Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciências da Computação

Análise de técnicas de Aprendizagem Ativa e Transfer Learning para reutilização de dados de outros domínios em Aprendizagem de Máquina

Proposta de Trabalho de Graduação

Aluno: Lucas de Souza Albuquerque (lsa2)
Orientador: Ricardo Bastos Cavalcante Prudêncio (rbcp)
Área: Aprendizagem de Máquina

Recife
Setembro de 2018

Sumário

1. Resumo

Um dos maiores problemas na área de Machine Learning se encontra na dificuldade de conseguir dados que representem competentemente a realidade de algum certo problema, seja pela falta dos mesmos, tempo para encontrá-los, complicações na mineração de dados, ou custo de acesso. Em muito dos casos, dados de outros domínios semelhantes são mais acessíveis, e classificadores destes domínios bem precisos - mas aplicar diretamente estes sistemas em outros contextos degrada a qualidade das previsões. As áreas de Transfer Learning e Active Learning almejam compensar os problemas de falta de dados etiquetados (e do tempo gasto para etiquetar todos os dados manualmente) de certo domínio utilizando informações e classificadores de domínios semelhantes que apresentam maior número de dados e facilidade de acesso aos mesmos.

Keywords: Transfer Learning, Active Learning, Domain, Machine Learning

2. Introdução

Apesar de técnicas de Machine Learning e Mineração de Dados terem sido altamente disseminadas e utilizadas na área de Ciências da Computação, tanto no mundo acadêmico e de pesquisas quanto em aplicações concretas e na solução de problemas do mundo real, elas ainda lidam com um grande problema. Essencialmente, o uso de Aprendizagem de Máquina tradicional assume que os domínios dos dados de treinamento e dados de testes são os mesmos - ou seja, para conseguir um avanço considerável em algum problema e extrair resultados e informações significantes, e, principalmente, para conseguir encontrar conclusões válidas, precisamos de uma base de dados válida e diretamente relevante ao problema. Para uma base ser válida, ela precisa ser grande o suficiente para uma execução concreta de ambos o treinamento e teste, precisa representar corretamente a situação em que ela está aplicada, e corretamente etiquetada, em especial para problemas de classificação. Em outras palavras, o 'output' Y previsto para um input X deve ter a garantia de que condiz com a realidade.

Infelizmente, no mundo real, as coisas não são simples assim - por exemplo, suponha que um profissional da área de Aprendizagem de Máquina construiu um sistema de Análise de Sentimento em cima de reviews de uma Câmera fotográfica. Tal profissional extraiu os dados de diferentes sites que vendem este produto; etiquetou-os de alguma forma não inteiramente automática (baseado no número de estrelas, por exemplo), que ainda precisou de configurações manuais; treinou o analisador em cima dos dados e conseguiu os resultados desejados. Agora, suponha que este mesmo profissional precise fazer uma análise de sentimentos em cima de reviews de livros, ou comida - algum produto de outro domínio.

Ora, apesar de estes novos produtos estarem no mesmo super-domínio de 'reviews' da aplicação de Câmeras Fotográficas, eles não são exatamente os mesmos - existem informações semelhantes entre os domínios, e palavras mais genéricas serão compartilhadas entre os mesmos, mas termos diferentes específicos podem ser usados, outros termos podem ter pesos diferentes em diferentes contextos (positivo em um e negativo em outro para análise de contextos). [1]

Aplicando o sistema de Análise de Sentimento de Câmeras Fotográficas no contexto de livros diretamente irá degradar os resultados - pois agora nossos dados de treinamento não condizem diretamente com a realidade da aplicação. Por outro lado, realizar novamente todo o processo de mineração, etiquetamento, entre outros, também não é a situação ideal. Isto não somente é um trabalho manual que toma tempo do profissional, mas também pode ser prejudicado no caso em que os dados de algum contexto sejam difíceis ou caros para serem coletados, ou o contexto seja relativamente novo e não apresente um número muito grande de dados. Se este contexto seja relacionado à um super-domínio com dados mais acessíveis, seria interessante aproveitar as informações já geradas e utilizadas em outras aplicações.

