Uma estratégia baseada no filtro de Kalman para monitoramento de epidemias de dengue a partir de redes sociais

Derick M. de Oliveira[2], Roberto C.S.N.P. Souza, Denise E.F. de Brito,
Rodrigo L. Cardoso e Wagner Meira Jr.

Universidade Federal de Minas Gerais, Brasil

Abstract.

Milhares de mensagens são propagadas em redes sociais diariamente, abordando os mais diversos assuntos. Com o grande crescimento desses canais de comunicação, os usuários tendem a discutir variados aspectos do seu cotidiano, preferências e até condições de saúde. A natureza dinâmica e de tempo real dessas mídias, tem transformado tais informações em uma importante ferramenta para pesquisas atuais. Entretanto, um dos grandes desafios na utilização de dados de redes sociais para previsão de eventos está relacionado ao ruído decorrente da ambiguidade da linguagem ou outros fatores presentes nos dados de entrada. Este trabalho propõe uma estratégia baseada no filtro de Kalman (FK) para monitoramento de epidemias de dengue a partir de redes sociais. O FK é uma solução recursiva do método de mínimos quadrados e apresenta-se como uma ferramenta poderosa para estimar estados futuros quando a precisa natureza de um sistema modelado é desconhecida. Os resultados mostram que a nossa proposta superou modelos de regressão linear e polinomial para todas as cidades avaliadas.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data mining

Keywords: Dengue, Filtro de Kalman, Mineração de Dados, Redes Sociais

1. INTRODUÇÃO

Milhares de mensagens são propagadas em redes sociais diariamente, abordando os mais diversos assuntos. Com o grande crescimento desses canais de comunicação, os usuários tendem a discutir variados aspectos do seu cotidiano, preferências e até condições de saúde. A natureza dinâmica e de tempo real dessas mídias tem transformado tais informações em uma importante ferramenta. Diversas pesquisas recentes fazem uso de dados obtidos a partir de redes sociais para mapear e prever eventos reais, como epidemias [Cullota 2010] [Lampos and Cristianini 2010] [Althouse et al. 2011] e fenômenos naturais [Sakaki et al. 2010], entre outros. Apesar do sucesso de tais pesquisas, ainda há espaço para propor melhorias na tentativa de aprimorar a acurácia dessas previsões.

Um dos grandes desafios na utilização de dados de redes sociais para previsão de eventos está relacionado ao ruído, que é inerente ao cenário das redes sociais, seja devido à ambiguidade da linguagem ou a outros fatores como ironia/sarcasmo, muitas vezes presentes no texto. Esse ruído inerente tende a causar variações significativas nos dados prejudicando a capacidade de previsão de diversos modelos. Desta forma, um ponto crucial na utilização de dados advindos dessas mídias sociais para monitoramento e previsão de eventos é a escolha do modelo adequado. Este modelo deve ser capaz de lidar com o ruído inerente aos dados, e também com o fator temporal do evento. Em geral, eventos de interesse, tais como surtos de epidemias, possuem uma localização temporal bem definida e as mensagens relacionadas a esse evento possuem uma correlação temporal com os dados reais. Assim, é importante que o modelo adotado seja capaz de se adequar a esse fator.

Tendo em vista essas características, este trabalho propõe a aplicação do filtro de Kalman (FK)

[2]Este trabalho foi financiado pelo CNPq, Fapemig, Capes e InWeb.

[Kalman 1960] para monitorar epidemias de dengue, ou seja, estimar a incidência da doença, a partir de dados coletados em redes sociais. O FK é um método simples e eficiente, que utiliza observações realizadas a respeito de uma determinada variável ao longo do tempo para recuperar os valores reais das grandezas envolvidas. O FK é capaz de lidar com medidas comprometidas pelo ruído e incertezas para estimar estados futuros mesmo quando a precisa natureza do sistema modelado é desconhecida. Ele tem sido aplicado em uma ampla gama de cenários, tais como navegação marítima, visão computacional e controle de sistemas, entre outros [Beravs et al. 2012].

