RODOLFO BOTTO DE BARROS GARCIA

APRENDIZADO SEMI-SUPERVISIONADO BASEADO EM BICLUSTERING EM BASES fMRI CEREBRAIS REDUZIDAS COM FOCO NO TDAH

CURITIBA-PR

2018

RODOLFO BOTTO DE BARROS GARCIA

APRENDIZADO SEMI-SUPERVISIONADO BASEADO EM BICLUSTERING EM BASES fMRI CEREBRAIS REDUZIDAS COM FOCO NO TDAH

apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de doutor em Informática.

Pontifícia Universidade Católica do Paraná - PUCPR

Programa de Pós-Graduação em Informática - PPGIa

Orientador: Prof. Dr. Júlio Cesar Nievola

Coorientador: Prof. Dr. Emerson Cabrera Paraiso

CURITIBA-PR

2018

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB 9 / 1636

G216a
2018

Garcia, Rodolfo Botto de Barros
Aprendizado semi-supervisionado baseado em biclustering em bases fMRI cerebrais reduzidas com foco no TDAH / Rodolfo Botto de Barros Garcia ; orientador, Júlio Cesar Nievola ; coorientador, Emerson Cabrera Paraiso. -- 2018
79 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2018.
Bibliografia: f. 75-79

1. Informática. 2. Imagem por ressonância magnética funcional. 3. Mapeamento do conectoma. 4. Mapeamento encefálico. 5. Transtorno do déficit de atenção com hiperatividade. I. Nievola, Júlio Cesar. II. Paraiso, Emerson Cabrera. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE SESSÃO PÙBLICA

DEFESA DE TESE DE DOUTORADO Nº 53/2018

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGIa
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ - PUCPR

Em sessão pública realizada às 09h00 de 23 de Março de 2018, no Auditório Guglielmo Marconi – Bloco 8, ocorreu a defesa da tese de doutorado intitulada **"Aprendizado Semi-Supervisionado Baseado em *Biclustering* Aplicado em Bases fMRI Cerebrais Reduzidas com Foco no TDAH** elaborada pelo aluno **Rodolfo Barros Botto Garcia**, como requisito parcial para a obtenção do título de **Doutor em Informática**, na área de concentração **Ciência da Computação,** perante a banca examinadora composta pelos seguintes membros:

Prof. Dr. Julio Cesar Nievola (orientador) - PUCPR

Prof. Dr. Emerson Cabrera Paraiso – PUCPR

Prof. Dr. Edson José Rodrigues Justino - PUCPR

Prof.ª Dr.ª Deborah Ribeiro Carvalho – PUCPR/PPGTS

Prof. Dr. Cassius Scarpin – UFPR

Após a apresentação da tese pelo aluno e correspondente arguição, a banca examinadora emitiu o seguinte parecer sobre a tese:

Membro	Parecer	
Prof. Dr. Julio Cesar Nievola	(X) Aprovada	( ) Reprovada
Prof. Dr. Emerson Cabrera Paraiso	(X) Aprovada	( ) Reprovada
Prof. Dr. Edson José Rodrigues Justino	(X) Aprovada	( ) Reprovada
Prof.ª Dr.ª Deborah Ribeiro Carvalho	(X) Aprovada	( ) Reprovada
Prof. Dr. Cassius Scarpin	(X) Aprovada	( ) Reprovada

Portanto, conforme as normas regimentais do PPGIa e da PUCPR, a tese foi considerada:

(X) **APROVADO**

(aprovação condicionada ao atendimento integral das correções e melhorias recomendadas pela banca examinadora, conforme anexo, dentro do prazo regimental)

( ) **REPROVADO**

E, para constar, lavrou-se a presente ata que vai assinada por todos os membros da banca examinadora.
Curitiba, 23 de Março de 2018.

Prof. Dr. Julio Cesar Nievola

Prof. Dr. Emerson Cabrera Paraiso

Prof. Dr. Edson José Rodrigues Justino

Prof.ª Dr.ª Deborah Ribeiro Carvalho

Prof. Dr. Cassius Scarpin

Programa de Pós-Graduação em Informática
PPGIa
PUCPR

Agradecimentos

Agradeço imensamente ao meu orientador Júlio Cesar Nievola e ao meu coorientador Emerson Cabrera Paraíso, primeiramente pelos ensinamentos durante todos esses anos e pela confiança depositada em mim e no trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, o CAPES, pelo suporte financeiro a este trabalho.

À minha mãe e ao meu irmão, por todo incentivo e apoio financeiro para que eu chegasse até aqui. À minha namorada, Franciele, por estar presente em todos os momentos felizes e muito tristes que surgiram no decorrer desta caminhada.

Aos queridos professores Andreia Malucelli e Edson José Rodrigues Justino por terem me aguentado todos os dias, aos professores Cassius Scarpin e Deborah Ribeiro Carvalho, demais membros da banca.

Aos amigos e companheiros de labuta que a PUCPR me deu: André Luiz, Cheila Cristina, Diogo Olsen, Estefânia Fuzyi, Flávio de Almeida, Gustavo Bonacina, Irapuru Florido, Jhonatan Geremias, Jurandir dos Santos, Luiz Giovanini (também como meu orientador de estágio em docência), Marcelo Pereira, Marcia Pascutti, Marcos Monteiro, Nicolas de Paula, Rodrigo Siega, Ronan Silva, Sandoval Ruppel, Sidnei Schuindt e Voncarlos Marcelo.

Aos amigos que fiz em Curitiba, à nova família que adquiri de Franciele e àqueles que me visitaram e compartilharam do agradável frio dessa cidade que já faz muita falta.

"Se vi mais longe
foi por estar de pé
sobre ombros de gigantes."
- Isaac Newton

Resumo

A partir do momento em que o TDAH foi considerado um transtorno que afeta o neurodesenvolvimento da pessoa, análises com fMRI vem sendo bastante usadas com o intuito de auxiliar futuramente em seu diagnóstico, diminuindo a probabilidade de condutas inapropriadas com relação a medicamentos e tratamentos. Com a evolução dos aparelhos de ressonância, a quantidade de imagens e seus detalhes aumentaram tanto a ponto de dificultar na rotulação manual de uma característica. Visando depender de poucas amostras rotuladas, a necessidade de auxílio computacional, como o uso de técnicas do aprendizado semi-supervisionado, se tornou indispensável. Além disso, o TDAH apresenta algumas particularidades que motivam aplicar técnicas de biclustering em análises de amostras obtidas pelo fMRI como: amostras incapazes de representar a classe a que pertence por presença de ruídos, falta de bases de dados com foco nesse transtorno e pouca informação útil vinda da maioria das regiões cerebrais. A fim de atender às particularidades supracitadas, este trabalho visa apresentar o SSBimax, um método de aprendizado semi-supervisionado baseado em biclustering, aplicado em bases fMRI com foco no TDAH. Essa combinação busca utilizar a quantidade mínima de amostras rotuladas suficiente para distinguir um portador do TDAH de uma pessoa com o desenvolvimento motor típico, auxiliada por amostras fMRI não rotuladas mas bem representativas e pelas regiões cerebrais que são afetadas no aparecimento do transtorno. Experimentos apresentando o poder de rotulação automática do SSBimax em quase todas as bases analisadas, reduzindo a quantidade de amostras previamente rotuladas sem diminuir o nível de qualidade, mostrou a vantagem de aplicar a estratégia de biclustering sobre outras estratégias utilizadas por métodos semi-supervisionados tradicionais como o S3VM, COP-Kmeans e a Minimização da Energia Harmônica, a partir do agrupamento simultâneo dos subconjuntos de atributos compartilhados com os subconjuntos das amostras mais representativas. Por fim, modificações pontuais no método clássico do Bimax fizeram do SSBimax um método mais robusto e que encontra biclusters mais significativos.

Palavras-chave: Aprendizado semi-supervisionado, biclustering, fMRI.

Abstract

Since ADHD was considered a disorder that affects the neurodevelopment of a individual, fMRI analyzes have been widely used to help in the diagnosis of this disorder, reducing the probability of inappropriate behavior regarding medications and treatments. With the evolution of the MRI machines, the amount of images and their details have increased so much that labeling manually one feature turned to be a hard task. In order to need few labeled samples, the need for computational assistance, such as the use of semi-supervised learning techniques, became indispensable. In addition, ADHD presents some peculiarities that motivate to apply biclustering techniques in analyzes of samples obtained by fMRI as: samples unable to represent the class to which it belongs due to presence of noise, the lack of databases focused on this disorder and little useful information coming from most of brain regions. In order to meet the above mentioned particularities, this work aims to present the SSBimax, a semi-supervised learning method based on biclustering, applied in fMRI databases focused on ADHD. This combination seeks to use the minimum amount of labeled samples which is sufficient to distinguish a person with ADHD from a person with typical motor development, aided by well-representative non-labeled fMRI samples and by the brain regions that are affected at activation of the disorder. Experiments presenting the automatic labeling power of SSBimax in almost all analyzed datasets, reducing the amount of previously labeled samples without decreasing the level of quality, showed the advantage of applying the biclustering strategy on other strategies used by traditional semi-supervised methods such as the S3VM, COP-Kmeans and Harmonic Energy Minimization, by the the simultaneous clustering of the subsets of features that shares subsets of representative samples. Finally, specific modifications in the classic Bimax method have made SSBimax a method more robust that found more significant biclusters.

Keywords: Semi-supervised learning, biclustering, fMRI.

Lista de ilustrações

Figura 1 – Relação de trabalhos relacionados à doenças psiquiátricas 16
Figura 2 – Passos para análise funcional do cérebro humano 19
Figura 3 – fMRI do cérebro humano 20
Figura 4 – Fronteira de decisão S3VM 26
Figura 5 – a) *Hinge Loss* e b) *Hat Loss* 27
Figura 6 – Iteração Bimax 38
Figura 7 – Etapas do projeto 44
Figura 8 – Relação amostras x atributos no SSBimax 51
Figura 9 – Média de atributos por experimento na base KKI 52
Figura 10 – Relação amostras x atributos no SSBimax 58
Figura 11 – Média de atributos por experimento na base NeuroIMAGE 59
Figura 12 – Relação amostras x atributos no SSBimax 65
Figura 13 – Média de atributos por experimento na base NYU 66

Lista de tabelas

Tabela 1 – Diferença entre agrupamento e *biclustering* 32
Tabela 2 – Quantidade de amostras nas bases de dados 42
Tabela 3 – Teste de Similaridade em Amostras Typ 42
Tabela 4 – Teste de similaridade em amostras TDAH 43
Tabela 5 – Resultados para base KKI original 49
Tabela 6 – Resultados da base KKI 49
Tabela 7 – Média de amostras rotuladas por biclustering 50
Tabela 8 – Resultados SVM para a base KKI 52
Tabela 9 – Resultados KNN para a base KKI 53
Tabela 10 – Resultados Perceptron para a base KKI 54
Tabela 11 – Resultados C4.5 para a base KKI 55
Tabela 12 – Resultados Naive Bayes para a base KKI 56
Tabela 13 – Resultados para base NeuroIMAGE original 56
Tabela 14 – Resultados da base NeuroIMAGE 57
Tabela 15 – Média de amostras rotuladas por biclustering 58
Tabela 16 – Resultados SVM para a base NeuroIMAGE 60
Tabela 17 – Resultados KNN para a base NeuroIMAGE 60
Tabela 18 – Resultados Perceptron para a base NeuroIMAGE 61
Tabela 19 – Resultados C4.5 para a base NeuroIMAGE 62
Tabela 20 – Resultados Naive Bayes para a base NeuroIMAGE 63
Tabela 21 – Resultados para base NYU original 64
Tabela 22 – Resultados da base NYU 64
Tabela 23 – Resultados SVM para a base NYU 66
Tabela 24 – Resultados KNN para a base NYU 67
Tabela 25 – Resultados Perceptron para a base NYU 68
Tabela 26 – Resultados C4.5 para a base NYU 69
Tabela 27 – Resultados Naive Bayes para a base NYU 70
Tabela 28 – Resultados para base Peking original 70
Tabela 29 – Resultados da base Peking 71

Lista de abreviaturas e siglas

TDAH	Transtorno de Déficit de Atenção e Hiperatividade
ADHD	Attention Deficit Hyperactivity Disorder
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
MRI	Magnetic Ressonance Imaging
fMRI	Functional MRI
rs-fMRI	Resting State fMRI
T-fMRI	Task-realted fMRI
BOLD	Blood Oxigenation Level Dependent
SSL	Semi-Supervised Learning
HCP	Human Connectome Project
SVM	Support Vector Machine
S3VM	Semi-Supervised SVM
TSVM	Transductive SVM
SSBimax	Semi-Supervised Bimax
LOOCV	Leave-one-out Cross Validation
PCC	Pearson Correlation Coefficient
KNN	K-Nearest Neighbors
RBF	Radial Bbasis Function
LAS	Large Average Submatrices
SAMBA	Statistical-Algorithmic Method for Bicluster Analysis

Sumário

1 INTRODUÇÃO 13
1.1 Trabalhos Relacionados 15
1.2 Conectoma Humano e fMRI 16
1.3 Hipóteses 19
1.4 Objetivos 19
1.5 Contribuição Científica 20
1.6 Contribuição Técnica 20
1.7 Estrutura do Trabalho 21

2 APRENDIZADO SEMI-SUPERVISIONADO 22
2.1 Aprendizado Supervisionado 23
2.2 Aprendizado Não-Supervisionado 23
2.3 Aprendizado Semi-Supervisionado 24
2.4 Métodos Semi-Supervisionado 25
2.4.1 S3VM 25
2.4.2 Minimização da Energia Harmônica 27
2.4.3 COP-KMeans 28
2.5 Considerações Finais 30

3 BICLUSTERING 31
3.1 Definição do Problema 32
3.2 Métodos Biclustering 33
3.2.1 Biclusters com Valores Constantes 34
3.2.2 Biclusters com Valores Constantes em Linhas e Colunas 34
3.2.3 Biclusters com Valores Coerentes 35
3.2.4 Biclusters com Evoluções Coerentes 36
3.3 SSBimax 36
3.4 Considerações Finais 39

4 EXPERIMENTOS 41
4.1 Bases de Dados 41
4.2 Detalhando os Experimentos 43
4.2.1 Coeficiente de Correlação de Pearson (PCC) 45
4.2.2 Configurações dos Métodos Semi-supervisionados 45
4.2.3 Configurações dos Classificadores 47

5 RESULTADOS EXPERIMENTAIS 48
5.1 Base KKI 49
5.1.1 Resultados SVM 51
5.1.2 Resultados KNN 53
5.1.3 Resultados Perceptron 54
5.1.4 Resultados C4.5 55
5.1.5 Resultados Naive Bayes 55
5.2 Base NeuroIMAGE 56
5.2.1 Resultados SVM 59
5.2.2 Resultados KNN 60
5.2.3 Resultados Perceptron 61
5.2.4 Resultados C4.5 62
5.2.5 Resultados Naive Bayes 63
5.3 Base NYU 63
5.3.1 Resultados SVM 65
5.3.2 Resultados KNN 67
5.3.3 Resultados Perceptron 68
5.3.4 Resultados C4.5 68
5.3.5 Resultados Naive Bayes 69
5.4 Base Peking 70

6 CONSIDERAÇÕES FINAIS 72

REFERÊNCIAS 75

Capítulo 1
Introdução

O TDAH, ou Transtorno de Déficit de Atenção e Hiperatividade, afeta 5,3% da população até 12 anos de idade, sendo o transtorno psiquiátrico mais comumente diagnosticado em crianças. Sintomas como falta de atenção, impulsividade e hiperatividade afetam o desenvolvimento cognitivo e dificultam a integração do indivíduo na sociedade, causando restrições sociais. Seu diagnóstico é tradicionalmente feito através de entrevistas com famílias ou professores sobre o comportamento do paciente, o que pode resultar em condutas inapropriadas com relação a medicamentos e tratamentos, principalmente com relação na especialização do tipo de TDAH (Desatento, Hiperativo-Impulsivo ou a combinação de ambos) (BANASCHEWSKI et al., 2015).

