Inferência bayesiana para modelos de classificação de Plackett-Luce

John Guiver joguiver@microsoft.com
Edward Snelson esnelson@microsoft.com
Microsoft Research Limited, 7 J J Thomson Avenue, Cambridge CB3 0FB, Reino Unido

Abstrair

Este artigo fornece um método bayesiano eficiente para inferir os parâmetros de um modelo de classificação de Plackett-Luce. Tais modelos são distribuições parametrizadas sobre classificações de um conjunto finito de objetos e têm sido tipicamente estudados e aplicados na literatura psicométrica, so-ciométrica e econométrica. O esquema de referência é uma aplicação de Power EP (propagação de expectativa). O esquema é robust e pode ser facilmente aplicado a conjuntos de dados em grande escala. O algoritmo de inferência se estende a variações do modelo básico de Plackett-Luce, incluindo classificações parciais. Mostramos uma série de vantagens da abordagem EP sobre o método tradicional de máxima verossimilhança. Aplicamos o método para agregar classificações de pilotos de corrida da NASCAR ao longo da temporada de 2002 e também para classificações de gêneros de filmes.

1. Introdução

Problemas Envolvendo Classificad Listas de Itens are amplamente difundidos e são passíveis da aplicação de métodos de aprendizado de máquina. Um exemplo é o subcampo de "aprender a classificar" no cruzamento de aprendizado de máquina e recuperação de informações (ver, por exemplo, Joachimset al., 2007). Outro exemplo é a agregação de classificação e meta-pesquisa (Dwork et al., 2001). A modelagem adequada de observações na forma de itens classificados exige que consideremos distribuições de probabilidade parametrizadas sobre classificações. Esta tem sido uma área de estudo em estatística há algum tempo (ver Marden, 1995 para uma revisão), mas grande parte desse trabalho não chegou à comunidade de aprendizado de máquina. Neste artigo, estudamos uma distribuição de classificação específica, a Plackett-Luce, que tem algumas propriedades muito boas. Embora a estimação de parâmetros no Plackett-Luce pode ser alcançado por meio de estimativa de máxima verossimilhança (MLE) usando métodos MM (Hunter, 2004), desconhecemos um tratamento bayesiano eficiente. Como mostraremos, o MLE é problemático para dados esparsos devido ao sobreajuste e nem mesmo pode ser encontrado para algumas amostras de dados que ocorrem em situações reais. Dados esparsos no contexto da classificação são um cenário comum para algumas aplicações e são caracterizados por ter um pequeno número de observações e um grande número de itens para classificar (Dwork et al., 2001), ou cada observação individual pode classificar apenas alguns dos itens totais. Portanto, desenvolvemos um procedimento bayesiano eficiente de referência aproximada para o modelo que evita o sobreajuste e fornece estimativas de incerteza adequadas nos parâmetros.

A distribuição Plackett-Luce deriva seu nome do trabalho independente de Plackett (1975) e Luce (1959). O Axioma da Escolha de Luce é um axioma geral que governa as probabilidades de escolha de uma população de 'escolhas', escolhendo um item de um subconjunto de um conjunto de itens. O axioma é melhor descrito por uma ilustração simples. Suponha que o conjunto de itens seja {A, B, C, D} e suponha que as probabilidades correspondentes de escolher a partir desse conjunto sejam (pA, pB, pC, pD). Agora considere um subconjunto {A, C} com probabilidades de escolha (qA, qC). Então o axioma da escolha de Luce afirma que qA/qC = pA/pC. Em outras palavras, a razão de probabilidade de escolha entre dois itens é independente de quaisquer outros itens do conjunto.

Suponha que consideremos um conjunto de itens e um conjunto de probabilidades de escolha que satisfaçam o axioma de Luce, e consideremos escolher um item de cada vez do conjunto, de acordo com as probabilidades de escolha. Essas amostras fornecem uma ordenação total dos itens, que pode ser considerada como uma amostra de uma distribuição sobre todas as ordenações possíveis. A forma de tal distribuição foi considerada pela primeira vez por Plackett (1975) para modelar probabilidades em uma corrida de cavalos K.

O modelo de Plackett-Luce é aplicável quando cada observação fornece uma classificação completa de todos os itens, ou uma classificação parcial de apenas alguns dos itens, ou uma classificação dos poucos itens mais importantes (consulte a seção 3.5 para o

Aparecendo em Anais da 26ª Conferência Internacional sobre Aprendizado de Máquina, Montreal, Canadá, 2009.

dois últimos cenários). As aplicações da distribuição de Plackett-Luce e das suas extensões têm sido bastante variadas, incluindo corridas de cavalos (Plackett, 1975), classificação documental (Cao et al., 2007), avaliação da procura potencial de automóveis eléctricos (Beggs et al., 1981), modelização de eleitorados (Gormley & Murphy, 2005) e modelização de preferências alimentares em vacas (Nombekela et al., 1994).

A inferência dos parâmetros da distribuição de Plackett-Luce é normalmente feita por estimativa de máxima verossimilhança (MLE). Hunter (2004) descreveu um método efficientMLE baseado em um algoritmo de minorização/maximização (MM). Nos últimos anos, novos e poderosos algoritmos de passagem de mensagens foram desenvolvidos para fazer inferência bayesiana determinística aproximada em grandes redes de crenças. Esses algoritmos são tipicamente precisos e altamente escaláveis para grandes problemas do mundo real. Minka (2005) forneceu uma visão unificada desses algoritmos e mostrou que eles diferem apenas pela medida da divergência de informações que eles minimizam. Aplicamos Power EP (Minka, 2004), um algoritmo neste framework, para realizar inferência bayesiana para modelos de Plackett-Luce.

