CONHECIMENTO FORMAÇÃO INOVAÇÃO
Salão UFRGS 2018

SALÃO DE INICIACAO CIENTIFICA XXX SIC

15 A 19 OUTUBRO CAMPUS DO VALE

UFRGS PROPESQ

Aprendizagem de Máquina Verticalmente Distribuída

Uma Abordagem Baseada em Métodos de Agregação

Bernardo Trevizan ⚛ Mariana Recamonde Mendoza

Introdução

- **Aprendizagem de Máquina Verticalmente Distribuída** é utilizada quando os atributos dos dados estão em diferentes locais e, por custos computacionais ou segurança, não podem ser compartilhados.
- Porém, quando os atributos estão distribuídos, os algoritmos clássicos não possuem desempenho tão satisfatório.
- Assim, os **métodos de agregação** podem ser aplicados para agrupar rankings de predições de modelos treinados com os dados locais, a fim de gerar um ranking global das predições.
- Entretanto, não existe uma clareza sobre o desempenho dos métodos de agregação e sua associação com as características do problema.

Objetivo

Analisar a relação de desempenho de diferentes métodos de agregação com as características dos dados a fim de explorar os diversos cenários apresentados pela aprendizagem de máquina distribuída, com interesse especial em partição vertical dos dados.

Metodologia

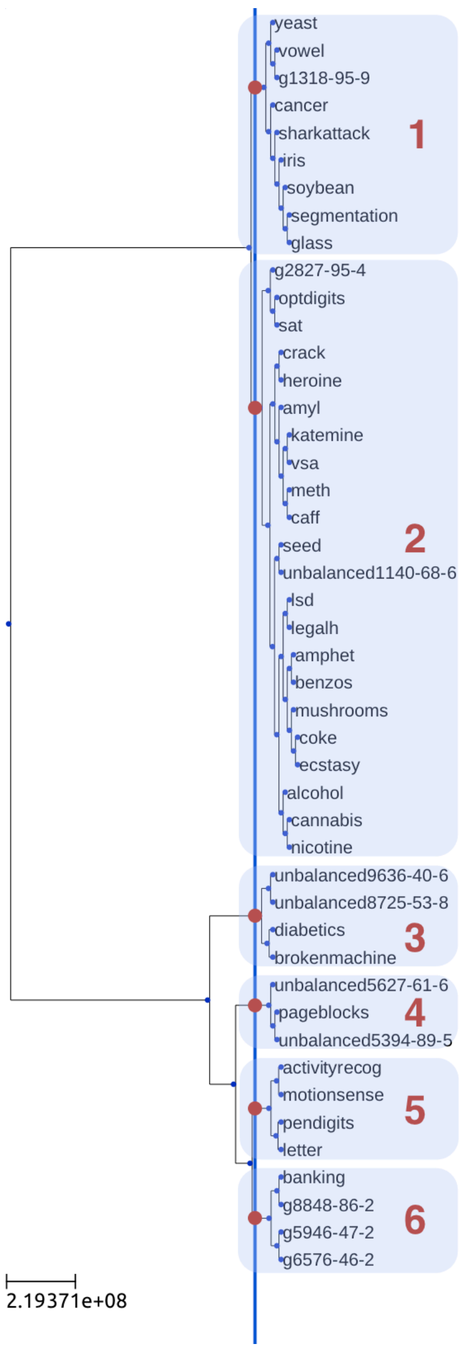

Figura 1. Agrupamento hierárquico dos conjuntos de dados analisados de acordo com suas propriedades e seus respectivos clusters conforme o corte realizado (demarcado em azul)

- O critério essencial para a coleta dos **dados** foi a diversidade em número de instâncias, de atributos, de classes e de atributos binários, silhueta e desbalanceamento.
- O **árbitro** é treinado com apenas um subconjunto das predições dos classificadores base, selecionado da seguinte forma: (i) conjunto de predições as quais discordam entre si; (ii) conjunto de predições discordantes em união com predições concordantes, porém incorretas; e (iii) união dos conjuntos (i) e (ii) com predições corretas e concordantes (iv).
- A **combinação** baseia-se na meta-aprendizagem. O método utiliza algoritmos de aprendizado para gerar um modelo de combinação a partir do conjunto de predições geradas pelos classificadores base.
- As **Funções de Escolha Social** utilizadas são: Borda, Simpson, Dowdall e Copeland. A entrada de uma função de escolha social é o conjunto de probabilidades de uma classe para todas as instâncias de todos os classificadores base. Gera-se um ranking das instâncias ao computar a pontuação de cada uma através de uma função de escolha social. Os rankings gerados são agrupados a partir das pontuações para determinar as classificações finais das instâncias.
- Para avaliar os modelos locais e globais em diferentes cenários, computou-se 10 repetições de 10-fold cross validation e utilizou-se o F1 score. Os atributos foram particionados aleatoriamente entre os modelos locais, sem sobreposição, a cada iteração.