A estratégia adotada para este trabalho é considerar pessoas postando publicamente em uma rede social como sensores. As mensagens desses usuários, que são relativas ao evento de interesse (especificadas por um conjunto de palavras-chave), funcionam então como indicador da ocorrência e/ou intensidade daquele evento específico. Dessa forma, a previsão da intensidade do evento é feita em função do número de mensagens postadas a respeito do determinado evento. No contexto deste trabalho, o evento de interesse são as epidemias de dengue no Brasil.

2. TRABALHOS RELACIONADOS

Nos últimos anos, diversas pesquisas tem feito uso de dados extraídos de redes sociais para monitoramento e previsão de epidemias e outros eventos do mundo real. Em [Cullota 2010] e [Lampos and Cristianini 2010], mensagens postadas publicamente no Twitter são utilizadas para monitorar taxas de gripe no Reino Unido e nos Estados Unidos, respectivamente. Em [Achrekar et al. 2011], um sistema coleta mensagens do Twitter para avaliar indicadores de epidemia de gripe. Em uma abordagem mais completa, [Paul and Dredze 2011] usa mensagens de redes sociais para uma política de vigilância em saúde pública. Em [Gomide et al. 2011], verifica-se que existe uma alta correlação entre o número de mensagens expressando experiência com a doença e a quantidade real de casos de dengue reportados pelo Ministério da Saúde. Dessa forma, os dados coletados são utilizados para prever níveis de incidência da doença. Este trabalho propõe uma abordagem semelhante a [Gomide et al. 2011], avaliando porém um modelo sensível a fatores temporais para estimar a incidência de dengue nas cidades brasileiras.

3. FILTRO DE KALMAN

Considere o problema geral de estimar um estado $Y_t \in \mathbb{R}^n$, que pode variar temporalmente, a partir de medições $X_t := \{x_t\}_{t=1}^m$. O FK [Kalman 1960] é um estimador que usa dados passados para corrigir seu atual estado. O FK é regido por um conjunto de equações lineares que soluciona o problema de forma recursiva através da minimização do erro quadrático. A cada etapa ele estima o valor do ponto real Y_t e obtém a variação da medição. O FK pode ser dividido em duas etapas principais: *atualização temporal* e *atualização de medida*. Na etapa de atualização temporal, os dados passados são utilizados para prever o valor do estado atual e calcular o ruído dessa medição. Na etapa de atualização da medida, calcula-se o ganho de Kalman e junto com o valor observado pelos sensores X_t ajusta-se o valor obtido na etapa anterior.

Uma fase importante para o uso do FK é adaptar as matrizes do sistema de equações para o contexto da aplicação. No cenário deste trabalho, a medição dos sensores pode se encontrar em uma escala diferente dos valores reais a serem estimados, nesse caso é possível contornar tal situação introduzindo uma uma matriz de medição. A matriz de medição é responsável pela forma como a escala dos sensores se adaptam à escala do fenômeno real.

4. EXPERIMENTOS

Esta seção apresenta resultados obtidos na aplicação do FK para monitoramento das epidemias de dengue no Brasil. Embora a análise completa compreenda 298 cidades (todas as cidades com mais de 100 mil habitantes), devido a restrições de espaço, selecionamos 10 cidades para analisar e discutir os resultados.

