Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Um novo método de avaliação de modelos de agrupamento baseado em Teoria de Resposta ao Item e concordância de modelos

Manuel Ferreira Júnior

Junho, 2023

Manuel Ferreira Júnior

Um novo método de avaliação de modelos de agrupamento baseado em Teoria de Resposta ao Item e concordância de modelos

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba como requisito fundamental para obtenção do Grau de Bacharel em Estatística.

Orientador: Prof. Dr. Marcelo Rodrigo Portela Ferreira

João Pessoa
Junho, 2023

Catalogação na publicação
Seção de Catalogação e Classificação

F383n Ferreira-Junior, Manuel.
Um novo método de avaliação de modelos de agrupamento baseado em Teoria de Resposta ao Item e concordância de modelos / Manuel Ferreira-Junior. - João Pessoa, 2023.
47 p. : il.

Orientação: Marcelo Rodrigo Portela Ferreira.
Coorientação: Telmo de Menezes e Silva Filho, Eufrásio de Andrade Lima Neto.
TCC (Graduação) - UFPB/CCEN.

1. Agrupamento. 2. Teoria de Resposta ao Item. 3. Métricas. I. Rodrigo Portela Ferreira, Marcelo. II. de Menezes e Silva Filho, Telmo. III. de Andrade Lima Neto, Eufrásio. IV. Título.

UFPB/CCEN CDU 311(043.2)

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

"Um Novo Método de Avaliação de Modelos de Agrupamento Baseado em Teoria de Resposta ao Item e Concordância de Modelos"

Manuel Ferreira Júnior

Ao sétimo dia do mês de Junho de 2023 às 09h00, de modo presencial, no Laboratório Joab Lima do Departamento de Estatística, realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso do(a) discente Manuel Ferreira Júnior, matrícula 20180008601, com a Banca Examinadora composta pelos professores: Dr. Marcelo Rodrigo Portela Ferreira, Presidente/Orientador (Departamento de Estatística - UFPB), Dra. Ana Hermínia Andrade e Silva, Examinadora (Departamento de Estatística - UFPB), Dr. Ricardo Bastos Cavalcante Prudêncio, Examinador (CIn - UFPE) e Dr. Hemílio Fernandes Campos Coelho, Examinador Suplente (Departamento de Estatística - UFPB). Iniciando-se os trabalhos, o presidente da Banca Examinadora cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que se fizesse, oralmente, a exposição do Trabalho de Conclusão de Curso intitulado **"Um Novo Método de Avaliação de Modelos de Agrupamento Baseado em Teoria de Resposta ao Item e Concordância de Modelos"**. Concluída a apresentação, a Banca Examinadora iniciou à arguição do(a) candidato(a). Encerrados os trabalhos de arguição os examinadores reuniram-se para avaliação e deram o parecer final sobre a apresentação e defesa oral do(a) candidato(a), tendo sido atribuída à sua apresentação a nota 10,0 (DEZ PONTOS), na disciplina de TCC II, resultante da média aritmética das notas atribuídas pelos membros da Banca Examinadora. A aprovação do(a) discente está condicionada a entrega da versão final do Trabalho de Conclusão de Curso com a inserção da ficha catalográfica e, as alterações sugeridas pelos examinadores, à Coordenação do Curso de Bacharelado em Estatística no prazo de 22 de Junho de 2023.

gov.br Documento assinado digitalmente
MARCELO RODRIGO PORTELA FERREIRA
Data: 17/06/2023 13:31:40-0300
Verifique em https://validar.iti.gov.br

Dr. Marcelo Rodrigo Portela Ferreira
(Professor Orientador)

gov.br Documento assinado digitalmente
RICARDO BASTOS CAVALCANTE PRUDENC
Data: 19/06/2023 11:22:23-0300
Verifique em https://validar.iti.gov.br

Dr. Ricardo Bastos Cavalcante Prudêncio
(Professor Examinador)

gov.br Documento assinado digitalmente
ANA HERMINIA ANDRADE E SILVA
Data: 19/06/2023 10:36:27-0300
Verifique em https://validar.iti.gov.br

Dra. Ana Hermínia Andrade e Silva
(Professora Examinadora)

gov.br Documento assinado digitalmente
MANUEL FERREIRA JUNIOR
Data: 19/06/2023 11:54:37-0300
Verifique em https://validar.iti.gov.br

Manuel Ferreira Júnior
(Discente)

João Pessoa, 07 de Junho de 2023.

Este trabalho é dedicado a ...

A Deus, por tudo. Aos meus pais, Manuel e Rosilene. As minhas irmãs, Roberta e Renata, que sempre me auxiliaram. Aos demais familiares e amigos, que direta ou indiretamente contribuíram para o meu progresso até aqui. Com amor, vos dedico.

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado a sabedoria e discernimento para conseguir chegar até aqui e concluir esta etapa da minha vida. Grato principalmente por aquele momento onde estava prestando vestibular, com a minha performace não sendo tão significante, não conseguindo entrar no curso que tinha interesse (Qualquer um da área de humanas, visto que não me considerava da exatas), o Senhor me guiou para esse mundo, ao qual eu fui muito bem recebido e almejo grandes coisas. Obrigado, Senhor.

Agradeço aos meus pais, em especial a minha mãe, por mais que nossa relação não seja das melhores, sou grato por tudo que a senhora deixou de fazer por minha causa, ou deixou de comprar por minha causa. Agradeço ao meu pai, exemplo de homem em sua paciência, me ensinou a ter calma e respeitar a todos. Me considero um ponto médio exato, entre ambos. Obrigado por tudo.

As minhas irmãs, Renata e Roberta, sou imensamente grato por tudo, a minha vontade de estudar, de aprender, de seguir em frente e pensar sempre alto, sempre foi alimentado também por vocês, que em muitas vezes me incentivaram de forma direta, e muitas vezes indiretamente, sem eu mesmo saber. Obrigado pelos inúmeros livros da turma da mônica, e por me apresentarem Harry Potter em especial. Tenho a minha vontade de aprender, graças a vocês, obrigado.

Aos meus avós, Narciso (*in memorian*) e Iraci, obrigado por todo o amor que ambos me deram, muitos encontros na casa da senhora vó, eu não pude ir, mas tudo culminou neste momento. Obrigado por tudo.

A minha tia Aparecida, obrigado sempre por cuidar e zelar sempre por mim, quando principalmente meus pais não podiam. Obrigado por ter estado sempre presente.

Em especial, agradeço ao meu primo Americo, agora teremos mais um estatístico na família (rsrsrs). Obrigado pelo apoio quando mais novo.

Agradeço imensamente aos meus amigos Yann e Igor, sempre se ajudando e se segurando para nenhum cair, sempre crescemos juntos e assim será, obrigado por tudo. Em

especial, agradeço a Evellyn, e seus pais, Edly e Wallac, durante muitos momentos da minha vida, vocês me incluiram como parte da "equipe"e sempre deram ótimos conselhos, obrigado.

Agradeço a cumplicidade e parceria da minha namorada, Hayse, a qual nunca desistiu de mim nesse periodo, o qual passamos por muitas coisas, muito obrigado, minha parceira.

Sem citar nomes para não ser injusto, gostaria de agradecer a todos os meus familiares, mesmo os que não encontraram-se presentes durante esse período, obrigado a todos.

Em especial, gostaria de agradecer ao professor Rodrigo, por ter me recusado no meu primeiro pibic (rsrs) mas por um bom motivo, sem isso, eu nunca teria me motivado a ser quem eu sou hoje. Obrigado professor.

Ao professor Telmo, muito obrigado por ter me aceitado como seu primeiro PIBIC e ultimo pela UFPB, graças a isso, eu tive oportunidades que jamais imaginei ter, e pude sonhar sonhos que jamais nem sequer imaginei, sempre tive esse mundo tão distante, hoje posso dizer que estou conquistando um sonho a cada passo. Muito obrigado por tudo! e minha sincera gratidão pela parceria de anos.

Além disso, agradeço imensamente aos professores Marcelo, meu orientador, que abraçou essa jornada comigo, agradecer também aos professores Eufrásio e Ricardo (UFPE), vocês são grandes exemplos a se inspirar na área, muito obrigado pelos conhecimentos passados.

Aos professores do departamento, um agradecimento especial a professora Ana Flavia, a qual me ajudou muito quando descobri um Câncer no meu pai, e também descobri inumeros problemas de saúde. Muito obrigado.

A todos os outros professores do departamento de estatística, a vocês, minha sincera gratidão, garanto que todo conhecimento adquirido a partir de vocês, foi e será muito bem aproveitado. Muito obrigado.

Aos meus amigos veteranos que adquiri ao longo desta graduação, Mateus, Kelfanio, Ullisses, Marina e Milleny, obrigado pelas experiências juntos ao longo do curso e dos aprendizados em conjunto. Em especial a turma de veteranos, agradecer ao Natan, por sempre ser um grande amigo e sempre estar disposto a dividir conhecimentos e sempre aprender em conjunto, obrigado irmão. Além disso, agradeço a minha segunda turma de veteranos, Wanusa, Leticia e Nicholas, obrigado também por tudo.

Por fim, mas não menos importante, agradecer aos amigos que adquiri ao longo dessa graduação, Beatriz, Marcos, Borges, Amanda. Lutamos por muito tempo juntos e assim continuaremos nessa caminhada, obrigado por todos os momentos.

Aos professores **Ana Herminia Andrade e Silva** e **Ricardo Bastos Cavalcante Prudêncio**, pela disposição em participar da banca avaliadora e por todas as sugestões e incentivos a esse trabalho.

Ao CNPQ por todo apoio financeiro durante a graduação.

"A esperança é o sonho do homem acordado".

— Aristóteles

"Não fostes vós que me escolhestes; fui eu que vos escolhi ".

— Bíblia Sagrada, Jó 15, 16

"Tentou. Falhou. Não importa. Tente de novo. Falhe de novo. Falhe melhor... Você erra 100% dos tiros que não dá ".

— Samuel Beckett e Wayne Gretzky

"Todas as vitórias ocultam uma abdicação.".

— Simone de Beauvoir

Resumo

A tarefa de avaliação de modelos de agrupamento é um problema complexo. Existem duas grandes divisões para mensuração do bom ajuste de um modelo baseado em agrupamento, sendo elas as medidas externas, como Acurácia ou Índice de Rand, não sendo aplicáveis, em problemas reais, para cenários em que não conhecemos o rótulo verdadeiro, ou seja, do tipo não supervisionados, e medidas internas, como o método de Calinski-Harabasz ou a medida de Silhueta, podendo ser aplicado para cenários não supervisionados, porém necessita da escolha de distâncias para serem calculados, ou seja, tornam-se significantes apenas para modelos que utilizaram-se das mesmas distâncias. Além disso, essas medidas de validação interna falham quando todas as instâncias do conjunto são agrupadas em um único grupo. Então, ambas as formas de validação de agrupamento não permitem a comparação direta entre modelos que possuem uma forma de estimação diferentes, por exemplo métodos *hard*, baseados em distâncias e métodos probabilísticos. O objetivo deste trabalho é propor um novo método de avaliação global para modelos de agrupamento, assumindo que bons modelos devem agrupar pares de instâncias em um mesmo grupo ou não. Para este fim, foi utilizado Teoria de Resposta ao Item para a estimação da habilidade do modelo e a dificuldade da instância, baseando-se na matriz de resposta obtida mensurando a concordância entre modelos. Experimentos foram conduzidos em conjuntos de dados simulados, clássicos da biblioteca do *scikit-learn* do *Python* para problemas de agrupamento, com diferentes números de grupos, diferentes graus de sobreposição dos grupos e distintos ruídos. Os resultados exibem uma forte correlação entre o valor da habilidade estimada pelo modelo de Teoria de Resposta ao Item e as medidas de validação externa, conseguindo recuperar corretamente o ranking dos melhores modelos para os conjuntos de dados utilizados.

Palavras-chave: Agrupamento; Teoria de Resposta ao Item; Métricas de avaliação.

Abstract

The task of clustering model evaluation is a complex problem. There are two major divisions for measuring the goodness of fit of a clustering model: external measures such as Accuracy or Rand Index, which are not applicable in real-world scenarios where the true labels are unknown, i.e., unsupervised scenarios; and internal measures such as the Calinski-Harabasz method or Silhouette score, which can be applied to unsupervised scenarios but require the choice of distance metrics to be calculated, making them meaningful only for models that used the same distances. Furthermore, these internal validation measures fail when all instances in the dataset are grouped into a single cluster. Hence, both forms of clustering validation do not allow for direct comparison between models that have different estimation approaches, such as hard methods based on distances and probabilistic methods. The objective of this work is to propose a new global evaluation method for clustering models, assuming that good models should group pairs of instances together or not. To this end, Item Response Theory (IRT) was used to estimate the model's ability and instance difficulty, based on the response matrix obtained by measuring agreement between models. Experiments were conducted on simulated datasets and classic datasets from the scikit-learn library in Python for clustering problems, with different numbers of clusters, varying degrees of overlap between clusters, and different levels of noise. The results show a strong correlation between the estimated ability value by the IRT model and external validation measures, successfully recovering the ranking of the best models for the utilized datasets.

Keywords: Clustering; Item Response Theory; Evaluation metrics.

Conteúdo

CAPÍTULO 1

Introdução

1.1 Introdução

Métodos de agrupamento são técnicas fundamentais na área de mineração de dados que visam identificar estruturas e padrões intrínsecos em conjuntos de dados não rotulados. O objetivo principal dos métodos de agrupamento é dividir os dados em grupos, de forma que os objetos dentro de um mesmo grupo sejam mais similares entre si do que com os objetos de outros grupos. Esses métodos são amplamente aplicados em diversas áreas, como análise de dados e reconhecimento de padrões, por exemplo. Além da seleção do método de agrupamento adequado, é importante avaliar a qualidade dos resultados obtidos.

Medidas de validação para métodos de agrupamento são utilizadas para mensurar o grau de ajuste de um modelo, com o intuito de encontrar o algoritmo que fornece a melhor partição para um particular conjunto de dados. Tais medidas de validação para algoritmos baseados em agrupamento podem ser divididas em duas categorias: validação interna e externa (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001a; HENNIG et al., 2015; DataNovia, Acessado em 2023; HENNIG, Acessado em 2023). Medidas de validação externa são utilizadas para comparar os grupos encontrados por um algoritmo quando se conhece o verdadeiro agrupamento para aquele conjunto de dados, assim como em tarefas supervisionadas (STREHL; GHOSH, 2002; ROSENBERG; HIRSCHBERG, 2007; HUBERT; ARABIE, 1985). Apesar de serem muito utilizadas em experimentos em que as classes são conhecidas, as medidas externas não podem ser utilizadas quando não se possui essas informações a priori sobre os dados, como vemos em casos práticos. Medidas de valida-

ção interna, pelo contrário, são utilizadas para estimar a qualidade de um agrupamento, baseando-se em medidas como a separação dos grupos retornados pelo algoritmo utilizado e a coesão das instâncias (linhas de um conjunto de dados) em cada grupo (CALINSKI; HARABASZ, 1974; ROUSSEEUW, 1987; DAVIES; BOULDIN, 1979). Essas medidas de validação não necessitam de nenhum pressuposto sobre a verdadeira informação acerca dos dados, contudo apresentam limitações, em especial com respeito à alta sensibilidade na escolha de uma distância única para um determinado algoritmo de agrupamento. Além disso, essa distância é comumente a mesma usada pelo algoritmo, dificultando a comparação de soluções produzidas por algoritmos distintos.

Nesse trabalho, é proposto um novo método de avaliação de algoritmos de agrupamento, CLAIRE (CLuster Agreement-based Item REsponses), baseado na estimação da habilidade de um método de agrupamento via modelo de Teoria de Resposta ao Item (TRI) (EMBRETSON; REISE, 2000). A TRI é largamente utilizada em áreas como Psicometria, com o intuito de estimar a habilidade latente de respondentes, considerando items de diferentes dificuldades. Um exemplo clássico de aplicação de TRI no Brasil é o Exame Nacional do Ensino Médio (ENEM) (Ministério da Educação (MEC), 21 de Março, 2023.). Altas habilidades, implicam em uma alta probabilidade de acerto para itens de diferentes tipos de dificuldades. Na TRI existem Curvas Características do Item, modelando a probabilidade da resposta para cada item, utilizando-se da habilidade e dificuldade como parâmetros. Todos os parâmetros obtidos são utilizados com o intuito para maximizar a máxima verossimilhança das respostas observadas em um teste. Recentemente, a TRI foi empregada para sistemas de Inteligência Artificial (MARTÍNEZ-PLUMED et al., 2016; MARTÍNEZ-PLUMED et al., 2019; MORAES et al., 2022; CHEN et al., 2019). Em avaliações tradicionais de métodos de agrupamento, é comumente usado medidas agregadas de performance, como por exemplo a contagem de acertos em provas ou a acurácia em aprendizagem de máquina, mas ambas as formas desconsideram a dificuldade latente da instância e a discriminação da mesma.

Ao considerar a dificuldade das instâncias, a TRI permite avaliação mais precisa das habilidades latentes. Dessa forma, ela produz medidas mais refinadas, levando em consideração, por exemplo, instâncias em região de fronteiras entre grupos, ou seja, instâncias que são problemáticas de serem agrupadas. Assim, a abordagem de TRI para a avaliação de métodos de agrupamento torna-se um meio mais robusto para estimação da habilidade latente de um modelo. De forma direta, não é possível a utilização de TRI para avaliação de métodos de agrupamento, uma vez que não se possuí a definição padrão de resposta correta para um determinado item.

Como solução proposta neste trabalho, TRI é utilizado para modelar a probabilidade de concordância entre pares de modelos de agrupamento para uma determinada instância

de um conjunto de dados. Passando para a nomenclatura adotada na TRI, definimos que cada respondente é um modelo de agrupamento, assim como cada item será uma instância em um conjunto de dados passado para o modelo. Cada resposta associada a um único modelo e a cada instância representa o quanto aquele modelo concorda com os outros modelos, ao agrupar essa instância em um mesmo grupo ou em grupos distintos das outras instâncias do conjunto. Considerando uma matriz de resposta criada pelas respostas obtidas a partir de modelos de agrupamento para um conjunto de dados, o modelo proposto estima duas informações distintas:

1 A habilidade de cada método de agrupamento;

2 A dificuldade de cada instância do conjunto de dados.

Para a interpretação dessa matriz, temos que altas habilidades serão obtidas na estimação para os modelos de agrupamento que concordam em como agrupar instâncias de níveis de dificuldades maiores. Dessa forma também, instâncias são consideradas difíceis quando apenas bons modelos para o determinado conjunto de dados concordam em como agrupar aquela instância, ou quando nenhum modelo concorda. Logo, bons modelos serão aqueles que apresentam um alto grau de consenso ao agrupar instâncias de alta complexidade adequadamente.

Para os experimentos realizados ao longo deste trabalho, foram considerados 35 diferentes modelos de agrupamento distintos, sendo variações de parâmetros advindos de diferentes grupos de algoritmos. No desenvolvimento do CLAIRE foi utilizado o modelo β^4-IRT (FERREIRA-JUNIOR et al., 2023) para a estimação das habilidades e dificuldades latentes dos modelos de agrupamento e das instâncias do conjunto de dados, respectivamente. Os experimentos foram conduzidos usando cinco conjuntos de dados característicos, variando em dimensões, sobreposições e presença de ruídos.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um novo método de avaliação de algoritmos de agrupamento, o CLAIRE, utilizando-se da TRI para a estimação das habilidades latentes de um conjunto de modelos, baseando-se em concordância entre eles e considerando uma dificuldade e discriminante associado as instâncias de um conjunto.

1.2.2 Objetivo Específicos

Dos objetivos específicos desse trabalho, temos:

- Revisar a literatura existente sobre a Teoria de Resposta ao Item, explorando seus princípios e conceitos fundamentais;
- Revisar a literatura existente sobre métodos para avaliação de modelos de agrupamento;
- Formulação da matriz de resposta, considerando um grau de concordância entre modelos, par a par, ao agrupar uma determinada instância de um conjunto de dados;
- Estimação dos parâmetros da TRI, utilizando-se do modelo β^4-IRT;
- Avaliação da estimação das habilidades pelo método do CLAIRE e o rankeamento do conjunto de modelos utilizados, para comparação de desempenho;
- Propor o CLAIRE como novo meio de avaliação para modelos de agrupamento.

1.3 Organização do trabalho

Além da introdução, este trabalho é organizado na seguinte sequência:

- **Capítulo 2:** Apresenta-se uma revisão sucinta sobre medidas de validação interna e externa para métodos de agrupamento.
- **Capítulo 3:** Apresenta sobre a Teoria de Resposta ao Item e o modelo β^3-IRT, bem como a formulação do modelo β^4-IRT (FERREIRA-JUNIOR et al., 2023), utilizado também na proposta do CLAIRE no Capítulo 4.
- **Catítulo 4:** Temos a principal contribuição deste trabalho: a formulação do CLAIRE, método para avaliação de modelos de agrupamento baseado em concordância entre eles. Aqui também definimos a matriz de resposta, bem como sua formulação matemática da mesma, para uso na estimação das habilidades do modelo β^4-IRT utilizado (FERREIRA-JUNIOR et al., 2023).

- **Capítulo 5:** É apresentado como foram realizados os experimentos ao longo do trabalho e sobre quais condições o mesmo pode ser replicado. Neste capítulo, é explicitado como os conjuntos de dados devem ser gerados, quais famílias de algoritmos e suas variações de parâmetros utilizadas e as especificações dos procedimentos para gerar os resultados.

- **Capítulo 6:** Neste Capítulo é apresentado os resultados obtidos a partir do CLAIRE, em comparação com outras medidas de validação interna e externa para agrupamentos, além da ordenação das habilidades estimadas pelo modelo proposto e a matriz de correlação da habilidade estimada para com as outras medidas de validação. Os resultados obtidos mostram que a habilidade estimada pelo CLAIRE apresenta um alto grau de relacionamento com medidas de validação, em especial, externas.

- **Capítulo 7:** Por fim, neste Capítulo é apresentado a conclusão do trabalho, em adição, direcionamentos para trabalhos futuros.

CAPÍTULO 2

Validação de Agrupamentos

Muitos métodos para avaliação de agrupamentos e comparação de modelos de agrupamento foram propostos na literatura (WU et al., 2020; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001b; KIM; RAMAKRISHNA, 2005; DEBORAH; BASKARAN; KANNAN, 2010). As medidas de avaliação podem ser organizadas em duas categorias: medidas internas, as quais utilizam-se de informações internas dos dados obtidos a partir das partições fornecidas por um algoritmo; medidas externas, as quais se baseiam na informação a priori do conjunto de dados, de modo similar às medidas de eficiência para métodos supervisionados. Neste Capítulo são apresentadas as principais medidas de validação interna e externa de agrupamentos, sendo as mesmas utilizadas para comparação com os resultados obtidos utilizando-se o método proposto neste trabalho.

2.1 Índices de validação interna

Índices de validação interna têm como intuito refletir o quão similar são as instâncias de um determinado conjunto de dados dentro de seu respectivo grupo; a conectividade, ou seja, até qual ponto as instâncias podem ser agrupadas em um mesmo grupo, bem como seus vizinhos em um espaço de dados definido; e o quão separáveis é um grupo dos seus outros grupos.

2.1.1 Índice de Calinski-Harabasz

O índice de Calinski-Harabasz (CH) tem por objetivo avaliar a validação do grupo baseando-se na média da soma de quadrados dentro do grupo obtido pelo agrupamento. A sua expressão pode ser definida a seguir (CALINSKI; HARABASZ, 1974):

$$CH = \frac{\sum_{k=1}^{K} n_k d^2(\mathbf{c}_k, \mathbf{c})}{\sum_{j=1}^{K} \sum_{\mathbf{x} \in C_k} d^2(\mathbf{x}, \mathbf{c}_k)} \times \frac{n-K}{K-1},$$

em que K é o número de grupos, n sendo o total de instâncias no conjunto de dados, n_k sendo o número de instâncias dentro do grupo C_k, $\mathbf{c}_k$ sendo o vetor de centróides referente ao grupo C_k, $\mathbf{x}$ sendo uma instância pertencente ao grupo C_k, $\mathbf{c}$ sendo o centróide geral, e por fim $d(\mathbf{x}, \mathbf{y})$ sendo a distância Euclidiana entre as instâncias $\mathbf{x}$ e $\mathbf{y}$. Altos valores desse índices indica um algoritmo com grupos mais bem definidos.

2.1.2 Índice de Silhueta

O Índice de Silhueta (S) é um índice de validação interna para mensurar a qualidade do agrupamento considerando a distância entre pares de instâncias entre e dentro do resultado obtido a partir do agrupamento. A expressão para definir essa medida é definida a seguir (ROUSSEEUW, 1987):

$$S(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}},$$

sendo $a(j)$ a média das distâncias entre a instância j e todas as outras instâncias no mesmo grupo, $b(j)$ é a menor distância média entre a instância j e qualquer outro grupo. Esta medida possui um suporte no intervalo $[-1, 1]$, quanto mais proximo de um, melhor foram agrupadas as instâncias e quanto mais proximo de menos um, pior é o agrupamento.

2.1.3 Índice de Davies-Bouldin

O índice de Davies-Bouldin (DB) tem como objetivo medir a similaridade entre cada grupo e o seu respectivo mais similar, sendo calculado baseado na razão entre a distância média intra-grupo e a distância interna dentro do grupo entre os centróides de cada

grupo. A representação matemática para esse índice pode ser definido a seguir (DAVIES; BOULDIN, 1979):

$$\mathrm{DB} = \frac{1}{K}\sum_{k=1}^{K}\max_{t\neq k}\left\{\frac{s_k + s_t}{d(c_k, c_t)}\right\},$$

sendo K o número de grupos, c_k o centróide do grupo k, s_k é a distância média entre todos os pontos no grupo k e o centróide c_k, e por fim, $d(c_k, c_t)$ é a distância entre os centróides do grupo k e t. Para interpretação desta métrica, quanto menor o valor desse índice, melhor será o agrupamento, com valores mais próximos de zero indicando grupos mais compactos e bem definidos.

2.2 Índices de validação externa

Índices de validação externa necessitam do conhecimento acerca dos verdadeiros rótulos das instâncias, sendo considerado como informação a priori sobre o conhecimento da partição dos dados. Esse tipo de métrica é mais similar à medidas de avaliação de métodos supervisionados, sendo mais utilizados em situações experimentais, em que se possui essa informação, sendo utilizado para comparar de forma geral a performace de algoritmos de agrupamento.

2.2.1 Índice de Informação Mútua

Essa medida tem como intuito mensurar o total de informação que o resultado do método de agrupamento obtém acerca do verdadeiro rótulo daquela instância. A formulação desta métrica é descrita abaixo (STREHL; GHOSH, 2002):

$$MI(T, P) = \sum_{r=1}^{|T|}\sum_{s=1}^{|P|} p_{rs}\log\frac{p_{rs}}{p_r p_s},$$

sendo $|T|$ o número de grupos únicos verdadeiros e $|P|$ o número de grupos únicos previstos pelo modelo, $p_{r,s}$ é a proporção de instâncias que pertencem ao verdadeiro grupo r e foi predito como s, p_r é a proporção de instâncias que pertencem ao grupo verdadeiro r e p_s é a proporção de instâncias que pertencem ao grupo predito como s. O suporte

definido dessa medida é o intervalo $[0, 1]$, sendo valores mais próximos de um indicando agrupamentos com melhores performaces.

2.2.2 V-measure

A V-measure busca mensurar a homogeneidade e completude dos resultados do modelo de agrupamento. Para tanto, é calculado a média harmônica do score obtido a partir da homogeneidade e da completude, obtidos levando em consideração tanto rótulos verdadeiros quanto os previstos pelo modelo para a instância. Definimos a expressão como (ROSENBERG; HIRSCHBERG, 2007):

$$V = 2 \times \frac{\text{homogeneidade} \times \text{completude}}{\text{homogeneidade} + \text{completude}}, \tag{2.1}$$

Sendo:

$$\text{homogeneidade} = 1 - \frac{H(C|P)}{H(C)}, \text{ e completude} = 1 - \frac{H(P|C)}{H(P)} \tag{2.2}$$

onde C é o conjunto de rótulos verdadeiros, P o conjunto de rótulos preditos, $H(C|P)$ é a entropia condicional do verdadeiro rótulo da instância dado o rótulo predito, $H(C)$ é a entropia do verdadeiro rótulo, $H(P|C)$ é a entropia condicional do rótulo predito dado o verdadeiro rótulo das instâncias e por fim, $H(P)$ é a entropia dos rótulos preditos. Essa medida está definida no intervalo $[0, 1]$, considerando valores maiores e próximos de um, indicando uma boa performace do agrupamento.

2.2.3 Índice de Rand Ajustado

Esta métrica busca mensurar o grau de concordância entre os grupos a priori e os grupos encontrados pelo método de agrupamento. Definimos a equação deste índice a seguir (HUBERT; ARABIE, 1985):

$$ARI = \frac{\sum_{rs} \binom{n_{rs}}{2} - [\sum_r \binom{a_r}{2} \sum_s \binom{b_s}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_r \binom{a_r}{2} + \sum_s \binom{b_s}{2}] - [\sum_r \binom{a_r}{2} \sum_s \binom{b_s}{2}]/\binom{n}{2}},$$

onde n é o total de números de instâncias, n_{rs} é o número de instâncias que foram colocadas em um mesmo grupo, tanto pelo modelo de agrupamento quanto pelo grupo verdadeiro,

a_r é o número de instâncias que foram colocadas no $r-$ésimo grupo verdadeiro, b_s é o número de instâncias que foram colocadas no $s-$ésimo grupo predito pelo modelo, e por fim $\binom{m}{2}$ é o número de formas diferentes de escolher duas observações de um conjunto de m instâncias. O suporte desta métrica é definida no intervalo $[-1, 1]$, em que um indicará uma perfeita concordância entre as partições valores proximos a zero ou negativos correspondem a concordância por acaso entre as partições.

2.2.4 Considerações Finais

Neste capítulo, exploramos diferentes métodos de validação externa e interna que podem ser aplicados para avaliar a eficácia e confiabilidade de modelos de agrupamento. Em resumo, os métodos de validação externa e interna são fundamentais para garantir a robustez e a generalização dos modelos analíticos. Ao utilizar essas técnicas, é possível obter uma avaliação mais confiável do desempenho do modelo e tomar decisões informadas com base nos resultados obtidos.

CAPÍTULO 3

Téoria de Resposta ao Item

Na TRI, o intuito é modelar a probabilidade da resposta correta de um respondente para um determinado item, considerando uma habilidade latente do respondente e uma dificuldade relacionada ao item. Em parte dos estudos de TRI, os modelos binários são adotados, nos quais existe apenas resposta correta ou incorreta. Esses modelos assumem uma resposta binária x_{ij}, representando a resposta do i-ésimo respondente para o j-ésimo item. Como extensão, temos o modelo com dois parâmetros modelados para o item (2PL), definindo $x_{ij} = 1$ sendo considerado como resposta correta, sendo modelada por uma função logística com parâmetro de localização δ_j e um parâmetro de forma a_j, com θ_i sendo a habilidade do respondente i, com as respostas modeladas por uma distribuição Bernoulli (BERNOULLI, 1738), com parâmetro p_{ij} como dado pela Equação (3.1).

$$x_{ij} = \mathcal{B}ern(p_{ij}),\ p_{ij} = \sigma(-\mathbf{a}_j d_{ij}),\ d_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\delta}_j, \tag{3.1}$$

sendo $\sigma(\cdot)$ a função logística, N é o número de itens e M é o número de respondentes. As respostas para cada x_{ij} irão compor uma matriz $N \times M$, denominada como matriz de respostas. A partir dessa formulação, é possivel extrair um conjunto de curvas características do item (CCI), com o intuito de mapear a habilidade à resposta esperada, como definido pela Equação (3.2).

$$\mathbb{E}[x_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] = p_{ij} = \frac{1}{1 + e^{-\mathbf{a}_j \times (\boldsymbol{\theta}_i - \boldsymbol{\delta}_j)}} \tag{3.2}$$

Note que se $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$, temos que $p_{ij} = 0,5$, ou seja, a resposta esperada do respondente i para o item j é igual a $0,5$. Considerando $\mathbf{a} = 1, \forall j = 1, \ldots, N$, o modelo pode ser reduzido a um mais simples, denominado como 1PL, também conhecido como modelo de Rasch (RASCH, 1960), descrevendo os itens por suas dificuldades.

