Estimativa de Esforço em Projeto de Software com Técnicas de Aprendizado de Máquina

Carlos Bitencourt
Faculdade de Computação - Universidade Federal de Mato Grosso do Sul.

Introdução

A necessidade do desenvolvimento de soluções e serviços digitais confiáveis e seguros alavancaram a busca pela melhoria de atividades de **Engenharia de Software**. Dentre elas a **Estimativa de Esforço de Software** vem sendo aprimorada com o uso de **Aprendizado de Máquina**.

Objetivo

O objetivo deste artigo é avaliar **a acurácia** de **cinco Algoritmos** de **Aprendizado de Máquina** em cinco bases de dados de domínio público relacionados à **Estimativa de Esforço de Software.**

Metodologia

- Levantamento da base de dados (Pesquisa por bases de domínio público)
- Análise dos parâmetros
- Limpeza da base da dados (Optou-se por remover linhas com atributos vazios e nulos)
- Transformação dos parâmetros (Textuais e Categóricos).
- Seleção de Atributos (Manual)
- Normalização (z-score)
- Treinamento (Definição dos Algoritmos, Seleção dos melhores parâmetros, *Scikit-learn*)
- Métrica de Acurácia: Erro Absoluto Médio (**MAE**) e Desvio Padrão.

Metodologia

Tabela 1. Base de Dados

ID	Nº Atr. Inicial	Nº Atr. Final	Nº Linhas Inicial	Nº Linhas Final
Cocomo81	17	17	63	63
Cocomonasa	17	17	60	60
Desharnais	12	17	81	60
Nasanumeric	24	41	93	93
Seera	76	33	120	111

Metodologia

Tabela 2. Algoritmos e componentes *Scikit-learn*

Algoritmos	Componentes
DT	DecisionTreeRegressor
MLP	MLPRegressor
KNN	KNeighborsRegressor
SVM	SVR
RF	RandomForestRegressor

Metodologia

1. Para uma base e algoritmo, roda o ***GridSearchCV*** com opções de variáveis escolhidas.
2. Caso, o “melhor parâmetro" para uma determinada variável pertença a fronteira.
    - Ajusta as opções do parâmetro estendendo a fronteira. Ex: opções iniciais [2,4,8], suponha que o resultado, de melhor parâmetro, após a execução, seja [2], então ajusta-se os parâmetros para novas opções [0, 2, 4].
3. Volta ao Passo 1, até que os melhores valores de parâmetros não pertençam a fronteira de opções.

Análise

Tabela 3. Algoritmos, base de dados, melhores parâmetros

Algoritmos	Base de Dados	Melhores Parâmetros
DT	Cocomo81	criterion: absolute_error,max_depth: 8, splitter: random
KNN	Cocomo81	algorithm: ball_tree, n_neighbors: 4, weights: distance
MLP	Cocomo81	activation: relu, hidden_layer_sizes: 60, solver: lbfgs
RF	Cocomo81	max_depth: 6, n_estimators: 15
SVM	Cocomo81	C: 10, epsilon: 0.023, kernel: linear
DT	Cocomonasa	criterion: absolute_error, max_depth: 9, splitter: random

Análise

Tabela 4. Algoritmos e Desempenho em Erro Absoluto Médio (MAE) e Desvio Padrão. Inverso do Erro Absoluto Médio, lê-se quanto maior melhor.

Algoritmos	Cocomo81	Cocomonasa	Desharnais	Nasanumeric	Seera	Média
DT	0.753 ±0.101	0.798 ±0.013	0.513 ±0.106	0.716 ±0.047	0.679 ±0.037	**0,691**
KNN	0.682 ±0.194	0.607 ±0.133	0.452 ±0.070	0.672 ±0.064	0.550 ±0.094	**0,592**
MLP	0.656 ±0.138	0.783 ±0.055	0.551 ±0.032	0.600 ±0.054	0.591 ±0.018	**0,636**
RF	0.709 ±0.142	0.808 ±0.047	0.477 ±0.029	0.748 ±0.041	0.655 ±0.012	**0,679**
SVM	0.696 ±0.140	0.759 ±0.026	0.558 ±0.142	0.666 ±0.062	0.666 ±0.034	**0,669**
Média	**0.699**	**0.751**	**0.510**	**0.680**	**0.628**	

Conclusão

Conclui-se que os algoritmos com os melhores desempenhos por ordem de classificação foram: **DT**, **RF** e **SVM**. Além disso, este estudo pode servir de apoio para pesquisas na área de Estimativa de Software com apoio da Aprendizagem de Máquina para auxiliar na fundamentação das relações produzidas na pesquisa.