UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

Algoritmo de Seleção de Instâncias Baseado em Ranking

Proposta de Trabalho de Graduação

Aluno: Nícolas Oliveira Gomes do Nascimento
Orientador: George Darmiton da Cunha Cavalcanti

Recife, Agosto de 2016

Contexto

Algoritmos de seleção de protótipos desempenham um papel fundamental em sistemas de aprendizagem baseados em conjunto de treinamento. Tais algoritmos reduzem o conjunto de dados, favorecendo de maneira expressiva problemas que possuem uma vasta quantidade de padrões.

Para tanto, algoritmos de seleção de instâncias visam preservar apenas as amostras determinantes para a classificação dos padrões e descartar outras que na prática correspondem a informação redundante ou ruído, podendo assim, diminuir espaço de armazenamento requerido e o tempo de execução para classificação, por exemplo. Os padrões selecionados são os protótipos do modelo que buscam no mínimo manter a qualidade de discriminação das instâncias representadas por eles.

Seleção de protótipos favorece diretamente classificadores em diversas áreas de aplicação como em mineração de dados, categorização de texto e análise de documentos da Web, por exemplo, em que o conjunto de dados original utilizado é bastante espeço. Também é válido mencionar que a eficiência de classificadores baseados em instâncias como o k-NN está relacionada ao tamanho do conjunto de treinamento utilizado, que calculam o grau de similaridade entre todas as amostras no processo de classifcação.

Um dos problemas enfrentados por algoritmos de seleção de protótipos é o de identificar padrões que deverão ser descartados e os que serão mantidos como representantes de certas classes, e este é um dos diferenciais entre os mais diversos métodos desta área de pesquisa. Estes algoritmos devem decidir, por exemplo, se os protótipos serão selecionados a partir das regiões centrais ou de fronteira entre as classes. Instâncias localizadas em fronteiras de classificação são mais importantes para a capacidade de classificação da máquina, enquanto que amostras mais internas não possuem grande influência, podendo então ser descartadas. Além disso, alguns algoritmos buscam remover ruído e *outliers* do conjunto de treinamento que atrapalham a classificação correta dos dados, pois são exemplos que não concordam com a mesma classe de seus vizinhos.

Motivação

Devido as suas distribuições no espaço, algumas instâncias são mais difíceis de serem corretamente classificadas pelos algoritmos de aprendizagem. A ocorrência de sobreposição entre classes em padrões próximos às fronteiras de classificação, *outliers* e ruídos, por exemplo,

é responsável por diminuir a precisão geral de algoritmos de aprendizagem. Estes grupos de instâncias possuem as mais altas probabilidades de serem mal classificados, sendo, então, bons candidatos a serem filtrados do conjunto de treinamento.

Incorporar métricas de avaliação da dificuldade em classificar cada instância do conjunto de treinamento, pode ajudar a descobrir quais padrões podem ser descartados, por representarem redundância de informação ou, devido a sua natureza, apresentar baixa probabilidade de serem corretamente classificados. Utilizando esta informação, o processo de seleção irá reduzir o conjunto de dados utilizado, bem como aumentar a precisão geral de classificação.

Objetivo

Neste trabalho, vamos propor um novo algoritmo de seleção de protótipos que inicialmente ranqueia cada instância do conjunto de treinamento, atribuindo um valor de importância, e posteriormente faz a seleção das instâncias priorizando as amostras mais significativas. Será feito também um estudo experimental para análise e comparação do desempenho do algoritmo proposto com outras técnicas do estado da arte em seleção de protótipos.

Cronograma

	Agosto	Setembro	Outubro	Novembro	Dezembro
Estudo do método proposto					
Implementação					
Execução dos testes e experimentos					
Elaboração da monografia					
Preparação para apresentação					
Apresentação					

Possíveis Avaliadores

Possível avaliador desse trabalho: Paulo Salgado Gomes de Mattos Neto.

Assinaturas

Nícolas Oliveira Gomes do Nascimento
Discente

George Darmiton da Cunha Cavalcanti
Orientador