1 Introdução

A utilização de técnicas baseadas em aprendizado estatístico vem sendo cada vez mais explorada em diversas áreas do conhecimento [48-50], [14], [15], [43], [44], [45], [18], [30], [31], [52]. O livro texto [32] é uma referência atualizada e abrangente no assunto.

Considere-se um ambiente constituído de (i) uma medida de desfecho, que pode ser quantitativa, e.g. o valor da demanda de carga elétrica em 24 horas, ou categórica, e.g. recaída (ou não) de uma determinada doença; e de (ii) um conjunto de atributos ou de variáveis independentes, que são características de alguma forma relacionadas ao desfecho (e.g. temperatura medida a cada meia hora ou resultados de exames de pacientes). Além disso, dispõem-se de um conjunto de observações – medidas dos atributos – associadas a um determinado desfecho. Estas observações e o desfecho associado são utilizados na construção de um modelo (fase dentro-da-amostra), que, em seguida, pode ser empregado para estimar medidas de desfechos desconhecidos associados a novas observações que não tenham sido utilizadas no ajuste nem na construção deste modelo (fase fora-da-amostra). O que se deseja é construir modelos que sejam eficientes no sentido de ter bom desempenho na fase fora-da-amostra. Este contexto, na área de inteligência computacional, é conhecido como Aprendizado Supervisionado. O conceito supervisão decorre da presença de medidas de desfecho conhecidas associadas aos vetores de atributos, gerando assim a possibilidade de se usar iterativamente uma determinada medida de erro no ajuste do modelo. Em problemas de Aprendizado Não-Supervisionado, o conjunto de observações não está associado a uma medida de desfecho e a tarefa neste caso é descobrir, a partir de alguma forma de similaridade, como as observações podem estar agrupadas.

A proposta desta tese é desenvolver e testar alguns novos algoritmos para ambientes supervisionados, embora técnicas não-supervisionadas sejam, por vezes, utilizadas como passos intermediários. O foco do trabalho está direcionado a problemas de Classificação de Padrões, onde as medidas de desfecho são

categóricas e a partir daqui serão chamadas de rótulos. Cada uma das classes (ou categorias) é representada por um rótulo diferente. Nesta tese nos restringiremos a problemas dicotômicos, i.e. com duas classes. A notação utilizada nesta tese encontra-se no Apêndice A.

Se as probabilidades à priori de cada uma das duas classes, além das respectivas densidades condicionais forem conhecidas, tem-se que a fronteira de decisão das classes C_1 e C_2 é a região do espaço definida por $p(x | C_1) = p(x | C_2)$ e é possível calcular a probabilidade posteriori de cada classe pela regra de Bayes [7]. A regra de classificação pode então ser estabelecida como: $x \in C_1$ se $P(C_1 | x) \geq P(C_2 | x)$ e $x \in C_2$ caso contrário. Utilizar o classificador Bayesiano significa atribuir a observação à classe estatisticamente mais provável [48], o que torna o classificador ótimo, no sentido que nenhum outro classificador possui uma taxa de acerto maior que o classificador Bayesiano. Entretanto, na maioria dos problemas de interesse, as funções de densidade de probabilidade (fdp's) envolvidas no cálculo da regra de Bayes não são conhecidas, tornando necessário o cálculo de estimativas. Estimativas não paramétricas de fdp's possuem conhecidas complicações numéricas, especialmente em espaços maiores que o $\mathfrak{R}^3$. Assim, há uma clara motivação para o desenvolvimento de classificadores sub-ótimos.

Uma forma de agrupar e diferenciar os métodos de classificação é analisar a forma de construção da função discriminante de classes, f . Um método é dito global quando f se constitui em uma única função – global – que é capaz de realizar a separação das classes. Outra possibilidade é construir f a partir de uma composição de funções locais, e neste caso diz-se que o método é local. Uma terceira possibilidade, que será explorada nesta tese, é a de um classificador local-global. Chamaremos um classificador de local-global quando os enfoques local e global estão conjugados na construção do classificador. Pode-se lançar mão de técnicas globais (aplicadas igualmente a todas as observações), e em seguida construir classificadores locais que são aplicados a apenas um pequeno subconjunto da amostra, para ao final gerar um esquema de classificação para todas as observações.

Uma maneira clássica de particionar o espaço é através de Quantização Vetorial (QV) [1], [16], [17], [4]. QV é um procedimento não supervisionado que tem como objetivo a construção de aproximações quantizadas da distribuição da

amostra, através dos chamados vetores-protótipo (*codebooks*). LVQ é uma versão de QV onde os vetores-protótipo são posicionados no espaço através de um processo supervisionado. Uma vez posicionados os protótipos, as observações são associadas a estes gerando o particionamento desejado. Em geral, esta associação se dá através da regra do vizinho mais próximo [42], [32]. Várias extensões relacionadas ao algoritmo original de LVQ vêm sendo publicadas na literatura, e.g. [61], onde um procedimento de transição entre medida de distância harmônica média e mínima distância é aplicado para que o algoritmo seja insensível a diferentes inicializações. Alguns algoritmos clássicos como, por exemplo, o K-means [40], os mapas auto-organizáveis de Kohonen [5], [20], e o LBG (Linde-Buzo-Gray) [1], que será usado nesta tese, se encontram dentro do contexto de QV.

Uma outra proposta desta tese está direcionada à seleção de um subconjunto da amostra. Existem, na literatura, alguns enfoques envolvendo seleção de observações, por exemplo, *Active Learning* [10], [35], [36], [11]; *Query-based learning* [37]; *Sequential design* [38] e os algoritmos DROP3 [57] e IB3 [56]. Estes dois últimos selecionam observações e realizam a classificação fora-da-amostra baseada em vizinhos mais próximos, utilizando apenas as observações selecionadas. O consagrado SVM [33], [34], [39], seleciona observações da amostra próximas à fronteira de decisão, os chamados vetores de suporte. Outras abordagens podem ser encontradas em [55].

Nesta tese, busca-se a seleção de um subconjunto das observações que sejam vizinhas às fronteiras de decisão. Uma vez realizada uma seleção das observações próximas à fronteira, algoritmos diversos de classificação podem ser aplicados exclusivamente a estas observações podendo gerar melhoras no desempenho ou desempenhos equivalentes com uma quantidade consideravelmente menor de observações. Propõe-se nesta tese um esquema para selecionar observações, de maneira independente de modelo. Nesta contribuição utiliza-se a distância entre as funções de densidade de probabilidade das classes, e selecionam-se as observações que estejam próximas à outra classe. Para isso, utiliza-se a divergência de *Cauchy-Schwartz* [3], [8], [12], uma medida da distância entre duas funções de densidade de probabilidade. Esta ferramenta está dentro do contexto de *Information Theoretic Learning* (ITL) [6], [8], [3], [19]. ITL recentemente introduzida em [6], é um procedimento para extrair medidas de

informação diretamente da amostra. Conceitos relacionados a ITL vêm sendo utilizados em diversos campos, como por exemplo: minimização de erro em treinamento supervisionado [29], clusterização [28] e quantização vetorial [3].

Dois dos métodos propostos nesta tese, os que chamamos de Classificador por Quantização Vetorial (CQV) e Zona de Risco Generalizada (ZRG) são procedimentos que executam a classificação em nível local em regiões próximas à fronteira. Em geral, a probabilidade posteriori é mais difícil de ser estimada para observações vizinhas às fronteiras de decisão, já que nesta região as observações possuem probabilidades condicionais as classes aproximadamente iguais. Entretanto, é justamente nesta região que a probabilidade posteriori deve ser bem estimada, e neste sentido, métodos com características locais podem ser relevantes.

De um outro ponto de vista, selecionar observações não representativas, como, por exemplo, *outliers* ou observações que possuam evidência de inversão de rótulo por mecanismo de ruído também pode ser um recurso interessante em associação com algoritmos de classificação. Um método para seleção de observações não representativas dentro do contexto de LVQ foi proposto em [46]. Já em [47], propõe-se uma aplicação que possui o mesmo objetivo e é baseada em quantização vetorial. Um número elevado de observações não representativas pode prejudicar o desempenho. Duas das contribuições desta tese, que originalmente foram desenvolvidas para classificação, possuem extensões aplicadas à identificação de observações não-representativas e a possível remoção destas observações nos procedimentos de classificação.

As principais contribuições desta tese são:

- Proposta de um método que denominamos ‘Classificador por Quantização Vetorial’ (CQV): Este método é baseado em QV com posterior classificação supervisionada local. O objetivo é resolver o problema de classificação estimando as probabilidades posteriores em regiões próximas à fronteira de decisão;
- Proposta de utilização de CQV como ferramenta para Seleção de Observações não Representativas. Esta ferramenta pode ser utilizada em um procedimento de preparação dos dados, onde observações não representativas ou que possuam evidências de terem o rótulo invertido podem ser excluídas da fase dentro-da-amostra;

- Proposta do que denominamos ‘Zona de Risco Generalizada’ (ZRG): Um método independente de modelo, para encontrar as observações vizinhas à fronteira de decisão. Não se trata de um método de classificação, mas sim de seleção de observações representativas que podem ser exclusivamente usadas na fase dentro-da-amostra para o treinamento do modelo;
- Proposta de utilização da ZRG como ferramenta para Seleção de Observações. A idéia central desta contribuição é categorizar as observações pertencentes à amostra, de acordo com sua relevância para a construção do modelo;
- Proposta do método que denominamos ‘Quantizador Vetorial das Fronteiras de Decisão’ (QVFD): Trata-se de um método de classificação que utiliza protótipos, cujo objetivo é construir uma aproximação quantizada das regiões vizinhas à fronteira de decisão.

Para validação das propostas, tem-se a implementação de todos os métodos e ferramentas propostos em diversos bancos de dados controlados e de dados reais publicamente disponíveis, além de discussão comparativa com métodos clássicos, implementados nos mesmos bancos. A tese está organizada da seguinte forma: No capítulo 2 descrevemos as contribuições metodológicas, divididas em diversas seções. Ao final deste capítulo, apresentamos os bancos de dados utilizados no capítulo de resultados. No capítulo 3 detalhamos os resultados dos experimentos realizados, utilizando os bancos descritos no capítulo 2, para cada uma das contribuições metodológicas. Além disso, neste mesmo capítulo, são apresentadas as discussões dos resultados alcançados. Finalmente, o capítulo 4 é dedicado as conclusões.