Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados

Ailton Fonseca Galvão[1], **Raul Fonseca Neto**[2], **Carlos Cristiano Hasenclever Borges**[2]

[1]Mestrando em Ciência da Computação – Área de Inteligência Computacional
Universidade Federal de Juiz de Fora – Juiz de Fora, MG – Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora – Juiz de Fora, MG – Brasil

ailton.mcc@gmail.com,raul.fonseca@ice.ufjf.br,cchb@lncc.br

Ingresso no programa: Março de 2011

Resumo. *Neste trabalho será proposta uma abordagem para o problema da seleção de itens em testes adaptativos computadorizados. Os modelos de seleção mais utilizados atualmente se baseiam apenas em métodos estatísticos para medir a quantidade de informação do item dada a habilidade do respondente no momento. Nossa proposta se baseia em modelos de aprendizado e otimização da área de inteligência computacional que possam selecionar o menor número de itens que produzam uma medida correta da habilidade minimizando o erro dessa medida.*

Palavras-chave: Teste Adaptativo Computadorizado, Seleção de Itens, Inteligência Computacional

1. Caracterização do Problema

Para determinarmos o nível de habilidade ou de conhecimento que um indivíduo possui em relação a uma determinada característica que não é possível medir diretamente é necessário que seja desenvolvido algum tipo de teste. Tradicionalmente, nos baseamos na quantidade de questões certas e erradas em um teste para avaliarmos o desempenho do indivíduo. Essa forma tradicional de medida é chamada de Teoria Clássica dos Testes (TCT) e é muito simples de interpretar. Porém, segundo [Baker 2001], é essa característica de simplicidade que limita a TCT. O resultado do indivíduo pode variar de teste para teste, dependendo dos conteúdos, fazendo com que seja difícil comparar o desempenho de alunos aplicando-se testes diferentes. [Thurstone 1959] definiu esse problema em relação aos conteúdos da seguinte forma: "Um instrumento de medida, na sua função de medir, não pode ser seriamente afetado pelo objeto de medida. Na extensão em que sua função de medir for assim afetada, a validade do instrumento é prejudicada ou limitada".

A partir da década de 1950, o desenvolvimento da Teoria da Resposta ao Item (TRI) trouxe uma nova forma de avaliar o conhecimento, superando as limitações do teste presentes na TCT. As principais características, e também vantagens, da TRI sobre a TCT são a localização da dificuldade do item e da proficiência do indivíduo na mesma escala e a independência dos resultados dos indivíduos examinados em relação ao teste utilizado. Essa possibilidade da análise dos parâmetros dos itens e da proficiência na mesma escala obtida com a TRI levaram ao desenvolvimento dos chamados testes adaptativos, que são testes sequenciais onde os itens são escolhidos um após o outro se adaptando ao conhecimento/habilidade do respondente [Linden and Glas 2000].

Desde meados dos anos 1980 a TRI vem se tornando a técnica predominante no campo dos testes e avaliações, notadamente em avaliações educacionais [Pasquali 1996]. O avanço da informática nesse período permitiu o desenvolvimento de *softwares* que tornaram a TRI, e seus métodos estatísticos complexos, mais acessível aos pesquisadores da área de avaliação.

Nesse contexto surgiram os Testes Adaptativos Computadorizados (TAC's), que, com a popularização da informática no início da década de 1990, tornaram-se frequentes nos países em que as avaliações e testes pela TRI já eram prática comum. Em um teste adaptativo, a ideia básica é que a seleção do próximo item depende do resultado do indivíduo até aquele momento, ou seja, a cada nova seleção de item e resposta deste por parte do examinando, a medida é reestimada e o próximo item selecionado deverá se adaptar a essa nova medida, chamada, na área de avaliação educacional, de proficiência. A figura 1 ilustra o método de funcionamento de um TAC.

