoritmo B0-conexo[2]. Este algoritmo não se necessita estabelecer a priori a quantidade de grupos, por tanto é um algoritmo de classificação não supervisionado livre, só depende de um limiar (B0) que indica em que medida se parecem dois objetos. Uma das características importantes de este algoritmo é que os grupos não são circulais [2,3]. Para obter os representantes de cada grupo implementou-se o seguinte algoritmo.

Algoritmo para o cálculo dos representantes da β_0-conexo.
Entrada:
O vector de características.
β_0 limiar de semelhança.
Saída:

R vector representante do grupo.
Método:
Passo 1: Se *O* tem um só elemento, fazer *O* o representante do grupo e terminar.
Passo 2: Fazer *R(1)* igual a *O(1)* e marcar *O(1)* visitado por *R(1)*.
Passo 3: Calcular a distância do objeto *O(i)* na interação *i* para cada um dos representantes.
Passo 4: Selecionar o representante *R(p)* de menor distância a *X(i)*.
Passo 5: Se no existem criar um novo representante com o valor de *X(i)* e marcar como visitado por *R(i)*. Senão marcar *X(i)* como visitado por *R(p)* e fazer *R(p)* igual à média dos elementos visitados por ele.
Passo 6: Ir ao passo 3 hasta chegar ao objeto *O(n)*.

Um objeto considera-se do grupo se é mais perto a uns de seus representantes. O algoritmo DS-Clusters é modificado para receber os representantes de cada grupo. Em teoria este método de clusters deveria brindar grupos diferentes do algoritmo k-means. Os resultados mostram-se em a Fig. 3. Um dos problemas do uso de B0, é a estimação do parâmetro B0. A Fig 3 mostra a comparação dos dois algoritmos, DS-Clusters e o proposto, variando o parâmetro K de 1 a 15 e o parâmetro B0 de 0.0001 a 0.0015. Como pode-se observar na gráfica da Fig. 3 b), o parâmetro B0, para BC dataset, não é bom ajustado.



a)WN b)BC

Figure 3. Resultados obtidos por DS-Clusters vs. Proposto.

Referências

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] J. R. Shulcloper, E. Alba, and M. Lazo. Introducción al Reconocimiento de Patrones (Enfoque Lóogico-Combinatorio). Grupo de Reconocimiento de Patrones Cuba-México, Centro de Investigación y de Estudios Avanzados del IPN, Dpto de Ingeniería. Electrica, Serie Verde No. 51, 1995.

[3] Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2002). On-line event and topic detection by using the compact sets clustering algorithm. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, *12*(3, 4), 185-194.