4
Conclusão

Nesta dissertação foi apresentado um método para seleção de dados em LVQ, inspirado em um resultado geral proposto em [15]. Utilizou-se uma técnica baseada em aprendizado exaustivo para calcular correlação entre o erro de cada padrão e o erro de toda a sua distribuição. Essa correlação mostrou ser uma boa indicadora de quais dados devem ou não ser mantidos no conjunto que irá ser utilizado para treinar o modelo.

Dois pontos foram importantes na adaptação da técnica ao LVQ. As regiões de sorteio de protótipos e a heurística para efetuar o corte no vetor de correlação, estipulando, de fato, a seleção de dados. Ambos influenciam fortemente no resultado final. É importante ressaltar que a política de corte no vetor de correlação foi conservadora, procurando eliminar o máximo de ruído possível, mas sem perder pontos que poderiam ser relevantes. Ao longo do processo de ajuste do algoritmo sempre houve uma tentativa de poupar esses pontos relevantes, principalmente aqueles pertencentes à zona de risco, para que um corte mais radical não os impedisse de participar do treinamento do modelo. Evidentemente, para cada problema pode ser feita uma análise diferente, mas o que é claro a partir do método, é que ele fornece claramente a dificuldade de classificação de todos os pontos das distribuições. A divisão efetiva para estabelecer quando a dificuldade de classificação de um ponto reflete que se trata de um ruído ou de um ponto relevante pode ser considerado o ponto central dessa heurística. No exemplo com o banco Íris, mostrou-se que, ao eliminar mais dados, a melhora ao classificar usando leave-one-out é muito grande. Entretanto, não há garantias de que os pontos eliminados são ruídos, e, é muito provável que existam relevantes eliminados também. Uma investigação sobre como tratar estes pontos se seguirá a esta dissertação.

A técnica proposta mostrou resultados muito interessantes para os experimentos realizados. Em particular, em casos muito ruidosos, a melhora é evidente, pois, atualizar os protótipos com todos os dados prejudica claramente a classificação.

Fechamos a dissertação deixando como propostas de trabalhos futuros as seguintes idéias:

Experimentos com múltiplos protótipos podem melhorar as classificações, como foi mencionado no experimento com o banco Íris. Além disso, regiões não convexas devem ser utilizadas para observar o comportamento do método nas zonas de risco e fronteiras de decisão.

Uma proposta geral para o corte no vetor de correlação ρ seria interessante, mas não é possível afirmar, nesse momento, se a análise deve ser feita caso a caso, ou se é possível estabelecer um critério para resolver este problema.

Finalmente, uma questão que deve ser abordada é a exaustividade intrínseca desse tipo de abordagem.