Combinação de Classificadores - Lista 2 - Diversidade
Nome: Pedro Diamel Marrero Fernández.

QUESTÃO 1

A arquitetura proposta é baseada na aplicação de técnicas de clustering incrementais para criar conjuntos de classificadores (pools), usando como medida de similaridade a diversidade. Neste caso se sugere uma variante de algoritmo K-means incremental.

Algoritmo 1

T: número de interações.
d: umbral de diversidade.

Inicializar:
$S = \{\}$, conjunto de pools.

Método
Passo 1. Para $i = 1,2 \ldots T$. Criar um classificador C_i no passo i usando Bagging.
Passo 2. Adicionar a C_i

> **Se** $S = \emptyset$, criar um pool com C_i e adicioná-lo ao conjunto S.
> **Se não**
> Para cada pool P_j de S, calcular a diversidade pareada de C_i com cada elemento de P_j. Se a diversidade de todos é maior ou igual que d, fazer $P_j = \{P_j \cup C_i\}$ e ir ao passo 3. Esta estratégia garante que a medida de diversidade do pool é menor o igual que d.
> Se não existe um pool de S que satisfaz tal condições, criar um novo pool $P = \{C_i\}$ e adicioná-lo ao S.

Passo 3. Se $i <= T$ ir ao passo 2.
Passo 4. Retornar o pool de P_j de S, de maior accuracy.

A vantagem desta arquitetura é que permite gerar diferentes pools com alta diversidade. A forma de adicionar os classificadores, garante que a medida de diversidade do pool é menor o igual que d.

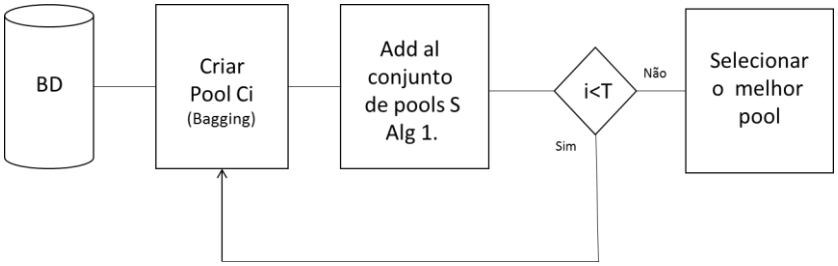


Figura 1. Arquitetura proposta.

QUESTÃO 2

Para elaborar a proposta de uma medida de diversidade utilizando Measure of "difficulty", teve-se em conta a forma que deveria ter um histograma diverso que garante um bom desempenho do pool. Dada a variável aleatória X, que toma valores $\{1,2,\ldots,L\}$. Seja $f(x)$ a função de densidade de probabilidade, que expressa a probabilidade de que exatamente x classificadores acertem. Um pool diverso poderia quedar caracterizado por $f(1) = 1$; lo cual significa que exatamente um classificador do pool acerta em cada caso. Esta interpretação do problema não é consistente com a regra de decisão por Voto Majoritário, na qual como máximo $L/2$ classificadores devem acertar para obter uma classificação correta. Por tanto para este caso um pool diverso ficaria caracterizado por $f(L/2) = 1$. Seguindo este razoamento desenha-se a seguinte medida de diversidade:

$$Dp = \sum_{x=1}^{L} f(x)w(x)$$

onde

$$w(x) = \begin{cases} -1 & 0 \le x \le \dfrac{L}{2} \\ \dfrac{(L-x)*2}{L} & \dfrac{L}{2} < x < L \end{cases}$$

A função Dp está definida em $[-1,1]$, 1 significa a máxima diversidade e -1 a mínima diversidade. A Fig. 2 mostra o comportamento da media frente aos exemplos conhecidos. Como pode-se observar, $w(x)$ pondera as diferentes probabilidades obtidas tendo em conta que um pool diverso deveria ter f(L/2) = 1. Esta medida é comparada com as propostas na literatura, na questão 3.

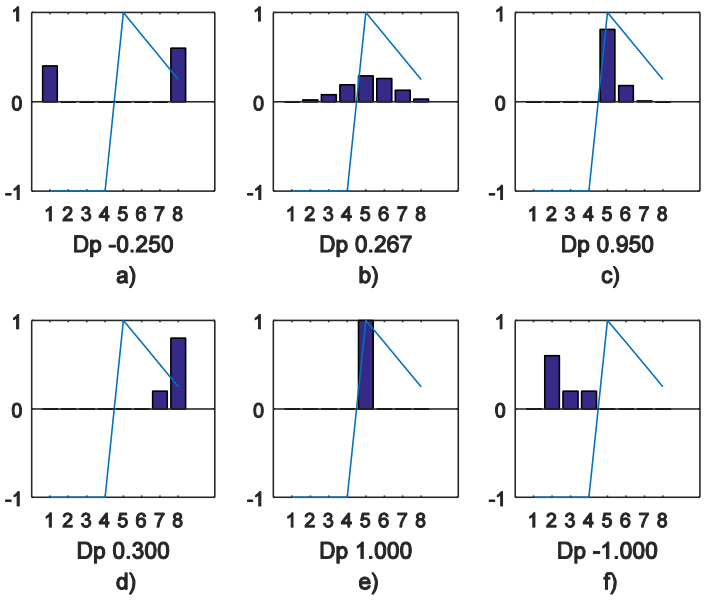


Figura 2. Resultados obtidos com a medida de diversidade proposta.