Por outro lado, o TDAH foi considerado um transtorno que afeta o neurodesenvolvimento da pessoa, segundo consta no DSM-5 (Manual de Diagnóstico e Estatística dos Transtornos Mentais da Associação Americana de Psiquiatria) (BANASCHEWSKI et al., 2015). Com isso, análise cerebral não invasiva usando imagens funcionais de ressonância magnética (fMRI) vem sendo bastante utilizada em estudos científicos com o intuito de poder auxiliar futuramente no diagnóstico de transtornos como o TDAH, o mal de Alzheimer e a Depressão (WOLFERS et al., 2015). Mais detalhes sobre a obtenção e as vantagens na utilização de fMRI estão na seção 1.2.

Com a constante evolução tecnológica, aparelhos de ressonância magnética conseguem tirar cada vez mais fotografias em um intervalo pequeno, aumentando tanto o detalhamento das imagens quanto a quantidade de dados analisadas. Porém, essa grande quantidade de dados dificulta a rotulação manual de uma característica por parte de especialistas, tornando a necessidade de auxílio computacional indispensável. Análises cerebrais demandam cada vez mais dos especialistas o domínio de diversas áreas como a celular, anatômica e computacional (AKIL; MARTONE; ESSEN, 2011). Estudos combinando abordagens computacionais, como o reconhecimento de padrões, e técnicas de MRI têm sido realizados para predição de doenças psiquiátricas (como pode ser visto na seção 1.1).

Na área da mineração de dados, em situação onde há conhecimento prévio sobre a rotulação de algumas amostras, viabilizando um aprendizado supervisionado, mas a praticidade de se obter uma quantidade de dados não classificadas é muito maior, encorajando um aprendizado não-supervisionado, é aconselhável o uso dos métodos semi-supervisionados (SSL - Semi-Supervised Learning). Dados não rotulados contêm menos informações que os dados rotulados mas, em grande quantidade, são capazes de trabalhar

em conjunto com técnicas supervisionadas na construção de modelos exatos e exigir menos esforço do que se fosse usada somente amostras rotuladas (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Além da dificuldade de obter grande quantidade de amostras fMRI rotuladas, diante de uma grande quantidade de imagens obtidas por essa técnica, o TDAH apresenta algumas particularidades que motivam a este trabalho aplicar técnicas do reconhecimento de padrões em análises de amostras obtidas pelo fMRI:

1. Nem todas as amostras são capazes de representar tão bem a classe a que pertence (portador do TDAH ou com desenvolvimento motor típico), em razão do TDAH se tratar de um transtorno em que a maior concentração de afetados sejam crianças, o tempo de concentração com a cabeça imóvel para quem tem o transtorno é considerado relevante e qualquer movimento afeta na qualidade das imagens, prejudicando na formação do modelo e, consequentemente, na classificação de novas amostras. Uma solução pode ser a identificação e uso apenas do conjunto de amostras que melhor representa a classe.

2. Apesar da popularidade do TDAH, revisões bibliográficas mostram que ainda há pouca exploração combinando esse transtorno e auxílio computacional comparado a outras doenças (mais detalhes sobre essas revisões na seção 1.1). Esse fato pode ser resultado da falta de bases com foco no transtorno, já que a base ADHD-200 é usada por todos os trabalhos da área.

3. Por fim, já é de conhecimento da literatura algumas, ainda poucas, regiões que são afetadas por causa do TDAH (ver na seção 1.1). Isso implica em pouca informação útil vinda das demais regiões cerebrais, além dos problemas causados pela chamada maldição de dimensionalidade.

Encontrar padrões locais é o objetivo dos métodos de agrupamento bi-direcional, ou biclustering, através de agrupamentos simultâneos nas amostras e em condições experimentais, a fim de encontrar os biclusters, que são subconjuntos de amostras que compartilham padrões de expressões similares sobre um subconjunto de condições (MADEIRA; OLIVEIRA, 2004). Em outras palavras, algoritmos biclustering podem identificar quais amostras fMRI são mais representativas ao TDAH e quais regiões cerebrais são afetadas por esse transtorno, amenizando a maldição da dimensionalidade.

A fim de atender às particularidades supracitadas, considerando a dificuldade de se obter amostras fMRI, este trabalho visa apresentar um método de aprendizado semi-supervisionado baseado em biclustering aplicado em bases fMRI com foco no TDAH. Essa combinação de técnicas procura utilizar a quantidade mínima de amostras rotuladas suficiente para distinguir um portador do TDAH de uma pessoa com o desenvolvimento

motor típico, auxiliada por amostras fMRI não rotuladas mas bem representativas e pelas regiões cerebrais que são afetadas no aparecimento do transtorno. A redução da quantidade de amostras rotuladas sem interferir na qualidade dos modelos gerados favorece na redução de tempo e custo por parte das análises manuais, contribuindo na aquisição de novos conhecimentos sobre o cérebro humano e principalmente sobre o processo de ativação desse transtorno que afeta tanta criança.

1.1 Trabalhos Relacionados

Este trabalho aborda o auxílio computacional em análise de bases MRI com foco no TDAH. O uso das técnicas MRI na predição de transtornos vem ganhando espaço desde que algumas doenças psiquiátricas podem ser caracterizadas pelo funcionamento desregulado de determinadas regiões cerebrais (ZENG et al., 2012), (LIM et al., 2013), (ZHU et al., 2017).

O TDAH, apesar de ser o transtorno mais comumente diagnosticado em crianças, não é tão explorado em trabalhos que utilizam auxílio computacional. Mwangi e colegas apresentaram em (MWANGI; TIAN; SOARES, 2014) uma revisão bibliográfica das técnicas de seleção de atributos em neuroimagens, onde reforçou a grande quantidade de aplicações em Alzheimer e o pouco uso em bases do TDAH.

Em (WOLFERS et al., 2015) foi feita uma pesquisa na plataforma PubMed[1] sobre trabalhos relacionados à doenças psiquiátricas. Foram considerados relevantes aqueles estudos que combinaram, entre outros requisitos, o uso de reconhecimento de padrões em imagens de Ressonância Magnética e apresentaram medidas de performance como resultados. Como pode ser visto na Figura 1, apenas 11 trabalhos relevantes abordando o TDAH foram incluídos, superior apenas a quantidade de trabalhos abordando a Ansiedade.

Este cenário não é pior pois, em 2011 foi lançado o ADHD-200 *Consortium* (MILHAM et al., 2012) e publicada a base ADHD-200[2], cuja finalidade foi de realizar um concurso para detecção de padrões relacionados ao TDAH, o que alavancou o interesse pelo estudo do transtorno com auxílios computacionais (LIANG et al., 2012), (SATO et al., 2012), (GARCIA; PARAISO; NIEVOLA, 2017).

Por outro lado, já é de conhecimento da literatura algumas regiões que são afetadas por causa do TDAH. Em (ZHU et al., 2008) são listadas algumas regiões encontradas por meio de estudos do MRI estrutural (Lobos Frontal, Parietal e Occipital, Gânglio Basal e Cerebelo) e obtidas em estudos por rs-fMRI (Córtex Cingulado Anterior, Lobos Frontal e Temporal, Cerebelo). Em (LIM et al., 2013) são listadas regiões relacionadas ao TDAH (Fronto-estriatal, Temporo-parietal e Fronto-cerebelar).

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] http:// fcon_1000.projects.nitrc.org/indi/adhd200/

Figura 1 – Relação de trabalhos relacionados à doenças psiquiátricas

Fonte: baseado em (WOLFERS et al., 2015)

Com a dificuldade de rotular muitas amostras MRI, técnicas de rotulação automática ganham espaço como soluções alternativas. A utilização de métodos semi-supervisionados para predição de doenças a partir de imagens de Ressonância Magnética do cérebro tem tido sucesso sobre o aprendizado supervisionado, em diferentes doenças: Alzheimer (MORADI et al., 2015),(FILIPOVYCH et al., 2011), Esquizofrenia, Depressão, Bipolaridade (BANSAL et al., 2012). O TDAH também é abordado em (BANSAL et al., 2012), tornando-o mais relevante por estudar esta doença tão pouco explorada.

Outra utilização do método semi-supervisionado é na descoberta de rede neurais que participam de uma determinada tarefa, como em (DU et al., 2014). Nele é proposto um método específico para cada indivíduo, a fim de identificar áreas de interesse contidas em imagens rs-fMRI. Este método SSL é baseado na teoria dos grafos e utiliza conhecimento anatômico prévio para localizar regiões iniciais de interesse. Nos resultados descritos, o método obtém vantagem sobre outros métodos comumente usados na identificação de redes cerebrais funcionais, como métodos de agrupamento.

Com relação à técnica de biclusteing, ela foi originalmente usada em análises de expressões gênicas. Um trabalho que utilizou biclustering em dados da neurociência foi (BUSYGIN et al., 2007), em que combinando com seleção de características supervisionada para análises de EEG, buscou realizar uma otimização de parâmetros para controle de convulsões epilépticas. Essa otimização objetivou encontrar a melhor configuração de variáveis na estimulação elétrica para reduzir custo, tempo e risco no tratamento contra epilepsia (exemplos de variáveis usadas: intensidade, duração e frequência da estimulação).

1.2 Conectoma Humano e fMRI

O cérebro, por ser um órgão altamente complexo e conter uma grande quantidade de elementos distintos, heterogêneos e interconectados, não nos permitiu ainda obter muito

conhecimento sobre a dinâmica da rede de neurônios que gera funcionalidades como nossos sentidos, memória e emoções.

Para conhecer o funcionamento do nosso cérebro, saber como suas regiões se comunicam e como se originam as funcionalidades, é necessário um modelo detalhado e compreensível dos elementos neurais e suas conexões. A esse modelo dá-se o nome de conectoma humano. O mapeamento do conectoma humano possibilita (SPORNS; TONONI; KÖTTER, 2005):

1. Aumentar nosso conhecimento de como os estados funcionais do cérebro emergem de seus substratos estruturais;

2. Prover compreensão de como uma função cerebral é afetada se seu substrato estrutural for corrompido. Dessa forma, o conectoma humano terá um grande impacto no nosso entendimento de lesão cerebral, doenças degenerativas e recuperação. Além disso, aumentará o nosso conhecimento sobre os efeitos da variação de desenvolvimento ou anormalidades individuais (assim como ocorre também no genoma humano), que é um dos maiores desafios para a análise do conectoma. O alto nível de variabilidade individual dificulta na descoberta de padrões na conectividade cerebral humana, tornando o conectoma muito menos exato que o genoma (ESSEN; UGURBIL, 2012);

3. Permitir a criação de estratégias de recuperação, novos tipos de terapias e novas formas de prevenção contra doenças;

4. Prover fontes de dados unificadas e legíveis disponíveis na neuroinformática, que podem ser utilizadas em todas as áreas da neurociência experimental e teórica, a fim de melhorar os modelos do cérebro humano.

Um dos problemas do conectoma é a grande quantidade de elementos estruturais e conexões entre eles. Diferentemente do genoma humano, em que os genes foram facilmente escolhidos como elementos estruturais, o conectoma pode considerar três níveis de detalhamento (microescalar, mesoescalar e macroescalar), cada um com elementos estruturais diferentes. O nível macroescalar é o favorito a ser usado nessa fase inicial de análise cerebral, como sugerido em (ESSEN; UGURBIL, 2012), pois é formado por menos detalhes e adota as regiões cerebrais como elementos estruturais, tendo aproximadamente 10^5 elementos estruturais e 10^7 conexões. Um mapeamento confiável dos elementos estruturais facilita na realização do objetivo principal do conectoma, que é de mostrar que a estrutura anatômica do cérebro pode servir de base para um maior entendimento das dinâmicas cerebrais e do comportamento humano (BEHRENS; SPORNS, 2012).

Quando se trata de técnicas para mapeamento de conexões em larga escala no cérebro podemos enfatizar as não invasivas. Apesar de serem baseadas em inferência e

estarem sujeitas a erro, sua natureza não invasiva e a facilidade de medição possibilitam a obtenção de respostas que não são possíveis por outros meios. Essas técnicas permitem a observação da anatomia de conexões por todo o cérebro, de forma simultânea, em seres ainda vivos, facilitando não somente a criação do conectoma, como também análises de estudos comparativos das áreas cerebrais entre humanos vivos (o que ajuda a responder algumas questões sobre a variabilidade individual) e investigação da importância da arquitetura estrutural para o processamento de funcionalidades (BEHRENS; SPORNS, 2012), (SPORNS, 2011). Técnicas para análise não invasiva da atividade funcional no cérebro que se destacam são as que usam Imagens de Ressonância Magnética, ou fMRI (*functional* MRI). No fMRI, dois métodos são bastante utilizados: o *Task-related* fMRI (T-fMRI), em que são verificadas as oscilações de frequência ao acontecer uma tarefa (por exemplo, quais regiões são ativadas ao visualizar uma fotografia), e o *resting state* fMRI (rs-fMRI), que utiliza das oscilações de baixa frequência exibidas pela atividade cerebral no momento de descanso para inferir conectividade funcional entre regiões que empenham oscilações coerentes (ESSEN et al., 2012).

Outra vantagem desses métodos é de utilizar as propriedades magnéticas do sangue para servir também como contraste. A frequência é medida pelo efeito BOLD (*Blood Oxigenation Level Dependent*), como resultado para a variação do fluxo sanguíneo no cérebro (alto fluxo na realização de atividades e baixo fluxo no descanso) (WESTBROOK; ROTH, 2011), (HEUVEL; POL, 2010).

Na Figura 2 são apresentados os passos realizados para possibilitar as análises funcionais do cérebro. No passo 1 são selecionadas regiões do cérebro para gravação funcional. A partir dessas regiões de referência, estudos com fMRI focam em medir a relação de ativação entre diferentes regiões cerebrais. No passo 2, o histórico funcional baseado no efeito BOLD entre regiões do cérebro são gravadas (Figura 3 retrata uma imagem fMRI com uma região referência contendo a *seed voxel*, que é um ponto representando uma determinada localização dentro de uma região cerebral, e uma região altamente correlacionada com ela). O passo 3 consiste na construção das matrizes de conectividade funcional para, enfim, realizar análises relacionando a atividade funcional ao mapeamento estrutural para uma determinada funcionalidade. Uma matriz de conectividade cerebral, como é descrita em (BROWN et al., 2012), corresponde aos dados de uma neuroimagem processada que armazenam níveis de conexão entre regiões de uma rede cerebral e que mantêm informações suficientes para capturar características pessoais como idade e estados de uma doença.

A grande evolução tecnológica dos métodos de neuroimagens motivou a criação do Projeto Conectoma Humano[3] (HCP), que vem realizando o mapeamento do conectoma humano por meio de técnicas estruturais e funcionais de MRI aplicadas em 1200 adultos

[3] http://humanconnectome.org/

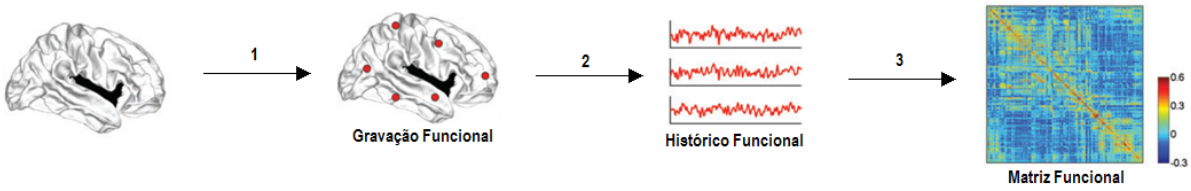
saudáveis. Com o HCP, grandes progressos estão sendo esperados em várias frentes metodológicas (por exemplo, melhorias nos métodos não invasivos de neuroimagens), combinados com a aquisição de grande quantidade de dados (ESSEN; UGURBIL, 2012).

1.3 Hipóteses

Este trabalho levanta duas hipóteses. Como já visto nos trabalhos relacionados (seção 1.1), o aprendizado semi-supervisionado leva vantagem sobre outras técnicas comumente usadas para rotulação na diferenciação de doenças como o Mal de Alzheimer. A primeira hipótese aqui levantada é que o SSL também seja capaz de trabalhar na rotulação de amostras TDAH e de pessoas com desenvolvimento motor típico, considerando quantidades reduzidas de amostras rotuladas sem interferir na qualidade dos modelos construídos.

A segunda hipótese é de que essa rotulação automática utilizando técnicas do biclustering seja superior às técnicas tradicionais SSL, uma vez que a exclusão de amostras não representativas e a seleção de atributos natural resultem em submatrizes somente com informações úteis, gerando modelos com maior poder de diferenciação entre amostras de classes diferentes em bases com rotulação reduzidas.

Figura 2 – Passos para análise funcional do cérebro humano



Fonte: baseado em (SPORNS, 2011)

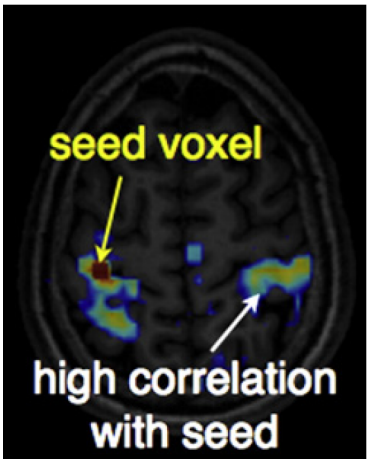
1.4 Objetivos

Este trabalho tem como objetivo principal apresentar um novo método de aprendizado semi-supervisionado baseado em biclustering, aplicado em bases fMRI reduzidas com foco no TDAH. Os objetivos específicos englobados são:

1. Adaptação das bases formadas por matrizes de conectividade com foco no TDAH e obtidas pelas técnicas de fMRI, originalmente publicadas na ADHD-200;

2. Diminuição da porcentagem de amostras rotuladas nas bases adaptadas, tornando-as reduzidas;

Figura 3 – fMRI do cérebro humano



Fonte: (HEUVEL; POL, 2010)

3. Definição da estratégia de rotulação e seleção das regiões por meio da abordagem de biclustering;

4. Avaliação e comparação do desempenho do método apresentado neste trabalho com diferentes métodos de aprendizado semi-supervisionado disponíveis na literatura.

1.5 Contribuição Científica

As comparações realizadas com diversos métodos semi-supervisionados contribuem para mostrar como diferentes modelos se comportam na geração de modelos a partir de bases com porcentagem de rotulação reduzidas. A combinação do aprendizado semi-supervisionado com a técnica bicluster, diferente das técnicas tradicionais, contribui na geração de modelos utilizando quantidades reduzidas de amostras rotuladas, auxiliadas por amostras não rotuladas representativas e atributos relevantes, o que ameniza os problemas causados pela maldição da dimensionalidade e aumenta o poder de diferenciação entre amostras de classes diferentes.

1.6 Contribuição Técnica

O desenvolvimento de um novo método de aprendizado semi-supervisionado baseado em biclustering visa primeiramente o uso de poucas amostras fMRI rotuladas, levando em consideração as dificuldades na obtenção dessas amostras, principalmente de pessoas com TDAH. Dessa forma, esse método contribui na redução de tempo e custo por parte das análises manuais por especialistas, favorecendo na aquisição de novos conhecimentos sobre o cérebro humano, não somente com relação ao TDAH, e no auxílio futuramente no diagnóstico de doenças neurológicas, assim como em tratamentos.

1.7 Estrutura do Trabalho

Este trabalho é dividido em 6 capítulos. No capítulo 2 é apresentado o aprendizado semi-supervisionado juntamente com suas definições e métodos que serão comparados ao método apresentado neste trabalho. O capítulo 3 apresenta a técnica de biclustering como uma alternativa para descobertas de padrões locais, assim como a descrição do novo método semi-supervisionado baseado em biclustering. Os experimentos, suas etapas realizadas e as bases de dados utilizadas estão no capítulo 4. Em seguida, no capítulo 5, são descritos os resultados dos experimentos. Por fim, no capítulo 6, algumas considerações finais sobre este trabalho são feitas e alguns trabalhos futuros são propostos.

Capítulo 2
Aprendizado Semi-supervisionado

Como visto na Introdução deste trabalho, rotular manualmente bases de imagens funcionais de Ressonância Magnética do cérebro é uma tarefa muito custosa por conta da grande quantidade de imagens. Como consequência são geradas bases com poucas amostras rotuladas representadas por uma grande quantidade de correlações entre regiões cerebrais, como exemplo da ADHD-200. Este cenário afeta diretamente na construção de modelos confiáveis para predição de doenças e, nessas situações em que há conhecimento prévio sobre a classe de algumas amostras, mas a praticidade de se obter amostras não rotuladas é muito maior, é aconselhável o uso de métodos semi-supervisionados (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Dados não rotulados contêm menos informações que os dados rotulados mas, em grande quantidade, são capazes de trabalhar em conjunto com técnicas supervisionadas na construção de modelos exatos e exigir menos esforço do que se fosse usada somente amostras rotuladas. O aprendizado semi-supervisionado (SSL - Semi-Supervised Learning) é uma combinação entre as aprendizagens supervisionada e não-supervisionada. Pelo fato de que poucas amostras rotuladas não são capazes de gerar modelos que caracterizam bem uma classe, a adição de algumas amostras não rotuladas pode guiar as amostras com rótulos na geração de grupos maiores para cada classe, e assim aumentar o nível de exatidão na predição destas (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Neste trabalho serão comparados métodos de aprendizado semi-supervisionado pertencentes à diferentes conceitos de rotulação, alguns populares na literatura e outro apresentado mais adiante, com objetivo de verificar seus comportamentos quando aplicados em bases fMRI com rotulações reduzidas. A finalidade dessa comparação é de reduzir à quantidade mínima de amostras rotuladas, sem comprometer a confiabilidade do modelo construído após a rotulação de novas amostras.

Este capítulo está organizado da seguinte forma: as seções 2.1 e 2.2 apresentam definições básicas do aprendizado supervisionado e não supervisionado, respectivamente. Na seção 2.3, a descrição do problema semi-supervisionado será vista. Os métodos SSL que serão comparados ao método proposto neste trabalho estão na seção 2.4 e, por fim, na seção 2.5 estão as considerações finais deste capítulo.

2.1 Aprendizado Supervisionado

O ato de conhecer a que classe (rótulo) pertence uma amostra caracteriza um processo de aprendizado supervisionado. No aprendizado supervisionado, as amostras rotuladas são submetidas à seguinte etapa, chamada de treinamento: sejam amostras dentro de um conjunto $X_i = x_1, ..., x_i$ cujas classes são conhecidas e estão contidas no vetor $Y = y_1, ..., y_i$. Nessa etapa, uma função $f : X \rightarrow Y$ é treinada com o objetivo de que $f(x)$ seja um modelo capaz de representar todas as amostras rotuladas e de predizer a verdadeira classe em futuros dados x não rotulados.

Uma função f é considerada boa se, ao aplicar as amostras rotuladas de treinamento, $f(x)$ minimize o erro de predição.

Se as classes em Y representam valores discretos, então o problema supervisionado é uma classificação e a função f é chamada de classificador. Caso as classes em Y sejam valores contínuos, o problema supervisionado é de regressão e f é chamada função regressão.

2.2 Aprendizado Não-Supervisionado

Contrário ao aprendizado supervisionado, a ausência do rótulo nas amostras de uma base impossibilita o treinamento de um modelo direcionado. Desta forma, o modelo tem que ser formado pela semelhança do conjunto de características, sendo então um processo de aprendizado não-supervisionado. A técnica não-supervisionada mais comum é o agrupamento, que tende a organizar um conjunto de objetos em grupos, de forma que aqueles de comportamentos mais parecidos fiquem no mesmo grupo, com a finalidade de revelar a estrutura natural das bases de dados (XU; WUNSCH, 2005), (DY, 2008).

Dada a matriz $A^{n \times m} = (X, Y)$ com o conjunto de linhas $X = x_1, ..., x_n$ e o conjunto de colunas $Y = y_1, ..., y_m$, um agrupamento C objetiva construir subconjuntos de C em que, obedecendo a um critério de similaridade aplicado em Y, as linhas de um subconjunto sejam mais similares entre eles do que as linhas pertencentes a outros subconjuntos. Cada subconjunto de linhas similares é chamado de grupo C_i e todos os subconjuntos recebem do nome de agrupamento $C = C_1, ..., C_k$, com k grupos.

Uma partição equivale ao conjunto de grupos de forma que:

1. $C_i \cap C_j = \emptyset$, com $i, j \in [1, k], i \neq j$
2. $C_1 \cup C_2 \cup ... \cup C_k = X$

A depender da quantidade de partições geradas pelo agrupamento, é possível classificar um método em particional (única partição) ou hierárquico (várias partições).

Um bom agrupamento ocorre quando os atributos que formam as bases de dados (conjunto Y) são relevantes e suas amostras (conjunto X) definem bem os grupos aos quais pertencem.

2.3 Aprendizado Semi-Supervisionado

Seja o conjunto de linhas $X = x_1, ..., x_n$ independentes e distribuídos identicamente. No aprendizado semi-supervisionado, as n amostras são divididas em dois subgrupos $X_l = x_1, ..., x_l$, correspondendo ao seu conjunto de rótulos $Y = y_i, ..., y_l$, e $X_u = x_{l+1}, ..., x_{l+u}$ sem nenhum conjunto de rótulos relacionado, com $l + u = n$. O objetivo do aprendizado semi-supervisionado, assim como do aprendizado supervisionado, é de construir um mapeamento de x para y, a partir de um conjunto de treinamento formado pelos pares (x_i, y_i), a fim de minimizar erros de classificação (SEEGER, 2006).

O aprendizado semi-supervisionado tem vantagem sobre o aprendizado supervisionado, caso o conhecimento adquirido em $p(x)$ por dados não rotulados contenha informação útil na inferência de $p(y|x)$.

Para que o aprendizado semi-supervisionado funcione, é necessário impor algumas restrições que são descritas abaixo (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

1. *Semi-supervised smoothness assumption*: Se dois pontos x_1 e x_2 são próximos dentro de uma região de alta densidade, então eles correspondem às classes y_1 e y_2. Isso quer dizer que uma classe é melhor definida em regiões de alta densidade, do que em regiões de baixa densidade. Consequentemente, se dois pontos estão na mesma região de alta densidade, então suas classes são próximas;

2. Existência de grupos: Se dois pontos estão no mesmo grupo, então eles fazem parte da mesma classe. Usando os dados não rotulados para modelar a borda de cada grupo, os dados rotulados podem ser usados para associar cada grupo a uma classe. Essa restrição é usada como base em métodos de agrupamento semi-supervisionado;

3. Separação em densidades baixas: A fronteira de um grupo é caracterizada por uma região de baixa densidade. A borda de um grupo em uma região de alta densidade pode resultar na divisão de uma classe em dois grupos distintos;

4. *Manifold assumption*: Dados com grande dimensionalidade concentram suas informações relevantes em poucos atributos. O crescimento exponencial de dimensões e do número de amostras requer tarefas estatísticas para estimativas de regiões densas. Se os dados se concentram em poucas dimensões, o algoritmo de aprendizagem pode operar em um espaço reduzido.

2.4 Métodos Semi-Supervisionado

Os métodos SSL podem atuar em dois cenários possíveis: o cenário indutivo e o transdutivo. No cenário indutivo, o classificador é treinado de forma que ele seja capaz de predizer corretamente as classes de dados futuros, além das amostras não rotuladas de treino. Imaginando um exame aplicado em sala de aula, o aprendizado indutivo faz com que o aluno aprenda e se prepare para todos os tipos de questões, sem ter conhecimento das questões presentes no exame (ZHU; GOLDBERG, 2009). Já no cenário transdutivo, o classificador é treinado de forma que ele seja capaz de predizer corretamente as classes das amostras não rotuladas na etapa de treinamento. É considerada uma função mais simples por não precisar realizar predição fora das amostras de treino. Imaginando um exame que o aluno pode realizar levando para casa, este aluno não somente conhece as questões que estão no exame, como também só irá estudar e aprender sobre estas questões (ZHU; GOLDBERG, 2009).

Da mesma forma, um método de aprendizado semi-supervisionado pode ser visto como um problema de aprendizado supervisionado com informações adicionais não-supervisionadas na distribuição dos dados, cujo objetivo é a predição de uma classe a cada amostra. Outra forma, o aprendizado semi-supervisionado pode ser visto como um aprendizado não-supervisionado guiado por constantes, sugerido em casos onde não se conhecem profundamente sobre a natureza das classes (CHAPELLE; SCHOLKOPF; ZIEN, 2006), (ZHU; GOLDBERG, 2009). Como o objetivo principal deste trabalho é apresentar um novo método de aprendizado semi-supervisionado, outros métodos já conhecidos da literatura que abordam esses dois pontos de vista e que obtiveram sucesso em seus experimentos foram comparados.

A seguir serão vistos tanto métodos transdutivos com base em aprendizado supervisionado quanto um método indutivo que utiliza o aprendizado não-supervisionado guiado por constantes como base. Primeiramente, o S3VM já vem sendo utilizado em estudos de predição de doenças psiquiátricas a partir de neuroimagens (MORADI et al., 2015), (FILIPOVYCH et al., 2011); em seguida o método de minimização de energia harmônica representando a teoria de grafos, que possui uma área muito ativa no aprendizado semi-supervisionado (CHAPELLE; SCHOLKOPF; ZIEN, 2006); por fim, o COP-Kmeans como representante do aprendizado semi-supervisionado baseado em agrupamento (WAGSTAFF et al., 2001).

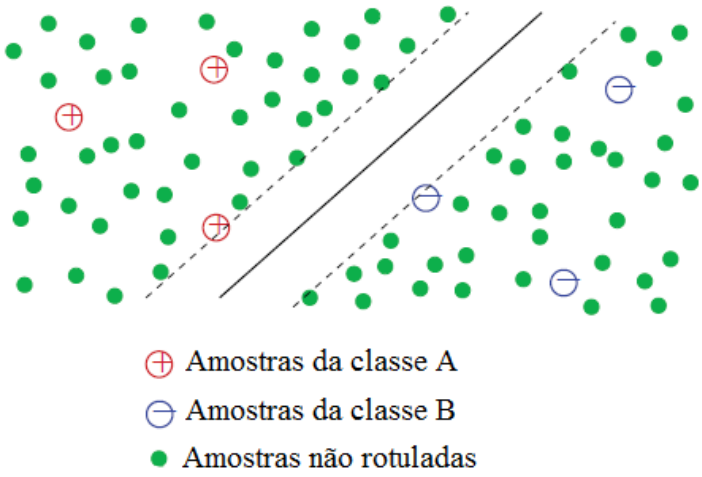
2.4.1 S3VM

O método S3VM (Semi-Supervised Support Vector Machines) foi originalmente chamado de SVM Transdutivo, TSVM (VAPNIK, 1998), pois foi desenvolvido para garantir performance às amostras não rotuladas de treinamento. A mudança de nome se deve ao

fato da função treinada ser naturalmente aplicada às amostras de teste.