Na seção 2, examinamos mais detalhadamente a distribuição de Plackett-Luce, motivando-a com algumas interpretações alternativas. Na seção 3, descrevemos o algoritmo em um nível de detalhe em que deve ser possível para o leitor implementar o algoritmo no código, fornecendo derivações quando necessário. Na seção 4, aplicamos o algoritmo aos dados gerados a partir de uma distribuição conhecida, a uma agregação dos resultados da corrida da NASCAR de 2002 e também à classificação de gêneros no conjunto de dados MovieLens. A Seção 5 fornece breves conclusões.

2. Modelos Plackett-Luce

Uma boa fonte para o material desta seção e para distribuições de classificação em geral é o livro de Mar-den (1995). Considere um experimento em que N juízes são solicitados a classificar K itens e assumem que não há empates. O resultado do experimento é um conjunto de N rankings{y(n) ≡ (y(n)1 , . . . , y(n)K) | n = 1, . . . , N } onde uma classificação é definida como uma permutação dos índices de classificação K; Em outras palavras, o juiz N classifica o item I na posição Y(N)I (onde a classificação mais alta é a posição 1). Cada classificação tem uma ordenação associada ω(n) ≡ (ω(n)1 , . . . , ω(n)K), onde uma ordenação é definida como uma permutação dos K índices de itens; Em outras palavras, o juiz N coloca o item ω(n)iin po-sition i. Rankings e ordenações estão relacionados por (drop-ping o índice de juízes) ωyi = i, yωi = i.O modelo de Plackett-Luce (PL) é uma distribuição superclassificada y que é melhor descrita em termos da ordenação associada ω. É parametrizado por um vetor v = (v1, . . . , vn) onde vi ≥ 0 está associado ao itemIndex I:

$$P\,L(\omega \mid v) = \prod_{k=1,\ldots,K} fk(v) \qquad (1)$$

onde

$$fk(v) \equiv fk(v\omega k\ ,\ .\ .\ .\ ,\ v\omega K)\ , \quad \frac{v\omega k v\omega k + \cdots}{+\ v\omega K} \qquad (2)$$

2.1. Interpretação do modelo de vaso

A metáfora do modelo de vaso se deve a Silverberg (1980). Considere um experimento de vários estágios em que, em cada estágio, estamos tirando uma bola de um vaso de bolas coloridas. O número de bolas de cada cor é proporcional ao vωk . Um vaso difere de uma urna apenas por ter um número infinito de bolas, permitindo assim proporções não racionais. No primeiro estágio, uma bola ω1 é retirada do vaso; A probabilidade dessa seleção é f1(v). No segundo estágio, outra bola é sorteada - se for da mesma cor da primeira, coloque-a de volta e continue tentando até que uma nova cor ω2 seja selecionada; A probabilidade desta segunda seleção é f2(v). Continue pelas etapas até que uma bola de cada cor tenha sido selecionada. É claro que a equação 1 representa a probabilidade dessa sequência. A interpretação do modelo de vaso também fornece um ponto de partida para extensões do modelo PL básico detalhado por Marden (1995), por exemplo, capturando a intuição de que os juízes fazem julgamentos mais precisos nos escalões mais altos.

2.2. Interpretação de Thurston

Um modelo de Thurstoniano (Thurstone, 1927) assume uma variável de pontuação aleatória não observada xi (normalmente independente) para cada item. Tirar das distribuições de pontuação e classificar de acordo com a pontuação amostrada fornece uma classificação de amostra - portanto, as pontuações de distribuição induzem uma distribuição sobre as classificações. Um resultado fundamental, devido a Yellott (Yellott, 1977), diz que se as variáveis de pontuação são independentes e as distribuições de pontuação são idênticas, exceto por suas médias, então as distribuições de pontuação dão origem a um modelo PL se e somente se as pontuações são distribuídas de acordo com uma distribuição de Gumbel β μ β μ.

$$G\,(x \mid m, b) = e{-}z \qquad (3)$$

$$g(x \mid m, b) = \frac{z\beta}{e{-}z} \qquad (4)$$

onde z(x) = e− x−μβ . Para um β fixo, g(x | μ, β) é a distribuição familiar anexponencial com o parâmetro natural v = e μβ que tem um conjugado de distribuição gama

prévio. O uso da notação v para este parâmetro natural é deliberado - verifica-se que vi = eμiβ é o parâmetro P-L para o i-ésimo item na distribuição de classificação induzida pelo modelo de Thurstoniano com distribuições de pontuação g(xi | μi, β). O sistema de classificação TrueSkill (Herbrich et al., 2007) é baseado em um modelo de Thurston com uma distribuição de pontuação gaussiana. Embora este modelo não satisfaça o Axioma da Escolha Luce, foi aplicado num sistema comercial de classificação em linha em larga escala com muito sucesso.

2.3. Estimativa da máxima verossimilhança

A maneira típica de ajustar um modelo PL é por estimativa de verossimilhança máxima (MLE) dos parâmetros v.Hunter (2004) descreve uma maneira de fazer isso usando um algoritmo de minimização / maximização (MM) (maximização de expectativa (EM) é um caso especial de um algoritmo MM), que se mostra mais rápido e mais robusto do que o método Newton-Raphson mais padrão. Fur-thermore Hunter fornece código MATLAB para este al-gorithm, juntamente com um exemplo interessante de aprender um PL para classificar os pilotos da NASCAR em toda a temporada de corridas de 2002. Retomamos este exemplo mais adiante na seção 4.2, demonstrando que, embora o MLE funcione bem em algumas configurações, ele se sobreajustará quando houver dados esparsos. Além disso, o algoritmo MM requer uma forte suposição (Suposição 1 de Hunter, 2004) para garantir a convergência: em todas as partições possíveis dos indivíduos em dois subconjuntos não vazios, alguns indivíduos no segundo conjunto são mais altos do que alguns indivíduos no primeiro conjunto pelo menos uma vez. Como veremos nos dados da NASCAR, essa suposição muitas vezes não é satisfeita em exemplos reais envolvendo dados esparsos e, de fato, o algoritmo MM não converge.

3. Inferência Bayesiana

Esta seção faz uso intenso das idéias, notação e algoritmos em (Minka, 2005), e não há espaço para resumi-los aqui. Portanto, embora demos uma descrição completa de nosso algoritmo, muitas informações de fundo de (Minka, 2005) são assumidas.

Suponha que temos um conjunto de ordenações completas observadasΩ = {ω(n)}. Gostaríamos de inferir os parâmetros do modelo aP-L, colocando prioris adequados neles. Pelo teorema de Bayes, a distribuição posterior sobre os parâmetros é proporcional a:

$$p(v) \equiv p(v \mid \Omega) = \prod_{n=0,\ldots,N} \prod_{k=1,\ldots,K} f(n)\, k\,(v) \quad (5)$$

onde f (0)k é um prior, e o restante f (n) k são como na equação (2), mas agora indexada também por um dado. Como os vi são valores positivos, é natural atribuir-lhes priores de distribuição gama, e isso é reforçado pela discussão na seção 2.2. Portanto, vamos supor que, para cada k,

$$f_{(0)k} = \mathrm{Gam}(vk \mid \alpha 0, \beta 0) \quad (6)$$

Em geral, estamos interessados em recuperar os marginais de p(v). Estaremos inferindo uma aproximação totalmente fatorada para p (v), de modo que os marginais serão uma saída direta do algoritmo de inferência. Quando a aproximação é totalmente fatorada, a passagem de mensagens tem uma interpretação gráfica como um gráfico de fatores, com mensagens passando entre variáveis e fatores.

3.1. Preliminares

O algoritmo de passagem de mensagens descrito abaixo fará uso de versões normalizadas e não normalizadas da distribuição Gama:

$$\mathrm{UGam}(x \mid \alpha, \beta)\ ,\ x\alpha-1e-\beta x \quad (7)$$

$$\mathrm{Gam}(x \mid a, b)\ ,\ \underline{\mathrm{UGam}}(x \mid a, b) \quad (8)$$

onde, para a versão normalizada, exigimos α > 0e β > 0. α é o parâmetro de forma e β é o parâmetro de taxa (ou seja, 1/escala). A família UGam é útil, pois nos permite lidar, de forma consistente, com distribuições inadequadas.

3.2. A fatoração

Vamos aproximar p(v) como um produto totalmente fatorado de distribuições gama:

$$p(v) \approx q(v) = \prod_{i=1,K} qi(vi) = \prod_{um} \tilde{}fa(v) \quad (9)$$

a = (n, k) resume o índice duplo de datum e rank em um único índice, de modo a manter a notação suc-cinct e consistente com (Minka, 2005), e qi (vi) = UGam (vi | αi, βi). Seguimos o tratamento de passagem de mensagens em (Minka, 2005, seção 4.1). Os fatores $\tilde{}fa(v)$, que aproximam os fatores P-L fa(v), transformam-se totalmente em mensagens ma→i do fator A para a variável vi:

$$\tilde{}fa(v) = \prod_{eu} ma\to i(vi) \quad (10)$$

onde ma→i(vi) = UGam(vi | αai, βai). Colete todos os termos envolvendo a mesma variável vi para definir mensagens da variável vi para o fator a

$$mi\to a(vi) = \prod_{b6=a} mb\to i(vi) \quad (11)$$

A lógica do algoritmo de passagem de mensagens é melhorar os fatores de aproximação $\tilde{}fa(v)$ um a um

tempo sob a suposição de que a aproximação do resto do modelo é boa - ou seja, assumindo que q \ a (v) = q (v) ∫ fa (v) é uma boa aproximação de p \ a (v) = p (v) / fa (v). Observe que

$$q\backslash a(v) = \prod_{b6=a} \prod_{eu} mb\to i(vi) = \prod_{eu} mi\to a(vi) \qquad (12)$$

3.3. A atualização da mensagem

A quantidade chave que precisamos calcular é:

$$q'i(vi) = proj[ma\to i(vi)1-\alpha mi\to a(vi) \times \int_{v\backslash vi} dv\, fa(v)\, \alpha \prod_{j6=i} ma\to j(vj)1-\alpha mj\to a(vj)] \qquad (13)$$

onde proj denota projeção K-L, e onde α é o parâmetro de divergência α que podemos escolher para tornar nosso problema tratável.1 Definir

$$ma\to j(vj)1-\alpha j\to a(vj) = UGam(vj \mid \gamma aj, \delta aj) \qquad (14)$$

O algoritmo de inferência falha se UGam(vj | γaj, δaj)se tornar impróprio para qualquer j — no entanto, não vimos isso acontecer na prática. Mensagens individuais, no entanto, são permitidas e muitas vezes se tornarão impróprias. Assumindo que UGam(vj | γaj, δaj) é adequado, podemos substituí-lo equivalentemente por sua versão normalizada gaj , Gam(vj | γaj, δaj) - isso simplifica as derivações. Escolhemos α = −1 como nossa α-divergência. Isso é necessário para tornar a integral tratável, já que o verdadeiro fator fac-tor fa(v) é invertido, levando a uma soma de produtos de integrais univariadas de distribuições gama.

3.3.1. Projeção para um fator P-L

Fixando a atenção em um fator específico fa (v) onde a = (n, k), com ordenação observada ω, temos fa (v) = vωk / ∑l = k ... K vωl. Assim, com uma α-divergência de −1,fa(v)α = ∑l=k... K(vωl/vωk). Ao calcular q'i(vi), se vi não aparecer em fa(v),então ma→i(v) = 1, então podemos restringir os cálculos a quando i = ωr para algum r. Observe, também, que podemos ignorar qualquer fator em ∏j6=i gaj(vj) para o qual j 6= ωl para algum l, porque eles se integram a 1. Consideraremos os casos r = k e r 6 = k separadamente.

1Para uma verdadeira projeção K-L, precisamos combinar as características da distribuição Gama - ou seja, ln(vi) e vi.No entanto, aproximaremos isso apenas combinando os dois primeiros momentos para evitar o procedimento iterativo não linear necessário para recuperar os parâmetros gama de E(ln(vi)) e E(vi).

Caso 1 (i = ωk):

$$\begin{aligned} &gai\,(vi)\int_{v\backslash vi} fa(v)-1\prod_{j6=i} gaj(vj)dv \\ &= gai(vi)\sum_{l=k\,...\,K}\int_{v\backslash vi}(v\omega l/v\omega k)\prod_{j6=i} gaj(vj)dv \\ &= desaparecido(v)\,(1+1v\sum_{l=k+1...K}\int v\omega l\,gaal(v\omega l)dv\omega l) \\ &= desaparecido(v)\,(1+1v\sum_{l=k+1...K}\text{Por exemplo,}) \\ &= (\frac{\Delta aiyai}{-1}\sum_{l=k+1...K}\text{Por exemplo,})Gam(vi \mid \gamma ai-1, \delta ai) \\ &\quad + Gam(vi \mid \gamma ai, \delta ai) \end{aligned} \qquad (15)$$

Caso 2 (i = ωr, r 6 = k):

$$\begin{aligned} &gai\,(vi)\int_{v\backslash vi} fa(v)-1\prod_{j6=i} gaj(vj)dv \\ &= (1+\frac{\Delta aiyai}{-1}\sum_{l=k+1...K,l6=r}\text{Por exemplo,})Gam(vi \mid \gamma ai, \delta ai) \\ &\quad + (\frac{\delta akya}{k-1}\frac{ya}{i\delta ai})Gam(vi \mid \gamma ai+1, \delta ai) \end{aligned} \qquad (16)$$

Observe que ambos se reduzem à forma geral

$$c\cdot Gam(vi \mid a, b) + d\cdot Gam(vi \mid a+1, b) \qquad (17)$$

Os dois primeiros momentos para uma expressão na forma da equação (17) são facilmente mostrados como:

$$E(vi) = \frac{ca + d(a+1)b}{(c+d)} E(v2i) = \frac{ca(a+1) + d(a+1)(a+2)b2}{(c+d)} \qquad (18)$$

A projeção não normalizada pode então ser calculada como

$$q'i(vi) = UGam(vi \mid \frac{E(vi)2E(v2i) - E(vi)2}{,}\ \frac{E(vi)E(v2i) - E(vi)2}{}) \qquad (19)$$

3.3.2. Atualização da mensagem para um fator P-L

Como um esclarecimento para (Minka, 2005), e correspondendo à descrição original do Power EP em (Minka, 2004), as atualizações marginais e as atualizações de mensagens são:

$$qnewi(vi) = qi(vi)2/q'i(vi) \qquad (20)$$

$$mnewa\to i = \frac{qnewi(vi)}{mi\to a(vi)} \qquad (21)$$

3.4. Resumo do algoritmo

1. Inicializar ma→i(vi) para que todos os a,i sejam uniformes, exceto quando a = (0, k), correspondendo às mensagens anteriores constantes. Definimos cada um deles como um broadprior de UGam(vi | 3.0, 2.0).

2. Repita até que todos os ma→i(vi) convenjam:(a) Escolha um fator a = (n, k). (b) Calcule as mensagens no fator usando a equação (11). (c) Calcule as projeções q′i(vi) usando a equação (19) por meio de equações. (15), (16), (17), (18). (d) Atualizar as mensagens de saída do fator usando as equações (20) e (21).

