MO829 Teoria dos Jogos Algorítmica

Resumo do artigo "Maximizing the Spread of Influence through a Social Network"

Autores Originais: David Kempe, Jon Kleinberg e Éva Tardos

Kleber Andrade Oliveira 102971

29 de junho de 2017

Resumo

Entender como ideias se difundem em redes sociais é um problema que existe há algum tempo e é estudado por campos distintos. Se uma ideia adotada desencadeia uma sequência de adoções na rede, de que conjunto de indivíduos deve partir a primeira adoção para maximizar o espalhamento da ideia? Formalizaremos o problema com dois modelos para o processo de difusão. Mostraremos que ele é NP-difícil neste contexto. Também oferecemos uma abordagem para uma classe específica de funções, através de um algoritmo guloso que consegue 63% do ótimo eficientemente. A partir do algoritmo, são conduzidos experimentos para comparar este com outros modelos de influência numa rede real de colaboração científica.

1 Problema

Redes sociais são o meio no qual se espalham a informação, as ideias, inovações, produtos ou influência entre os seus membros. Algumas ideias se espalham de maneira profunda pela rede, enquanto outras morrem rapidamente. Para entender como este espalhamento funciona, é necessário captar a dinâmica por trás da adoção de ideias. O quão suscetível um membro da rede é em relação à influência exercida pelos seus amigos ou colegas, ou em outras palavras, o efeito do "boca-a-boca"?

O processo de difusão na rede conta com uma longa história de investigação nas ciências sociais e em outros contextos. O problema, para os efeitos de marketing viral, foi posto de maneira algorítmica por Pedro Domingos e Matthew Richardson na conferência *Knowledge Discovery and Data Mining* em 2001 e 2002 ([Domingos and Richardson 2001] e [Richardson and Domingos 2002]). Neste primeiro momento, trata-se de encontrar a melhor escolha de ações de marketing sobre os jogadores que maximiza adoções mapeadas em variáveis aleatórias numa rede de Markov, isto é, a adoção de cada indivíduo depende da sua vizinhança.

O trabalho que resumiremos aqui vem de outros dois trabalhos anteriores do mesmo trio de autores: [Kempe et al. 2003] e [Kempe et al. 2005]. Nestes, o problema atacado está posto de outra maneira. Considere que um nó da rede é ativado se a ideia ou inovação é adotada. De acordo com alguma regra probabilística de ativação, queremos escolher um conjunto de usuários que inicia o processo capaz de maximizar o número esperado de adoções ao final. Formalmente:

O problema de maximização de influência. Considere uma rede social caracterizada por um grafo direcionado $G(V, E)$. Dado um timestamp t, há um conjunto $A_t \in V$ de nós *ativos*.
De t para $t+1$, mais nós podem se tornar ativos, de acordo com o processo de difusão. Se de t' para $t'+1$ nenhum nó se tornou ativo, diremos que a difusão parou e o resultado do processo é o conjunto $A_{t'}$.
Definimos a *influência* $\sigma(A)$ sobre um conjunto inicial $A = A_0$ como o tamanho esperado do resultado $A_{t'}$, ou $\sigma(A) = \mathbb{E}[|A_{t'}|]$.
O problema de maximização da influência consiste em, dado um parâmetro k, encontrar um conjunto inicial A de tamanho k que maximiza $\sigma(A)$.

