# Brazil Around the World: Characterizing and Detecting Brazilian Emigrants Using Google+

Johnnatan Messias
Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil
johnnatan@dcc.ufmg.br

Gabriel Magno
Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil
magno@dcc.ufmg.br

Fabrício Benevenuto
Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

Adriano Veloso
Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil
adrianov@dcc.ufmg.br

Virgílio Almeida
Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil
virgilio@dcc.ufmg.br

## ABSTRACT

Currently available data about people whose left their home country to live in a foreign country does not adequately capture the standards of contemporary global migration flows. A new trend for migration studies is to study the data from the Internet, either by Social Networks or other data in the *WEB*. In this study, we collected users data from the social network Google+ to investigate which *features* of Brazilian users are relevant to classify them as a possible emigrant. Our study uses machine learning techniques, *SVM*. We selected some *features* to compose our *dataset*. Our results show that the network features were the ones that had greater capacity for discrimination. The most relevant for the prediction of Brazilian emigrants users are, in order: reciprocity, PageRank, in-degree, clustering coefficient and ratio of incoming foreigners.

## Categories and Subject Descriptors

H.4 [**Social Media**]: Brazilian Emigration; D.2.8 [**Machine Learning**]: SVM

## Keywords

Social Media, Brazilian Emigration, Machine Learning, Google+

## 1. INTRODUÇÃO

Devido a globalização, o processo de migração internacional tem crescido nos últimos anos em muitos países estrangeiros desenvolvidos [4]. Muitos desses países tornaram-se sociedades multiétnicas devido aos altos índices de imigração [6]. Entretanto os dados atualmente disponíveis sobre o número de pessoas que deixaram o seu país de origem para morar em um país estrangeiro não captura adequadamente os padrões de fluxos migratórios globais contemporâneos [1]. Ainda, segundo os últimos dados estatísticos disponíveis do Censo Demográfico de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), o volume de brasileiros no exterior é uma questão controvérsia. Um exemplo disso é que, de acordo com o Ministério das Relações Exteriores, seriam de aproximadamente entre 2 a 3.7 milhões. No entanto, para a Organização Internacional Para As Migrações, *International Organization for Migration - IOM*, teriam, em 2010, entre 1 a 3 milhões de brasileiros no exterior [2].

Estudar o comportamento humano de migração é um campo multi-disciplinar e por isso diferentes pesquisadores de diferentes áreas atuam nesse estudo. Sociólogos, Cientistas Sociais, entre outros, tentam explicar esse comportamento. Uma nova tendência para os estudos de migração, para capturar adequadamente os padrões de fluxos migratórios, é estudar os dados obtidos da Internet, seja por Redes Sociais ou outros dados obtidos na *WEB*.

Assim, nesse trabalho, coletamos os dados dos usuários da rede social Google+ para investigarmos quais *features* dos usuários brasileiros do Google+ são relevantes para classificalos como possíveis emigrantes. O tempo de coleta de toda a rede social durou 3 meses, março a junho de 2012. Queremos, portanto, investigar quais características são determinantes para fazer com que o brasileiro deixe o Brasil para morar em um país estrangeiro. Para isso, nosso estudo faz o uso de técnicas de aprendizado de máquina supervisionado, *SVM*, para predizer se um usuário brasileiro morou ou morará, em algum momento, no exterior. Foram selecionadas algumas *features* para compor o nosso *dataset* e foi utilizado o algoritmo *SVM* para criarmos um modelo de classificação dos dados.

Nossos resultados mostram que as features de rede foram as que tiveram maior capacidade de discriminação. As mais relevantes para a predição de usuários brasileiros emigrantes são, em ordem: reciprocidade, PageRank, in-degree, coeficiente de clusterização e proporção de estrangeiros de entrada.

Esse artigo está organizado da seguinte forma. A seção 2 descreve os trabalhos relacionados. Em seguida, na seção 3 apresentamos o dataset e a metodologia utilizada juntamente com a classificação dos dados. Na seção 4 apresenta-


*WebMedia'15* , October 27-30, 2015, Manaus, Brazil

DOI: http://dx.doi.org/10.1145/2820426.2820447

mos a caracterização dos dados bem como as *features* extraídas. Os resultados da detecção são apresentados na seção 5. Por último, na seção 6, a conclusão.

## 2. TRABALHOS RELACIONADOS

Estudar o comportamento humano de migração é um campo multi-disciplinar e por isso diferentes pesquisadores de diferentes áreas atuam nesse estudo. Sociólogos, Cientistas Sociais, entre outros tentam explicar esse comportamento. Uma nova tendência para os estudos de migração é estudar os dados obtidos da Internet, seja por Redes Sociais ou outros dados obtidos na *WEB*.