QUESTÃO 3

Para o analise das mediadas de diversidade usou-se a base de dados *Breast Cancer Wisconsin* (Original) da *UCI Machine Learning Repository* [1]. Para geral os pools aplicou-se *Bagging* com *Árvore de Decisão*. Foram criados 10 pools de $L = 100$ classificadores. A Tabela 1 mostra os valores das medidas de diversidade e accuracy para cada pool. Em amarelo mostram-se as medidas que correspondem com o melhor accuracy y em verde as que correspondem com o pior acurracy. A medida Q-statistic es a que mostra os piores resultados. A medida proposta Dp mostra um bom comportamento com respeito as demais medidas. A Fig. 3 mostra o comportamento das medidas para os 10 pools.

Tabela 1. Resultados de as medidas.

No	Qst↓	ρ ↓	Dis↑	DF↓	E↑	kw↑	k↓	θ ↓	GD↑	CFD↑	Dp↑	Acc
1	0,990	0,791	0,032	0,062	0,045	0,014	0,777	0,058	0,206	0,444	0,090	0,928
2	1,000	0,978	0,003	0,057	0,003	0,001	0,974	0,054	0,024	0,200	0,117	0,943
3	0,991	0,681	0,026	0,023	0,041	0,012	0,617	0,023	0,369	0,648	0,171	0,971
4	0,988	0,733	0,026	0,037	0,035	0,012	0,723	0,036	0,263	0,556	0,152	0,957
5	0,983	0,650	0,028	0,016	0,035	0,012	0,525	0,017	0,460	0,819	0,194	0,986
6	0,972	0,680	0,028	0,029	0,038	0,013	0,659	0,028	0,326	0,556	0,147	0,957
7	0,840	0,559	0,030	0,016	0,044	0,014	0,499	0,017	0,485	0,704	0,199	0,986
8	1,000	0,848	0,015	0,030	0,019	0,007	0,787	0,029	0,205	0,630	0,164	0,971
9	1,000	0,916	0,010	0,043	0,013	0,005	0,890	0,042	0,105	0,356	0,142	0,957
10	0,966	0,756	0,033	0,054	0,041	0,015	0,749	0,050	0,234	0,506	0,113	0,929

Qst: Q-statistic, **ρ** Correlation coefficient, **Dis** Disagreement Measure, **DF** Double-Fault Measure, **E** Entropy, **kw** Kohavi-Wolpert Variance, **k** Measure of Interrater Agreement, **θ** Measure of difficulty, **GD** Generalized Diversity, **CFD** Coincident Failure Diversity, **Dp** Medida proposta.