Na versão original do SVM, uma função $f(x_i)$ é treinada para cada amostra rotulada x_i composta por m dimensões. Essa função visa construir uma linha reta chamada fronteira de decisão (representada pela linha contínua da Figura 4) e suas margens (representadas pelas linhas tracejadas da Figura 4), a fim de separar amostras de classes Y iguais no espaço de atributos. A fronteira de decisão é definida por $w^T x + b = 0$, em que $w^T \in R^m$ é o vetor que especifica sua orientação e escala, e $b \in R$ é o *offset*.

Figura 4 – Fronteira de decisão S3VM



Fonte: baseado em (ZHU; GOLDBERG, 2009)

A função objetivo 2.1 busca minimizar a chamada *Hinge Loss Function* (Equação 2.2) e a função reguladora $\Omega(f)$ (Equação 2.3) usando o peso λ que balanceia os dois objetivos. Consequentemente, o classificador $f(x_i)$ maximiza a distância entre amostras de classes distintas.

$$min_{w,b} \sum_{i=1}^{l} c(x_i, y_i, f(x_i)) + \lambda\Omega(f) \tag{2.1}$$

$$c(x_i, y_i, f(x_i)) = max(1 - y_i(w^T x_i + b), 0) \tag{2.2}$$

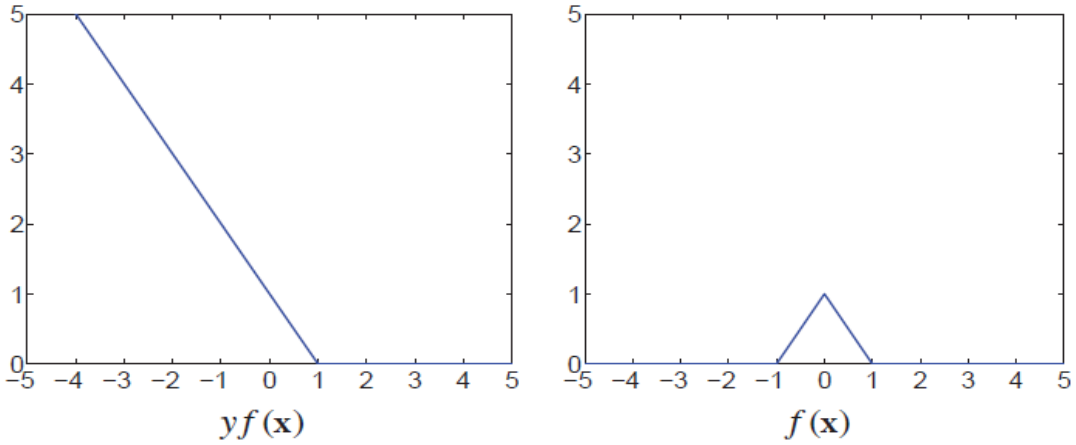
$$\Omega(f) = \|w\|^2 \tag{2.3}$$

O rótulo estimado para um ponto não rotulado x_i é $\hat{y}_i = sign(f(x_i))$. Incorporando x_i na função *hinge loss* (Figura 5a) do SVM é gerada a função *hat loss* (Equação 2.4).

$$c(x_i, \hat{y}_i, f(x_i)) = max(1 - sign(w^T x_i + b)(w^T x_i + b), 0) \tag{2.4}$$

Os valores preferíveis são aqueles que estiverem além das extremidades do *hat*, ou seja, $f(x) \geq 1$ ou $f(x) \leq -1$ (Figura 5b), pois são pontos longe da margem de separação de classes. Equivalentemente, a fronteira de decisão precisa estar em uma região de baixa densidade (considera o pressuposto de existência de grupos).

Figura 5 – a) *Hinge Loss* e b) *Hat Loss*



Fonte: baseado em (ZHU; GOLDBERG, 2009)

Incorporando o *hat loss* das amostras não rotuladas na função objetivo do SVM e definindo $sign(z) = |z|$, obtemos a função objetivo do S3VM (Equação 2.5):

$$min_{w,b} \sum_{i=1}^{l} max(1 - y_1(w^T x_i + b), 0) + \lambda_1 \|w\|^2 + \lambda_2 \sum_{i=l+1}^{l+u} max(1 - |w^T x_j + b|, 0) \quad (2.5)$$

A Equação 2.5 é considerada desbalanceada porque a predição de quase todas as amostras não rotuladas indica a mesma classe. Para isso, uma heurística obriga todos os dados não rotulados a realizar uma predição proporcional de classe, que é igual para os dados rotulados (Equação 2.6).

$$\frac{1}{u} \sum_{j=l+1}^{l+u} w^T x_i + b = \frac{1}{l} \sum_{i=1}^{l} y_i \quad (2.6)$$

2.4.2 Minimização da Energia Harmônica

Este método foi proposto por (ZHU; GHAHRAMANI; LAFFERTY, 2003) e se trata de um método de aprendizado semi-supervisionado baseado na teoria de grafos. O aprendizado é formulado como um problema de campos aleatórios de Gauss no grafo, cujo o meio do campo é descrito em termos de uma função harmônica.

Seja um grafo com pesos $G = (V, E)$, em que os vértices V são as l amostras rotuladas $(x_1, y_1), ..., (x_l, y_l)$ e u amostras não rotuladas $x_{l+1}, ..., x_{l+u}$ (com $l \ll u$ e $n =$

$l + u$). O objetivo é atribuir rótulos às amostras em u dentro do conjunto binário $y \in 0, 1$. O peso de cada aresta E representa a similaridade entre duas amostras na matriz simétrica W, de tamanho $n \times n$ (Equação 2.7).

$$w_{ij} = exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2} \right) \tag{2.7}$$

Na Equação 2.7, x_{id} é o d-ésimo atributo da amostra $x_i \in R^m$ e os $\sigma_1, ..., \sigma_m$ são os tamanhos das escalas dos hiperparâmetros de cada atributo. Amostras próximas no espaço Euclidiano têm alto peso de aresta.

Zhu e colegas dividiram a estratégia em duas etapas: primeiro é computada uma função $f : V \rightarrow R$ em G e depois atribuir rótulos baseados em f. As amostras rotuladas recebem $f(i) = y_i$, $i = 1, ..., l$ e é esperado que as amostras não rotuladas recebam rótulos iguais das amostras próximas, a partir da função de energia quadrática (Equação 2.8).

$$E(f) = \frac{1}{2} \sum_{i=j} w_{ij}(f(i) - f(j))^2 \tag{2.8}$$

Uma vez que a função de distribuição de probabilidade é formada por um campo Gaussiano e a função quadrática mínima $f = argmin_{f|L=y} E(f)$ seja harmônica, ela tem a propriedade de que o valor de f em cada amostra não rotulada é a média dos valores de f nas amostras vizinhas, como mostrada na Equação 2.9, para $j \in Ue0 < f(j) < 1$.

$$f(j) = \frac{1}{d_j} \sum_{ij} w_{ij} f(i) \tag{2.9}$$

A rotulação de amostras em casos de classes bem separadas se dá por: se $f(i) > \frac{1}{2}$, $y_i = 1$, caso contrário, $y_i = 0$. Quando as classes não são bem separadas, para evitar classificação desbalanceada, estima-se a proporção de classes dos dados rotulados e sua normalização de massa. Sabendo que a proporção da classe 1 seja q e a da classe 0 seja $1 - q$, suas massas são dadas como $\sum_i f_u(i)$ e $\sum_i (1 - f_u(i))$, respectivamente. Desta forma, uma amostra rotulada i é classificada como classe 1 se, e somente se, satisfizer a Equação 2.10.

$$q \frac{f_u(i)}{\sum_i f_u(i)} > (1 - q) \frac{1 - f_u(i)}{\sum_i (1 - f_u(i))} \tag{2.10}$$

2.4.3 COP-KMeans

Como dito anteriormente, o aprendizado semi-supervisionado pode ser visto como um problema do aprendizado não-supervisionado guiado por restrições. Estes casos são

chamados de agrupamento semi-supervisionado (semi-supervised clustering) (FATEHI et al., 2014).

Um método de agrupamento muito popular para versões semi-supervisionadas é o K-means (FATEHI et al., 2014), (WAGSTAFF et al., 2001). Neste trabalho será abordado o COP-Kmeans por ter sido a versão que melhor se comportou dentre as demais nos experimentos que serão vistos mais adiante.

O funcionamento do K-means tem início ao selecionar k centroides. O valor k pode ser imposto pelo usuário e os centroides podem ser amostras escolhidas de forma aleatória na base de dados. Na sequência, o K-means executa dois passos até que se atinja o critério de parada: o primeiro passo é encaixar cada amostra da base no grupo cujo centroide está mais próximo, com base em uma função de distância (por exemplo, a distância Euclidiana). Após encaixar todas as amostras, o segundo passo se dá pela atualização dos centroides de cada grupo. O critério de parada pode ser o número de iterações ou a não modificação dos grupos entre as iterações.

Proposto por (WAGSTAFF et al., 2001) e representado pelo Algoritmo 1, o COP-Kmeans é baseado por algumas restrições responsáveis por guiar as amostras não rotuladas durante a execução do aprendizado não-supervisionado. Essas restrições são extraídas das amostras rotuladas e podem ser de 2 tipos: *must-link*, que indica quando duas amostras devem estar no mesmo grupo e *cannot-link*, que indica quando duas amostras não podem pertencer ao mesmo grupo.

Após a seleção dos centroides iniciais de forma tradicional, a única diferença entre o COP-Kmeans e o K-means clássico é que cada amostra será encaixada no grupo mais próximo sem violar nenhuma das restrições aplicadas. Após alcançar o critério de parada, o agrupamento com todas as amostras rotuladas é retornado.

Algoritmo 1: COP-KMeans

Entrada : amostras rotuladas X_l, amostras não rotuladas X_u, restrições *must-link*, restrições *cannot-link*
Saída : $X_l + X_u$ com rótulos

```
while parada = falso do
    Selecionar k centroides iniciais C_1, ..., C_k
    for d in X_l + X_u do
        Achar centroide mais próximo sem violar restrição
        if Não achar centroide then
            return vazio
        end
    end
    Atualizar centroides
end
```

Em outras versões semi-supervisionadas do K-means, (BASU; BANERJEE; MOO-

NEY, 2002) propõe o Seeded-Kmeans, em que as amostras rotuladas são usadas tanto para formar os centroides iniciais quanto para compor o agrupamento; e o Constrained-KMeans, cuja diferença é que as amostras rotuladas são excluídas do agrupamento após formação dos centroides iniciais. Em (FATEHI et al., 2014), uma evolução do Constrained-KMeans é proposta, em que os dados a serem rotulados são escolhidos por um pré-processamento realizado pelo K-means tradicional.

2.5 Considerações Finais

Neste capítulo foi apresentada a técnica de aprendizado semi-supervisionado. Foram vistos métodos pertencentes à diferentes abordagens, se comportando tanto como problemas supervisionados quanto como problemas não-supervisionados. Com isso, a rotulação de poucas amostras diminui o custo requerido por parte de especialistas e, ao mesmo tempo, a rotulação automática das demais amostras pode diminuir as chances de um diagnóstico errado.

Por outro lado, a quantidade de atributos em uma base fMRI é muito maior que o número de amostras. Porém, sabe-se que o número de regiões que são afetadas pelo TDAH é pequeno e necessita de uma redução de dimensionalidade para obter modelos mais representativos.

Por fim, nem todas as amostras não rotuladas são capazes de representar tão bem uma classe (portador do TDAH ou com desenvolvimento motor típico) pois, além de se tratar de crianças na maioria dos casos, o tempo de concentração para manter a cabeça imóvel para quem tem o transtorno é considerado relevante e qualquer movimento afeta na qualidade das imagens. Isso prejudica na formação do modelo e, consequentemente, na classificação de novas amostras.

Nos casos em que é necessário utilizar apenas as amostras que representam bem o problema, simultaneamente direcionando ao subconjunto de atributos de interesse, deve-se usar a técnica de biclustering, visto no capítulo a seguir.

Capítulo 3
Biclustering

Em bases de fMRI cerebrais a quantidade de expressões entre regiões nas matrizes de conectividade é muito maior que a quantidade de amostras que as compõem. Como é de conhecimento, poucas são as regiões que são afetadas pelo TDAH e o excesso de informação que não ajuda na construção do modelo resulta na maldição de dimensionalidade. Na maldição de dimensionalidade, a quantidade de atributos é tão superior ao número de amostras que a construção de modelos requer alto custo computacional e decrementa a acurácia de classificadores por dificultar a distinção entre diferentes classes, necessitando diminuir a dimensionalidade da base (JOHNSTON et al., 2014), (MWANGI; TIAN; SOARES, 2014).

Outro problema é que as amostras não rotuladas podem não ser representativas para nenhuma classe, muitas vezes por causa dos ruídos gerados por movimentos da cabeça na aquisição da fMRI. Desta forma, para analisar bases fMRI cerebrais com ênfase no TDAH é necessário também reconhecer padrões de expressões locais incluindo somente as amostras que representam bem o problema. Os métodos de agrupamento bi-direcional, ou biclustering, são capazes de encontrar padrões locais através de agrupamentos simultâneos nas amostras e em atributos, a fim de encontrar os chamados biclusters, que são subconjuntos de amostras que compartilham padrões de expressões similares sobre um subconjunto de atributos, formando uma submatriz homogênea (PADILHA; CAMPELLO, 2017), (MADEIRA; OLIVEIRA, 2004). Em outras palavras, algoritmos biclustering podem identificar quais amostras fMRI são mais representativas ao TDAH, que compartilham das mesmas regiões cerebrais afetadas por esse transtorno, amenizando a maldição da dimensionalidade e revelando padrões mais complexos que não podem ser revelados pelo agrupamento tradicional. Na Tabela 1 podem ser vistas algumas diferenças entre o agrupamento tradicional e o biclustering.

Provavelmente, o primeiro algoritmo *biclustering* encontrado na literatura foi descrito por (HARTIGAN, 1972) e batizado de "*direct clustering*" ou *Block Clustering*. Entretanto, somente em (CHENG; CHURCH, 2000) o termo "biclustering"foi adotado para esta técnica de reconhecimento de padrões. A partir do método apresentado por (CHENG; CHURCH, 2000), vários métodos biclustering focaram na identificação de grupos em dados com expressões gênicas (PADILHA; CAMPELLO, 2017), (OGHABIAN et al., 2014), (EREN et al., 2012), (BARKOW et al., 2006). Isto se deve à grande quantidade de informações disponíveis sobre processos e funções relacionadas a genes em coleções de anotações gênicas, como o Gene Ontology Consortiom (ASHBURNER et al., 2000).

Tabela 1 – Diferença entre agrupamento e *biclustering*

Característica	**Agrupamento**	**Biclustering**
Aplicação	Linhas OU colunas, separadamente	Linhas E colunas, simultaneamente
Modelo produzido	Global	Local
Amostras	Cada amostra é agrupada usando todos os atributos	Cada amostra é agrupada usando um subconjunto de atributos
Atributos	Cada atributo é agrupado usando todas as amostras	Cada atributo é agrupado usando um subconjunto de amostras

No decorrer deste capítulo, as definições e formulação do biclustering são descritas na seção 3.1. Os tipos de biclustering, assim como os tipos de biclusters estão na seção 3.2. Seção 3.3 apresenta o SSBimax, método semi-supervisionado baseado em biclustering que será testado e comparado nos experimentos deste trabalho. Por fim, a seção 3.4 apresenta as considerações finais.

3.1 Definição do Problema

A definição de biclustering descrita por (MADEIRA; OLIVEIRA, 2004) é a que segue: dada a matriz $A^{n\times m} = (X, Y)$, com o conjunto de linhas $X = x_1, ..., x_n$ e o conjunto de colunas $Y = y_1, ..., y_m$. O elemento a_{ij} corresponde ao valor representando a relação entre a linha x_i e coluna y_i.