O trabalho [9] estudou o comportamento de migração humana utilizando dados de geolocalização dos *IPs* de mais de 100 milhões de usuários anônimos obtidos de logins do *Yahoo! Web Services*. Eles consideraram usuários que, pelo período de um ano, passou 3 meses em um país diferente do de origem (migrantes) ou menos do que um mês (turistas). Na predição dos fluxos de migração e turismo, utilizaram, como atributos, laços coloniais, localização geográfica, desenvolvimento econômico e controle de visto. Suas análises mostraram a persistência de padrões de migração tradicionais, como o surgimento de novas rotas. Ainda, as migrações tendem a ser mais pendular entre países com fronteiras. Observaram também níveis particularmente altos de pendularidade dentro do espaço econômico europeu mesmo considerando restrições territoriais.

Em [11] foi utilizado dados de geolocalização de 500.000 usuários da rede social *Twitter*. Foram considerados somente usuários pertencentes ao países membros do *OECD*[1]. Eles analisaram uma subamostra de usuários que postaram *tweets* regularmente com dados de geolocalização. Propuseram a aproximação *difference-in-differences* para reduzir o viés de seleção quando houver tendência nas taxas de emigração para países individuais. Seus resultados mostram que a abordagem utilizada é relevante para abordar duas questões na literatura de migração: (1) os métodos podem ser usados para prever pontos de *turning* na tendência de migração; (2) dados de geolocalização do "Twitter"podem melhorar o entendimento das relações entre migração interna e internacional.

Para investigar a tendência de migração internacional de trabalhadores, o [8] analisou dados de milhões de usuários, mais precisamente do histórico de carreira, com dados de geolocalização fornecido pelo *LinkedIn*. Confirmaram que os EUA é absolutamente o principal destino para imigrantes profissionais internacionais. Entretanto, no período de 2000 a 2012 houve um decaimento na taxa de imigração no percentual de imigrantes profissionais em todo o mundo que possuem os EUA como país de destino. Isso, somente para pessoas com diplomas de graduação, mestrado ou doutorado. Suas análises também revelaram um crescimento da taxa de migração profissional asiática. Seus resultados, ainda mostraram um crescimento na taxa de migração estudantil que escolheram os EUA como país de destino.

Diferentemente dos trabalhos abordados na literatura, nosso estudo faz o uso de técnicas de aprendizado de máquina supervisionado, *SVM* para predizer se um usuário brasileiro morou ou morará, em algum momento, no exterior. Também investigamos quais características são determinantes para fazer com que o brasileiro deixe o Brasil para morar em um país estrangeiro. Para isso utilizamos dados da rede social *Google+*. Foram selecionadas algumas *features* para compor o nosso *dataset* e foi utilizado o algoritmo *SVM* para criarmos um modelo de classificação dos dados. Maiores detalhes serão explicados posteriormente nas seções seguintes.

[1]Organisation for Economic Co-operation and Development

## 3. METODOLOGIA

Nessa seção vamos explicar o processo metodológico necessário para a realização da coleta e filtragem dos dados. Para a coleta dos dados do Google+ utilizamos a linguagem de programação *Python*. Foi realizada a coleta dos perfis de usuários bem como do campo "Places lived" ("Lugares onde morei"). Esse último campo foi necessário para inferirmos a geolocalização, isto é, os países em que os usuários moraram. Após a coleta foi preciso realizar uma filtragem dos dados coletados. Essa filtragem é necessária para selecionarmos os usuários brasileiros da nossa base.

### 3.1 Coleta dos Dados

O sistema de coleta foi implementado em *Python*, usando uma abordagem de servidor-cliente, onde o servidor gerencia a lista de IDs a serem coletados e os clientes solicitam novos IDs para serem coletados. Foram utilizadas 11 máquinas com IPs diferentes para realizar a coleta em paralelo, de forma que grandes quantidades de dados pudessem ser coletadas eficientemente. A informação do perfil foi obtida realizando uma requisição HTTP à página do perfil público do usuário no Google+. A informação do grafo é coletada realizando requisições às duas "listas de círculos" correspondentes no perfil do usuário.

Para coletar IDs de usuários, inspecionamos o arquivo `robots.txt`, que contém o mapa do site, que por sua vez contém as URLs de todos os perfis do Google+. Como coletamos a lista completa de perfis fornecida pelo Google, acreditamos que coletamos todos os usuários que possuíam perfil público no Google+ na data em que a coleta foi realizada.

A coleta de dados foi realizada do dia 23 de março até o dia 1 de junho de 2012. Ao inspecionar o mapa do site encontramos 193,661,503 IDs de usuários. No total, fomos capazes de coletar a informação de 160,304,954 perfis, pois alguns IDs foram deletados ou não fomos capazes de extrair suas informações. Através dos links sociais contidos nos perfis dos usuários (amigos e seguidores), construímos um grafo social direcionado que possui 61,165,224 vértices e 1,074,088,940 arestas. O número de vértices (61 milhões) foi menor do que o número de perfis (160 milhões). Isso ocorre porque, como não dependemos do grafo para encontrar usuários, boa parte deles podem não ter disponibilizado publicamente suas listas de círculos, apesar de terem outras informações do perfil disponíveis.

Também foram coletadas as postagens (posts) públicas dos usuários. Entre os 160 milhões de perfis coletados, apenas 8,564,462 (5%) têm pelo menos uma postagem pública. Foi possível coletar apenas as 10 últimas (mais recentes) postagens de cada usuário, totalizando 29,366,310 posts.

### 3.2 Geolocalização dos Usuários

Para estudar o efeito de emigração dos usuários no Google+ exploramos o campo "Places lived" ("Lugares onde morei"). Nesse campo o usuário pode listar lugares do mundo onde já morou. Os items da lista são de texto livre, então o usuário pode escrever lugares em níveis variados ('Brasil', 'Maranhão' ou 'São Luís'), em línguas diferentes ('United

States' ou 'Estados Unidos') ou até mesmo em variações da mesma língua ('São Paulo' ou 'Sampa'). Felizmente, o Google+ associa o texto digitado pelo usuário a um ponto no Google Maps, então foi possível coletar a lista de coordenadas geográficas associadas aos lugares. Outra característica importante é que a lista não tem nenhuma ordem definida ou sugerida pelo sistema, então não há garantia de cronologia entre dois lugares presentes na lista. Dessa forma, para cada um dos usuários que compartilharam o campo "Places lived", obtivemos uma lista de países convertendo as coordenadas geográficas para o país correspondente. No total, temos 22,578,898 (14.08%) usuários com localização válida.

Além dos países dos usuários também analisamos a língua em que escrevem suas postagens. Para identificar a língua dos *posts* utilizamos o `langid.py` [2], uma solução para identificação de línguas que fornece a probabilidade de um determinado texto pertencer a uma certa língua, funcionando bem tanto para documentos longos quanto para documentos curtos, incluindo microblogs [5]

## 3.3 Filtragem dos Dados

Após realizada a coleta e o pré-processamento dos dados foi necessário selecionar os dados dos usuários brasileiros de toda a base de dados. Devido ao fato de o Google+ não possuir uma ordem definida ou sugerida para a inserção desses dados, não tendo, portanto, garantia de cronologia, tivemos que fazer uma consideração para identificar usuários brasileiros. Consideramos como usuários brasileiros aqueles usuários que tenham morado em algum momento no Brasil e que tenham como idioma principal o Português. Com essa abordagem podemos ter algum ruído como, por exemplo, selecionar usuários estrangeiros de países onde a língua oficial seja o Português e que tenham morado no Brasil, como no caso da Angola. Todavia, para contornar o problema da cronologia tivemos que adotar essa abordagem. Para selecionar esses usuários foi necessário realizar os seguintes passos: (1) Selecionar os usuários que tenham morado no Brasil, ou seja, que possuam na lista de países morados a sigla "BR"; (2) Desses usuários, selecionamos apenas aqueles que, dentre os idiomas utilizados nos 10 (dez) posts disponíveis no Google+, possuam o idioma Português utilizado em, pelo menos, 90% dos posts.

Em seguida, para compor um dataset de usuários brasileiros com features mais consistentes incluímos mais características dos usuários à nossa base de dados. Foram incluídas a lista de países, em ordem alfabética, que os brasileiros moraram; o número de idiomas utilizados; o gênero; relacionamento; ocupação; grupo de ocupação; confirmação da conta; interesses na rede social; características de grafo (in-degree, out-degree, reciprocidade, coeficiente de clusterização, índice pagerank). Ainda, incluímos a taxa de amigos estrangeiros que o usuário brasileiro possui, ou seja, a taxa de estrangeiros que ele segue e a taxa de estrangeiros que o seguem.

Após o processo de filtragem a base foi reduzida para 50.946 usuários considerados brasileiros, portanto, os estudos realizados nesse artigo utilizará essa nova base de dados.

# 4. CARACTERIZAÇÃO

Nessa seção resolvemos investigar quais são os países em que os brasileiros, usuários do Google+, moraram ou moram. Dentre os usuários brasileiros considerados, tivemos que 3.615(7%) brasileiros moraram em algum momento em um país estrangeiro, ou seja, usuários brasileiros emigrantes. Ainda, utilizamos diversas características dos usuários do Google+ que pudessem gerar alguma evidência para a separação das classes dos usuários em emigrantes e não-emigrantes. As características também podem ser classificadas em relação a sua natureza, podendo ser numéricas, categóricas ou lógicas.

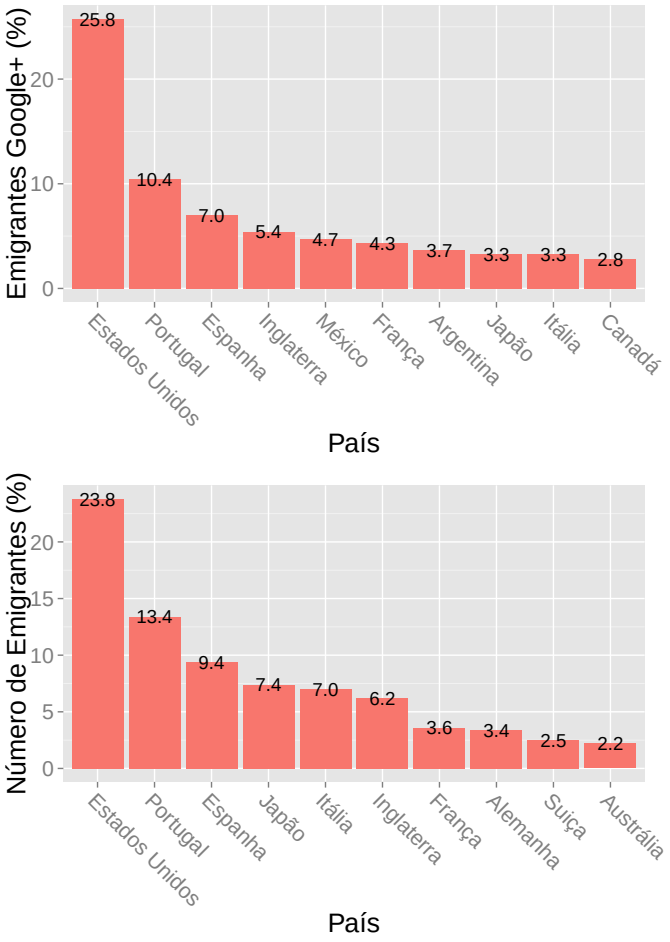


**Figure 1: Países com maior imigração brasileira da base filtrada**

## 4.1 Caracterização dos Emigrantes

A partir da nossa base filtrada, que possui 50.946 usuários considerados brasileiros, resolvemos investigar quais são os países em que esses brasileiros moraram ou moram. Dentre os usuários brasileiros considerados, tivemos que 3.615(7%) moraram em algum momento em um país estrangeiro, ou seja, usuários brasileiros do Google+ considerados emigrantes. A Figura 1 mostra o top 10 dos países escolhidos pelos brasileiros da nossa base, sendo: (1) a figura acima referente aos dados coletados do Google+; (2) a figura abaixo referente aos dados do Censo Demográfico de 2010 do IBGE [2], esses são os dados mais recentes disponibilizado pelo IBGE referentes a emigração brasileira.

De fato, EUA e Portugal já eram esperados como prin-

[2] `https://github.com/saffsd/langid.py`

cipais destinos dos brasileiros, quando considerado que os EUA são o centro global de imigração, possuindo o maior fluxo imigratório. Ainda, uma pessoa possui 4 vezes mais probabilidade de migrar para um país que possua laços coloniais com o seu país de origem e possui, também, maior facilidade em migrar para um país com língua em comum, como é o caso de Portugal e também explicado no trabalho [9].

Podemos observar que os países EUA, Portugal, Espanha, Japão, Itália e Inglaterra possuem aproximadamente 55.18% e 70% dos usuários brasileiros emigrantes, para Google+ e IBGE, respectivamente. Ao verificar a correlação entre os dados disponíveis do IBGE e os obtidos da nossa base do Google+, para os países do top 10 da nossa base, com exceção do México (por não termos os dados do IBGE para ele), obtemos um coeficiente de Pearson de 0.9452, significando que há uma forte correlação entre as bases.

Sobre a diferença dos países em termos de posições no top 10 pelo Google+ e IBGE pode ser devido ao fato de que, nesses países com posicionamento distinto, o Google+ não seja muito popular.

## 4.2 Caracterização das Features

A fim de identificar emigrantes e não-emigrantes utilizamos diversas características dos usuários do Google+ que pudessem gerar alguma evidência para a separação das classes. Extraímos características das três possíveis fontes de dados do dataset: perfil, grafo social e texto. As características também podem ser classificadas em relação a sua natureza, podendo ser numéricas, categóricas ou lógicas.

### 4.2.1 Features Categóricas

Todas as features categóricas foram extraídas do perfil do usuário. O gênero e o relacionamento são campos com opções pré-estabelecidas. O campo ocupação é um campo aberto, onde os usuários podem digitar o que quiserem para descrever sua atividade profissional. Nesse caso, foi necessário realizar uma tarefa de sumarização das informações introduzidas pelos usuários. Primeiro, contamos as 100 strings com maior ocorrência no dataset. Em seguida, como uma mesma profissão pode ser escrita de maneiras diferentes, inclusive em línguas diferentes (ex: *student*, *study*, *estudante*, *go to school*), agregamos manualmente as relacionadas. Em seguida, selecionamos as 30 ocupações mais populares e usamos o *Standard Occupational Classification - SOC* do *U.S. Bureau of Labor Statistics* [10] para classificar as ocupações dentre as categorias profissionais estabelecidas no SOC. Também incluímos as ocupações "estudante" e "aposentado", pois apesar de não estarem no SOC, tiveram grande ocorrência em nosso dataset.