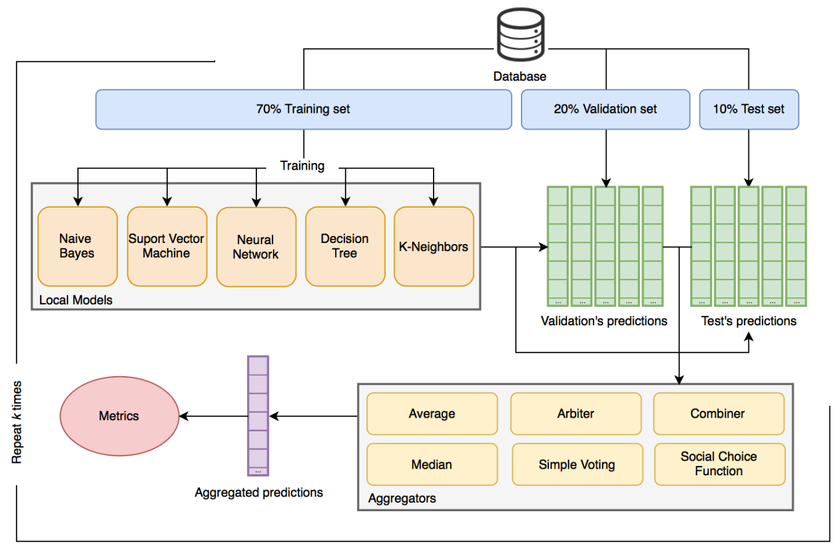

Figura 2. Framework para treinamento, agregação e teste. A cada repetição, geram-se novas partições de treino, validação e teste com os atributos distribuídos. Cada classificador base recebe todas as instâncias com um subconjuntos dos atributos. Eles são treinados com a partição de treino e geram predições com as partições de validação e teste. Os métodos de agregação utilizam as predições para gerar o ranking de predições agregadas. A partir dos ranking agregados, extrai-se o F1 score dos métodos de agregação e classificadores locais.

Resultados

- Separamos os métodos de agregação por similaridade das características dos dados que mais influenciam nos seus desempenhos.
- Observamos que diferentes características nas bases de dados podem influenciar no desempenho dos métodos de agregação.

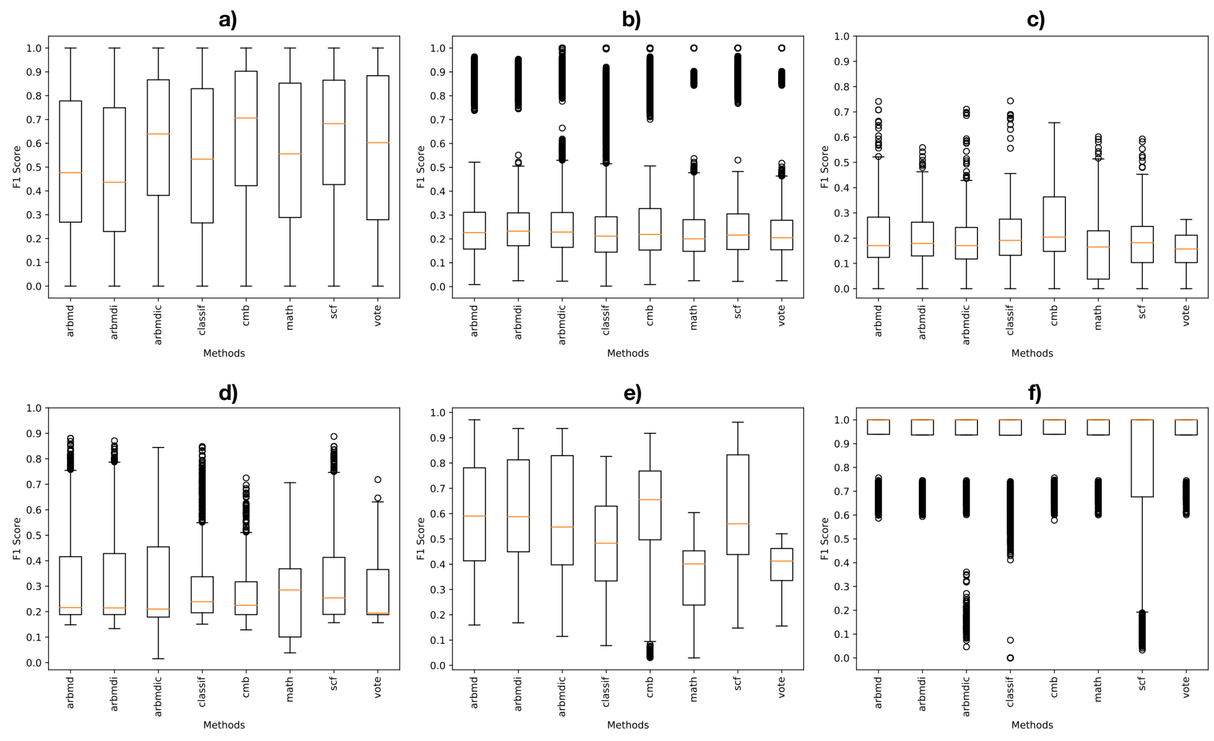

Figura 3. Distribuição do F1 score dos grupos de métodos de agregação e classificadores base para cada cluster de conjuntos de dados com características comuns: a) cluster 1, b) cluster 2, c) cluster 3, d) cluster 4, e) cluster 5 e f) cluster 6.

Conclusão

O entendimento das relações entre as características das bases de dados, agrupados através de agrupamento hierárquico, e o desempenho de métodos de agregação pode auxiliar na tomada de decisão acerca de estratégias para lidar com aprendizagem de máquina descentralizada. Apesar de preliminares, os resultados demonstram que a eficiência dos métodos de agregação é sensível a variações na estrutura dos dados analisados. Para os clusters 2 e 3, por exemplo, observamos que os métodos de agregação não possuem bom desempenho. Uma análise mais aprofundada destes resultados será realizada para criar uma relação das características dos dados com o desempenho dos métodos, visando tornar possível, de acordo com o cenário, identificar a melhor solução para um problema de Aprendizado de Máquina com atributos distribuídos.

Bolsista Bernardo Trevizan
E-mail btrevizan@inf.ufrgs.br

Orientadora Mariana Recamonde Mendoza
E-mail mrmendoza@inf.ufgrs.br

Auxílio Financeiro: FAPERGS

INSTITUTO DE INFORMÁTICA UFRGS

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL