CENTRO UNIVERSITÁRIO DA FEI

DOUGLAS DE RIZZO MENEGHETTI

METODOLOGIA DE SELEÇÃO DE ITENS EM TESTES ADAPTATIVOS INFORMATIZADOS BASEADA EM AGRUPAMENTO POR SIMILARIDADE

São Bernardo do Campo

2015

DOUGLAS DE RIZZO MENEGHETTI

METODOLOGIA DE SELEÇÃO DE ITENS EM TESTES ADAPTATIVOS INFORMATIZADOS BASEADA EM AGRUPAMENTO POR SIMILARIDADE

Dissertação de Mestrado, apresentada ao Centro Universitário da FEI para obtenção do título de Mestre em Engenharia Elétrica. Orientado pelo Prof. Dr. Plinio Thomaz Aquino Junior.

São Bernardo do Campo
2015

Meneghetti, Douglas De Rizzo
Metodologia de seleção de itens em testes adaptativos informatizados baseada em agrupamento por similaridade / Douglas De Rizzo Meneghetti. São Bernardo do Campo, 2015.
96 f. : il.

Dissertação de Mestrado - Centro Universitário da FEI.
Orientador: Prof. Dr. Plinio Thomaz Aquino Junior

1. Agrupamento por Similaridade. 2. Teoria da Resposta ao Item. 3. Testes Adaptativos Informatizados. I. Aquino Junior, Plinio Thomaz, orient. II. Título.

CDU 37:681.3

CENTRO UNIVERSITÁRIO DA FEI

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno: Douglas de Rizzo Meneghetti **Matrícula:** 113302-4

Título do Trabalho: Metodologia de seleção de itens em testes adaptativos informatizados baseada em agrupamento por similaridade.

Área de Concentração: Inteligência Artificial Aplicada à Automação

Orientador: Prof. Dr. Plinio Thomaz Aquino Júnior

Data da realização da defesa: 31/08/2015

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

São Bernardo do Campo, / / .

MEMBROS DA BANCA EXAMINADORA	
Prof. Dr. Plinio Thomaz Aquino Júnior	Ass.: _____
Prof. Dr. Carlos Eduardo Thomaz	Ass.: _____
Prof. Dr. Ocimar Munhoz Alavarse	Ass.: _____

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

VERSÃO FINAL DA DISSERTAÇÃO

APROVO A VERSÃO FINAL DA DISSERTAÇÃO EM QUE FORAM INCLUÍDAS AS RECOMENDAÇÕES DA BANCA EXAMINADORA

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

Dedico este trabalho aos meus pais.

AGRADECIMENTOS

Agradeço primeiramente a meus pais, os quais me apoiaram em minha decisão de ingressar no programa de Mestrado e continuaram a me apoiar no decorrer dele. Esse trabalho não existiria, não fosse por eles.

Agradeço a meu orientador, por aceitar acompanhar-me durante este período de aprendizado, contribuindo para minha evolução como aluno, pesquisador e pessoa.

Agradeço a todos os meus professores do programa de Mestrado, cujas aulas eu assisti e peço desculpas pelas inconveniências que causei. Agradeço especialmente ao prof. Carlos Eduardo Thomaz, por me encorajar a ingressar e continuar no programa de Mestrado, e ao prof. Reinaldo Bianchi, por me apontar tanto o caminho do método científico como o das boas epígrafes.

Agradeço a meus colegas do Grupo de Estudos e Pesquisa em Avaliação Educacional da Faculdade de Educação da Universidade de São Paulo pelas oportunidades que me deram e pela paciência. Agradeço especialmente ao prof. Ocimar Munhoz Alavarse por ter aberto as portas a mim e tê-las mantido abertas pelos últimos 4 anos.

Agradeço aos colegas que me acompanharam e aconselharam no dia-a-dia do programa de Mestrado, os quais o mesmo programa me deu a oportunidade de conhecer, sejam eles colegas do mundo da pesquisa ou não. Um agradecimento especial ao Vagner do Amaral, o qual conheci graças a uma sucessão pouco provável de eventos e graças ao qual não estaria aqui hoje, não fosse por um convite dele; ao Leonardo Anjoletto Ferreira, que estimulou-me de maneira sem igual na busca pela solução do problema dessa dissertação; e ao Andrey Araujo Masiero, pela companhia, paciência e ajuda durante todo o período do programa de Mestrado.

Por fim, um agradecimento aos órgãos de apoio à pesquisa que me garantiram a oportunidade de participar do programa de Mestrado. Em ordem cronológica: o Centro Universitário da FEI, CAPES, FINEP e CAPES novamente. Agradeço especialmente à Secretaria de Pós-Graduação Stricto Sensu da FEI, por se aventurar comigo em todas as burocracias que essa sucessão de trocas de bolsas de estudos causou.

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be."
William Thomson, 1º Barão Kelvin

RESUMO

Testes Adaptativos Informatizados são avaliações educacionais aplicadas em um computador, cuja escolha dos itens é feita em tempo real, baseada nas respostas do examinando aos itens anteriores. Esta característica garante aos testes adaptativos maior precisão na estimativa da proficiência do examinando com a aplicação de menos itens. Diversas metodologias foram criadas para balancear a precisão na estimativa da proficiência com o uso homogêneo do banco de itens, a maioria delas fundadas na Teoria da Resposta ao Item, uma série de modelos psicométricos criados especificamente para mensurar traços latentes dos indivíduos e modelar probabilisticamente o comportamento desses indivíduos durante a aplicação de um teste. Este trabalho apresenta uma nova metodologia de seleção de itens para aplicação de Testes Adaptativos Informatizados, baseada no agrupamento por similaridade de um banco de itens, representados por seus parâmetros sob o modelo logístico de 3 parâmetros da Teoria da Resposta ao Item. Diversos algoritmos de agrupamento tradicionais são aplicados, seus resultados são avaliados utilizando medidas de validação disponíveis na literatura e, por fim, são alimentados à metodologia proposta, a qual é validada através de um estudo de simulações de aplicações de Testes Adaptativos Informatizados sob diversas restrições de exposição dos itens. O estudo de simulações demonstra que o número de grupos no qual o banco de itens é separado possui influência sobre a precisão da estimativa das proficiências dos indivíduos e também sobre a taxa de sobreposição do teste, permitindo que precisão e segurança do banco de itens possam ser controlados através da variação no número de partições.

Palavras-chave: Agrupamento por Similaridade. Testes Adaptativos Informatizados. Teoria da Resposta ao Item.

ABSTRACT

Computerized Adaptive Tests are educational evaluations applied using a computer, whose choice of items is done in real time, based on responses from the previous items answered by the examinee. This feature ensures greater accuracy in estimating examinees' abilities with the use of fewer items. Several methodologies were created to balance the accuracy in estimating abilities with the homogeneous use of the item database, most of them based on Item Response Theory, a series of psychometric models created specifically to measure latent traits of individuals and probabilistically model the behavior of these individuals during the application of a test. This work presents a new item selection methodology for the application of Computerized Adaptive Tests, which is based on the clustering of the items in the item pool, represented by their parameters under the three-parameter logistic model of Item Response Theory. Several traditional clustering algorithms are applied, their results are evaluated using validation measures available in the literature and finally are fed to the proposed method, which is validated by a simulation study of adaptive testing applications under various item exposure restrictions. The simulation study showed that the number of groups in which the item bank is separated has influence on the accuracy of the estimation of examinees' skills and on the test overlap rate, allowing both variables to be controlled by varying the number of partitions.

Keywords: Clustering. Computerized Adaptive Tests. Item Response Theory.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ASVAB	*Armed Services Vocational Aptitude Battery*.
CISM	*Cluster-based Item Selection Method*.
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*.
EAP	Expectativa a Posteriori.
ETS	*Educational Testing Systems*.
GA	Algoritmo Genético.
GRNN	*Generalized Regression Neural Network*.
MAP	Máximo a Posteriori.
MIS	*Maximum Information Stratification*.
MIS-B	*Maximum Information Stratification with Blocking*.
ML1	modelo logístico de 1 parâmetro.
ML2	modelo logístico de 2 parâmetros.
ML3	modelo logístico de 3 parâmetros.
MVC	Máxima Verossimilhança Conjunta.
MVM	Máxima Verossimilhança Marginal.
SVM	*Support-Vector Machines*.
TAI	Teste Adaptativo Informatizado.
TCT	Teoria Clássica dos Testes.
TRI	Teoria da Resposta ao Item.

LISTA DE SÍMBOLOS

a	Discriminação de um item sob os modelos logísticos da Teoria da Resposta ao Item.
b	Dificuldade de um item sob os modelos logísticos da Teoria da Resposta ao Item.
c	Probabilidade de acerto ao acaso de um item sob os modelos logísticos da Teoria da Resposta ao Item.
D	Índice de Dunn.
g	Grupo de elementos.
I	Função de informação de um item.
k	Número de grupos; Número de sementes para o *k-means*.
P	Função de resposta ao item.
r	Taxa de exposição de um item.
r^{max}	Taxa de exposição máxima.
$RMSE$	Raíz da soma dos erros quadráticos médios.
T	Taxa de sobreposição do teste.
θ	Proficiência.
$\hat{\theta}$	Proficiência estimada.
W	Soma dos erros quadráticos; soma das variâncias intra-grupos.
X	Banco de itens.
$\boldsymbol{\zeta}$	Conjunto de parâmetros de um modelo da Teoria da Resposta ao Item.

SUMÁRIO

1 INTRODUÇÃO

O uso do instrumento de avaliação é uma forma comum de se medir fatores e características comumente não mensuráveis, como conhecimento em determinado assunto, xenofobia e usabilidade de software. Usualmente, um instrumento de avaliação é composto de uma série de itens, os quais devem ser respondidos pela população que se deseja avaliar, possibilitando que os indivíduos sejam posteriormente posicionados em uma escala que determina até qual ponto um indivíduo da população domina ou concorda com a característica sendo avaliada. Devido ao fato dessas características não serem diretamente mensuráveis, dá-se a elas o nome de *traços latentes*.

Diversas ferramentas psicométricas existem para auxiliar o avaliador na tarefa de se medir os traços latentes. A Teoria Clássica dos Testes (TCT), desenvolvida inicialmente por Spearman (1950), visava principalmente suprir a necessidade da área psicológica em testes de aptidão, por isso possui um viés significativo na medição da inteligência dos indivíduos (PASQUALI, 2003). Apesar de primitiva, a TCT é capaz de proporcionar grande quantidade de informação referente aos resultados dos indivíduos. Ela utiliza como medida principal a pontuação total do examinando no teste, ou seja, a soma de todos os seus acertos, possuindo como medida suplementar a variância entre as pontuações de examinandos diferentes.

A Teoria da Resposta ao Item (TRI) (LORD; NOVICK; BIRNBAUM, 1968; RASCH, 1966; RASCH, 1980), criada na década de 70, foge dos preceitos da TCT em medir a aptidão intelectual do indivíduo, sendo elaborada para medir traços latentes em geral, ao mesmo tempo aprimorando as medidas já presentes na TCT. A TRI permite descobrir diversas características de cada item, como sua dificuldade, o quão bem o item discrimina indivíduos de níveis de proficiência próximos e, inclusive, a probabilidade de se acertar o item ao acaso. O conhecimento desses parâmetros permite ao avaliador criar instrumentos de avaliação mais especializados, como, por exemplo, um teste voltado somente para indivíduos de alta proficiência em determinado assunto, aumentando a precisão com a qual a proficiência desses indivíduos é medida e, consequentemente, discriminando melhor cada um na escala que mede o traço latente.

Contudo, apesar de munidas dessas ferramentas, as avaliações em formato fixo, como provas em lápis-e-papel, sofrem de um problema: por serem padronizadas, elas obrigam todos os indivíduos a responderem todos os itens; ora, se um indivíduo de baixa proficiência é obrigado a responder itens de dificuldade muito elevada, as respostas erradas a estes itens geram pouca informação relacionada à estimativa de sua proficiência. O mesmo acontece com indi-

víduos de alta proficiência diante de uma prova com itens que exigem baixa proficiência: seus acertos aos itens mais fáceis revelam pouco com relação a suas proficiências. Esta situação se agrava no contexto das avaliações em larga escala, como vestibulares ou avaliações da qualidade do ensino de determinada região geográfica, na qual uma quantidade muito heterogênea de indivíduos é sujeita ao mesmo tipo de teste (VAN DER LINDEN; GLAS, 2000).

Observando isso, na década de 70 surgiu a hipótese da utilização de um teste cujos itens se adaptem às respostas do indivíduo para aferir seu nível de proficiência com maior precisão e menos itens (LORD, 1977). Neste novo modelo, batizado de teste adaptativo, novos itens são escolhidos à medida que o examinando responde aos itens anteriores. Desta forma, um examinando que demonstre alto grau de proficiência pode ser apresentado a itens de dificuldade gradualmente mais elevada, evitando assim que itens cujas respostas sejam de pouco valor, ou até mesmo irrelevantes, interfiram no teste, alongando-o desnecessariamente e cansando o examinando. Com o avanço da tecnologia e a maior disponibilidade de computadores e outros dispositivos de apresentação de mídia, este novo paradigma de testes ganhou novos horizontes, passando a se chamarem Teste Adaptativo Informatizado (TAI). Aliados a bancos de itens cujas características já são conhecidas de acordo com algum modelo estatístico da TRI, os TAI são capazes de gerar testes personalizados automaticamente através dos parâmetros dos itens e estimativas em tempo real da proficiência dos indivíduos.

No entanto, em situações reais, os itens que compõem a base que será utilizada para aplicação do teste adaptativo não possuem características ótimas (MOREIRA JUNIOR, 2012). O banco de itens pode conter menos itens de determinadas dificuldades, impedindo que todos os indivíduos sejam avaliados com a mesma precisão, ou pode conter itens cujas respostas não sejam informativas, por exemplo, itens que são respondidos correta ou incorretamente por indivíduos de diversos níveis de proficiência. Devido a essas limitações, critérios para a escolha automática dos itens que serão aplicados em tempo real devem ser estudados cuidadosamente. Esses critérios devem assegurar a maior precisão possível na estimativa da proficiência dos indivíduos, ao mesmo tempo em que utiliza de forma homogênea os itens disponíveis para seleção no banco de itens.

Diversas metodologias de aplicação de TAI foram criadas (CHANG; QIAN; YING, 2001; CHANG; VAN DER LINDEN, 2003; CHANG; YING, 1996; BARRADA; MAZUELA; OLEA, 2006; BARRADA; OLEA; PONSODA, 2007; BARRADA; OLEA; ABAD, 2008; BARRADA et al., 2009; BARRADA; ABAD; VELDKAMP, 2009; BARRADA et al., 2010), visando balancear duas variáveis que, teoricamente, são inversamente proporcionais: a precisão com a qual a proficiência dos examinandos é aferida, a qual depende diretamente da esco-

lha dos melhores itens possíveis a cada instante do teste, e a utilização homogênea de todos os itens disponíveis para seleção no banco de itens, garantindo que haja pouca repetição nos itens utilizados. Na literatura, no entanto, não foram encontrados trabalhos que realizassem o agrupamento de itens utilizando medidas de similaridade, com o propósito de selecionar itens semelhantes durante a aplicação de um TAI, garantindo a diversidade na escolha dos itens ao mesmo tempo que itens semelhantes aos itens ótimos sejam aplicados a cada instante do teste. Este trabalho apresenta uma análise da aplicação de diferentes algoritmos de agrupamento em diferentes bases de itens calibradas pela TRI para utilização durante a aplicação de um TAI.

1.1 JUSTIFICATIVA

A aplicação de técnicas de agrupamento por similaridade em bases de itens calibrados pela TRI pode ser justificada pelo fato dos itens não possuírem nenhuma classificação a priori, nem fazerem parte de grupos similares, quando seus parâmetros são estimados. Caso alguma classificação exista (e.g, por níveis de dificuldade), ela foi fruto de demasiado trabalho por parte dos criadores do banco de itens. Uma vez que as técnicas de agrupamento por similaridade são de natureza não-supervisionada, não necessitando de dados pré-agrupados ou classificados como base de treinamento ou informação a priori, qualquer tipo de informação a priori relacionada ao conteúdo ou tema da avaliação se mostra desnecessária para a aplicação do método proposto, dependendo apenas dos parâmetros dos itens, parâmetros estes que são pré-requisitos para a aplicação de qualquer TAI. Adicionalmente, da mesma forma que o agrupamento por similaridade não foi estudado no agrupamento de itens de acordo com seus parâmetros sob os modelos logísticos da TRI, o uso das informações provenientes do processo de agrupamento na aplicação de um teste adaptativo também não o foi, sendo este um objetivo secundário deste trabalho.

1.2 OBJETIVO

Este trabalho tem por objetivo propor uma nova metodologia de seleção de itens para uso em Teste Adaptativo Informatizado que utiliza itens calibrados pelo modelo logístico de 3 parâmetros da Teoria da Resposta ao Item e unidos por diferentes algoritmos de agrupamento.

1.2.1 Objetivos específicos

Analisar os principais métodos de aplicação de testes adaptativos disponíveis na literatura.

Realizar o agrupamento de um banco de itens através de seus parâmetros, de acordo com o modelo logístico de 3 parâmetros da Teoria da Resposta ao Item e avaliar os resultados dos agrupamentos utilizando as medidas de avaliação apropriadas.

Utilizar os resultados do processo de agrupamento em um processo de validação por simulação de Testes Adaptativos Informatizados.

1.3 METODOLOGIA

Primeiramente, foi realizada a revisão bibliográfica do principal tema do trabalho, os testes adaptativos. Visto que virtualmente todas as implementações de testes adaptativos dependiam de algum modelo disponibilizado pela TRI, foi feita uma revisão bibliográfica da mesma. Também foram compilados os principais trabalhos da área da Inteligência Artificial que propunham métodos de aplicação de testes adaptativos.

Terminada esta fase e visto que não foram encontradas tentativas de utilizar algoritmos de agrupamento por similaridade em bases de itens para uso em testes adaptativos, foi realizada uma revisão bibliográfica do tema, assim como a escolha de algoritmos que representassem as principais abordagens de agrupamento descritas na literatura.

O próximo passo consistiu na utilização de uma base sintética de itens, cujas características simulam aquelas de itens reais calibrados sob o modelo logístico de 3 parâmetros da Teoria da Resposta ao Item e são utilizadas na comparação de diferentes técnicas de aplicação de testes adaptativos. A seguir, diferentes algoritmos de agrupamento foram utilizadas para agrupar estes itens, de acordo com os 3 parâmetros disponibilizados pelo modelo. Os grupos de itens foram então utilizados em simulações de testes adaptativos para avaliar a precisão do método proposto, assim como a homogeneidade na seleção dos itens da base.

1.4 ORGANIZAÇÃO DO TRABALHO

A seção 2 faz uma revisão dos principais temas abordados no trabalho: a Teoria da Resposta ao Item, uma coleção de modelos probabilísticos utilizados na modelagem do comportamento de respondentes a uma série de itens; os Testes Adaptativos Informatizados, tes-

tes em cuja escolha dos itens é feita em tempo real, com o intuito de melhorar a precisão da estimativa das proficiências dos respondentes; e os algoritmos de agrupamento, técnicas não-supervisionadas que organizam dados em grupos baseadas em medidas de similaridade entre os dados. A seção 3 descreve a metodologia do trabalho: a realização do agrupamento de itens, utilizando como dados de entrada os parâmetros dos itens extraídos através do modelo logístico de 3 parâmetros da Teoria da Resposta ao Item e a subsequente utilização dos agrupamentos na aplicação de uma série de Testes Adaptativos Informatizados, comparando a precisão da estimativa das proficiências dos respondentes feitas pelo método proposto com os resultados de experimentos semelhantes disponíveis na literatura. A seção 4 apresenta os resultados do trabalho e a seção 5 discorre sobre as conclusões dos experimentos e possíveis trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Esta seção discorre acerca das principais técnicas utilizadas no decorrer deste trabalho: a Teoria da Resposta ao Item, os Testes Adaptativos Informatizados e os algoritmos de agrupamento.

2.1 TEORIA DA RESPOSTA AO ITEM

Criada em paralelo pelos pesquisadores da *Educational Testing Systems* (ETS) Lord, Novick e Birnbaum (1968) e Rasch (RASCH, 1966; RASCH, 1980), a TRI é um conjunto de modelos matemáticos da Psicometria criados com o intuito de mensurar características não observáveis de um indivíduo, denominadas traços latentes (ANDRADE; TAVARES; VALLE, 2000). Exemplos de traços latentes os quais a TRI possibilita mensurar incluem a capacidade de realizar somas, proficiência em um idioma estrangeiro, usabilidade de sistemas (TEZZA, 2009; TEZZA; BORNIA; MOREIRA JUNIOR, 2009; TEZZA, 2012) etc.

Uma vez que caracteriza um processo de medição, a TRI necessita obrigatoriamente de um instrumento de medição, assim como uma unidade de medida, ambos relacionados ao traço latente que se deseja medir. Como exemplo de um instrumento capaz de medir a proficiência de um indivíduo em realizar somas, pode-se citar uma prova de matemática, sendo cada questão da prova (assim como de qualquer instrumento de medida utilizado pela TRI) denominado item.

Um item pode ser definido de maneira mais formal como uma tarefa que exteriorize o traço latente que se deseja medir. Os itens, por sua vez, podem ser ordenados entre si, formando uma escala, a qual pode ser visualizada em primeira instância como a régua que mede o traço latente. Itens que exteriorizem o traço latente de forma menos acentuada são localizados em um extremo desta escala, enquanto itens que representem o traço de forma mais densa são posicionados no outro extremo. De forma análoga, em uma prova de matemática, itens menos desafiadores constituem os primeiros trechos da escala, enquanto aqueles mais difíceis se localizam ao final da escala. Tem-se formada, portanto, a unidade de medida necessária para medir o traço latente.

Para que a qualidade e interpretação dos resultados finais sejam fidedignos, a TRI atua sob duas suposições:

a) **Unidimensionalidade:** As respostas aos itens são influenciadas majoritariamente por apenas um traço latente, e;

b) **Independência local:** As respostas aos itens são independentes entre si.

Sob a perspectiva das avaliações educacionais, é comum chamar o traço latente de "proficiência", nomenclatura esta que será utilizada pelo decorrer do texto.

No âmbito matemático, a TRI permite o cálculo da probabilidade de acerto de um respondente a determinado item, dados alguns parâmetros deste item e o nível de proficiência do respondente, todos representados numericamente. Para isso, a TRI requer que os parâmetros dos itens sejam estimados, seguindo algum de seus modelos estatísticos (ANDRADE; TAVARES; VALLE, 2000; PASQUALI, 2003). Os modelos mais difundidos são os logísticos, os quais serão apresentados a seguir.

2.1.1 Modelos logísticos

O modelo logístico de 3 parâmetros (ML3) (3PL, em inglês *three-parameter logistic model*), proposto por Lord (1980), calcula a probabilidade de um respondente acertar um item dados três parâmetros deste item: sua discriminação, dificuldade e probabilidade de acerto ao acaso. Sua fórmula é a seguinte:

$$P_i(X_i = 1|\theta) = c_i + \frac{(1 - c_i)}{1 + e^{a_i(\theta - b_i)}} \tag{1}$$

onde:

a) $\theta \in (-\infty; \infty)$: Proficiência do respondente;

b) $P_i(X_i = 1|\theta)$: probabilidade de um indivíduo com um determinado θ responder corretamente ao item X_i. Geralmente denotada simplesmente por $P_i(\theta)$;

c) $a_i > 0$: discriminação do item i. Indica o quanto o item discrimina respondentes de valores θ altos daqueles com θ mais baixos;

d) $b_i \in (-\infty; \infty)$: dificuldade do item i. Indica em qual ponto da escala de dificuldade ocorre maior discriminação;

e) $c_i \in (0; 1)$: probabilidade de acerto ao acaso do item i.

Caso c seja fixo em 0, o modelo de 3 parâmetros se reduz ao de 2 parâmetros (ML2), indicando que o modelo desconsidera a probabilidade de acerto ao acaso. Adicionalmente, se o

parâmetro a de todos os itens possuírem um valor fixo, o modelo de 2 parâmetros se reduz ao de 1 parâmetro (ML1), ou seja, ele assume que todos os itens discriminam indivíduos igualmente. Por último, caso os valores de a sejam fixados em 1 para todos os itens, obtêm-se o modelo de Rasch (RASCH, 1980).

Essa fórmula gera um gráfico cartesiano, representado por uma curva sigmoidal, visível na Figura 1, onde a caracteriza o ângulo da curva no ponto de inflexão, b evidencia o valor do ponto de inflexão no eixo das ordenadas e c diminui a diferença entre todos os valores de θ e seus valores b correspondentes, demonstrando a chance de acerto ao acaso.

Adicionalmente, a TRI disponibiliza uma função de informação para o item, cujo valor também é condicionado pelos parâmetros do modelo. Formalizada inicialmente em Lord, Novick e Birnbaum (1968), a função de informação modela o comportamento esperado de um item quando aplicado a indivíduos de diferentes níveis de proficiência: caso a dificuldade do item esteja próxima do nível de proficiência do examinando, a resposta (independente de ser correta ou incorreta) resulta na aquisição de mais informação referente à proficiência deste indivíduo do que itens cujas dificuldades sejam muito superiores ou inferiores à proficiência do indivíduo, casos nos quais há maior previsibilidade na resposta.

No modelo de um parâmetro, por exemplo, a função de informação é dada por $I(\theta) = P(\theta)Q(\theta)$ (DE AYALA, 2009), onde $Q(\theta) = 1 - P(\theta)$, ou seja a probabilidade de acerto ao item multiplicada pela probabilidade de errá-lo. No modelo de dois parâmetros, onde a discriminação a não é fixa, a separação entre respondentes que dominam a proficiência medida pelo item daqueles que não a dominam é refletida diretamente pelo valor de a. Essa medida de aleatoriedade nas respostas se reflete na função de informação do modelo de dois parâmetros, onde a possui característica exponencial (DE AYALA, 2009): $I(\theta) = a^2 P(\theta)Q(\theta)$. Para o modelo de 3 parâmetros, a função de informação é dada por (DE AYALA, 2009):

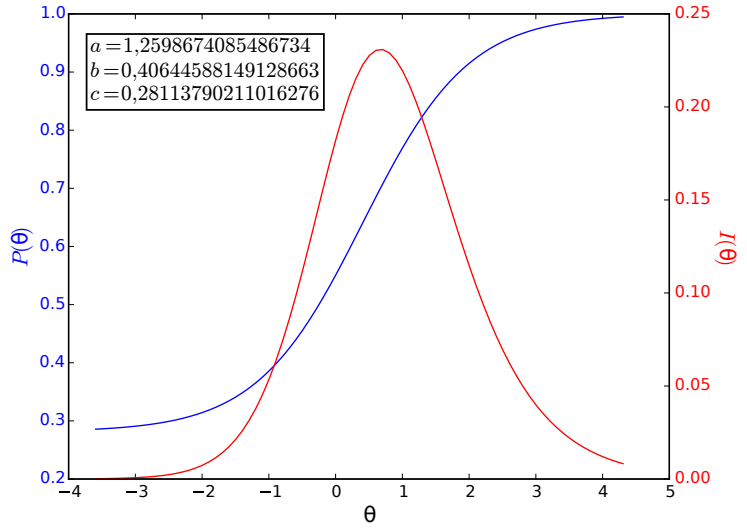
$$I(\theta) = a^2 \frac{(P(\theta) - c)^2}{(1 - c)^2} \cdot \frac{Q(\theta)}{P(\theta).} \tag{2}$$

As funções de informação têm como característica comum formarem um gráfico de distribuição normal, como demonstra a figura 1.

Analogamente, o total de informação de um teste é dado por $\sum_{i=1}^{N} I_i(\theta)$, sendo N o total de itens no teste, o que significa que, quantos mais itens um teste possui, maior a quantidade de informação que pode ser extraída sobre as proficiências do examinando.