A lista das categorias de cada atributo está listada na Tabela 1. É importante observar que para o caso das ocupações, como foi realizada uma filtragem dos dados, algumas categorias da ocupação não aparecem no dataset de brasileiros (marcadas com um '*'). Também apresentamos a distribuição das categorias entre emigrantes e não-emigrantes na Figura 2.

**Table 1: Lista das categorias das features categóricas**

| Gênero | Relacionamento | Ocupação |
|---|---|---|
| Masculino | Solteiro(a) | Alimentos |
| Feminino | Comprometido(a) | Arquitetura e Engenharia |
| Outro | Noivo(a) | Arte e Design |
| | Casado(a) | Ciência* |
| | É complicado | Educação |
| | Em um relac. aberto | Gerência* |
| | Viúvo(a) | Informática e Matemática |
| | Moro com alguém | Jurídico* |
| | Em uma união civil | Mída |
| | | Negócios e Finanças* |
| | | Religião |
| | | Saúde |
| | | Vendas* |
| | | Estudante |
| | | Aposentado(a)* |

* Categorias não presentes no dataset filtrado

### 4.2.2 Features Lógicas

As features lógicas são features "booleanas", ou seja, podem assumir o valor TRUE ou FALSE. Temos cinco características lógicas, todas extraídas do perfil do usuário. O campo "confirmed" indica se o perfil foi verificado como autêntico por uma equipe interna do Google. Geralmente é utilizado por celebridades e pessoas públicas em geral, funcionando como um "selo de autenticidade" que confirma que é um perfil oficial.

Os outros quatro campos estão relacionados ao interesse do usuário dentro do Google+: amigos, namoro, networking e relacionamentos. Para cada um desses atributos o usuário pode marcar o que é de interesse para ele ou não, no ambiente do Google+. A proporção de usuários emigrantes e não-emigrantes das features lógicas está apresentada na Figura 2.

### 4.2.3 Features Numéricas

Temos ao todo 8 features numéricas, extraídas de três diferentes fontes de dados. A feature "número de línguas" é de natureza textual, e indica a quantidade de línguas diferentes presentes nos posts do usuário. Espera-se que emigrantes tenham maior susceptibilidade de escrever textos em mais de uma língua. É importante observar que, como utilizamos a língua para realizar a filtragem de brasileiros, temos que o número de línguas faladas é no mínimo uma, e o máximo teórico é 10, já que temos apenas os últimos 10 posts do usuário. Como explicado anteriormente, utilizamos a ferramenta `langid.py` para detectar a língua dos textos.

O grafo social do Google+ é um grafo direcionado, similar ao Twitter, em que as pessoas podem colocar usuários em seus círculos (seguir) e podem ser colocadas nos círculos dos outros (seguidas). Utilizando esse grafo de conexões, medimos cinco métricas de rede. O "in-degree" e o "out-degree" de cada usuário medem respectivamente o número de seguidores e o número de seguidos. A reciprocidade mede a porcentagem de arestas recíprocas em relação às arestas de saída, ou seja, mede a proporção de pessoas que "seguiram de volta" o usuário. O "Clustering Coefficient" mede o grau de conectivade na vizinhança do usuário. O PageRank é uma métrica que mede a "importância" de cada vértice do grafo levando em consideração apenas a estrutura global do grafo e a posição do usuário no mesmo.

Além das características estruturais do grafo, também medimos duas métricas que levam em consideração as informações do perfil dos amigos de um usuário. Para cada usuário em nosso dataset, classificamos seus amigos e seguidores como estrangeiros ou não estrangeiros. Um amigo é considerado estrangeiro caso ele não faça parte da base de usuários brasileiros filtrada, caso contrário é considerado não es-