Modelos de TRI binárias, apesar do forte uso na área de psicometria, apresentam uso limitado quando tratamos de respostas que estão definidas naturalmente como contínuas, em especial, na área de aprendizagem de máquina. A Equação (3.3) define o modelo β^3-IRT (CHEN et al., 2019), em que a resposta do respondente i para o item j (p_{ij}) pode ser modelada por uma distribuição Beta.

$$\begin{aligned}
&p_{ij} \sim \mathcal{B}(\alpha_{ij}, \beta_{ij}), \\
&\alpha_{ij} = \mathcal{F}_\alpha(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{\boldsymbol{\theta}_i}{\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\
&\beta_{ij} = \mathcal{F}_\beta(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{1 - \boldsymbol{\theta}_i}{1 - \boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\
&\boldsymbol{\theta}_i \sim \mathcal{B}(1,1),\ \boldsymbol{\delta}_j \sim \mathcal{B}(1,1),\ \mathbf{a}_j \sim \mathcal{N}(1, \sigma_0^2)
\end{aligned} \tag{3.3}$$

Para este modelo, a CCI é modelada pela esperança da distribuição Beta, com parâmetros α_{ij} e β_{ij}, definida na Equação (3.4).

$$\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] = \frac{\alpha_{ij}}{\alpha_{ij} + \beta_{ij}} = \frac{1}{1 + \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j}} \tag{3.4}$$

Recentemente, foi proposto uma fatoração do parâmetro $\mathbf{a}_j$ para o modelo β^3-IRT (CHEN et al., 2019), estendendo-o para o modelo β^4-IRT (FERREIRA-JUNIOR et al., 2023). Essa divisão tem como interesse melhorar a estimação do parâmetro $\mathbf{a}_j$, uma vez que o modelo com 3 parâmetros apresentava dificuldade para recuperar o sinal do discrimante quando $p_{ij} \approx 1$, frequentemente indicando $\alpha_{ij} > 1$ e $\alpha_{ij} < 1$, podendo implicar em dois cenários: $\boldsymbol{\theta}_i > \boldsymbol{\delta}_j$ para $\mathbf{a}_j > 0$ ou $\boldsymbol{\theta}_i < \boldsymbol{\delta}_j$ para $\mathbf{a}_j < 0$. O problema da recuperação do sinal para o discriminante pode ser visto no experimento do artigo do β^4-IRT, em que é vista a mudança de sinal na Figura 3.1, representando uma alta inversão do sinal da instância. A CCI para o modelo β^4-IRT pode ser modelado pela Equação (3.5).

$$E[p_{ij}|\theta_i, \delta_j, \omega_j, \tau_j] = \frac{1}{1 + \left(\frac{\delta_j}{1-\delta_j}\right)^{\tau_j \cdot \omega_j} \cdot \left(\frac{\theta_i}{1-\theta_i}\right)^{-\tau_j \cdot \omega_j}}. \tag{3.5}$$

Tendo a substituição do discriminante $\mathbf{a}_j$ da Equação (3.4) pelos parâmetros ω_j e τ_j, representando o sinal e o valor absoluto do discriminante, respectivamente. Além disso, a separação do discriminante em dois parâmetros a serem estimados, é proposta também uma forma de inicialização dos parâmetros para a melhor estimação, dados pela Equação (3.6).

$$\begin{aligned}
\theta_i &= N^{-1} \times \left(\sum_{j=1}^{N} p_{ji} \right), \\
\delta_j &= 1 - M^{-1} \times \left(\sum_{i=1}^{M} p_{ji} \right), \\
\tau_j &= \rho(\boldsymbol{\theta}, \boldsymbol{p}_j), \\
\omega_j &= 1
\end{aligned} \tag{3.6}$$

sendo θ_i como a resposta média do i-ésimo respondente. Para os parâmetros relacionados ao discriminante, ω_j é inicializado como $\omega_j = 1$ e τ_j é o coeficiente de correlação Spearman, em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ e $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jM})$. Os sinais do discriminante do item são aproximadas pelas correlações entre as probabilidades das respostas e as habilidades, uma vez que para instâncias discriminadas positivamente, a habilidade deve aumentar, bem como o contrário deve acontecer para instâncias discriminadas negativamente, considerados como pontos ruídosos.

Para demonstrar o impacto dessa estimação do parâmetro $\mathbf{a}_j$, no artigo do β^4-IRT (FERREIRA-JUNIOR et al., 2023) são realizados experimentos para avaliar a recuperação do sinal referente ao discriminante, analisando a proporção de mudança de sinal entre a implementação do β^3-IRT e o modelo proposto, sendo comparadas as variações:

- β^3-IRT (3.1);
- β^3-IRT com 1000 inicializações fixas para o discriminante (3.2);
- β^4-IRT com 1000 inicializações fixas para o discriminante (3.3), sem a informação a priori;
- β^4-IRT com 1000 inicializações fixas e a utilização da informação a priori para todos os parâmetros (3.4).

Os conjuntos de dados foram amostrados 1000 habilidades, dificuldades e discriminações, definindo $\sigma_0^2 = 1$. Em seguida, para cada respondente i-ésimo e item j-ésimo, geramos a resposta p_{ij} tomando a média de 100 amostras da distribuição correspondente $B(\alpha_{ij}, \beta_{ij})$. O modelo original β^3-IRT foi treinado usando o conceito de Máxima Verossimilhança utilizando-se do gradiente descendente, diferente da implementação antiga, utilizando-se de Inferência Variacional.

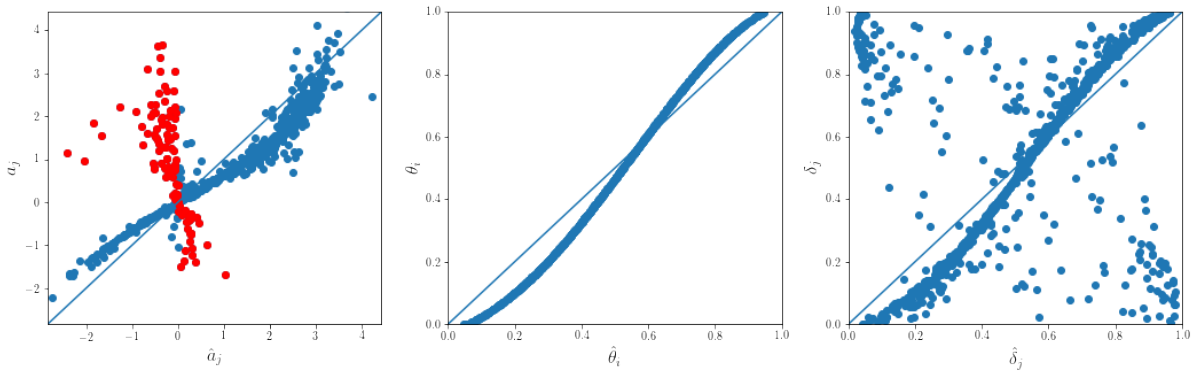

Figura 3.1: Gráficos de dispersão mostrando discriminações amostradas (esquerda), habilidades (centro) e dificuldades (direita) usadas para gerar uma matriz de resposta 1000×1000 e suas estimativas produzidas pelo β^3-IRT. Pontos vermelhos no gráfico de discriminação representam discriminações estimadas com sinais invertidos (109 de 1000 discriminações).

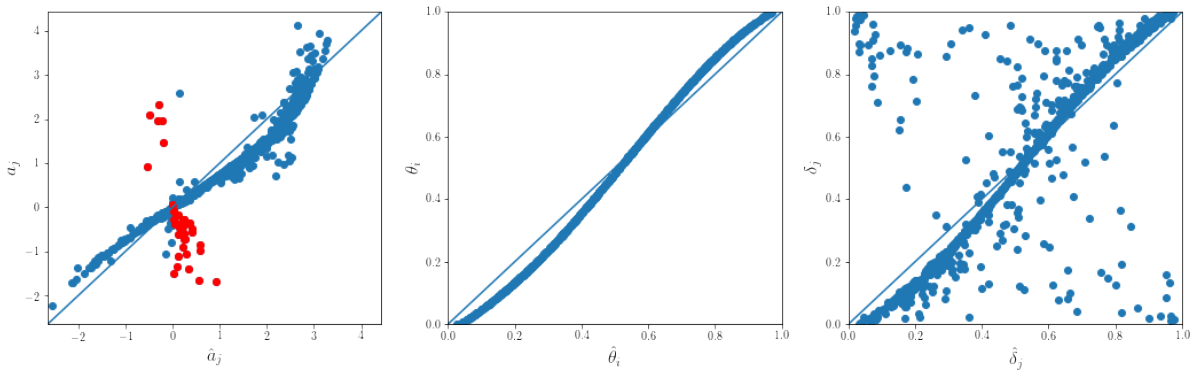

Figura 3.2: Gráficos de dispersão mostrando discriminações amostradas (esquerda), habilidades (centro) e dificuldades (direita) usadas para gerar uma matriz de resposta 1000×1000 e suas estimativas produzidas pelo β^3-IRT com 1000 iterações iniciais com discriminações fixas. Pontos vermelhos no gráfico de discriminação representam discriminações estimadas com sinais invertidos (37 de 1000 discriminações).

Na Figura 3.4 é possível notar que não houve nenhuma mudança de sinal nas estimativas para o conjunto de dados sintético, em comparação aos outros modelos no experimento. Dessa forma, é possível notar que a informação a priori incrementada para a inicialização dos parâmetros apresentou uma melhora na estimação do modelo, sendo o melhor o β^4-IRT, em especial para o parâmetro de discriminante.

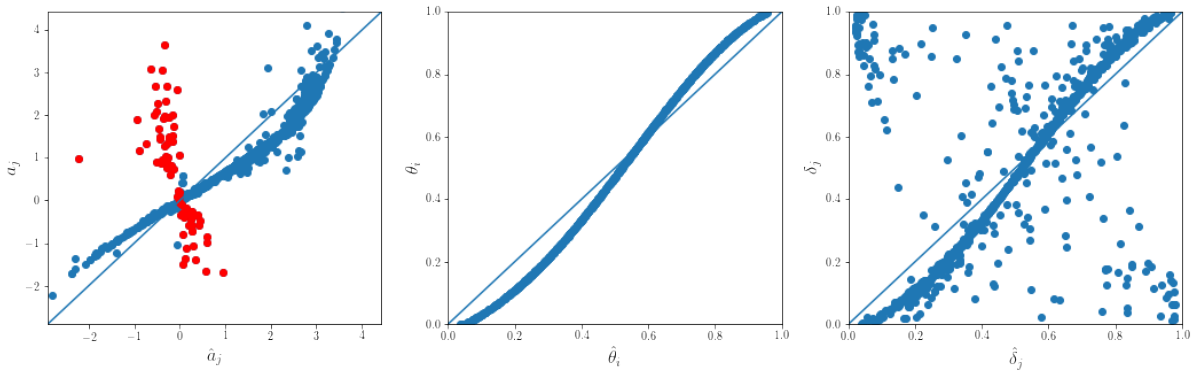

Figura 3.3: Gráficos de dispersão mostrando discriminações amostradas (esquerda), habilidades (centro) e dificuldades (direita) usadas para gerar uma matriz de resposta 1000×1000 e suas estimativas produzidas pelo β^4-IRT com 1000 iterações iniciais com discriminações fixas. Pontos vermelhos no gráfico de discriminação representam discriminações estimadas com sinais invertidos (77 de 1000 discriminações).

