Universidade Federal de Pernambuco

Sistemas de Múltiplos Classificadores

Relatório - Lista 2

Hartur Barreto Brito

Sumário

1 Introdução

Para implementação dos algoritmos de mensuração de diversidade, foi utilizada a linguagem *Python*, assim como no algoritmo de geração de *ensembles* de classificadores, aplicando o *Bagging*, que foi reaproveitado da solução desenvolvida durante a resolução da *Lista 1*.

O classificador utilizado no *ensemble* foi o SGD devido a sua simplicidade e velocidade de execução. Tendo em vista que o intuito dessa lista é o de observar o comportamento de várias medidas de diversidade, não foi feita uma análise aprofundada a respeito do classificador utilizado.

1.1 Análise da base utilizada

1.1.1 Forest type mapping Data Set

Essa base pode ser encontrada na base de dados UCI [1], mais especificamente em [3]. Originalmente, as classes dos dados são: 'd' (Mixed deciduous forest), 'h' (Hinoki forest), 'o' (Other non-forest land) e 's' (Sugi forest) mas, para possibilitar a utilização do *Sci-kit Learn*, elas foram convertidos em números, conforme mostrado na tabela 1. A tabela 1 também mostra como estão distribuídos as 523 amostras dessa base.

Código Original	Código Utilizado	Classe	Amostras
d	1	Mixed deciduous forest	159
h	2	Hinoki forest	86
o	3	Other non-forest land	83
s	4	Sugi forest	195

Tabela 1: Distribuição de amostras de cada classe da base *Forest type mapping*

1.2 Metodologia

Para obtenção de resultados foi seguida a mesma metodologia da primeira lista, que consiste em executar o algoritmo 10 vezes e utilizar a média dos resultados como resultado final. Os passos executados nessa lista são:

1. Mistura dos dados lidos da base;
2. Separação de dados de treino e teste;
3. Geração do *ensemble*;
4. Cálculo da diversidade entre os classificadores do *ensemble*;
5. Teste do *ensemble* para obter taxa de acerto.

2 Questão 1

Para o desenvolvimento desse experimento foram implementadas as medidas de diversidade pareada *Double Fault Measure* e *Disagreement Measure*, e as medidas de diversidade não pareada *Entropy Measure* e *Coincident Failure*.

A *Double Fault Measure* é obtida a partir da razão entre a quantidade de vezes em que os dois classificadores erram pela soma das quantidades de vezes que os dois erram, apenas 1 deles erra e os dois acertam, sendo quanto menor o resultado, mais diverso os classificadores são entre si.

A *Disagreement Measure* é obtida a partir da razão entre a soma da quantidade de vezes em que apenas 1 dos classificadores erram pela soma das quantidades de vezes que os dois erram, apenas 1 deles erra e os dois acertam, sendo quanto maior o resultado, mais diverso os classificadores são entre si.

A *Entropy Measure* busca definir a diversidade quando os classificadores acertam em locais diferentes, e é definido pela equação 1, sendo N a quantidade de amostras e L a quantidade de classificadores. Para essa medida, quanto maior o valor, maior é a diversidade.

$$E = \frac{2}{NL} \sum_{j=1}^{N} min\{(\sum_{i=1}^{L} y_{j,i}), (L - \sum_{i=1}^{L} y_{j,i})\} \tag{1}$$

A *Coincident Failure Measure* busca definir a diversidade a partir da quantidade de vezes em que os classificadores acertam e erram juntos. Ela é definida pela equação 2, sendo L a quantidade de classificadores e p_i a probabilidade de i classificadores acertarem alguma amostra. Para essa medida, quanto maior o valor, maior é a diversidade.

$$\begin{cases} 0 \ se \ p_0 = 1 \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i \ se \ p_0 < 1 \end{cases} \tag{2}$$

Para comparação do comportamento das medidas de acordo com a taxa de acerto, foi traçado o gráfico da figura 3 seguindo o pseudocódigo descrito pelo algoritmo 1. Ele faz com que todos os gráficos sejam deslocados para terem o mesmo ponto máximo, a fim de analisar apenas seu comportamento.

Os gráficos das figuras 1, 2 e 3 mostram que as medidas de diversidade implementadas não são suficientes para definir se taxa de acerto de um *pool* é maior ou não do que a de outro, com exceção apenas do *Double Fault Measure*, que apresentou um bom comportamento em relação à taxa de acerto, diminuindo quando a taxa de acerto era mais alta e aumentado quando a taxa de acerto é mais baixa. No gráfico da figura 3 pode ser verificado que o *pool* com maior precisão (80 classificadores) obteve um dos piores resultados na maioria das medidas de diversidade implementadas, com exceção apenas do *Double Fault Measure*.

Algorithm 1 Deslocamento de gráficos

$precision \leftarrow$ *Vetor de taxa de acerto média dos pools*
$DFM \leftarrow$ *Vetor de medidas aplicando Double Fault Measure*
$DM \leftarrow$ *Vetor de medidas aplicando Disagreement Measure*
$EM \leftarrow$ *Vetor de medidas aplicando Entropy Measure*
$CFD \leftarrow$ *Vetor de medidas aplicando Coincident Failure Diversity*
$p \leftarrow max(precision)$
$df \leftarrow max(DFM)$
$d \leftarrow max(DM)$
$e \leftarrow max(EM)$
$c \leftarrow max(CFD)$
$deslocamento \leftarrow max(p, df, d, e, c)$
$precision \leftarrow \forall$ *x in precision x do:* $+ deslocamento - p$
$DFM \leftarrow \forall$ *x in DFM do:* $x + deslocamento - df$
$DM \leftarrow \forall$ *x in DM do:* $x + deslocamento - d$
$EM \leftarrow \forall$ *x in EM do:* $x + deslocamento - e$
$CFD \leftarrow \forall$ *x in CFD do:* $x + deslocamento - c$

2.1 Resultados

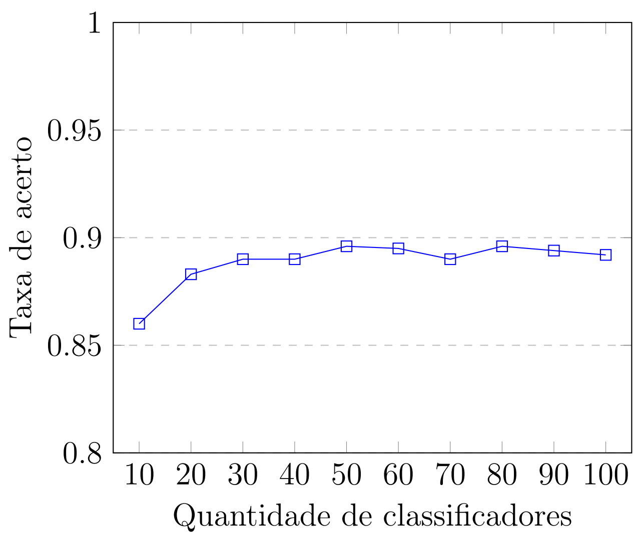

Figura 1: Taxa de acerto de acordo com a quantidade de classificadores no *Bagging*

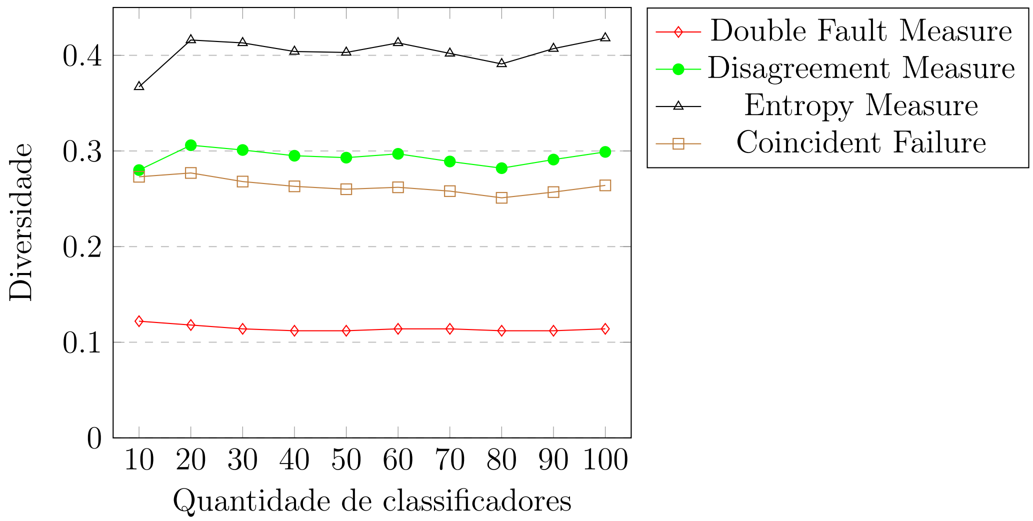

Figura 2: Medidas de diversidade de acordo com a quantidade de classificadores no *Bagging*

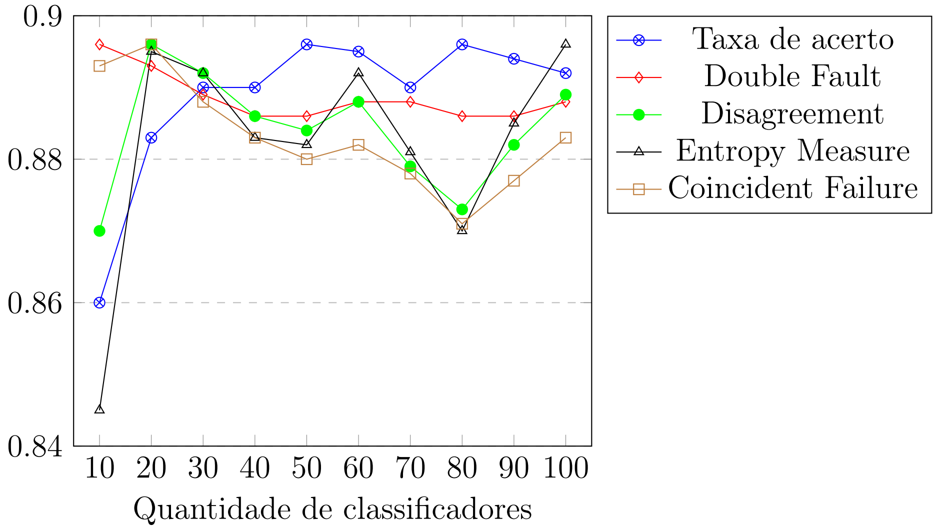

Figura 3: Comparação entre medidas de diversidade e taxa de acerto do pool de acordo com a quantidade de classificadores no *Bagging*.

3 Questão 2

O algoritmo proposto (algoritmo 3) busca selecionar uma mesma quantidade de classificadores especialistas para cada tipo de classe. Por tender a ter uma mesma quantidade de especialistas para cada classe, torna-se necessário a adição de classificadores medianos (com boa taxa de acerto média) para que sirvam de critério de desempate caso os classificadores especialistas entrem em conflito ao ser executado o voto majoritário.

Antes de adicionar um classificador no *pool*, é verificado também se ele possui ao menos 0.5 pontos de diversidade com relação aos classificadores já adicionados utilizando a função abstrata descrita pela questão ($D_{\alpha,\beta}$). O intuito dessa seleção é o de evitar que sejam adicionados classificadores semelhantes no *pool*, prejudicando seu desempenho, tendo em vista que, teoricamente, quanto mais diverso for o *pool*, melhor seu desempenho. Essa verificação é feita pelo algoritmo 2. A função $D_{\alpha,\beta}$ é referenciada na chamada da função $getDiversity(p, c)$ no pseudocódigo.

Algorithm 2 IsDiverso

Require: *P, c*
  $diverso \leftarrow True$
  for *p in P* **do**
    if *getDiversity(p, c) < 0.5* **then**
      $diverso \leftarrow False$
      $Break$
    end if
  end for
  return $diverso$

Algorithm 3 SelectClassifiers

```
C ← Lista de classificadores elegíveis
X ← Matriz de features das amostras
y ← Vetor de respostas utilizando as features de X
P ← Vetor de classificadores selecionados
E ← Matriz de classificadores especialistas em cada classe
Q ← Vetor de quantidade de especialistas para cada classe
for c in C do
   vetorPrecisao ← getPrecisionScore(c, X, y)
   precisaoMedia ← sum(vetorPrecisao) / len(vetorPrecisao)
   if precisaoMedia > 0.5 then
      if IsDiverso(P,c) then
         P.append(c)
      end if
   else if existe algum p em vetorPrecisao que p > 0.8 then
      for indice in range(0, len(vetorPrecisao)) do
         if vetorPrecisao[indice] > 0.8 then
            E[indice].append(c)
            Q[indice] ← Q[indice] + 1
         end if
      end for
   end if
end for
especialistas ← min(Q)
for e in E do
   q = especialistas
   while q > 0 and len(e) > 0 do
      c ← e.pop()
      if IsDiverso(P,c) then
         P.append(c)
         q ← q − 1
      end if
   end while
end for
P ← P.unique()
```

4 Questão 3

A técnica de medida de diversidade *The Measure Of Difficulty* θ se baseia na variância do histograma da porcentagem de acerto de acordo com a quantidade de classificadores que acertaram. Alguns exemplos de histogramas, podem ser visualizados nos gráficos da figura 4, retirada de [4].

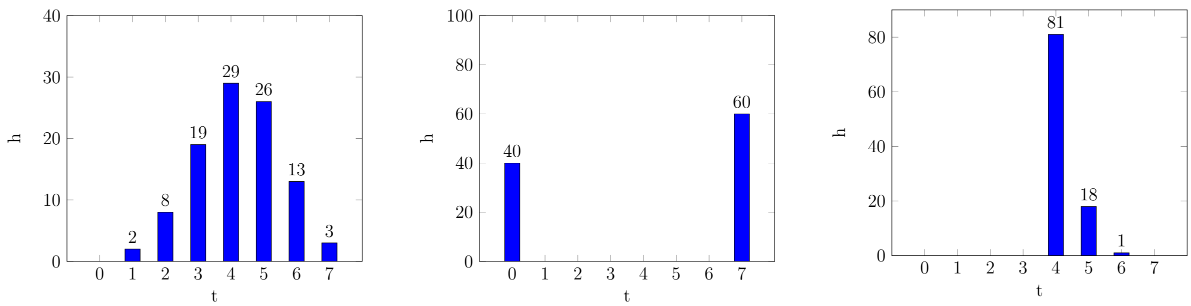

Figura 4: Padrões de "*difficulty*" para três *ensemble* de classificadoras com $L = 7$, $p = 0.6$ e $N = 100$. O eixo x é "proporção correta", isto é, i/L.

Ao analisar os gráficos, pode ser observado que:

- O gráfico mais à esquerda da figura 4 possui uma variância baixa (0.034), e isso ocorre pois uma quantidade diferente de classificadores acertam em locais diferentes, o que caracteriza o *pool* como diverso.

- No gráfico central da figura 4, onde em 40% dos exemplos nenhum classificador acerta e todos acertam 60% das amostras, é obtido uma grande variância (0.240). Isso indica um baixo índice de diversidade, de acordo com *The Measure Of Difficulty* θ.

- Já no gráfico mais à direita da figura 4, 4 classificadores acertam 81% das amostras, enquanto 5 acertam 18% e 6 1%, o que daria uma baixa variância (0.004) e, como consequência, um alto índice de diversidade.

Essa técnica apresenta problemas quando a variância é pequena, mas o gráfico se concentra próximo à região onde nenhum classificador acerta as amostras, como exemplificado no gráfico da figura 4.

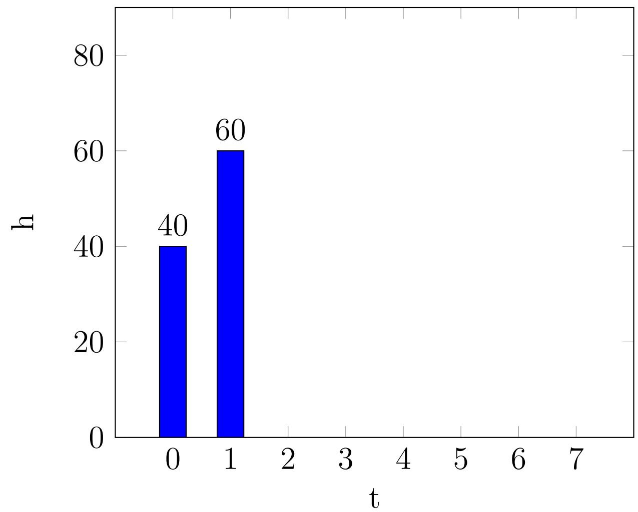

Figura 5: Histograma com baixa variância e baixa média.

Para solucionar o problema descrito acima, seria necessário a adição de informações no cálculo da medida de diversidade, para representar a região na qual o gráfico está localizado.

Uma possível solução, seria somar os resultados obtidos entre 0 e $\frac{L}{2}$, sendo L a quantidade total de classificadores no *pool*, para que seja analisada a concentração dos resultados onde poucos classificadores acertam as amostras e dividir por 100, para que o valor obtido seja entre 0 e 1. Como está sendo calculada a concentração de valores onde poucos classificadores acertam, quanto menor, melhor será, seguindo a mesma linha de raciocínio do cálculo da variância.

Essa alteração iria modificar o intervalo de resultados, ficando entre 0 e 2, e melhoraria o relacionamento entre a medida e a taxa de acerto. A equação 3 demonstra como seria feita a nova medida.

$$M = Var(X) + \frac{\sum_{i=0}^{\frac{L}{2}} h(i)}{100} \tag{3}$$

5 Referências

1. https://archive.ics.uci.edu/ml/datasets.html
2. http://scikit-learn.org/stable/index.html
3. Forest type mapping Data Set - https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping
4. Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Mach. Learn. 51, 2 (May 2003), 181-207.