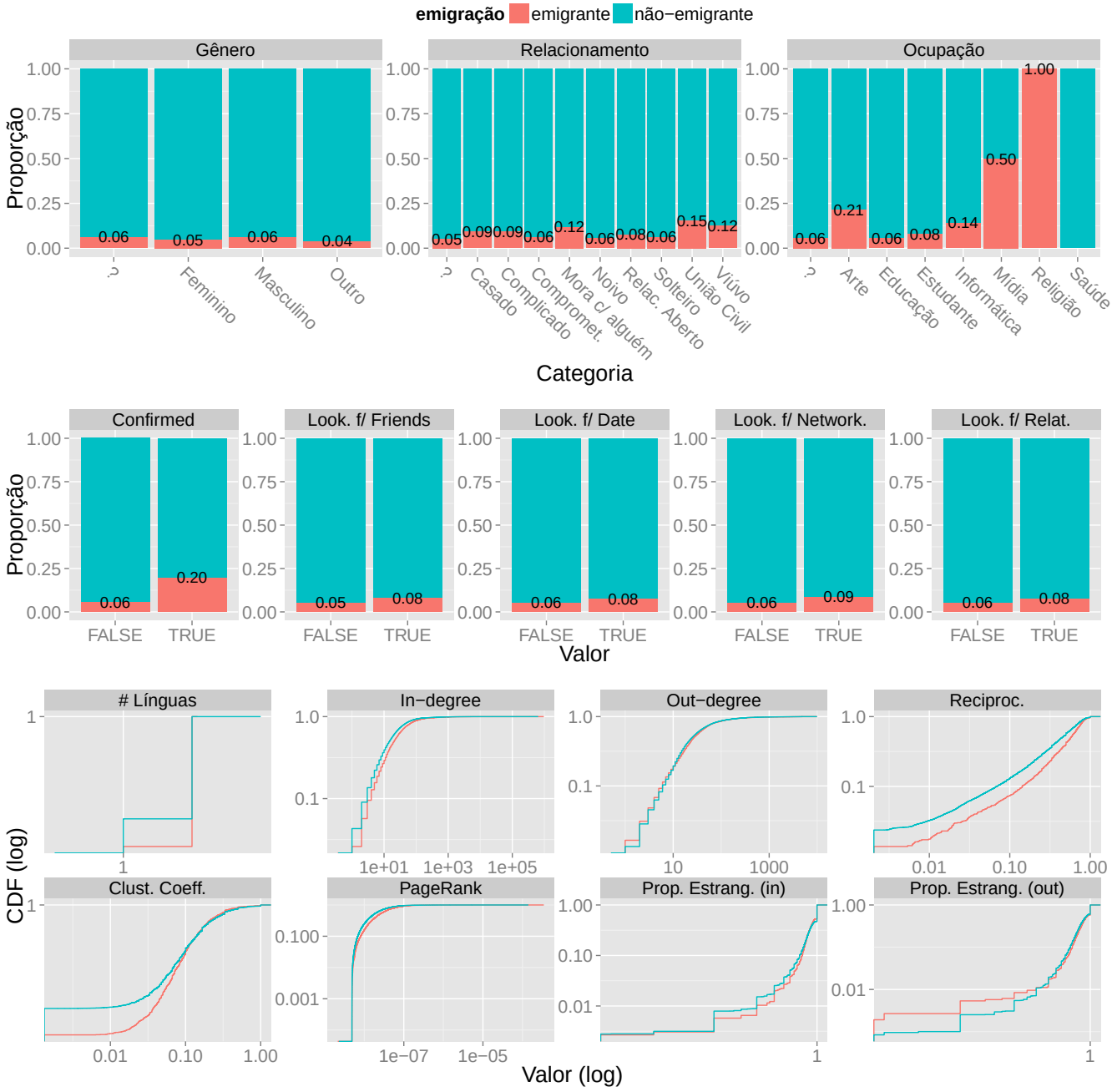


**Figure 2: Distribuição das features do dataset em relação aos emigrantes e não-emigrantes**

trangeiro. Medimos então a proporção dos amigos que são estrangeiros, tanto para os seguidores (in) quanto para os seguidos (out). Espera-se que usuários emigrantes tenham mais amigos estrangeiros do que uma pessoa não-emigrante que nunca viajou.

A Figura 2 mostra o gráfico da distribuição acumulada (CDF) das features numéricas. É interessante observar que algumas features, como o "out-degree", têm distribuições bem similares para emigrantes e não-emigrantes. Por outro lado, existem features, como a reciprocidade, que apresentam distribuições relativamente diferentes entre as classes, o que é um indício de uma boa característica discriminativa para detectar emigrantes.

## 5. DETECÇÃO DE EMIGRANTES

Nesta seção vamos avaliar a possibilidade de se detectar emigrantes brasileiros usuários do Google+ utilizando as características coletadas de seus perfis de usuário. Utilizamos um *bagging* de SVM como processo de aprendizado supervisionado para realizar a tarefa de classificação entre emigrantes e não-emigrantes.

## 5.1 Ambiente de experimentação

As tarefas de pré-processamento dos dados, bagging, SVM, e avaliação dos modelos de classificação foram implementadas na linguagem *Python* v3.4.0, utilizando a biblioteca *scikit-learn* [7] v0.16.1. Para o cálculo de *Information Gain* foi utilizado o Weka [3] v3.7.12. Os experimentos foram realizados numa máquina virtual com 8 núcleos e 16GB de memória RAM. O sistema operacional é o Linux com kernel 3.13.