Existe então uma necessidade da construção de sistemas com alta performance e precisão em um contexto **target** que consiga ser treinado com a database de mais fácil acesso de domínios diferentes **source**. Técnicas com este intuito são denominadas como técnicas de **Transfer Learning**, e esta mudança na distribuição de inputs e outputs entre os dados de treinamento e teste é chamada de **Dataset Shift** em certos contextos. Estas técnicas também almejam melhorar o desempenho de sistemas, por exemplo, que demoraram um tempo considerável para serem desenvolvidos, e no qual os dados do mundo real foram afetados, apesar da aplicação ter permanecido a mesma. Uma situação detalhada por [2] seria a criação de um sistema de filtragem de e-mails spam, já que novas formas de spam estão sendo constantemente desenvolvidas, e os dados originais em que o sistema foi treinado podem estar obsoletos e não refletir mais o domínio do problema de maneira adequada.

Associado às técnicas de **Transfer Learning** está a ideia de **Active Learning**, isto é, a ideia que um sistema de Aprendizagem de Máquina pode ser mais preciso com menos labels etiquetadas no conjunto de treinamento, se o mesmo sistema consegue escolher quais subsets de data que eles usarão para aprender. No caso, um sistema de Active Learning é interativo, podendo realizar perguntas para um usuário (chamado de oracle [3]) - potencialmente na forma de instâncias de dados não etiquetadas X que precisam de um output Y correspondente.

Apesar de ainda necessitar de inputs manuais de um usuário com conhecimento da área e acesso a um número de dados, a motivação da área de Active Learning está nos problemas em que dados não classificados/etiquetados são fáceis de serem conseguidos, enquanto as etiquetas são difíceis e laboriosas de conseguir. Os inputs manuais do usuário também se tornam pontuais em diferentes etapas do aprendizado, em vez de serem uma grande etapa de preparo - e isto ajuda, quando por exemplo, estamos mudando de contexto e precisamos reorientar nosso sistema para trabalhar em cima do novo domínio, sem ter que manualmente etiquetar dados ou evitando problemas quando os novos dados não são fáceis de serem gerados ou minerados. [4]

3. Objetivos

Este trabalho propoem uma análise de diferentes técnicas das áreas de Transfer Learning e Active Learning, comparando metodologias, fundamentos teóricos, e resultados quando aplicados em situações do mundo real - almejando entender quais aspectos e variáveis dos domínios **source, target** (particularmente nas diferenças entres os domínios e features), e das técnicas em si afetam as predições das distribuições de P(Yt|Xt), isto é, da função do domínio alvo que prediz o output Yt para o input Xt. Ao longo do projeto também serão identificadas oportunidades de implementações de ferramentas em cima de dados simulados para se conseguir um maior entendimento da área.

4. Cronograma

As atividades serão desenvolvidas de acordo com o cronograma a seguir:

Atividade	Ago.				Set.				Out.				Nov.				Dez.			
Elaboração da Proposta																				
Revisão da Literatura																				
Desenvolvimento																				
Análise de Resultados																				
Escrita do Trabalho																				
Apresentação (Defesa)																				
Revisão																				

5. Possíveis Avaliadores

Listados abaixo estão alguns possíveis avaliadores do trabalho. A lista poderá ser expandida se necessário.

George Darmiton (Cin/UFPE)
Germano Crispim (Cin/UFPE)
Tsang Ing Ren (Cin/UFPE)

6. Referências

[1] WEISS, Karl; KHOSHGOFTAAR, T.m.; WANG, DingDing. **A survey of transfer learning.** Journal of Big Data, May 2016
https://doi.org/10.1186/s40537-016-0043-6

[2] QUIÑONERO-CANDELA, Joaquin; SUGIYAMA, Masashi; SCHWAIGHOFER, Anton; LAWRENCE, N.d. **Dataset Shift in Machine Learning.** Massachusetts Institute of Technology, December 2008

[3] SETTLES, Burr; **Active Learning Literature Survey.** Computer Sciences Technical Report 1648, University of Wisconsin–Madison, January 2010
http://burrsettles.com/pub/settles.activelearning.pdf

[4] WANG, Xuezhi; **Active Transfer Learning.** CMU-CS-16-119, School of Computer Science, Pittsburgh, PA, June 2016
http://reports-archive.adm.cs.cmu.edu/anon/anon/usr0/ftp/2016/CMU-CS-16-119.pdf

7. Assinaturas

_____

Lucas de Souza Albuquerque

lsa2@cin.ufpe.br

Orientando

_____

Ricardo Bastos Cavalcante Prudêncio

rbcp@cin.ufpe.br

Orientador