Figure 1. Esquema de representação de um Teste Adaptativo Computadorizado

O cálculo correto da proficiência depende diretamente que os itens aplicados no teste

meçam adequadamente na região da escala de habilidades em que o indivíduo se encontra. Quando um teste é desenvolvido na forma tradicional, ou seja, um teste com as questões definidas previamente, um ponto importante a se observar é a necessidade da presença de itens que cubram todos os níveis de dificuldade. Assim, é estabelecido um teste grande para garantir que indivíduos com proficiências em pontos diferentes da escala possam ser avaliados com precisão.
Como em um teste adaptativo os itens não são previamente definidos, o processo de seleção se torna o ponto principal para garantir a precisão da medida [Linden and Glas 2000]. Diversas propostas de modelos para seleção de itens surgiram desde que os testes adaptativos começaram a ser desenvolvidos. Os modelos de seleção mais utilizados atualmente se baseiam em métodos estatísticos para medir a quantidade de informação do item, sendo o principal deles o critério de Máxima de Informação de Fisher [Veldkamp 2010].
Um modelo computacional inteligente se tornaria uma solução melhor para o problema da seleção de itens, uma vez que pode ser desenvolvido de forma a analisar as possibilidades buscando uma solução ótima global. Dessa forma, esse modelo se basearia em aprendizado buscando minimizar o número de itens selecionados para atingir uma medida de proficiência com precisão.

2. Seleção de Itens em Testes Adaptativos

2.1. Teoria da Resposta ao Item (TRI)

A Teoria da Resposta ao Item (TRI) trabalha com diversos modelos probabilísticos que atendem aos diversos testes e avaliações aplicados em todo tipo de área: testes psicológicos, avaliações educacionais, indicadores socioeconômicos, escalas de concordância ou satisfação e várias outras medidas. Devido ao grande número de modelos existentes, cada área de aplicação deve avaliar quais modelos melhor se adaptam às suas necessidades [Baker 2001]. Por exemplo, um questionário para um indicador socioeconômico ou uma prova dissertativa utiliza itens politômicos, ou seja, itens em que há graduações de valores para cada resposta, logo, somente os modelos específicos para esses tipos de itens poderão ser aplicados.
Na TRI os itens apresentam determinadas características chamadas de parâmetros, os quais, em conjunto com a habilidade ou proficiência dos indivíduos, geram uma função de probabilidade de acerto, ou de graduações maiores, quando os itens são respondidos [Baker 2001]. Para avaliações que utilizam questões objetivas, caso dos testes adaptativos, os modelos mais utilizados são os logísticos de um, dois ou três parâmetros. Para avaliações educacionais o modelo logístico de três parâmetros tem sido o mais amplamente utilizado [Linden and Hambleton 1996].
Os três parâmetros desse modelo são chamados de discriminação (parâmetro a), dificuldade (parâmetro b) e acerto ao acaso (parâmetro c). O parâmetro a mostra o quanto o item discrimina entre indivíduos em diferentes regiões da escala de habilidade. O parâmetro b mostra o ponto em que o item se localiza na escala de habilidade, ou seja, o nível de conhecimento que um indivíduo precisa ter para que, provavelmente, consiga acertar o item. O parâmetro c, apesar de chamado de acerto ao acaso (e muitas vezes de ”chute”), é, na verdade, a probabilidade de um indivíduo de habilidade muito baixa acertar o item.

A função de probabilidade de acerto de um item, dada uma proficiência θ de um indivíduo, é definida por

$$P = c + \frac{1 - c}{1 + \exp^{-a(\theta - b)}} \tag{1}$$

de onde obtemos a Curva Característica do Item (figura 2) [Baker 2001].

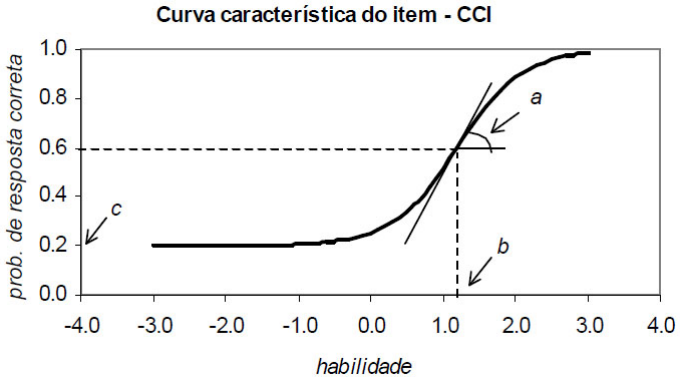

Figure 2. Curva Característica do Item

2.2. Seleção Baseada em Estratificação

Os critérios de seleção mais simples e primordiais nos testes adaptativos foram os de análise de dificuldade. A ideia se resume a classificar os itens em faixas de dificuldade e selecionar o próximo item aleatoriamente em uma faixa de dificuldade acima da anterior quando o indivíduo acerta a resposta ou em uma faixa de dificuldade abaixo da anterior quando o indivíduo erra. Assim, o indivíduo atinge uma sequência de acertos e erros que indica a região em que se encontra sua proficiência.

