Previsão de Engajamento de Usuários Durante Transmissão Adaptativa de Vídeo ao Vivo

Thiago Guarnieri[1], **Alex Vieira**[2], **Ítalo Cunha**[1], **Jussara Almeida**[1]

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

[2]Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)

{thiagoguarnieri,cunha,jussara}@dcc.ufmg.br, alex.borges@ufjf.edu.br

Abstract. *The recent efforts in developing adaptive algorithms and user allocation management tools have contributed significantly to the increase on the quality of experience (QoE) in Internet live video. However, a considerable fraction of sessions still suffer from low QoE, which may reduce user engagement. This problem persists because service providers cannot predict when users will leave the system and attempt to prevent their departure. In this work, we propose a pipelined model for predicting user engagement based on independent variables historically related to QoE. User sessions are clustered by their performance similarities. For each cluster, regression and decision tree based models are built to predict (1) the remaining session time and (2) whether the user will remain watching for the next n minutes. Experiments with real datasets show significant accuracy in the prediction of remaining session time and user permanence, which demonstrates the feasibility of using performance metrics to predict user engagement.*

Resumo. *Os recentes esforços no desenvolvimento de algoritmos de adaptação e alocação de usuários têm contribuído de forma significativa para o aumento da qualidade de experiência (QoE) na distribuição de vídeo ao vivo na Internet. Entretanto, uma considerável parcela de sessões ainda sofre com baixa QoE, o que pode implicar em queda de engajamento dos usuários. Esse problema persiste uma vez que os provedores de serviço não conseguem prever a saída de um usuário e tomar medidas para evitá-la. Neste trabalho, propomos um modelo multi-estágio para predição de engajamento baseado em variáveis historicamente relacionadas a QoE. As sessões são agrupadas de acordo com suas similaridades de desempenho. Para cada grupo de sessões, modelos baseados em árvores de decisão são criados para prever (1) o tempo restante da sessão e (2) se o usuário permanecerá ou não no sistema pelos próximos n minutos. Experimentos com um conjunto de dados reais mostram uma significativa acurácia na predição de tempo restante e permanência, o que evidencia a viabilidade do uso das métricas de desempenho para prever engajamento de usuários.*

1. Introdução

Aplicações para transmissão de mídia contínua têm uma visível penetração na *Internet*. Dia a dia, observamos um crescente número de companhias e empresas que disponibilizam conteúdo de vídeo pela *Internet*, o que reflete em um crescente tráfego relacionado a essas aplicações. De fato, estimativas apontam que o tráfego de vídeo já totaliza 73% de

todo tráfego na *Internet* e que esta proporção pode chegar a 83% até 2020 [Cisco 2017]. Em parte, o sucesso de tais aplicações se deve aos baixos custos associados às transmissões pela *Internet* e à crescente velocidade média de conexão doméstica.

Até 10 anos atrás, essas aplicações eram associadas a baixa qualidade e executadas sobre protocolos específicos como RTP/UDP.[1] Atualmente, transmissões de mídia contínua são tipicamente realizadas por HTTP, com protocolos que se adaptam às condições do acesso à Internet de seus usuários (*Dynamic Adaptive HTTP Streaming*, DASH [Stockhammer 2011]). Adaptar a taxa (*bitrate*) do fluxo de vídeo à capacidade de recepção do usuário é essencial para a sua Qualidade de Experiência (QoE). Pode-se abaixar a taxa em cenários com restrições de banda para evitar interrupções na reprodução (*stalls*), ou aumentá-la em cenários com banda suficiente para melhorar a qualidade do vídeo. A redução de interrupções e o aumento na qualidade do vídeo são fatores que aumentam a qualidade de experiência (QoE) dos usuários [Seufert et al. 2015, Hossfeld et al. 2012, Guarnieri et al. 2017, Mao et al. 2017, Yin et al. 2015].

Para manter o engajamento dos usuários, os provedores de conteúdo devem, antecipadamente, intervir na qualidade do vídeo que um cliente recebe. Por vezes, é melhor reduzir a qualidade do vídeo evitando interrupções indesejadas. De fato, a correlação entre interrupções e engajamento é mais acentuada do que a correlação entre qualidade do vídeo e engajamento [Guarnieri et al. 2017]. Assim, prever o engajamento de seus usuários permite que provedores de serviço acionem mecanismos de adaptação, o que evitaria redução do engajamento e possível desconexão prematura. No entanto, uma área em especial—a das transmissões ao vivo—ainda não possui métodos capazes de prover QoE de forma plena [Ahmed et al. 2016]. Em geral, algoritmos de adaptação necessitam de um tempo para reagir às instabilidades da rede; os limites rígidos de latência e tamanho de *buffer* na distribuição ao vivo reduzem a eficácia de tais abordagens.

Há um grande número de trabalhos que caracterizam a qualidade de serviço que um usuário obtém em transmissões de mídia contínua (seção 2). Esses trabalhos avaliam o impacto de métricas de desempenho—como banda de rede, taxa de perdas de pacote e taxa de transmissão—no desempenho de aplicações específicas, incluindo vídeo [Casas et al. 2013, Gill et al. 2007, Shafiq et al. 2014, Chen et al. 2015, Ahmed et al. 2017]. Entretanto, previsões, quando realizadas, são direcionadas a descoberta da opinião subjetiva dos usuários a respeito da qualidade de experiência (*Mean Opinion Score*, MOS) a partir das métricas de desempenho [da Costa Filho et al. 2016, Balachandran et al. 2013]. Isso se deve à existência de diversos obstáculos à predição de engajamento. O principal deles é a existência de fatores de confusão (*confounding factors*), como interesse no assunto e contexto, que interferem significativamente na permanência do usuário e são difíceis (ou impossíveis) de medir. Por exemplo, um usuário pode abandonar uma sessão precocemente, mesmo que ela tenha boa qualidade.

Neste trabalho, propomos um modelo multi-estágio para predição de engajamento dos usuários baseado em métricas de desempenho independentes historicamente relacionadas a QoE (seção 4). Enquanto os usuários assistem um vídeo ao vivo, sua sessão é particionada em *janelas* não sobrepostas de duração fixa. No primeiro nível, janelas de sessões são agrupados por similaridade de métricas de desempenho. Em seguida, modelos

[1] Real-time transport protocol: protocolo que implementa controle de fluxo e sequenciamento de pacotes específico para fluxos multimídia sobre UDP.

específicos para cada grupo de períodos de sessões são usados para se obter estimativas do engajamento dos usuários. Duas tarefas de predição, que exploram técnicas de modelagem diferentes, são consideradas: (1) modelos de regressão são usados para prever por quanto tempo o usuário permanecerá no sistema e (2) modelos de classificação baseados em árvore de decisão são usados para prever se o usuário irá ou não permanecer no sistema pelos próximos n minutos. Em ambos os casos, a previsão é atualizada continuamente ao longo do tempo para refletir variações de desempenho nas sessões dos usuários.

Nós avaliamos a nossa proposta utilizando como base um conjunto de dados reais (seção 3) coletado durante a transmissão de um grande evento esportivo transmitido em larga escala pela *Internet* [Almeida et al. 2016, Guarnieri et al. 2017]. Apesar da relação entre o desempenho de uma sessão e o engajamento do usuário ser complexa, nossos resultados (seção 5) mostram que os modelos propostos para ambas tarefas atingem significativa acurácia. Para a tarefa de regressão, os modelos produziram estimativas com uma correlação entre duração prevista e duração real de até 0.63 (erro médio de 22%). Para a tarefa de previsão, foi obtida uma acurácia de 81% e uma área sob a curva ROC de no mínimo 70%. Note que, trabalhos semelhantes [Balachandran et al. 2013] só atingiram 45% de acurácia para previsão de duração no caso geral e 70% de acurácia quando ignorando sessões erráticas e muito curtas (*early-quitters*). Por fim, o modelo também manteve uma acurácia alta (próxima a 70%) quando treinado em sessões de uma partida e aplicado a outras partidas, demonstrando sua capacidade de generalização.

No geral, nossos resultados podem ser utilizados por provedores de conteúdo para instruir e direcionar sistemas de adaptação de vídeo visando melhorar a QoE dos usuários. Em cenários de restrições de banda no servidor, nossos modelos podem ser utilizados, por exemplo, para priorizar recursos de banda aos clientes com pior engajamento e mais propensos a deixar o sistema; ou para redirecionar recursos como banda de rede de clientes com altos índices de engajamento para novos clientes entrando no sistema.

2. Trabalhos Relacionados

Nas últimas décadas, sistemas de transmissão de mídia contínua têm-se popularizado. Os perfis das aplicações e de seus usuários têm se alterado e, hoje em dia, um número considerável de pessoas já preferem a *Internet* como meio de acesso a vídeos. Há um grande número de trabalhos de caracterização de sistemas de mídia contínua na *Internet*: alguns caracterizam os sistemas de transmissão de mídia contínua e o perfil de seus usuários [Costa et al. 2004, Almeida et al. 2016], enquanto outros caracterizam a qualidade de serviço prestada [Gill et al. 2007].

Vários trabalhos também têm dado atenção à qualidade de experiência (QoE) que um usuário obtém de um sistema de transmissão de mídia contínua. Nesta direção, foram criadas propostas que monitoram a QoE de usuários em sistemas populares, como o *Youtube* [Casas et al. 2013], além de terem sido conduzidas diversas caracterizações a respeito do engajamento dos usuários de sistemas de transmissão ao vivo [Shafiq et al. 2014, Chen et al. 2015, Guarnieri et al. 2017].

Claramente, há uma relação entre métricas desempenho e experiência percebida pelos usuários. Mais ainda, o engajamento do usuário está intimamente ligado a estas métricas [Guarnieri et al. 2017, Ahmed et al. 2016]. Os primeiros trabalhos para estimação de QoE seguiam a linha qualitativa: o conteúdo de mídia era reproduzido e os entrevistados emitiam uma opinião sobre ele. Essa opinião é então mapeada em uma nota

média (*Mean Opinion Score*, MOS), que quantifica a QoE.

Esta abordagem qualitativa tem limitações de escalabilidade e alto custo devido à necessidade de considerar diversos parâmetros de configuração, cenários de desempenho e tipos de vídeo para um grande número de usuários. Assim, recentemente, um número de trabalhos se dedicam a estimar QoE (capturada via MOS ou outros índices) a partir de métricas alternativas. Tipicamente, os métodos existentes usam a taxa do vídeo ou estatísticas globais de eventos de *stalls* como previsores de QoE. Por exemplo, [da Costa Filho et al. 2016] estimam QoE através de uma árvore de decisão, tomando como variáveis independentes métricas de desempenho obtidas em um ambiente controlado. Essa abordagem não leva em conta transmissões adaptativas. Ainda nessa linha, [Vriendt et al. 2013] desenvolveram um modelo para estimação de QoE usando uma representação sintética dos diversos perfis de qualidade possíveis para um vídeo. Finalmente, [Duanmu et al. 2016] preveem QoE a partir de métricas de desempenho, levando em consideração a degradação instantânea na qualidade da taxa de vídeo recebida.

