Evento	Salão UFRGS 2018: SIC - XXX SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2018
Local	Campus do Vale - UFRGS
Título	Aprendizagem de Máquina Verticalmente Descentralizada: Uma Abordagem Baseada em Métodos de Agregação
Autor	BERNARDO TREVIZAN
Orientador	MARIANA RECAMONDE MENDOZA GUERREIRO

Aprendizagem de Máquina Verticalmente Descentralizada: Uma Abordagem Baseada em Métodos de Agregação

Universidade Federal do Rio Grande do Sul

Bernardo Trevizan [1], Mariana Recamonde Mendoza [2]

A maioria dos problemas de aprendizagem de máquina caracterizam-se por dados centralizados em um banco de dados único, e para este cenário existe uma grande variedade de algoritmos bem estabelecidos e com bom desempenho no treinamento de classificadores. No entanto, quando os atributos estão distribuídos (particionamento vertical) em bancos de dados distintos, e por questões de custos computacionais ou privacidade das informações não podem ser compartilhados e centralizados em um local, os algoritmos clássicos já não possuem desempenho tão satisfatório. Em trabalhos anteriores do grupo foi proposto o uso de funções de escolha social (SCFs) para agregar rankings de predição gerados a partir de estimadores locais, cada qual analisando uma partição vertical dos dados. Entretanto, há uma ampla gama de agregadores apresentados na literatura, como métodos de arbitragem e combinadores. Assim, este trabalho tem como objetivo a análise comparativa de desempenho de diferentes métodos de agregação em cenários de particionamento vertical dos dados, e sua relação com as características dos dados, a fim de melhor compreender em que situações diferentes métodos são mais promissores para tratar problemas de classificação em aprendizagem de máquina distribuída.

A fim de abranger um grande conjunto de possibilidades, o critério essencial para a coleta dos dados foi a diversidade em número de instâncias, de atributos, de classes, e o grau de balanceamento entre as classes. Cinco algoritmos de classificação (modelos locais) foram utilizados de forma a representar cinco sítios distintos, cuja escolha visou contemplar métodos com diferentes vieses indutivos: naïve Bayes, support vector machine, redes neurais, árvores de decisão e K-nearest neighbors. O desempenho destes modelos foi comparado com o desempenho dos modelos globais gerados por SCFs (Borda, Copeland, Downdall, Simpson), métodos arbitrários e combinadores. Como parâmetro de comparação, também foram utilizadas média, mediana e votação simples dentre as predições locais. Para avaliar os modelos locais e globais em diferentes cenários, computou-se 10 repetições de *10-fold cross validation* e utilizou-se a métrica F1-Measure. Os atributos foram redistribuídos a cada iteração, alterando, assim, os parâmetros dos modelos locais para garantir a variação nos dados de treinamento e teste.

Observou-se que a maioria dos métodos de agregação apresenta grande variação em seu desempenho para os bancos de dados testados. Dois métodos possuem um desempenho mais consistente, sendo um deles gerado por arbitragem e outro pela SCF Simpson, com valores sistematicamente dentre as últimas posições do ranking, o que indica desempenho bastante inferior aos demais métodos testados. Os experimentos realizados com as diversas bases de dados e algoritmos de classificação distintos permitiu identificar um grande conjunto de possíveis cenários para Aprendizagem de Máquina com dados verticalmente distribuídos, sendo difícil extrair um consenso acerca do melhor método de agregação. Portanto, um melhor entendimento das relações entre as características dos dados e desempenho dos métodos de agregação para estes cenários será buscado através da análise destes resultados por árvores de decisão. Construindo estes modelos, visamos auxiliar na tomada de decisão acerca da solução mais promissora para problemas de Aprendizado de Máquina distribuído de acordo com suas propriedades intrínsecas (e.g., características dos dados e descentralização dos atributos).

[1] Bolsista IC - btrevizan@inf.ufrgs.br

[2] Orientadora - mrmendoza@inf.ufrgs.br