## 5.2 Pré-processamento dos Dados

Devido às features categóricas presentes em nosso dataset, foi necessário realizar um pré-processamento dos dados. Como os classificadores assumem dados contínuos, uma representação direta com inteiros não faria sentido, pois as classes das features categóricas seriam interpretadas como ordenadas. Resolvemos utilizar uma codifição *One Hot*, que cria uma feature binária para cada valor possível de uma feature categórica. Como nosso dataset é esparso, ou seja, poucos usuários têm todos os campos preenchidos, resolvemos considerar a classe "desconhecido" para nossas 3 features categóricas. Com isso, o gênero será representado por 4 features, o relacionamento por 10 e a ocupação por 8. No caso das features numéricas resolvemos desconsiderar os usuários que não tivessem preenchido ou que não fosse possível calcular a métrica (ex.: clustering coefficient para um usuário de 1 amigo). Ao final, temos 35 features, e 45,429 usuários, sendo 42,817 não-emigrantes e 2,612 emigrantes.

## 5.3 Ranking de Features

Algumas features são mais importantes na predição do que outras. Por esse motivo, desejamos avaliar o potencial discriminativo de cada feature e criar um ranking de importância das mesmas. Para isso calculamos o *Information Gain - IG* (ganho de informação) para nossas 16 características dos usuários. Como resultado, a Tabela 2 lista as features e os respectivos ganhos de informação.

**Table 2: Ranking do ganho de informação das features**

| IG Rank | IG | Feature |
|---|---|---|
| 1 | 0.00444037 | Reciprocidade |
| 2 | 0.00423303 | PageRank |
| 3 | 0.00256617 | In-degree |
| 4 | 0.00179552 | Clust. Coeff. |
| 5 | 0.00136334 | Prop. Estrang. (in) |
| 6 | 0.00103249 | Gênero |
| 7 | 0.00090245 | Look. f/ Networking |
| 8 | 0.00089999 | Look. f/ Friends |
| 9 | 0.00062809 | Out-degree |
| 10 | 0.00048379 | Prop. Estrang. (out) |
| 11 | 0.0001631 | Look. f/ Relationship |
| 12 | 0.00014254 | Relacionamento |
| 13 | 0.00011345 | Look. f/ Dating |
| 14 | 0.00003754 | Confirmed |
| 15 | 0.00000443 | Ocupação |
| 16 | 0 | langnumber |

É interessante observar que as features de rede foram as que tiveram maior capacidade de discrimanação por terem um maior ganho de informação. Assim, dentre as features listadas essas são as mais importantes para a classificação dos usuários em emigrantes e não-emigrantes. Esse resultado é consistente com as distribuições da Figura 2. Emigrantes têm reciprocidade e PageRank maiores. Isso pode ser explicado pelo fato de emigrantes terem oportunidade de contato com pessoas de outros países, aumentando assim seu "networking". Outra boa feature é a Proporção de Estrangeiros de entrada, indicando que existe homofilia na relação de amizade entre emigrantes. Ou seja, emigrantes têm mais susceptibilidade de fazerem amizades com outros emigrantes, sejam eles do Brasil ou de outro país.

## 5.4 Classificação dos Dados

Nessa seção vamos abordar os algoritmos utilizados para a realização da classificação dos perfis dos usuários em emigrantes e não-emigrantes. Para realizar a tarefa de classificação utilizamos o classificador SVM, um algoritmo de aprendizado de máquina supervisionado.

### 5.4.1 SVM (Support Vector Machine)

Para realizar a tarefa de classificação utilizamos um classificador SVM. Nessa técnica de aprendizado de máquina supervisionado cada usuário é representado por um vetor $U = (f_{1u}, f_{2u} \cdots f_{nu})$ de features e uma classe objetivo $y$. A classe objetivo possui a identificação do usuário, isto é, emigrante para caso o usuário seja um emigrante e não-emigrante, caso contrário. São usuários emigrantes brasileiros aqueles usuários que moraram em, pelo menos, dois países (incluindo o Brasil), possuindo na lista de países, pelo menos, dois países.

A parametrização do SVM foi feita utilizando um processo de validação cruzada com 5 partes (5-fold). A otimização dos parâmetros do *SVM* foi feita avaliando a métrica AUC (*Area Under Curve*). Trata-se de uma métrica para classificação binária. A acurácia não foi utilizada como métrica objetivo porque nosso dataset é desbalanceado em número de classes. Utilizamos um normalizador gaussiano para normalizar (scaling) os dados. Avaliamos o desempenho do SVM linear tradicional e do SVM com kernel RBF separadamente.

Como resultado da classificação, o *SVM* com kernel RBF se saiu melhor com o AUC de $\mathbf{0.6275} \pm 0.0161$.

Vale ressaltar que, a partir de agora, possuímos o modelo resultante do treino realizado pelo SVM. Assim, para classificarmos um novo usuário do Google+, selecionado ao acaso, em emigrante ou não-emigrante brasileiro será necessário obtermos suas características e processá-las no modelo gerado. Como resultado obteremos se esse usuário será classificado como emigrante ou não-emigrante brasileiro.