Considerar o engajamento como métrica indicadora de QoE é desafiador. Apesar da sua correlação com desempenho, o engajamento do usuário está sujeito a fatores de confusão não mensuráveis como, por exemplo, o interesse no conteúdo ou a velocidade de conexão Internet usada [Juluri et al. 2016]. Por essa razão, poucos trabalhos se dedicam à previsão dessa métrica. Um exemplo é o trabalho de [Balachandran et al. 2013], que usa uma árvore de decisão para estimar o engajamento—discretizando em *faixas de permanência*—tomando como parâmetros quantitativos o tempo gasto com preenchimento de *buffer* e a qualidade (*bitrate* médio) do vídeo. Ainda nesse trabalho, os autores sugerem algumas abordagens para tratar os fatores de confusão no engajamento. Para treinar seu modelo e prever o engajamento de uma sessão em andamento, os autores usam um conjunto completo de sessões já concluídas. Por outro lado, eles não consideram adaptações da taxa de transmissão nas suas estimações. Por fim, o método proposto atinge apenas uma acurácia de 45% para previsão de duração da sessão. Mesmo retirando parte das amostras erráticas e sessões muito curtas (*early quiters*), eles não superam 70% de acurácia.

Com base no exposto, nosso trabalho procura oferecer mais uma contribuição para a previsão de engajamento em mídia contínua, seja o *tempo remanescente* de uma sessão ou *se um usuário irá abandonar o sistema* em uma janela de tempo futura. Além das métricas objetivas, nossa abordagem avalia a importância de fatores de confusão, o que aumenta a acurácia da previsão de engajamento. Além disso, consideramos diferentes perfis de usuários atendidos por um sistema de mídia contínua. Por fim, destacamos que nossa abordagem é adaptativa e funciona em tempo real, sendo executado em menos de 1 minuto para dezenas de milhares de usuários. Além disso, destacamos que a nossa abordagem se generaliza para outros sistemas desse tipo, tendo em vista que as métricas utilizadas são comuns e podem ser extraídas de diversos sistemas para qualquer tipo de conteúdo. Ao contrário dos trabalhos existentes, nosso treinamento não necessita de um conjunto de sessões completas. Podemos treiná-lo à medida que usuários assistem vídeos (*online*), o que expande a aplicabilidade do modelo. Em particular, nosso previsor pode auxiliar decisões de mecanismos de adaptação, planejamento de publicidade e controle de alocação de usuários em CDNs.

3. Cenário alvo e conjunto de dados

Antes de apresentar os modelos de previsão de engajamento propostos, descrevemos o sistema de transmissão de mídia ao vivo a partir do qual os dados usados no treino e avaliação dos modelos foram coletados.

3.1. Infraestrutura de transmissão de mídia contínua HTTP adaptativa

O sistema de transmissão de mídia ao vivo alvo deste estudo é o da Globo.com, um dos maiores provedores de conteúdo da América Latina. Durante a coleta de dados, a Globo.com distribuía vídeo usando a tecnologia *Apple HTTP Live Streaming* (HLS) [Apple 2016]. Em linhas gerais, a mídia é obtida de dispositivos de gravação e codificada em diferentes taxas (*bitrates)*. A Globo.com codificava mídia em até seis taxas, variando entre 264–2564 Kbps. Em seguida, a mídia é dividida em segmentos que são indexados para posterior transmissão via protocolo HTTP. A transmissão dos segmentos é sequencial e considera a capacidade do cliente (i.e., largura de banda).

3.2. Conjunto de dados

Nós utilizamos em nossas análises registros dos *logs* de todas as requisições HTTP atendidas pelos servidores de mídia ao vivo da Globo.com, coletados durante a Copa do Mundo de Futebol FIFA de 2014. Nessa etapa do trabalho, utilizamos como base o jogo com maior número de usuários.

Cada requisição HTTP para um segmento de vídeo é registrada pelo provedor de conteúdo em seus servidores nginx, no formato padrão. Estes registros contêm a data e hora da requisição, IP do cliente, URL requisitada, *status* HTTP, total de *bytes* enviados e a identificação do navegador (*user agent*). O descritor do navegador nos permite identificar o tipo de dispositivo (e.g., *PC, tablet ou smartphone*). Identificamos clientes combinando o descritor do agente de navegação e o endereço IP do cliente.

Para capturar métricas de desempenho de um cliente, agrupamos suas requisições em sessões. Para isso, observamos que os clientes solicitam, em 98,9% dos casos, segmentos consecutivos de mídia em um intervalo médio de 3 segundos. Dado esse comportamento, definimos de forma conservadora uma sessão como sendo uma sequência de requisições de um mesmo cliente que não esteja separada por mais de 120 s. Diferentes limiares (de 30 s a 180 s) levaram a resultados qualitativamente similares.

Muitos usuários acessam a Internet através de NAT, o que pode dificultar a identificação de múltiplos clientes advindos de uma mesma rede. Optamos pela exclusão das sessões com possível influência de NAT, ignorando sessões onde foram observadas recebimento duplicado de mais de 5% do total de segmentos transmitidos ou a taxa de envio de segmentos superior a 1 pacote a cada 2 segundos. Todavia, menos de 5% das sessões apresentaram um desses tipos de comportamento.

No geral, observamos até 1,1 milhão de clientes únicos em um único dia (com 4 partidas) e até 470 mil sessões simultâneas durante uma única partida. Ao analisar os descritores de agentes de navegação, identificamos que 81% das sessões são iniciadas a partir de PCs, 12% de *smartphones* e 6% de *tablets*. Algumas sessões de tipos de dispositivos não identificados e robôs completam o conjunto.

4. Previsão de engajamento de usuários

De forma resumida, a nossa abordagem para predição de engajamento consiste de quatro etapas, como apresentada na figura 1. Na primeira etapa, requisições por blocos de vídeo de um determinado usuário são agrupadas em sessões como descrito na seção 3. Esse agrupamento poderia ser feito em tempo de execução enquanto os *logs* das requisições são armazenados.