Esse tipo de seleção está sujeito à subjetividade, dependendo da opinião da pessoa que classifica os itens, e é considerado ineficiente, uma vez que, de acordo com a amplitude da faixa, a diferença entre os itens do início e fim da faixa pode ser grande o suficiente para que não sejam considerados como pertencentes ao mesmo nível. Além disso, os itens que estão nos limites das faixas têm um comportamento mais parecido com os das faixas adjacentes do que com os itens do interior da sua própria faixa. Quando a proficiência do indivíduo se encontra perto do limite de uma faixa esse mesmo problema se repete, pois o próximo item selecionado pode estar no outro limite da faixa de seleção.

Para tentar amenizar esse tipo de problema, um método de estratificação do banco de itens foi proposto [Chang and Ying 1999], onde os itens são divididos em diferentes estratos baseado nos valores dos parâmetros dos itens e o teste adaptativo é dividido em estágios. O banco de itens é dividido em alguns estratos de acordo com a ordem ascendente do parâmetro de discriminação. Em seguida, o banco de itens é ordenado pelo parâmetro de dificuldade dos itens e dividido em pequenos estratos, formando vários subgrupos de itens classificados por parâmetro de dificuldade e discriminação.

No início do teste, são administrados itens do estrato de menor discriminação selecionados dentro do estrato de dificuldade que está sendo utilizado. À medida que o teste evolui, serão selecionados itens com melhor discriminação que pertencem aos diferentes estratos

de dificuldade. O que justifica o uso dessa estratégia é a imprecisão da estimação de proficiência no início do teste, assim, os itens com melhores valores de discriminação são deixados para as fases finais do teste onde a precisão do resultado deve ser maior.

2.3. Seleção Baseada em Informação do Item

Devido à necessidade de estabelecer métodos que não dependessem de uma classificação baseada em conceitos subjetivos, como na classificação por estratos, e que tomassem como base a matemática, a seleção baseada na medida de informação do item tornou-se a mais utilizada nos testes adaptativos e, entre eles, destaca-se o método baseado no critério de Máxima Informação de Fisher [Veldkamp 2010]. Para o modelo logístico de três parâmetros, a medida de informação de um item, dada uma proficiência θ de um indivíduo, é definida por

$$I(\theta) = D^2.a^2.[\frac{Q(\theta)}{P(\theta)}].[\frac{P(\theta) - c^2}{1 - c^2}] \tag{2}$$

onde *P(θ)* é a probabilidade de acerto do indivíduo pelo modelo de três parâmetros (1) e *Q(θ)* = 1 - *P(θ)* [Baker 2001] e *D* é uma constante de aproximação da curva logística à curva de distribuição normal .

O método de máxima informação trabalha com uma solução local para a seleção de itens, levando em consideração apenas a estimação atual da proficiência para selecionar o próximo item. Não há nenhuma indicação de aprendizado ou, pelo menos, de uma sequência de escolhas para adicionar dados auxiliares que possam auxiliar na decisão da seleção do item. Dessa forma, o método garante apenas que, para a proficiência atual, o próximo item é o melhor a ser aplicado. Porém, como a proficiência verdadeira do indivíduo é desconhecida e pode estar longe do ponto atual, a informação que esse item agrega ao teste, muitas vezes, acaba sendo ineficaz.

Esse método também mostra um problema devido à seleção repetitiva dos itens que produzem maior informação em determinados pontos da escala. Quando a estimação de proficiência de indivíduos estiver em uma mesma região, o mesmo item será selecionado. Esse tipo de problema, chamado de superexposição do item, deve ser controlado, garantindo que o número de vezes em que os itens são aplicados seja equilibrado.

Um método de controle de exposição, chamado método de Sympson-Hetter, define uma probabilidade de exposição, sendo essa probabilidade inversamente proporcional ao número de vezes em que esse item é selecionado para aplicação. Através de várias simulações de seleção de itens, e mesmo pelos testes realizados posteriormente, são definidos quais são os itens mais selecionados e, assim, são calculadas as probabilidades que limitam se esse item será aplicado ou não [Hetter and Sympson 1997]. Quanto mais um item é selecionado nessas simulações, menor será a probabilidade dele ser aplicado em um teste futuro. Quando um indivíduo participa de um teste e um item é selecionado, esse item só será aplicado caso passe em um teste simples de resultado dentro dessa probabilidade de exposição. Normalmente, 20% é o valor mínimo utilizado para controle na aplicação do item.

Assim, estabelecemos a seleção por critério de Máxima Informação como:

$$x = 1(itemselecionado) \tag{3}$$

se

$$Ix(\theta) = MaxI(\theta)$$
$$P_{random} <= P_{aplic}$$

Entretanto, quando se impede a seleção de um item por superexposição há interferência direta no critério de Máxima Informação, uma vez que o item de maior informação deixa de ser utilizado.

3. Sistema de Simulação de TAC's

Nossa proposta se baseia em solucionar o problema de seleção de itens em TAC's através de modelos de aprendizado e otimização da área de inteligência computacional. Acreditamos que um modelo que contemple essas duas características funcionaria de forma mais eficiente e conseguiria atingir os problemas definidos em relação à seleção baseada em medida de informação.

Inicialmente o foco do trabalho é simular o funcionamento dos métodos já existentes, analisar seus resultados e tentar estabelecer metas as quais um sistema computacional inteligente deva alcançar. Para isso, o primeiro passo foi a seleção de aproximadamente 600 itens de Língua Portuguesa para compor o banco de itens utilizado nas simulações. Essa seleção é realizada segundo determinados critérios para garantir a qualidade e a precisão do TAC: abranger todos os tipos de habilidade em toda a escala de dificuldade, atingir um valor mínimo para o parâmetro de discriminação de 0,5 e um valor máximo para o parâmetro de acerto casual de 20%.[Flaugher 2000]

O segundo passo no desenvolvimento do sistema foi definir o modelo de estimação da proficiência do indivíduo. Atualmente, na área de avaliação educacional, os modelos mais utilizados na estimação das proficiências são Bayesianos [Costa 2009]. Esses modelos combinam a função de verossimilhança que relaciona a proficiência e as respostas do examinando com uma distribuição a priori para os valores desconhecidos da proficiência, e o escolhido para as simulações foi o da Média a Posteriori (EAP).

Como o objetivo de um TAC é produzir uma medida de proficiência com precisão, devemos também definir um critério para essa precisão, o que acaba por se tornar a meta que o sistema deve alcançar. O erro-padrão definido em 0,25 é suficiente para considerar que o teste adaptativo de um indivíduo convergiu para uma estimativa correta de proficiência.[Linden and Glas 2000]

Entre as definições básicas para o funcionamento de um sistema de TAC, o limite do número de itens a serem aplicados no teste é uma das principais características a serem definidas. Um TAC que necessite de um número muito alto de itens para atingir a precisão de resultado desejada mostra claramente que é ineficiente. Um limite máximo de 30 itens por teste foi definido para as nossas simulações, uma vez que esse valor é considerado razoável para testes do tipo papel-e-caneta.

3.1. Simulação por Estratificação

No procedimento de estratificação os itens foram divididos em grupos de mesmo número de itens, sendo 10 grupos de acordo com a dificuldade e 3 grupos de acordo com a

discriminação, gerando assim 30 subgrupos que são acessados de acordo com o estágio do teste e a estimação atual de proficiência, escolhendo um item aleatoriamente dentro do subgrupo indicado.
Após uma simulação de mil testes, foi constatado que 62,1% deles atingiram o valor esperado do erro-padrão e, dentro desse grupo, a média de itens necessários para isso foi de aproximadamente 29, com um desvio-padrão de 0,99. Podemos observador que o ganho médio em relação ao limite estabelecido de trinta itens praticamente inexiste, além de aproximadamente 38% dos testes não conseguiram atingir o critério, o que demonstra uma grande ineficiência do método.