Outros métodos podem ser utilizados para a seleção dos itens, contanto que posicionem a proficiência do respondente em uma escala que possibilite a escolha de itens de dificuldade

Figura 1 – Curva característica de um item (CCI) e curva de informação de um item e os valores de seus parâmetros



Fonte: Autor

relacionada a esta proficiência. Os métodos mais relevantes para este trabalho, relacionados aos TAI, são apresentados na seção 2.2.1.2.

2.1.2 Estimativa dos parâmetros

Para se utilizar a TRI, é preciso realizar a estimativa dos parâmetros dos itens. Para tal, é necessário que estes itens sejam utilizados em testes antes de serem calibrados, fase denominada pré-teste. Nesta fase, um teste pode ser composto por uma mistura de itens calibrados com itens não-calibrados, sendo que os parâmetros dos itens calibrados ajudam a estimar os parâmetros dos novos itens. Alternativamente, um teste composto completamente por itens cujos parâmetros são desconhecidos pode ser aplicado. Analogamente, é possível que estes testes sejam respondidos por examinandos cujos valores de θ já sejam conhecidos, facilitando a estimativa dos parâmetros. Todos os valores desconhecidos, sejam eles parâmetros de itens ou proficiências de examinandos, são descobertos através de métodos de estimativa, como por exemplo, as estimativas por máxima verossimilhança e a bayesiana (ANDRADE; TAVARES; VALLE, 2000; MOREIRA JUNIOR, 2012).

A estimativa por máxima verossimilhança tem por objetivo encontrar os valores dos parâmetros do modelo que maximizem a seguinte função de verossimilhança (ANDRADE; TAVARES; VALLE, 2000, cap. 3):

$$L(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P_{ij}(\theta)^{X_{ij}} Q_{ij}(\theta)^{1-X_{ij}} \tag{3}$$

onde:

a) $\boldsymbol{X}$: a matriz binária de respostas de todos os J respondentes a todos os N itens, onde $X_{ij} = 1$ denota que o respondente j concordou, ou respondeu corretamente ao item i. Caso contrário, $X_{ij} = 0$;

b) $\boldsymbol{\theta}$: o vetor das proficiências dos examinandos;

c) $\boldsymbol{\zeta}$: os parâmetros do modelo escolhido. No caso do ML3, estudado neste trabalho, estes parâmetros são a, b e c;

d) $P_{ij}(\theta)$: a probabilidade do respondente j responder corretamente ao item i. No caso da ML3, esta é a função (1);

e) $Q_{ij}(\theta)$: $1 - P_{ij}(\theta)$;

f) X_{ij}: resposta dada pelo respondente j ao item i Caso o respondente responda o item corretamente, $X_{ij} = 1$, caso contrário, , $X_{ij} = 0$.

Para fins computacionais, é comum utilizar o logaritmo da função de máxima verossimilhança:

$$\log L(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta}) = \sum_{i=1}^{N}\sum_{j=1}^{J} \{x_{ij}\log P_{ij}(\theta) + (1-x_{ij})\log Q_{ij}(\theta)\} \tag{4}$$

Os parâmetros que maximizam essa função são aqueles que igualam as derivadas da função de verossimilhança com relação a cada parâmetro a zero. Andrade, Tavares e Valle (2000, cap. 3) dão as equações para maximização dos parâmetros da ML3:

$$\frac{\partial \log L(\boldsymbol{\zeta})}{\partial a_i} : D(1-c_i)\sum_{j=1}^{J}(x_{ij} - P_{ij}(\theta))(\theta_j - b_i)W_{ij} = 0 \tag{5}$$

$$\frac{\partial \log L(\boldsymbol{\zeta})}{\partial b_i} : -Da_i(1-c_i)\sum_{j=1}^{J}(x_{ij} - P_{ij}(\theta))W_{ij} = 0 \tag{6}$$

$$\frac{\partial \log L(\boldsymbol{\zeta})}{\partial c_i} : \sum_{j=1}^{J}(x_{ij} - P_{ij}(\theta))\frac{W_{ij}}{P^*_{ij}} = 0 \tag{7}$$

$$\frac{\partial \log L(\boldsymbol{\zeta})}{\partial \theta_j} : D\sum_{i=1}^{N} a_i(1-c_i)(x_i j - P_{ij}(\theta))W_{ij} = 0 \tag{8}$$

onde:

$$W_{ij} = \frac{P^*_{ij} Q^*_{ij}}{P_{ij} Q_{ij}}$$

$$P^*_{ij} = 1 + e^{-Da_i(\theta_j - b_j)-1}$$

$$Q^*_{ij} = 1 - P^*_{ij}$$

É importante notar que as funções (5) a (8) não possuem solução analítica, sendo que seus resultados devem ser encontrados utilizando métodos iterativos de aproximação, como o de Newton-Raphson.

Quando tanto os parâmetros dos itens como as proficiências dos respondentes são estimados em conjunto, o método utilizado chama-se Máxima Verossimilhança Conjunta (MVC). No entanto, é possível se utilizar de uma abordagem mais simples computacionalmente, denominada Máxima Verossimilhança Marginal (MVM). A MVM separa o processo de estimativa dos parâmetros em duas etapas: na primeira, assume-se uma distribuição normal para os valores de θ, permitindo que os parâmetros dos itens sejam estimados de forma independente das proficiências dos respondentes. Na segunda etapa, os valores de θ dos respondentes são calculados utilizando os parâmetros dos itens estimados na etapa anterior (ANDRADE; TAVARES; VALLE, 2000; MOREIRA JUNIOR, 2012). O uso da MVM justifica-se pelo fato das proficiências de uma população assumir distribuições que aproximam a normal.

É importante ressaltar a incapacidade das abordagens por máxima verossimilhança em estimar as proficiências de respondentes que acertaram ou erraram todos os itens, assim como itens que foram acertados ou errados por todos os respondentes. Dado um indivíduo j com vetor de respostas $\boldsymbol{X_j}$, sua proficiência será estimada da seguinte forma:

$$\theta_i = \left\{ \begin{array}{rl} -\infty & \text{se } \boldsymbol{X_{ij}} = 0 \\ \infty & \text{se } \boldsymbol{X_{ij}} = 1 \end{array} \right\}, j = 1, \ldots, J \tag{9}$$

Existem abordagens Bayesianas que resolvem este problema. Nelas, são assumidas distribuições a priori para os parâmetros do modelo e é escolhida a combinação de parâmetros que satisfaça uma função a posteriori pré-determinada. Andrade, Tavares e Valle (2000) argumentam que, para o parâmetro a, uma distribuição aceitável seria a log-normal, uma vez que ela se restringe a valores positivos; para b, utiliza-se a distribuição normal, pois condiz com as situações práticas, nas quais há maior quantidade de itens nas dificuldades médias do que nas

dificuldades extremas; já c assume a distribuição *Beta*, pelo fato de seus valores se situarem no intervalo [0; 1].

Descobertos os valores dos parâmetros, o próximo passo consiste em estimar os θ dos examinandos. Na abordagem bayesiana, assim como nas abordagens por máxima verossimilhança, assume-se que a distribuição que mais se aproxima do comportamento de θ é a normal (ANDRADE; TAVARES; VALLE, 2000). Ela é, então, assumida como a distribuição a priori para esta variável. A posteriori de θ para o respondente i é então definida em termos do vetor de respostas $\boldsymbol{X_i}$, do conjunto de parâmetros recém-descobertos $\boldsymbol{\zeta}$ e dos parâmetros da distribuição a priori dos valores de θ, dada por η, como:

$$P(\theta|\boldsymbol{X_i},\boldsymbol{\zeta},\eta) = \frac{P(\boldsymbol{X_i}|\theta,\boldsymbol{\zeta})G(\theta|\eta)}{P(\boldsymbol{X_i}|\boldsymbol{\zeta},\eta)} \tag{10}$$

Essa posteriori é utilizada nos métodos Máximo a Posteriori (MAP) ou Expectativa a Posteriori (EAP) para encontrar valores de θ que satisfaçam seus critérios. No MAP, escolhe-se o valor de θ que maximize a equação (10), dada por

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|\boldsymbol{X_i},\boldsymbol{\zeta},\eta) = \frac{P(\boldsymbol{X_i}|\theta,\boldsymbol{\zeta})P(\theta|\eta)}{P(\boldsymbol{X_i}|\boldsymbol{\zeta},\eta)} = P(\boldsymbol{X_i}|\theta,\boldsymbol{\zeta})P(\theta|\eta).$$

O denominador da operação pode ser omitido, uma vez que não depende de θ, não fazendo parte do processo de otimização.

O EAP utiliza a expectativa dessa equação, ou seja,

$$\theta \equiv E[\theta|\boldsymbol{X},\boldsymbol{\zeta},\eta] \approx \frac{\sum_{k=1}^{q} Q_k L(Q_k) W(Q_k)}{\sum_{k=1}^{q} L(Q_k) W(Q_k)},$$

onde $Q_{1...q}$ simbolizam q pontos de quadratura que aproximam as distribuições de θ e $W(Q_k)$, os respectivos pesos para cada ponto, indicando sua representatividade na amostra (BOCK; MISLEVY, 1982).

Esta seção discorreu acerca do ML3 da TRI, explicando os parâmetros do modelo e demonstrando suas funções de probabilidade e informação, assim como o método de estimativa dos parâmetros dos itens e proficiências dos respondentes, dada a matriz binária de respostas dos respondentes aos itens. Posteriormente, os parâmetros deste modelo serão utilizados pelos algoritmos de agrupamento na classificação de duas bases de itens. Porém, é necessário antes discorrer sobre os Testes Adaptativos Informatizados, outro ponto-chave do trabalho.

2.2 TESTES ADAPTATIVOS INFORMATIZADOS

De acordo com Tyler (2013), avaliação educacional pode ser definida como “o processo de determinar até qual ponto os objetivos educacionais estão sendo realizados”. Já Nevo (2006) defende que a avaliação educacional tem como objetivos “prover informação para auxílio na tomada de decisão” ou “estimar mérito ou valor”. Várias características do ambiente educacional são passíveis de avaliação: alunos podem ser avaliados para ter seu conhecimento aferido; professores, a fim de demonstrar sua didática ou melhor capacitá-los para seu ofício; instituições, para comprovar que aderem a determinados padrões educacionais, e materiais didáticos e metodologias de ensino, garantindo que estas estejam adequadas para uso em sala de aula (NEVO, 2006).

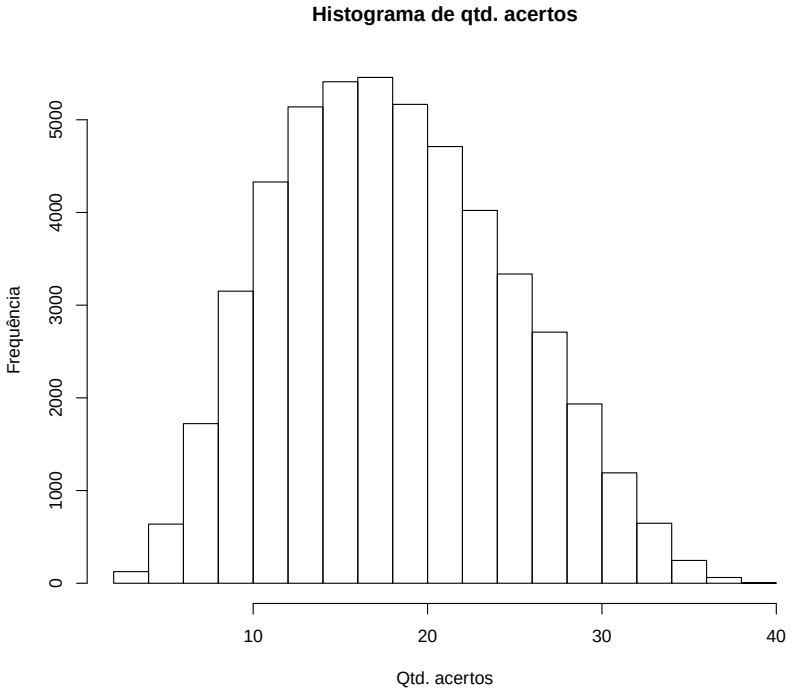
As avaliações educacionais podem ter caráter diagnóstico, quando, em um primeiro contato com um respondente, têm por objetivo aferir a situação atual deste com o objeto de avaliação; formativo, quando é aplicada durante o período letivo, verificando deficiências de aprendizado a tempo de corrigi-las; e somativo, quando, ao término do período letivo, coleta dados relacionados ao resultado do processo de aprendizagem, tendo fins tanto informativos como classificatórios (SCRIVEN, 1967).

Neste contexto, entende-se por objeto de avaliação aquilo que a avaliação deseja aferir. O objeto de uma avaliação de matemática, por exemplo, é o conhecimento dos avaliados no campo da matemática. Este grau de conhecimento, por outro lado, é chamado de proficiência ou, mais formalmente, proficiência.

Em um contexto avaliativo, é comum assumir que a proficiência de uma população de respondentes, dado um determinado objeto de avaliação, segue a distribuição normal: indivíduos de proficiências extremas, tanto altas quanto baixas, são raros, sendo que a proficiência da maior parte da população se concentra gradualmente na média (vide figura 2) . Por este motivo, é comum que, ao criar um instrumento de avaliação (um teste com uma série de questões, por exemplo), este seja composto por itens cujas dificuldades também estejam distribuídas de acordo com a normal. Essa abordagem permite que haja alguns itens de baixa/alta dificuldade capazes de discriminar indivíduos de baixa/alta proficiência entre si, assim como grande quantidade de itens para discriminar os indivíduos de proficiência mediana. Todos os itens são, então, apresentados aos respondentes, e suas respostas a estes comporão, ao final, o nível de sua proficiência no objeto de avaliação em questão.

Esta abordagem de realização de teste, aqui denominada de abordagem tradicional, possui uma desvantagem. Ela necessita que o teste contenha grande quantidade de itens para que

Figura 2 – Histograma de quantidade de acertos para o caderno de Linguagens e Códigos do Enem de 2012, contendo 45 itens e uma amostra de 50 mil respondentes



Fonte: Autor

este seja capaz de discriminar adequadamente os respondentes em todo o espectro de proficiência, tornando o teste extenso demais e obrigando cada indivíduo da população a responder itens de dificuldades muito discrepantes de suas proficiências.

Em resposta a esta e outras características negativas dos testes tradicionais, surgiram os TAI (LORD; NOVICK; BIRNBAUM, 1968; LORD, 1977; LORD, 1980; RASCH, 1966), testes aplicados por meio eletrônico, apoiados por um banco de itens calibrados através de algum método estatístico, onde os itens a serem respondidos pelo examinando são escolhidos em tempo real, de acordo com estimativas da proficiência feitas a cada nova resposta que o examinando devolve ao sistema. Esta abordagem possibilita a escolha de itens mais próximos da proficiência estimada do respondente e, consequentemente, a convergência mais rápida e precisa para seu nível de proficiência, utilizando menos itens que um teste convencional (LORD, 1977).

van der Linden e Glas (2000) citam Binet e Simon (1904) como a primeira aparição dos TAI na literatura. No teste de inteligência de Binet, era recomendado ao avaliador que estimasse a idade mental dos examinandos antes da aplicação do teste, para que o conjunto de itens específico para aquela faixa de idade mental fosse aplicado ao examinando.

Como exemplo de um teste adaptativo aplicado em larga escala, pode-se citar a *Armed Services Vocational Aptitude Battery* (ASVAB), uma bateria de exames aplicada anualmente para candidatos ao alistamento nas forças armadas norte-americanas (POMMERICH, 2009). A ASVAB teve sua pesquisa iniciada em 1970, com primeira aplicação em larga escala em 1985, sendo utilizada até hoje.

2.2.1 Algoritmo Iterativo

De forma geral, os TAI são regidos por um algoritmo iterativo, apresentado na figura 1. O algoritmo é dividido em quatro partes distintas, as quais diferenciam as diversas técnicas existentes para aplicação de testes adaptativos:

a) Uma estimativa inicial da proficiência do respondente;

b) Uma regra de seleção de itens do banco;

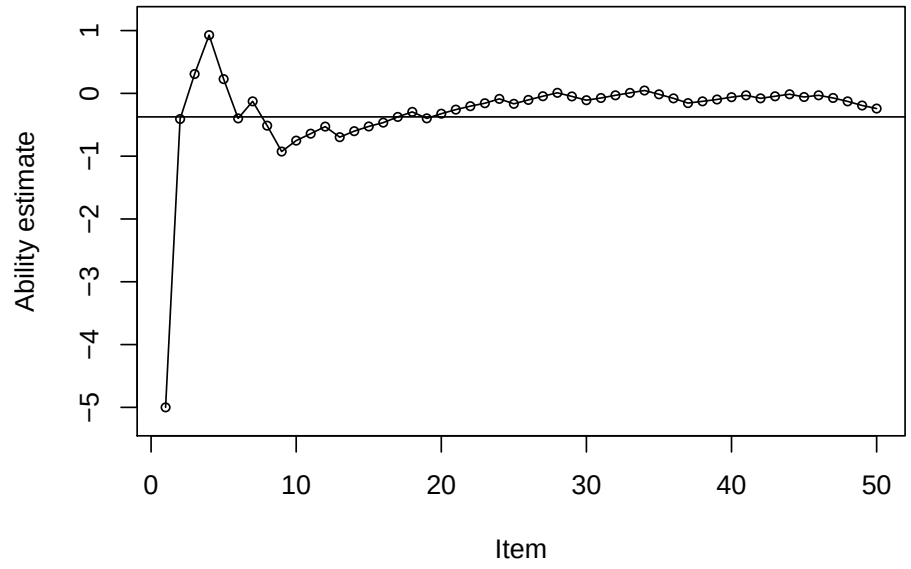
c) Uma forma de reestimar a proficiência do examinando após uma resposta (correta ou incorreta);

d) Um critério de parada.

A figura 3, gerada com o pacote catR (MAGIS; RAÎCHE, 2012), demonstra graficamente o processo de aplicação de um TAI e a convergência de $\hat{\theta}$ para θ. A linha sólida denota a proficiência real do indivíduo, com $\theta \approx -0{,}3$. Durante o teste, o respondente é apresentado a uma série de itens, representados no gráfico pelos pontos. Caso a dificuldade do item seja inferior à proficiência do respondente, ele possuirá alta probabilidade de respondê-lo corretamente. Caso contrário, ele possuirá probabilidade maior de errá-lo. A cada resposta, tanto correta quanto incorreta, $\hat{\theta}$ é re-estimado levando em consideração a última resposta do respondente e um novo item cuja dificuldade está mais próxima do novo valor de $\hat{\theta}$ é apresentado. O processo termina quando o examinando responde uma quantidade fixa de itens (neste caso, 50).

Ao iniciar o teste para um indivíduo de $\theta \approx -0{,}3$, o algoritmo inicia $\hat{\theta} = -5$ e apresenta o item da base que maximiza a informação para um indivíduo com tal nível de proficiência. Devido a $\theta > \hat{\theta}$, o examinando responde corretamente ao item apresentado; $\hat{\theta}$ é reestimado e um novo item é apresentado. É possível perceber que, quando o teste alcança aproximadamente 15 itens, as estimativas posteriores de $\hat{\theta}$ deixam de oscilar, indicando que o TAI encontrou um valor de $\hat{\theta} \approx \theta$. O teste poderia ser terminado com 15 itens sem prejudicar a estimativa final

Figura 3 – Convergência da proficiência estimada para o valor da proficiência real de um examinando



Fonte: Autor

de $\hat{\theta}$, não se estendendo até os 50 itens que foram originalmente apresentados ao indivíduo, representando uma redução de aproximadamente $2/3$ no número de itens presentes no teste.

Algoritmo 1 – Os passos de um algoritmo genérico para aplicação de um TAI

Entrada: Banco de itens X
Saída: Habilidade estimada $\hat{\theta}$
$\hat{\theta} =$ estimativa inicial
enquanto *critério de parada não alcançado* **faça**
    selecionar item ainda não administrado X_i
    aplicar X_i ao examinando
    reestimar $\hat{\theta}$ de acordo com resposta dada a X_i
fim
retorna $\hat{\theta}$

As próximas seções explicarão mais a fundo cada uma das partes que compõem um TAI. No entanto, pode-se perceber o vínculo que os TAI possuem com a TRI, uma vez que esta última fornece os valores necessários para que o teste funcione, uma escala de proficiências θ e de itens (simbolizada pelo parâmetro b de cada item) e a função de informação do item $I(\theta)$ para escolha do item mais informativo dado um $\hat{\theta}$.

2.2.1.1 Estimativa inicial da proficiência

Idealmente, quanto mais próximo de θ o valor inicial de $\hat{\theta}$, mais rápida a convergência do segundo para o primeiro (CONEJO et al., 2001). As técnicas para alcançar estes resultados diferenciam entre autores, sendo as mais comuns: posicionar $\hat{\theta}$ no centro da escala de proficiências (SUKAMOLSON, 2002); encontrar seu valor através da avaliação do currículo do aluno; posicionar $\hat{\theta}$ onde $I(\theta)$ do banco de itens é maior (DODD, 1990) ou inicializar $\hat{\theta}$ aleatoriamente.

2.2.1.2 Regras de seleção de itens

Após a estimativa inicial de $\hat{\theta}$ (e as posteriores re-estimativas), é necessário aplicar um item ao examinando para poder re-estimar $\hat{\theta}$, preferencialmente mais próximo de θ. Neste ponto, possui-se também a preocupação de que a regra de seleção escolha de forma relativamente homogênea os itens presentes no banco, por dois motivos principais. O primeiro é econômico: uma vez que a produção e calibração de um banco de itens demanda trabalho considerável, espera-se que todos os itens sejam usados no decorrer dos testes. Já o segundo é ligado à segurança. Se uma regra de seleção de itens aplica itens semelhantes à maioria dos examinandos, as respostas destes itens podem ser comprometidas, de modo que novos respondentes podem fazer o teste sabendo as respostas dos itens mais comuns, comprometendo assim a qualidade final da medida de proficiência. Este efeito possui o nome de viés de exposição e, no âmbito dos TAI, é uma das questões centrais relacionadas à segurança dos testes.

A regra mais comumente utilizada para selecionar itens é a de aplicar aquele que maximize $I(\theta)$ dado o $\hat{\theta}$ atual (LORD, 1977), consequentemente maximizando a informação do teste. Seu defeito mais notável, no entanto, é o de utilizar apenas os poucos itens que possuem altos picos na função de informação, ignorando aqueles que possuam funções de informação com distribuições mais uniformes dentre os diferentes valores de θ. A estratégia de escolha de itens por máxima informação é, portanto, a estratégia que estima $\hat{\theta}$ com máxima precisão, ao mesmo tempo em que (e justamente porque) ignora o viés de exposição dos itens, selecionando indiscriminadamente o item mais informativo a cada passo do teste.

É possível prever, portanto, que técnicas que forcem a escolha de itens sub-ótimos na tentativa de controlar a exposição dos itens prejudicam na estimativa final de $\hat{\theta}$ (DAVIS, 2004; DAVIS; DODD, 2008). Diversas técnicas, como a proposta por este trabalho, visam balancear essas características inversamente proporcionais.

Veerkamp e Berger (1997) criaram dois novos métodos de escolha de itens: função de informação por intervalo e função de informação ponderada por verossimilhança. No primeiro método, é escolhido o item que possua a maior área de informação entre dois valores pré-determinados; no segundo, o valor da função de informação é ponderado por uma função de verossimilhança. As duas abordagens visam refletir a incerteza da estimativa da proficiência do examinando nos resultados da função de informação, uma vez que o método de escolha da máxima função de informação desconsidera tal incerteza. Para "driblar" os defeitos da função de máxima verossimilhança (a qual não pode ser usada em caso de padrões de resposta completamente certos ou errados), foi usado o EAP ao re-estimar as proficiências dos examinandos. Os métodos foram validados em simulações, nas quais o método da função de informação por intervalo teve resultados inferiores à máxima função de informação e o método da função de informação ponderada por verossimilhança teve desempenho satisfatório.

Georgiadou, Triantafillou e Economides (2007) categorizam as regras em cinco tipos: regras aleatórias, de seleção condicional, estratificação, regras combinadas e testes adaptativos de múltiplos estágios. Para os fins deste trabalho, apenas as três primeiras categorias serão descritas, uma vez que as duas últimas são combinações das anteriores:

a) Métodos aleatórios: consistem na aplicação de itens de forma aleatória ao início do teste, aleatoriedade essa que diminui gradativamente conforme o teste progride e $\lim_{i \to N} \hat{\theta} = \theta$, sendo i o número de itens aplicados e N o total de itens no teste. Um exemplo destes é a estratégia 5-4-3-2-1 (MCBRIDE; MARTIN, 1983), a qual escolhe aleatoriamente um dos cinco itens mais informativos do banco para ser aplicado primeiro; o segundo item é escolhido dentre os quatro mais informativos e assim por diante. Revuelta e Ponsoda (1998) propuseram a estratégia progressiva, a qual adiciona um peso às funções de informação dos itens ao início do teste, dando menos valor à informação dos itens ao início do teste e aumentando o peso gradativamente, até que o valor original da função de informação seja utilizado na escolha dos itens.

b) Seleção condicional: nesta categoria, pode-se afirmar que a maioria das regras de seleção de itens sejam derivadas do método de Sympson-Hetter (SYMPSON; HETTER, 1985). O objetivo deste método está em manter a taxa de exposição dos itens abaixo de um limiar pré-determinado r^{max}. Para isso, os autores definem a probabilidade de um item i ser aplicado a um respondente como $P(A_i) = P(A_i|S_i)P(S_i)$, onde $P(A_i)$ é a probabilidade do item ser aplicado, $P(A_i|S_i)$ é a probabilidade do item ser aplicado tal que ele foi selecionado para aplicação e $P(S_i)$ é a probabilidade do item ser selecionado.

Uma vez que $P(S_i)$ depende da regra de seleção de itens, uma série de simulações pode ser realizada para descobrir o valor de $P(A_i|S_i)$ tal que $P(A_i) \leq r^{max}$.

c) Regras de estratificação: consistem na estratificação do banco de itens em camadas, de acordo com alguma regra ou conjunto de regras específicas, a fim de utilizar todos os itens de forma mais homogênea. Um exemplo é a regra de estratificação pelo parâmetro a (CHANG; YING, 1996), que consiste em ordenar o banco de itens em ordem crescente pelo parâmetro de discriminação, estratificando-o em K camadas, sendo K o número de itens a serem aplicados no teste. O primeiro item é escolhido da camada 1; o segundo, da camada 2, até o último, que é escolhido da camada k. Esta regra parte do princípio de que, ao início do teste, a estimativa $\hat{\theta}$ é muito diferente de θ, não justificando o uso de itens que possuam alto valor informativo para $\hat{\theta}$. A regra, então, dá preferência a funções que possuam distribuições mais uniformes de informação ao início do teste, escolhendo itens com maiores picos conforme o teste progride e $\hat{\theta}$ converge para θ. Posteriormente, após descoberta uma correlação entre as variáveis de discriminação e dificuldade, foi proposta a estratificação pelos parâmetros a e b (CHANG; QIAN; YING, 2001), garantindo que todas as camadas com valores de a crescentes possuam itens em todos os níveis de dificuldade.

O objetivo deste trabalho foca nesta etapa de aplicação dos TAI: a apresentação de um método de seleção de itens baseada no agrupamento dos itens pela similaridade entre seus parâmetros.

2.2.1.3 Re-estimativa da proficiência

Re-estimar a proficiência de um examinando consiste em encontrar um novo valor para a proficiência estimada ($\hat{\theta}$) deste examinando que se aproxime do valor de sua proficiência real (θ), dadas suas respostas até o momento. Este processo pode ser dividido em dois momentos: re-estimar $\hat{\theta}$ enquanto o padrão de respostas do examinando não muda (apenas acertos ou apenas erros) e re-estimar $\hat{\theta}$ quando este padrão se modifica.

Vários métodos já foram testados para re-estimar $\hat{\theta}$ enquanto o padrão de respostas do examinando não muda. No método de Dodd (1990, Cf. BARRADA; OLEA; PONSODA; BARRADA; OLEA; ABAD; BARRADA et al.; BARRADA et al., 2007, 2008, 2009, 2009), o valor de $\hat{\theta}$ é atualizado de acordo com a seguinte fórmula (11), onde b_{max} significa o valor máximo de b no banco de itens e x_t o acerto ao item x aplicado no instante t.

$$\hat{\theta}_{t+1} = \left\{ \begin{array}{ll} \hat{\theta}_t + \frac{b_{max} - \hat{\theta}_t}{2} & \text{se } x_t \\ \hat{\theta}_t - \frac{\hat{\theta}_t - b_{min}}{2} & \text{se } \neg x_t \end{array} \right\} \tag{11}$$

É possível, adicionalmente, utilizar os métodos descritos na seção 2.1.2 para realizar essa re-estimativa, sendo que aqueles que se utilizam da máxima verossimilhança só podem ser utilizados após mudança no padrão de respostas do examinando, enquanto os Bayesianos funcionam em todos os casos.

2.2.1.4 Critérios de parada

Após a aplicação de cada item, o algoritmo do TAI utiliza um critério de parada para verificar se o teste deve ser encerrado, ao invés de aplicar um novo item. O critério de parada mais comumente usado é o de restringir o teste a um número máximo de itens (BARRADA; MAZUELA; OLEA, 2006; BARRADA; OLEA; ABAD, 2008). Uma proporção de 1:12 entre o tamanho do teste e do banco de itens é defendido por alguns autores (CHANG; VAN DER LINDEN, 2003; CHAJEWSKI; LEWIS, 2009). Dodd (1990) argumenta que um critério de parada aceitável para um TAI seria o momento em que o banco de itens se esgotasse de itens com um valor de informação relevante, ou seja, no instante em que nenhum item i disponível no banco possua $I_i(\theta) \geq I_{min}$, sendo I_{min} um limiar pré-definido.

Outro critério comumente utilizado consiste em terminar o teste assim que o erro padrão da proficiência do examinando diminuir além de um certo limiar pré-determinado, garantindo também que a pontuação dos examinandos seja equiprecisa. De acordo com Moreira Junior (2012, p. 131), "essa precisão normalmente está relacionada com a informação do teste, no caso da estimação por MV, ou com a variância da posteriori, no caso da estimação Bayesiana".

2.2.2 Medidas de desempenho

Para se aferir o desempenho do algoritmo de um TAI, algumas medidas de desempenho podem ser utilizadas. A raiz dos erros quadráticos médios permite medir o quanto o algoritmo do TAI consegue aproximar $\hat{\theta}$ de θ,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{N}}, \tag{12}$$

onde N representa o número de examinandos (BARRADA et al., 2009; BARRADA; ABAD; VELDKAMP, 2009; BARRADA et al., 2009). A medida é comumente utilizada por diversos autores (CHAJEWSKI; LEWIS, 2009; DAVIS; DODD, 2008; BARRADA et al., 2010; BARRADA; ABAD; OLEA, 2014) para comparar a precisão das diferentes metodologias de aplicação de TAI.

Já a taxa de exposição de um item, comumente denotada por r_i, indica a quantidade de testes nos quais o item i figurou, dividida pela quantidade de testes aplicados no total. É o alvo principal dos métodos de seleção condicional supracitados, os quais tentam manter $r_i \leq r^{max}$, sendo r^{max} um limiar superior pré-determinado (SYMPSON; HETTER, 1985; VAN DER LINDEN, 2003).

Por último, a taxa de sobreposição de itens indica a proporção de itens que dois ou mais examinandos compartilham entre si em um teste e é dada por

$$T = \frac{N}{Q} S_r^2 + \frac{Q}{N}, \tag{13}$$

onde N indica a quantidade de itens no banco de itens, Q indica o número de itens no teste e S_r^2, a variância das taxas de exposição de todos os itens (BARRADA et al., 2010). É possível notar que, sendo $\frac{Q}{N}$ um valor constante, T alcança seu menor valor quando S_r^2 tende a 0. Para isso, é necessário que todos os itens do banco sejam utilizados de maneira uniforme, homogeneizando os valores de suas taxas de exposição e minimizando a variância deste parâmetro.

Ambas as equações (12) e (13) serão utilizadas neste trabalho para o avaliar o desempenho do método proposto para aplicação de um TAI.

2.3 AGRUPAMENTO POR SIMILARIDADE

Denomina-se agrupamento o conjunto de técnicas e algoritmos utilizados no agrupamento de elementos de forma não-supervisionada. Como resultado, obtêm-se grupos de elementos, de forma que elementos situados em um grupo sejam semelhantes entre si e diferentes de elementos de grupos distintos. Apesar desta definição, muitos algoritmos de agrupamento não são determinísticos, ou seja, diferentes aplicações de um mesmo algoritmo na base de dados podem resultar em grupos diferentes. Da mesma forma, diferentes algoritmos de agrupamento aplicados na mesma base de dados também podem resultar em grupos diferentes, resultando em controvérsia quanto à definição do termo grupo (IZENMAN, 2008).

Os primeiros relatos referentes ao agrupamento de elementos por suas semelhanças tiveram origem na antropologia e psicologia, sendo suas primeiras utilizações em testes psicológicos (ZUBIN, 1938; CATTELL, 1943). Atualmente, as técnicas de agrupamento são estudadas dentro da grande área de *data mining* (WITTEN; FRANK; HALL, 2011; MIRKIN, 2012), porém, devido à generalidade de seus algoritmos, o agrupamento é utilizado em diversas áreas do conhecimento humano, como na segmentação de imagens (SHI; MALIK, 2000; JAIN, 2010), agrupamento de cidades, imagens de faces, micro-arranjos de DNA (FREY; DUECK, 2007), materiais bibliográficos (JAIN; MURTY; FLYNN, 1999), perfis de usuários no projeto de softwares (MASIERO, 2013), criação de personas para projetos de interface de software (AQUINO JUNIOR, 2008).

Jain, Murty e Flynn (1999) defendem que o processo de agrupamento é dividido nas seguintes fases:

a) Representação de dados, que engloba seleção e extração de características

b) Definição de medida de similaridade

c) Agrupamento

d) Abstração e interpretação dos dados

No processo de representação de dados, as melhores características dentre aquelas disponíveis são escolhidas para compor a base de dados sobre a qual o agrupamento será executado (seleção de características). Opcionalmente, novas características podem ser extraídas da base de dados original através da execução de transformações nestes dados, a fim de se conseguir novas informações que otimizem o processo de agrupamento (extração de características) (JAIN; MURTY; FLYNN, 1999; MIRKIN, 2012).

Para que elementos possam ser comparados entre si, é necessário definir uma medida de similaridade. A medida de similaridade pode ser uma medida de distância entre os elementos, como também pode ser uma função específica do domínio da aplicação, a qual utiliza as características selecionadas/extraídas no passo anterior para representar numericamente a similaridade entre dois elementos da base de dados (JAIN; MURTY; FLYNN, 1999; MIRKIN, 2012).

Medidas de distância comumente usadas como medidas de dissimilaridade são a distância de Minkowski (LEIGH, 2007), apresentada na equação (14), e suas derivadas. Dado um parâmetro p, a distância de Minkowski entre dois elementos x e y com n características cada é dada por

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_n - y_n|^p \right)^{\frac{1}{p}}. \tag{14}$$

Quando $p = 1$ a distância de Minkowski se comporta como a distância de Manhattan (LEIGH, 2007):

$$d_1(x, y) = \sum_{i=1}^{n} |x_n - y_n| \tag{15}$$

Para $p = 2$ ela equivale à distância Euclideana:

$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_n - y_n)^2} \tag{16}$$

Quando $\lim_{p\to\infty}$, ela corresponde à distância de Chebyshev.

$$\lim_{p\to\infty} d_p(x, y) = \max_i |x_i - y_i| \tag{17}$$

Para $p > 1$, as distâncias são métricas (IZENMAN, 2008), ou seja,

$d(x, y) \geq 0$, positividade;
$d(x, x) = 0$, identidade;
$d(x, y) = d(y, x)$, simetria;
$d(x, z) \leq d(x, y) + d(y, z)$, desigualdade dos triângulos.

Uma outra distância que deve ser mencionada é a distância de Mahalanobis (MAHALANOBIS, 1936), a qual leva em consideração a covariância entre os dados no cálculo da similaridade. Sejam x e y dois vetores de características e S a matriz de covariância entre x e y. A distância de Mahalanobis entre x e y é dada por

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}. \tag{18}$$

A fase de agrupamento em si consiste na utilização de um algoritmo de agrupamento, o qual utilizará a medida de similaridade em questão para organizar os elementos em grupos, de forma que elementos pertencentes ao mesmo grupo sejam mais similares entre si do que elementos que estejam grupos distintos.

Por fim, os grupos de dados podem ser então usados para se compreender melhor o fenômeno em questão. Elementos no mesmo grupo podem, por exemplo, ter suas características generalizadas por apenas um elemento do grupo, denominado representante (JAIN; MURTY; FLYNN, 1999; FREY; DUECK, 2007; MIRKIN, 2012), para que as similaridades entre si possam ser melhor compreendidas. No presente trabalho, são priorizados os algoritmos de agrupamento que possuem como parâmetro de entrada o número de grupos, representados por

k. Tais algoritmos são apresentados nas seções seguintes: algoritmos de particionamento por aproximação de centroides e algoritmos hierárquicos. Também são apresentados algoritmos de agrupamento por densidade e baseados em grafos, inclusos na pesquisa preliminar porém não utilizados nos experimentos.

2.3.1 Agrupamento de particionamento por aproximação de centroides

Criado em 1967 por James MacQueen (MACQUEEN, 1967), o *k-means* é um dos algoritmos de agrupamento mais conhecidos. O *k-means* foi criado com o propósito de encontrar soluções para o problema da soma dos erros quadráticos: dado um número de grupos k e um conjunto de dados multivariados X no espaço $\boldsymbol{R}^n$, onde cada x é representado pela tupla de valores $\{x_1, x_2, \ldots, x_n\}$, agrupar X em k grupos de forma a minimizar a soma dos erros quadráticos de cada grupo. A contribuição de um dado x_i para o erro de seu grupo j é representada pela distância de x_i com relação à média μ do grupo j ao qual x_i foi vinculado.

Usando a terminologia supracitada, a soma dos erros quadráticos de um agrupamento pode ser representada por

$$W = \sum_{j=1}^{K} \sum_{i=1}^{n} a_{ij} |x_i - \mu_j|^2, \tag{19}$$

onde $a_{ij} = \left\{ \begin{array}{ll} 1 & \text{se } x_i \in j \\ 0 & \text{se } x_i \notin j \end{array} \right\}$, $\mu_j = \sum_{x_i \in j} (x_i / n_j)$ e $n_j = \sum_{x_i \in j} a_{ij}$.

O *k-means* é um caso específico do algoritmo de Lloyd, cujo objetivo é realizar o particionamento do espaço Euclideano, gerando o diagrama de Voronoi de tal espaço, composto de partições representadas por formas geométricas convexas. No *k-means*, os elementos a serem agrupados são tratados como uma discretização do espaço geométrico sobre o qual o algoritmo de Lloyd é aplicado. Como consequência dessa semelhança, os grupos gerados pelo *k-means* têm formatos elipsoidais (ou hiper-elipsoidais, em casos de muitas dimensões) (LLOYD, 1982).

De acordo com Jain (2010), a minimização da soma dos erros quadráticos sobre todos os grupos é um problema NP-difícil. O objetivo do *k-means* é minimizar o resultado da função-objetivo (19) em um processo iterativo de escolha dos valores de a_{ij} e reestimativa de μ_j, dado o número desejado de grupos k. O algoritmo é inicializado posicionando aleatoriamente os centroides $\mu_{1\ldots k}$ desses grupos no espaço de dados, iniciando um processo iterativo que consiste nos seguintes passos:

a) associar cada dado x ao grupo j cujo centroide μ_j esteja mais próximo de x;

b) atualizar $\mu_{1...k}$ de acordo com os novos dados associados a cada μ.

O processo termina quando nenhum elemento for designado a um centroide diferente em uma dada iteração.

Algoritmo 2 – *K-means*

Entrada: Matriz de dados X, número de grupos k
Inicializar $\mu_{1...k}$ nas coordenadas de um dado x aleatório
enquanto a_{ij} *e* μ_j *não estabilizarem* **faça**
    para cada $x_i \in X$ **faça**
        $a_{ij} = \left\{ \begin{array}{ll} 1 & \text{se } j = \arg\min_l |x_i - \mu_l| \\ 0 & \text{senão} \end{array} \right\}$
    fim
    para $j = 1, 2 \ldots k$ **faça**
        $\mu_j = \sum_{x_i \in j} (x_i / n_j)$
    fim
fim

Sob esta definição, o *k-means* possui uma complexidade assintótica de $O(i \cdot k \cdot n \cdot d)$, onde i indica o número de iterações até a convergência; k, o número de grupos; n, o número de elementos; e d o número de dimensões, ou características dos dados.

Duas características do *k-means* devem ser ressaltadas. A primeira tem relação a sua inicialização: dependendo dos valores iniciais escolhidos para $\mu_{1...k}$, o algoritmo pode convergir para mínimos locais. A segunda característica importante é a capacidade do *k-means* de gerar apenas grupos elipsoidais, um viés causado pelo uso da distância dos elementos aos centroides para designar o elemento ao grupo (JAIN, 2010).

2.3.2 Agrupamento hierárquico

No agrupamento hierárquico, os elementos são organizados em uma estrutura hierárquica, denominada dendrograma, utilizando-se a medida de similaridade entre os dados (JAIN; MURTY; FLYNN, 1999; MIRKIN, 2012). O agrupamento hierárquico pode ser abordado de forma aglomerativa ou divisiva. No agrupamento aglomerativo, todos os elementos fazem partes de grupos com apenas um item. No próximo passo, grupos que possuem a menor distância entre si são unidos. Este novo grupo pode ser então utilizado para futuras aglomerações, gerando a estrutura binária supracitada. No agrupamento divisivo ocorre o processo inverso: todos os elementos fazem parte de um grande grupo, o qual é dividido de forma a minimizar a similaridade entre os dois novos grupos gerados.

Em ambos os casos, com o dendrograma formado, a geração de grupos é realizada através de um corte horizontal na árvore: dependendo da altura em que o corte é feito, mais ou menos grupos são gerados ao final, dependendo da situação.

A principal diferença entre os diferentes algoritmos hierárquicos está na medida de similaridade utilizada por cada um.

Nos exemplos a seguir, sejam g_i e g_j dois grupos, $g_i \neq g_j$. Sejam também g_k e g_l dois grupos, onde $g_i = g_k \cup g_l$.

No agrupamento por ligação simples (JAIN; MURTY; FLYNN, 1999), a distância entre g_i e g_j é a distância entre os dois elementos mais próximos de cada grupo,

$$d(g_i,g_j) = \min_{x_i \in g_i; x_j \in g_j} d(x_i, x_j).$$

No agrupamento por ligação completa (JAIN; MURTY; FLYNN, 1999), utiliza-se a distância entre os dois elementos mais distantes de g_i e g_j:

$$d(g_i,g_j) = \max_{x_i \in g_i; x_j \in g_j} d(x_i, x_j).$$

No agrupamento por ligação média (*Unweighted Pair-Group Method using Arithmetic Averages*, UPGMA) (MOUCHET; GUILHAUMON; MASON, 2008), é calculada a média entre as distâncias de todos os elementos de ambos os grupos:

$$d(g_i,g_j) = \frac{\sum_{x_i \in g_i} \sum_{x_j \in g_j} d(x_i,x_j)}{n_i + n_j}.$$

No agrupamento por ligação média ponderada (*Weighted Pair-Group Method using Arithmetic Averages*, WPGMA) (MOUCHET; GUILHAUMON; MASON, 2008), é calculada a média ponderada entre as distâncias de todos os elementos de ambos os grupos:

$$d(g_i,g_j) = \frac{d(g_j,g_k) + d(g_j \in g_l)}{2}.$$

Na ligação pelo centroide (*Unweighted Pair-Group Method using Centroids*, UPGMC) (MOUCHET; GUILHAUMON; MASON, 2008), é utilizada a distância entre os centroides de ambos os grupos,

$$d(g_i,g_j) = ||\mu_{g_i} - \mu_{g_j}||,$$

onde μ_i é a média aritmética dos dados do grupo g_i.

Na ligação pela mediana (*Weighted Pair-Group Method using Centroids*, WPGMC) (MOUCHET; GUILHAUMON; MASON, 2008), também é utilizada a distância entre os cen-

troides de ambos os grupos, porém μ_i é definido recursivamente como a média ponderada entre os centroides dos grupos g_r e g_s que compõem o grupo g_i,

$$g_i = \frac{(g_r + g_s)}{2}.$$

Outra medida utilizada no agrupamento hierárquico é a função de Ward (MIRKIN, 2012), a qual visa minimizar a variância entre os dois grupos sendo unidos,

$$d(g_i,g_j) = W(C(g_i \cup g_j)) - W(C(g_i,g_j)),$$

onde $W(C(g_i,g_j))$ representa a soma das variâncias intra-grupos de um agrupamento C no qual os grupos g_i e g_j não foram unidos e $W(C(g_i \cup g_j))$, a soma das variâncias intra-grupos de um agrupamento C após a aglomeração dos grupos g_i e g_j.

O algoritmo 3 descreve o funcionamento do método aglomerativo de agrupamento, o qual recebe como entrada a matriz de dados X com N elementos e dimensão d. Durante a inicialização do método, N grupos são criados, sendo que cada grupo contém um elemento de X. O vetor G recebe, em sua primeira iteração, os N grupos unitários.

Em cada iteração i do algoritmo, as distâncias entre os grupos gerados até a iteração anterior são calculadas. Este processo leva tempo $O(n^2d)$. A seguir, o algoritmo seleciona os dois grupos m e n que possuam distância mínima entre si, aglomera-os em um único grupo, excluindo m e n da lista de grupos na iteração i e adicionando o grupo recém formado $m \cup n$. O processo se repete por N iterações até que todos os dados sejam aglomerados em um único grupo. Considerando-se que as distâncias entre os grupos devem ser calculadas nas n iterações, o processo completo do algoritmo possui complexidade $O(n^3)$.

Algoritmo 3 – Agrupamento hierárquico aglomerativo

Entrada: Matriz de dados X com N elementos
$G[0] = X$
para $i = 1$ **até** $N - 1$ **faça**
    para cada $m \in G[i-1]$ **faça**
        para cada $n \in G[i-1]$ **faça**
            $D[m,n] = d(m,n)$
        fim
    fim
    $(m,n) = \arg\min_{m,n} D[m,n]$
    $G[i] = G[i-1] - \{m,n\} + (m \cup n)$
fim
retorna G

Tabela 1 – Valores dos parâmetros da fórmula recursiva de Lance-Williams

Critério de agrupamento	α_m	β	γ
Ligação simples	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação completa	$\frac{1}{2}$	0	$\frac{1}{2}$
Média	$\frac{n_m}{n_m+n_n}$	0	0
Média ponderada	$\frac{1}{2}$	0	0
Centroide	$\frac{n_m}{n_m+n_n}$	$\frac{-n_m n_n}{(n_m+n_n)^2}$	0
Mediana	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_m+n_g}{n_m+n_n+n_g}$	$\frac{-n_g}{n_m+n_n+n_g}$	0

Fonte: Autor “adaptada de” Everitt (2011)

Todas as medidas de distância supracitadas podem ser implementadas no algoritmo aglomerativo através da fórmula recursiva de Lance-Williams (EVERITT, 2011). Dados dois grupos m e n que foram aglomerados na iteração i, as distâncias entre qualquer grupo g ($g \neq m$ e $g \neq n$) podem ser atualizadas com relação ao novo grupo $m \cup n$ utilizando a seguinte fórmula,

$$D[g,m \cup n] = \alpha_m D[g,m] + \alpha_n D[g,n] + \beta D[m,n] + \gamma |D[g,m] - D[g,n]|,$$

onde os valores dos parâmetros α, β e γ são dados na tabela 1. Nela, n_m, n_n e n_g indicam o número de elementos nos grupos m, n e g, respectivamente.

2.3.3 Agrupamento baseado em densidades

O agrupamento baseado em densidades teve origem da definição de que grupos são espaços de dados de alta densidade separados por espaços de baixa densidade. Desta forma, algoritmos foram criados para realizar o agrupamento dos dados nestas zonas de alta densidade. Como exemplo desta classe de algoritmos pode-se citar o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) (ESTER et al., 1996). No DBSCAN, caso um elemento possua $MinPts$ vizinhos a uma distância ϵ dele mesmo, ele é considerado um elemento de núcleo, localizado em um espaço de alta densidade e um grupo é formado incluindo ele e seus vizinhos. Caso estes elementos vizinhos também forem considerados elementos de núcleo, o processo se repete e o grupo aumenta de tamanho. Esta abordagem de geração de grupos permite ao DBSCAN gerar grupos de formatos arbitrários, ao contrário dos grupos elipsoidais formados por algoritmos como o *k-means*.

2.3.4 Agrupamento espectral

Shi e Malik (2000) apresentam o conceito de corte normalizado aplicado ao agrupamento de regiões de imagens. Para ser realizado, é necessário que os elementos sejam representados como um grafo totalmente conexo, no qual as arestas indicam os graus de similaridade entre os elementos. O corte mínimo é aquele que remove as arestas do grafo de forma que a soma dos valores das arestas removidas seja o menor possível, indicando que a massa de dados foi separada em dois grupos na área onde a similaridade que unia os elementos é a mínima possível. Devido ao viés que o corte mínimo possui em formar grupos pequenos, foi criado o corte normalizado, que consiste no valor do corte mínimo dividido pela soma de todas as similaridades do grafo.

Devido à escolha da posição ótima do corte normalizado ser um problema NP-difícil, (SHI; MALIK, 2000), diversos autores utilizam uma técnica, denominada agrupamento espectral, que realiza uma diminuição no número de dimensões da matriz de similaridades, seguida pelo agrupamento da matriz resultante utilizando um algoritmo mais trivial, como *k-means* (MEILA; SHI, 2001; NG; JORDAN; WEISS, 2001; VON LUXBURG, 2007).

2.3.5 Validação de grupos

Por padrão, a maioria das técnicas de agrupamento se comportam de maneira não-supervisionada, o que significa que os algoritmos não recebem nenhuma informação *a priori* relacionada aos dados que devem classificar. No caso do agrupamento por similaridade, isso se reflete no fato de os dados não possuírem qualquer classificação prévia (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002). Por esse motivo, é necessário avaliar os resultados dos algoritmos através do uso de medidas de avaliação, que utilizam características como compactação dos grupos e distância entre grupos para representar numericamente a qualidade do processo de agrupamento.

Diversos autores defendem a existência de três paradigmas para se avaliar os resultados de um processo de agrupamento (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002; KOVÁCS; LEGÁNY; BABOS, 2005): avaliação através do uso de critérios externos, internos ou relativos. Na avaliação por critérios externos, ou avaliação externa, são utilizadas na validação dos grupos características que não foram utilizadas na criação dos mesmos, como as classificações originais dos elementos (MIRKIN, 2012). Essas características dependem de fatores como a base de dados sendo avaliada e o conhecimento do domínio. Na avaliação por critérios

internos, as características utilizadas para se avaliar o processo de agrupamento são extraídas diretamente da base de dados e dos resultados do agrupamento, como por exemplo, a matriz de similaridades. Por último, a avaliação por critérios relativos utiliza os resultados de diversos processos de agrupamento realizados com parâmetros de entrada diferentes, na tentativa de eleger aquele que possui os melhores resultados.

2.4 TRABALHOS RELACIONADOS

Nesta seção são apresentadas os trabalhos relacionados com este. Primeiramente, são apresentados trabalhos que utilizam abordagens semelhantes na aplicação de Testes Adaptativos Informatizados. Em seguida, é disponibilizada uma revisão de trabalhos que utilizam técnicas de inteligência artificial na construção de TAI. A seção foi dividida entre as principais técnicas encontradas na literatura que trouxeram contribuições significativas para o tema.

Phuvipadawat et al. (2008) criaram um modelo de TAI separando os itens do teste em 5 níveis. O examinando tem sucesso no teste se conseguir responder itens até chegar ao nível pré-determinado, o qual corresponde ao seu escore. Durante o teste, cada nível é representado pela mesma quantidade n de itens. Para subir um nível, o examinando deve responder uma quantidade predeterminada dos n itens do nível anterior (por exemplo, 3 itens para $n = 4$). O teste termina quando, após ter avançado pelo menos um nível, o respondente falha em passar para o nível seguinte ou, de forma inversa, quando o respondente, após falhar em um ou mais níveis, finalmente acerta uma quantidade satisfatória dos n itens em determinado nível. O último nível em que o usuário teve sucesso identifica sua proficiência.

O algoritmo de similaridade de itens utiliza uma matriz A que combina item/palavra-chave, onde $A_{ij} = 1$ indica que o item i é associado à palavra-chave j (caso contrário, $A_{ij} = 0$). A similaridade do cosseno[1] entre os vetores de dois itens na matriz A indica a similaridade entre eles. Isso impede que examinandos tenham de responder vários itens de temas semelhantes e também diminui o viés de exposição. Eles alegaram ter uma precisão de estimativa de proficiência dos examinandos de 93%. É importante notar que os testes foram simulados usando o modelo de 3 parâmetros da TRI, utilizando proficiências aleatórias e a função de probabilidade de acerto da TRI (PHUVIPADAWAT et al., 2008), pontos que este trabalho também visa utilizar.

[1] A similaridade do cosseno é uma medida de distância entre dois vetores dada por $cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$, onde $\boldsymbol{x}\cdot\boldsymbol{y}$ é o produto escalar $\sum_{n=1}^{N} x_n y_n$ e $\|\boldsymbol{x}\|$ é o tamanho do vetor $\boldsymbol{x}$, $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}\cdot\boldsymbol{x}}$ (TAN; STEINBACH; KUMAR, 2005). A medida varia entre $[-1; 1]$,

Utilizando um banco de itens cujas proficiências e hierarquia entre elas foram indicadas por especialistas na área, Robles Pedrozo e Rodriguez-Artacho (2013) utilizaram o pacote *Latent Trait Models*, da linguagem R, para estimar os parâmetros de alunos e itens em diversos modelos diferentes da TRI, selecionando o mais adequado através da ANOVA. Duas matrizes foram então utilizadas: a primeira identifica quais proficiências são inter-relacionadas, formando uma hierarquia entre elas. Na segunda, um valor de 0 a 5 é associado para cada par Item/Habilidade, sendo que 0 indica que o item não afere aquela proficiência e 5, que ele afere de maneira muito acentuada.

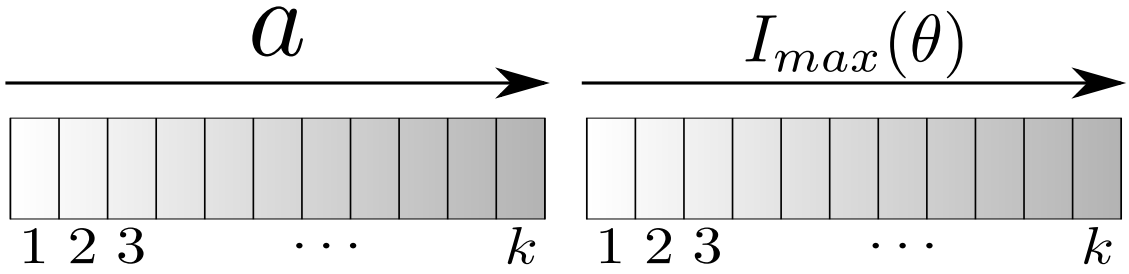
Assumindo-se um limiar que denota se o respondente dominou uma proficiência (ex.: 0,6), é possível calcular, através das duas matrizes e das respostas dos respondentes, quais proficiências este dominou. Estas informações são então utilizadas para realizar o agrupamento dos alunos de acordo com a soma dos pesos das proficiências que o respondente não dominou (ROBLES PEDROZO; RODRIGUEZ-ARTACHO, 2013).

Em Chang e Yang (2009) e Chang et al. (2009), os autores calcularam o valor da proficiência de alunos em um teste online de acordo com os modelos logísticos de 1, 2 e 3 parâmetros da TRI, posteriormente realizando o agrupamento dos alunos de acordo com o valor de suas proficiências utilizando o algoritmo *k-means*. Eles observaram que os centroides dos grupos gerados pelo modelo de 1 parâmetro são mais próximos que nos modelos de 2 e 3 parâmetros, indicando que os dois posteriores oferecem resultados de melhor qualidade durante o processo de agrupamento.

Nos trabalhos supracitados, foram realizados o agrupamento de itens com informações disponíveis a priori (PHUVIPADAWAT et al., 2008) e o agrupamento de indivíduos de acordo com os valores de suas proficiências (CHANG et al., 2009; CHANG; YANG, 2009; ROBLES PEDROZO; RODRIGUEZ-ARTACHO, 2013). Apesar de o presente trabalho propor o agrupamento de itens e não de indivíduos, os trabalhos de Chang e Yang e Chang et al. demonstram melhores resultados dos modelos logísticos da TRI que apresentam maiores números de parâmetros, característica incorporada neste trabalho.

Existem trabalhos que não utilizam a lgoritmos de agrupamento para agregar itens, porém segregam os itens em camadas em um processo denominado “estratificação” do banco de itens. Chang e Ying (1996) apresenta o processo de estratificação-α. Neste procedimento, os itens da base são ordenados em ordem crescente por seu parâmetro a e, após isso, o banco de itens é dividido em k camadas, sendo que o valor de k coincide com o número de itens do teste. O primeiro item no teste é então selecionado da camada 1, a camada onde se encontram os itens de menor discriminação; o segundo item a ser aplicado no teste é selecionado da camada 2 e

Figura 4 – Estratificação-α (CHANG; YING, 1996) e MIS (BARRADA; MAZUELA; OLEA, 2006).



(a) Estratificação-α (b) MIS

Fonte: Autor

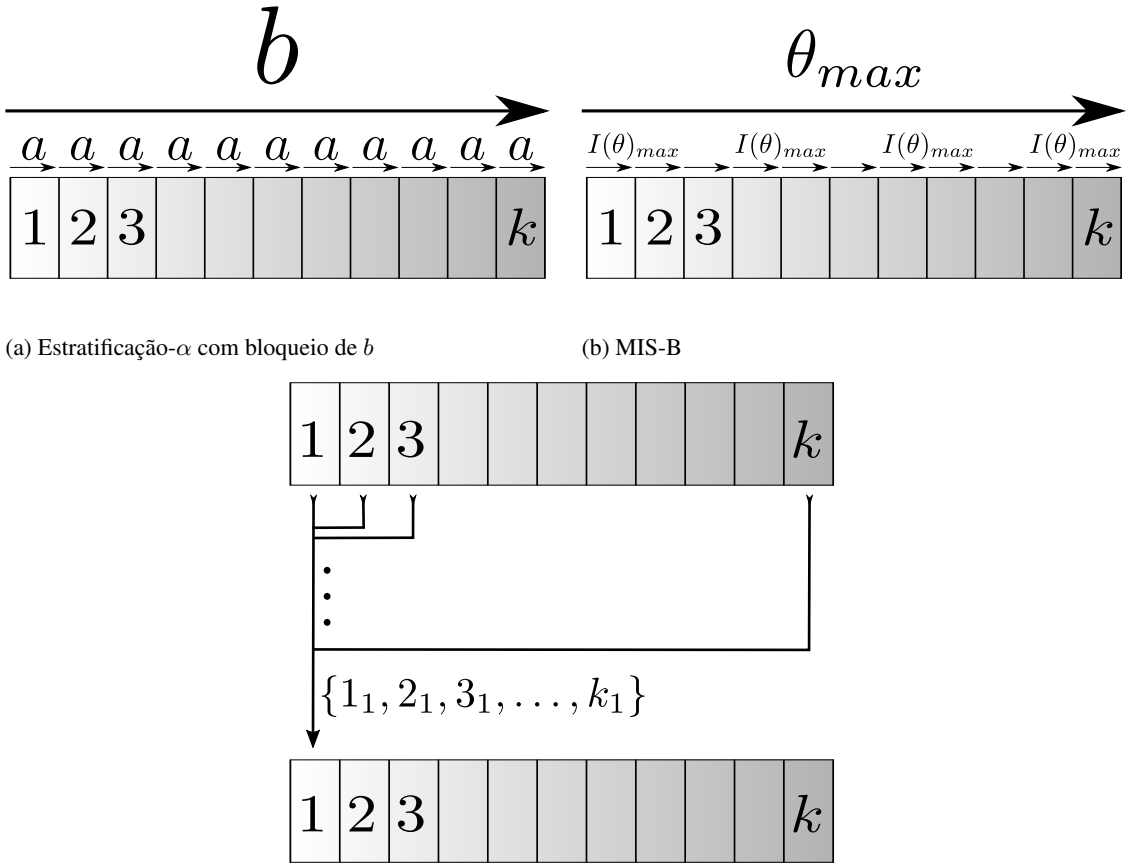
o processo se repete, até que o item k é selecionado da camada k. Esta estratégia garante que itens de baixo valor de informação sejam utilizados nos testes adaptativos, concentrando-os no início do teste, onde a incerteza com relação à estimativa de $\hat{\theta}$ é maior e itens de alto valor de informação podem não ser tão úteis.

Em Chang, Qian e Ying (2001), é descrita a técnica de estratificação-α com bloqueio do parâmetro b. Nela, o banco de itens é ordenado de maneira crescente pelo parâmetro b e estratificada em k camadas. Cada camada é então ordenada de forma crescente de acordo com o parâmetro a. Um novo conjunto de camadas é então formado, onde os itens da primeira camada do novo conjunto é formado pelos primeiros itens de todas as k camadas originais. O processo se repete, de forma que cada camada formada possua itens de discriminação média crescente, porém com uma distribuição aproximadamente uniforme de dificuldades.

Apesar de realizar o particionamento do banco de itens, os métodos de Chang e Ying e Chang, Qian e Ying diferem do presente trabalho pela abordagem utilizada na realização do particionamento. Enquanto os trabalhos citados particionam o banco de itens se utilizando do conhecimento do domínio, i.e. o conhecimento do impacto da discriminação dos itens nas estimativas de $\hat{\theta}$ em diferentes pontos do teste, o método proposto se utiliza de algoritmos de agrupamento genéricos que não se utilizam de conhecimento do domínio na fase de agrupamento dos itens.

Barrada, Mazuela e Olea (2006) apresentam um novo método de estratificação, similar aos dois métodos supracitados, que leva o parâmetro c em consideração na separação da base. Na abordagem adotada, a estratificação pelo parâmetro a (CHANG; YING, 1996) é trocada pela estratificação realizada por $I(\theta)_{max}$), o máximo valor de informação alcançado por cada item e o bloqueio pelo parâmetro b (CHANG; QIAN; YING, 2001) é substituído por θ_{max}, o

Figura 5 – Estratificação-α com bloqueio de b (CHANG; QIAN; YING, 2001) e MIS-B (BARRADA; MAZUELA; OLEA, 2006).



(a) Estratificação-α com bloqueio de b

(b) MIS-B

(c) Organização final dos itens

Fonte: Autor

valor de θ onde o cada alcança $I(\theta)_{max}$. Os dois métodos são denominados *Maximum Information Stratification* (MIS) (Estratificação por Máxima Informação) e *Maximum Information Stratification with Blocking* (MIS-B) (Estratificação por Máxima Informação com Bloqueio).

Já em Barrada, Abad e Olea (2014), os métodos MIS e MIS-B são estudados para que o número de camadas ótimo seja encontrado. Foi concluído que o número de camadas ótimo é igual ao número de itens aplicados no teste. Neste trabalho, o número de grupos no qual a base é particionada é estudado de maneira similar.

2.4.1 Redes Bayesianas

Redes Bayesianas, ou redes probabilísticas, são estruturas de dados capazes de representar as dependências condicionais entre diferentes variáveis probabilísticas. Uma vez que um

problema ou domínio é descrito na forma de uma rede Bayesiana, uma série de técnicas existem para realizar inferências e descobrir os valores de variáveis desconhecidas (BARBER, 2014).

Millán et al. (2000) definem uma topologia de rede Bayesiana para aferir os conhecimentos de alunos em avaliações educacionais. Semelhante a uma árvore, as folhas representam itens de uma avaliação e os níveis superiores representam agregados cada vez menos granulares de conhecimentos (como conceitos, tópicos e assuntos). Os valores probabilísticos das conexões são escolhidos por especialistas. Essa topologia permite extrair quanto um aluno domina determinados conteúdos: uma vez que um conceito, tópico ou assunto é representado por um nó na rede, extrai-se a pontuação do examinando em determinada área do conhecimento somando-se os valores de todos os filhos deste nó.

Baseado em Millán et al. (2000), Kim e Choi (2012) criaram uma rede Bayesiana na qual sua topologia representava a hierarquia de tópicos de conhecimento em um TAI. A técnica foi testada utilizando as respostas de 160 alunos em dois testes diferentes, cada um contando com 13 e 6 categorias de tópicos. Como a base de treinamento estava incompleta (pois possuíam as respostas dos examinandos aos itens, mas não suas classificações) o problema foi contornado utilizando o algoritmo EM para estimar as probabilidades das variáveis remanescentes.

Os resultados da rede Bayesiana foram validados calculando seu coeficiente de correlação com os resultados dos alunos nos testes em lápis-e-papel. Os resultados de ambos os testes, assim como simulações adicionais baseadas na TRI e em uma rede *naïve Bayes*, foram normalizados somando-se a pontuação que cada pergunta valia com a probabilidade de acertá-la, sendo que esta probabilidade divergia nos diferentes modelos (KIM; CHOI, 2012). No final, os autores observaram que o modelo proposto precisava de menos itens para que a precisão da pontuação convergisse: 17 no primeiro teste, contra 21 pelo modelo que utilizava a TRI; e 14 no segundo teste, contra 23 no modelo que utilizava a TRI.

Vomlel (2004) criou uma plataforma para construir estratégias de decisão utilizando redes Bayesianas e aplicou-a em um teste adaptativo. Para isso, um modelo de estudante foi criado utilizando um algoritmo para descoberta de estruturas causais e uma série de respostas de alunos a um teste no modelo lápis-e-papel. O modelo foi refinado por um especialista na área de conhecimento do teste e os valores das novas probabilidades foram descobertos utilizando o algoritmo EM. Com o modelo pronto, o autor encontrou a sequência de itens a serem respondidos por um examinando de forma a diminuir a entropia de Shannon a cada passo, dadas as evidências até o momento. Dado que todas as sequências possíveis de itens em um teste adaptativo formam uma árvore, foi utilizada uma abordagem de programação dinâmica, calculando-se

a entropia dos nós-folha da árvore e utilizando os resultados para compor a entropia dos nós superiores.

Utilizando um limite inferior para o valor da entropia de cada nó, foi criada uma heurística admissível, a qual foi implementada no algoritmo de busca AO* para selecionar o nó com a menor entropia de cada nível da árvore, diminuindo o espaço de busca. O modelo foi validado realizando uma simulação baseada nos dados coletados de um teste específico. 90% das proficiências foram estimadas corretamente após 7 itens, de um total de 20. Utilizando uma abordagem miópica (escolha do item que diminui a entropia para o próximo nível da árvore) a entropia decresceu logaritmicamente (VOMLEL, 2004).

Desmarais, Pu e Blais (2007) apresentam um novo modelo de teste adaptativo informatizado que utiliza uma rede Bayesiana genérica (modelo *naïve Bayes*) para inferir as probabilidades de acertos futuros a itens dadas as respostas até o momento. Os itens são representados por nós na rede, enquanto as conexões entre os nós é feita calculando-se a distribuição de frequências de respostas corretas aos itens: itens que possuem uma alta correlação entre suas respostas têm uma conexão criada entre si.

Os itens futuros são escolhidos através da fórmula de informação de Fisher, da mesma forma que a TRI o faz. Esta e outras semelhanças entre os modelos possibilitaram a comparação dos resultados entre si, demonstrando que a rede Bayesiana obteve desempenho equivalente e até superior em certas partes dos experimentos (DESMARAIS; PU; BLAIS, 2007).

Millán et al. (2013) modelaram uma rede Bayesiana genérica para aferir o conhecimento de alunos em tópicos de matemática. Semelhante a uma árvore, as folhas representavam itens de uma avaliação e os níveis superiores representavam níveis cada vez menos granulares de conhecimentos (como conceitos, tópicos e assuntos). Os valores probabilísticos das conexões foram escolhidos por especialistas e a validação foi feita utilizando uma prova de matemática já existente. 152 alunos fizeram testes tanto escritos quanto digitais e seus resultados finais (o valor do nó raiz) foram comparados com as notas dadas por três especialistas. Houve certas discrepâncias nos resultados, as quais foram justificadas pelas características da prova (os alunos eram obrigados a responder lotes de 4 perguntas simultaneamente) e das perguntas (verdadeiro ou falso, dando probabilidade de acerto ao acaso de 0,5).

2.4.2 Redes Neurais

Uma rede neural é uma estrutura de processamento paralelo utilizada na classificação de instâncias. A rede recebe uma série de amostras classificadas como base de treinamento e

cada unidade de processamento da rede, denominada neurônio, é responsável pela adequação dos pesos da rede no geral, sendo que, ao final do treinamento, a rede com os pesos ajustados será responsável pela classificação de instâncias da qual ela não possui informações *a priori* (HAYKIN, 1999).

Benitez Rochel, Trella Lopez e Conejo Muñoz (2000) utilizaram duas redes neurais no modelo *competitive learning* para determinar a proficiência de respondentes a testes de acordo com a TRI. A primeira rede neural utiliza o conceito *winner-takes-all* não supervisionado comum, enquanto a segunda utiliza o algoritmo supervisionado *Learning Vector Quantization* de Kohonen, que se baseia no princípio de aprendizado "reforçar-ou-punir".

As entradas consistem nas respostas do examinando a um determinado teste e as saídas são uma forma discretizada de θ, que ao invés de ser representado por valores de $-\infty$ a ∞, começa a pertencer a um espaço de valores $[0, N]$, sendo que um valor $0 \leq m \leq N$ pode representar sucesso no teste. Apesar dos esforços dos autores, os testes mais promissores classificaram 86,8% dos examinandos corretamente (BENITEZ ROCHEL; TRELLA LOPEZ; CONEJO MUÑOZ, 2000).

El-Alfy e Abdel-Aal (2008) utilizaram aprendizado de máquina abdutivo na construção de testes educacionais e avaliação dos examinandos. Um mecanismo indutivo abdutor (tradução livre de *abductory inductive mechanism*) é uma técnica supervisionada de aprendizado de máquina capaz de realizar regressões não-lineares nos dados. O treinamento da ferramenta consiste na criação de uma rede onde cada nó é responsável pela solução de uma função polinomial, sendo que novos nós são adicionados iterativamente ao modelo, possibilitando à rede classificar dados através de polinômios cada vez mais complexos, respeitando alguns parâmetros pré-definidos que evitam o *overfitting* da rede.

Utilizando uma base com 45 itens e 2000 indivíduos, o objetivo do experimento era classificar os indivíduos em dois grupos distintos de acordo com seu desempenho no teste: os que possuíam proficiência acima da média e aqueles abaixo da média. Foram utilizados 1500 indivíduos como treinamento, resultando em uma rede que utilizava 12 dos 45 itens para realizar a classificação. Os outros 500 indivíduos foram utilizados na validação da técnica, a qual alcançou um erro de 9,4% na classificação, comparado a 6,15% quanto utilizado o teste inteiro (EL-ALFY; ABDEL-AAL, 2008).

Baylari e Montazer (2009) criaram um sistema multi-agente utilizado na aplicação de Teste Adaptativo Informatizado baseados na TRI a alunos, avaliação dos resultados dos testes e subsequente recomendação de estudos.

Inicialmente, um banco de itens é calibrada através da aplicação dos itens aos alunos no formato lápis-e-papel. Os alunos utilizam um Sistema Tutorial Inteligente para assistir ao conteúdo da matéria, conteúdo este separado em Objetos de Aprendizado (*Learning Objects, LO*) e, posteriormente, são sujeitos a testes no formato adaptativo, os quais utilizam a maximização da função de informação como parâmetro de escolha de novos itens durante o curso do teste. O desempenho do aluno é então alimentado a uma rede neural perceptron com *back-propagation*, cuja saída consiste nos códigos das *LOs* recomendadas para melhorar ainda mais o desempenho do aluno (BAYLARI; MONTAZER, 2009).

O sistema foi validado comparando as saídas da rede neural com recomendações feita por especialistas, sendo que a rede recomendou as mesmas *LOs* que o especialista em 83,3% dos casos (BAYLARI; MONTAZER, 2009).

Yu (2009) utilizou trinta redes neurais do modelo *Generalized Regression Neural Network* (GRNN) para estimar os parâmetros dos itens de acordo com a TRI. As entradas consistiam na dificuldade e discriminação dos itens de acordo com a TCT e as saídas, dificuldade e discriminação de acordo com a TRI. A média das 30 redes foi utilizada como saída final do sistema. Os resultados foram validados pelo cálculo do desvio médio da raiz quadrada entre o sistema criado e a saída do programa BILOG-MG, sendo que as 30 GRNN tiveram desvios médios menores.

2.4.3 Árvores de decisão

Árvores de decisão são estruturas de classificação que utilizam o conceito de entropia para a criação da hierarquia de características, nas quais características mais próximas da raiz da árvore são responsáveis pela classificação da maioria das instâncias (MITCHELL, 1997).

Em Ueno e Songmuang (2010), foi criado um TAI no qual todos os padrões de resposta possíveis dos respondentes são pré-computados utilizando-se o algoritmo de aprendizado ID3 (QUINLAN, 1986). Dessa forma, foi criada uma árvore de decisão na qual nós internos representam itens, arestas representam a probabilidade do respondente acertar um item (levando-o a outro nó) e nós externos, as proficiências. Respostas erradas fazem o examinando tender a itens mais fáceis, os quais levam a nós externos com θ menor.

2.4.4 Abordagens evolucionárias

Um Algoritmo Genético (GA) é uma técnica de busca por resultados ótimos ou subótimos em um espaço de estados. Os GA se utilizam de procedimentos que emulam o processo

de seleção, como a seleção dos indivíduos mais aptos de uma população para a próxima população (através do uso de uma função de aptidão); o cruzamento entre membros de uma população; e a mutação de indivíduos. A programação genética, por sua vez, utiliza conceitos de otimização semelhantes aos dos GA na criação evolucionária de programas. Nela, árvores são criadas, nas quais nós terminais representam valores e nós intermediários, funções (AFFENZELLER et al., 2009).

Chen e Doong (2008) utilizaram programação genética para descobrir a relação entre os valores dos parâmetros dos itens de acordo com o ML3 e o valor da taxa de exposição (r) de cada item calculado através do método de Sympson-Hetter (SYMPSON; HETTER, 1985). Foram utilizadas três bases de itens sintéticas com 360 itens cada e simulados 1000 testes adaptativos, com um limite máximo de 20 itens. Os itens eram selecionados de forma a maximizar sua função de informação $I(\theta)$, porém itens cuja taxa de exposição se aproximava, ou ultrapassava, a taxa de exposição r^{max} pré-determinada tinham probabilidade reduzida de serem administrados. Eles eram, então, substituídos pelo item segundo com o segundo maior valor de $I(\theta)$.

Após o término dos testes, e com os valores de r devidamente estimados, foi utilizada a técnica de programação genética de forma a encontrar a função que mapeasse os valores dos parâmetros dos itens às suas respectivas taxas de exposição. O conjunto de funções disponível foi $(+, -, *, /, \exp, \log, \sqrt{})$. A função resultante foi validada através da raiz dos erros quadráticos médios entre r e r_{GP} (sendo r_{GP} a taxa de exposição calculada através da função resultante), a qual variou entre $0{,}11$ e $0{,}15$ (CHEN; DOONG, 2008).

Li et al. (2012) criaram um sistema de construção automática de cursos online utilizando duas abordagens evolucionárias, algoritmos genéticos e *particle swarm optimization*. O processo, dividido em quatro etapas, consistia na criação de um grafo direcionado de conceitos, os quais esperava-se ensinar aos alunos. Em seguida, um banco de itens era construído, baseada no grafo de conceitos, e um teste era aplicado aos alunos para descobrir quais conceitos deveriam ser reforçados. Por último, uma das duas abordagens evolucionárias era responsável por selecionar entre os materiais disponíveis em um repositório de conteúdo e criar um curso online personalizado para cada aluno. A função de aptidão levava em consideração os erros do aluno no teste, a dificuldade dos conceitos que deveriam ser reforçados e o tempo para aprendê-los.

Para validar a estratégia, dois grupos de alunos realizaram o curso online, sendo que ao primeiro grupo foi apresentado conteúdo predeterminado, enquanto o segundo realizou o teste e, em seguida, realizou o curso personalizado. A taxa de aprovação do primeiro grupo com

relação ao conteúdo do curso foi de 63%, enquanto no segundo grupo a aprovação foi de 93,5% (LI et al., 2012).

Lotito e Pirlo (2013) utilizaram um GA que selecionava itens de um banco calibrado pela TRI para montar questionários para alunos de faixas de proficiências utilizando distintas. A população era inicializada aleatoriamente, sendo que cada indivíduo da população do GA representava uma prova candidata, que era, por sua vez, representada por um vetor binário V, onde $V_i = 0$ indicava a ausência do item i do banco na prova candidata. Era realizado cruzamento do tipo *one-point* através do método de seleção de roleta, na qual a probabilidade de um item ser selecionada é diretamente proporcional à sua aptidão, e mutação com probabilidade de 2%.A função de aptidão era $F = P(\theta_{max}) - P(\theta_{min})$, onde θ_{max} indica a maior proficiência entre os examinandos para os quais a avaliação era voltada, e θ_{min}, a menor proficiência do conjunto.

Utilizando uma base de 100 itens, foram criados diversos questionários de tamanhos $n = 5, 10$ e 15, os quais foram utilizados tanto em testes simulados quanto empíricos. Nos testes simulados, a estimativa de proficiência foi mais precisa utilizando os questionários gerados pelo GA do que questionários construídos com itens aleatórios. Nos dados reais, foi comprovada uma queda no desvio padrão das estimativas de proficiência dos examinandos, demonstrando que o questionário gerado pelo GA também foi mais preciso (LOTITO; PIRLO, 2013).

2.4.5 Modelos mistos

El-Alfy e Jafri (2007) investigaram a plausibilidade de se usar diferentes tipos de redes neurais na estimativa das proficiências de examinandos em Teste Adaptativo Informatizado. Foram testados três tipos de redes neurais: perceptron multi-camadas; perceptron multi-camadas processando apenas componentes principais das respostas dos examinandos e função de base radial, além de *Support-Vector Machines* (SVM). As entradas consistiam nas respostas dos examinandos e as saídas, em seu nível de proficiência.

Foram realizados testes em duas bases de dados: uma base com 20 itens dicotômicos, 500 respondentes e 2 níveis de proficiência e outra com 20 itens dicotômicos, 500 respondentes e 5 níveis de proficiência. 75% da base foi usada como treinamento e 25% como teste. Os tempos de convergência das redes foram similares, assim como o grau de precisão. A quantidade de neurônios na cama escondida afetou apenas o desempenho do MLP, diminuindo seu erro quadrático médio (EL-ALFY; JAFRI, 2007).

Kastrin (2009) aplicou uma série de técnicas para simplificar o processo de estimativa da expressão de amostras de um micro-arranjos de DNA. Utilizando duas bases de amostras de DNA, foi realizado o agrupamento hierárquico de ambas as bases, utilizando-se a distância Euclidiana entre as amostras e o algoritmo de Ward para minimizar a variância dos grupos. A quantidade de grupos experimentada foi k = 2, 3, 4, 5.

As amostras, que são representadas por valores contínuos, foram então discretizadas em vetores binários, os quais serviram de entrada para o modelo de Rasch da TRI. No modelo de Rasch, cada valor binário do vetor representa a expressão do gene, sendo que um valor de 1 indica um gene de expressão alta, e a probabilidade calculada pelo modelo logístico indica a probabilidade de uma amostra ter alta expressão. Os valores estimados pelo modelo foram então utilizados como entrada para dois modelos de predição, SVM e árvore de decisão, para que novas amostras pudessem ter seu grau de expressão calculado de maneira mais simples (KASTRIN, 2009).

Apesar de se desviar das pesquisas relacionadas a TAI, o trabalho de Kastrin demonstra a união de duas técnicas de interesse para este trabalho, que são o agrupamento hierárquico e a aplicação da TRI de forma inovadora, organizando e agrupando amostras biológicas ao invés de itens e respondentes de um teste.

3 TESTE ADAPTATIVO INFORMATIZADO BASEADO EM AGRUPAMENTO POR SIMILARIDADE

Este trabalho propõe um novo método de seleção de itens para ser utilizado durante a aplicação de Teste Adaptativo Informatizado. O método, denominado *Cluster-based Item Selection Method* (CISM), se utiliza do agrupamento por similaridade de um banco de itens calibrados pela Teoria da Resposta ao Item. Como características de entrada para o processo de agrupamento, são utilizados os parâmetros dos itens extraídos de acordo com o modelo logístico de 3 parâmetros da TRI. Esses grupos de itens são então utilizados durante a aplicação de um TAI, de forma que itens que estejam em um mesmo grupo possam ser substituídos entre si durante a etapa de escolha de itens do TAI, gerando maior variabilidade no uso dos itens, o que, hipoteticamente, visa diminuir a taxa de exposição de cada item. As consequências da substituição de itens, mais precisamente o impacto na taxa de sobreposição de itens e na estimativa da proficiência dos examinandos, serão utilizados na validação e comparação do método proposto com métodos já disponíveis na literatura.

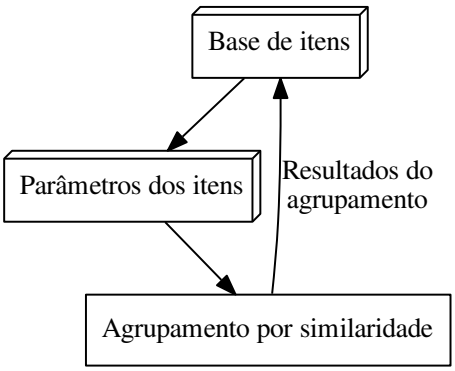
A principal justificativa para a escolha do agrupamento por similaridade é sua capacidade de agrupar dados de maneira não-supervisionada, uma vez que itens em uma base não são comumente pré-classificados. Utilizando o modelo logístico de 3 parâmetros da TRI, os itens, antes descritos de forma textual, passam a ser representados por características numéricas, as quais são então utilizadas como parâmetros de entrada para um algoritmo de agrupamento.

3.1 AGRUPAMENTO

Na fase de agrupamento, os parâmetros dos itens são utilizados como entrada para um algoritmo de agrupamento por similaridade, de forma que itens que possuam parâmetros com valores similares se situem no mesmo grupos, e itens cujos valores dos parâmetros sejam distantes sejam designados a grupos distintos. Sob esta descrição, foi decidido estudar técnicas de agrupamento que visem minimizar a soma das variâncias intra-grupos, uma vez que o objetivo destas técnicas é justamente agrupar dados cujos valores das características aproximem-se da média do grupo.

A figura 6 demonstra o processo de aplicação de um algoritmo de agrupamento por similaridade no CISM. Dado um banco de itens cujos parâmetros já foram estimados de acordo com o modelo de 3 parâmetros da TRI, o algoritmo de agrupamento utiliza apenas os parâmetros

Figura 6 – Fluxograma do processo de agrupamento de itens



Fonte: Autor

dos itens para realizar o agrupamento destes. Os resultados do agrupamento são então salvos no banco de itens, de forma que a informação referente ao grupos aos quais cada item foi designado sejam mantidas.

3.2 SELEÇÃO DE ITENS

Durante a etapa de seleção de itens de um TAI, utilizar-se-á o CISM, o qual recebe como parâmetros de entrada: o banco de itens, com todos seus parâmetros e taxas de exposição; a proficiência estimada $\hat{\theta}$ no ponto atual; a taxa de exposição máxima permitida para os itens, r^{max}; e o conjunto de itens que já foram aplicados ao examinando atual até o momento.

Primeiramente seleciona-se o item x de todo o um banco de itens, cujos parâmetros maximizem o valor da função de informação da TRI, apresentada na seção 2.1 e repetida abaixo.

$$I(\theta) = a_x^2 \frac{(P(\theta) - c_x)^2}{(1 - c_x)^2} \cdot \frac{Q(\theta)}{P(\theta)}$$

Seja g_x o grupo ao qual x pertence. Na segunda etapa, é selecionado o item em g_x que:

a) não tenha sido aplicado;

b) possua máxima informação;

c) tenha taxa de exposição menor que r^{max}.

Caso nenhum item satisfaça as três condições, é selecionado aquele que satisfaça apenas as duas primeiras. Se, mesmo assim, nenhum item satisfizer as duas condições, então o processo se inicia novamente, dessa vez selecionando o segundo item da base que maximize a informação, dado $\hat{\theta}$. O processo continua até que um grupo seja selecionado.

Ao selecionar o grupo que possua o item x de máxima informação, dado $\hat{\theta}$ atual, os itens presentes no mesmo grupo que x devem possuir parâmetros semelhantes aos de x, graças ao agrupamento por similaridade dos itens baseado nesses mesmos parâmetros. A priorização pela seleção de itens deste mesmo grupo aumenta as chances de alcançar alta precisão na estimativa das proficiências dos indivíduos; por outro lado, a preferência pelos itens que possuam baixa taxa de exposição visa utilizar de forma homogênea o banco de itens, de forma que itens similares àquele que maximize a informação sejam priorizados, focando primeiramente aqueles que não foram superexpostos.

O algoritmo 4 formaliza essa descrição.

O algoritmo possui algumas características importantes. A primeira diz respeito ao seu espaço de busca: uma vez que visita todos os grupos em busca daquele que possua itens ainda não aplicados, ele garante encontrar, a cada iteração do teste adaptativo, pelo menos um item para aplicação, contanto que o banco de itens seja maior que o número de itens presentes no teste.

A segunda característica diz respeito às circunstâncias nas quais o CISM se comporta como o método de seleção de itens por máxima informação.

a) **Não há restrições no uso dos itens:** caso o valor selecionado de r^{max} não restrinja o uso dos itens, o CISM se comporta como o método de seleção por máxima informação, descrito na seção 2.2.1.2, garantindo a maior precisão possível nas estimativas das proficiências;

b) **Todos os itens são super-expostos:** quando $\forall\ x \in X, r_x > r^{max}$, o CISM não possui candidatos com baixa taxa de exposição para selecionar, aplicando então o item que maximiza a função de informação. Este comportamento pode ser evitado com a contínua inclusão de itens na base, ou garantindo a escolha de um valor de r^{max} compatível com o uso dos itens;

c) **O número de itens por grupo é baixo:** sejam k o número de grupos, N o número de itens na base e seja N/k uma estimativa não-rigorosa do número médio de elementos por grupo. Então, $\lim_{k \to N} N/k = 1$. Quando esta situação ocorre, o CISM também se comporta como o método de seleção por máxima informação. Isso se dá devido à redução

Algoritmo 4 – CISM, o método de seleção de itens proposto

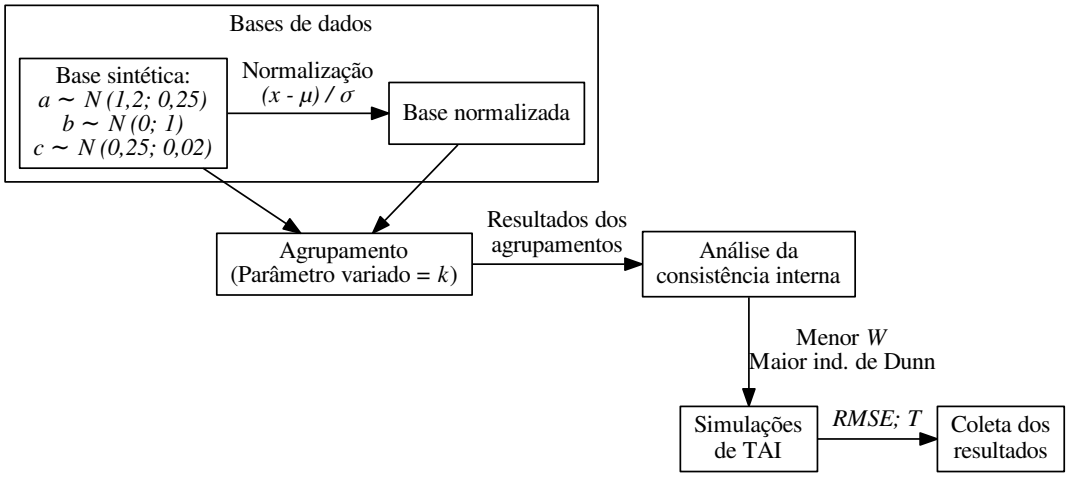
Entrada: Base de itens agrupados X, proficiência estimada $\hat{\theta}$, taxa de exposição máxima r^{max}, conjunto com itens já aplicados a este examinando $apresentados$
Função `busca_grupo` (*itens X, proficiência estimada $\hat{\theta}$, conjunto $apresentados$*)
  $U = \{\}$
  para cada $x \in X$ **faça**
    $x = \arg\max_{x \in X-U} I(\hat{\theta}|a_x, b_x, c_x)$
    $g_x =$ grupo ao qual x pertence
    se $\exists\, i \in g_x, i \notin apresentados$ **então**
      retorna g_x
    senão
      $U = U + x$
    fim
  fim
Função `busca_item` (*itens X, proficiência estimada $\hat{\theta}$, taxa de exposição máxima r^{max}, conjunto $apresentados$*)
  $g_x =$ `busca_grupo` (X, $\hat{\theta}$, $apresentados$)
  se $\exists\, j \in g_x, r_j < r^{max} \wedge j \notin apresentados$ **então**
    $j = argmax_{j \in g_x} I(\hat{\theta}|a_j, b_j, c_j), r_j < r^{max} \wedge j \notin apresentados$
  senão
    $j = argmax_{j \in g_x} I(\hat{\theta}|a_j, b_j, c_j), j \notin apresentados$
  fim
  Atualizar r_j
  $apresentados = apresentados \cup j$
  retorna j
retorna `busca_item` (X, $\hat{\theta}$, r^{max}, $apresentados$)

do número de itens por grupo: quanto menos itens existirem para que seja feito o controle da taxa de exposição, mais rapidamente os itens do grupo serão superexpostos. No limite, caso um grupo composto de apenas um item seja selecionado, este item será aplicado independente de sua taxa de exposição.

3.3 EXPERIMENTOS

A figura 7 descreve os experimentos propostos. Para realização dos experimentos, foi utilizado um banco de itens sintético, contendo 500 itens. Os três parâmetros de cada item foram extraídos das seguintes distribuições: $a \sim N(1{,}2; 0{,}25); b \sim N(0; 1); c \sim N(0{,}25; 0{,}02)$. A figura 8 apresentada as densidades resultantes da extração dos parâmetros de suas respectivas distribuições. Através dessa base, é possível verificar o desempenho do CISM em bancos de

Figura 7 – Fluxograma explicativo dos experimentos propostos



Fonte: Autor

tamanho maior que os disponíveis empiricamente, assim como comparar os resultados com os de outros autores que utilizaram bancos gerados de forma similar em seus experimentos (BARRADA; ABAD; VELDKAMP, 2009; BARRADA et al., 2009; BARRADA et al., 2009).

A figura 9a exibe a base em três dimensões, cada dimensão representa um dos três parâmetros. A figura 9b apresenta a mesma base redimensionada para duas dimensões utilizando análise de componentes principais, para melhor visualização. Em ambas as representações, é possível perceber a alta concentração dos dados.

Uma segunda versão da base, denominada de "base normalizada", foi criada através da normalização dos parâmetros dos itens, subtraindo-se de cada variável sua média μ e dividindo-os pelo desvio padrão σ,

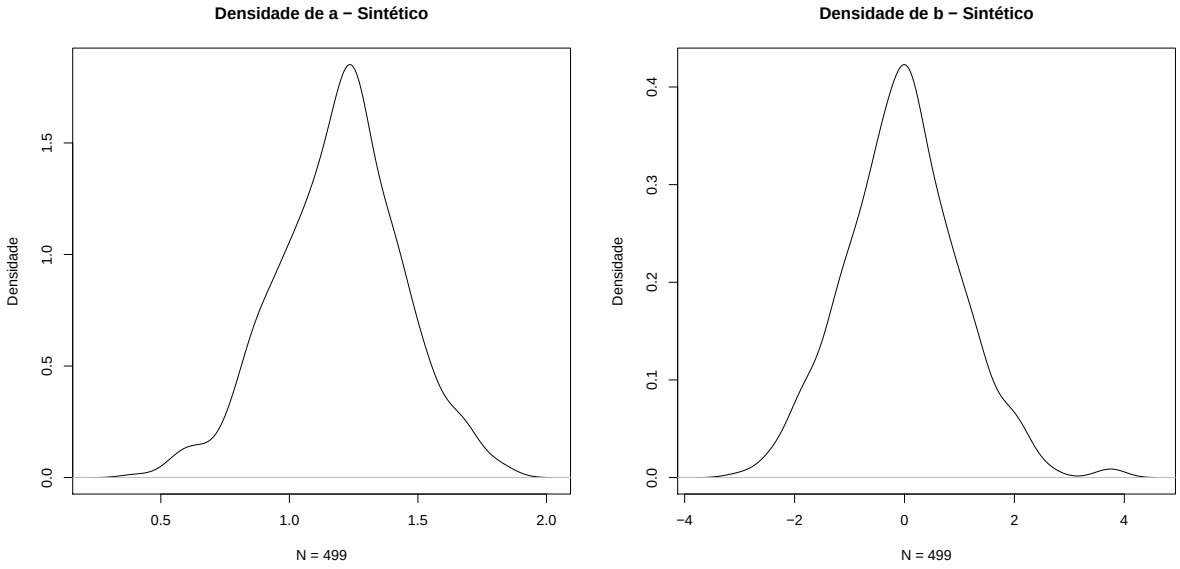
$$y = \frac{x - \mu}{\sigma},$$

para que os efeitos dos algoritmos de agrupamento sob os parâmetros possam ser observados.

3.3.1 Agrupamento dos itens

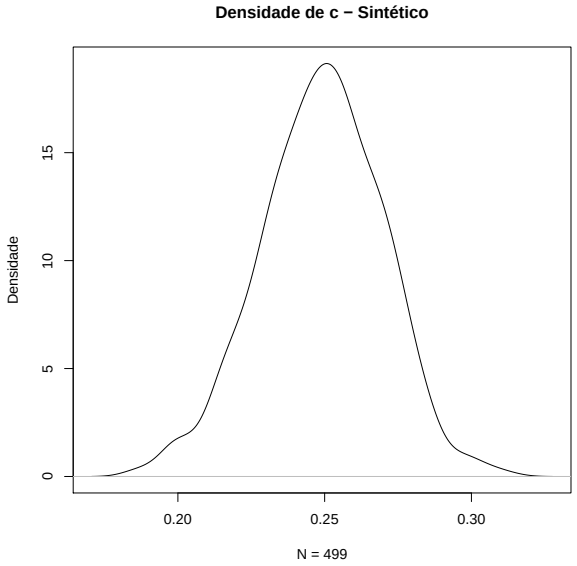
No primeiro experimento, os parâmetros dos itens de cada base foram utilizados como entrada para diferentes algoritmos de agrupamento, listados na tabela 2, com o intuito de gerar diferentes grupos de itens para serem utilizados na próxima etapa do experimento. Os algorit-

Figura 8 – Distribuições dos valores dos parâmetros a, b e c da base utilizada



(a) Discriminação (a)

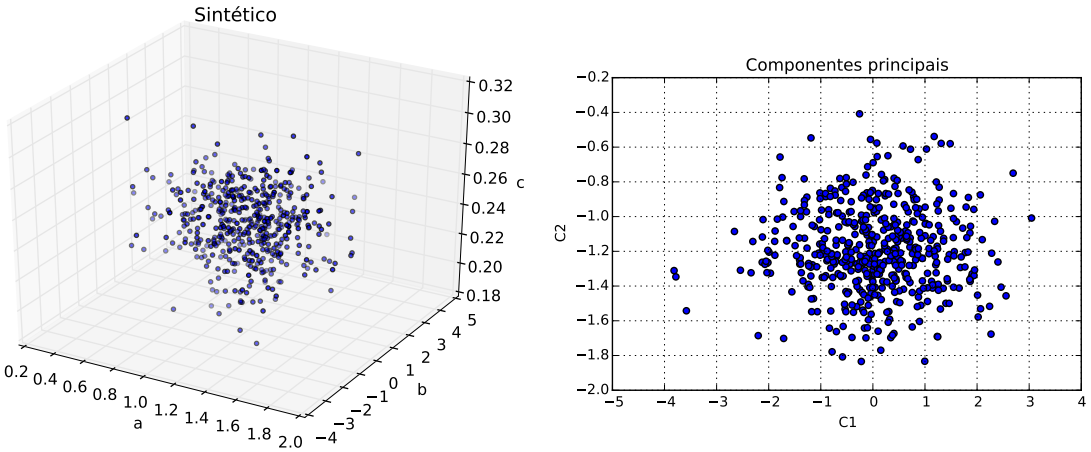
(b) Dificuldade (b)



(c) Probabilidade de acerto ao acaso (c)

Fonte: Autor

Figura 9 – Representação gráfica da base sintética utilizada nos experimentos



(a) Representação tridimensional

(b) Representação bidimensional das duas componentes principais da base

Fonte: Autor

mos foram selecionados por terem um parâmetro de entrada em comum: o número de grupos k, o que permite controlar a quantidade de grupos que serão utilizados pelo CISM na aplicação dos testes adaptativos. Adicionalmente, dada a natureza dos valores dos parâmetros do banco de itens (representado graficamente na figura 9), o controle sobre o valor de k permite a seleção do número de grupos gerados nos dados de alta densidade.

O *K-means* foi escolhido por seu foco na minimização da variância intra-grupos. Foi implementada uma versão em Python do algoritmo que se utiliza do início aleatório das k sementes, selecionadas aleatoriamente entre os dados disponíveis na base. Para se garantir resultados de agrupamento de baixa variância apesar do início aleatório, o algoritmo foi executado 100 vezes para cada valor de k, mantendo-se o resultado com menor variância. Os agrupamentos foram realizados utilizando as distâncias Euclideana e Mahalanobis: a primeira, por ser a distância padrão utilizada pelo algoritmo e a segunda, por levar em consideração a matriz de covariância dos dados, permitindo a geração de grupos elipsoidais mais acentuados.

Também foram utilizados algoritmos aglomerativos. A inclusão dos algoritmos aglomerativos se justifica por eles se basearem exclusivamente nas distâncias entre os dados, permitindo o uso de medidas de distância diferentes para o agrupamento dos dados. Os algoritmos de ligação simples, média, por média ponderada e completa realizaram os agrupamentos utilizando as distâncias Manhattan, Euclideana, Mahalanobis e Chebyshev. O algoritmo aglomerativo por função de Ward trabalha exclusivamente com a distância Euclideana, sendo a única utilizada

Tabela 2 – Algoritmos utilizados no experimento e as medidas de distância utilizadas. Em todos os algoritmos, o parâmetro de entrada é o número de grupos k.

Algoritmo	**Distância**
K-means	Euclideana, Mahalanobis
Ward K-means	Euclideana
Aglomerativo – Ward	Euclideana
Aglomerativo – Ligação simples	Manhattan, Euclideana, Mahalanobis, Chebyshev
Aglomerativo – Ligação completa	Manhattan, Euclideana, Mahalanobis, Chebyshev
Aglomerativo – Ligação média	Manhattan, Euclideana, Mahalanobis, Chebyshev
Aglomerativo – Ligação média ponderada	Manhattan, Euclideana, Mahalanobis, Chebyshev

Fonte: Autor

por ele. Nos experimentos realizados, foram utilizadas as implementações dos algoritmos aglomerativos na suíte de computação científica SciPy (JONES; OLIPHANT; PETERSON, 2001–).

Por último, foi incluído o algoritmo Ward *k-means* (KWEDLO, 2011). Nele, as sementes do *k-means* convencional são inicializadas através do cálculo dos centroides do agrupamento produzido pelo algoritmo aglomerativo por função de Ward. Uma vez que um dos desafios do *k-means*, além do número de sementes k, está na escolha das posições dos centroides iniciais, o uso do algoritmo aglomerativo por função de Ward pode ser utilizado como uma heurística para a execução de tal inicialização, uma vez que ambos os algoritmos visam a minimização da variância intra-grupos. A inclusão deste método de inicialização no trabalho é justificada por resultados preliminares de agrupamentos por função de Ward favoráveis, os quais serão apresentados na seção 4.

O valor de k foi incrementado unitariamente no intervalo $[2; 250]$. O valor 250 foi escolhido por ser a metade do número de itens na base, sob a perspectiva de que cada grupo gerado deve ter, no mínimo, dois itens, de forma que o CISM seja capaz de selecionar substitutos durante a aplicação dos testes adaptativos. A análise dos resultados dos agrupamentos é feita através da variação dos valores de k e da extração das seguintes medidas de validação para cada agrupamento realizado:

a) Soma das variâncias intra-grupos;

b) índice de Dunn.

A soma das variâncias intra-grupos, apresentada na seção 2.3.1, indica o quão distante da média de seus grupos cada dado está. Quanto menor o valor da soma das variâncias, mais compactos são os grupos, indicando maior similaridade entre seus membros.

Os índices de Dunn (KOVÁCS; LEGÁNY; BABOS, 2005) são uma família de índices de avaliação interna de agrupamentos. A equação base dos índices de Dunn é a seguinte,

$$D = \frac{\min_{1 \leq i < j \leq n_g} \delta\left(g_i, g_j\right)}{\max_{1 \leq k \leq n_g} diam\left(g_k\right)}, \tag{20}$$

onde $\delta(g_i,g_j)$ representa a distância entre os grupos g_i e g_j e $diam(g_k)$, o diâmetro do grupo g_k. Quanto maior o índice de Dunn, melhor é considerado o resultado do agrupamento, pois um alto valor do índice significa que os grupos são compactos e distantes entre si. Os diferentes índices na família dos índices de Dunn são dados pelas diferentes representações possíveis de distâncias entre grupos e seus diâmetros. Neste trabalho, $\delta\left(g_i,g_j\right)$ foi dado pela distância entre seus pontos mais próximos e $diam(g_k)$, pela distância entre os pontos mais distantes de g_k

3.3.2 Aplicação de testes

No segundo experimento, foi utilizada a metodologia proposta por Barrada et al. (2010) para realizar a simulação de Teste Adaptativo Informatizado, com o propósito de se avaliar o desempenho do CISM, utilizando como entrada os resultados dos agrupamentos do experimento anterior.

Dado um conjunto de valores $r_{1...10}^{max}$ linearmente separados entre $[0{,}1; 1]$, denominados taxas de exposição máximas dos itens, uma série de simulações de testes adaptativos é executada, com o intuito de aferir a precisão e o uso homogêneo dos itens da base sob diferentes restrições para a taxa de exposição máxima dos itens.

a) θ: 5000 valores extraídos de uma distribuição $N(0; 1)$;

b) Estimativa inicial de $\hat{\theta}$: extraído de uma distribuição $U(-5; 5)$;

c) Seleção de itens: Nesta etapa, é utilizado o CISM, descrito no início da seção;

d) Aplicação de itens: Para simular a resposta de um examinando a um item, utiliza-se a fórmula (1), que indica a probabilidade de um examinando com proficiência θ responder corretamente a um item com parâmetros a, b e c,

$$P_i(X_i = 1|\theta) = c_i + \frac{(1 - c_i)}{1 + e^{Da_i(\theta - b_i)}}$$

e) Re-estimativa de $\hat{\theta}$: a re-estimativa é feita utilizando-se duas equações diferentes. A primeira é a equação (11), proposta por Dodd (1990), a qual é utilizada quando o padrão de respostas do examinando se mantém constante, impedindo o uso dos métodos de máxima verossimilhança. A segunda é a função de log-verossimilhança para estimativa de $estheta$ quando os parâmetros dos itens já são conhecidos,

$$\log L(\boldsymbol{X}|\theta_j, \boldsymbol{\zeta}) = \sum_{i=1}^{N} \{x_{ij} \log P_{ij}(\theta) + (1 - x_{ij}) \log Q_{ij}(\theta)\} .$$

Maximizando-se esta função, é encontrado o valor de $\hat{\theta}$ mais provável, dado o vetor de respostas $\boldsymbol{X}$ e os parâmetros dos itens respondidos $\boldsymbol{\zeta}$. A maximização da função deu-se através do uso de um algoritmo de busca binária (DE AYALA, 2009, p. 347–348). A figura 10 exibe as curvas para a função de log-verossimilhança para vetores de respostas de 20 itens, com 5, 10 e 15 acertos, assim como os valores de $\hat{\theta}$ estimados através do algoritmo 5 para cada vetor.

f) Critério de parada: Decidiu-se utilizar como critério de parada a aplicação de um número fixo de 20 itens, a fim de estudar os efeitos dos diferentes resultados gerados pelos algoritmos de agrupamento modificando o mínimo possível de variáveis adicionais.

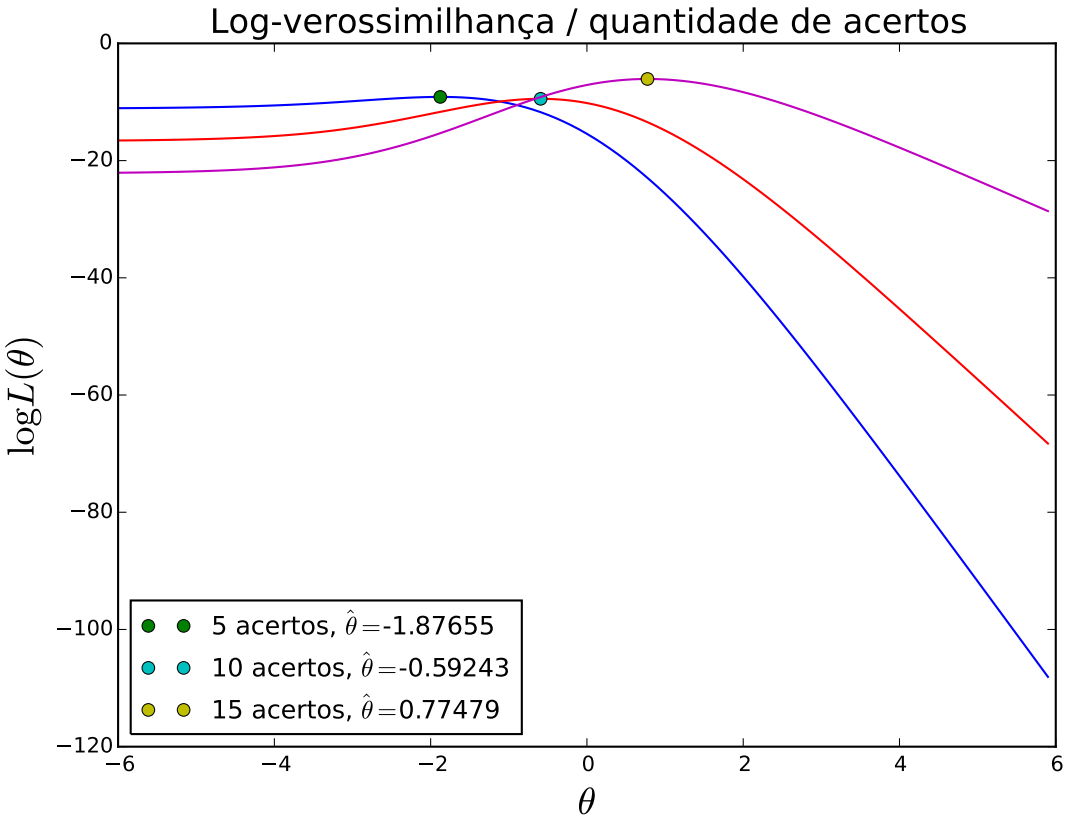
Algoritmo 5 – Busca binária utilizada na estimativa de $\hat{\theta}$

Entrada: Vetor de respostas $\boldsymbol{X}$, parâmetros dos itens administrados $\boldsymbol{\zeta}$, número máximo de iterações $iter$
Saída: $\hat{\theta}_j$
$l^- = \min_b(\boldsymbol{\zeta})$
$l^+ = \max_b(\boldsymbol{\zeta})$
para $i = 1 \rightarrow iter$ **faça**
  se $\log L(\boldsymbol{X}|l^+, \boldsymbol{\zeta}) > \log L(\boldsymbol{X}|l^-, \boldsymbol{\zeta})$ **então**
    $l^- = l^- + \frac{l^+ - l^-}{2}$
  senão
    $l^+ = l^+ - \frac{l^+ - l^-}{2}$
  fim
fim
$\hat{\theta} = \arg\max_{l \in \{l^-, l^+\}} \log L(\boldsymbol{X}|l, \boldsymbol{\zeta})$
retorna $\hat{\theta}$

A plataforma de simulação foi implementada na linguagem Python, versão 3.4 (MENEGHETTI, 2015).

A estratégia proposta por Barrada et al. (2010) tem como objetivo comparar os desempenhos de diferentes métodos de aplicação de Teste Adaptativo Informatizado sob diferentes

Figura 10 – Gráfico da função de log-verossimilhança para 5, 10 e 15 acertos em vetores de respostas de 20 itens e os respectivos máximos encontrados com o algoritmo de subida de encosta



Fonte: Autor

valores de r^{max}. Fixando-se o valor desta variável, é possível avaliar a precisão do método, calculando-se a raiz dos erros quadráticos médios entre θ e $\hat{\theta}$ através da equação (12); assim como avaliar o quão seguro é o método, calculando-se a taxa de sobreposição dos itens, dada pela equação (13).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{S}} \tag{12}$$

$$T = \frac{N}{Q}S_r^2 + \frac{Q}{N} \tag{13}$$

Dessa forma, os diferentes valores de $RMSE$ e T alcançados por cada método para cada valor de r^{max} podem ser comparados, evidenciando o comportamento de cada método sob diferentes restrições para o uso dos itens.

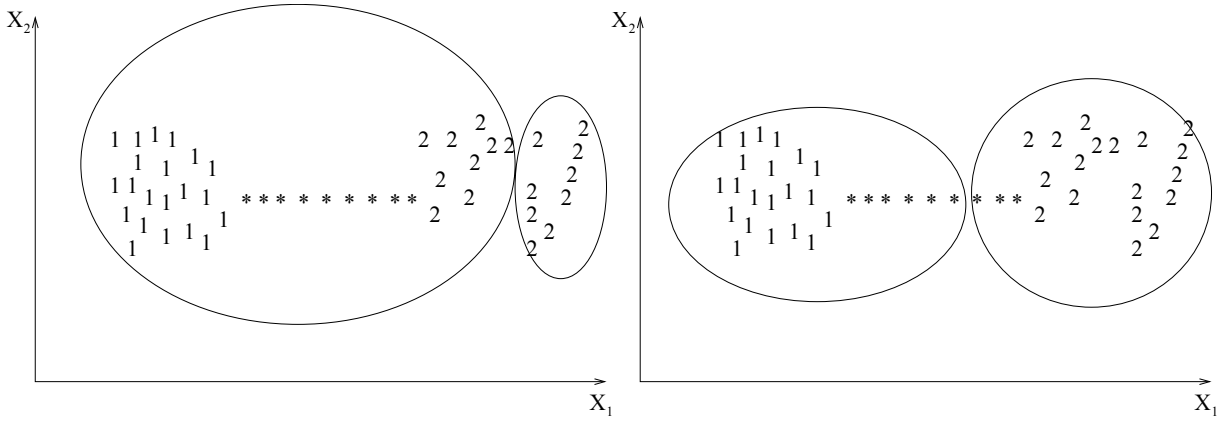
4 RESULTADOS

A figura 12 exibe graficamente os valores da soma das variâncias intra-grupos (W) de acordo com o aumento do parâmetro k para o banco de itens é sintético e sua versão normalizada. Em ambas as bases, houve diminuição de W conforme o aumento de k: devido ao maior número de grupos, grupos menores foram gerados, os quais, consequentemente, possuem menor variância. Os algoritmos que alcançaram menor W foram aqueles que têm como objetivo a minimização desse valor, ou seja, o algoritmo aglomerativo por função de Ward e o Ward *k-means*. A busca por mínimos locais do *k-means* aliada à inicialização dos centroides pelo método de Ward garantiu valores menores de variância para o Ward *k-means* do qua para o algoritmo aglomerativo por função de Ward.

Quanto ao índice de Dunn, apresentado na figura 13, é possível observar comportamentos diferentes para os algoritmos. Primeiramente, o algoritmo hierárquico por ligação completa foi o que obteve os melhores resultados de acordo com o índice. Isso se deve ao fato de sua predisposição em gerar grupos compactos e evitar a criação de grupos "alongados" (JAIN; DUBES, 1988), característica que eleva o valor do índice. Os algoritmos hierárquicos por ligação média, média ponderada e função de Ward tiveram desempenho mediano, por não possuírem nenhuma característica que leve efetivamente à diminuição do índice. O *k-means* por sua vez, não possui como foco a geração de grupos separados; com o aumento do valor de k, houve queda do índice de Dunn, indicando a presença de grupos próximos. Por último, o agrupamento por ligação simples teve o menor valor do índice de Dunn, exatamente por sua predisposição em gerar grupos de formato "alongado" ao invés de separar os mesmos grupos ao meio, como feito pelo algoritmo por ligação completa. A figura 11 demonstra graficamente este fenômeno.

O apêndice A apresenta os gráficos com os valores das medidas de validação dos agrupamentos individualmente para cada algoritmo, utilizando cada medida de distância. Neles, é possível perceber comportamento semelhante aos supracitados nos valores da variância e do índice de Dunn, independente da medida de distância utilizadas pelos algoritmos. Esta observação indica que, dentre as medidas de distância utilizadas nos experimentos, não há uma que possua desempenho superior, quando aplicados à base em questão.

Figura 11 – Diferenças nos agrupamentos aglomerativos por ligação simples e completa



(a) Ligação simples

(b) Ligação completa

Fonte: Autor “adaptado de” Jain, Murty e Flynn (1999)

4.1 AVALIAÇÃO DO CISM

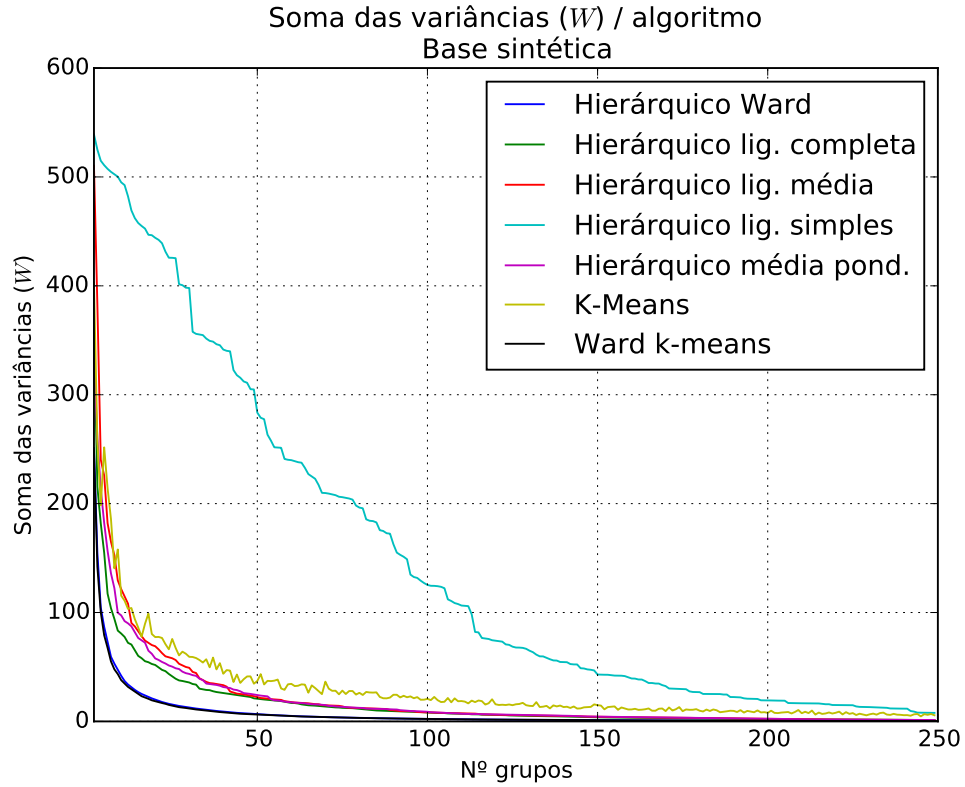
Para o segundo experimento, foram selecionados os resultados de agrupamento com menor variância ou maior índice de Dunn para a realização das simulações de Teste Adaptativo Informatizado. Por motivos computacionais, foram selecionados apenas os agrupamentos com valores de k múltiplos de 10, afim de se estudar a relação entre o número de grupos e os resultados dos testes adaptativos.

4.1.1 SIMULAÇÕES COM AGRUPAMENTOS DE MENOR VARIÂNCIA

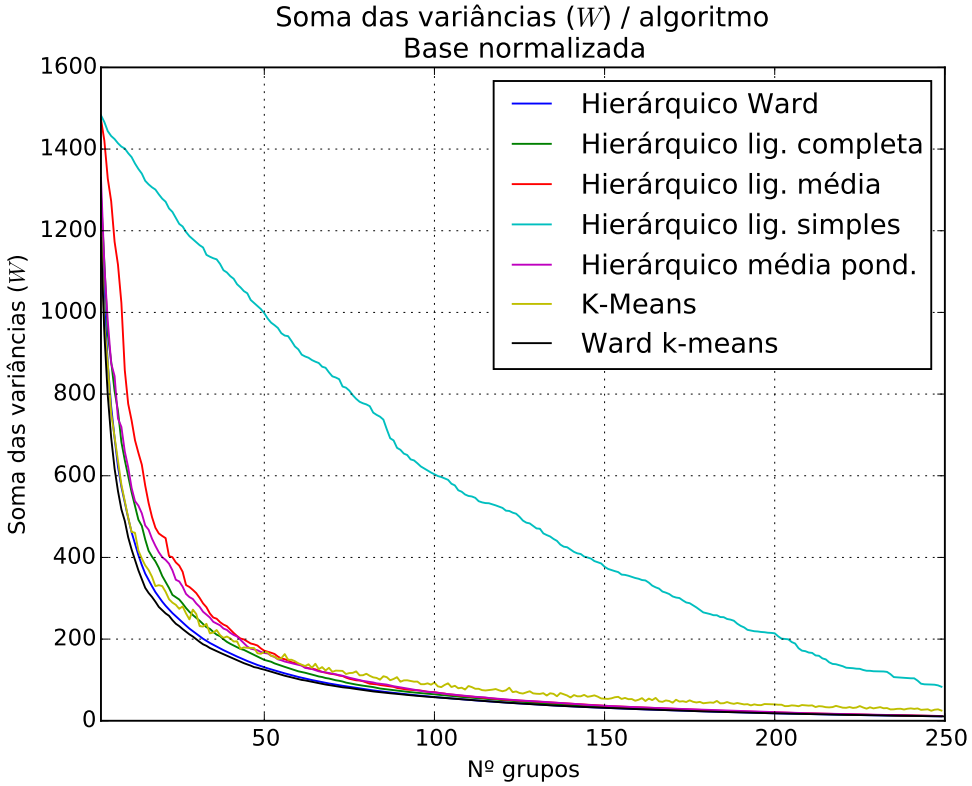
A tabela 3 exibe os agrupamentos que tiveram o menor valor de variância. Como pôde ser observado na figura 12, o algoritmo de agrupamento que obteve os menores valores de variância para todos os valores de k foi o Ward *k-means*.

A figura 14 exibe os valores da taxa de sobreposição (T) para simulações com diferentes valores de r^{max}. Para os menores valores de k (10, 20, 30 e 40), T inicia acima de $0{,}1$ e estabiliza quando $r^{max} \approx 0{,}7$. Para todos os outros valores de k, T manteve comportamento linear, indicando que, exceto para os menores valores de k testados, o número de grupos nos quais o banco de itens é separado não afeta a taxa de sobreposição da base. Este comportamento pode ser explicado pela relação entre k e o número médio de itens por grupo: quanto menos itens um grupo possui, mais rápido as taxas de exposição dos itens presentes no grupo alcançarão

Figura 12 – Soma das variâncias intra-grupos para os diferentes algoritmos utilizados, de acordo com a variação do número de grupos k



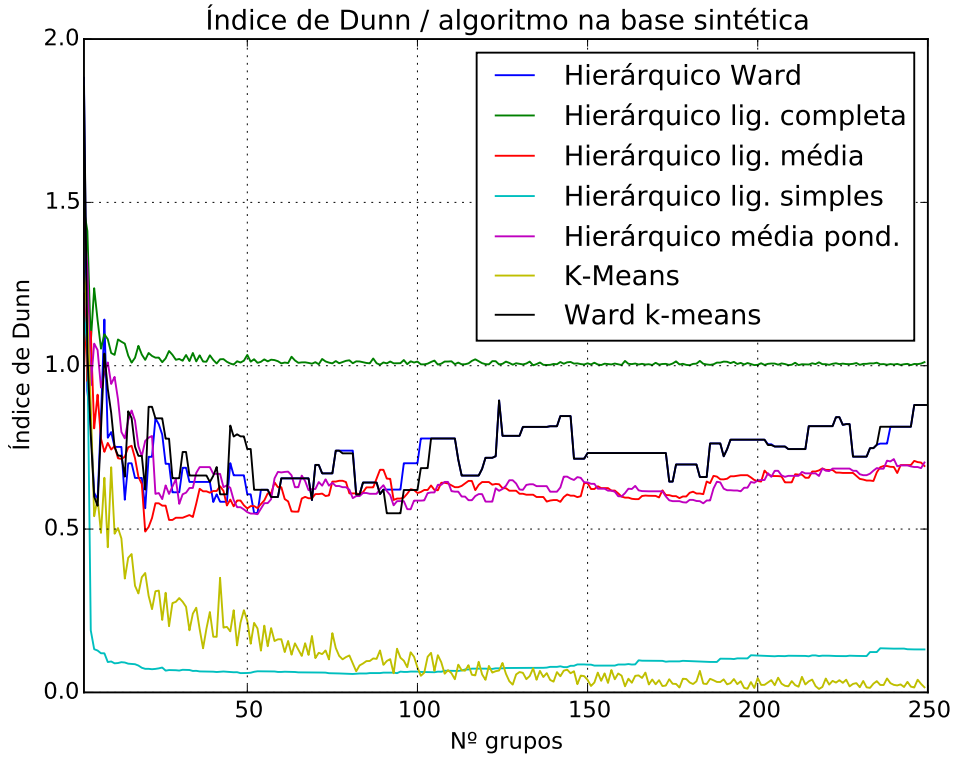
(a) Variância / algoritmo na base sintética



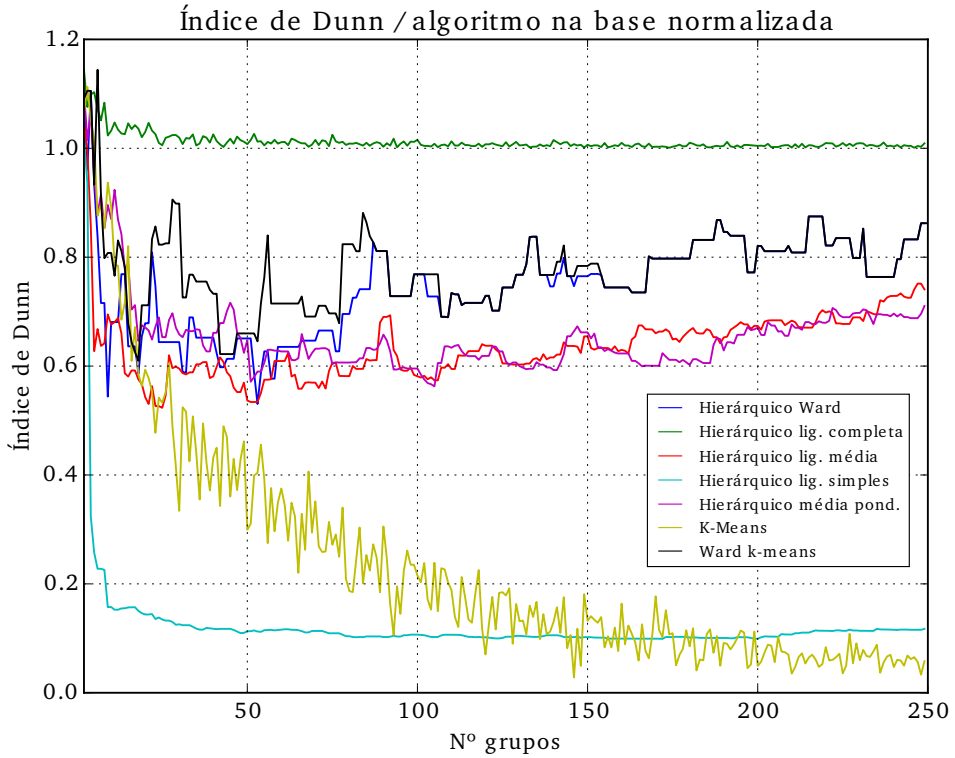
(b) Variância / algoritmo na base normalizada

Fonte: Autor

Figura 13 – Soma das variâncias intra-grupos para os diferentes algoritmos utilizados, de acordo com a variação do número de grupos k



(a) Índice de Dunn / algoritmo na base sintética



(b) Índice de Dunn / algoritmo na base normalizada

Fonte: Autor

Tabela 3 – Resultados de agrupamento com menor valor de variância escolhidos para execução das simulações de TAI. Todos os agrupamentos desta tabela foram extraídos com o algoritmo Ward *k-means* e a distância Euclideana.

Nº grupos	**Variância**	**Dunn**
10	37.88055	0.855691
20	18.512313	0.62069
30	11.831354	0.655638
40	8.312832	0.663084
50	6.156749	0.745014
60	4.849996	0.655282
70	3.846389	0.696388
80	3.182602	0.731726
90	2.647662	0.620803
100	2.249982	0.619565
110	1.919735	0.777255
120	1.642267	0.718186
130	1.435613	0.78538
140	1.256987	0.814713
150	1.110444	0.732679
160	0.984276	0.732679
170	0.866881	0.732679
180	0.768545	0.697711
190	0.682204	0.722682
200	0.607957	0.773602
210	0.542756	0.744836
220	0.486042	0.81498
230	0.433806	0.721679
240	0.390015	0.813216
250	0.34808	0.879754

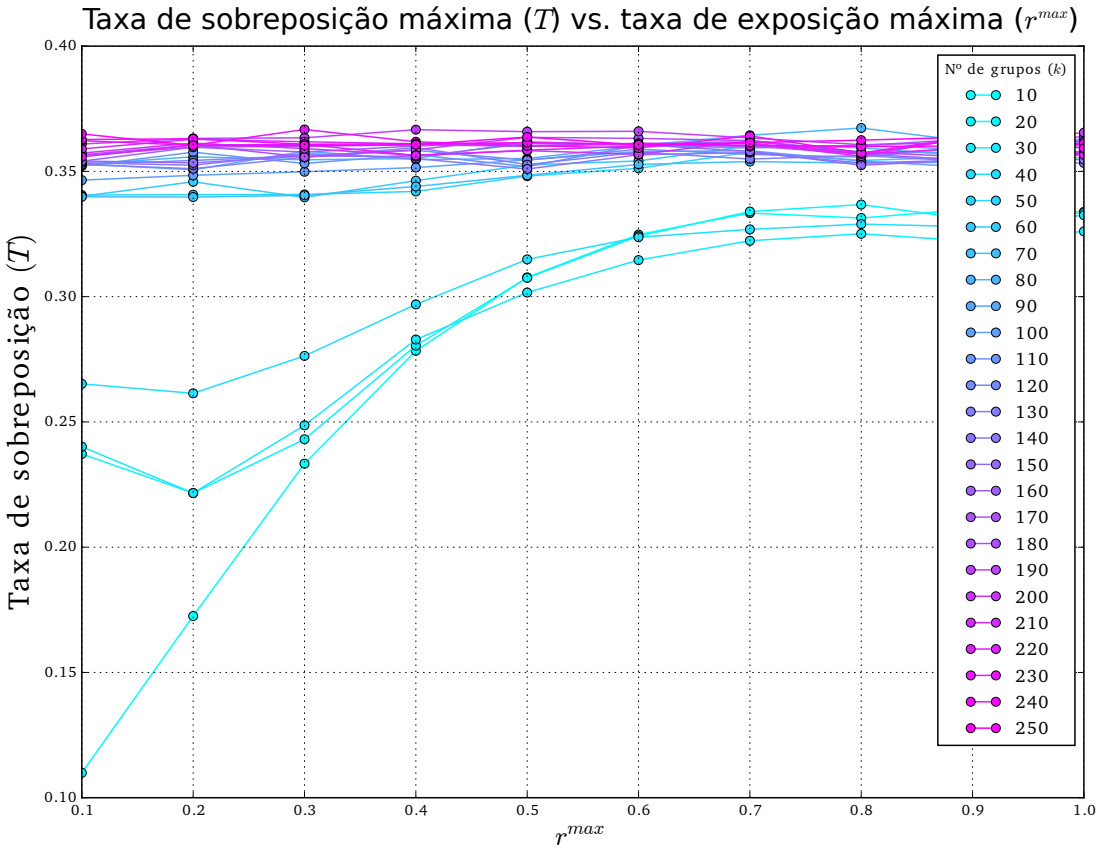
Fonte: Autor

r^{max}, fazendo com que o CISM se comporte como o método de seleção de itens por máxima informação.

Na figura 15 é possível observar a existência de maior variação nas taxas de sobreposição para valores pequenos de k. Conforme k aumenta, os valores de T passam a se tornar mais próximos. Isso demonstra duas coisas diferentes: a primeira, de que o método proposto é capaz de controlar a taxa de exposição do teste caso menos grupos sejam utilizados e a segunda, de que, para valores altos de k, restringir a taxa de exposição máxima dos itens surte menos efeitos na supressão da taxa de exposição do teste.

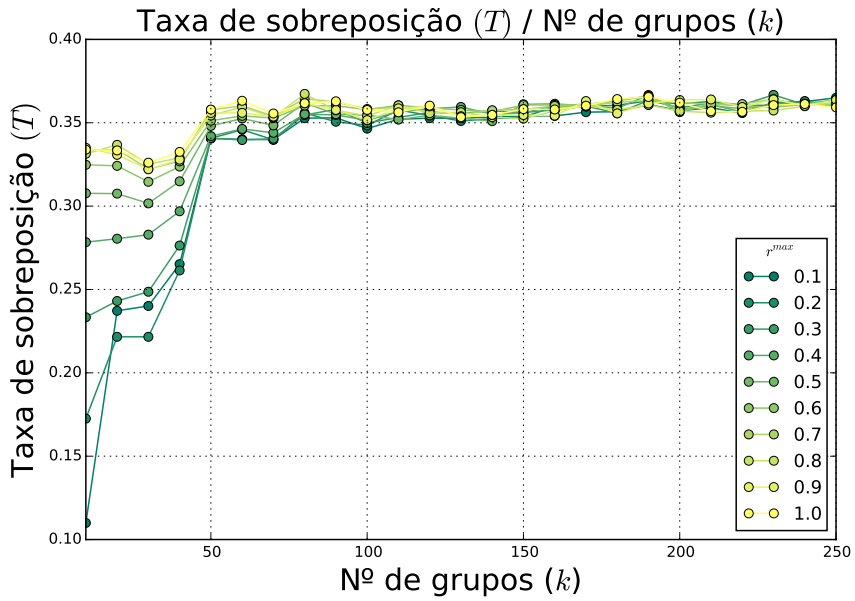
A figura 16 exibe os valores da raiz dos erros quadráticos médios ($RMSE$) para diversos valores de k e sob diferentes restrições de r^{max}. Aqui, é possível observar os resultados do controle de exposição demonstrado nas figuras anteriores: testes realizados com $k <= 40$

Figura 14 – Mudança da taxa de sobreposição de acordo com o aumento de r^{max} para diferentes números de grupos (agrupamentos de menor variância)



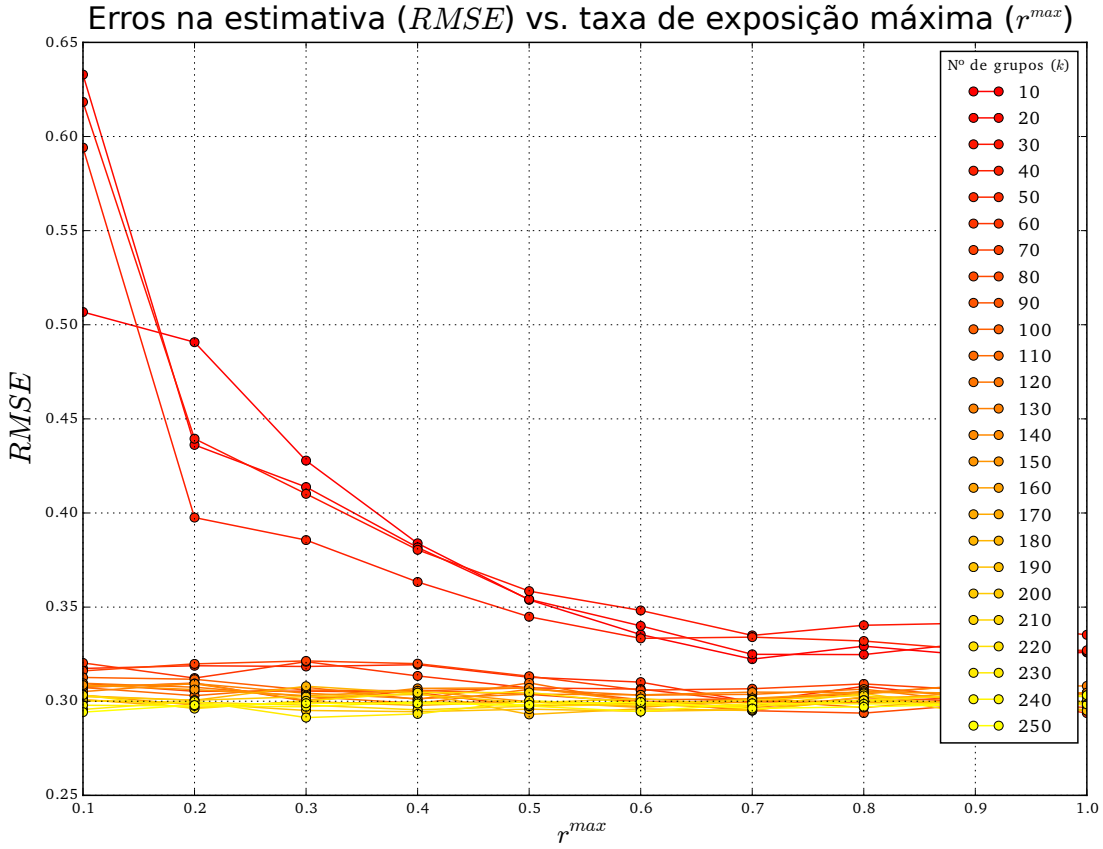
Fonte: Autor

Figura 15 – Relação entre a taxa de exposição e o número de grupos, para diferentes valores de r^{max} (agrupamentos de menor variância)



Fonte: Autor

Figura 16 – Mudança da raiz dos erros quadráticos médios de acordo com o aumento de r^{max} para diferentes números de grupos (agrupamentos de menor variância)



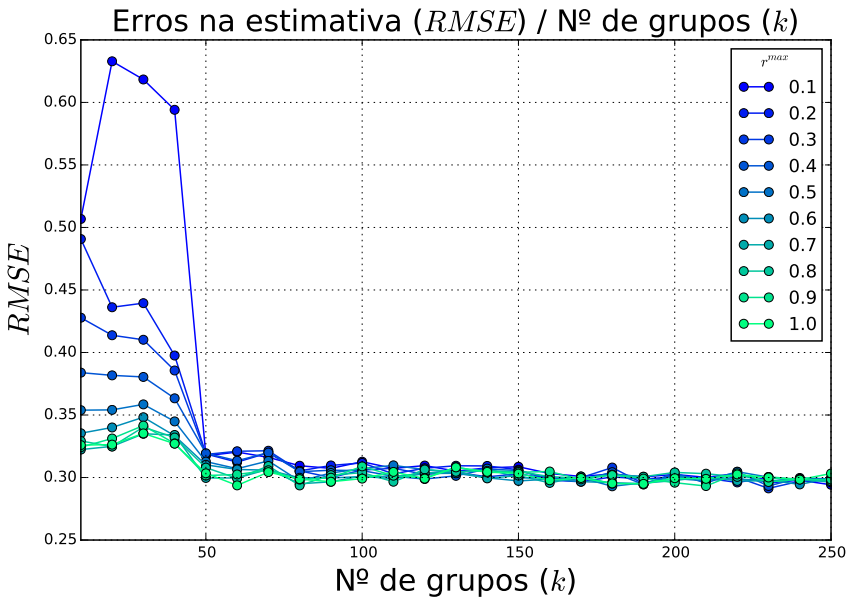
Fonte: Autor

resultaram não só em menores taxas de exposição como maiores erros. Para valores superiores de k, no entanto, é possível observar comportamento linear de $RMSE$, independente dos valores de r^{max}. Este comportamento, descrito na seção 3.2, tem relação ao número de itens em cada grupo: valores maiores de k resultam em grupos com menos itens, o que, por sua vez; com menos itens, estes grupos possuem menos controle da taxa de exposição, resultando em superexposição mais rápida dos itens.

A figura 17 confirma esta observação, demonstrando que valores maiores de k acarretam em menor erro na estimativa das proficiências. Porém, o método não se beneficia de valores muito altos de k: valores de $k > 50$ não demonstraram maior diminuição do erro na estimativa além da já alcançada.

