**An extensive experimental survey of regression methods**

**(Fernández-Delgado et al., 2019)**

Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., & Febrero-Bande, M. (2019). An extensive experimental survey of regression methods. *Neural Networks : The Official Journal of the International Neural Network Society*, *111*, 11–34. https://doi.org/10.1016/j.neunet.2018.12.010

**Resumo:**

A regressão é um problema muito relevante na aprendizagem de máquinas, com muitas abordagens diferentes disponíveis. O trabalho atual apresenta uma comparação de uma grande coleção composta por 77 modelos populares de regressão que pertencem a 19 famílias: modelos lineares e lineares generalizados, modelos aditivos generalizados, menos quadrados, métodos de projeção, LASSO e regressão de cristas, modelos Bayesianos, processos Gaussianos, regressão de quantis, vizinhos mais próximos, árvores e regras de regressão, florestas aleatórias, ensacamento e impulso, redes neuronais, aprendizagem profunda e apoio à regressão vectorial. Estes métodos são avaliados utilizando todos os conjuntos de dados de regressão do repositório de aprendizagem da máquina UCI (83 conjuntos de dados), com algumas exceções devido a razões técnicas. O trabalho experimental identifica vários modelos de regressão excepcionais: o modelo baseado em regras M5 com correções baseadas nos vizinhos mais próximos (cubista), a máquina de reforço do gradiente (gbm), o conjunto de reforço das árvores de regressão (bstTree) e a árvore de regressão M5. Cubista atinge a melhor correlação quadrática ( R2) em 15,7% dos conjuntos de dados muito próximos, com diferença abaixo de 0,2 para 89,1% dos conjuntos de dados, e a mediana destas diferenças sobre a recolha de conjuntos de dados é muito baixa (0,0192), em comparação, por exemplo, com a regressão linear clássica (0,150). No entanto, o cubista é lento e falha em vários grandes conjuntos de dados, enquanto outros modelos de regressão semelhantes ao M5 nunca falham e a sua diferença para o melhor R2 é inferior a 0,2 para 92,8% dos conjuntos de dados. Outros modelos de regressão com bom desempenho são o comité de redes neurais (avNNet), árvores de regressão extremamente aleatórias (extraTrees, que alcança o melhor R2 em 33,7% dos conjuntos de dados), floresta aleatória (rf) e ε-suportam a regressão vectorial (svr), mas são mais lentas e falham em vários conjuntos de dados. O modelo de regressão mais rápido é o de menor ângulo de regressão lars, que é 70 e 2.115 vezes mais rápido do que M5 e cubista, respectivamente. O modelo que requer menos memória é o não-negativo menos quadrados (nnls), cerca de 2 GB, à semelhança do cubista, enquanto o M5 requer cerca de 8 GB. Para 97,6% dos conjuntos de dados existe um modelo de regressão entre os 10 melhores, que está muito próximo (diferença abaixo de 0,1) do melhor R2, que aumenta para 100%, permitindo diferenças de 0,2. Portanto, desde que o nosso conjunto de dados e a nossa coleção de modelos sejam suficientemente representativos, a principal conclusão deste estudo é que, para um novo problema de regressão, algum modelo do nosso top-10 deverá atingir R2 próximo do melhor alcançável para esse problema.

4. Conclusão

O trabalho atual desenvolve uma comparação exaustiva de 77 métodos de regressão, 73 implementados em R e outros 4 em C++, Matlab e Python, sobre toda a coleção de 83 conjuntos de dados de regressão do repositório de aprendizagem de máquinas UCI, incluindo grandes conjuntos de dados até 2 milhões de padrões e 640 entradas. A coleção de modelos de regressão, que pertencem a 19 famílias diferentes, pretende ser uma amostra representativa dos métodos mais populares e conhecidos atualmente disponíveis para tarefas de regressão. Os resultados foram avaliados em termos de R2, RMSE e MAE, sendo semelhantes com as três medições, e dependendo das propriedades do conjunto de dados (tamanho e dificuldade, medidos pelo desempenho alcançado pela regressão linear clássica). Para conjuntos de dados pequenos e difíceis, a regressão linear penalizada alcança os melhores resultados, seguida de árvores de regressão aleatórias (rf) e extremamente aleatórias (extraTrees).

Para conjuntos de dados pequenos e fáceis, o modelo baseado na regraM5 com correcções baseadas nos vizinhos mais próximos (cubista) alcança o melhor resultados, seguido do comité de retropropagação neural redes (avNNet) e o conjunto impulsionador de árvores de regressão (bstTree). Finalmente, tanto para conjuntos de dados grandes-difíceis como para grandes-easy a árvore de regressão M5 é a melhor, seguindo-se o aumento do gradiente máquina (gbm) e cubista. Considerando globalmente todos os conjuntos de dados, cubista, gbm, bstTree e M5 alcançam as melhores posições, e o as diferenças entre eles estão principalmente relacionadas com: (1) o número de casos em que ultrapassam a memória e os limites de tempo (128 GB e 48 h, respectivamente): cubista e bstTree falham em 8% e 6% dos conjuntos de dados, respectivamente, gbm apenas para 1% e M5 nunca falha; e (2) a velocidade (gbm, M5 e bstTree são 70, 30 e 10 vezes mais rápidas do que cubista). Em termos de R2, gbm e M5 nunca diminuem mais do que 0,35 abaixo do melhor R2 para qualquer conjunto de dados, e R2 melhor -R2 M5 > apenas 0,25 em 2,4% dos conjuntos de dados. Outros modelos com bons resultados são extremamente árvores de regressão aleatória (extraTrees), que alcança o melhor R2 em 33,7% dos conjuntos de dados, suporta regressão vectorial (svr) e floresta aleatória (rf), mas são muito lentas, ultrapassando o máximo permitido tempo (48 h) para mais de 20% dos conjuntos de dados. Um post-thoc Teste Friedman-Nemenyi comparando o cubista e o restante modelos dá p < 0,05 (i.e., diferença estatisticamente significativa) excepto gbm, bstTree e extraTrees.

De acordo com a posição dos seus melhores modelos de regressão no ranking R2, as melhores famílias são as regras de regressão (cujos melhores modelos são o cubista e o M5), os conjuntos de reforço (gbm e bstTree), redes neuronais (avNNet), florestas aleatórias (extraTrees e rf), métodos de projeção (perseguição de projecção, ppr) e regressão vectorial de apoio (svr).Outras famílias com modelos incluídos no top-20 são os ensacadores (ensacador de modelos MARS, bagEarth), modelos de aditivos generalizados (MARS, terra), vizinhos mais próximos (kknn), modelos lineares generalizados (penalizados) e aprendizagem profunda (dlkeras). As restantes famílias apresentam desempenhos mais pobres: cumeeira e LASSO, modelos Bayesianos, regressão linear, árvores de regressão, processos Gaussianos e regressão de quantis.

O R2 supera melhor 0,5625, considerado o limiar para R2 muito bom a excelente, de acordo com a escala de Colton (Colton, 1974), para 76,2% dos conjuntos de dados. Considerando o tempo decorrido, o modelo mais rápido é o de menor regressão angular (lars), enquanto M5 e cubista são 30 e 2000 vezes mais lentos, respectivamente. No que diz respeito à memória, a regressão não negativa dos mínimos quadrados (nnls) nunca requer mais de 2 GB, enquanto que o cubista e o M5 requerem em média cerca de 3 e 8 GB, respectivamente, e o conjunto impulsionador de modelos lineares generalizados (rndGLM) requer cerca de 78 GB, ultrapassando 128 GB em cerca de metade dos conjuntos de dados. O trabalho futuro inclui o estudo das relações entre o problema da regressão e os melhores modelos, a fim de prever o melhor modelo e o seu desempenho para um dado conjunto de dados.