Na segunda etapa, que também pode ser realizada em tempo de execução, as sessões são submetidas a um segmentador que as divide em janelas de 1 minuto. Optamos por essa duração de janela tendo como base artigos que utilizam metodologia similar para segmentação (i.e., [Ahmed et al. 2017]). Além disso, a grande maioria de artigos de caracterização de QoE fracionam as métricas de desempenho em intervalos de 1 minuto. O intuito da segmentação é prover fotografias da sessão ao longo de sua existência, o que captura melhor as oscilações de desempenho.

Na terceira etapa, as janelas são submetidas a um algoritmo de agrupamento. Esse algoritmo procura agrupar janelas com base em similaridades de desempenho. A finalidade do agrupamento é capturar o impacto de fatores de confusão agrupando como, por exemplo, usuários que abandonam precocemente a sessão mesmo tendo bom desempenho ou aqueles que permanecem mesmo em situações em que este desempenho é baixo.

Por fim, na quarta etapa, janelas de cada grupo de sessões são submetidas a ferramentas de classificação e regressão para treinamento de modelos de previsão. Na classificação, resolvemos um problema de decisão de duas classes para decidir se uma sessão em andamento vai ou não durar pelos próximos n minutos. Já a regressão é usada para obter uma estimativa do tempo restante de sessão. Nas próximas subseções apresentaremos os detalhes de cada uma das etapas mencionadas.

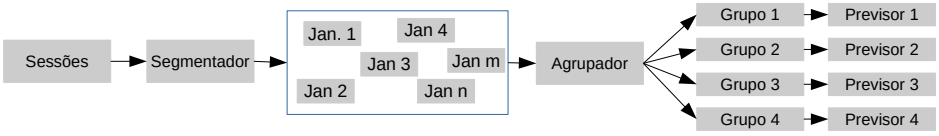

Figura 1. Sequência de etapas para treinamento de modelo de previsão.

4.1. Estrutura das janelas

Para estimar o engajamento de usuários, sessões em andamento são segmentadas em janelas de 1 minuto. Cada janela captura métricas de desempenho agregadas, desde o *início da sessão* até o *término da janela* em questão. A agregação é feita em cada janela em razão da escala do treinamento. Somente o conteúdo utilizado para treino do modelo possui mais de 10 milhões de janelas, sendo impossível manter todas elas em memória para consultar o desempenho nas janelas anteriores. As métricas calculadas para cada janela são descritas a seguir:

Atraso de inicialização é o tempo de espera necessário para que o usuário possa iniciar a reprodução. Em outras palavras, é o tempo para carregar o *buffer* de reprodução pela primeira vez. [Hossfeld et al. 2012] mostram que um *delay* acima de 16 segundos já pode provocar um impacto na permanência do usuário.

Taxa média do vídeo: Em transmissões adaptativas, a taxa média do vídeo (*bitrate*) pode variar ao longo do tempo. Em cada janela, é calculada a média aritmética da taxa de todos os segmentos. [Balachandran et al. 2013, Krishnan and Sitaraman 2013] mostram que há uma relação não-monotônica entre o esta métrica e o engajamento.

Interrupções: Também chamadas de *stalls*, ocorrem quando ocorre esvaziamento do *buffer* de reprodução, fazendo com que a reprodução seja interrompida até que o *buffer* seja re-preenchido. No trabalho corrente, caso um usuário fique mais de 12 segundos sem receber dados de mídia, contabiliza-se um *stall*; utilizamos este limite pois é o tamanho padrão de *buffer* em clientes HLS [Apple 2016]. Além disso, observamos que após um *stall* o *player* requisita pacotes em uma taxa mais elevada para preenchimento rápido do *buffer* e o número de pacotes requisitados é equivalente a 12 segundos. Tanto a duração de um *stall* [Qi and Dai 2006] quanto a sua taxa [Krishnan and Sitaraman 2013, Guarnieri et al. 2017] tem um impacto significante no engajamento.

Adaptações: Define-se como adaptação o evento de *troca* da taxa de transmissão (qualidade) do vídeo requisitado e transmitido a um cliente. Em cada janela, contabilizamos o número de adaptações. Em nosso trabalho anterior [Guarnieri et al. 2017] observamos que reduções da taxa de transmissão têm influencia negativa na permanência.

Tempo de sessão: Define-se como tempo de sessão, o tempo decorrido desde a inicialização do vídeo até a janela atual.

O atraso de inicialização é calculado uma única vez no início da sessão. A taxa média do vídeo, o número de interrupções e o número de adaptações é calculado de duas formas para cada janela: calculamos a média aritmética e o valor agregado desde o início da sessão até a janela atual e uma média exponencial móvel.[2] Como em trabalhos anteriores, usamos a média aritmética esperando capturar o desempenho médio da sessão do usuário. Usamos a média exponencial móvel para capturar a propriedade de que o desempenho de uma sessão no passado distante tem menor influência nas decisões de um usuário comparadas ao desempenho da sessão no passado recente. Por exemplo, se um usuário sofreu com má qualidade do vídeo recentemente, a probabilidade dele abandonar o sistema é alta.

Considerações sobre desbalanceamento de amostras: Em nosso conjunto de dados, a tendência natural, a cada minuto, é de permanência em detrimento do abandono de sessões. Por essa razão, existe uma fração majoritária (90%) de janelas cuja resposta para o problema de decisão *"vai permanecer pelos próximos n minutos"* é *sim*. Por outro lado, para que um modelo seja bem treinado, é necessário que a quantidade de itens da classe *"sim"* seja similar a da classe *"não"*. Para isso, efetuamos o ajuste os pesos das classes de forma que itens da classe *"não"* tenham peso maior. Isso gera por consequência um modelo que favorece às duas classes igualmente.