Para a regra probabilística de ativação, foram propostos dois modelos: o de limiar e o de cascata. O primeiro, originado na sociologia matemática, supõe que cada vizinho ativo de algum nó exerce uma influência e que, quando a soma destas influências ultrapassa um dado limite, o nó se torna ativo. Já no modelo de cascata, construído por dois teóricos da probabilidade, cada nó tem uma chance de ativar os vizinhos quando se torna ativo. São eles:

Modelo de Limiar. O nó u recebe influência de um vizinho v de acordo com um peso $b_{u,v} \in [0, 1]$. Cada nó tem um limiar θ_u (também entre 0 e 1) e é ativado se e somente se:

$$\sum_{v \in S} b_{u,v} \geq \theta_u,$$

onde S é o conjunto dos nós *ativos* com arestas incidentes em u. Este modelo é chamado de modelo de limiar linear, pois a influência é acumulada linearmente. Os limiares são sorteados uniformemente - se são dados como entrada ou mesmo constantes, tornam o problema NP-difícil de aproximar a menos de um fator multiplicativo de $n^{1-\varepsilon}$ (esta prova é oferecida na seção 3.2).

Modelo de Cascata. Para um conjunto inicial A_0 de nós ativos, o processo segue em passos discretos; se o nó u se tornou ativo no tempo t, então ele pode ativar cada um dos seus vizinhos v com probabilidade $p_{u,v}$. Quando as *probabilidades de ativação* são parâmetros do sistema, i.e. independem do histórico de tentativas de ativação, chamamos este modelo de modelo de cascata independente.
Caso obtenha êxito em ser ativado, um vizinho se tornará ativo no tempo $t+1$. Do tempo $t+1$ em diante, o nó u não poderá ativar mais ninguém. A difusão segue até que não haja mais nenhuma ativação.

Agora, mostraremos que resolver o problema da maximização de influência é NP-difícil em cada modelo.

Teorema 1. *O problema de maximização de influência é* NP*-difícil para o modelo de limiar linear.*

Prova. Considere uma instância do problema de *cobertura de vértices* definida pelo grafo não-direcionado $G(V, E)$ e um inteiro $k > 0$. Reduziremos este problema ao de maximização de influência. Para obter uma instância correspondente do problema de maximização de influência, faça:

- Todas as arestas $\{u, v\}$ vão nas duas direções
- $b_{u,v} = \frac{1}{grau(v)}$
- Se há uma cobertura de vértices S de tamanho k, tome o conjunto inicial $A = S \implies \sigma(A) = n$ deterministicamente

Agora iremos de uma instância a outra no sentido oposto. Se algum par de nós adjacentes não tem nó no conjunto inicial A, então conforme limiares próximos de 1 são sorteados, estes nós poderão ficar inativos $\implies \sigma(A) < n$. Nessa situação, ou não há cobertura S na instância do problema de cobertura de vértices, ou A não é ótimo $\therefore$ Contradição ■

A seguir, além de provar que resolver o problema de maximização de influência pelo modelo de cascata, também demonstraremos a dificuldade da $(1 - \frac{1}{e})$-aproximação.

Teorema 2. *A aproximação do problema de maximização de influência por um fator melhor que* $1 - \frac{1}{e}$ *é* NP*-difícil pelo modelo de cascata independente.*

Prova: Tome uma instância do problema de *máxima cobertura* definida pelos subconjuntos S_j e elementos $u_i \in U$ e parâmetro k, com $j \in \{1, ..., m\}$, $i \in \{1, ..., n\}$ e $k < n < m$. Reduziremos este problema ao de maximização de influência. Vamos construí-la do modelo de cascata independente. Definimos um grafo bipartido de $m + n^2$ nós de maneira que:

- Para cada subconjunto S_j, há um nó j
- Para cada elemento u_i, há n nós $v_1, v_2, ..., v_n$
- Se $u_i \in S_j$, então existe uma aresta direcionada $\{j, v_l\}, \forall l$ e $p_{j,v_l} = 1$

Ao escolhermos o conjunto inicial A para X na partição dos S_J e $T \subseteq U$ na outra partição, onde X é formado por k subconjuntos S_j e T união dos v_l adjacentes aos S_j , obtemos deterministicamente $k + n|T|$ nós ativos. Dessa maneira, r elementos serão cobertos por k subconjuntos apenas se $k+nr$ nós são ativados por um conjunto inicial de tamanho k na instância do modelo de cascata. No sentido oposto, se já temos um conjunto inicial A podemos dizer que:

Os nós j de A correspondentes aos S_j são tantos quanto os k subconjuntos S_j; se mais subconjuntos fossem selecionados para a cobertura, mais nós seriam ativados. As arestas farão com que estes nós j ativem mais $n|T|$ nós com probabilidade 1,

onde T são os elementos cobertos pelos S_j.

Assim, se r nós são cobertos, exatamente $k + nr$ nós foram ativados na instância do modelo de cascata. Como o problema da cobertura máxima é difícil de se aproximar por um fator melhor que $1 - 1/e$, segue que o problema da maximização de influência também o é. ■

Não é possível avaliar a função de influência $\sigma(A)$ em tempo polinomial (em particular, este é um problema #P-completo). Mas é possível estimá-la com alta probabilidade, após um número suficiente de simulações. Vamos enunciar que, se esta função pertence a à classe das funções submodulares, podemos aproximá-la por um fator de $(1 - \frac{1}{e})$ pelo algoritmo 1.

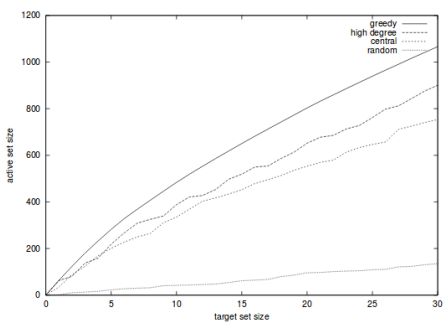
Algorithm 1 Aproximação Gulosa

1: Comece com $A = \emptyset$.
2: **enquanto** $|A| \leq k$ **faça**
3: Para cada nó x, obtenha uma $(1 \pm \varepsilon)$-aproximação de $\sigma(A \cup \{x\})$ com alta probabilidade através da amostragem repetida.
4: Adicione o nó com a melhor estimativa para $\sigma(A \cup \{x\})$ em A.
5: **termine enquanto**
6: Retorne A.

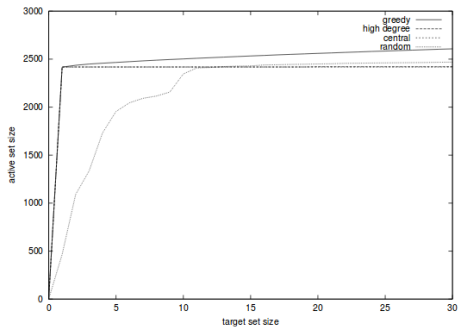
Teorema 3: Nemhauser-Wolsey-Fisher (1978) *Se $\sigma(\cdot)$ é uma função submodular, monótona e não-negativa, então o algoritmo 1 produz um conjunto A de k elementos tal que*

$$\sigma(A) \geq (1 - 1/e) max_{|B|=k} \sigma(B)$$

Este é um resultado da área de otimização de funções submodulares que os autores não se preocupam em mostrar. São oferecidos vários casos nos quais a função de influência é submodular. O trabalho também realizou um experimento com este algoritmo guloso natural. Numa rede de colaboração científica real, com 10748 nós e quase 53000 arestas, a performance do algoritmo é comparada com três heurísticas interessantes da área, incluindo o sorteio uniformemente aleatório do conjunto inicial.



(a) Modelo de limiar.



(b) Modelo de cascata com $p = 0,1$.

Figura 1: Total de ativações por tamanho k do conjunto inicial para o (a) modelo de limiar e (b) modelo de cascata.

Referências

[Domingos and Richardson 2001] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM.

[Kempe et al. 2003] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

[Kempe et al. 2005] Kempe, D., Kleinberg, J. M., and Tardos, É. (2005). Influential nodes in a diffusion model for social networks. In *ICALP*, volume 5, pages 1127–1138. Springer.

[Richardson and Domingos 2002] Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM.