1
Introdução

No estudo de aprendizagem de máquinas, muitas vezes busca-se desenvolver um classificador ou estimar uma função a partir de um conjunto de dados. Em muitas situações, há um grande número de exemplos disponíveis com muitas características envolvidas. Embora a abundância de exemplos possa ser útil para generalização do algoritmo, a existência de muitas características, algumas das quais podem ser irrelevantes, pode acarretar em um custo computacional maior. Dessa maneira, muitos algoritmos têm sido desenvolvidos com o objetivo de reduzir a quantidade de características no conjunto de dados, isto é, reduzir a dimensionalidade dos dados.

Os métodos de aprendizagem de variedades são métodos não lineares de redução da dimensionalidade e também baseiam-se na ideia de que os dados são na verdade uma amostra de uma variedade de baixa dimensão em um espaço de alta dimensão, assim, o objetivo é encontrar uma representação de baixa dimensionalidade desses dados. Dentre os métodos de aprendizagem de variedades existentes, estamos interessados no método de votação por tensores.

O método de votação por tensores (*Tensor Voting*, TV) não realiza uma redução da dimensionalidade e todas as operações ocorrem no espaço de entrada original. Este método representa os dados de entrada como tensores simétricos de segunda ordem, que codificam a presença de uma estrutura perceptual em cada ponto, indicando a saliência de cada tipo de estrutura (curva e superfície, por exemplo) a qual pertence o dado e sua orientação normal e tangente preferida. O tipo de estrutura a qual pertencem os dados pode ser inferido baseado na configuração de seus vizinhos. O método fornece a estimativa da dimensionalidade intrínseca local dos dados; permite obter estimativas confiáveis do espaço tangente e normal em cada amostra; pode ser utilizado para realizar interpolação não linear e gerar saídas a partir de entradas não observadas, mesmo na presença de ruídos. Em Medioni e Mordohai (47), é exibida uma implementação do método para N dimensões, a qual será utilizada neste trabalho.

O método proposto, diferente dos métodos de aprendizagem convenci-

onais, assim como o método de votação por tensores, não realiza o mapeamento dos dados de entrada para um espaço de baixa dimensionalidade. O método a ser apresentado busca aproximar uma variedade implícita utilizando informações da dimensionalidade intrínseca e estimativas dos vetores normais e tangentes, fornecidos pelo método de votação por tensores, o qual utiliza como dados de entrada pontos de um espaço euclidiano. Uma vez identificado o tipo de estrutura a qual pertencem os dados, uma aproximação para a variedade implícita é realizada.

Métodos de aproximação de formas implícitas, tais como curvas e superfícies, apresentam algumas vantagens sobre a representação paramétrica e, por essa razão têm motivado muitos estudos, como pode ser verificado nos trabalhos de Arouca (2), Azevedo (3), Li et al. (38), Macedo (1), Mederos et al. (42), Ohtake et al. (49), Tasdizen et al. (70), dentre outros. Algumas vantagens da representação implícita incluem o fato de que, ao modelar um objeto como uma variedade implícita, torna-se fácil projetar um ponto qualquer sobre a variedade, pois as informações sobre o espaço normal estão disponíveis diretamente. Além disso, a distância algébrica a variedade é bastante útil.

Para obter uma boa aproximação das formas implícitas utiliza-se o conceito de partição da unidade, como sugerido nos trabalhos de Li et al. (38) e Ohtake et al. (49). Desse modo, o domínio é dividido em subconjuntos e, em cada um desses subconjuntos, funções implícitas simples aproximam os dados. Neste trabalho, a subdivisão do domínio é realizada tanto usando malha 2^n-ádica, onde n representa tanto o número de características dos dados de entrada, quanto árvore de divisão binária com funções de transição suave. Em cada subdivisão, utilizam-se funções polinomiais multivariadas para aproximar os dados e, um dos critérios de parada da subdivisão espacial inclui uma precisão definida pelo usuário para o erro dessa aproximação. Essas funções simples locais podem ser combinadas para obter a forma global.

Considerando os métodos de regressão baseados em árvores, conforme trabalhos de Kubrusly (30) e Lage et al. (31), realizamos uma regressão utilizando a árvore obtida na aproximação da variedade implícita. Desse modo, temos o método de regressão construtiva em variedades implícitas, onde cada folha k da árvore guarda uma estimativa y_k, assim, a estimativa dada pela árvore para o valor de y associado a uma nova entrada $\mathbf{x}$ é expressa pelo valor de y_k obtido em cada folha k, ponderado pelo peso da entrada $\mathbf{x}$ em cada uma dessas folhas.

1.1 Contribuições

Nesta tese um novo método de regressão construtiva é proposto. O método é chamado de Regressão Construtiva em Variedades Implícitas (RCVI), utiliza redução da dimensionalidade para obter um melhor entendimento da estrutura da qual os dados fazem parte, e em seguida, aproxima uma variedade implícita empregando funções polinomiais multivariadas e partição do domínio. Essa partição do domínio resulta em uma estrutura de árvore, na qual se baseia a regressão. Vamos elencar as contribuições do RCVI.

- Primeiramente, obtém-se uma boa estimativa da dimensionalidade da variedade a partir de pontos esparsos, utilizando o método de votação por tensores. A informação dos vetores tangentes e normais obtidas pelo método TV também serão necessárias para o passo seguinte. Nesta etapa, a fim de avaliar o potencial do método TV em fornecer informações confiáveis, testamos o método com variedades diversas, a saber, curva e superfície no $\mathbb{R}^3$, curva, superfície e volume no $\mathbb{R}^4$. Logo, para avaliar as informações obtidas, precisamos utilizar ferramentas matemáticas mais sofisticadas, como no caso da avaliação do erro de orientação obtido na estimativa do plano normal em cada ponto da superfície no $\mathbb{R}^4$. Para este exemplo, precisamos utilizar a teoria de álgebra exterior e trabalhar com o produto exterior. Tal teoria também foi utilizada para avaliação da orientação da normal obtida ao trabalhar com o exemplo do volume no $\mathbb{R}^4$, onde foi preciso calcular o produto vetorial entre três vetores no $\mathbb{R}^4$. Portanto, contribuições foram dadas na verificação da confiabilidade das informações fornecidas pelo método TV;
- Aproximam-se variedades implícitas, utilizando-se funções polinomiais multivariadas e partição da unidade. O grau do polinômio a ser utilizado em cada subconjunto de dados é definido pelo usuário e a aproximação é controlada pelo erro, cuja precisão também é fornecida pelo usuário. Com relação a partição da unidade, deve-se optar por usar malha 2^n-ádica ou árvore de partição binária com funções de transição suaves. A subdivisão do domínio permite aproximar localmente a variedade implícita por funções simples. A ideia nesta etapa é utilizar os vetores tangentes e normais para obter tal aproximação, e a contribuição dada é devido a generalização realizada dos métodos de aproximação implícita existentes, geralmente empregados para aproximar curvas no $\mathbb{R}^2$ e superfícies no $\mathbb{R}^3$, fornecendo assim uma nova proposta para empregar tais métodos a outras estruturas mais complexas;

- A cada um dos dados de entrada tem-se uma saída associada. O método proposto assume que os dados de entrada se encontram em uma variedade e busca aproximar essa variedade implícita, também levando em consideração as saídas a que estão associados os dados de entrada. Para isso, inclui no processo de construção da árvore um critério baseado na dispersão das saídas, assumindo que dados de entrada próximos devem apresentar valores de saída próximos. Desse modo, em cada folha da árvore obtida, realiza-se uma regressão. Para cada novo dado de entrada $\mathbf{x}$, a saída y deve ser obtida considerando-se as estimativas dadas para y e o peso desse novo dado $\mathbf{x}$ em cada uma das folhas da árvore construída;
- Testamos o método RCVI em dados reais e realizamos uma aplicação na área de dados de poços de petróleo. Essas aplicações com o método RCVI baseiam-se na hipótese de que os dados de entrada podem fazer parte de um espaço de baixa dimensionalidade. Dessa maneira, utiliza-se redução da dimensionalidade para identificar a estrutura da qual os dados fazem parte e desse modo aproximar essa estrutura e, em seguida, realizar uma regressão.

1.2 Estrutura do trabalho

No capítulo 2, uma breve definição sobre variedades é dada e os teoremas que serão utilizados para aproximação da variedade serão apresentados. No capítulo 3, expõe-se a ideia de redução da dimensionalidade, bem como alguns métodos de redução da dimensionalidade e aproximação de funções. Já no capítulo 4, é exibido o algoritmo de votação por tensores. No capítulo 5, explicamos como aproximar a variedade utilizando partição da unidade. No capítulo 6, expõe-se o método RCVI. No capítulo 7, são apresentados resultados da aplicação do método RCVI a dados reais e em dados de poços de petróleo. Por fim, o capítulo 8 apresenta as conclusões e identifica possíveis trabalhos futuros.