Observe que os marginais podem ser recuperados a qualquer momento byqi(vi) = ∏a ma→i. Como há um grau de liberdade no vi, o parâmetro de taxa dos marginais pode ser dimensionado coletivamente de modo que, por exemplo, as médias da soma do vi para um valor especificado; Isso é útil, por exemplo, se você estiver tentando identificar parâmetros conhecidos como fazemos na seção 4.1. Finalmente, não há espaço para mostrar o cálculo da evidência aqui, mas pode ser facilmente derivado das mensagens não normalizadas convergidas conforme mostrado (Minka, 2005, seção 4.4). Este cálculo das evidências é uma vantagem adicional da abordagem totalmente bayesiana, pois nos permite construir modelos de mistura com diferentes números de componentes de mistura e avaliar seus fatores de Bayes (seleção de modelo).

3.5. Rankings incompletos

Uma das boas propriedades da distribuição PL é que ela é internamente consistente: a probabilidade de uma ordenação particular não depende do subconjunto do qual os indivíduos são assumidos como extraídos (ver Hunter, 2004 para um esboço de uma prova e relação com o Axioma da Escolha de Luce). Suponha que temos dois conjuntos de itens A e B onde B ⊂ A. Isso significa que a probabilidade de uma determinada ordenação dos itens em B, marginalizando todas as possíveis posições desconhecidas dos itens restantes em A, é exatamente a mesma que a probabilidade P-L de simplesmente ordenar esses itens em B completamente independente de A. A consequência da consistência interna é que cada dado pode ser uma ordenação incompleta do conjunto total de itens e, no entanto, eles ainda podem ser combinados de forma consistente, com a probabilidade de cada dado sendo um simples produto dos fatores f (n) k dos itens que são classificados nesse dado particular. Isso é extremamente útil na prática, pois em muitas aplicações um "juiz" individual pode classificar apenas alguns dos itens. Um exemplo são os dados da NASCAR da seção 4.2, em que um diferente, mas sobreposto, conjunto de pilotos competem em cada corrida. Em termos de nosso algoritmo de inferência, o caso de classificação incompleto diminui simplesmente o número de fatores que devem ser incluídos no gráfico de passagem de mensagens.

Outra variação é onde as classificações dos primeiros S foram dadas. Um exemplo pode ser quando os usuários são solicitados a classificar seus 10 principais filmes, ou em meta-pesquisa em que cada mecanismo de pesquisa relata seus 10 principais (ou 100 melhores, etc.) documentos para uma determinada consulta. Novamente, essa situação pode ser tratada de forma consistente e, neste caso, os fatores f (n) k para os quais k > S são removidos da probabilidade(1). Isso equivale a marginalizar todas as posições desconhecidas dos outros itens, mas assumindo que eles estão classificados em algum lugar abaixo dos itens S principais.

4. Exemplos

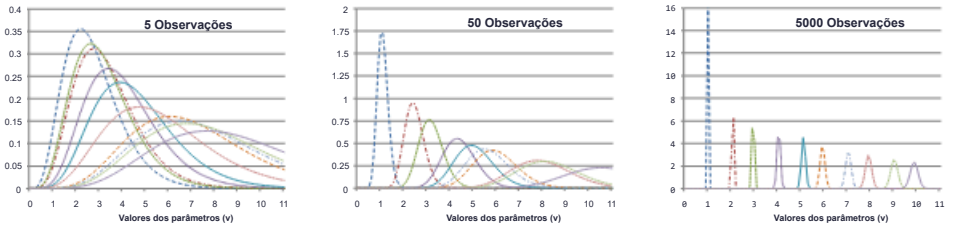
4.1. Inferir parâmetros conhecidos

Para verificar se o algoritmo está fazendo a coisa certa, podemos gerar dados de uma distribuição PL com parâmetros conhecidos e, em seguida, tentar inferir os parâmetros. A Figura 1 mostra os parâmetros inferidos de um modelo PL com vetor de parâmetro v = (1,0, 2,0, . . . , 10,0). Os marginais em 1a são inferidos a partir de 5 observações, os de 1b de 50 e os de 1c de 5000 observações. Como esperado, a dispersão dos marginais diminui à medida que os dados aumentam, e os verdadeiros valores dos parâmetros são razoavelmente representados pelos marginais.

É interessante observar que as estimativas se tornam menos certas para parâmetros maiores. Isso talvez seja esperado, já que a proporção v10/v9 neste exemplo é muito menor do que a proporção de v2/v1, então as escolhas de classificação superior são decisões menos claras do que as inferiores.

4.2. Classificação dos pilotos de corrida da NASCAR

Hunter (2004) realiza um estudo de caso de ajuste de um modelo P-L aos resultados das corridas de carros da temporada de 2002 da NASCAR. Nesta seção, também estudamos esses dados porque servem como uma comparação e destacam uma série de vantagens de nossa abordagem totalmente bayesiana para o método MM MLE. A temporada de 2002 da NASCAR consistiu em 36 corridas nas quais um total de 87 pilotos diferentes competiram. No entanto, qualquer corrida envolveu apenas 43 pilotos. Isso variou de alguns pilotos competindo em todas as corridas e alguns apenas em uma corrida na temporada. Este é, portanto, um bom exemplo do caso de classificação incompleta discutido na seção 3.5. Conforme discutido na seção 2.3, o algoritmo MM de Hunter requer uma suposição bastante forte para convergência. Em muitos casos, e de fato neste caso, essa suposição não será satisfeita. Nos dados da NASCAR, 4 pilotos ficaram em último lugar em todas as corridas em que participaram, violando assim essa suposição. Ali-



a) 5 Observações b) 50 observações c) 5000 observações

Figura 1. Distribuições marginais de parâmetros PL inferidas a partir de dados gerados a partir de PL (ω | (1.0, 2.0, . . . , 10.0))

antes que Hunter tivesse que simplesmente remover esses drivers do modelo. Em contraste, nosso método bayesiano pode ser aplicado a todos os dados sem problemas devido aos prioris adequados que são colocados nos parâmetros PL. No entanto, para efeitos de comparação directa com o seu trabalho, seguimos este procedimento e removemos estes condutores de modo a utilizarem exactamente o mesmo conjunto de dados, com um total de 83 condutores.

A Tabela 1 mostra os 10 principais e inferiores motoristas ordenados por posição média, bem como sua classificação atribuída por ambos Para máxima verossimilhança, a ordenação é feita pelo parâmetro MLE PL e, para EP, a ordenação é feita pelo parâmetro PL médio. Existem algumas diferenças claras entre os dois métodos. O método MLE coloca Jones e Pruett em primeiro e segundo lugar, respectivamente - isso certamente se relaciona com sua posição média muito alta (numericamente baixa). No entanto, eles só correram em uma corrida cada, em comparação com alguns pilotos que correram toda a temporada de 36 corridas. Este é um exemplo do sobreajuste do algoritmo MLE - uma corrida não é evidência suficiente para julgar a habilidade desses pilotos, e ainda assim o MLE os coloca bem no topo. Em contraste, a inferência EP coloca esses drivers no meio do ranking, e também seus parâmetros PL têm alta incerteza em comparação com outros drivers. Com mais evidências, é possível que esses motoristas subam no ranking. O método EP classifica Mark Martin em primeiro lugar, seguido por Rusty Wallace: pilotos que correram todas as 36 corridas. Da mesma forma, na parte inferior da tabela, o método EP coloca Morgan Shepherd na parte inferior, em vez de alguns dos outros pilotos com posição média semelhante, mas que correram em apenas uma ou duas corridas. Morgan Shepherd correu em 5 corridas e, portanto, evidências suficientes se acumularam de que ele consistentemente se sai mal. Observe que, mesmo quando o número de corridas disputadas é o mesmo (por exemplo, Martin, Stewart, Wallace, Johnson correu 36 corridas), nem MLE PL ou EP PL equivalem a simplesmente ordenar por lugar médio - o modelo PL leva em consideração exatamente quem está correndoem cada corrida: é melhor ter vencido em uma corrida cheia de bons pilotos do que em uma corrida de pilotos ruins.

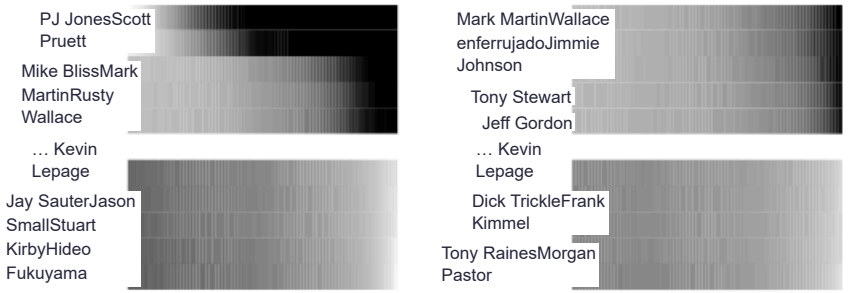
A Figura 2 é uma maneira alternativa de visualizar as inferências sobre os pilotos selecionados da NASCAR - os 5 pilotos superiores e inferiores, conforme ordenado pelo MLE (2a) e pelo EP (2b). Em vez de mostrar os parâmetros PL inferidos, que são um pouco difíceis de interpretar em si mesmos, mostramos as distribuições marginais de classificação inferidas implícitas nos parâmetros PL inferidos para cada driver. Esta é a visualização em escala de cinza da probabilidade de cada piloto chegar a um determinado lugar em uma corrida envolvendo todos os 83 pilotos. Como vemos, o enredo da MLE é dominado pelo excesso de ajuste para os dois pilotos PJ Jones e Scott Pruett, que têm distribuições altamente distorcidas para os escalões superiores. Em contraste, o EPplot mostra distribuições marginais de classificação muito mais amplas e razoáveis, refletindo o fato de que, mesmo para os melhores pilotos, há alta incerteza em qualquer corrida sobre onde eles se colocarão.

4.3. Classificação dos gêneros cinematográficos

O conjunto de dados do MovieLens foi coletado e é de propriedade do Projeto de Pesquisa GroupLens da Universidade de Minnesota. O conjunto de dados consiste em 100.000 classificações (1-5) de 943 usuários em 1682 filmes. Estes dados são interessantes na medida em que (a) fornecem informações demográficas simples para cada utilizador e (b) fornecem informações sobre cada filme como uma lista de vectores de género — um filme pode ter mais do que um género — por exemplo, a comédia romântica. Obtivemos dados de classificação criando, para cada usuário, uma classificação média de cada gênero em todos os filmes vistos pelo usuário em particular. Cada usuário avaliou pelo menos 20 filmes para que cada um veja muitos gêneros, mas não há garantia de que um usuário verá todos os tipos de gênero. Isso significa que os rankings de gênero são parciais e a ausência de um determinado gênero em uma observação não é uma indicação de que um usuário está dando a ele um

Tabela 1. Classificações PL posteriores para os dez primeiros e últimos pilotos da NASCAR de 2002, conforme dado pela posição média. As estimativas de parâmetros v foram normalizadas para somar 1 para MLE e EP, de modo que sejam comparáveis (para EP, sua soma média é 1). A coluna EP SDev v mostra o desvio padrão da distribuição gama posterior sobre v.

Motorista	Corridas	De. Lugar	Classificação MLE	MLE v	Classificação EP	EP Média vEP SDev v	
PJ Jones	1	4.00	1	0.1864	18	0.0159	0.0079
Scott Pruett	1	6.00	2	0.1096	19	0.0156	0.0078
Marcos Martin	36	12.17	4	0.0235	1	0.0278	0.0047
Tony Stewart	36	12.61	7	0.0184	4	0.0229	0.0040
Wallace enferrujado	36	13.17	5	0.0230	2	0.0275	0.0046
Jimmie Johnson	36	13.50	6	0.0205	3	0.0250	0.0042
Espadim Esterlino	29	13.86	9	0.0167	6	0.0207	0.0040
Mike Bliss	1	14.00	3	0.0274	23	0.0146	0.0073
Jeff Gordon	36	14.06	8	0.0168	5	0.0213	0.0036
Kurt Busch	36	14.06	12	0.0153	8	0.0198	0.0034
... Carl Longo	2	40.50	75	0.0021	73	0.0062	0.0029
Christian Fittipaldi	1	41.00	77	0.0019	68	0.0075	0.0039
Hideo Fukuyama	2	41.00	83	0.0014	77	0.0054	0.0028
Jason Pequeno	1	41.00	81	0.0017	71	0.0067	0.0036
Morgan Pastor	5	41.20	78	0.0019	83	0.0041	0.0016
Kirk Shelmerdine	2	41.50	76	0.0021	75	0.0059	0.0028
Austin Cameron	1	42.00	68	0.0029	62	0.0083	0.0043
Dave Marcis	1	42.00	67	0.0030	61	0.0083	0.0043
Dick Gotejamento	3	42.00	74	0.0022	80	0.0050	0.0022
Joe Varde	1	42.00	71	0.0025	66	0.0078	0.0041



a) Máxima verossimilhança (b) Inferência EP bayesiana

Figura 2. Distribuições marginais de classificação posterior para os 5 pilotos superiores e inferiores, conforme ordenado por (a) MLE ou (b) EP. Brancoindica alta probabilidade e as classificações são da esquerda (1º lugar) para a direita (83º lugar).

Tabela 2. Parâmetros PL normalizados para classificar gêneros MovieLens, com no. de pontos de dados em cada categoria entre parênteses.

Todos (943)			Idade 25-29 (175)			Idade 55-59 (32)		
Gênero	Significar	SDev	Gênero	Significar	SDev	Gênero	Significar	SDev
Foi	0.0968	0.0036	Filme Noir	0.0920	0.0101	Foi	0.0873	0.0165
Drama	0.0902	0.0032	Drama	0.0911	0.0075	Thriller	0.0805	0.0147
Filme Noir	0.0828	0.0039	Documentário	0.0867	0.0117	Drama	0.0741	0.0137

Romance0.07090.0026Guerra0.08200.0070Filme-Noir0.06810.0153Crime0.06190.0023Romance0.07300.0060Mistério0.06760.0131Mistério0.06070.0023Crime0.05700.0050Crime0.0.0050Crime0.0.0050Crime0.0050Crime0.0050Crime0.0050Crime0.0050.0050.0050.0050.0050.0050.0050.0050.0050.0050.0050.0050.0050.0050Crime0.0050.0050.0050Crime0.0050.0050.0050Crime0.0050.0050.0050.0050.0050.0050.0023Crime0.0023.0023.0023Crime0.05700.0050Crime0.0050.0050.0050Crime0.0050.0050.0050Crime0.00500.00500 06550.0124Suspense0.05630.0020Ficção Científica0.05330.0045Aventura0.06070.0119Ficção Científica0.05450.0020Animação0.05130.0049Western0.06030.0149Documentário0.05380.0034Thriller0.05010.0041Ação0.05950.0112Ação0.05140.0018Mistério0.04870.0043Romance0.05690.0104Western0.05110.0027Ação0.04790.0039Sci-Fi0.05350.0113Aventura0.04890.0018Western0.04610.0053Documentário0.04590.0139Animação0.04780.04780.0022Comédia0.04500.0037Comédia0.04500.0083Comédia0.04280.0015Aventura0.04460.0038Animação0.04180.0119Musical0.03970.0017Musical0.04110.0039Fantasia0.04180.0148Infantil0.03480.0014Infantil0.03860.0036Musical0.03650.0081Horror0.03130.0013Horror0.02850.0027Horror0.02780.0065Fantasia0.02440.0013Fantasia0.02290.0026Infantil0.02720.0064

classificação baixa. Em seguida, construímos um modelo PL usando essas observações. A vantagem de usar classificações de usuários em vez de classificações é que isso remove o viés do usuário na escala de classificações e, de fato, ordenar os gêneros por classificação média fornece resultados significativamente diferentes. Observe que não estamos classificando a popularidade do gênero aqui - em vez disso, estamos classificando o quão bem um determinado gênero foi recebido, embora seja provável que haja um viés dependente do gênero na seleção de filmes. Assim, por exemplo, o algoritmo colocou o gênero War no topo do ranking; Embora os filmes de guerra não fossem o tipo de filme mais assistido, quando assistidos, eles eram bem classificados. A Tabela 2 mostra as médias dos parâmetros posteriores e as classificações correspondentes para toda a população de usuários; Estes são comparados com as estimativas/classificações de parâmetros para as subpopulações de usuários de 25 a 29 anos e usuários de 55 a 59 anos. Não apenas os rankings são diferentes, com a categoria mais jovem preferindo o Film-Noir aos filmes de guerra da categoria mais antiga, mas também as incertezas são maiores para a categoria mais antiga, devido à existência de apenas 32 pontos de dados de 55 a 59 anos. A divisão dos utilizadores em diferentes categorias sugere uma extensão directa do modelo P-L básico — uma mistura de distribuições P-L. Uma vantagem da inferência EPBayesiana é que a evidência do modelo pode ser usada para determinar o número ideal de componentes em uma mistura. O modelo de mistura resultante pode então ser usado como base para um sistema de recomendação. Deixamos essa extensão para trabalhos futuros.

5. Conclusões

Descrevemos um algoritmo de passagem de mensagens para parâmetros de entrada de uma distribuição de classificação PL. Mostramos que isso pode aprender com precisão os parâmetros e suas incertezas a partir de dados gerados a partir de um modelo PL conhecido. Mostramos a escalabilidade do algoritmo executando-o em conjuntos de dados do mundo real e demonstramos vantagens significativas sobre a abordagem de máxima verossimilhança, especialmente a prevenção de sobreajuste a dados esparsos.

O trabalho futuro envolve a extensão do algoritmo para aprender misturas desses modelos. Um tratamento bayesiano de misturas deve fornecer informações sobre grupos de usuários - por exemplo, dados de classificação de filmes, como o MovieLens. A interpretação de Thurston desses modelos fornece insights sobre como podemos construir modelos mais complexos onde os parâmetros PL são saídas de outros modelos baseados em recursos, estendendo assim a gama de aplicações. Por exemplo, em um aplicativo de "aprender a classificar", poderíamos construir um modelo de regressão baseado em recursos para vincular recursos de documento de consulta a parâmetros de classificação PL. O método EP descrito também é diretamente Seguimos as extensões do modelo P-L básico brevemente discutido na secção 2.1. Esses modelos de "vários estágios" têm muito mais parâmetros e, portanto, provavelmente se beneficiarão ainda mais de um tratamento bayesiano.

Referências

Beggs, S., Cardell, S., & Hausman, J. (1981). Avaliar- a procura potencial de veículos eléctricos. Journ.Econometria, 17, 1–19.Cao, Z., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007).

Aprendendo a classificar: da abordagem em pares à abordagem em lista (Relatório Técnico). Microsoft Research.Dwork, C., Kumar, R., Naor, M., & Sivakumar, D.

(2001). Métodos de agregação de classificação para a Web. World Wide Web (WWW) (pp. 613–622). Gormley, I., & Murphy, T. (2005). Explorando a eletricidade irlandesa

Dados de ção: Uma abordagem de modelização de misturas (Relatório Técnico). Trinity College Dublin, Dept. Stat.Herbrich, R., Minka, T., & Graepel, T. (2007).

TrueSkill(TM): Um sistema de classificação de habilidades bayesiano. InAdv. em Neur. Inf. Proc. Sys. (NIPS) 19, 569–576.Hunter, D. R. (2004). Algoritmos MM para

Modelos Bradley-Terry. Ann. de Estatísticas., 32, 384–406.Joachims, T., Li, H., Liu, T.-Y., & Zhai, C. (2007).

Aprendendo a classificar para recuperação de informações. Spec.Int. Grp. Info. Retr. (SIGIR) Fórum, 41, 58–62.Luce, R. D. (1959). Comportamento de escolha individual: Um o-

análise teórica. Wiley.Marden, J. (1995). Análise e modelagem de dados de classificação.

Chapman e Hall.Minka, T. (2004). Power EP (Relatório Técnico). Mi-

crosoft Research.Minka, T. (2005). Medidas e mensagem de divergência

aprovação (Relatório Técnico). Microsoft Research.Nombekela, S., Murphy, M., Gonyou, J., & Marden, J.

(1994). Preferências alimentares em vacas no início da lactação afetadas por gostos primários e alguns sabores alimentares comuns. Jornal de Ciência de Laticínios, 77, 2393–2399.Plackett, R. (1975). A análise de permutações. Ap-

dobrado Stat., 24, 193–

202.Silverberg, A. (1980). Modelos estatísticos para q-

Permutações. Tese de doutorado, Princeton Univ., Dept. Stat.Thurstone, L. (1927). Uma lei de julgamento comparativo

Mento. Revisões Psicológicas, 34, 273–286.Yellott, J. (1977). A relação entre Luce's

axioma da escolha, teoria do julgamento comparativo de Thurstone e a distribuição exponencial dupla. Journ. Matemática. Psych., 15, 109–144.