Análise de métodos de Inferência Ecológica em dados de redes sociais

Gustavo Penha[1,2], **Thiago N. C. Cardoso** [2], **Ana Paula Couto da Silva** [1], **Mirella M. Moro**[1]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)

[2]Hekima - Belo Horizonte – MG – Brazil

{guzpenha,ana.coutosilva,miella}@dcc.ufmg.br, {thiago.cardoso}@hekima.com

Abstract. *Online Social Networks have recently become extremely popular and generate a huge volume of spontaneous data. Knowing demographics of these users can provide useful information (e.g. marketing campaign segmentation). Unlike most approaches, we propose the use of Ecological Inference for understanding demographics of groups of people. Our results show that it is possible to infer gender and age using only a census and aggregated information obtained from a social network such as aggregated support for a political candidate.*

Resumo. *Redes Sociais Online se tornaram extremamente populares e têm gerado um enorme volume de dados. Saber as características demográficas desses usuários pode ser útil, por exemplo, para direcionamento de campanhas de marketing. Ao contrário da maioria das abordagens, nós propomos a utilização de Inferência Ecológica para entender características demográficas para grupos de pessoas. Nossos resultados mostram que é possível inferir gênero e idade utilizando apenas o censo do IBGE e informações agregadas obtidas de uma rede social, como a quantidade agregada de pessoas que apoiam um candidato político.*

1. Introdução

As mídias sociais se tornaram extremamente populares recentemente e têm gerado um grande volume de conteúdo espontâneo, tornando possível obter um feedback rápido sobre diversos assuntos como, por exemplo, a fatia de mercado de uma marca ou a opinião em relação a um candidato político entre inúmeras outras aplicações [Tumitan and Becker 2013]. Utilizar os dados gerados de redes sociais se tornou uma opção de baixo custo para estimar a opinião pública, em detrimento de métodos tradicionais como pesquisas de rua em períodos de eleição que podem ser feitas diariamente para acompanhar intenções de voto para cada candidato segmentada por classe social, idade e gênero. Dada a preocupação recente de proteger dados pessoais dos usuários das redes sociais (levando à baixa disponibilidade de atributos públicos), abordagens para inferir suas características demográficas já foram propostas na literatura possuindo diversas aplicações como o direcionamento de campanhas e serviços. Entretanto, elas possuem o objetivo de inferir características de cada usuário individualmente.

Este trabalho propõe a análise de modelos capazes de inferir características de grupos de usuários de redes sociais, sendo assim de menor custo (em atributos necessários e computacionalmente) do que inferir tais cacterísticas demográficas individualmente. Para

Tabela 1. Exemplo do problema de Inferência Ecológica. Dado uma seção eleitoral sabemos quantos homens e mulheres votaram, assim como quantos votos cada um dos candidatos receberam. Neste caso, conseguimos inferir os valores interiores da tabela?

	Dilma	Aécio	
Homem	?	?	52%
Mulher	?	?	48%
	65%	35%	

tal, utilizamos técnicas de **Inferência Ecológica**, que é o processo de extrair pistas sobre o comportamento individual a partir de informações relatadas no nível de grupo ou agregado [King et al. 2004]. Este problema surge em diversas áreas na qual pesquisadores precisam de informação de um grupo de indivíduos mas não conseguem obtê-la diretamente por motivos de privacidade, custo ou indisponibilidade, observe uma ilustração do problema na Tabela 1. A principal contribuição deste trabalho é a aplicação e comparação de modelos de Inferência Ecológica em uma base de dados gerada por **redes sociais**, inferindo o gênero e a faixa etária de grupos de usuários.

2. Trabalhos Relacionados

O objetivo da Inferência Ecológica é inferir comportamento individual a partir de dados agregados. As variáveis de interesse que tentamos extrair geralmente são de difícil acesso ou inacessíveis. Dizemos que uma tabela de inferência é 2X2 (duas colunas e duas linhas) quando temos apenas duas características e apenas dois grupos, mas existem também tabelas nas quais existem mais linhas e mais colunas, que são chamadas de casos R X C. Modelos foram desenvolvidos para ambos os casos, sendo alguns menos flexíveis [Imai et al. 2008, Wakefield 2004] ao aceitar apenas tabelas 2X2 e outros mais flexíveis ao aceitar tabelas RXC [Flaxman et al. 2015]. Inferir os valores de tais tabelas pode ser útil para uma série de problemas diferentes. Dessa maneira, técnicas de Inferência Ecológica podem ser utilizadas em diversas áreas: ciência política [King 1997, Flaxman et al. 2015], sociologia, história, epidemiologia espacial, saúde, entre outras como marketing e publicidade [King et al. 2004].

