Undersampling Representativo de Classe Dominante por Fator de Qualidade Baseado em Multiplicadores de Lagrange

Paulo Cirino

October 8, 2017

Abstract

1 Introdução

Esse trabalho é fundamento em um algorítimo de *fuzzy clustering*, à ser publicado, que foi criado para acelerar o método *Fuzzy C Means*. No trabalho feito, se atinge o objetivo por meio da remoção de pontos da bordas com o auxilio de um fator de qualidade, que diz respeito a pertinência de cada amostra à todos os *clusters*.

Os algorítimo de *fuzzy clustering* permitem que uma amostra de um *data set* pertença, ao mesmo tempo, à múltiplos agrupamentos. O nível que uma amostra pertence a cada *cluster* é tradicionalmente chamado de **pertinência** $\mu_i(x_j)$, que é a pertinência da amostra x_j para o *cluster* i.

A função de custo J, associada à problemas de *fuzzy clustering*, pode ser definida em 1.

$$\begin{aligned} &\min && J \\ &\text{sujeito a} && \sum_{k=1}^{c} u_{ik} = 1, \quad k = 1, 2, ..., N \end{aligned} \tag{1}$$

Onde J é definido em 2.

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^2 d_{ik}^2 \tag{2}$$

Nessa situação μ_{ik}, é a pertinência da amostra k em relação ao centro i. Adotando a solução de Multiplicadores de Lagrange, a nova função de custo assume a forma descrita em 3, com derivadas parciais 4 e 5.

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} \left[u_{ik}^2 d_{ik}^2 - \lambda (\sum_{m=1}^{c} u_{mk} - 1) \right] \tag{3}$$

$$\frac{\partial J}{\partial \lambda} = \sum_{m=1}^{c} u_{mk} - 1 : \frac{\partial J}{\partial \lambda} = 0 \implies \sum_{i=1}^{c} u_{ik} = 1 \tag{4}$$

$$\frac{\partial J}{\partial u_{st}} = 2u_{st}d_{st}^2 - \lambda : \frac{\partial J}{\partial u_{st}} = 0 \implies u_{st} = \frac{\lambda}{2d_{st}^2} \tag{5}$$

Assim, a equação 6, representa cada um dos multiplicadores de Lagrange do *data set*.

$$\lambda_k = \frac{2}{\sum_{j=1}^{c} \frac{1}{d_{jk}^2}}, \quad k = 1, 2..., N \tag{6}$$

Assim é possível definir uma medida de qualidade para cada amostra, descrita na equação, 7. A medida q_k de qualidade, é obtida para cada amostra $\mathbf{x}_k$ de $\mathbf{X} = \{x_i \in \mathbb{R} | i = 1...N\}$, e representa uma medida de incerteza da pertinência μ_{ik}.

$$q_k = c^c \prod_{i=1}^{c} \frac{1}{\mu_{ik}} \tag{7}$$

Substituindo a equação 5 em 7, podemos representar q_k em 8

$$q_k = \frac{2}{\lambda_k} c^c \prod_{i=1}^{c} d_{ik}^2 \tag{8}$$

É importante notar que amostras fortemente ligadas a um determinado centro, terão q_k muito próximo à 0, aquelas que estão distantes terão valores tendendo à ∞.

Uma forma alternativa de enxergar o índice de qualidade é substituindo a equação 6 em 8.

$$q_k = c^c \frac{\prod_{i=1}^{c} d_{ik}^2}{\sum_{i=1}^{c} d_{ik}^2} \tag{9}$$

Uma forma de visualizar essa qualidade é por meio do gráfico abaixo:

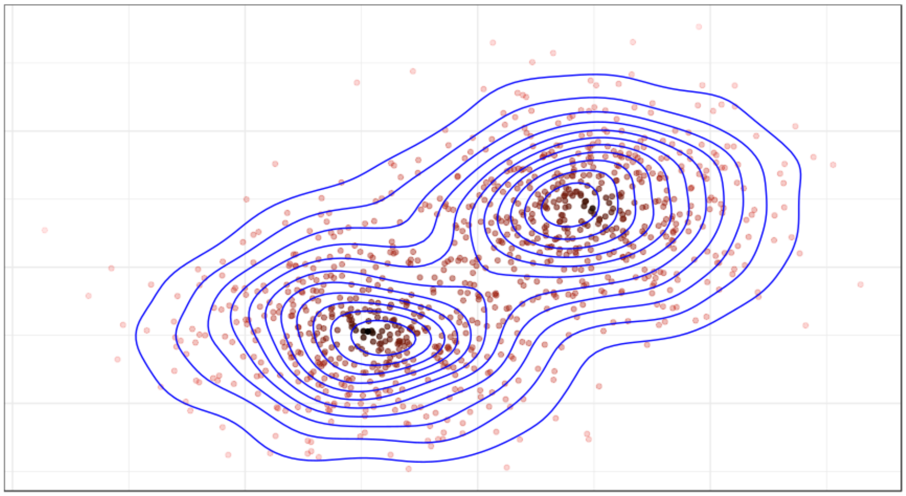

Figure 1: Gráfico de 2 normais, onde cor e as curvas de nível representam a qualidade

Visto que, é possível calcular um índice de qualidade para cada amostra que diz respeito à quanto um ponto faz parte ou não de um agrupamento. É possível utilizar a informação da qualidade para fazer seleção de amostras pertinentes.

Em um senário de classificação desbalanceada, é possível então, fazer uma subamostragem da classe dominante, no sentido de balancear o problema.

2 Método

Foram propostos um total de 5 abordagens de amostragem utilizando essa qualidade. Para esse método, foram considerados apenas problemas binários desbalanceados.

O que é comum em todas as abordagens é que foram selecionadas $N_{minority}$ das $N_{majority}$ amostras da classe majoritária, tal que $N_{minority} < N_{majority}$.

2.1 briSelection

A primeira abordagem proposta foi a de *briSelection*, onde são selecionadas as $N_{minority}$ amostras da classe majoritária com maior fator de qualidade.

Essa abordagem têm o efeito de selecionar os pontos marginais da classe dominante. Isso pode ser positivo no sentido de amostras apenas os pontos que definem o contorno da superfície de decisão, porém, pode ser muito negativo em situações onde os dados possuem muitos *outliers*.

2.2 briSelection++

s abordagem *briSelection++* é inspiradas na metodologia de inicialização do método **FCM++**.

Isso é feito de forma que quanto maior a qualidade de uma amostra da classe dominantes, maior a probabilidade de essa amostra ser selecionada, assim como mostra a equação 10.

$$P(x_k \mid q_k) = N_{minority} \frac{q_k}{\sum_{k=1}^{N_{majority}} q_k} \quad k = 1, 2, ..., N_{majority} \tag{10}$$

Essa metodologia possui a vantagem de amostrar a distribuição da classe dominante como um todo, nas regiões centrais por conta da alta densidade de pontos e nas periferias por causa do alto índice de qualidade. Isso possui uma vantagem de representar todos os dados e não só as bordas.

2.3 briSelection−−

A metodologia *briSelection−−* é exatamente igual ao $briSelection++$, só que q_k é substituído por $\frac{1}{q_k}$, como mostra a formula 11.

$$P(x_k \mid q_k) = N_{minority} \frac{\frac{1}{q_k}}{\sum_{k=1}^{N_{majority}} \frac{1}{q_k}} \quad k = 1, 2, ..., N_{majority} \tag{11}$$

Essa alteração na formulação permite que o centro da distribuição da classe dominante seja muito bem representado, mas ainda criando uma probabilidade de os pontos da margem também serem amostrados.

2.4 briSelectionLog++ e briSelectionLog−−

3 Testes e Resultados

4 Conclusão