UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
JORNADA ACADÊMICA
ISSN: 2674-6670

Avaliação de desempenho, modelagem e implementação de métodos de imputação bioinspirados na classificação multirrótulo.

Fabrício Almeida Do Carmo e Fábio Manoel França Lobato

A incompletude de informação é um problema recorrente nas análises de dados do mundo real. Quando não tratados de maneira adequada, os resultados finais podem ser seriamente comprometidos. Atuando nessa problemática, os métodos de imputação buscam valores plausíveis para preencher esses dados faltantes. Dependendo do desempenho do método, por exemplo, o classificador também tem ganho de performance quando tais analises são aplicados em tarefas de classificação. O presente trabalho foca justamente nesse cenário, analisando principalmente o contexto multirrótulo. A classificação multirrótulo (CM) é um problema de aprendizado supervisionado onde uma instância pode estar associada a múltiplos rótulos, diferente da classificação tradicional (monorrótulo) que associa um exemplo a uma única classe. Devido ao número crescente de novas aplicações, como classificação semântica de vídeos e imagens, categorização de música e diagnósticos médicos, o aprendizado multirrótulo é considerado um tópico de pesquisa emergente e promissor. A despeito da relevância da CM, a literatura dispõe de poucos trabalhos que endereçassem o tratamento de valores ausentes neste contexto. Visando contornar esta lacuna, o presente trabalho tem por objetivo desenvolver e avaliar um método de imputação múltipla de dados baseado em algoritmos genéticos aplicado no aprendizado multirrótulo. A avaliação é feita por meio de um *benchmarking* com várias bases de dados, submetendo o método a diversos cenários de distribuição de valores ausentes. Para fins de comparação e validação, o método, nomeado EvoImp, foi comparado com outras estratégias de imputação consolidadas na literatura. Em um experimento secundário, o EvoImp também foi adaptado para no aprendizado monorrótulo. Os resultados mostram-se promissores em ambos os cenários de classificação avaliado. Como trabalhos futuros pretende-se ampliar os testes para mais bases de dados e para outros mecanismos de ausência de dados além do *Missing Complete At Random* utilizado nos experimentos anteriores, além de aplicar testes estatísticos para avaliar criteriosamente a significância dos resultados.