4.1 Coleta dos dados e pré-processamento

Com o objetivo de coletar as mensagens relacionadas a casos de dengue, definiu-se como palavras-chave os termos "dengue" e "Aedes". A base de dados é composta de mensagens postadas no Twitter de janeiro de 2011 até Dezembro de 2013. Essas mensagens são restritas a 140 caracteres e a API do Twitter disponibiliza a localização do usuário, quando informada. No Brasil, as políticas de vigilância relativas a epidemias de dengue são delegadas a cada cidade. Assim, torna-se necessário filtrar as mensagens coletadas com base em sua localização. Além disso, na fase de pré-processamento do texto são filtrados acentos, URL's e *stopwords*. Em seguida, com esses dados em mãos, é preciso classificar os tweets em diferentes categorias de acordo com o sentimento e opinião que eles expressam. De maneira similar ao que é feito em [Gomide et al. 2011], os tweets são classificados em 5 categorias: experiência pessoal; ironia/sarcasmo; informação; opinião e campanha. Essa classificação é feita usando algoritmos de classificação associativa sob demanda [Veloso et al. 2006]. Para fins deste trabalho, apenas as mensagens expressando experiência pessoal do usuário com a doença são usadas para a previsão.

4.2 Estratégia de monitoramento baseada no filtro de Kalman

Para a aplicação do FK é preciso então definir as matrizes de transição e de medição. Nesta etapa do trabalho, tais matrizes foram definidas de forma empírica. Utiliza-se uma janela deslizante sobre os dados, observando períodos de epidemia e não epidemia e quantas semanas cada período desse dura, para construir a matriz de transição. A matriz de medição é gerada com razão entre a escala dos dados coletados e a escala dos casos reais apenas no intervalo de aprendizagem. Como a escala dos dados coletados e o número de casos reportados são diferentes, optou-se pela normalização dos dados, dividindo cada série por sua respectiva média.

4.3 Resultados experimentais

Para analisar os resultados foram escolhidas as seguintes cidades: Belo Horizonte, Divinópolis, Florianópolis, Goiania, Macaé, Magé, Petrópolis, Recife, Rio Branco e São Paulo. As cidades foram escolhidas por apresentarem diferentes características em relação ao tamanho da população e nível de incidência de dengue. A tabela I apresenta a correlação entre o número de mensagens expressando experiência pessoal e o total de casos reportado pelo Ministério da Saúde para cada cidade. É interessante notar que em alguns casos a correlação é bastante alta, o que favorece o uso dos dados coletados para monitoramento da doença. Os resultados obtidos pelo filtro de Kalman são comparados com 3 diferentes *baselines*: regressão linear; regressão polinomial de grau 5 e regressão polinomial de grau 7. Para os *baselines* foram usados os dados das semanas 60 a 120 como dados de treinamento e as demais semanas para teste. Com o objetivo de comparar os resultados gerados apresenta-se o RMSE (raiz do erro médio quadrado) calculado nas semanas de teste. Para viabilizar a comparação, o erro do FK é calculado para as mesmas semanas. É importante destacar que os *baselines* apresentados não levam em consideração a informação temporal. A partir da tabela I observa-se que o filtro de Kalman foi superior aos outros modelos em todos os casos. Isso valida a hipótese de que essa informação é de suma importância nesse contexto, justificando assim uma maior investigação a respeito de modelos que incorporam a informação temporal em suas previsões. A figura 1 ilustra a aplicação do FK para 3 das cidades escolhidas para a análise. A partir da figura é possível verificar que o algoritmo capta de forma satisfatória o comportamento da série e, apesar do ruído inerente ao cenário, o resultado é promissor.

5. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou uma aplicação do FK no monitoramento e previsão de epidemias de dengue a partir de dados coletados em redes sociais. O FK se mostrou robusto ao ruído inerente aos dados. Além disso, sua característica de utilizar a informação temporal o torna uma importante ferramenta para o cenário avaliado. O trabalho ainda está em desenvolvimento e como próximo passo pretende-se utilizar todas as categorias de mensagens coletadas, cada uma funcionando como um sensor diferente.

Tabela I. Correlação entre mensagens expressando experiência pessoal e número de casos de dengue; RMSE da aplicação do FK; da regressão linear; da regressão polinomial de grau 5; da regressão polinomial de grau 7.

Cidade	Estado	Correlação	FK	Linear	Poli5	Poli7
Belo Horizonte	MG	0.97030121	0.789358	4.409653	4.850795	5.451746
Divinópolis	MG	0.95933919	1.019865	4.425649	4.540357	5.553869
Florianópolis	SC	0.38141271	1.140881	6.980695	6.511162	6.583076
Goiania	GO	0.64208859	0.915780	5.816474	5.635499	5.776894
Macaé	RJ	0.72845594	1.167469	6.262760	6.311506	6.258486
Magé	RJ	0.35774936	2.320582	18.131181	13.968961	14.233556
Petrópolis	RJ	0.66675824	1.564222	17.089006	15.009303	14.903315
Recife	PE	0.82164502	0.746346	9.345993	7.808170	7.744510
Rio Branco	AC	0.82047324	1.458492	23.035014	10.038641	10.263870
São Paulo	SP	0.86233596	0.747278	8.971457	8.186809	8.369131

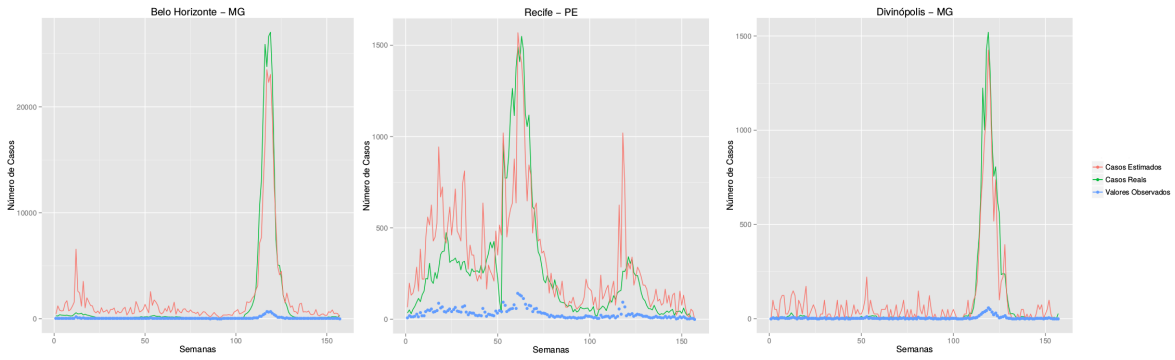

Fig. 1. Resultados obtidos com a aplicação do filtro de Kalman para as cidades de Belo Horizonte, Divinópolis e Recife.

Assim, incrementando a quantidade de dados disponíveis para realizar a previsão espera-se aumentar ainda mais a acurácia do modelo.

REFERÊNCIAS

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., and Liu, B. Predicting flu trends using twitter data. In *Proceedings of the Communications Workshops*. IEEE, 2011.

Althouse, B. M., Ng, Y. Y., and Cummings, D. A. T. Prediction of dengue incidence using search query surveillance. *Plos Neglected Tropical Disease* vol. 8, pp. 1258–4, 2011.

Beravs, T., Podobnik, J., and Munih, M. Three-axial accelerometer calibration using kalman filter covariance matrix for online estimation of optimal sensor orientation. *Instrumentation and Measurement, IEEE Transactions on* 61 (9): 2501–2511, 2012.

Cullota, A. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of 1st workshop on social media analytics*. ACM, 2010.

Gomide, J., Veloso, A., Meira Jr., W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the ACM WebSci Conference*, 2011.

Kalman, R. A new approach to linear filtering and prediction problems. *Transaction of the ASME: Journal of Basic Engineering*, 1960.

Lampos, V. and Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *Proceedings of 2nd Workshop on Cognitive Information Processing*. IAPR, 2010.

Paul, M. J. and Dredze, M. Analyzing twitter for public health. In *Proceedings ICWSM*, 2011.

Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of International Conference on World Wide Web*. ACM, pp. 851–860, 2010.

Veloso, A., Jr., W. M., and Zaki, M. J. Lazy associative classification. In *Proceedings of the International Conference on Data Mining*. pp. 645–654, 2006.