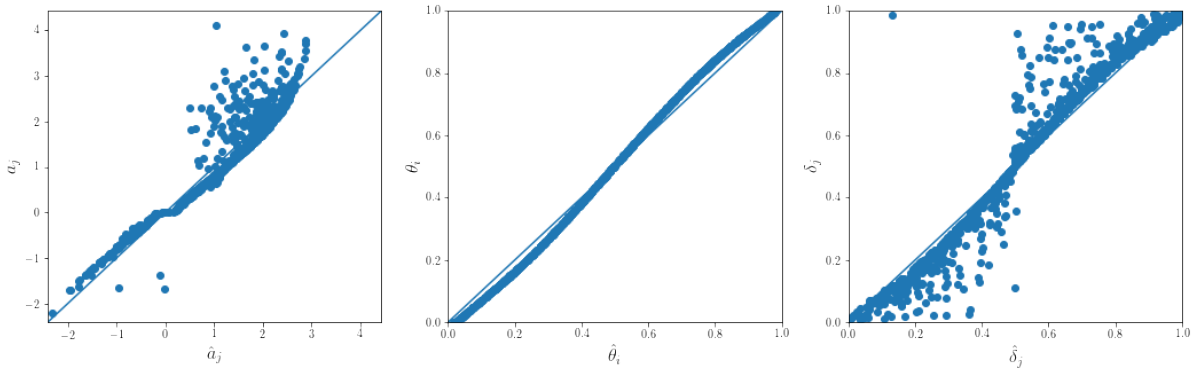

Figura 3.4: Gráficos de dispersão mostrando discriminações amostradas (esquerda), habilidades (centro) e dificuldades (direita) usadas para gerar uma matriz de resposta 1000×1000 e suas estimativas produzidas pelo β^4-IRT com 1000 iterações iniciais com discriminações fixas e prioris melhores para todos os parâmetros. Todas as discriminações foram estimadas com os sinais corretos.

3.0.1 Considerações Finais

Neste capítulo, exploramos a Teoria de Resposta ao Item (TRI), uma abordagem estatística utilizada na área de avaliação educacional e psicológica para medir habilidades, conhecimentos ou traços latentes dos indivíduos. Durante a discussão sobre a TRI, vimos que essa teoria é baseada em modelos matemáticos que relacionam as respostas dos indivíduos a itens específicos com seus níveis de habilidade. Esses modelos permitem estimar as habilidades individuais de forma mais precisa e confiável, levando em consideração as características dos itens e as respostas dadas pelos participantes.

Em resumo, a Teoria de Resposta ao Item é uma ferramenta poderosa para medir habilidades e traços latentes. Por meio dos modelos TRI e dos procedimentos de estimação dos parâmetros, é possível obter resultados mais precisos e confiáveis, proporcionando uma compreensão mais profunda das capacidades individuais e auxiliando na tomada de decisões informadas em diversos campos.

CAPÍTULO 4

CLAIRE: IRT para validação de métodos de agrupamento

A estrutura do método CLAIRE tem como intuito avaliar o desempenho de métodos de agrupamentos com diferentes tipos de estimação, associando diferentes modelos em uma matriz de resposta para a estimação utilizando-se do β^4-IRT. Considerando a nomenclatura do IRT, os modelos de agrupamento serão os respondentes e as instâncias do conjunto de dados serão os itens. Para os modelos contínuos de IRT, o valor de p_{ij} modela a resposta do respondente i para o item j, porém em agrupamentos, não é tão intuitivo mensurar se um modelo esta agrupando corretamente uma instância ou não. Entretanto, na proposta do CLAIRE, é possível definir uma noção similar, a partir da concordância entre modelos, ao agrupar uma determinada instância i. A interpretação da resposta de um modelo de agrupamento para uma determinada instância será medida como o total de modelos concordantes ao alocar uma mesma instância em um determinado grupo ou concordarem em agruparem em grupos distintos.

A Figura 4.1 apresenta a distribuição de um conjunto de dados sintético, com a presença de dois grupos bem definidos e linearmente separáveis, aonde foram criados modelos totalmente artificiais para o problema, além disso, ambos os grupos possuem a mesma quantidade de instâncias ($n_{c_1} = 10$ e $n_{c_2} = 10$, com n_{c_k} sendo o total de instâncias para o grupo c_k). Para o cenário (A), nota-se que o modelo consegue separar bem os dois grupos, não apresentando alguma instância aparentemente alocada de forma errada. Para o cenário (B), o modelo consegue separar bem os dois conjuntos, porém duas instâncias (1

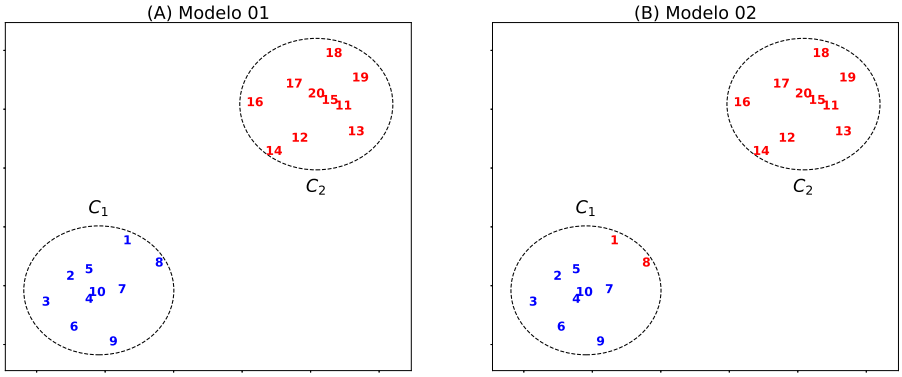

Figura 4.1: Cenário gerado com a resposta do agrupamento referente a dois modelos para um conjunto de dados sintético, com a presença de concordância e discordância entre ambos. Números do intervalo 1 a 10 pertencem ao grupo C_1 e números do intervalo 11 a 20 pertencem ao grupo C_2.

e 8) foram agrupadas de forma errada. Dessa forma, a definição de concordância e discordância entre pares de modelos definido pelo CLAIRE é ilustrada acima, em que ambos os modelos concordaram em como agrupar as instâncias 2 a 7, 9 a 20, porém analisando em pares existem certas discordâncias. Em resumo, algumas informações importantes para entender o conceito de concordância e discordância pode ser listado abaixo:

- **Concordância**:
  - Ambos modelos concordam em alocar as instâncias 5 e 7 em um mesmo grupo;
  - Ambos modelos concordam que devem alocar em grupos distintos as instâncias 5 e 20;
  - Ambos modelos concordam em alocar as instâncias 11 e 19 em um mesmo grupo.
- **Discordância**:
  - O modelo 1 aloca as instâncias 1 e 5 em um mesmo grupo, porém o modelo 2 coloca o par em grupos distintos;
  - Para o modelo 1, o par de instâncias 1 e 16 é alocado em grupos distintos, porém o modelo 2 coloca o par em grupos iguais.

Seja N sendo o número de instâncias de um conjunto de dados e M sendo o número de modelos de agrupamento a serem testados, podemos definir a matriz de resposta como $P = \{p_{ji}\}_{i \geq M, j \geq N}$, com M linhas e N colunas. Cada entrada da matriz de resposta será definida como o grau de concordância, dado aquele conjunto de modelos, para o modelo i para a instância j. A Equação (4.1) indica $a_{i,i'}^{j,j'}$ se um par de modelos (i, i') concorda em alocar um par de instâncias (j, j') em um mesmo grupo ou em grupos distintos.

$$a_{i,i'}^{j,j'} = \begin{cases} 1, & \text{se } \mathbb{I}(i,j,j') = \mathbb{I}(i',j,j'), \\ 0, & \text{caso contrário.} \end{cases} \tag{4.1}$$

Considere $\mathbb{I}(i,j,j')$ como a função indicadora, em que $\mathbb{I}(i,j,j') = 1$ se ambas instâncias j e j' foram agrupadas em um mesmo grupo pelo modelo i, e $\mathbb{I}(i,j,j') = 0$ caso contrário.

Se pares de modelos se ajustam bem aos dados, é esperado que ao agrupar um determinado conjunto de instâncias similares, ambos coloquem as mesmas em um mesmo grupo, de maneira análoga, é esperada que ambos os modelos posicionem instâncias não similares em grupos distintos. Dessa forma, é possível afirmar que o grau de concordância obtido ao agrupar esses pares de instâncias impacta diretamente na habilidade dos modelos. Portanto, podemos definir as entradas da matriz P pela Equação (4.2).

$$p_{ji} = \frac{1}{(M-1) \times (N-1)} \times \sum_{i' \neq i} \sum_{j' \neq j} a_{i,i'}^{j,j'} \tag{4.2}$$

Com o intuito de estimar as habilidades dos modelos, será utilizado a matriz de resposta formulada na Equação (4.2), sendo utilizada para modelar os parâmetros do β^4-IRT para os experimentos a se seguir deste trabalho. Para interpretação dos parâmetros estimados, CLAIRE irá retornar um conjunto de parâmetros, sendo: θ_i a habilidade estimada para o modelo i, sendo alto para modelos que apresentam um alto consenso com os outros modelos testados, δ_j mostrando o quão difícil para o conjunto de modelos concordarem em agrupar um determinada instância j.

j	Modelo 1	Modelo 2
1	0	1
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	1
9	0	0
10	0	0
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	1
20	1	1

Tabela 4.1: Resposta dos modelos sintéticos referentes a Figura 4.1.

A Tabela 4.1 apresenta a resposta de ambos os modelos sintéticos para cada instância j no conjunto de dados sintético. Ao aplicarmos a matriz acima na Equação (4.2), temos:

$$P = \{p_{ji}\}_{1\geq j\geq 20, 1\geq i\geq 2} = \begin{bmatrix} \frac{1}{(2-1)\times(20-1)} \times \sum_{i'\neq 1}\sum_{j'\neq 1} a_{1,i'}^{1,j'} & \frac{1}{(2-1)\times(20-1)} \times \sum_{i'\neq 2}\sum_{j'\neq 1} a_{2,i'}^{1,j'} \\ \vdots & \vdots \\ \frac{1}{(2-1)\times(20-1)} \times \sum_{i'\neq 1}\sum_{j'\neq 20} a_{1,i'}^{20,j'} & \frac{1}{(2-1)\times(20-1)} \times \sum_{i'\neq 2}\sum_{j'\neq 20} a_{2,i'}^{20,j'} \end{bmatrix} \tag{4.3}$$

Após o cálculo das entradas para obtenção da matriz de resposta, obtemos a matriz P, representada pela Equação (4.4) final abaixo:

$$P = \begin{bmatrix} 0,052632 & 0,052632 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,052632 & 0,052632 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \\ 0,894737 & 0,894737 \end{bmatrix} \tag{4.4}$$

Por fim, obtemos a matriz de resposta apos o processo do CLAIRE, como visto na Equação (4.4). É importante notar que o CLAIRE é uma metodologia desenvolvida baseada na concordância entre modelos de agrupamento, sobre um conjunto de dados. O desempenho direto desses modelos dependem da qualidade dos modelos passados. Dessa forma, para resultados mais eficientes, o método CLAIRE espera que os modelos passados na matriz de resposta apresentem um alto número de pares de instâncias, para os quais os modelos concordaram em agrupar em um mesmo grupo ou grupos distintos, i.e., modelos com grupos de instâncias similares em um mesmo grupo.

Nota-se que o CLAIRE é um modelo agnóstico, ou seja, a matriz de resposta pode conter qualquer tipo de modelo que possa particionar os dados em um ou mais grupos. Dessa forma, modelos capazes de detectar ruídos bem como o DBSCAN (ESTER et al., 1996) e o Optics (ANKERST et al., 1999). Assim, uma vez que o modelo DBSCAN agrupa instâncias como ruídos, os modelos concordantes são aqueles que alocaram as mesmas instâncias em um mesmo grupo ou também detectaram que são instâncias ruídosas.

CAPÍTULO 5

Experimentos

Neste capítulo serão apresentados os modelos de agrupamento a serem utilizados nos experimentos para o CLAIRE, selecionados de um diverso grupo de famílias de algoritmos. Os algoritmos utilizados foram: KMeans (MACQUEEN, 1967), DBSCAN (ESTER et al., 1996), SpectralClustering (NG; JORDAN; WEISS, 2002), MeanShift (COMANICIU; MEER, 2002), OPTICS (ANKERST et al., 1999), e Kernel KMeans (DHILLON; GUAN; KULIS, 2004). Além disso, a Seção 5.1 irá abordar a geração dos dados sintéticos e as características de cada um deles individualmente. Em seguida, a Seção 5.3 mostra os procedimentos seguintes para reprodução dos experimentos deste trabalho e a definição de dois modelos sintéticos para compor a matriz de resposta. O experimento foi realizado em um computador com as seguintes especificações: processador Intel Core i7-10510U 1.80GHz, 16 GB de memória RAM DDR4, sistema operacional Linux 22.04.1-Ubuntu e utilizando a versão 3.10.6 do Python. O ambiente de desenvolvimento bibliotecas Pandas 1.5.0, Scikit-learn 1.1.2, Numpy 1.23.3 e Birt-gd 0.1.47.

5.1 Conjuntos de dados

Os experimentos foram realizados considerando os conjuntos de dados sintéticos da biblioteca Scikit-learn (PEDREGOSA et al., 2011), sendo eles: Anisotropic, Blobs Varied,

Noisy Circles e Noisy Moons. Além destes conjuntos com comportamento distintos, foi gerado um conjunto de dados uniformamente dentro de um hipercubo, No Structure, gerado com o intuito de similar dados homogenous, em outras palarvras, que não exista um grupo definido. A Figura 5.1 mostra o comportamento dos conjuntos gerados. A descrição sobre cada conjunto segue abaixo.

- **Anisotropic:** Gerado aleatoriamente em um espaço 2D. Os pontos são rotacionados pela matriz definida como:

$$A = \begin{bmatrix} 0,6 & -0,6 \\ -0,4 & 0,8 \end{bmatrix}.$$

- **Blobs:** Conjunto de dados gerado aleatoriamente, considerando uma distribuição Gaussiana em um espaço 2D, com dois grupos próximos e um grupo linearmente separável.
- **Varied:** Grupos com diferentes variâncias, sendo o conjunto mais similar com o problema real, tendo um grupo com alta variabilidade entre dois grupos distantes.
- **Noisy Circles:** Conjunto composto por dois círculos concêntricos, gerados a partir de uma distribuição Gaussiana, sendo uma das estruturas mais complexas e tarefa não linearmente separável.
- **Noisy Moons:** Grupos definidos como duas meias luas, com comportamentos opostos, sendo também um conjunto não linear com uma estrutura complexa.
- **No Structure:** Conjunto gerado uniformemente distribuído em torno de um hipercubo.

A escolha dos conjuntos de dados citados dar-se devido a abrangência de diferentes tipos de cenários que podem ser encontrados em problemas reais. Além disso, para cada conjunto, é esperado que um determinado modelo performe melhor, devido a forma de estimação de um dado algoritmo utilizado. Por fim, para todos os conjuntos foi considerado um tamanho de amostra de 500 instâncias, com exceção do No Structure, em que foi considerado um amostra de 1000 instâncias.

5.2 Algoritmos de Agrupamento

- **K-means (MCQUEEN, 1967):** Este modelo busca dividir um conjunto de dados em k grupos distintos, onde cada grupo é representado pelo seu centroide, que é

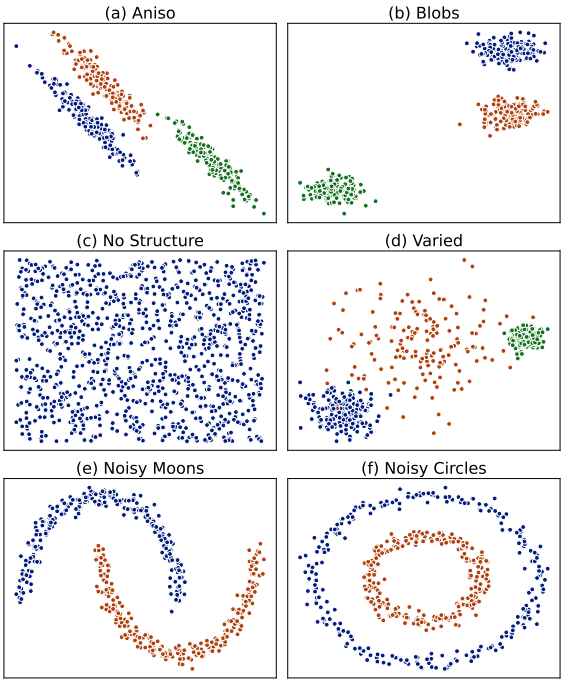

Figura 5.1: Distribuição dos dados utilizados no experimentos. Os parâmetros para replicar os experimentos encontram-se na documentação da biblioteca do Scikit-Learn (PEDREGOSA et al., 2011).

calculado como a média dos pontos pertencentes ao grupo. O objetivo do K-means é minimizar a distância entre os pontos dentro de cada grupo e maximizar a distância entre os grupos. Serão utilizados variações do número de grupos. Sendo K o número de grupos, então:

$$K \in \{1, 2, 3, 4, 5, 6, 7, 8\}$$

- **DBSCAN (ESTER et al., 1996):** Este modelo não requer que o número de grupos seja especificado a priori. Em vez disso, ele identifica regiões densas de pontos no espaço de dados e os agrupa em grupos. O DBSCAN define pontos como núcleos se eles têm um número mínimo de pontos dentro de uma determinada distância (chamada de epsilon) ao seu redor. Em seguida, expande esses núcleos encontrando outros pontos próximos e adicionando-os ao mesmo grupos. Serão utilizado modelos com variações do parâmetro eps, onde:

$$\text{eps} \in \{0,1,\ 0,2,\ 0,3,\ 0,4,\ 0,5,\ 0,6,\ 0,7,\ 0,8,\ 0,9\}$$

  Esse parâmetro referece-se a máxima distância entre duas amostras para que seja considerado um vizinho. O número de amostras em uma vizinhança foi fixado igual a 2;

- **Spectral Clustering (NG; JORDAN; WEISS, 2002):** Este modelo utiliza a análise espectral de uma matriz de similaridade dos dados para encontrar estruturas de grupos, assim calculando autovalores e autovetores da matriz de similaridade para identificar as características principais dos dados. O parâmetro referente ao número de grupos foi variado. Sendo K o número de grupos, então:

$$K \in \{1, 2, 3, 4, 5, 6, 7, 8\}$$

  Para a estratégia de decomposição dos autovalores, foi fixado o método arpack considerando o método para construir a matriz de afinidade sendo nearest_neighbors;

- **Mean Shift (COMANICIU; MEER, 2002):** Este modelo busca encontrar os modos ou centros de densidade dos dados, sendo capaz de identificar grupos de diferentes tamanhos e formas, sem a necessidade de especificar o número de grupos antecipadamente. Será considerado os parâmetros default da implementação da biblioteca Scikit-learn;

- **Optics (ANKERST et al., 1999):** Permite a descoberta de estruturas de grupos com diferentes densidades e tamanhos, sem a necessidade de especificar o número de grupos a priori, calculando a densidade de cada ponto em relação aos seus vizinhos e constrói um gráfico de alcance que ordena os pontos com base em sua densidade e proximidade. Os parâmetros referente ao número de amostras em uma vizinhança de um ponto, para que este seja considerado um núcleo, será configurado igual a 2;

- **Kernel K-means (DHILLON; GUAN; KULIS, 2004):** Este modelo permite o agrupamento de dados em espaços de características não lineares, utilizando um conjunto de funções de kernel para medir as similaridades entre os pontos no espaço transformado. Será considerado variações do número de grupos. Sendo K o número de grupos, então:

$$K \in \{1, 2, 3, 4, 5, 6, 7, 8\}$$

5.3 Configuração dos Experimentos

Cada experimento realizado neste trabalho, considerando todos os modelos da Seção 5.2 para cada conjunto de dados da Seção 5.1 foi executado um processo para agrupar os dados. Após o processo de agrupamento, é aplicada a Equação (4.2) para a obtenção da matriz de resposta. Além disso, foi incluído dois modelos artificiais a matriz de resposta:

(1) a resposta média para cada instância, com o nome de **Average Response** e (2) a resposta máxima para cada instância, **Best Response**. Dessa forma, para cada conjunto de dados é obtido uma matriz $M \times N$, sendo $M = 37$ (Número de modelos utilizados) e N é o número de instâncias para um específico conjunto de dados.

Por fim, com a matriz de resposta obtida, o modelo β^4-IRT é executado para a estimação da habilidade para cada modelo, assim como a dificuldade e discriminante para cada instância do conjunto. Para fim de comparação entre as métricas de validação e a habilidade estimada, é calculado a correlação de Spearman (SPEARMAN, 1904). Para o conjunto **No Structure**, não é possível calcular medidas de validação externas, além disso não é calculado os índices de validação interna para modelos que identificaram apenas um grupo, uma vez que essas métricas precisam de ao menos dois grupos para serem calculados. Por fim, nenhuma métrica foi calculada para os modelos artificiais, por ambos serem gerados diretamente após a aplicação da Equação (4.2), ou seja, não possuem partições definidas sobre os conjuntos de dados.

CAPÍTULO 6

Resultados

Os modelos utilizados foram ranqueados baseados na habilidade estimada para cada conjunto de dados, como vistos na Tabela 6. Como era esperado, o modelo **Best Response** sempre foi colocado como primeiro lugar e o **Average Response** nunca foi o pior.

Na Tabela 6.1, temos para cada conjunto de dados a habilidade estimada para cada modelo utilizado, com seu pré-definido conjunto de hiper-parâmetros. Para o conjunto de dados Aniso, Spectral Clustering, DBSCAN e Kernel K-Means, todos tiveram versões que superaram K-Means e Mean Shift, o que era esperado, visto que K-Means e Mean Shift tendem a se ajustar bem quando o conjunto de dados consiste principalmente em grupos esféricos de tamanho semelhante . Este é o caso de Blobs, para o qual Mean Shift, K-Means, DBSCAN e Kernel K-Means estão todos empatados no topo com o Best Model.

Para o conjunto de dados Varied, embora os grupos tenham tamanhos diferentes, a maior variância do grupo do meio cria áreas difusas de sobreposição com os grupos menores. Portanto, K-Means ($n\_clusters = 3$), Mean Shift e Kernel K-Means ($n\_clusters = 3$) simplesmente dividem os dados em 3 grupos de tamanho semelhante, aparentemente uma boa abordagem para ajustar os dados.

Para o Noise Moons, com seus grupos que não são linearmente separáveis, é um bom

exemplo de quando DBSCAN e Spectral Clustering devem ter um bom desempenho. Isso é realmente visto na Tabela 6.1, com quatro versões do DBSCAN e duas versões do Spectral Clustering entre os melhores modelos.

Para o conjunto Noisy Circles, CLAIRE coloca corretamente Spectral Clustering e Kernel K-means no topo e espera-se que esses métodos se ajustem bem a esse conjunto de dados, movendo os dados para um novo espaço onde os dois grupos são mais separáveis. DBSCAN e Optics também devem ser capazes de ajustar este conjunto de dados, no entanto, eles tiveram uma classificação baixa. Na verdade, o Optics não fez um bom trabalho para nenhum dos conjuntos de dados, provavelmente porque os valores testados para seus hiper-parâmetros levaram muitas instâncias a serem rotuladas como ruído. Isso também pareceu afetar algumas versões do DBSCAN.

Para No Structure, espera-se que os modelos com $n\_clusters = 1$ tenham a melhor classificação. No entanto, a maioria dos modelos não concordou com uma única partição de grupo e, de fato, os melhores modelos foram aqueles que dividiram os dados em 4 grupos. Acreditamos que isso ocorre porque apenas 3 dos 37 modelos (35 modelos, em adição os modelos Best Response e Average Response) ajustaram explicitamente um único grupo ao conjunto de dados, portanto, seria difícil obter consenso sobre suas decisões. Por fim, é importante salientar que, apesar das habilidades estimadas parecerem todas próximas, isso não prejudica o CLAIRE em classificar os modelos adequadamente, considerando os conjuntos de dados e modelos utilizados na estrutura do experimento.

Tabela 6.1: Habilidades estimadas e ranking (sub-escrito) dos modelos para cada conjunto de dados.

Modelo	Versão	Aniso	Blobs	Varied	Moons	Circles	No Structure
Best Response	-	$0,506_{1,0}$	$0,500_{4,0}$	$0,518_{1,0}$	$0,489_{1,0}$	$0,500_{1,0}$	$0,511_{1,0}$
Average Response	-	$0,461_{26,0}$	$0,422_{28,0}$	$0,456_{25,0}$	$0,444_{29,0}$	$0,463_{29,0}$	$0,471_{23,0}$
Mean Shift	-	$0,492_{11,0}$	$0,500_{4,0}$	$0,514_{4,5}$	$0,468_{22,0}$	$0,466_{28,0}$	$0,436_{31,5}$
Optics	min_samples = 2	$0,429_{27,0}$	$0,345_{34,0}$	$0,402_{27,0}$	$0,417_{30,0}$	$0,458_{30,0}$	$0,444_{25,0}$
DBSCAN	eps=0,1	$0,480_{23,0}$	$0,491_{8,0}$	$0,484_{21,0}$	$0,448_{28,0}$	$0,471_{25,0}$	$0,449_{24,0}$
	eps=0,2	$0,502_{3,0}$	$0,500_{4,0}$	$0,503_{9,0}$	$0,479_{4,5}$	$0,479_{21,5}$	$0,436_{31,5}$
	eps=0,3	$0,401_{32,5}$	$0,500_{4,0}$	$0,384_{28,0}$	$0,479_{4,5}$	$0,479_{21,5}$	$0,436_{31,5}$
	eps=0,4	$0,401_{32,5}$	$0,500_{4,0}$	$0,381_{29,0}$	$0,479_{4,5}$	$0,479_{21,5}$	$0,436_{31,5}$
	eps=0,5	$0,401_{32,5}$	$0,439_{20,0}$	$0,380_{30,5}$	$0,479_{4,5}$	$0,479_{21,5}$	$0,436_{31,5}$
	eps=0,6	$0,401_{32,5}$	$0,439_{20,0}$	$0,380_{30,5}$	$0,354_{34,0}$	$0,393_{34,0}$	$0,436_{31,5}$
K-means	n_clusters=1	$0,401_{32,5}$	$0,236_{36,0}$	$0,379_{34,5}$	$0,354_{34,0}$	$0,393_{34,0}$	$0,436_{31,5}$
	n_clusters=2	$0,491_{12,0}$	$0,439_{20,0}$	$0,489_{18,5}$	$0,465_{24,0}$	$0,469_{27,0}$	$0,492_{13,0}$
	n_clusters=3	$0,499_{5,0}$	$0,500_{4,0}$	$0,515_{2,0}$	$0,471_{17,0}$	$0,481_{17,5}$	$0,500_{8,5}$
	n_clusters=4	$0,495_{6,0}$	$0,483_{9,5}$	$0,513_{7,0}$	$0,472_{13,5}$	$0,487_{12,0}$	$0,508_{2,5}$
	n_clusters=5	$0,487_{14,0}$	$0,457_{14,0}$	$0,499_{10,0}$	$0,472_{13,5}$	$0,484_{14,0}$	$0,502_{5,0}$
	n_clusters=6	$0,487_{14,0}$	$0,435_{25,0}$	$0,496_{13,0}$	$0,474_{11,0}$	$0,483_{15,0}$	$0,495_{11,0}$
	n_clusters=7	$0,482_{18,5}$	$0,426_{27,0}$	$0,494_{14,0}$	$0,468_{22,0}$	$0,482_{16,0}$	$0,490_{15,0}$
	n_clusters=8	$0,481_{21,0}$	$0,415_{31,0}$	$0,492_{15,5}$	$0,461_{27,0}$	$0,481_{17,5}$	$0,484_{19,0}$
Kernel K-means	n_clusters=1	$0,401_{32,5}$	$0,236_{36,0}$	$0,379_{34,5}$	$0,354_{34,0}$	$0,393_{34,0}$	$0,436_{31,5}$
	n_clusters=2	$0,493_{8,5}$	$0,439_{20,0}$	$0,489_{18,5}$	$0,469_{20,0}$	$0,470_{26,0}$	$0,491_{14,0}$
	n_clusters=3	$0,500_{4,0}$	$0,500_{4,0}$	$0,514_{4,5}$	$0,470_{19,0}$	$0,479_{21,5}$	$0,501_{6,5}$
	n_clusters=4	$0,493_{8,5}$	$0,483_{9,5}$	$0,514_{4,5}$	$0,472_{13,5}$	$0,488_{10,0}$	$0,508_{2,5}$
	n_clusters=5	$0,486_{16,5}$	$0,461_{13,0}$	$0,498_{11,0}$	$0,471_{17,0}$	$0,486_{13,0}$	$0,500_{8,5}$
	n_clusters=6	$0,487_{14,0}$	$0,440_{16,0}$	$0,509_{8,0}$	$0,472_{13,5}$	$0,492_{5,0}$	$0,494_{12,0}$
	n_clusters=7	$0,482_{18,5}$	$0,450_{15,0}$	$0,492_{15,5}$	$0,468_{22,0}$	$0,488_{10,0}$	$0,488_{16,0}$
	n_clusters=8	$0,478_{24,0}$	$0,478_{11,0}$	$0,497_{12,0}$	$0,464_{25,0}$	$0,488_{10,0}$	$0,484_{19,0}$
Spectral Clustering	n_clusters=1	$0,401_{32,5}$	$0,236_{36,0}$	$0,379_{34,5}$	$0,354_{34,0}$	$0,393_{34,0}$	$0,436_{31,5}$
	n_clusters=2	$0,493_{8,5}$	$0,364_{33,0}$	$0,486_{20,0}$	$0,479_{4,5}$	$0,479_{21,5}$	$0,483_{21,0}$
	n_clusters=3	$0,503_{2,0}$	$0,420_{29,5}$	$0,514_{4,5}$	$0,479_{4,5}$	$0,489_{8,0}$	$0,501_{6,5}$
	n_clusters=4	$0,481_{21,0}$	$0,477_{12,0}$	$0,491_{17,0}$	$0,476_{9,5}$	$0,492_{5,0}$	$0,507_{4,0}$
	n_clusters=5	$0,493_{8,5}$	$0,384_{32,0}$	$0,468_{23,0}$	$0,476_{9,5}$	$0,495_{2,0}$	$0,499_{10,0}$
	n_clusters=6	$0,486_{16,5}$	$0,438_{24,0}$	$0,472_{22,0}$	$0,477_{8,0}$	$0,494_{3,0}$	$0,484_{19,0}$
	n_clusters=7	$0,481_{21,0}$	$0,428_{26,0}$	$0,464_{24,0}$	$0,471_{17,0}$	$0,492_{5,0}$	$0,485_{17,0}$
	n_clusters=8	$0,477_{25,0}$	$0,420_{29,5}$	$0,452_{26,0}$	$0,463_{26,0}$	$0,491_{7,0}$	$0,482_{22,0}$

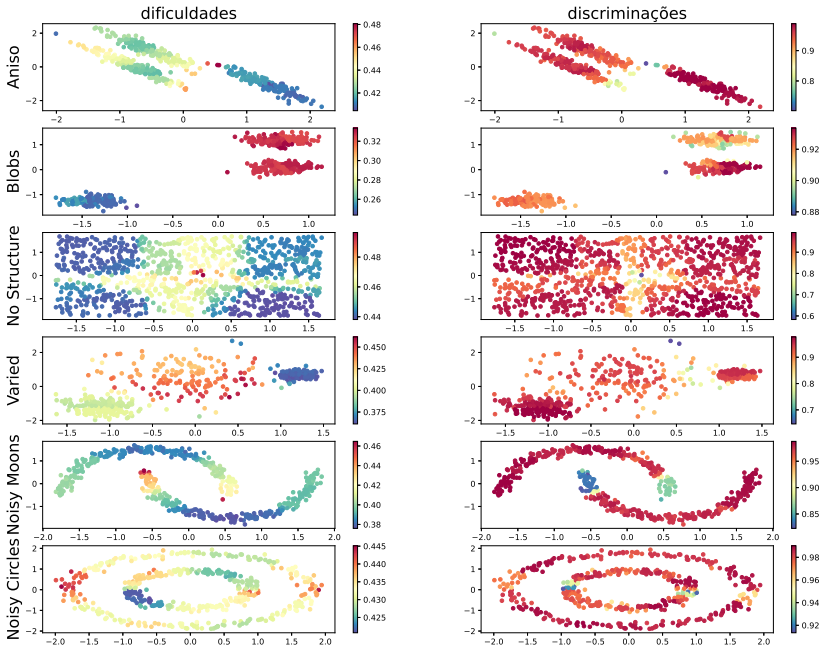

Figura 6.1: Distribuição das dificuldades e discriminações estimadas pelo β^4-IRT para cada conjunto de dados sintético.

A Figura 6.1 mostra as dificuldades e discriminações para as instâncias de todos os conjuntos de dados. Em geral, as instâncias internas de cada grupo tendem a ser mais fáceis (tons mais escuros de azul), pois os modelos tendem a concordar que devem ser agrupados juntos. A dificuldade aumenta ao longo das bordas do grupo, onde se espera que os modelos discordem mais. Para o conjunto No Structure, foi traçado uma cruz com centro do conjunto, quanto mais proximo do centro, ou da cruz, dificuldades mais altas, cenário analogo para o discriminante. Os modelos tentaram fazer o possível para encontrar alguma estrutura ali, então havendo uma alta discordância em especial nas fronteiras dos 4 grupos que foram divididos. Em relação às discriminações, embora os valores sejam todos muito semelhantes, existe uma ligeira tendência para maior discriminação nas fronteiras, pois estas regiões são melhores a separar os melhores modelos para aquele conjunto de dados dos restantes modelos que não tiveram um bom desempenho. Esse resultado é diferente daqueles obtidos em aplicações anteriores da TRI em aprendizado de máquina (CHEN et al., 2019; MORAES et al., 2020; MORAES et al., 2022), que apresentaram maior discriminação para itens mais fáceis.