### 5.4.2 Bagging Balanceado

Nosso dataset é bem desbalanceado na frequência das classes: temos 16 vezes mais não-emigrantes do que emigrantes. Nesse caso o classificador poderia classificar todos como não-emigrantes que seu erro seria mínimo. Como queremos classificar bem os usuários das duas classes, resolvemos abordar uma estratégia para tentar contornar o problema do desbalanceamento.

Utilizamos um bagging com bootstrap (reposição) que usa 10 classificadores SVM. A diferença de um bagging tradicional é que a seleção é feita sempre igualando o número de exemplos de cada classe. Nesse caso, o conjunto da classe menor é fixado, e selecionamos aleatoriamente o mesmo número de exemplos da outra classe. Espera-se que isso gere uma robustez no modelo, levando em consideração o desbalanceamento.

Vamos comparar o desempenho de um Bagging tradicional e do Bagging balanceado. Assim como anteriormente,

**Table 3: Resultado dos Modelos de Bagging com SVM**

| | Bagging Tradicional | | | Bagging Balanceado | | |
|---|---|---|---|---|---|---|
| | precision | recall | f1-score | precision | recall | f1-score |
| não-migrante | 0.96 | 0.72 | 0.82 | 0.97 | 0.58 | 0.73 |
| migrante | 0.11 | 0.55 | 0.18 | 0.09 | 0.70 | 0.16 |
| avg / total | 0.91 | 0.71 | 0.79 | 0.92 | 0.59 | 0.70 |
| Accuracy | 70.95 % | | | 58.89% | | |
| AUC | 0.6332 | | | 0.6426 | | |

utilizamos o SVM como classificador base. Para efeitos de comparação, avaliamos o AUC score empírico dos dados. Os resultados são apresentados na Tabela 3. Os resultados mostram que o efeito do Bagging balanceado foi aumentar o AUC, mas a acurácia diminuiu consideravelmente em relação ao Bagging tradicional. Concluímos então que o Bagging balanceado não traz uma vantagem tão grande, sendo então recomendado utilizar o bagging tradicional.

# 6. CONCLUSÃO

Mensurar dados sobre o número de pessoas que deixaram o seu país de origem para morar em um país estrangeiro é uma tarefa difícil. Muitos dados atualmente disponíveis e mensurados pelo governo para tentar estimar o fluxo de migração não captura os padrões de fluxos migratórios globais contemporâneos. Por isso, nesse trabalho, coletamos os dados dos usuários da rede social Google+ para investigarmos quais *features* dos usuários brasileiros do Google+ são relevantes para classifica-los como possíveis emigrantes. Para isso, nosso estudo fez o uso de técnicas de aprendizado de máquina, *SVM*, para predizer se um usuário brasileiro morou ou morará, em algum momento, no exterior.

Nossos resultados mostram que as features de rede foram as que tiveram maior capacidade de discriminação. As mais relevantes para a predição de emigrantes brasileiros usuários do Google+ são, em ordem: reciprocidade, PageRank, "in-degree", coeficiente de clusterização e proporção de estrangeiros de entrada. Concluímos, também, que o Bagging balanceado não traz uma vantagem tão grande, sendo então recomendado utilizar o Bagging tradicional. No SVM, a acurácia não foi utilizada como métrica objetivo porque nosso dataset é desbalanceado em número de classes. Utilizamos um normalizador gaussiano para normalizar (scaling) os dados. Avaliamos o desempenho do SVM linear tradicional e do SVM com kernel RBF separadamente. O *SVM* com kernel RBF se saiu melhor com o AUC de $\mathbf{0.6275} \pm 0.0161$.

Como trabalho futuro esperamos avaliar novas características para obter resultados AUC melhores. Também, verificar se os padrões se mantém para os demais países da nossa base de dados, ou seja, procurar entender o fluxo migratório de outros países.

# 7. AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

# 8. REFERENCES

[1] G. J. Abel and N. Sander. Quantifying global international migration flows. *Science*, 343(6178):1520–1522, Março 2014.

[2] I. B. de Geografia e Estatística IBGE. Censo demográfico, 2010.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[4] R. Lee. The outlook for population growth. *Science*, 333(6042):569–573, July 29 2011.

[5] M. Lui and T. Baldwin. langid.py: an off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 25–30, 2012.

[6] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor. Theories of international migration: A review and appraisal. *Population and Development Review*, 19(3):431–466, Setembro 1993.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] B. State, M. Rodriguez, D. Helbing, and E. Zagheni. Migration of professionals to the U.S. - evidence from linkedin data. In *Social Informatics - 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, pages 531–543, 2014.

[9] B. State, I. Weber, and E. Zagheni. Studying inter-national mobility through ip geolocation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 265–274, New York, NY, USA, 2013. ACM.

[10] U.S. Bureau of Labor Statistics. Standard occupational classification and coding structure. `http://1.usa.gov/14INxmQ`, February 2010.

[11] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State. Inferring international and internal migration patterns from twitter data. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 439–444, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.