3.2. Simulação por Medida de Informação

Para os testes de medida de informação, o simulador se baseava na probabilidade de exposição de Sympson-Hetter para permitir ou não que o item selecionado fosse aplicado. Os valores dessa probabilidade variavam entre 20% e 70% aproximadamente, garantindo que os itens de maior informação tivessem suas chances de aplicação reduzidas.
Também foram simulados mil testes, em que 95,4% deles atingiram o valor esperado do erro-padrão com uma média de aproximadamente 20 itens para isso, e um desvio-padrão de 2,5. Podemos observar que o ganho médio em relação ao limite estabelecido de trinta itens foi de um terço do teste, um valor considerável. Porém, quase 5% das simulações não conseguiram atingir o critério, o que demonstra uma pequena ineficiência do método em alguns casos.

3.3. Simulação por Inteligência Computacional

Para o desenvolvimento de um sistema inteligente de seleção de itens em TAC's, será necessária uma análise dos métodos já existentes. Como o método de estratificação não se mostrou eficiente, essa análise deve se basear nos resultados do método de medida de informação. Assim, um novo sistema deve atingir números parecidos com os obtidos nessa simulação.
Primeiramente devemos observar que aproximadamente 5% dos testes não atingiram o erro-padrão desejado. Dentre os testes que atingiram o objetivo, cerca de 95% deles (dois desvios-padrões) utilizaram até 25 itens para isso. Assim, podemos estabelecer uma primeira restrição para o novo método: os testes devem atingir o erro-padrão definido utilizando até 25 itens.
A principal restrição para o modelo seria o próprio erro-padrão definido inicialmente, porém, sabemos que esse erro só pode ser atingido se a seleção de itens for feita de forma correta. Caminhar na direção errada no momento da seleção dos itens não traz precisão para a medida. Assim, uma solução possível para o problema seja estabelecer uma taxa de redução do erro-padrão a cada seleção de item, lembrando que essa redução não é linear, ou seja, nas fases iniciais do teste o erro-padrão diminui mais rapidamente do que nas fases finais do teste.
Além disso, é necessário definir um espaço de busca da solução, isto é, o próximo item selecionado deve ser limitado a uma determinada região, para que o fator de aleatoriedade na seleção não prejudique o sistema. Utilizar a própria estimação atual de proficiência e do erro-padrão como limite dessa região é uma opção viável para a seleção.

O modelo também deve restringir a repetição de itens que meçam o mesmo tipo de habilidade, garantindo uma variedade no teste de muitos tipos de conhecimentos necessários ao avaliando.
Assim, podemos definir que o um novo modelo deve minimizar o número de itens de forma a obter um erro-padrão final menor ou igual ao erro-padrão definido para a convergência, buscando o próximo item dentro de uma região de limite pré-estabelecido maximizando a queda do erro a cada passo (4).

$$\textit{Min(i)} \tag{4}$$

$$ep_i \leq ep_{conv}$$

$$Max(\frac{ep_{i-1}}{ep_i})$$

$$(\theta_{i-1} - ep_{i-1}) \leq i \leq (\theta_{i-1} + ep_{i-1})$$

A solução dos problemas de seleção de itens em Testes Adaptativos Computadorizados ainda é uma área quase que exclusiva da estatística. Assim, o desenvolvimento de um novo modelo deve, no mínimo, visar os mesmos resultados obtidos com os métodos atuais sem abrir mão da eficiência computacional.

4. Referências

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, 1st edition.

Chang, H. H. and Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23.

Costa, D. R. (2009). Métodos estatísticos em testes adaptativos informatizados. Master's thesis, Universidade Federal do Rio de Janeiro.

Flaugher, R. (2000). Item pool. In Wainer, H., editor, *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates.

Hetter, R. D. and Sympson, B. (1997). Item exposure control in cat-asbav. In W. A. Sands, B. K. W. and McBride, J. R., editors, *Computerized Adaptive Testing: from inquiry to operation*. American Psychological Association.

Linden, W. J. V. D. and Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Kluwer, 1st edition.

Linden, W. J. V. D. and Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory*. Springer, 1st edition.

Pasquali, L. (1996). *Teoria e métodos de medida em ciências do comportamento*. UnB : Inep, 1st edition.

Thurstone, L. (1959). *The measurement of values*. Chicago University Press, 1st edition.

Veldkamp, B. P. (2010). Bayesian item selection in constrained adaptive testing using shadow tests. *Psicologica*.