Um bicluster $B^{k\times s} = (G, C)$ é uma submatriz de A formada pelo subconjunto de linhas $G = g_1, ..., g_k$ ($G \subseteq X$ e $k \leq n$) que possuem comportamentos similares sob um subconjunto de colunas $C = c_1, ..., c_s$ ($C \subseteq Y$ e $s \leq m$).

O objetivo do biclustering é identificar em A um conjunto de bicluster $B_{opt} = B_1, B_2, ..., B_l$, tal que cada bicluster $B_k = (G_k, C_k)$ satisfaça características de homogeneidade específicas do método. Os critérios usados para qualificar um bicluster e manter os padrões de seus elementos são determinados pelas funções objetivo (por exemplo, baixa variância com relação à matriz original).

Em (FREITAS et al., 2013), uma função objetivo que visa encontrar B_{opt} que maximixe o grau de coerência dentro do conjunto $BC(A)$ de todos os possíveis grupos de biclusters associados a A é dada pela Equação 3.1.

$$f(B_{opt}) = max_{B\in BC(A)} f(B) \tag{3.1}$$

Uma matriz de dados pode ser vista como um grafo bipartido com peso. Um grafo $F = (V, E)$, em que V é o conjunto de vértices e E o conjunto de arestas, é considerado bipartido se os vértices podem ser particionados em dois conjuntos L e R, tal que toda aresta em E tem exatamente um fim em $a \in L$ e o outro fim em $b \in R$, com $V = L \cup R$. O par de vértices $a, b \in V$ é chamado de biclique e as arestas entre a e b formam um subgrafo bipartido de F. A matriz de dados $A^{n \times m} = (X, Y)$ é um grafo bipartido com peso, em que cada nó $n_i \in L$ corresponde a uma linha e cada nó $n_j \in R$ corresponde a uma coluna. As arestas entre n_i e n_j são os elementos a_{ij} e o conjunto de linhas e colunas o qual estas arestas estão contidas formam um bicluster. Em (PEETERS, 2003), é mostrado que a busca por um biclique, assim como pelo maior biclique, é um problema NP-Completo. Por este fato, um método biclustering geralmente é um problema de complexidade NP-Completo.

3.2 Métodos Biclustering

Os métodos biclustering têm suas próprias características de homogeneidade. Estas características envolvem o tipo de heurística em que se baseiam e o tipo de biclusters que são encontrados, além da forma como múltiplos biclusters se relacionam. Na revisão bibliográfica escrita por (PADILHA; CAMPELLO, 2017), os tipos de heurísticas podem ser divididos em:

1. Algoritmos que executam busca gulosa, geralmente obtêm a melhor solução local em cada iteração a fim de ser guiada para a melhor solução global;
2. Algoritmos de divisão e conquista, divide o problema em instâncias menores que são resolvidas recursivamente e cujas soluções são combinadas à solução do problema original;
3. Algoritmos de enumeração exaustiva, acreditam que as melhores submatrizes são identificadas apenas após geração de todas as combinações possíveis entre linhas e colunas;
4. Algoritmos que identificam parâmetros de distribuição, utilizam modelos estatísticos relacionados a estrutura dos biclusters e então aplica procedimentos iterativos para adaptar seus parâmetros.

Os tipos de biclusters que podem ser encontrados pelos métodos biclustering, segundo (MADEIRA; OLIVEIRA, 2004), são classificados em:

1. Biclusters com valores constantes;
2. Biclusters com valores constantes nas linhas ou colunas;

3. Biclusters com valores coerentes;

4. Biclusters com evoluções coerentes.

Os três primeiros tipos de biclusters levam em consideração os valores numéricos reais dos elementos na matriz de dados. O último tipo de bicluster objetiva encontrar comportamentos coerentes independentemente dos valores exatos dos elementos na matriz, uma vez que estes elementos são vistos como símbolos. Desta forma, biclusters com evoluções coerentes são úteis na identificação de subconjuntos de atributos que mantêm efeitos iguais ou opostos em um subconjunto de amostras.

3.2.1 Biclusters com Valores Constantes

Os métodos biclustering mais simples identificam biclusters com valores constantes. Biclusters constantes revelam subconjuntos de amostras em que todos os valores de expressão dentro de um subconjunto de condições experimentais são similares.

Métodos biclustering que procuram biclusters constantes tendem a ordenar as linhas e colunas da matriz, de modo que valores similares fiquem juntos. Um bicluster constante perfeito é a submatriz (G, C), em que todos os valores são iguais como segue na Equação 3.2:

$$b_{ij} = \mu, (\forall i \in G e \forall j \in C) \tag{3.2}$$

Os elementos b_{ij} de um bicluster encontrados pela Equação 3.2 são tidos como perfeitos porque não apresentam ruídos e fornecem os melhores resultados para análise. Porém, as vezes os elementos em um bicluster constante podem ser mascarados com uma quantidade n_{ij} de ruído (ou seja, $b_{ij} = \mu + n_{ij}$). Quando isso acontece, a função mérito que geralmente é usada para avaliar biclusters constantes é a variância, e um bicluster constante considerado perfeito é aquele cuja a variância é 0.

3.2.2 Biclusters com Valores Constantes em Linhas e Colunas

Também é possível encontrar biclusters com valores constantes somente nas linhas, ou somente nas colunas. Em biclusters com valores constantes nas linhas, os valores de expressões de um subconjunto de amostras continuam similares para um subconjunto de condições experimentais porém, o nível de expressão é diferente para cada amostra. Já para os biclusters com valores constantes nas colunas, cada amostra dentro do bicluster tem o nível de expressão diferente para cada condição experimental mas, para cada condição, o subconjunto de amostras tem expressão similar.

Um bicluster com linhas constantes perfeito é uma submatriz (G, C) em que todos os valores no bicluster são encontrados da seguinte forma:

$$b_{ij} = \mu + \alpha_i \tag{3.3}$$

$$b_{ij} = \mu \times \alpha_i \tag{3.4}$$

Nas Equações 3.3 e 3.4, μ é o valor típico no bicluster e α_i é o ajuste para a linha $i \in G$. Este ajuste pode ser obtido de forma aditiva (Equação 3.3) ou da forma multiplicativa (Equação 3.4).

Similarmente, um bicluster com colunas constantes perfeito é uma submatriz (G, C) em que todos os valores no bicluster são encontrados sob influência de um ajuste β_j para a coluna $j \in C$ que pode ser obtido de forma aditiva (Equação 3.5) ou de forma multiplicativa (Equação 3.6).

$$b_{ij} = \mu + \beta_j \tag{3.5}$$

$$b_{ij} = \mu \times \beta_j \tag{3.6}$$

Para identificar biclusters não constantes, a aplicação da variância nos valores da matriz, como apresentada na seção 3.2.1, não retornará bons resultados. Uma técnica utilizada envolve normalizar as linhas ou colunas da matriz, de forma que transforme os biclusters encontrados em biclusters constantes (seção 3.2.1). Outra técnica, utilizada em (CALIFANO et al., 2000), trabalha com a descoberta de biclusters não perfeitos por causa da presença de ruídos.

3.2.3 Biclusters com Valores Coerentes

Para revelar relações mais complexas entre amostras e condições experimentais, há métodos biclustering que visam encontrar submatrizes com valores coerentes nas linhas e nas colunas simultaneamente. O bicluster (G, C) com valores coerentes pode ter seus elementos definidos conforme um modelo aditivo (Equação 3.7) ou multiplicativo (Equação 3.8).

$$b_{ij} = \mu + \alpha_i + \beta_j \tag{3.7}$$

$$b_{ij} = \mu' \times \alpha'_i \times \beta'_j \tag{3.8}$$

Nestes modelos, b_{ij} é definido usando seus valores típicos no bicluster, μ e μ', seus ajustes de linha $i \in G$, α e α'_i, e seus ajustes de coluna $j \in C$, β_j e β'_j. Os dois modelos são equivalentes quando $\mu = log(u')$, $\alpha_i = log(\alpha'_j)$ e $\beta_j = log(\beta'_j)$.

Como pode ser visto, as equações 3.3 e 3.5 são casos especiais da equação 3.7, quando $\alpha_i = 0$ e $\beta_j = 0$, respectivamente. Da mesma forma, as equações 3.4 e 3.6 são casos especiais da equação 3.8, quando $\alpha'_i = 0$ e $\beta'_j = 0$, respectivamente.

3.2.4 Biclusters com Evoluções Coerentes

Métodos biclustering nem sempre realizam análises usando diretamente os valores numéricos na matriz de dados. Quando o objetivo é identificar biclusters com evoluções coerentes, a análise é feita independente dos valores exatos dos elementos pois, eles são tratados como símbolos. Esses símbolos podem ser puramente nominais, podem corresponder a uma ordem estabelecida, ou representar mudanças positivas e negativas com relação ao valor real do elemento.

Biclusters com evoluções coerentes são úteis quando o objetivo é de encontrar subconjuntos de amostras que tenham comportamentos acima ou abaixo da média em toda a matriz, assim como para identificar subconjuntos de condições que tenham sempre efeitos iguais ou opostos em um subconjunto de amostras.

3.3 SSBimax

Esta seção apresentará o SSBimax, método semi-supervsionado baseado no popular método biclustering Bimax, proposto em (PRELIĆ et al., 2006). A escolha de propor uma versão semi-supervisionada do Bimax se deu primeiramente pela rapidez em gerar biclusters, pela simplicidade de seu algoritmo e pela flexibilidade de inserir alterações. Em (PADILHA; CAMPELLO, 2017), o Bimax foi identificado como método já referenciado em revisões bibliográficas e frequentemente referenciado na literatura, foi classificado como método que reconhece bem a quantidade existente de biclusters, independente dos tamanhos, e que se comporta bem na presença de sobreposição.

O método biclustering *Binary Inclusion-maximal*, ou Bimax, utilizado originalmente para análise gênica, usa um simples modelo binário de dados, ou seja, as expressões entre regiões cerebrais só podem assumir dois possíveis valores: 0 e 1. A binarização como fase de pré-processamento, no método original, é feita através de um limiar calculado pela Equação 3.9, em que max e min representam os valores máximo e mínimo, respectivamente, na matriz original. Em uma matriz de dados binários $P_{n \times m} = (Q, I)$, o elemento p_{ij} é igual a 1 sempre que a amostra x_i responde a condição y_i na matriz original não binária $A_{n \times m} = (X, Y)$. Caso contrário, é 0. Em outras palavras, $p_{ij} = 1$ sempre que o valor do

elemento a_{ij} é igual ou maior que um limiar.

$$limiar = min + \frac{(max - min)}{2} \tag{3.9}$$

Por se tratar de um limiar global, não são consideradas as particularidades de cada atributo e a análise pode ser prejudicada pelo excesso de valor 1 (no caso das bases utilizadas neste trabalho, atributo é cada expressão entre duas regiões cerebrais).

No SSBimax, esse limiar global é substituído por um conjunto de intervalos entre valores calculados com base em conhecimento prévio, já que se trata de um método semi-supervisionado e existe tanto amostras rotuladas quanto as não rotuladas. Para cada condição y_i haverá um intervalo específico $\lfloor min_j, max_j \rfloor$ formado pelo seu maior e menor valor, respectivamente. Atualizando, $p_{ij} = 1$ sempre que o valor do elemento a_{ij} estiver dentro do intervalo e p_{ij} em caso contrário. Desta forma, cada atributo terá seu intervalo de binarização específico e em região de interesse.

O Bimax conceitua como bicluster (G, C) um subconjunto de amostras que respondem em conjunto a um subconjunto de condições, ou seja, (G, C) é submatriz de P em que todos os elementos $p_{ij} = 1$. Isso o caracteriza como método que busca biclusters de valores constantes. Para evitar biclusters de um único elemento, Bimax objetiva encontrar todos os biclusters que são *inclusion-maximal*. Para isso, o par (G, C) é definido se, e somente se:

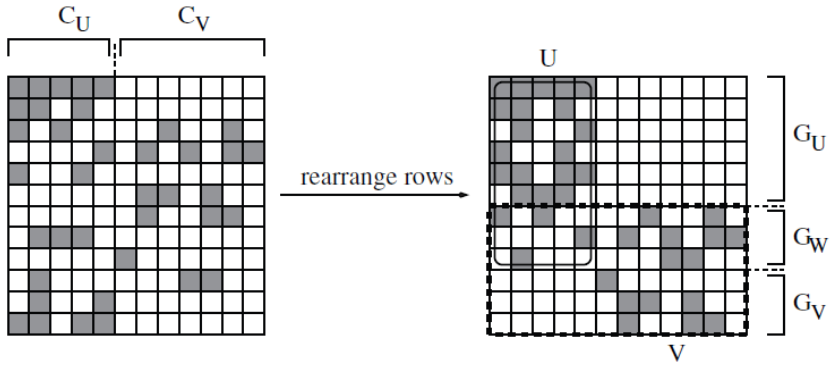
1. $\forall i \in G, \forall j \in C : e_{ij} = 1$
2. $\nexists (G', C') \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ com
   a) $\forall i' \in G', \forall j' \in C' : e_{i'j'=1}$ e
   b) $C \subseteq G' \wedge C \subseteq C' \wedge (G', C') \neq (G, C)$

O algoritmo do Bimax realiza um método da classe de divisão e conquista, que é classificado por (MADEIRA; OLIVEIRA, 2004) como sendo potencialmente rápido. O tempo de execução que o Bimax fornece para matrizes contendo somente biclusters separados é da ordem $O(nm\beta)$, em que β é o número de *inclusion-maximal* na matriz de dados, e para biclusters arbitrários o tempo de execução é da ordem $O(nm\beta min\{n, m\})$. A memória requerida é da ordem $O(nmmin\{n, m\})$.

Dada a matriz binária $P^{n \times m} = (Q, I)$, a ideia do Bimax é particionar P em três submatrizes: uma submatriz composta somente pelo valor 0, que sempre será o alvo para exclusão da análise, e as submatrizes U e V, que participam do processo recursivo do algoritmo. O ponto de parada da recursividade se dá quando a matriz corrente for um bicluster, ou seja, quando todos os seus elementos forem 1.

O primeiro passo para particionar P é selecionar uma linha que contenha tanto elementos 0 quanto elementos 1 para servir de referência (caso só haja linhas contendo 1, P é um único bicluster, mas se somente houver 0, não existe bicluster em P). Essa referência serve para dividir o conjunto de colunas em dois subconjuntos, C_u e C_v, em que C_u corresponde as colunas com elementos 1 e C_v corresponde as colunas com elementos 0 da linha referência. Na parte esquerda da Figura 6 é possível ver que a primeira linha foi tomada como referência para construção de C_u e C_v.

Figura 6 – Iteração Bimax



Fonte: (PRELIĆ et al., 2006)

Depois disso, as linhas de P são reorganizadas do modo a seguir: primeiramente vêm as linhas cujos os elementos 1 estejam somente no subconjunto de colunas C_u e chamá-los de G_u; depois as linhas cujos elementos 1 pertençam tanto ao subconjunto de colunas C_u quanto ao C_v e chamá-los de G_w; por último, as linhas cujos elementos 1 estejam somente no subconjunto C_v e chamá-los G_v. Essas três novas submatrizes (G_u, G_w e G_v), como mostra a parte direita da Figura 6, combinadas a C_u e C_v, geram as submatrizes U e V. A submatriz $U = (G_u \cup G_w, C_u)$ é formada pelas amostras contidas em G_u e G_w mas usando apenas as colunas de C_u, de forma que exclui as amostras pertencentes a G_v. Já a submatriz $V = (G_w \cup G_v, C_u \cup C_v)$ é formada pelas amostras contidas em G_w e G_v.

Ainda na Figura 6, tanto U quanto V não podem ser considerados biclusters porque contêm elementos 0, portanto o ponto de parada não foi atingido e a recursividade é necessária. Contudo, mesmo na primeira iteração do Bimax já é possível perceber na submatriz U sua capacidade de reduzir a dimensionalidade nos biclusters para um subconjunto de amostras.

O processo recursivo pode seguir por duas direções a depender da existência de G_w: se não houver G_w, então U e V não têm sobreposição e são processados independentemente; se houver G_w, pode correr o risco de um bicluster em V não ser *inclusion-maximal* na matriz original P e não interessar ao Bimax. Neste caso, (PRELIĆ et al., 2006) adverte para que sejam gerados apenas os biclusters em V que compartilham ao menos uma coluna com C_v.

Por fim, os biclusters encontrados pelo Bimax não são exaustivos, ou seja, nem todas as amostras poderão estar presentes nos biclusters, e não são exclusivos porque uma linha ou uma coluna pode pertencer a mais de um bicluster.

Além de utilizar esse conceito de bicluster, o SSBimax também aplica um procedimento pós-processamento para filtrar somente os chamados biclusters válidos. Por se tratar de um método semi-supervisionado, algumas restrições devem ser respeitadas, e aqueles biclusters que não violarem nenhuma restrição *must_link* e *cannot_link* são considerados biclusters válidos limpos. Ainda assim, em casos onde há uma quantidade considerável de amostras rotuladas, os intervalos de binarização tendem a ficar maiores e os biclusters encontrados tendem a aceitar falsos positivos. Como solução, o SSBimax também leva em consideração os biclusters que tenham um limiar pequeno de falso positivo, batizado de limite de impureza. Esses biclusters recebem o nome de biclusters válidos impuros. No Algoritmo 2 estão representadas todas as etapas realizadas pelo SSBimax até a lista de biclusters válidos.

Algoritmo 2: SSBimax

Entrada : amostras rotuladas X_l, amostras não rotuladas X_u
Saída : bicValidos

$listaLimiares = calculaLimiares(X_l)$
$matrizBinaria = binariza(X_l, X_u, listaLimiares)$
$listaBiclusters = Bimax(matrizBinaria)$
$must\_link = restMust\_link(X_l)$
$cannot\_link = restCannot\_link(X_l, X_u)$
$bicValidos = \{\}$
for *bic in listaBiclusters* **do**
  if *bic não viola restrições must_link e cannot_link* **then**
    $bicValidos = \{bicValidos, bic\}$
  else if $FalsoPositivo < Impureza$ **then**
    $bicValidos = \{bicValidos, bic\}$
end

3.4 Considerações Finais

Neste capítulo foi apresentado o SSBimax, método semi-supervisionado baseado no Bimax (tradicional método biclustering). Alterações pontuais com relação ao método original, a capacidade de combinar amostras rotuladas e amostras não rotuladas a partir da aplicação de restrições *must_link* e *cannot_link*, fazem do SSBimax capaz de amenizar a árdua tarefa de rotular manualmente bases de fMRI, ao mesmo tempo realizar uma seleção de atributos específica para um determinado grupo de amostras que representam tão bem o problema abordado, a fim de construir modelos confiáveis para predição do TDAH.

A seguir serão apresentados os experimentos, as etapas realizadas para análise de bases fMRI com quantidade de amostras rotuladas reduzidas, em que serão aplicados os métodos semi-supervisionados tradicionais e o SSBimax.

Capítulo 4
Experimentos

Neste capítulo serão apresentadas as especificações dos experimentos que foram realizados neste trabalho. Como o objetivo principal aqui é de apresentar um método semi-supervisionado e aplica-lo em bases fMRI reduzidas com foco no TDAH, o poder da rotulação automaticamente de amostras deve ser avaliado através da redução da quantidade de amostras rotuladas disponível.

Na seção 4.1 serão apresentadas as bases de matrizes de conectividades obtidas por técnicas de fMRI utilizadas neste trabalho. Em seguida, na seção 4.2, será feito um detalhamento do processo de experimentos cujos resultados estão no Capítulo 5.

4.1 Bases de Dados

As bases utilizadas neste trabalho foram originalmente publicadas por (MILHAM et al., 2012) como um conjunto de matrizes de conectividade chamada ADHD200_CC200, que foi extraída da base ADHD-200, usada em uma competição que avaliou a performance de vários classificadores na predição de classes (BROWN et al., 2012). Ela consiste em 520 matrizes de 190 pessoas com TDAH (ou ADHD, do inglês *Attention Deficit Hyperactivity Disorder*) e 330 pessoas consideradas com desenvolvimento motor típico (ou *Typically Developing*), extraídas da técnica do rs-fMRI. Cada matriz de conectividade, ou cada amostra, representa o cérebro de uma pessoa com idade entre 7 e 21 anos e é formada por níveis de expressão entre 190 regiões cerebrais, totalizando 17955 atributos, que foram medidas usando o coeficiente de correlação de Pearson, cujos valores pertencem ao intervalo [-1,1]. Todas as matrizes de conectividade estão disponíveis na página do USC *Multimodal Connectivity Database*[1].

Pelo fato de subconjuntos de matrizes terem sido adquiridas em lugares diferentes, por máquinas de ressonância magnética diferentes e manuseadas por pessoas diferentes, este trabalho considera que elas sejam independentes. Além disso, foi-se verificado um nível de dissimilaridade entre as amostras obtidas pela classe com desenvolvimento motor típico (ou Typ, de forma reduzida), como mostra o experimento mais adiante nesta seção. Como consequência, foram adaptados grupos de matrizes de conectividade, ou quatro bases de dados, batizados de acordo com o local onde foram adquiridas as rs-fMRI: KKI, NeuroIMAGE, NYU e Peking. As amostras de cada base de dados estão detalhadas na

[1] http://umcd.humanconnectomeproject.org/

Tabela 2.

Tabela 2 – Quantidade de amostras nas bases de dados

Bases de Dados	**Typ**	**TDAH**	**Total**
KKI	58	20	78
NeuroIMAGE	22	17	39
NYU	91	96	187
Peking	93	57	150

Dessa forma, para avaliar a rotulação automática dos métodos de aprendizado semi-supervisionado, estarão disponíveis bases balanceadas (NeuroIMAGE e NYU) e desbalanceadas (KKI e Peking), todas com a quantidade de amostras muito inferior à quantidade de atributos. Para validar a independência dessas bases, as Tabelas 3 e 4 mostram a sensibilidade e a especificidade quando amostras da mesma classe obtidas por diferentes lugares são agrupadas usando o método clássico de agrupamento K-means.

As medidas $sensibilidade = (VerdadeirosPositivos/TotalPositivos)$ e $especificidade = (VerdadeirosNegativos/TotalNegativos)$ calculam a quantidade de acertos para a classe positiva e negativa, respectivamente. Seus valores variam no intervalo [0,1], sendo o valor máximo significado de boa classificação.

Tabela 3 – Teste de Similaridade em Amostras Typ

Base de Dados	**Sensibilidade**	**Especificidade**
KKI x NeuroIMAGE	1,00	0,81
KKI x NYU	1,00	1,00
KKI x Peking	0,97	0,94
NeuroIMAGE x NYU	0,40	0,70
NeuroIMAGE x Peking	0,04	0,91
NYU x Peking	0,20	0,81

Considerando as amostras de desenvolvimento motor típico, a Tabela 3 mostra que a base KKI tem amostras com altas dissimilaridades com relação às demais bases de dados. Nesse caso, tanto a sensibilidade quanto a especificidade tiveram suas medidas próximas de 1.0, o que minimiza os índices de falso positivo e falso negativo. As bases NeuroIMAGE e Peking demonstraram ter amostras similares de acordo com o alto nível de especificidade e baixo nível de sensibilidade. A causa dessa baixa semelhança pode ser a pequena quantidade de amostras na base NeuroIMAGE. As demais comparações entre bases fMRI não mostraram nível de semelhança, nos permitindo analisa-las cada uma de forma independente.

Tabela 4 – Teste de similaridade em amostras TDAH

Base de Dados	**Sensibilidade**	**Especificidade**
KKI x NeuroIMAGE	0,95	0,35
KKI x NYU	0,93	0,05
KKI x Peking	0,80	0,10
NeuroIMAGE x NYU	0,76	0,05
NeuroIMAGE x Peking	0,94	0,49
NYU x Peking	0,98	0,01

Por outro lado, a Tabela 4 mostra que as amostras TDAH são mais homogêneas e podem compor um grande grupo com a união delas. Esse fato pode revelar alguns padrões relacionados aos subtipos do TDAH (Desatento, Hiperativo-Impulsivo ou a combinação de ambos), com a única exceção para a baixa similaridade nas combinações das amostras da base NeuroIMAGE. Ainda sobre a base NeuroIMAGE, tanto os baixos valores de especificidade e os altos valores de sensibilidade quando comparados às bases KKI e Peking, quanto a baixa sensibilidade e especificidade quando comparados à base NYU, fazem do uso das amostras NeuroIMAGE desvantajoso na identificação dos subtipos do TDAH.

4.2 Detalhando os Experimentos

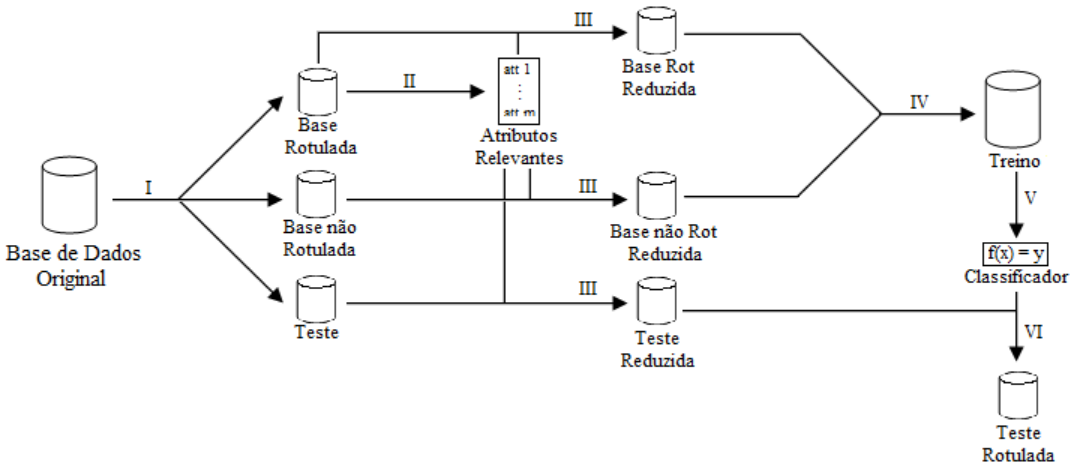
Nesta seção serão detalhadas as etapas realizadas nos experimentos para avaliar a capacidade de rotulação automática do SSBimax e dos métodos semi-supervisionados tradicionais, quando aplicados às bases fMRI apresentadas na seção 4.1. Essas bases terão a quantidade de amostras rotuladas reduzida até a menor quantidade suficiente para que o SSBimax consiga gerar um modelo dito confiável com relação às medidas de qualidade (estabelecida mais adiante) resultadas por um conjunto de classificadores. Desta forma, os resultados do capítulo a seguir vão mostrar se a estratégia de encontrar biclusters pode levar vantagem sobre as técnicas usadas por métodos semi-supervisionados tradicionais, assim como se o uso de intervalos como limiar é capaz de criar bons modelos a partir de poucas amostras rotuladas.

As etapas realizadas nos experimentos são mostradas pela Figura 7. Para reduzir a perda de informação causada pela pouca quantidade de amostras nas bases fMRI originais, principalmente em comparação ao tamanho do conjunto de atributos, uma base formada por n amostras é particionada usando *leave-one-out cross validation* (LOOCV), onde cada amostra é usada sozinha como base de teste em uma das n iterações. Por se tratar de um experimento semi-supervisionado, as $n - 1$ amostras que sobraram são escolhidas de forma aleatória entre base rotulada e base não rotulada. No intuito de amenizar o efeito de aleatoriedade da escolha das amostras para cada base, são realizadas dez repetições

em cada iteração do LOOCV. Com relação à quantidade de amostras na base rotulada, a porcentagem inicial será de 90% da base original. Como o critério de avaliação aqui é de obter um modelo confiável usando a menor quantidade de amostras rotuladas, essa porcentagem será alterada de forma decrescente. Cada conjunto de iteração terá uma diferença de 10% de amostras inicialmente rotuladas do conjunto de iterações anterior, até que a medida de qualidade do SSBimax seja inferior à medida obtida pela base original.

O processo de particionar a base original é representada pela etapa I.

Figura 7 – Etapas do projeto



Fonte: Autoria própria

A etapa II representa uma seleção de atributos aplicada à base rotulada usando o Coeficiente de Correlação de Pearson (PCC), apresentada na seção 4.2.1, com finalidade de tornar os experimentos mais rápidos, diminuindo a gravidade da maldição de dimensionalidade ao retirar atributos irrelevantes na diferenciação de amostras pertencentes a diferentes classes. A partir dos experimentos em (GARCIA; NIEVOLA; PARAISO, 2016) e (GARCIA; PARAISO; NIEVOLA, 2017), que avaliaram métodos de seleção de atributos na mesmas bases fMRI usadas neste trabalho de formas distintas, constataram que o PCC foi quem teve o melhor comportamento na predição do TDAH. Como resultado, a partir dessa etapa todas as bases de dados usaram o subconjunto de m atributos considerados relevantes pelo PCC, adotando o nível de significância em 0.01 (etapa III). Como o PCC trabalha com classes numéricas, as amostras pertencentes às pessoas com TDAH recebem o rótulo 1, enquanto que as amostras pertencentes às pessoas com desenvolvimento motor típico recebem rótulo 0. A versão do PCC usado neste trabalho está disponível no Matlab[2].

O aprendizado semi-supervisionado está representado na etapa IV. Nela, as bases rotulada e não rotulada são combinadas de forma a dar origem à base de treino. Como já vistos no capítulo 2, os métodos utilizados e comparados neste trabalho, cujas configurações

[2] www.mathworks.com/products/matlab/

aparecem na seção 4.2.2, são: S3VM, a Minimização da Energia Harmônica o COP-Kmeans, o Bimax e o SSBimax.

Com a base de treino formada, esta deve ser submetida ao processo de treinamento de classificadores (etapa V). Neste trabalho serão utilizados 5 classificadores pertencendo a diferentes abordagens: SVM, KNN, C4.5, Voted Perceptron e Naive Bayes. As configurações dos respectivos parâmetros são vistas na seção 4.2.3. Ao final de cada iteração do LOOCV, cada base de teste (composta por uma única amostra) é avaliada por cada classificador treinado (etapa VI).

Como forma de reduzir a influência do desbalanceamento das bases de dados e evitar acurácias que não têm significados úteis, a avaliação de um método semi-supervisionado para um determinado classificador é calculada por $(sensibilidade + especificidade)/2$. Ao fim das n iterações, a medida de qualidade para cada classificador usando a mesma porcentagem de amostras originalmente rotuladas é dada pela média das avaliações calculadas, considerando também as 10 repetições. Por fim, os métodos semi-supervisionados serão comparados através da média das medidas de qualidade obtidos por todos os classificadores. Os experimentos são repetidos diminuindo cada vez mais a quantidade de amostras originalmente rotuladas até que o SSBimax obtenha média de qualidade inferior à média da medida de qualidade obtida pela base original, onde não houve redução de rotulação.

Todos os experimentos foram realizados no Java Eclipse IDE[3], Matlab ou Weka[4], em uma máquina Intel i7 CPU, 1.73GHz com 6GB de memória RAM.

4.2.1 Coeficiente de Correlação de Pearson (PCC)

O PCC é um método de filtro que constrói um *ranking* de relevância baseado na correlação linear entre cada atributo e a sua classe, de forma univariada. Considerando que os atributos e os rótulos das classes sejam numéricos, a função *score* do PCC para o j-ésimo atributo é obtida pela equação 4.1 e a relevância é avaliada pelo valor absoluto de P_j. Altos valores indicam que o atributo tem grande relevância em distinguir as classes (LAZAR et al., 2012).

$$P_j = \frac{\sum_i (x_{ij} - \bar{x_j})(y_i - \bar{y})}{\sqrt{\sum_i (x_{ij} - \bar{x_j})^2 \sum_i (y_i - \bar{y})^2}} \tag{4.1}$$

4.2.2 Configurações dos Métodos Semi-supervisionados

A versão do S3VM utilizada faz parte do SVM *light*[5] e se baseia no algoritmo de (JOACHIMS, 1999). O SVM *light* executa o método semi-supervisionado automaticamente

[3] https://eclipse.org/
[4] www.cs.waikato.ac.nz/ml/weka/
[5] http://svmlight.joachims.org/

se a base tiver amostras não rotuladas e foi configurado pelos seguintes parâmetros: a proporção de rotulação obedece a mesma da base que já está rotulada e o tipo da função kernel é linear. O método de Minimização da Energia Harmônica está presente no *Semi-supervised Learning Software* (SemiL)[6] e foi configurado pelos seguintes parâmetros: *standard Laplacian* e *hard label* representam Gaussianas harmônicas (como especificado na documentação do método), nosso problema se trata de rotular duas classes, o tipo da função kernel é RBF com distância Euclidiana e a porcentagem de amostras rotuladas é especificada pelo usuário. O COP-Kmeans foi desenvolvido em Java, as restrições *must-link* e *cannot-link* foram obtidas das amostras previamente rotuladas, o número de grupos gerados foi 2 (representam TDAH e desenvolvimento motor típico), com as amostras associadas aos grupos com base na distância Euclidiana e o critério de parada é alcançado sempre que não houver modificações entre iterações do agrupamento.

Os últimos métodos semi-supervisionados comparados neste trabalho são o SSBimax e o Bimax, apresentados na seção 3.3. O código fonte do Bimax está disponível na linguagem de programação C, na página do trabalho onde foi proposto (PRELIĆ et al., 2006) [7]. Os parâmetros exigidos pelo algoritmo para buscar biclusters são as menores e as maiores quantidades de linhas e colunas que podem formar um bicluster. O número mínimo de amostras em um bicluster no SSBimax é o número de amostras TDAH rotuladas, já o número máximo é a quantidade total de amostras da base rotulada mais a base não rotulada. Para o Bimax original, os mesmos dois parâmetros precisam variar para tornar possível encontrar biclusters válidos.

Com relação ao número de colunas, para ambos os métodos, a quantidade máxima não pode ultrapassar a quantidade de atributos considerados relevantes pelo PCC. Por último, a quantidade mínima de atributos é iterativamente aumentada para os dois métodos, a partir de 5 atributos, de forma que o tempo de processamento desde a geração de todos os biclusters até o cálculo da medida de qualidade dos classificadores não seja muito maior que o tempo levado pela base original para qualificar o conjunto de amostras completamente rotulado. A maior medida de qualidade para cada classificador é comparada com os demais métodos semi-supervisionados. Desta forma, além de mostrar se o SSBimax é capaz de gerar bons modelos através da redução da quantidade de amostras rotuladas, também será possível ver se a tática de encontrar biclusters leva vantagem sobre as demais técnicas tradicionais semi-supervisionadas.

No lugar do limiar global único utilizado pelo Bimax, o SSBimax utiliza para cada atributo um intervalo entre o menor valor e o maior valor das amostras pertencentes à classe TDAH. Desta forma, espera-se que aquelas amostras fora dos intervalos pertençam aos que tenham desenvolvimento motor típico, e que os biclusters encontrados façam parte

[6] http://www.learning-from-data.com/te-ming/semil.htm
[7] http://people.ee.ethz.ch/ sop/bimax/

de uma região específica e de interesse. Assim como no COP-Kmeans, as restrições *must-link* e *cannot-link* também foram obtidas das amostras previamente rotuladas. O limiar de impureza foi fixado em 15% da quantidade total de amostras sem TDAH agrupadas em um bicluster. Após a fase de pós-processamento, onde são filtrados os biclusters de interesse, a escolha do melhor bicluster pode ser guiada de várias maneiras (HORTA; CAMPELLO, 2014), (LIU; WANG, 2006). Neste trabalho, essa escolha é dada por: primeiro são selecionados aqueles biclusters formados pela quantidade média de atributos dentro do universo de biclusters encontrados; por fim, aquele cujas amostras na base não binarizada tenham a menor variância média para o conjunto de atributos.

Um bicluster é específico para uma classe, no caso daqui, é sempre específico para o TDAH por mais que haja impurezas. Para que a base de treino tenha representantes de ambas as classes, ela será composta da seguinte forma:

1. As amostras do melhor bicluster pertencentes à classe TDAH mais as amostras da classe Typ que não pertencem ao bicluster;
2. As amostras pertencentes ao conjunto de impureza são excluídas;
3. Para completar, as amostras rotuladas automaticamente.

Nota-se que as amostras TDAH que não fazem parte do melhor bicluster também não fará parte da base de treino, já que essas amostras não representam a classe tão bem quanto as presentes no bicluster.

4.2.3 Configurações dos Classificadores

Os métodos semi-supervisionados são avaliados através de medidas estatísticas obtidas por cinco classificadores pertencentes a diferentes abordagens e disponíveis no Weka. O libSVM usando kernel linear com 1.0 como parâmetro de custo; KNN considerando 5 vizinhos mais próximos; C4.5 considerando ao menos 5 amostras por folha; Voted Perceptron e Naive Bayes. Múltiplos classificadores são usados para prover melhores comparações, uma vez que cada base de dados tem seu comportamento observado por diferentes princípios de classificação. As configurações estabelecidas para os classificadores já foram utilizadas em (GARCIA; NIEVOLA; PARAISO, 2016) e (GARCIA; PARAISO; NIEVOLA, 2017). Por este trabalho não se tratar de um estudo sobre os classificadores e pela pequena quantidade de amostras, não foram realizados particionamentos das bases para ajustes das melhores configurações dos classificadores supracitados.

Capítulo 5
Resultados Experimentais

Neste capítulo serão apresentados os resultados dos experimentos descritos no Capítulo 4. O objetivo deste capítulo é de mostrar o comportamento de um novo método de aprendizado semi-supervisionado baseado em biclustering, aplicado em bases fMRI com foco no TDAH. O já apresentado SSBimax terá sua capacidade de rotulação automática testada usando a menor quantidade de amostras previamente rotuladas suficiente para gerar um modelo confiável. Além disso, também será mostrada se a sua estratégia de encontrar biclusters leva vantagem sobre as técnicas usadas por métodos semi-supervisionados tradicionais, e se o uso de intervalos como limiar é mais poderoso que o uso de um limiar único utilizado pelo método original do Bimax.

O SSBimax será comparado a outros quatro métodos semi-supervisionados, com diferentes conceitos de rotulação: O S3VM, baseado em SVM, a Minimização de Energia Harmônica, baseado em teoria dos grafos, o COP-Kmeans, baseado no K-means e o método usando a versão original do Bimax. Nessa mesma comparação entrarão duas versões da base de dados parcialmente rotulada sem rotulação automática: a base reduzida pelo PCC e a base sem sofrer redução, usando todo o conjunto de atributos.

Como já dito anteriormente, a confiabilidade de um modelo é dado pela média das qualidades obtidas por um conjunto de classificadores. Um modelo é confiável se a sua média de qualidade for igual ou superior à média de qualidade obtida pela base original, onde somente a amostra destinada à base de teste não é rotulada. Desta forma, as tabelas apresentadas nas próximas seções mostrarão os níveis de confiança para cada método comparado, em cada experimento realizado.

Por convenção e para tornar a leitura mais fácil de entendimento, os experimentos serão referenciados da seguinte forma:

- Experimento 1: 90% da base original é previamente rotulada;
- Experimento 2: 80% da base original é previamente rotulada;
- Experimento 3: 70% da base original é previamente rotulada;
- Experimento 4: 60% da base original é previamente rotulada;
- Experimento 5: 50% da base original é previamente rotulada;
- Experimento 6: 40% da base original é previamente rotulada;

- Experimento 7: 30% da base original é previamente rotulada;
- Experimento 8: 20% da base original é previamente rotulada;
- Experimento 9: 10% da base original é previamente rotulada.

Além das tabelas, gráficos mostrarão as médias de atributos usados nos melhores biclusters encontrados tanto pelo SSBimax quanto pelo Bimax, comparadas com as médias de atributos considerados relevantes pelo PCC no nível de significância em 0.01. As seções a seguir estão divididas pelas bases fMRI: KKI, NeuroIMAGE, NYU e Peking.

5.1 Base KKI

A primeira base cujos resultados são apresentados nesta seção é o KKI. Além de ser uma das menores bases, ela também é desbalanceada e a geração de modelos que conseguem diferenciar as classes se torna em um trabalho árduo. A Tabela 5 mostra as medidas de qualidade de cada classificador para a base KKI original (composta por 77 amostras de treino e todos os atributos), resultando em uma média de qualidade geral de 0.51 (0.63 como a maior média obtida por C4.5 e 0.47 como a menor média obtida por SVM e Naive Bayes). Na Tabela 6 estão as médias de qualidade para a base KKI com a quantidade reduzida de amostras inicialmente rotuladas, considerando a média de qualidade da base original como critério de parada dos experimentos.

Tabela 5 – Resultados para base KKI original

SVM	**KNN**	**Perceptron**	**C4.5**	**NB**
0.47	0.51	0.49	0.63	0.47

Tabela 6 – Resultados da base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.48	0.48	0.48
Com PCC	0.50	0.48	0.48
S3VM	0.48	0.48	0.48
COP-KMeans	0.47	0.47	0.47
Grafo	0.48	0.48	0.49
Bimax	0.49	0.49	0.49
SSBimax	0.53	0.51	0.50

A Tabela 6 mostra que o SSBimax permitiu reduzir a base KKI para até 80% dela rotulada inicialmente, mas usando 70% sua média de qualidade foi inferior à obtida

pela base original por 1%. Em geral, a redução da rotulação inicial não afetou muito a tarefa de classificar pois, do experimento 1 até o experimento 3, a média de qualidade se manteve constante e inferior apenas 0.03 à média original. Da mesma forma aconteceu após a redução de dimensionalidade pelo PCC, em que a única diferença foi o aumento da qualidade no experimento 1, de 0.48 para 0.50.

Tabela 7 – Média de amostras rotuladas por biclustering

Métodos	**1**	**2**	**3**
SSBimax	69	62	57
Bimax	54	50	43

Apesar da pequena queda por causa da redução de amostras rotuladas, os métodos semi-supervisionados, juntamente com a redução de dimensionalidade, conseguiram aumentar as médias de qualidade em alguns casos. Entre os destaques temos: as médias 0.53 e 0.51 obtidas pelo SSBimax nos experimentos 1 e 2, respectivamente, que foram maiores que a qualidade original; a média 0.49 obtida pelo Bimax em todos os experimentos, assim como o método baseado em grafo para o experimento 3. Já o COP-Kmeans foi o pior método nos três experimentos.

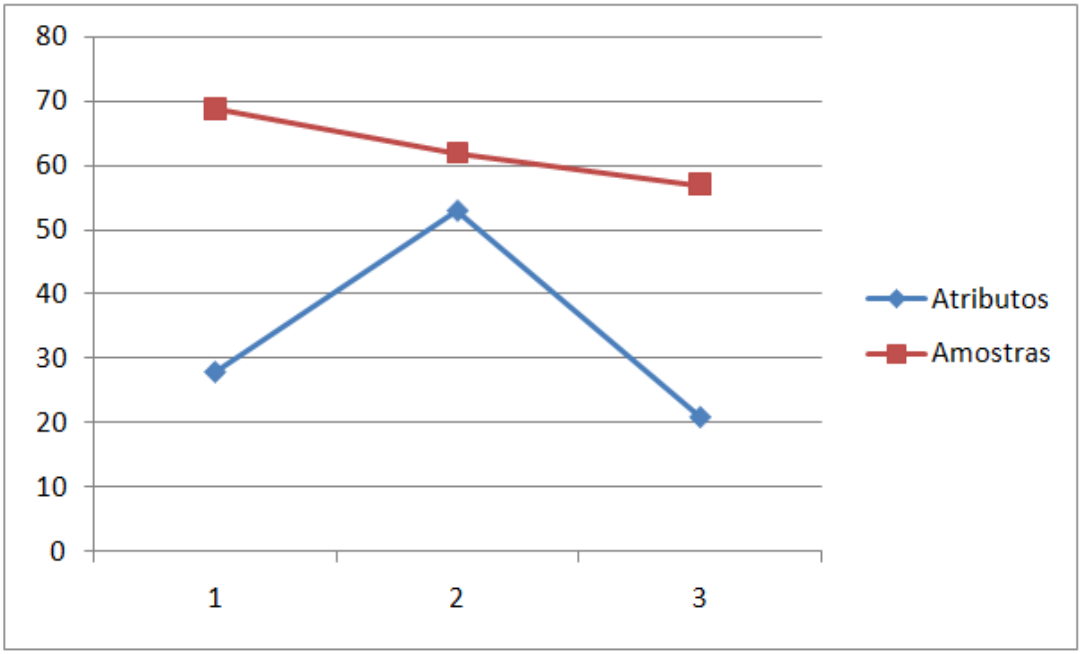
Além de terem sidos os melhores métodos para a base KKI, o SSBimax e o Bimax mostraram que mais amostras rotuladas nem sempre é vantagem porque nem todas as amostras são representativas. Como pode ser visto na Tabela 7, o SSBimax (único a ter melhorado a média das qualidades dos classificadores na base original) teve média de 62 amostras rotuladas após a geração da base de treino para os três experimentos e o Bimax, 49.

Todos os métodos foram comparados após a base original ter passado pelo PCC. A Figura 9 mostra nas barras azuis que a dimensionalidade da base KKI foi reduzida para médias de 125, 118 e 112 atributos relevantes nos três experimentos e mostrou ser vantajoso por ter aumentado a qualidade no experimento 1 e não ter sido inferior nos experimentos 2 e 3. A média de atributos utilizados pelo SSBimax nessa base, como pode ser vista nas barras vermelhas da Figura 9, foi de 28, 53 e 21 nos experimentos 1, 2 e 3, respectivamente. Pelo fato da quantidade de atributos ser menor que a quantidade de amostras em todos os experimentos, como ser visto na Figura 8, pode-se confirmar que o SSBimax foi capaz também de eliminar a maldição da dimensionalidade no base KKI.

No caso do Bimax, representado pelas barras verdes da Figura 9, ele encontrou biclusters com médias de atributos bem menores. Juntamente com os baixos números de amostras rotuladas, que são menos que as amostras inicialmente rotuladas, representam a dificuldade de encontrar biclusters válidos, sejam com impurezas ou não.

