AVALIAÇÃO DA SELEÇÃO DE ATRIBUTOS EM PROCESSO DE CONSTRUÇÃO DE CLASSIFICADORES

Marta Duarte de Barros
Universidade Federal Fluminense (UFF)
Rua Passo da Pátria 156, São Domingos, 24210-240, Niterói, RJ, Brasil
marta_uff@hotmail.com

Glauco Barbosa da Silva
Centro de Análises de Sistemas Navais (CASNAV – UFF)
Praça Barão de Ladário S/N – Centro – Rio de Janeiro
glaucos@id.uff.br

Altina Silva Oliveira
Universidade Federal Fluminense (UFF)
Rua Passo da Pátria 156, São Domingos, 24210-240, Niterói, RJ, Brasil
altinaadm@gmail.com

Helder Gomes Costa
Universidade Federal Fluminense (UFF)
Rua Passo da Pátria 156, São Domingos, 24210-240, Niterói, RJ, Brasil
Helder.hgc@gmail.com

Resumo

O presente trabalho avalia técnicas de seleção de atributos na construção de classificadores a partir de dados do Censo 2010.

O processo de descoberta do conhecimento em banco de dados (KDD) é um processo onde técnicas estatísticas e computacionais são aplicadas para identificar padrões em grandes volumes de dados. São etapas do KDD: Seleção, Pré-processamento, Transformação, Data Mining e Interpretação. A classificação é uma tarefa de *data mining* e é definida como a ordenação de objetos em classes estabelecidas a priori.

Operando com uma grande quantidade de dados, é comum a existência de dados redundantes, que tornam lento ou confuso o KDD. Como alternativa, estratégias para seleção de atributos auxiliam a melhorar os resultados do processo.

Um conjunto de 27 instâncias com 15 atributos foi selecionado para compor a amostra para o experimento. O primeiro passo foi a avaliação da correlação de Pearson(r) entre os atributos, que identificou a existência de correlações de diferentes intensidades.

Seguindo a estratégia, foi feita a avaliação da importância dos atributos para determinação das classes com o método ReliefF, que trabalha a partir da seleção aleatória de uma instância e da localização dos vizinhos mais próximos da mesma classe e de classes opostas.

O terceiro passo foi a avaliação da correlação dentre os atributos e a determinação da classe, heurística Correlation-based Feature Selection (CFS), que busca pelo melhor conjunto de atividades com baixa correlação entre si e com alta correlação com a classe. A CFS destacou um conjunto de 4 atributos principais(g1, g2, g3,g4)

Na tarefa de classificação, com 15 atributos e as classes definidas a priori, definiu-se uma baseline com um classificador (algoritmo ZeroR) que alcançou uma precisão de 33%.

Em continuidade ao experimento, face o pequeno número de instâncias, o método de treinamento *Leave-One-Out* foi escolhido. A partir daí, sem seleção dos atributos, foi construído um classificador baseado em árvore de decisão(J48), que atingiu uma precisão de 37%. Com base na seleção da heurística CFS, preservados os parâmetros da execução anterior, o

classificador(J48) atingiu 59% de precisão. A tarefa de classificação foi repetida para os algoritmos Naive Bayes e KNN.

Para o Naive Bayes, sem seleção de atributos a precisão foi de 48% e com a seleção de atributos 56%. Para o KNN(k=5), sem seleção de atributos a precisão foi de 33% e com a seleção 52 %. Uma análise dos resultados e a baixa precisão dos classificadores sugerem uma heterogeneidade intraclasses, o que dificulta a construção e precisão dos classificadores. Assumindo a heterogeneidade entre as classes, uma ação recomendada é tornar a base não-supervisionada e aplicar clusterização para que novos classificadores possam ser avaliados.

KEYWORDS. Classificação, Seleção de atributos, Data Mining.

Área principal. AO – Outras Aplicações de PO