Em seguida, calculamos a correlação de Spearman entre as habilidades estimadas e os índices de validação interna e externa para cada conjunto de dados (exceto Uniforme, pois alguns índices não podem ser calculados para apenas um grupo). A Figura 6.2 mostra

um mapa de calor das correlações.

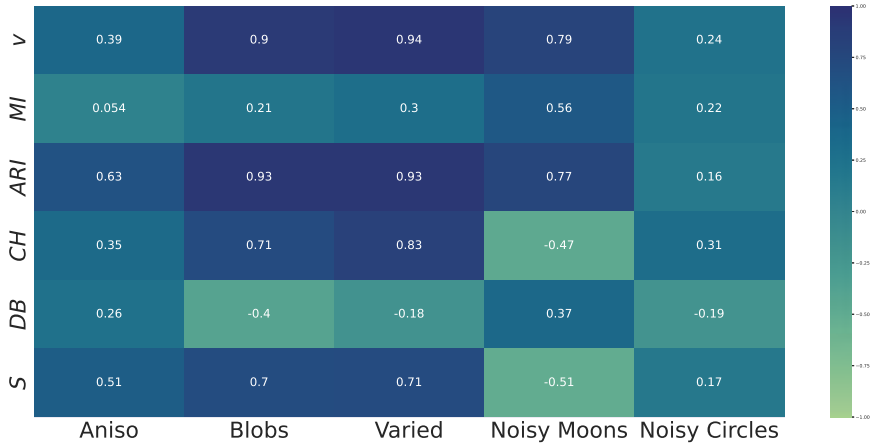

Figura 6.2: Correlação da habilidade estimada com cada método de validação de agrupamento utilizado.

Os resultados indicam uma forte relação positiva entre as habilidades estimadas e o ARI para todos os conjuntos de dados, exceto Noisy Circles. Da mesma forma, as habilidades e a medida V foram fortemente e positivamente correlacionadas para três dos cinco conjuntos de dados. Particularmente, as correlações mais fortes foram observadas para Blobs, Varied e Noisy Moons, para as quais também houve uma forte correlação positiva entre habilidades e MI. Esses resultados são importantes e surpreendentes. O CLAIRE se baseia simplesmente na concordância entre o pool de modelos e ainda é capaz de produzir rankings que se aproximam daqueles obtidos a partir de índices de validação externa, pois, sendo índices de validação externa, como ARI, MI e V-measure, que realmente usam a verdade básica para avaliar modelos de agrupamento.

Considerando os índices de validação interna, as habilidades estimadas pelo CLAIRE também estão, pelo menos, moderadamente correlacionadas positivamente (fortemente correlacionadas para Blobs e Varied) com CH e Silhouette, e negativamente correlacionadas com DB. Isso é invertido para Noisy Moons, o que pode indicar que esses índices não são adequados para avaliar modelos nas condições impostas por esse conjunto de dados, considerando que o CLAIRE tende a classificar os modelos de maneira semelhante aos índices externos para esse conjunto de dados.

Para Noisy Circles, as habilidades de CLAIRE não foram particularmente correlacionadas com nenhum dos outros índices de validação. Isso pode indicar que os outros índices não conseguem classificar corretamente o desempenho em tal conjunto de dados. O CLAIRE, por outro lado, identifica corretamente os melhores modelos para este conjunto

de dados, conforme mencionado acima.

CAPÍTULO 7

Conclusões

Neste trabalho foi apresentado uma nova forma para a avaliação de métodos de agrupamento, utilizando-se da Teoria de Resposta ao Item com o modelo β^4-IRT (FERREIRA-JUNIOR et al., 2023). O método proposto é baseado no grau de concordância em como agrupar um instância j em um determinado grupo ou em grupos distintos, dado pela formula da Equação (4.2). O método proposto visa cobrir um grupo maior de algoritmos de agrupamento para avaliação, de uma forma unificada. Com respeito a pressupostos do método novo, é considerado que seja utilizados modelos minimamente bons para aquele específico conjunto de dados.

Após a realização dos experimentos, foram armazenados os rankings, ordenado pela habilidade estimada, para cada modelo considerando cada conjunto de dados. A habilidade estimada apresentou uma alta correlação com os índices de validação de agrupamentos, em específico para medidas de mensuração externa, ainda que nenhuma informação externa seja fornecida ao CLAIRE. Outro fator importante advinda pela utilização da Teoria de Resposta ao Item, é a consideração de uma dificuldade para cada instância sendo utilizada internamente para estimação de uma habilidade para um dado modelo de agrupamento, ajudando também a identificar como, por exemplo, a criação áreas de incerteza como as fronteiras entre grupos.

Por fim, a solução proposta utiliza-se da resposta de um conjunto de modelos, utilizando-

se do consenso entre eles sobre a estrutura do grupo. Além disso, a TRI atribui mais importância as instâncias mais difíceis a serem agrupadas. Considerando a dificuldade da instância do conjunto, uma análise mais profunda dos dados pode ser realizada dos métodos de agrupamento, por exemplo, criação de áreas de incerteza como as fronteiras entre grupos.

7.1 Trabalhos futuros

Alguns trabalhos podem ser desenvolvidos considerando o tema abordado neste trabalho, dentre os possíveis assuntos podemos listar:

- Caracterização de áreas de incerteza no conjunto de dados:
  - Identificar instâncias difíceis para serem agrupadas em um conjunto de dados;
  - Utilizar da dificuldade estimada a partir do processo do CLAIRE para definir instâncias consideradas difíceis e assim definir regiões de incertezas no conjunto de dados.
- Extensão do CLAIRE para introduzir modelos probabilísticos ou difusos no conjunto de modelos para a composição da matriz de resposta:
  - Elaboração de uma formulação para a matriz de resposta do CLAIRE, de tal forma que seja possível a introdução de modelos probabilísticos ou difusos na sua composição.
- Utilização do CLAIRE para diferentes tarefas não supervisionadas, por exemplo, detecção de anomalias não supervisionadas:
  - Utilizar da informação obtida a partir do processo do CLAIRE com respeito a instância do conjunto de dados, para verificar possíveis instâncias que possuem uma mudança de sinal do discriminante, podendo ser identificada como uma possível anomalia.
- Disponibilizar implementação em formato de pacote, além do pacote ja disponibilizado na linguagem Python [1], para a linguagem R.

[1] https://pypi.org/project/birt-gd/

Bibliografia

ANKERST, M. et al. Optics: Ordering points to identify the clustering structure. In: ACM. Proceedings of the ACM SIGMOD International Conference on Management of Data. [S.l.], 1999. p. 49–60.

BERNOULLI, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole. Mémoires de l'Académie Royale des Sciences de Paris, v. 1, p. 1–45, 1738.

CALINSKI, T.; HARABASZ, J. Dendrite method: a new graph-based heuristic for clustering. Communications in Statistics - Theory and Methods, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.

CHEN, Y. et al. β^3-irt: A new item response model and its applications. In: CHAUDHURI, K.; SUGIYAMA, M. (Ed.). Proceedings of Machine Learning Research. [S.l.: s.n.], 2019. (Proceedings of Machine Learning Research, v. 89), p. 1013–1021.

COMANICIU, D.; MEER, P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, v. 24, n. 5, p. 603–619, 2002.

DataNovia. Cluster Validation: Statistics Must-Know Methods. Acessado em 2023. Website. Disponível em: <https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#:~:text=The\%20term\%20cluster\%20validation\%20is,to\%20compare\%20two\%20clustering\%20algorithms.>

DAVIES, D.; BOULDIN, D. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, v. 1, n. 2, p. 224–227, 1979.

DEBORAH, L. J.; BASKARAN, R.; KANNAN, A. A survey on internal validity measure for cluster validation. International Journal of Computer Science & Engineering Survey, v. 1, n. 2, p. 85–102, 2010.

DHILLON, I. S.; GUAN, Y.; KULIS, B. Kernel k-means: Spectral clustering and normalized cuts. In: ACM. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.], 2004. p. 551–556.

EMBRETSON, S. E.; REISE, S. P. Item response theory: principles and applications. [S.l.]: Psychology Press, 2000.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, v. 96, n. 34, p. 226–231, 1996.

FERREIRA-JUNIOR, M. et al. β^4-IRT: A New β^3-IRT with Enhanced Discrimination Estimation. [S.l.]: arXiv, 2023.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity assessment: Finding the optimal partitioning of a data set. IEEE Transactions on Knowledge and Data Engineering, IEEE, v. 13, n. 4, p. 499–516, 2001.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. Journal of Intelligent Information Systems, v. 17, n. 2-3, p. 107–145, 2001.

HENNIG, C. Cluster Validation Methods in Computational Statistics. Acessado em 2023. Slide presentation. Disponível em: <http://www.homepages.ucl.ac.uk/~ucakche/presentations/compstatvalidation.pdf>.

HENNIG, C. et al. (Ed.). Handbook of Cluster Analysis. [S.l.]: Chapman and Hall/CRC, 2015.

HUBERT, L.; ARABIE, P. Comparing partitions. Journal of classification, Springer, v. 2, p. 193–218, 1985.

KIM, M.; RAMAKRISHNA, R. New indices for cluster validity assessment. Pattern Recognition Letters, v. 26, n. 15, p. 2353–2363, 2005. ISSN 0167-8655.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, v. 1, p. 281–297, 1967.

MARTÍNEZ-PLUMED, F. et al. Making sense of item response theory in machine learning. In: European Conference on Artificial Intelligence, ECAI. [S.l.: s.n.], 2016. p. 1140–1148.

MARTÍNEZ-PLUMED, F. et al. Item response theory in ai: Analysing machine learning classifiers at the instance level. Artificial Intelligence, Elsevier, v. 271, p. 18–42, 2019.

MCQUEEN, J. Some methods for classification and analysis of multivariate observations. In: Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. [S.l.: s.n.], 1967. p. 281–297.

Ministério da Educação (MEC).
Teoria de Resposta ao Item Avalia Habilidade e Minimiza o Chute.
21 de Março, 2023. Disponível em: <http://portal.mec.gov.br/ultimas-noticias/389-ensino-medio-2092297298/17319-teoria-de-resposta-ao-item-avalia-habilidade-e-minimiza-o-chute>.

MORAES, J. V. et al. Evaluating regression algorithms at the instance level using item response theory. Knowledge-Based Systems, v. 240, p. 108076, 2022. ISSN 0950-7051.

MORAES, J. V. C. et al. Item response theory for evaluating regression algorithms. In: 2020 International Joint Conference on Neural Networks (IJCNN). [S.l.: s.n.], 2020. p. 1–8.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems (NIPS). [S.l.: s.n.], 2002. v. 14, p. 849–856.

PEDREGOSA, F. et al. scikit-learn: Machine Learning in Python. 2011. Accessed on: 2023-03-31. Disponível em: <https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html>.

RASCH, G. Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, v. 1, n. 3, p. 1–20, 1960.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. Journal of Machine Learning Research, v. 8, n. Nov, p. 2937–2958, 2007.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, v. 20, p. 53 – 65, 1987. ISSN 0377-0427.

SPEARMAN, C. The proof and measurement of association between two things. The American Journal of Psychology, v. 15, n. 1, p. 72–101, 1904.

STREHL, A.; GHOSH, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, v. 3, n. Dec, p. 583–617, 2002.

WU, W. et al. Decision-making support for the evaluation of clustering algorithms based on mcdm. Complexity, Hindawi Limited, v. 2020, p. 1–17, 2020.