Por outro lado, a inferência de atributos dos usuários de sistemas pode ser extremamente útil, já que conhecer o perfil de uma pessoa contribui significantemente com recomendação de itens, propaganda direcionada, marketing, predição de *links* entre outros. Estudos foram feitos utilizando diferentes iterações de usuários com sistemas para inferir os seus atributos demográficos. Por exemplo, [Bi et al. 2013] reportou que, baseado somente no histórico de pesquisas em um sistema de busca, obteve uma alta acurácia de predição dos atributos de gênero e idade, assim como visão política e afiliação com o judaísmo. [Zhong et al. 2015] mostrou via um modelo de decomposição de tensores que é possível inferir uma série de características como gênero, idade, formação acadêmica, orientação sexual a partir de check-ins. Diferente das abordagens existentes de inferência dos atributos demográficos que utilizam algoritmos supervisionados de aprendizado máquina, propomos neste trabalho a aplicação de métodos de Inferência Ecológica para inferir os atributos de grupos de usuários em redes sociais.

Tabela 2. Os métodos inferem W1 e W2 a partir das outras variáveis.

Variável	Dados gênero	Dados idade
Y_i	% de homems na cidade *i*	% de pessoas com menos de 40 anos na cidade *i*
X_i	% de usuários que são favoráveis a Dilma na cidade *i*	% de usuários que são favoráveis a Dilma na cidade *i*
$W1_i$	% de homems que são favoráveis a Dilma na cidade *i*	% de pessoas com menos de 40 anos que são favoráveis a Dilma na cidade *i*
$W2_i$	% de mulheres que são favoráveis a Dilma na cidade *i*	% de pessoas com mais de 40 anos que são favoráveis a Dilma na cidade *i*
N_i	número de usuários coletados na cidade *i*	número de usuários coletados na cidade *i*

3. Configuração de Experimentos

Nesta seção é resumida a metodologia utilizada no trabalho, através dos seguintes temas: a base de dados, os métodos de Inferência Ecológica e o procedimento de avaliação.

3.1. Base de dados social

A base foi construída a partir de dados do Twitter no intervalo entre 25 de novembro de 2015 e 25 de março de 2016. Utilizando o Zahpee Monitor[1] foram coletadas somente *tweets* contendo conteúdo relacionado à Dilma Roussef. Além disso, uma equipe de cientistas político, que é composta de um doutor em Ciência Política e dois cientistas políticos da empresa Hekima[2], utiliza o software para, em conjunto com o seu algoritmo de aprendizado de máquina, definir qual o sentimento de cada *tweet*: negativo, neutro ou positivo. Utilizamos nos experimentos informações de gênero, idade e geolocalização dos autores das publicações. O gênero de cada autor é obtido a partir do cruzamento do seu nome com uma base de dados com 57209 nomes masculinos e femininos coletada da internet. A idade é extraída da biografia informada pelo próprio autor e a partir de um algoritmo supervisionado de classificação treinado com outras características públicas dos usuários. Por fim, selecionamos apenas os autores que contenham posts geolocalizados. Com isso a base de dados utilizada nos experimentos contém os seguintes microdados: id do usuário, posição em relação à Dilma (predominância de posts positivos ou negativos), gênero, idade e cidade. Além disso, utilizamos o censo do IBGE de 2010[3] que contém informações demográficas sobre cada cidade do país. A notação das tabelas de Inferência Ecológica geradas estão resumidas na Tabela 2.

3.2. Modelos de Inferência Ecológica

A escolha dos três métodos de Inferência Ecológica avaliados pelo trabalho foi baseada em critérios de reproducibilidade do modelo e relevância do artigo que o propôs (citações). Apesar disso, outros modelos de Inferência Ecológica de tabelas $2X2$ podem ser aplicados nos dados coletados e modelados neste trabalho. Os métodos escolhidos foram [King 1997], [Wakefield 2004] e [Imai et al. 2008]. As implementações de cada um dos métodos pode ser obtida nos pacotes em R disponíveis no CRAN [4], respectivamente: ei, MCMCPack e eco.

3.3. Procedimento de avaliação

Métricas de avaliação: Para avaliar a eficácia dos algoritmos de Inferência Ecológica são utilizadas duas métricas que capturam o quanto as predições se distanciam do valor

[1] `https://www.zahpee.com`
[2] `http://hekima.com/`
[3] `http://www.censo2010.ibge.gov.br/`
[4] `https://cran.r-project.org/web/packages/`

Tabela 3. Resultados de RMSE e MAE dos métodos para as duas bases sociais.

Base de gênero				
Modelo	MAE W1 (+-IC)	MAE W2 (+-IC)	RMSE W1 (+-IC)	RMSE W2 (+-IC)
King	**0.0233** +-0.0039	**0.0256** +-0.0042	0.0355 +-0.0030	0.0381 +-0.0025
Imai	0.0391 +-0.0034	0.0425 +-0.0039	0.0347 +-0.0077	0.0404 +-0.0067
Wakefield	0.0553 +-0.0043	0.0982 +-0.0062	0.0683 +-0.0008	0.0991 +-0.0014
Base de idade				
Modelo	MAE W1 (+-IC)	MAE W2 (+-IC)	RMSE W1 (+-IC)	RMSE W2 (+-IC)
King	**0.0293** +-0.0018	0.0688 +-0.0048	**0.0319** +-0.0038	0.0763 +-0.0064
Imai	0.0488 +-0.0028	0.1040 +-0.0067	0.0583 +-0.0050	0.1245 +-0.0092
Wakefield	0.0688 +-0.0042	0.0618 +-0.0065	0.0845 +-0.0832	0.0793 +-0.0026

real: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c}_i)^2}$ e $MAE = \frac{1}{N}\sum_{i=1}^{N}|c_i - \bar{c}_i|$ [Willmott and Matsuura]. Sendo que c_i é a predição do algoritmo e $\bar{c}_i$ é o valor real, para uma cidade i.

Otimização de hiperparâmetros: Inicialmente, para avaliar o impacto dos hiperparâmetros no resultado dos métodos é realizado um projeto fatorial 2^k [Jain 2008]. Os seus resultados são utilizados para dar foco nos hiperparâmetros que explicam a maior variação do RMSE. É realizada então uma busca em um intervalo maior de valores possíveis para estes hiperparâmetros, com o objetivo de encontrar aqueles que minimizam o valor do RMSE.

Comparação dos métodos: Após a otimização de hiperparâmetros, os algoritmos são comparados a partir de intervalos de confiança das métricas e do teste pareado (já que podemos medir o erro para a mesma cidade i nos diferentes modelos) *t-test* [Jain 2008] para os erros em diferentes configurações da base de dados. Para comparar os resultados dos métodos em diferentes bases de dados utilizamos o *t-test* não pareado [Jain 2008], uma vez que a representação de uma cidade i não tem vínculo entre as diferentes bases.

4. Análise e Resultados

Esta seção apresenta a análise e comparação dos métodos em diferentes configurações da base de dados.

4.1. Comparação dos métodos

Utilizando os hiperparâmetros que minimizaram o RMSE, vemos que as correlações entre a predição e o valor observado são maiores para os métodos de King e Imai. Os resultados das duas métricas de avaliação para os modelos nas duas características demográficas estão apresentados na Tabela 3. Os valores em negrito são aqueles em que os intervalos de confiança da média das métricas permitem dizer que os erros são diferentes dos outros 2 modelos com 95% de confiança. Os resultados também apontam para o melhor resultado no método de King, empatando com o de Imai em apenas uma configuração.

Os resultados dos testes pareados entre os erros absolutos dos modelos estão apresentados na Tabela 4. Os resultados em negrito são os testes que com 95% de confiança apresentam erros diferentes. Os resultados reforçam que os modelos que apresentam os melhores resultados são King e Imai, sendo que na base de idade o modelo de King apresenta os melhores resultados e na base de gênero não conseguimos refutar a hipótese nula de que os dois são iguais com uma confiança alta.

Tabela 4. P-values para o t-test pareado utilizando os erros absolutos.

	Base de gênero		Base de Idade	
Modelo	W1 pvalue	W2 pvalue	W1 pvalue	W2 pvalue
King e Imai	0.5652	**0.0248**	**2.7167e-15**	**2.3207e-11**
Imai e Wakefield	**7.1609e-23**	**4.5786e-68**	**1.4149e-90**	**5.9135e-06**
Wakefield e King	**9.0796e-30**	**1.7818e-60**	**1.5641e-60**	0.14589

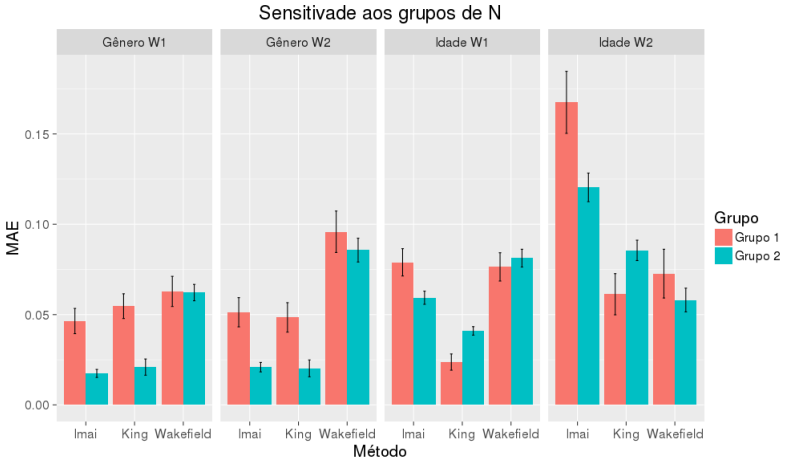

Figura 1. Gráfico mostra o resultado da métrica MAE dos métodos para os dois grupos $N < 200$ (Grupo 1) e $N > 200$ (Grupo 2) em ambas bases de dados.

4.2. Sensitividade em relação à variável N_i

A caracterização da base de dados realizada mostra que existem poucas cidades com um número grande de usuários distintos e várias cidades com número pequeno de usuários geolocalizados. Especificamente, após aplicar um algoritmo de clusterização e analisar os resultados dividimos as cidades i nos seguintes grupos: Grupo 1 com N menor que 200 e Grupo 2 com N maior que 200 (sendo N_i o número de usuários distintos na cidade i). O primeiro grupo apresenta 72 cidades e o segundo apresenta 111 cidades para os dados de idade (61 e 92 para os dados de gênero). A Figura 1 apresenta o MAE e o intervalo de confiança dos métodos nos dois grupos. Os resultados mostram que na maioria dos casos o Grupo 2 que possui amostragem maior de usuários apresenta erros menores. Tais indícios corroboram com a intuição de que amostras maiores levam a erros menores.

4.3. Sensitividade em relação ao tipo de base de dados

Para verificar se os erros obtidos na base de dados sociais são diferentes dos erros obtidos nos *benchmarks* tradicionais de dados eleitorais, a Tabela 5 mostra os resultados do t-teste não pareado entre os erros dos algoritmos na base de dados sociais e na base de dados de registro *reg*, disponibilizada por [King 1997]. Os resultados mostram que para todas as bases de dados sociais os erros são menores do que quando utilizamos uma base de dados eleitoral. Conseguimos refutar a hipótese nula de que são amostras iguais com alta confiança. Apesar disso, não podemos concluir que os algoritmos de Inferência Ecológica apresentam melhores resultados para dados sociais de maneira geral, pois, na base de dados social utilizada de apoio à Dilma, possuímos limites superiores e inferiores determinísticos mais justos se aplicarmos o método dos limites [Goodman].

Tabela 5. T-testes não pareados para os métodos de Inferência Ecológica.

	Base social (gênero)		Base eleitoral (*reg*)		t-teste não pareado dos erros entre as bases	
Modelo	MAE W1 médio	MAE W2 médio	MAE W1 médio	MAE W2 médio	p-value W1	p-value W2
King	**0.0233**	**0.0256**	0.3883	0.3442	5.5512e-73	3.8638e-57
Imai	**0.0391**	**0.0425**	0.3917	0.3116	2.363e-64	6.0851e-43
Wakefield	**0.0553**	**0.0982**	0.5399	0.1243	3.4523e-106	0.0067

5. Conclusão

Neste trabalhos abordamos o problema de inferir características demográficas de grupos de usuários em redes sociais a partir de métodos de Inferência Ecológica. Até onde sabemos essa é uma aplicação nova que apresenta vantagens sobre algoritmos de classificação supervisionada que inferem características dos usuários individualmente. Após a coleta da base de dados social, foi realizada uma avaliação experimental comparando três algoritmos do estado-da-arte em Inferência Ecológica, mostrando que os melhores resultados são do [King 1997]. Como trabalho futuro pretendemos abordar outros aspectos do problema: avaliar a utilização de um censo específico da internet e comparar os resultados dos métodos de Inferência Ecológica com resultados agregados de algoritmos de classificação supervisionada.

Referências

Bi et al., B. (2013). Inferring the demographics of search users: social data meets search queries. In *WWW*.

Flaxman et al., S. R. (2015). Who supported obama in 2012?: Ecological inference through distribution regression. In *SIGKDD*.

Goodman, L. A. Some alternatives to ecological correlation. *American Journal of Sociology*.

Imai, K., Lu, Y., and Strauss, A. (2008). Bayesian and likelihood inference for $2\times$ 2 ecological tables: an incomplete-data approach. *Political Analysis*.

Jain, R. (2008). *The art of computer systems performance analysis*. John Wiley & Sons.

King, G. (1997). *A solution to the ecological inference problem*. Princeton, NJ: Princeton University Press.

King, G., Tanner, M. A., and Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge University Press.

Tumitan, D. and Becker, K. (2013). Tracking sentiment evolution on user-generated content: A case study on the brazilian political scene. In *SBBD*.

Wakefield, J. (2004). Ecological inference for $2\times$ 2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Willmott, C. and Matsuura, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. In *WSDM*.