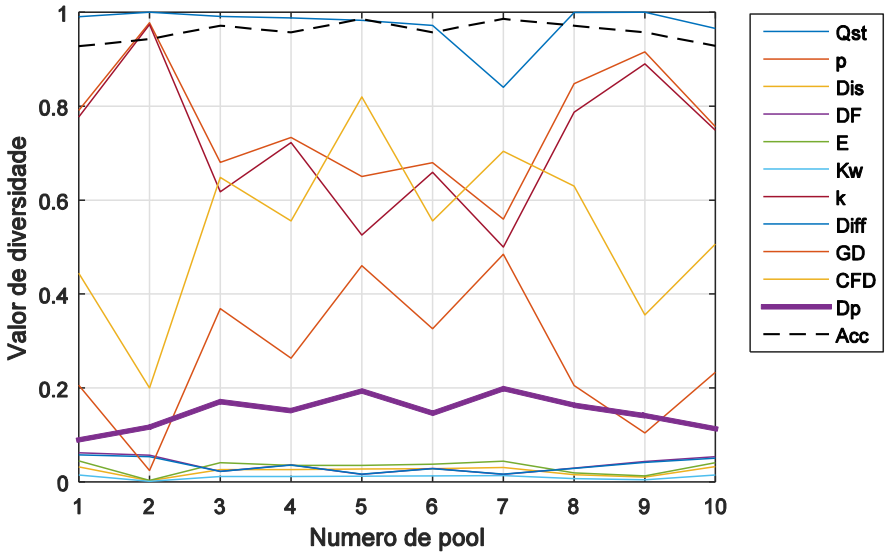
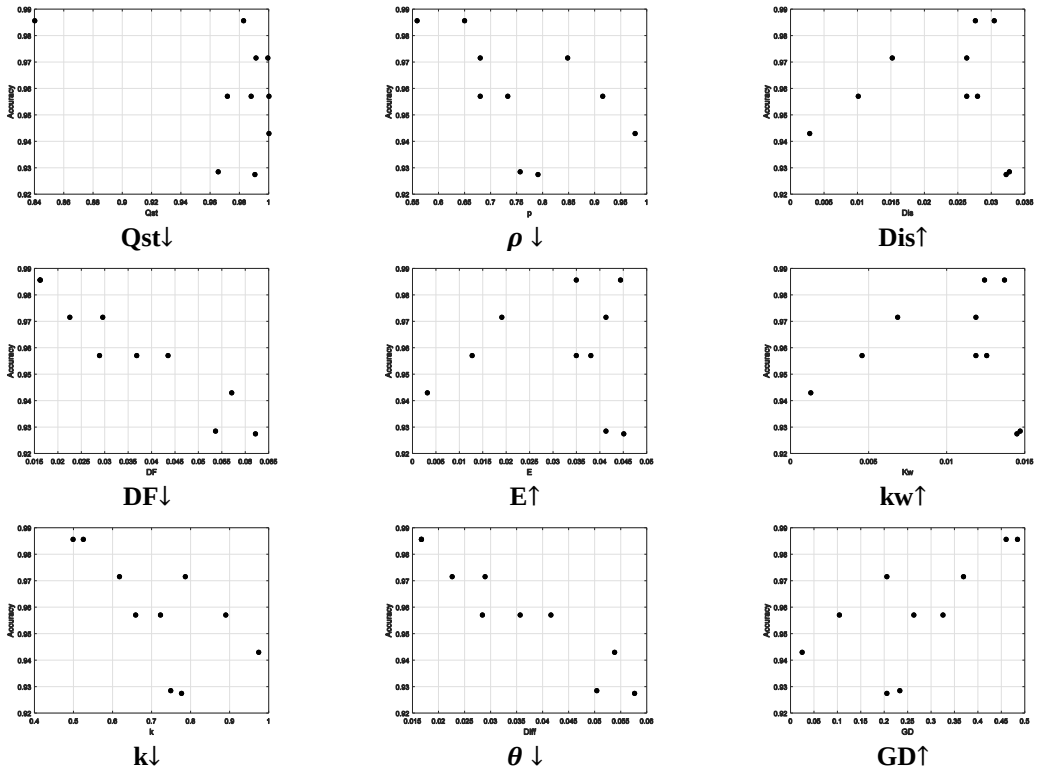
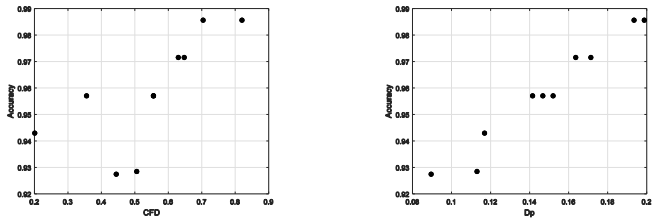


Figura 3. Comportamento de as medidas de diversidade para os 10 pools.

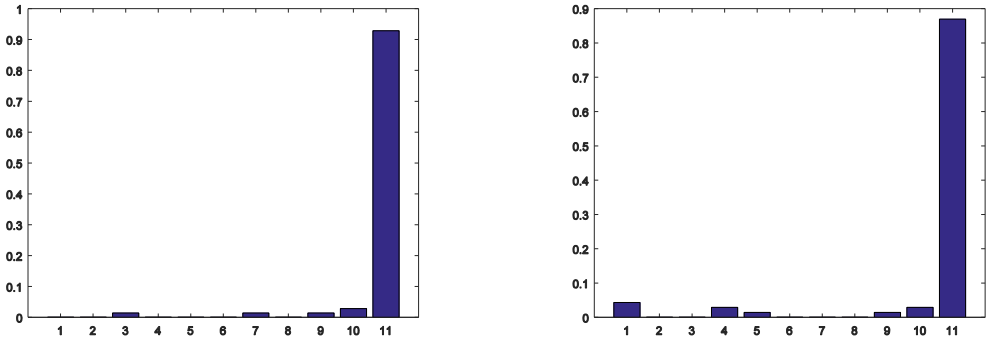
A Fig 4. mostra a relação entre o accuracy e o valor de cada medida. Como pode-se observar a medida Q-statistic presenta um comportamento contrário al esperado para esta base de dados. A Fig 5. mostra o histograma dos pools 1 e 7, correspondentes al pior y al melhor valor de accuracy respectivamente. A medida *Dp*, em ambos casos oferece valores baixos em comparação com GD e CFD para o pools 7. Embora este pool presenta um accuracy elevado, tem diversidade não, dado que na maioria dos acertos todos os classificadores concordam.





CFD↑ **Dp↑**

Figura 5. Relação entre cada medida e accuracy.



a) Histograma del pool 7
GD =0,48; **CFD** =0,70; **Dp** =0,2.
Acc=0.99.

b) Histograma del pool 1
GD=0,21; **CFD**=0,44; **Dp**=0,09.
Acc = 0.92.

Figura 6. Histogramas dos pools 1 y 7. Os histogramas oferecem o pior e o maior resultado respectivamente.

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.