8
Conclusão e trabalhos futuros

Nesta tese um novo método de regressão em árvores foi proposto. O método de Regressão Construtiva de Variedades Implícitas. A partir de dados esparsos, ele utiliza redução da dimensionalidade para obter informação da estrutura da qual os dados fazem parte. Para isso, ele emprega o método de votação por tensores para obter a dimensionalidade intrínseca dos dados e informações dos vetores tangentes e normais. A confiabilidade das informações dadas pelo método TV foram verificadas. Foram gerados alguns exemplos de variedades, como curva e superfície no $\mathbb{R}^3$, curva, superfície e volume no $\mathbb{R}^4$ e, verificou-se o percentual da estimativa da dimensionalidade correta bem como os erros de orientação dos vetores tangentes/normais obtidos. Tais resultados mostram-se estáveis quanto a escolha da escala σ, mas dependem da escolha do número vizinhos e do número de amostras disponíveis. Os erros na orientação das tangentes/normais, conforme cada caso, foram calculados e para efetuar essa verificação, foi preciso utilizar algumas ferramentas matemáticas, dentre elas, empregamos a teoria de álgebra exterior.

Após obter informações sobre a estrutura a qual os dados pertencem, o método RCVI aproxima uma variedade implícita utilizando partição da unidade, generalizando o método MPU de Othake et al. (49) para estruturas mais complexas. A fim de efetuar a partição da unidade utilizamos malha 2^n-ádica, no entanto, para dados com muitos atributos, tal subdivisão espacial não pode ser empregada e, para estes casos, utilizou-se árvores de partição binária do espaço. Foi realizada uma partição do domínio, controlada pelo erro, e em cada região do domínio, a aproximação da variedade foi efetuada, considerando-a como o conjunto de nível de uma função polinomial multivariada. Essas soluções locais podem ser unidas para se obter uma aproximação global usando funções de transição ou funções peso. Essa técnica permite ajustar um conjunto de dados grande pelo ajuste de pequenos conjuntos de dados, permitindo-se ajustar estruturas complexas e trabalhar em paralelo. Para aproximar localmente a variedade foram utilizados polinômios de graus 1, 2 e 3. Por fim, optou-se por trabalhar com polinômios de grau 1, pois para dados

PUC-Rio - Certificação Digital Nº 0710710/CA

com muitos atributos precisa-se de muitos pontos para ajustar um polinômio de grau maior e além disso, com o método de partição da unidade podemos ficar confiantes em utilizar estruturas mais simples e conseguir um bom ajuste.

Com a estrutura de árvore obtida, em cada folha da árvore uma regressão foi realizada e estimativas locais e globais fornecidas. Os experimentos realizados mostraram que as estimativas globais oferecem melhores resultados. O método RCVI assume que os dados de entrada estão em um espaço de alta dimensionalidade, mas pertencem a um espaço de dimensionalidade baixa, assim, o método procura aproximar a variedade da qual os dados fazem parte e, nessa aproximação efetuada, uma árvore é construída e uma regressão realizada. O método foi aplicado a alguns conjuntos de dados reais e constatou-se que o método teve um desempenho satisfatório, dentro das limitações no número de amostras disponíveis. Para alguns casos, não foi possível obter uma boa representação da variedade, como no conjunto de dados *Boston Housing* e no conjunto de dados *Computer Activity*. Neste último, embora houvesse 8192 amostras disponíveis, o conjunto de dados possuía 21 atributos e apenas 2000 dados eram usados para treino com a finalidade de comparar com outros métodos, que adotaram tal estratégia. Entretanto, ao se aumentar o número de dados disponíveis para treino, o método RCVI melhorou consideravelmente seu desempenho.

O método RCVI também foi aplicado a dados de poços de petróleo e teve seu desempenho comparado com o tradicional método de regressão em árvores, o CART, foi comparado com o SVR, método que mostra bons resultados nessa área, e também utilizou-se o método de aprendizagem supervisionada *Bayesian Ridge Regression*. Para todos os poços utilizados, verificamos a superioridade do método RCVI sobre os demais métodos, principalmente quando a amostra disponível possuía uma quantidade maior de dados.

Com relação a trabalhos futuros, podem-se enumerar algumas propostas:

- estender o método para tratar saídas vetoriais;
- na aproximação da variedade, em cada região do domínio utilizamos polinômios de grau fixo, definido no início pelo usuário. Seria interessante que se pudesse utilizar polinômios de graus diferenciados em cada região do domínio e mesmo dentro de cada região, pois no caso de se ter mais de uma função a ser aproximada, poderiam ser empregados polinômios de graus diferentes para cada uma dessas funções, segundo algum critério a ser definido;
- tentar paralelizar o algoritmo, tanto o TV quanto a geração da árvore;

PUC-Rio - Certificação Digital Nº 0710710/CA

- testar o desempenho do método para outras aplicações, como por exemplo, na área de finanças e economia.

PUC-Rio - Certificação Digital Nº 0710710/CA