4.2. Agrupamento de sessões

Como mencionado na seção 1, uma parte considerável do engajamento do usuário não pode ser mensurada por se tratar de aspectos subjetivos. Por exemplo, pode ocorrer o

[2]Calculamos a média exponencial móvel da métrica X na janela $t+1$ como $X_{t+1} = 0.5X_t + 0.5M_t$, onde X_t é o valor da média móvel na janela t e M_t é o valor medido da métrica na janela t.

abandono precoce de uma sessão com alta qualidade se o conteúdo não for de interesse do usuário. Analogamente, uma sessão de baixo desempenho pode ter alto engajamento caso o conteúdo seja de alta relevância para o usuário.

A classificação dos usuários levando em conta o impacto de tais fatores de confusão é efetuada por *clusterização*: as janelas são agrupadas por similaridade de métricas de desempenho e tempo de sessão usando o algoritmo de agrupamento *k-means* [Macqueen 1967]. O valor de k foi definido automaticamente empregando o *x-means* [Pelleg and Moore 2000]. Nesta técnica, atribui-se um número máximo e mínimo de grupos. Em cada iteração do algoritmo, a quantidade de grupos é incrementada e a solução é comparada com a anterior. Se for melhor (conforme um critério qualquer), a solução é registrada. Isso é feito até que o limite superior de grupos seja atingido. Para decidir qual número de grupos é o melhor, é usado um critério de seleção de modelos (um modelo nesse contexto são os elementos e seus centroides). No caso do *x-means*, o critério de escolha é o *Bayesian Information Criterion* [Schwarz 1978]. Além do *k-means*, outros modelos foram considerados, tais como os modelos hierárquicos, descartados devido a sua alta complexidade computacional, e os baseados em densidade, que não agrupam todos os elementos, sendo por isso inadequados ao nosso problema.

4.3. Modelos de previsão

A última etapa na sequência é a submissão das janelas para treinamento nos algoritmos de classificação e regressão. Neste trabalho, efetuamos o teste com diversos classificadores e regressores, e o que obteve melhor desempenho foi o *random-forest* [Ho 1995]. Modelos baseados em árvore são amplamente usados e são considerados o estado da arte em problemas de previsão. No *random-forest*, as predições de diversas árvores, cada uma com parte da amostra, são combinadas de forma a gerar um resultado mais acurado. A partição em cada nível da árvore é feita escolhendo-se o subconjunto das métricas que propicia o melhor particionamento segundo uma função objetivo específica. O *random-forest* em particular possui poucos parâmetros a serem ajustados e o único modificado foi o número de árvores que passou de 100 pra 200, seguindo orientação de [Oshiro et al. 2012]. Por fim, para cada grupo de janelas, um modelo específico foi treinado para refletir suas particularidades.

5. Avaliação

Após a etapa de treinamento, obtemos para os centroides de cada grupo seus modelos de previsão. Quando um servidor de vídeo desejar obter uma previsão de engajamento, ele submete as janelas para o agrupador, que associa cada janela ao centroide já calculado. A seguir, cada janela é submetida ao modelo associado ao seu grupo e a previsão é efetuada. Para avaliar o desempenho do modelo, submetemos um conjunto de teste para os previsores e contabilizamos a acurácia através de diversas métricas bem estabelecidas na literatura. A divisão entre conjunto de treino e teste se dá de duas formas:

1. Validação cruzada em 10 partições (*10-fold cross-validation*): Nessa abordagem janelas de uma partida são divididas em 10 partes, sendo 9 utilizadas para treino e 1 para teste. O processo é repetido 10 vezes trocando, a cada repetição, as partições de treino e teste. Ao final a média dos resultados em todas as repetições é apresentada.

2. Conjunto de teste de outra partida: Nessa modalidade de teste treinamos um modelo com janelas de um jogo e avaliamos a acurácia da previsão em uma amostra de janelas de outra partida arbitrária. Isso permite avaliar a generalização do modelo para outros conteúdos de natureza similar. Um modelo com essa característica de generalização permite que o treino possa ser feito em intervalos de tempo maiores, reduzindo custos de processamento. A amostra para essa análise não é balanceada.

A acurácia, por sua vez, é medida em função das taxas de verdadeiros e falsos positivos (TP e FP, respectivamente) e das taxas de verdadeiros e falsos negativos (TN e FN, respectivamente). Abaixo apresentamos as métricas utilizadas para avaliação de acurácia:

- Taxa de acertos: Calculada para a classificação, quantifica a fração de verdadeiros positivos e negativos em relação ao total de predições: $(\text{TP}+\text{TN}) \div (\text{TP}+\text{TN}+\text{FP}+\text{FN})$.

- *F-measure*: A média ponderada entre a precisão[3] e a revocação[4] do classificador. É uma medida útil para estimar a acurácia do previsor para as possíveis classificações (SIM/NÃO) de forma independente, mas é menos robusta quando as classes são desbalanceadas. Nesses casos, é recomendado utilizar outras medidas como a área sobre a curva ROC. A *F-measure* é calculada como $2rp/(r+p)$, onde r e p representam revocação e precisão, respectivamente.

- Macro F1: Dados os valores de *f-measure* para as classes positiva (FP) e negativa (FN), esta medida pode ser calculada como $(FP + FN)/2$. Ela é reportada para conjuntos de teste onde o desbalanceamento entre as classes é mais acentuado. Assume uma faixa de valores entre 0 e 1.