Por fim, a figura 18 exibe a relação entre a raiz dos erros quadráticos médios e a taxa de sobreposição do teste para todas as simulações. É possível perceber uma relação linear entre as variáveis, da forma $RMSE = 0{,}733884081744 - 1{,}20735651462 \cdot T$, com coeficiente

Figura 17 – Relação entre a raiz dos erros quadráticos médios e o número de grupos, para diferentes valores de r^{max} (agrupamentos de menor variância)



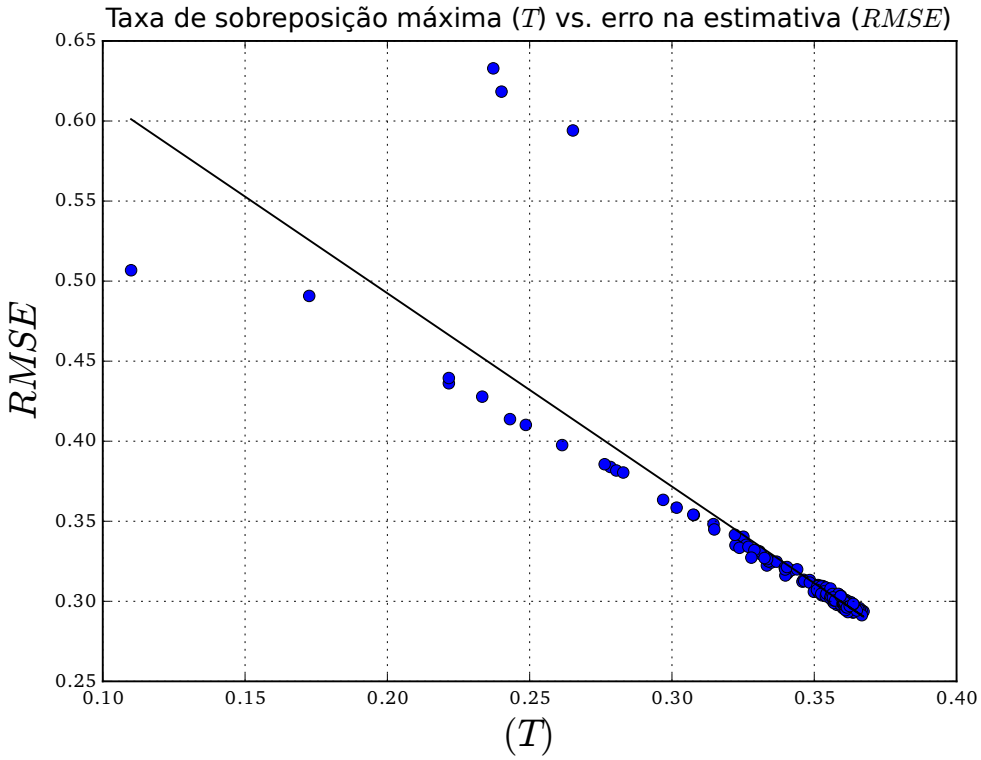
Fonte: Autor

de correlação $r = -0{,}873389310066$ e coeficiente de determinação $r^2 = 0{,}762808886937$, indicando alta correlação negativa entre as variáveis.

Estes resultados permitem observar as características previstas do CISM: para valores altos de k, onde $\lim_{k \to N}$, os grupos formados são constituídos por número cada vez mais reduzido de elementos. Dessa forma, não há itens suficientes no mesmo grupo para serem utilizados durante o processo de seleção do CISM, o que faz com que o método se comporte como o método comum de seleção de itens, o de máxima informação. Adicionalmente, para valores altos de r^{max} (nos experimentos, $r^{max} > 0{,}7$), não existem restrições suficientes sobre a taxa de exposição dos itens para que o CISM selecione itens de baixo r.

Nesta etapa dos experimentos, também foi possível observar que, para os valores mais baixos de k ($k \leq 50$) testados, houve aumento em $RMSE$ e diminuição de T. Essa observação demonstra que k pode ser utilizado como variável de controle, equilibrando os valores de $RMSE$ e T de acordo com as necessidades do teste. No entanto, os valores de k extraídos nos experimentos expostos são condicionados ao número de elementos no banco de itens utilizado, onde $N = 500$. Bases com diferentes números de elementos podem necessitar que k seja redimensionado de acordo.

Figura 18 – Relação entre a taxa de sobreposição e a raiz dos erros quadráticos médios, para diferentes valores de r^{max} (agrupamentos de menor variância)



Fonte: Autor

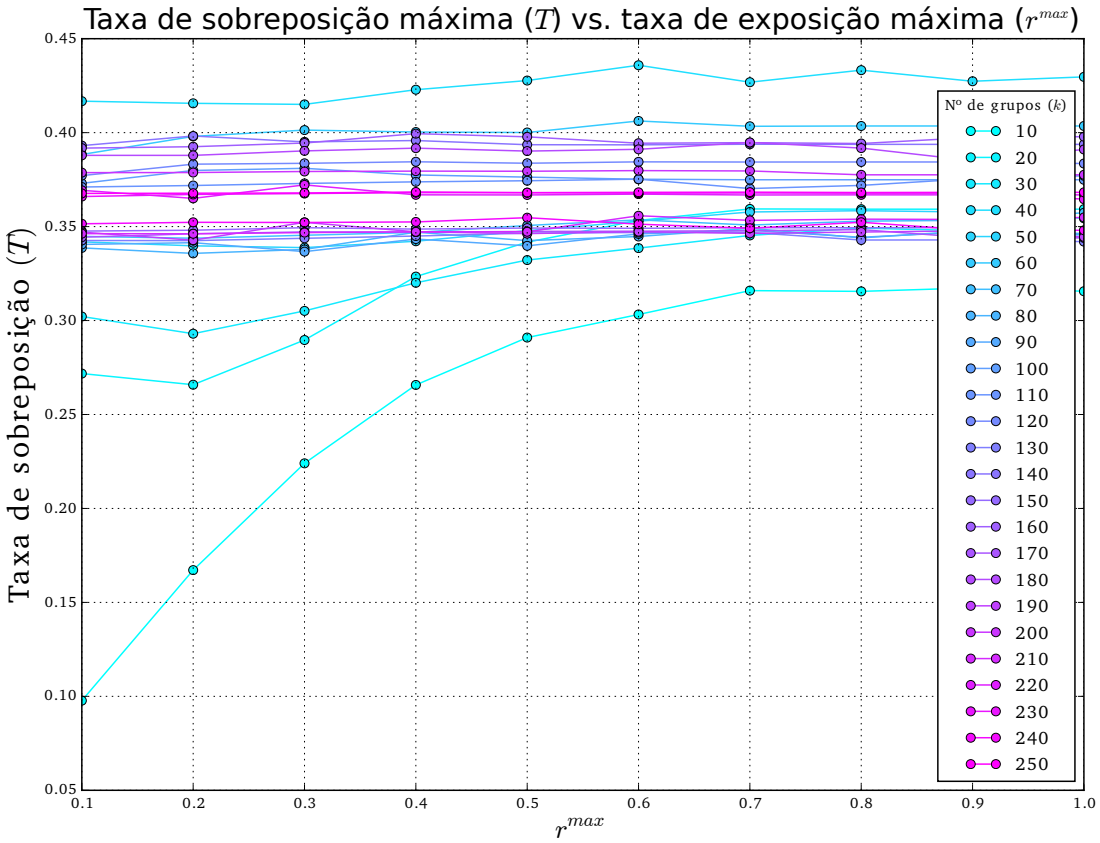
4.1.2 SIMULAÇÕES COM AGRUPAMENTOS DE MAIOR ÍNDICE DE DUNN

A tabela 4 exibe os resultados de agrupamento que tiveram os maiores índices de Dunn. Como observado na figura 13, o algoritmo de agrupamento hierárquico por ligação completa teve os resultados com os maiores índices de Dunn para todos os valores de k.

As figuras 19 a 22 apresentam resultados de mesma natureza daqueles apresentados na seção anterior, porém para os agrupamentos descritos na tabela 4. É possível perceber, na figura 19, maior variação nos valores de T, tendo maior concentração no intervalo $[0{,}34; 0{,}4]$. Também é possível observar, na figura 20, o mesmo comportamento do experimento anterior: quanto maior o valor de k, menor a capacidade do método proposto de controlar a taxa de sobreposição para diferentes valores de r^{max}.

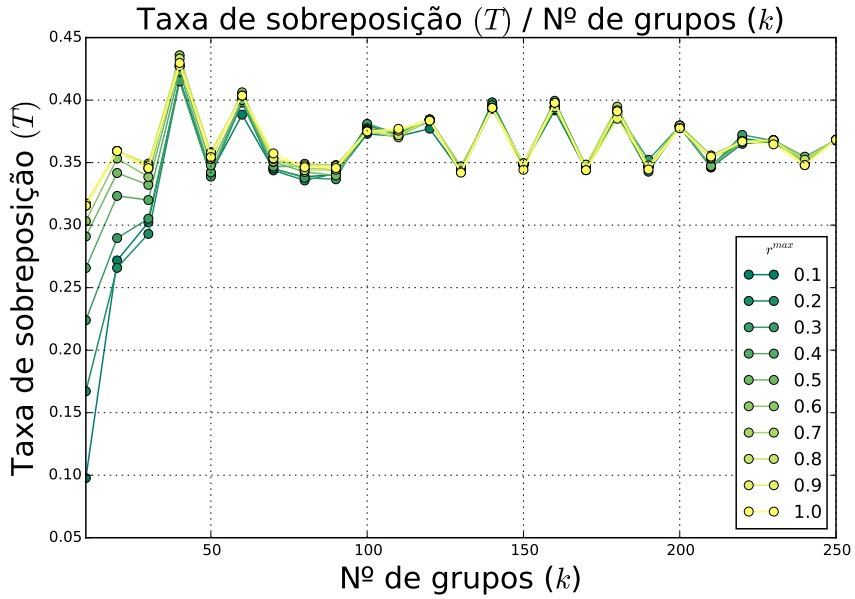
A figura 21 exibe maior variação no cálculo dos erros, indicando que agrupamentos cuja qualidade é julgada pelo índice de Dunn resultam em estimativas de proficiências menos previsíveis que agrupamentos validados pela minimização da soma das variâncias intra-grupos. Os valores da raiz dos erros quadráticos médios também tiveram o mesmo comportamento daqueles extraídos através de simulações que utilizaram os agrupamentos de menos variância.

Figura 19 – Mudança da taxa de sobreposição de acordo com o aumento de r^{max} para diferentes números de grupos (agrupamentos de maior índice de Dunn)



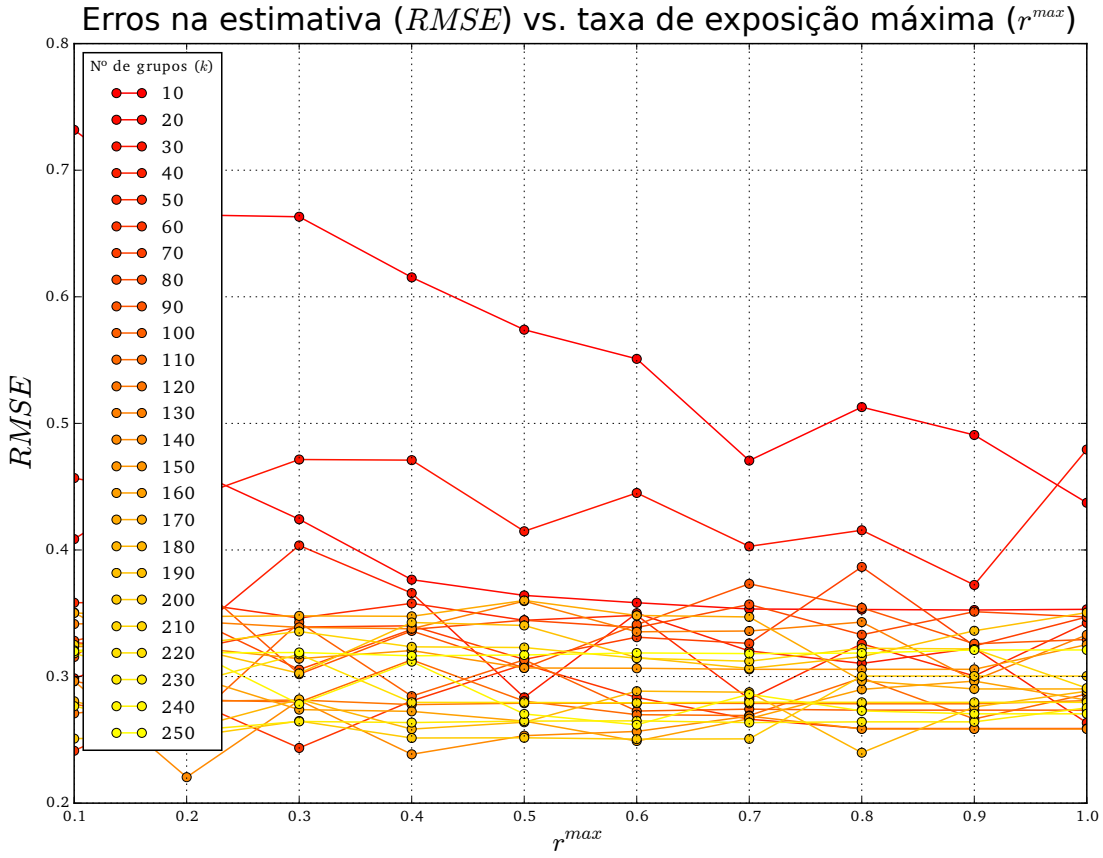
Fonte: Autor

Figura 20 – Relação entre a taxa de exposição e o número de grupos, para diferentes valores de r^{max} (agrupamentos de maior índice de Dunn)



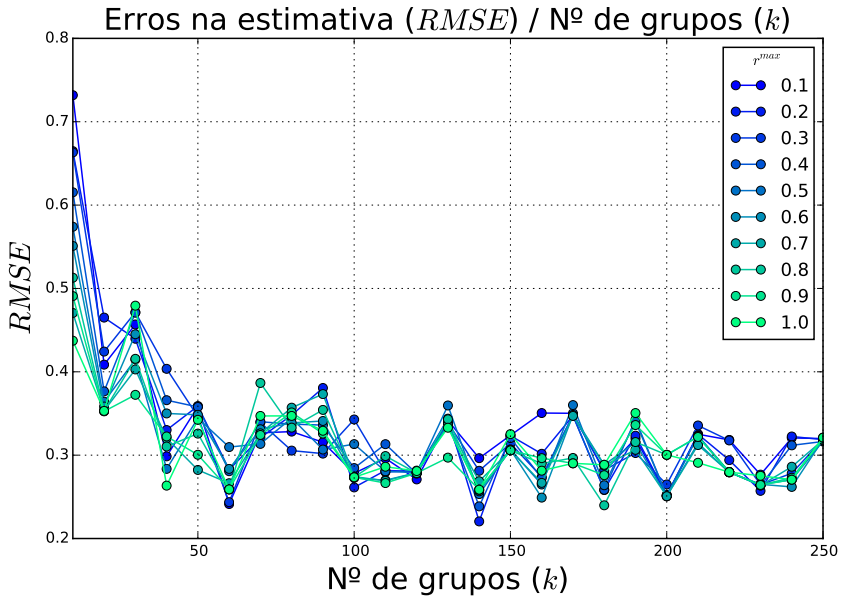
Fonte: Autor

Figura 21 – Mudança da raiz dos erros quadráticos médios de acordo com o aumento de r^{max} para diferentes números de grupos (agrupamentos de maior índice de Dunn)



Fonte: Autor

Figura 22 – Relação entre a raiz dos erros quadráticos médios e o número de grupos, para diferentes valores de r^{max} (agrupamentos de maior índice de Dunn)



Fonte: Autor

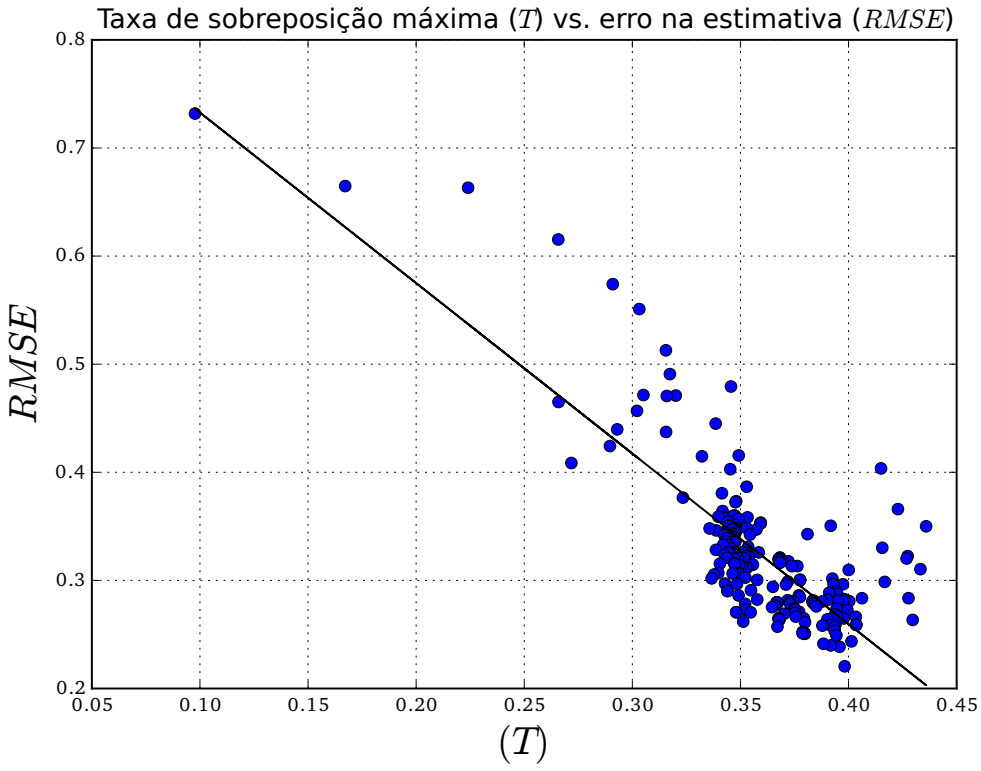
Tabela 4 – Resultados de agrupamento com maior índice de Dunn escolhidos para execução das simulações de TAI. Todos os agrupamentos desta tabela foram extraídos com o algoritmo hierárquico por ligação completa.

Nº grupos	**Distância**	**Variância**	**Dunn**
10	Euclideana	8814.309916	1.060165
20	Chebyshev	14329.480018	1.047157
30	Chebyshev	23324.866009	1.050349
40	Manhattan	28535.130887	1.043059
50	Manhattan	38381.291276	1.060434
60	Euclideana	43788.249202	1.025196
70	Chebyshev	51652.051637	1.028443
80	Chebyshev	57414.932003	1.055119
90	Manhattan	66018.774724	1.021112
100	Mahalanobis	71004.066112	1.028866
110	Mahalanobis	78963.96325	1.017863
120	Euclideana	85301.394856	1.020554
130	Chebyshev	92715.653673	1.009594
140	Chebyshev	97305.490881	1.047053
150	Euclideana	105429.765347	1.017068
160	Manhattan	110148.724788	1.020295
170	Chebyshev	118548.576964	1.016565
180	Chebyshev	123479.358254	1.018063
190	Chebyshev	130483.59005	1.025648
200	Chebyshev	136759.624044	1.00992
210	Euclideana	143542.34744	1.006011
220	Euclideana	149513.344961	1.024699
230	Euclideana	156374.939349	1.02089
240	Euclideana	162244.227039	1.00597
250	Mahalanobis	167704.732769	1.022984

Fonte: Autor

A figura, 23 exibe a relação entre a raiz dos erros quadráticos médios e a taxa de sobreposição do teste para todas as simulações. É possível perceber uma relação linear entre as variáveis da forma $RMSE = 0{,}890399647814 - 1.57702077289 \cdot T$, com coeficiente de correlação $r = -0{,}796061173087$ e coeficiente de determinação $r^2 = 0{,}633713391297$,

Figura 23 – Relação entre a taxa de sobreposição e a raiz dos erros quadráticos médios, para diferentes valores de r^{max} (agrupamentos de maior índice de Dunn)



Fonte: Autor

5 CONCLUSÕES

Este trabalho apresentou o CISM, uma metodologia de seleção de itens para aplicação de Testes Adaptativos Informatizados que se utiliza do agrupamento por similaridade de itens. Na forma descrita neste trabalho, o CISM independe do método de agrupamento por similaridade empregado, assim como de qualquer informação referente ao tema do teste, contanto que os itens nela empregados possuam seus parâmetros sob o modelo logístico de 3 parâmetros da Teoria da Resposta ao Item estimados.

Representados por tais parâmetros, o agrupamento dos itens foi realizado utilizando algoritmos de agrupamento tradicionais, como o *k-means*, o Ward *k-means* e algoritmos aglomerativos por ligação simples, média, por média ponderada e completa, sob diversas medidas de distância. Os resultados dos diversos agrupamentos foram avaliados através da soma da variância intra-grupos e do índice de Dunn, dois índices de validação de agrupamento. Foi constatado que as diferentes medidas de distância tiveram resultados semelhantes no processo de agrupamento, devido à alta densidade dos parâmetros dos itens.

Em seguida, os agrupamentos que obtiveram os melhores valores nos índices de validação foram utilizados pela metodologia proposta de seleção de itens, o CISM, em uma série de simulações de testes adaptativos, a fim de avaliar seu desempenho. Os agrupamentos que tiveram a menor variância intra-grupos foram decorrentes do uso do Ward *k-means*, uma variante do *k-means* cuja inicialização dos centroides é feita através do uso do algoritmo aglomerativo por função de Ward na realização da do particionamento inicial dos dados. Os resultados de agrupamento que obtiveram o maior índice de Dunn foram aqueles decorrentes do algoritmo aglomerativo por ligação completa, devido a sua tendência em separar grupos que tendem a se alongar.

Durante as simulações de testes adaptativos, onde os agrupamentos com maior índice de Dunn e menor soma de variâncias intra-grupos foram utilizados como entrada para o CISM, foram fixados diversos valores gradativamente maiores para a taxa de exposição máxima dos itens, com o objetivo de mensurar a precisão nas estimativas das proficiências dos indivíduos e a homogeneidade de uso do banco de itens sob diferentes restrições de exposição dos itens. Os resultados dos experimentos demonstraram que os agrupamentos de menor soma de variâncias intra-grupos resultaram em testes adaptativos com medidas mais previsíveis que os agrupamentos com maior índice de Dunn.

O controle da taxa de sobreposição, por sua vez, resultou no aumento do erro das estimativas das proficiências dos examinandos, sendo demonstrada a alta correlação negativa entre essas duas variáveis. Desta forma, foi observado que o número de grupos no qual o banco de itens é particionado pode ser utilizado como parâmetro para controle da taxa de sobreposição do teste, tendo como efeito colateral o aumento no erro da estimativa das proficiências. Quanto menor o número de partições feita na base, menor a taxa de sobreposição do teste e maior o erro na estimativa; aumentando-se o número de partições, a taxa de sobreposição aumenta, diminuindo-se o erro. Caso o número de partições alcance um valor muito alto, o CISM começa a se comportar como o método de escolha de itens por máxima informação, tendo precisão máxima nas estimativas, porém desconsiderando o controle das taxas de exposição dos itens e, consequentemente, da taxa de sobreposição do teste.

Em uma base de 500 itens, o CISM foi capaz de controlar a taxa de sobreposição do teste quando os itens são agrupados em até 50 grupos, ou seja, $\frac{1}{10}$ do tamanho total da base. Esta fração não pode ser considerada uma regra na escolha do número de grupos para bases com diferentes números de itens, porém espera-se que o número máximo de grupos passível de uso pelo CISM seja dependente do tamanho da base.

Como trabalhos futuros, aponta-se a possibilidade do uso do agrupamento por similaridade dos itens na criação de blocos de itens, os quais podem ser aplicados em conjunto aos examinandos, aumentado a precisão da estimativa de suas proficiências nas imediações da escala de proficiências que aquele grupo de itens representa. Também aponta-se a incorporação ao CISMde métodos de controle de exposição de itens disponíveis na literatura.

O estudo do uso do agrupamento por similaridade de itens representados por parâmetros de diferentes modelos da TRI (e.g. multidimensionais ou com menos parâmetros) também é uma área que merece ser estudada, assim como o uso desses agrupamentos na aplicação de Testes Adaptativos Informatizados, em conjunto com o CISM ou não.

REFERÊNCIAS

AFFENZELLER, M. et al. **Genetic Algorithms and Genetic Programming**: Modern concepts and practical applications. 1. ed. Chapman and Hall/CRC, 2009. Disponível em: <http://www.crcpress.com/product/isbn/9781584886297>.

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. d. C. **Teoria da Resposta ao Item**: Conceitos e aplicações. [S.l.]: AVALIA Educacional, 2000.

AQUINO JUNIOR, P. T. **PICaP**: padrões e personas para expressão da diversidade de usuários no projeto de interação. 2008. Tese (Doutorado) — Universidade de São Paulo, São Paulo.

BARBER, D. **Bayesian reasoning and machine learning**. Cambridge University Press, 2014. Disponível em: <http://www.cs.ucl.ac.uk/staff/d.barber/brml/>.

BARRADA, J. et al. Test overlap rate and item exposure rate as indicators of test security in cats. In: **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing**. [s.n.], 2009. Disponível em: <http://iacat.org/sites/default/files/biblio/cat09barrada.pdf>.

BARRADA, J. R.; ABAD, F. J.; OLEA, J. Optimal number of strata for the stratified methods in computerized adaptive testing. **The Spanish Journal of Psychology**, v. 17, 2014. Disponível em: <http://www.journals.cambridge.org/abstract_S113874161400050X>.

BARRADA, J. R.; ABAD, F. J.; VELDKAMP, B. P. MetodologÍa: Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. **Psicothema**, v. 21, n. Número 2, p. 313–320, 2009. Disponível em: <http://www.unioviedo.es/reunido/index.php/PST/article/view/8858>.

BARRADA, J. R.; MAZUELA, P.; OLEA, J. Maximum information stratification method for controlling item exposure in computerized adaptive testing. **Psicothema**, v. 18, n. 1, p. 156–159, 2006. Disponível em: <http://156.35.33.98/reunido/index.php/PST/article/view/8411>.

BARRADA, J. R.; OLEA, J.; ABAD, F. J. Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. **The Spanish journal of psychology**, v. 11, n. 2, p. 618–625, 2008. Disponível em: <http://revistas.ucm.es/index.php/SJOP/article/view/29891>.

BARRADA, J. R.; OLEA, J.; PONSODA, V. Methods for restricting maximum exposure rate in computerized adaptative testing. **Methodology: European Journal of Research Methods for the Behavioral and Social Sciences**, v. 3, n. 1, p. 14, 2007. Disponível em: <http://psycnet.apa.org/journals/med/3/1/14/>.

BARRADA, J. R. et al. Item selection rules in computerized adaptive testing: Accuracy and security. **Methodology: European Journal of Research Methods for the Behavioral and**

Social Sciences, v. 5, n. 1, p. 7–17, 2009.

____. A method for the comparison of item selection rules in computerized adaptive testing. **Applied Psychological Measurement**, v. 34, n. 6, p. 438–452, 2010. Disponível em: <http://apm.sagepub.com/cgi/doi/10.1177/0146621610370152>.

BAYLARI, A.; MONTAZER, G. Design a personalized e-learning system based on item response theory and artificial neural network approach. **Expert Systems with Applications**, v. 36, n. 4, p. 8013–8021, 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095741740800777X>.

BENITEZ ROCHEL, R.; TRELLA LOPEZ, M.; CONEJO MUÑOZ, R. Neural networks applied to item response theory. In: **Proceeding of the ICSC Symposia on Neural Computation**. [s.n.], 2000. p. 23–26. Disponível em: <http://www.lcc.uma.es:8080/repository/fileDownloader?rfname=LCC559.pdf>.

BINET, A.; SIMON, T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. **L'année psychologique**, v. 11, n. 1, p. 191–244, 1904. Disponível em: <http://www.persee.fr/web/revues/home/prescript/article/psy_0003-5033_1904_num_11_1_3675>.

BOCK, R. D.; MISLEVY, R. J. Adaptive eap estimation of ability in a microcomputer environment. **Applied Psychological Measurement**, v. 6, n. 4, p. 431–444, 1982. Disponível em: <http://apm.sagepub.com/content/6/4/431.short>.

CATTELL, R. B. The description of personality: basic traits resolved into clusters. **The Journal of Abnormal and Social Psychology**, v. 38, n. 4, p. 476–506, 1943.

CHAJEWSKI, M.; LEWIS, C. Optimizing item exposure control algorithms for polytomous computerized adaptive tests with restricted item banks. In: **Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing**. [s.n.], 2009. Disponível em: <http://www.psych.umn.edu/psylabs/catcentral/pdffiles/cat09chajewski.pdf>.

CHANG, H.-H.; QIAN, J.; YING, Z. a-stratified multistage computerized adaptive testing with b blocking. **Applied Psychological Measurement**, v. 25, n. 4, p. 333–341, 2001. Disponível em: <http://apm.sagepub.com/content/25/4/333.short>.

CHANG, H.-H.; VAN DER LINDEN, W. J. Optimal stratification of item pools in α-stratified computerized adaptive testing. **Applied Psychological Measurement**, v. 27, n. 4, p. 262–274, 2003. Disponível em: <http://apm.sagepub.com/content/27/4/262.shorthttp://www.utwente.nl/gw/omd/medewerkers/artikelen/APM2003,262-274.pdf>.

CHANG, H.-H.; YING, Z. A global information approach to computerized adaptive testing. **Applied Psychological Measurement**, v. 20, n. 3, p. 213–229, 1996. Disponível em: <http://apm.sagepub.com/content/20/3/213>.

CHANG, W.; YANG, H. Applying irt to estimate learning ability and k-means clustering in web based learning. **Journal of Software**, v. 4, n. 2, p. 167–174, 2009. Disponível em: <https://academypublisher.com/~academz3/ojs/index.php/jsw/article/view/0402167174>.

CHANG, W.-C. et al. Integrating irt to clustering student's ability with k-means. In: **Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on**. IEEE, 2009. p. 1045–1048. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5412633>.

CHEN, S.-Y.; DOONG, S.-H. Predicting item exposure parameters in computerized adaptive testing. **British Journal of Mathematical and Statistical Psychology**, v. 61, n. 1, p. 75–91, 2008. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1348/000711006X129553/full>.

CONEJO, R. et al. Modelado del alumno: un enfoque bayesiano. **Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial**, v. 5, n. 12, p. 50–58, 2001. Disponível em: <http://www.researchgate.net/publication/220071595_Modelado_del_alumno_un_enfoque_bayesiano/file/9fcfd50bc794edfe57.pdf>.

DAVIS, L. L. Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. **Applied Psychological Measurement**, v. 28, n. 3, p. 165–185, 2004. Disponível em: <http://apm.sagepub.com/content/28/3/165.short>.

DAVIS, L. L.; DODD, B. G. Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. **Journal of applied measurement**, v. 9, n. 1, p. 1, 2008. Disponível em: <http://www.pearsonemsolutions.com/downloads/research/PartialCreditModel_rr0501.pdf>.

DE AYALA, R. J. **The Theory and Practice of Item Response Theory**. New York: Guilford Press, 2009. 448 p.

DESMARAIS, M. C.; PU, X.; BLAIS, J. G. Partial order knowledge structures for cat applications. In: **Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.** [s.n.], 2007. Disponível em: <http://iacat.org/sites/default/files/biblio/cat07desmarais.pdf>.

DODD, B. G. The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. **Applied psychological measurement**, v. 14, n. 4, p. 355–366, 1990. Disponível em: <http://apm.sagepub.com/content/14/4/355.short>.

EL-ALFY, E.-S. M.; ABDEL-AAL, R. E. Construction and analysis of educational tests using abductive machine learning. **Computers & Education**, v. 51, n. 1, p. 1–16, 2008. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0360131507000218>.

EL-ALFY, E.-S. M.; JAFRI, S. S. A neural network approach for estimating examinees\textquoterightproficiency levels in computerized adaptive testing. In: **Proceedings of the sixth conference on IASTED International Confe-**

rence Web-Based Education. [s.n.], 2007. v. 2, p. 505–510. Disponível em: <http://www.actapress.com/PaperInfo.aspx?paperId=30006>.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Second International Conference on Knowledge Discovery and Data Mining**. [s.n.], 1996. v. 96, p. 226–231. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2930>.

EVERITT, B. **Cluster analysis**. Chichester, West Sussex, U.K.: Wiley, 2011.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. **Science**, v. 315, n. 5814, p. 972–976, 2007. Disponível em: <http://www.sciencemag.org/content/315/5814/972.short>.

GEORGIADOU, E. G.; TRIANTAFILLOU, E.; ECONOMIDES, A. A. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. **The Journal of Technology, Learning and Assessment**, v. 5, n. 8, 2007.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part i. **ACM Sigmod Record**, v. 31, n. 2, p. 40–45, 2002.

HAYKIN, S. S. **Neural Networks**: A comprehensive foundation. [S.l.]: Prentice Hall, 1999. 842 p.

IZENMAN, A. J. **Modern Multivariate Statistical Techniques**: Regression, classification, and manifold learning. 1. ed. New York, NY, USA: Springer-Verlag New York, 2008. v. 1. 733 p. Disponível em: <http://www.springer.com/us/book/9780387781884>.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, v. 31, n. 3, p. 264–323, 1999. Disponível em: <http://dl.acm.org/citation.cfm?id=331504>.

JONES, E.; OLIPHANT, T.; PETERSON, P. **SciPy: Open source scientific tools for Python**. 2001–. Disponível em: <http://www.scipy.org/>. Acesso em: 2015-06-25.

KASTRIN, A. Item response theory modeling for microarray gene expression data. **Applied Statistics Conference**, v. 6, n. 1, p. 51–67, 2009. Disponível em: <http://mrvar.fdv.uni-lj.si/pub/mz/mz6.1/kastrin.pdf>.

KIM, K. S.; CHOI, Y. S. Bayesian network approach to computerized adaptive testing. 2012.

Disponível em: <http://www.sersc.org/journals/IJSH/vol6_no3_2012/10.pdf>.

KOVÁCS, F.; LEGÁNY, C.; BABOS, A. Cluster validity measurement techniques. In: **6th international symposium of Hungarian researchers on computational intelligence**. [S.l.]: Citeseer, 2005.

KWEDLO, W. A clustering method combining differential evolution with the k-means algorithm. **Pattern Recognition Letters**, Elsevier B.V., v. 32, n. 12, p. 1613–1621, 2011. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2011.05.010>.

LEIGH, J. R. **Functional Analysis and Linear Control Theory**. [S.l.]: Dover Publications, 2007. 176 p.

LI, J.-W. et al. A self-adjusting e-course generation process for personalized learning. **Expert Systems with Applications**, v. 39, n. 3, p. 3223–3232, 2012. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417411013182>.

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982.

LORD, F. M. A broad-range tailored test of verbal ability. **Applied Psychological Measurement**, v. 1, n. 1, p. 95–100, 1977. Disponível em: <http://apm.sagepub.com/content/1/1/95.short>.

____. **Applications of Item Response Theory to Practical Testing Problems**. [S.l.]: Routledge, 1980.

LORD, F. M.; NOVICK, M. R.; BIRNBAUM, A. **Statistical theories of mental test scores**. Oxford, England: Addison-Wesley, 1968.

LOTITO, G.; PIRLO, G. Item response theory for optimal questionnaire design. **Journal of e-Learning and Knowledge Society**, v. 9, n. 3, 2013. Disponível em: <http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/820>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. Oakland, CA, USA.: [s.n.], 1967. v. 1, n. 14, p. 281—-297. Disponível em: <http://projecteuclid.org/euclid.bsmsp/1200512992>.

MAGIS, D.; RAÎCHE, G. Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. **Journal of Statistical Software**, v. 48, n. 8, p. 1–31, 2012. Disponível em: <http://www.jstatsoft.org/v48/i08/>.

MAHALANOBIS, P. C. On the generalized distance in statistics. **Proceedings of the National Institute of Sciences (Calcutta)**, v. 2, p. 49–55, 1936.

MASIERO, A. A. **Algoritmo de Agrupamento por Similaridade aplicado a Criação de Personas**. 2013. Tese (M. Sc. in Electrical Engineering) — Centro Universitário da FEI, São Bernardo do Campo.

MCBRIDE, J. R.; MARTIN, J. T. Reliability and validity of adaptive ability tests in a military setting. In: **New horizons in testing: Latent trait test theory and computerized adaptive testing**. [S.l.]: Academic Press New York, 1983. p. 224–236.

MEILA, M.; SHI, J. A random walks view of spectral segmentation. In: . [S.l.: s.n.], 2001.

MENEGHETTI, D. D. R. **catsim: Computerized Adaptive Testing methodology assisted by SIMilarity algorithm**. 2015. Disponível em: <http://douglasrizzo.github.io/catsim/>. Acesso em: 2015-07-01.

MILLÁN, E. et al. Using bayesian networks to improve knowledge assessment. **Computers & Education**, v. 60, n. 1, p. 436–447, jan. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360131512001649>.

MILLáN, E. et al. Using bayesian networks in computerized adaptive tests. In: **Computers and Education in the 21st Century**. Springer Netherlands, 2000. p. 217–228. Disponível em: <http://link.springer.com/content/pdf/10.1007/0-306-47532-4_20.pdf>.

MIRKIN, B. **Clustering**: A data recovery approach. 2. ed. [S.l.]: CRC Press, 2012. 374 p.

MITCHELL, T. M. **Machine learning**. Boston, MA: McGraw-Hill, 1997.

MOREIRA JUNIOR, F. J. Sistemática para a implantação de testes adaptativos informatizados baseados na Teoria da Resposta ao Item. 2012. Disponível em: <http://repositorio.ufsc.br/handle/123456789/95506>.

MOUCHET, M.; GUILHAUMON, F.; MASON, N. Towards a consensus for calculating dendrogram‐based functional diversity indices. **Oikos**, v. 117, p. 794–800, 2008. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.0030-1299.2008.16594.x/full>.

NEVO, D. Evaluation in education. In: SHAW, I. F.; GREENE, J. C.; MARK, M. M. (Ed.). **Handbook of Evaluation: Policies, Programs and Practices**. [S.l.]: SAGE Publications, 2006. p. 441–460.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**. [S.l.]: MIT Press, 2001. p. 849–856.

PASQUALI, L. **Psicometria**: Teoria dos testes na psicologia e educação. 5. ed. [S.l.]: Vozes, 2003.

PHUVIPADAWAT, S. et al. A comparability approach to item reduction in computerized adaptive testing. In: **4th IEEE International Conference on Management of Innovation and Technology**. [S.l.: s.n.], 2008. p. 1456–1460.

POMMERICH, M. **The nine lives of CAT-ASVAB: Innovations and revelations**. 2009. Disponível em: <http://publicdocs.iacat.org/cat2010/cat09pommerich.pdf>. Acesso em: 2015-04-26.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, v. 1, n. 1, p. 81–106, 1986. Disponível em: <http://link.springer.com/article/10.1023/A:1022643204877>.

RASCH, G. An item analysis which takes individual differences into account. **British Journal of Mathematical and Statistical Psychology**, v. 19, n. 1, p. 49–57, 1966. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8317.1966.tb00354.x/abstract>.

____. **Probabilistic Models for Some Intelligence and Attainment Tests**. [S.l.]: University of Chicago Press, 1980. v. 1. 199 p.

REVUELTA, J.; PONSODA, V. A comparison of item exposure control methods in computerized adaptive testing. **Journal of Educational Measurement**, v. 35, n. 4, p. 311–327, 1998. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1998.tb00541.x/abstract>.

ROBLES PEDROZO, L. S.; RODRIGUEZ-ARTACHO, M. A cluster-based analisys to diagnose students' learning achievements. In: **IEEE Global Engineering Education Conference, EDUCON**. IEEE, 2013. p. 1118–1123. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6530248>.

SCRIVEN, M. S. The methodology of evaluation. **Perspectives of Curriculum Evaluation**, Chicago, n. 1, 1967.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 22, n. 8, p. 888–905, 2000. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=868688>.

SPEARMAN, C. J. L. W. S. C. **Human ability**: A continuation of "the abilities of man". [S.l.]: London: Macmillan, 1950.

SUKAMOLSON, S. Computerized test/item banking and computerized adaptive testing for teachers of lecturers. **Information Technology and Universities in Asia – ITUA**, 2002. Disponível em: <http://www.stc.arts.chula.ac.th/ITUA/Papers_for_ITUA_Proceedings/Suphat2.pdf>.

SYMPSON, J. B.; HETTER, R. D. Controlling item-exposure rates in computerized adaptive testing. In: . [S.l.: s.n.], 1985. p. 973–977.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction To Data Mining**. 1. ed. Boston, MA: Addison-Wesley, 2005. 568 p. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract>.

TEZZA, R. Proposta de um construto para medir usabilidade em sites de e-commerce utilizando a teoria da resposta ao item. 2009. Disponível em: <http://repositorio.ufsc.br/handle/123456789/92424>.

TEZZA, R. **Modelagem multidimensional para mensurar qualidade em website de e-commerce utilizando a teoria da resposta ao item**. 2012. Tese (D. Sc. in Production Engineering) — Universidade Federal de Santa Catarina. Disponível em: <https://repositorio.ufsc.br/handle/123456789/103430>.

TEZZA, R.; BORNIA, A. C.; MOREIRA JUNIOR, F. d. J. Avaliação de usabilidade em sites de e-commerce: Uma aplicação da teoria da resposta ao item. In: . Curitiba, PR: [s.n.], 2009. Disponível em: <http://www.custosemedidas.ufsc.br/70.pdf>.

THOMSON, W. **Popular lectures and addresses**. Macmillan London, 1889. 490 p. Disponível em: <https://archive.org/details/popularlecturesa01kelvuoft>.

TYLER, R. W. **Basic principles of curriculum and instruction**. [S.l.]: University of Chicago press, 2013.

UENO, M.; SONGMUANG, P. Computerized adaptive testing based on decision tree. In: **2010 IEEE 10th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.: s.n.], 2010. p. 191–193.

VAN DER LINDEN, W. J. Some alternatives to sympson-hetter item-exposure control in computerized adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 28, n. 3, p. 249–265, 2003. Disponível em: <http://jeb.sagepub.com/content/28/3/249.short>.

VAN DER LINDEN, W. J.; GLAS, C. a. W. **Computerized Adaptive Testing**: Theory and practice. Dordrecht; Boston: Kluwer Academic, 2000. 323 p. Disponível em: <http://link.springer.com/book/10.1007\%2F0-306-47531-6>.

VEERKAMP, W. J. J.; BERGER, M. P. F. Some new item selection criteria for adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 22, n. 2, p. 203–226, 1997. Disponível em: <http://jeb.sagepub.com/content/22/2/203.short>.

VOMLEL, J. Building adaptive tests using bayesian networks. **Kybernetika**, v. 40, n. 3, p. 333–348, 2004. Disponível em: <http://dml.cz/handle/10338.dmlcz/135599>.

VON LUXBURG, U. A tutorial on spectral clustering. **Statistics and computing**, v. 17, n. 4, p. 395–416, 2007. Disponível em: <http://link.springer.com/article/10.1007/s11222-007-9033-z>.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3. ed. Burlington, MA: Morgan Kaufmann, 2011.

YU, J. Neural networks ensemble-based irt parameter estimation. In: **Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on**. [s.n.], 2009. p. 1–3. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5365773>.

ZUBIN, J. A technique for measuring like-mindedness. **The Journal of Abnormal and Social Psychology**, v. 33, n. 4, p. 508, 1938.

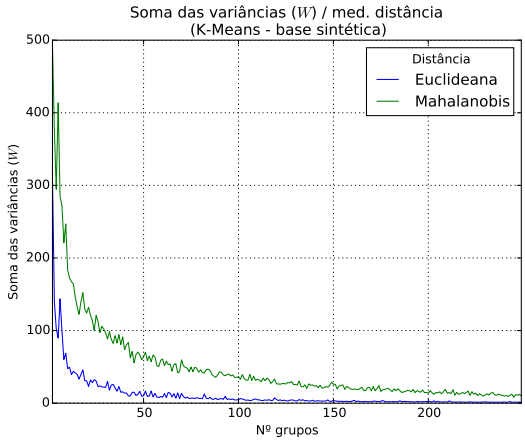
APÊNDICE A – MEDIDAS DE VALIDAÇÃO DE AGRUPAMENTO POR ALGORITMO E MEDIDA DE DISTÂNCIA

Neste apêndice, são exibidos os valores da soma das variâncias intra-grupos e do índice de Dunn para os diferentes algoritmos de agrupamento aplicados no trabalho, separados em diferentes medidas de similaridade.

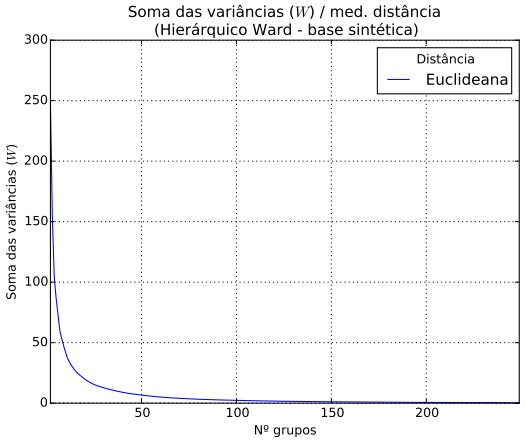
A figura 24 exibe as somas das variâncias intra-grupos (W) para os diferentes algoritmos empregados no experimento de agrupamento, separados pelas diferentes medidas de distância utilizada na base de parâmetros em questão. Como os resultados para a base normalizada foram muito semelhantes aos resultados na base original, os resultados para a base normalizada foram omitidos. É possível perceber a queda nos valores de W conforme o aumento do número de grupos. Também é possível observar maior variância em todos os algoritmos quando a distância de Mahalanobis é utilizada, devido à sua propensão em gerar grupos elipsoidais. Não houve combinação de medida de distância e número de grupos que resultasse em queda significativa da soma das variâncias. É possível perceber, no entanto, desempenho superior do algoritmo aglomerativo por função de Ward e do Ward *k-means*, devido à tendência dos algoritmos que utilizam a função de Ward em minimizar a variância intra-grupos.

a figura 25 exibe os valores do índice de Dunn para os agrupamentos realizados com os diferentes algoritmos, utilizando diferentes medidas de distância. Novamente, a medida de distância empregada teve pouco efeito na busca por valores maiores do índice de Dunn. Os algoritmos, por sua vez, demonstraram resultados diferentes entre si: o *k-means* e o algoritmo aglomerativo por ligação simples tiveram os resultados mais baixos do índice. O *k-means* tem como objetivo a minimização das variâncias intra-grupo, porém um valor alto do índice de Dunn depende adicionalmente da separação entre os grupos, algo que o *k-means* ignora em seu processo de convergência. Já o algoritmo aglomerativo por ligação simples tende a criar grupos "alongados" (JAIN; DUBES, 1988), aumentando o diâmetro dos grupos e, consequentemente, diminuindo o índice de Dunn. O algoritmo que teve o melhor resultado foi o algoritmo aglomerativo por ligação completa, devido à sua tendência em não criar grupos alongados, diminuindo assim o diâmetro dos grupos e alcançando maiores valores no índice de Dunn. Os demais algoritmos não possuem como característica principal a criação de grupos compactos ou separados entre si, portanto tiveram valores intermediários no índice.

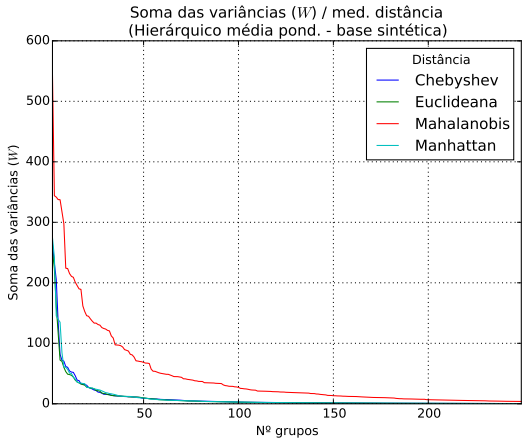
Figura 24 – Soma das variâncias intra-grupos dos diferentes algoritmos aplicados, utilizando diferentes medidas de distância na base estudada



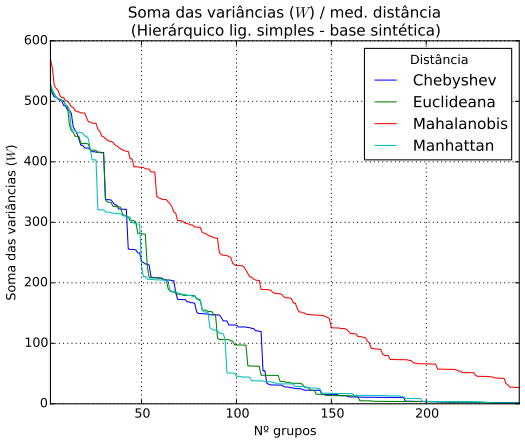
(a)



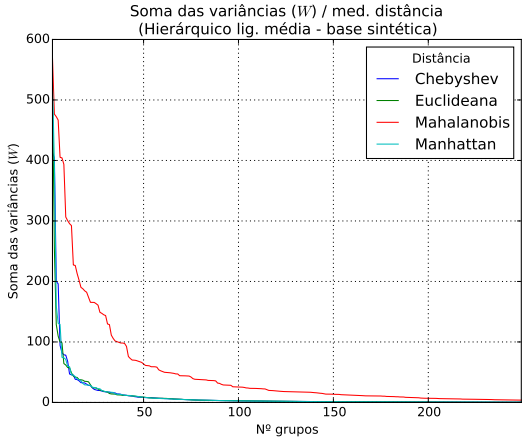
(b)



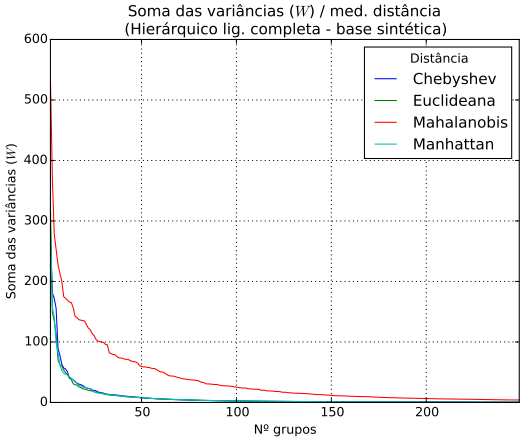
(c)



(d)



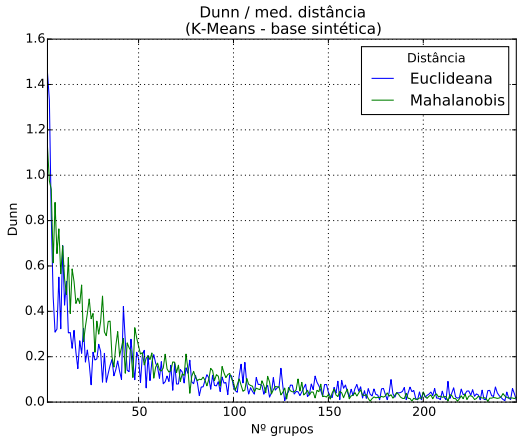
(e)



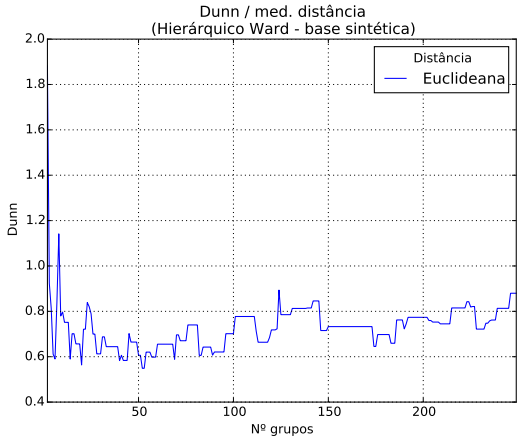
(f)

Fonte: Autor

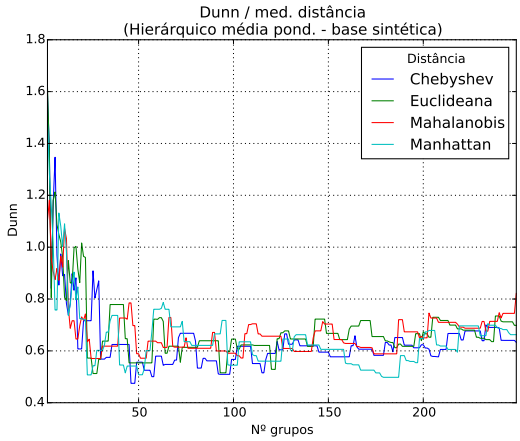
Figura 25 – Índice de Dunn dos diferentes algoritmos aplicados, utilizando diferentes medidas de distância na base estudada



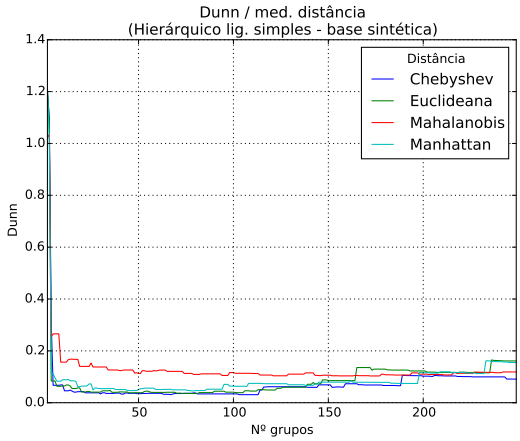
(a)



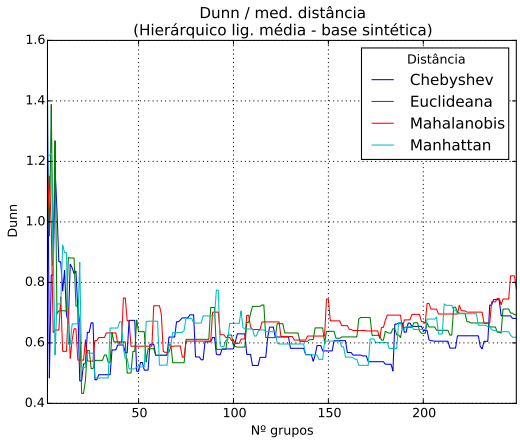
(b)



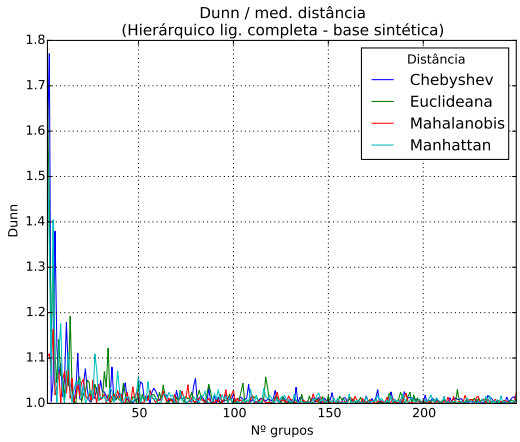
(c)



(d)



(e)



(f)

Fonte: Autor

ÍNDICE