Matriz de Dissemelhança Entrópica para Classificação Não Supervisionada

Palavras Chave: *Clustering*, Entropia, Matriz de dissemelhança.

Resumo: A classificação hierárquica é um processo de classificação usualmente baseado em medidas de semelhança ou dissemelhança entre objectos ou conjuntos de objectos de um determinado conjunto de dados. A medida de dissemelhança mais comum é a métrica pesada l_p (a distância Euclidiana é um caso particular da métrica não pesada l_p) que serve de suporte para a construção de matrizes de dissemelhança, elemento base dos algoritmos de classificação hierárquica. Estas medidas de dissemelhança não fornecem nenhuma informação sobre a estrutura dos dados, a forma dos grupos (*clusters*), facto pelo qual a grande maioria dos algoritmos de classificação não supervisionada (*clustering*) produz *clusters* globulares (o *K-means* é um bom exemplo deste tipo de algoritmos). A medida de dissemelhança "ideal" para um algoritmo de *clustering* deveria providenciar informação sobre a estrutura dos dados de forma a facilitar a obtenção de soluções óptimas. Neste trabalho mostramos como podemos obter uma medida de dissemelhança com as características referidas usando uma medida de dissemelhança entrópica.

Neste trabalho usamos a entropia quadrática de Renyi, $H_{R2}(X)$, pois é, entre todas as medidas entrópicas, a de menor complexidade computacional. A estimação da entropia quadrática de Renyi é efectuada combinando o método de Parzen para estimar a função densidade de probabilidade com a aplicação de um "kernel" Gaussiano, $G(x, 0, I)$. A estimativa da entropia quadrática de Renyi de uma variável X é obtida através da fórmula $\widehat{H}_{R2}(X) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G\left(x_i - x_j, 0{,}2h^2 I\right)\right)$, sendo $x_i,\ i = 1{,}2, \ldots, N$, as concretizações da variável X, e h um parâmetro de suavização.

Vamos ver como usar a entropia para obter uma nova matriz de dissemelhança. Seja um conjunto de vectores $X = \{x_1, x_2, \ldots, x_N\}, x_i \in \mathbb{R}^m$, correspondentes a uma série de elementos de um dado conjunto de dados. Cada elemento da matriz de dissemelhança $A,\ A \in \mathbb{R}^{N\times N}$, é calculado usando uma medida de dissemelhança usual, $A_{i,j} = d(x_i, x_j)$. Se construirmos um sub-grafo unindo cada elemento com o elemento mais próximo de acordo com a medida de dissemelhança usada este sub-grafo não contém nenhuma informação sobre a estrutura local dos dados. Vamos, de seguida, aplicar uma medida entrópica para conseguir a informação pretendida. Consideremos, para cada elemento P, os seus M vizinhos mais próximos de acordo com a medida de dissemelhança inicialmente usada. Tomemos esses M+1 elementos e computemos os vectores diferença (ligações) entre cada par de elementos. Calculemos, de seguida, os valores da entropia $E_k,\ k = 1{,}2, .., M$, sendo que cada valor E_k é obtido calculando a entropia do conjunto reunião de todos os vectores diferença entre os M vizinhos mais próximos de P com o vector diferença entre P e o elemento k. Estes M valores vão-nos permitir fazer o ranking das ligações de acordo com um critério entrópico: vamos escolher para ligação principal (mais forte) aquela que, num conjunto de ligações possíveis entre P e os seus vizinhos mais próximos, introduz menor complexidade no sistema, menor desordem, por conseguinte, menor entropia. Ao usar este princípio as ligações entre elementos passam a reflectir a estrutura local dos dados. A matriz de dissemelhança construída com este novo critério, ao contrário da matriz de dissemelhança inicial, vai conter em cada linha (ou coluna) apenas M valores não nulos.
Esta matriz de dissemelhança entrópica pode ser usada como base para um algoritmo de clustering. Com esta nova matriz as ligações entre objectos são sensíveis à estrutura local dos dados o que permite (sem conhecimento *à priori*) a obtenção de *clusters* que reflectem essa mesma estrutura.

Foram realizadas um grande número de experiências sobre dados artificiais e reais que confirmam a validade desta nova medida.