- Área sob a curva ROC: Medida de precisão utilizada para problemas onde as classes são também desbalanceadas, em particular quando há prevalência de positivos. A curva padrão ROC consiste num *plot* da taxa de verdadeiros positivos (TVP) contra a taxa de falsos positivos (TFP). No cálculo é usado o teste Wilcoxon–Mann–Whitney [Mann and Whitney 1947], que emprega um ranqueamento das instâncias de acordo com a probabilidade de pertencer a uma das duas classes. Quanto mais próximo de 1, mais preciso é o modelo, e um valor próximo de 0,5 indica que o preditor funciona de maneira aleatória.

- Coeficiente de *Pearson* (ρ): Utilizado na regressão para medir a correlação linear entre os tempos de permanência previstos e os reais. Os valores 1 e -1 indicam, respectivamente, uma correlação máxima positiva e negativa. Já o valor zero indica que não existe correlação linear entre as duas variáveis.

5.1. Caracterização dos grupos

A figura 2 apresenta a caracterização dos grupos encontrados pelo *x-means*. Foram encontrados 4 grupos distintos. Abaixo discutimos propriedades de cada um.

- Grupo de baixo desempenho (31.07% das janelas): Neste grupo estão as janelas das sessões com duração média pequena, com 23% do tempo decorrido para 80% das sessões.

[3]Conta as previsões certas efetuadas dentre todas as previsões efetuadas (certas e erradas).

[4]Conta as previsões certas efetuadas dentre todas as previsões certas possíveis.

Ou seja, supondo que tenha se passado 10 minutos de jogo, essas janelas indicam duração menor que 3 minutos. As janelas desse grupo também são as que possuem adaptações mais frequentes, sendo duas por minuto para mais de 60% das janelas. É possível que os usuários desse grupo sejam mais sensíveis às variações de qualidade durante a sessão ou que sejam apenas usuários que assistam por períodos mais curtos. Por fim, o *bitrate* das sessões é menor que 1300 Kbps para 85% das janelas.

- Grupo de médio desempenho (9.73% das janelas): Composto pelas janelas com duração menor que 28% do tempo decorrido, para 80% das sessões. Têm janelas com uma taxa de quedas e *stalls* moderada e um *bitrate* acima de 1500 Kbps para 75% das sessões.

- Grupo de bom desempenho (17.29% das janelas): Grupo cuja ocupação do tempo decorrido alcança até 45% para 80% das janelas. Também tem as menores taxas de *stall*, quedas e menor *delay* de inicialização. O fato de apenas 17% dos usuários desfrutarem dessa qualidade pode ser outro indicativo de que melhorias na QoE podem ser implementadas.

- Grupo errático (41.91% das janelas): similar ao grupo de duração média, este grupo é composto por janelas de tempo decorrido menor que 27%, para 80% das sessões. No entanto sua taxa de *stalls* é significativamente alta, com 23% das sessões com pelo menos um *stall*. O *bitrate* também é baixo, com 70% dos slices indicando uma taxa de 330 Kbps. Este grupo é um exemplo de como os fatores subjetivos influenciam a duração dos usuários: mesmo com desempenho ruim, os usuários se mantêm por um tempo considerável na transmissão.

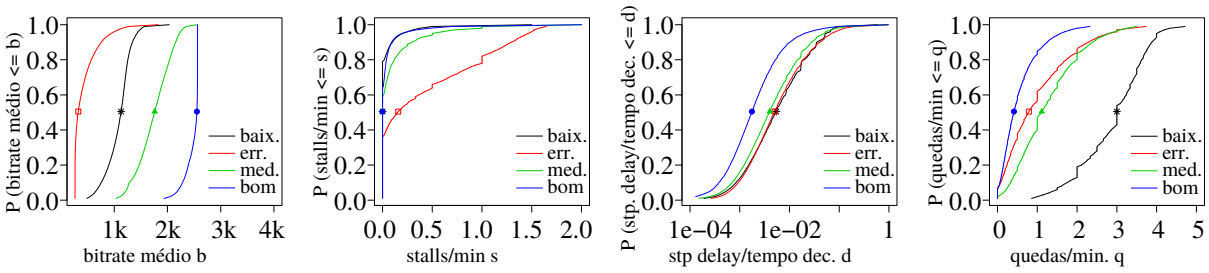

Figura 2. Características dos grupos

5.2. Predição de horizonte de permanência

Nesta seção apresentaremos os resultados das predições do modelo para as amostras consideradas. A tabela 1 mostra a precisão para cada um dos grupos considerados para horizontes de classificação h variando de 1 a 5 minutos, i.e., quando tentamos prever se o usuário vai deixar (ou não) o sistema nos próximos h minutos.

Como esperado, a precisão reduz à medida em que o horizonte de classificação fica mais distante. Por exemplo, na análise via validação cruzada a precisão máxima, que é de 81% para um horizonte de 1 minuto, cai para uma precisão máxima de 75% para um horizonte de 5 minutos. Observamos também que, para horizontes de 2 a 5 minutos, o grupo de maior precisão foi o que tem as sessões de maior qualidade (grupo *grd*). Como visto na subseção 5.1, as distribuições das métricas de desempenho mostram uma

variação menor, principalmente na taxa de *stalls* e quedas. Essa uniformidade pode ter ajudado a melhorar a acurácia da previsão. Já o grupo de qualidade baixa tem uma queda mais acentuada a partir do horizonte de 2 minutos. Isto é em razão de as saídas precoces de sessões de até 1 minuto, com baixa qualidade, terem mais a ver com desempenho do que o abandono de sessões em outros grupos. A área sob a curva ROC também apresenta valores acima de 0,7 para todos os grupos e todos os horizontes de classificação, indicando um desempenho melhor do que atribuição aleatória.

Ao aplicarmos o modelo em outras partidas, observamos que a acurácia em geral diminui quando comparada à validação cruzada. Isto pode ser explicado por uma perda de precisão do modelo treinado em um jogo e aplicado em outro. De qualquer forma, esta perda de precisão é pequena em relação ao conjunto de treino. Também é interessante observar que a curva ROC mantém acurácia de 0.7 até 2 minutos, mesmo no cenário desbalanceado, o que é um resultado encorajador dado o fato de eventos reais seguirem tal desbalanceamento.

Em relação ao valor F (*f-measure*), podemos perceber que a precisão é similar tanto para a classe de valores positivos quanto para a de negativos. No entanto, para o conjunto de teste onde usamos o macro f1, percebemos uma queda de desempenho, embora a medida de área sob a curva se mantenha alta. O fato dos dados de teste serem altamente desbalanceados pode estar interferindo nesta medida. Finalmente, uma análise adicional cabe com relação às classes desse problema: é possível que para algumas aplicações, a importância de se saber quando um usuário vai sair seja maior. Nesse caso devemos considerar um modelo sensível a custo, dando pesos diferentes aos erros de predição pra cada classe. Em nossos trabalhos futuros pretendemos adicionar essa característica.

Tabela 1. Desempenho do previsor para diferentes horizontes de classificação

	10-fold CV			Teste		
cluster	Acur.%	F-meas.(s/n)	AUC ROC	Acur. %	Macro F1	AUC ROC
1 min						
baixo	81.57	0.838/0.786	0.893	78.95	0.527	0.815
errático	73.91	0.720/0.756	0.816	72.94	0.542	0.839
médio	72.98	0.650/0.780	0.802	65.37	0.495	0.797
bom	73.26	0.793/0.624	0.783	68.63	0.561	0.769
2 min						
baixo	68.46	0.701/0.667	0.747	64.85	0.512	0.688
errático	66.95	0.611/0.713	0.728	61.79	0.519	0.705
médio	66.11	0.675/0.646	0.721	65.29	0.506	0.673
bom	75.66	0.826/0.595	0.792	86.61	0.590	0.713
3 min						
baixo	66.38	0.676/0.651	0.722	64.49	0.541	0.664
errático	66.37	0.601/0.709	0.719	61.18	0.541	0.681
médio	65.77	0.680/0.632	0.713	66.62	0.546	0.663
bom	76.62	0.831/0.621	0.805	85.05	0.612	0.701
4 min						
baixo	65.35	0.626/0.678	0.710	60.13	0.543	0.644
errático	64.98	0.666/0.633	0.703	64.41	0.554	0.647
médio	68.22	0.643/0.713	0.745	64.68	0.593	0.697
bom	76.71	0.899/0.542	0.809	83.52	0.625	0.701
5 min						
baixo	65.82	0,676/0,639	0,718	61.32	0.557	0,644
errático	68.95	0,658/0,716	0,754	65.12	0.604	0,681
médio	65.69	0,580/0,710	0,708	61.17	0.568	0,652
bom	75.13	0,811/0,638	0,803	79.92	0.633	0,702

5.3. Predição de tempo de permanência restante

O objetivo maior de nossa tarefa de previsão é prever com acurácia o tempo que o usuário ainda permanecerá assistindo algum conteúdo. Como já explorado nesse artigo, a tarefa é desafiadora. Um usuário pode abandonar o sistema por inúmeras razões, que nada tem a ver com o desempenho observado, como por exemplo o interesse no conteúdo. Além disso, o problema tratado neste trabalho é de granularidade mais fina: prever engajamento ao longo de uma sessão é no mínimo tão difícil quanto prever o engajamento somente no final da mesma.

Tendo tais observações em mente, a tabela 2 apresenta as correlações entre os valores preditos e reais para a predição de duração do tempo de permanência do usuário no sistema. O treinamento foi efetuado com o tempo de permanência normalizado pelo tempo restante. Novamente, o grupo de sessões de alto desempenho foi o que proporcionou a melhor previsão ($\rho = 0.634$). Já o erro médio absoluto foi de 22%, o que significa que o previsor estimou uma permanência 22% maior (ou menor) para os usuários. Para o grupo de sessões erráticas obtivemos a menor correlação ($\rho = 0.417$), indicando o forte impacto que fatores alheios ao desempenho têm nesse caso. Observe que a diferença nas correlações é maior que a diferença dos erros pois a clusterização divide as janelas em faixas de duração (subseção 5.1) e o tempo restante de permanência está normalizado. Como exemplo podemos imaginar dois cenários: em um o usuário ainda permanecerá mais 5 minutos e o tempo restante é de 15 minutos. No outro cenário o usuário ainda permanecerá mais 10 minutos e o tempo restante é de 30 minutos. Em ambos os casos a permanência será de 33%, mas o impacto de um erro no segundo caso é maior.

Os valores obtidos na previsão podem ser usados também de forma indireta. Se o provedor de conteúdo sabe que o usuário se encaixa no grupo de sessões erráticas, ele pode reduzir a qualidade de transmissão sem que isso gere um impacto significativo na experiência do usuário. Assim os recursos podem ser alocados com mais prioridade para os usuários cuja sessão tem maior correlação com as métricas de desempenho.

Tabela 2. Desempenho do previsor para regressão

Cluster	(ρ) 10-fold CV	(ρ) Erro médio absoluto (tempo decorrido)
Baixo	0.541	27%
Errático	0.417	29%
Médio	0.461	28%
Bom	0.634	22%

6. Conclusões e trabalhos futuros

Um desafio ainda presente no contexto de vídeos para Internet é a ausência métricas que orientem o desenvolvimento de soluções sensíveis à QoE. O presente trabalho contribui nesta direção, propondo um modelo de previsão de engajamento de usuários em sessões de vídeo ao vivo. O modelo utiliza como variáveis dependentes as métricas de desempenho de uma sessão para prever o engajamento do usuário, em particular a duração restante de sua sessão. O modelo é flexível e pode ser usado para orientar escolhas de taxa de transmissão em algoritmos de adaptação ou alocar recursos da infraestrutura de distribuição de forma a maximizar QoE em um sistema.

Mais ainda, nosso modelo avalia o impacto de fatores de confusão agrupando usuários pelo desempenho instantâneo de suas sessões. Assim, os usuário cuja relação

entre desempenho e engajamento é fraca são agrupados à parte. O processo de agrupamento também melhora a acurácia da previsão ao agrupar usuários cujas sessões tem desempenho e engajamento parecido.

Nossos resultados mostram que o impacto dos fatores de confusão é significativo, mas de prevalência diferente para diferentes contextos de sessões. Existem usuários que abandonam a transmissão quando observam desempenho ruim e outros que, na mesma situação, permanecem. Mesmo neste cenário desafiador, nosso modelo alcança precisão de previsão considerável. Como trabalhos futuros pretendemos avaliar o uso de outras métricas de desempenho ou características das sessões dos usuários, expandir as bases de treino e testes e criar algoritmos adaptativos em sistemas reais.

Referências

Ahmed, A., Shafiq, Z., Bedi, H., and Khakpour, A. (2017). Suffering from buffering? detecting QoE impairments in live video streams. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*, pages 1–10.

Ahmed, A., Shafiq, Z., and Khakpour, A. (2016). QoE analysis of a large-scale live video streaming event. In *Proc. of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pages 395–396.

Almeida, B., Carnivalli, G., de Almeida Junior, W., Almeida, J., Cunha, I., and Vieira, A. B. (2016). Caracterização do comportamento dos clientes de um sistema de vídeo ao vivo durante um evento de larga escala na internet. In *34o. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*.

Apple (2016). HTTP Live Streaming (HLS) - Apple Developer. `http://developer.apple.com/streaming`.

Balachandran, A., Sekar, V., Akella, A., Seshan, S., Stoica, I., and Zhang, H. (2013). Developing a predictive model of quality of experience for internet video. *SIGCOMM Comput. Commun. Rev.*, 43(4):339–350.

Casas, P., Seufert, M., and Schatz, R. (2013). Youqmon: A system for on-line monitoring of youtube QoE in operational 3g networks. *ACM SIGMETRICS Performance Evaluation Review*, 41(2):44–46.

Chen, Y., Chen, Q., Zhang, F., Zhang, Q., Wu, K., Huang, R., and Zhou, L. (2015). Understanding viewer engagement of video service in wi-fi network. *Computer Networks*, 91:101–116.

Cisco, V. N. I. (2017). The zettabyte era: Trends and analysis. *Cisco whitepaper (june, 2017), https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html*.

Costa, C., Cunha, I., Borges, A., Ramos, C., Rocha, M., Almeida, J., and Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *Proceedings of the 13th ACM international conference on World Wide Web*, pages 534–543.

da Costa Filho, R. I. T., Lautenschlager, W., Kagami, N., Roesler, V., and Gaspary, L. P. (2016). Network fortune cookie: Using network measurements to predict video streaming performance and QoE. In *Proc. of the IEEE GLOBECOM*.

Duanmu, Z., Rehman, A., Zeng, K., and Wang, Z. (2016). Quality-of-experience prediction for streaming video. In *Proc. of the IEEE ICME*.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28.

Guarnieri, T., Ítalo Cunha, Almeida, J., Drago, I., and Vieira, A. B. (2017). Characterizing QoE in large-scale live streaming. In *Proc. of the IEEE GLOBECOM*.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, pages 278–.

Hossfeld, T., Egger, S., Schatz, R., Fiedler, M., Masuch, K., and Lorentzen, C. (2012). Initial delay vs. interruptions: Between the devil and the deep blue sea. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 1–6.

Juluri, P., Tamarapalli, V., and Medhi, D. (2016). Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys Tutorials*, 18(1):401–418.

Krishnan, S. S. and Sitaraman, R. K. (2013). Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. *IEEE/ACM Transactions on Networking*, 21(6):2001–2014.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Mao, H., Netravali, R., and Alizadeh, M. (2017). Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, pages 197–210.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). *How Many Trees in a Random Forest?*, pages 154–168. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Seventeenth International Conference on Machine Learning*, pages 727–734. Morgan Kaufmann.

Qi, Y. and Dai, M. (2006). The effect of frame freezing and frame skipping on video quality. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, pages 423–426.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Seufert, M., Egger, S., Slanina, M., Zinner, T., Hoßfeld, T., and Tran-Gia, P. (2015). A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys Tutorials*, 17(1):469–492.

Shafiq, M. Z., Erman, J., Ji, L., Liu, A. X., Pang, J., and Wang, J. (2014). Understanding the impact of network dynamics on mobile video user engagement. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 367–379.

Stockhammer, T. (2011). Dynamic adaptive streaming over http–: standards and design principles. In *Proceedings of the second annual ACM conference on Multimedia systems*, pages 133–144. ACM.

Vriendt, J. D., Vleeschauwer, D. D., and Robinson, D. (2013). Model for estimating QoE of video delivered using http adaptive streaming. In *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, pages 1288–1293.

Yin, X., Jindal, A., Sekar, V., and Sinopoli, B. (2015). A control-theoretic approach for dynamic adaptive video streaming over http. In *Proc. of the ACM Conference on Special Interest Group on Data Communication*, SIGCOMM '15, pages 325–338.