A seguir, os resultados detalhados de cada classificador são apresentados para a base

Figura 8 – Relação amostras x atributos no SSBimax



Fonte: Autoria própria

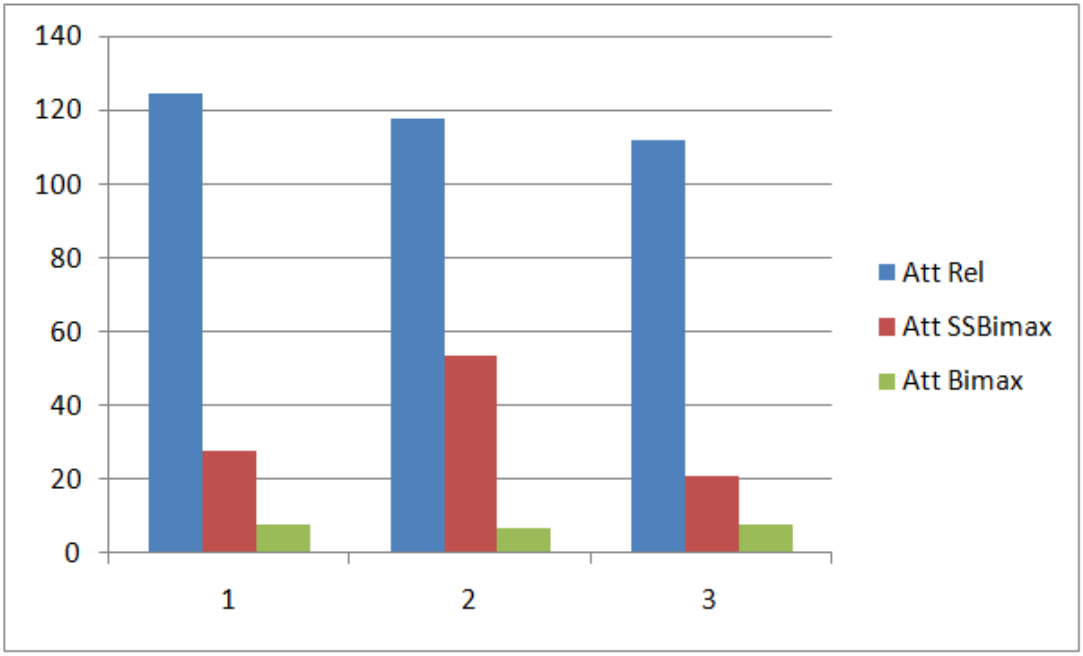
KKI nos três experimentos realizados, mostrando quais classificadores foram responsáveis pela interrupção no experimento 3 e quais possibilitaram diminuir para 80% a quantidade de amostras inicialmente rotuladas.

5.1.1 Resultados SVM

Com o classificador SVM, como mostra a Tabela 8, a base não foi afetada pela redução da rotulação pois, mesmo se tratando de 70% das amostras originalmente rotuladas, manteve os 0.47 de qualidade obtido pela sua versão original. Uma vez que essas bases tiveram suas dimensões reduzidas, os subconjuntos de atributos considerados relevantes pelo PCC não foram capazes de melhorar a média de qualidade da classificação, provavelmente por causa do desbalanceamento da base. Mesmo assim, a sua queda não foi grande (0.01 de diferença) e se manteve constante até o experimento três.

Por outro lado, a rotulação automática foi vantajosa na maioria das vezes. Primeiro por ter aumentado as médias de quando foram usadas somente as amostras inicialmente rotuladas, independente se as bases foram reduzidas pelo PCC. Consequentemente, a segunda vantagem do aprendizado semi-supervisionado usando o SVM para a base KKI fica por conta dos métodos S3VM, Bimax e SSBimax, por obterem médias acima da original em todos os experimentos, mostrarem bom uso das amostras não rotuladas e inibirem a queda por parte do PCC. Os métodos biclustering se destacaram por terem sido os melhores, o SSBimax foi melhor por ter conseguido 0.51 no experimento 1 e o Bimax

Figura 9 – Média de atributos por experimento na base KKI



Fonte: Autoria própria

Tabela 8 – Resultados SVM para a base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.47	0.47	0.47
Com PCC	0.46	0.46	0.46
S3VM	0.48	0.48	0.48
COP-KMeans	0.45	0.46	0.46
Grafo	0.46	0.47	0.48
Bimax	0.50	0.49	0.50
SSBimax	0.51	0.49	0.49

foi o melhor método semi-supervisionado tradicional. Dessa forma, ao considerar apenas este classificador, o SSBimax seria capaz de reduzir a base KKI para 60% de amostras rotuladas.

Diferente do que ocorreu com o SSBimax, onde as medidas de qualidade foram diminuindo junto com a quantidade de amostras rotuladas, o método baseado em grafo recuperou a qualidade perdida pelo PCC de forma contrária, já que teve medições crescentes com o passar dos experimentos. Por fim, o COP-Kmeans aumentou sua média em 0.01 quando passou do experimento 1 para o 2, a manteve e não ultrapassou o 0,46 obtido pelo PCC.

5.1.2 Resultados KNN

A média de classificação correta usando o KNN para a base original foi de 0.51. Um fato interessante, apresentado na Tabela 9, que aconteceu ao reduzir a quantidade de amostras rotuladas foi a grande queda no experimento 1 (0.41) seguida de uma recuperação ainda maior ao reduzir para 80% e logo após 70%, quando atingiu uma média 0.54 de qualidade, maior até que a média de qualidade original.

Diferente do que aconteceu com o SVM, a redução da dimensionalidade por parte do PCC aumentou bastante a média de qualidade no experimento 1 e ainda foi superior ao quando foi usado todos os atributos no experimento 2. No experimento 3 o PCC manteve os 0.46 de média do experimento anterior e já foi inferior ao não reduzir a dimensionalidade.

A melhora que o PCC provocou nas média de qualidade no experimento 1 foi fundamental para o SSBimax obter média 0.52, único valor maior que a média original. Se dependensse apenas do KNN, a base KKI só seria reduzida a 90% das amostras rotuladas sem prejudicar a qualidade do modelo gerado. Outro método que aumentou a média do PCC no mesmo experimento foi o Bimax, com 0.50. A rotulação automática se mostrou benéfica de forma geral apenas nos experimentos 2 e 3, quando ninguém foi pior que a média obtida pelo PCC.

O SSBimax foi superior aos demais métodos semi-supervisionados e, apesar de ter conseguido superar a média original deste classificador somente no experimento 1, suas médias nos experimentos 2 e 3 foram bem próximas do sucesso. Entre os métodos tradicionais, o Bimax foi quem mais se aproximou do SSBimax e foi considerado o melhor. Apesar dos demais métodos terem comportamentos bem próximos, o baseado em grafo levou vantagem nos experimentos 2 e 3, deixando os métodos S3VM e COP-Kmeans como os piores.

Tabela 9 – Resultados KNN para a base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.41	0.44	0.54
Com PCC	0.49	0.46	0.46
S3VM	0.47	0.46	0.46
COP-KMeans	0.47	0.46	0.46
Grafo	0.47	0.47	0.47
Bimax	0.50	0.50	0.49
SSBimax	0.52	0.49	0.50

5.1.3 Resultados Perceptron

Com relação ao classificador Perceptron (Tabela 10), a redução da quantidade de amostras rotuladas nos três experimentos afetou de forma que as médias de qualidade foram menores que a média da base original (0.49), mas não acarretou em grandes mudanças entre eles. Como aconteceu com o classificador KNN, o experimento 3 obteve a melhor média com 0.48.

Ao reduzir também a quantidade de atributos com o PCC, somente o experimento 1 conseguiu melhorar a média de qualidade, mesmo assim ficando abaixo da média original. Nos demais experimentos, quando a base inicialmente rotulada tinha 80% da base original, a média de qualidade caiu 0.04 e quando tinha 70% caiu 0.03.

Mais uma vez, os métodos semi-supervisionados aumentaram todas as médias de quando o PCC foi aplicado nos experimentos 2 e 3. No experimento 1, os métodos biclustering foram os únicos a aumentar a qualidade obtida pelo PCC, assim como igualar ou aumentar a média de qualidade original, 0.49 para o Bimax e 0.50 para o SSBimax. No experimento 2, os mesmos métodos ainda foram os únicos a alcançar a qualidade original, mas desta vez a rotulação automática se mostrou sempre vantajosa por causa da baixa qualidade obtida pela redução de dimensionalidade. Por fim, o experimento 3 mostra que as médias de qualidade obtidas pelos métodos semi-supervisionados beiraram a qualidade original, com exceção do COP-Kmeans, em que apenas o Bimax foi capaz de atingir os 0.49.

Considerando apenas o Perceptron, somente o Bimax possibilitaria reduzir a base KKI para 60% de amostras inicialmente rotuladas. O SSBimax, usado como critério de parada dos experimentos, foi inferior à média original por 0.01. Dessa forma, o Bimax pode ser considerado o melhor método semi-supervisionado para este classificador. O COP-Kmeans foi o pior nos três experimentos.

Tabela 10 – Resultados Perceptron para a base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.47	0.46	0.48
Com PCC	0.48	0.42	0.45
S3VM	0.47	0.47	0.48
COP-KMeans	0.42	0.43	0.44
Grafo	0.45	0.45	0.47
Bimax	0.49	0.49	0.49
SSBimax	0.50	0.49	0.48

5.1.4 Resultados C4.5

O classificador C4.5 foi quem obteve a maior média de qualidade para a base KKI original, com 0.63. A Tabela 11 mostra que nos três experimentos houve perda de qualidade de forma crescente, chegando a 0.11 de queda no experimento 3.

Por outro lado, o PCC melhorou as médias das bases com rotulação reduzidas nos experimentos 1 e 2, chegando aos 0.61 de média, mas não serviu de ajuda para quase nenhum método semi-supervisionado. Somente o SSBimax superou a média obtida pelo PCC no experimento 1. Já no experimento 2 ninguém foi superior, apesar do SSBimax ter tido boas médias dos outros classificadores para reduzir a base no experimento 3. No experimento 3 houve uma queda muito grande por parte do PCC, resultando em 0.50, mas foi justamente nesse experimento que a rotulação automática foi superior em 3 dos 5 métodos.

A boa média obtida pelo PCC no primeiro experimento ajudou o SSBimax a conseguir 0.63, que foi a média obtida pela base original. A partir do segundo experimento, nenhum método alcançou essa média e o SSBimax continuou sendo o melhor entre eles. Somente no experimento 3 que o método baseado em grafo empatou com ele, sendo considerado o melhor entre os métodos semi-supervisionados tradicionais. Ainda no experimento 3, a diferença de 0.09 entre a média original e a média obtida pelo SSBimax (0.54) foi o grande responsável pelo não prosseguimento dos experimentos na base KKI.

Tabela 11 – Resultados C4.5 para a base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.57	0.56	0.52
Com PCC	0.59	0.61	0.50
S3VM	0.50	0.49	0.49
COP-KMeans	0.56	0.54	0.51
Grafo	0.56	0.55	0.54
Bimax	0.48	0.48	0.48
SSBimax	0.63	0.59	0.54

5.1.5 Resultados Naive Bayes

Por fim, o classificador Naive Bayes, que obteve 0.47 para a base KKI original. Pela primeira vez, reduzir a quantidade de amostras inicialmente rotuladas melhorou a qualidade da classificação. Como pode ser visto na Tabela 12, nos experimentos 1 e 2 foram obtidos 0.49 de médias e no experimento 3, 0.50. No caso do PCC, as médias de qualidade nos três experimentos caíram quando foi aplicada a redução de dimensionalidade. Porém, somente no experimento 1 o PCC não alcançou ou melhorou a média original.

Ainda na Tabela 12, os métodos semi-supervisionados mostraram que a rotulação automática igualou ou melhorou quase todas as médias obtidas pelo PCC em todos os experimentos, com exceção do COP-Kmeans no experimento 3. Melhor ainda, 100% das vezes foram pelo menos iguais à média original. Pelas médias do SSBimax, a base KKI poderia ser reduzida para 60% das amostras rotuladas. Porém, no experimento 3, o Bimax obteve 0.49 de média e foi considerado o melhor método semi-supervisionado para o Naive Bayes.

Tabela 12 – Resultados Naive Bayes para a base KKI

Métodos	**1**	**2**	**3**
Sem PCC	0.49	0.49	0.50
Com PCC	0.46	0.47	0.48
S3VM	0.48	0.49	0.49
COP-KMeans	0.47	0.47	0.47
Grafo	0.47	0.48	0.49
Bimax	0.50	0.49	0.49
SSBimax	0.50	0.49	0.48

A diferença para os demais métodos foi o padrão decrescente das médias, já que o S3VM, baseado em grafo e o COP-Kmeans não pioraram suas médias ao longo da redução de rotulação inicial. Até o pior método, o COP-Kmeans, se manteve na média original, com 0.47 nos três experimentos.

5.2 Base NeuroIMAGE

A base NeuroIMAGE é a menor base analisada neste trabalho (39 amostras) mas, diferente da base KKI, esta é balanceada (22 amostras Typ e 17 TDAH) e a geração dos modelos devem resultar médias maiores de acertos, assim como a redução de dimensionalidade deve selecionar subconjuntos de atributos melhores. Como pode ser visto na Tabela 13, os classificadores obtiveram 0.58 como média de qualidade geral na base original, tendo como maior média original novamente para o C4.5, 0.78, e a menor média original 0.45 obtida pelo KNN.

Tabela 13 – Resultados para base NeuroIMAGE original

SVM	**KNN**	**Perceptron**	**C4.5**	**NB**
0.6	0.45	0.47	0.78	0.58

Ao considerar a média 0.58 como critério de parada, a Tabela 14 mostra que o SSBimax foi capaz de superar essa média usando pelo menos 80% da base original rotulada,

Tabela 14 – Resultados da base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.50	0.52	0.49
Com PCC	0.53	0.51	0.50
S3VM	0.55	0.53	0.51
COP-KMeans	0.51	0.52	0.49
Grafo	0.49	0.49	0.50
Bimax	0.49	0.49	0.50
SSBimax	0.63	0.59	0.57

enquanto que no experimento 3 ele ficou abaixo por 0.01. Ainda na Tabela 14 é possível ver a grande perda de qualidade que a redução de amostras rotuladas provocou na base original, caindo para 0.5 no experimento 1 e 0.49 no experimento 3. Mesmo assim, o SSBimax foi capaz de reduzir a base para 70% de amostras inicialmente rotuladas com sucesso. O motivo dessa queda passa pela pequena quantidade de amostras na base, o que dificulta bastante na construção de um modelo confiável, apesar de se tratar de uma base balanceada.

A redução de dimensionalidade realizada pelo PCC se mostrou bastante eficaz no experimento 1, ao aumentar em 0.03 a média de qualidade e auxiliar nos valores expressivos do S3VM e SSBimax. Apesar dos demais métodos semi-supervisionados tradicionais não conseguirem aumentar a média de qualidade do PCC e ainda ficarem abaixo da qualidade sem o PCC, com exceção do COP-Kmeans, o S3VM aumentou em 0.02 a média do PCC e o SSBimax superou a média original geral com 0.63.

Ainda por causa da pequena quantidade de amostras na base NeuroIMAGE, a eficácia do PCC foi diminuindo quando a base rotulada ficou cada vez mais reduzida. No experimento 2, a média do PCC já foi menor que a média usando todo o conjunto de atributos. Neste caso, os métodos S3VM e SSBimax ainda foram superiores mas em fluxo decrescente (este último com média 0.59, possibilitando o executar o próximo experimento), enquanto que o COP-Kmeans foi superior e em fluxo crescente. No experimento 3, quando as bases originais com e sem influência do PCC obtiveram as menores médias, somente o COP-Kmeans não melhorou nenhuma das médias e o SSBimax obteve média 0.57, menor que a média original geral e encerrou o processo de redução de rotulação inicial.

Entre os métodos semi-supervisionados tradicionais, o S3VM foi o melhor e, mesmo assim, sempre ficou abaixo da qualidade média original. As baixas médias das qualidades obtidas pelo método Bimax mais as altas médias obtidas pelo SSBImax tornam visível que a estratégia de encontrar biclusters para rotular automaticamente novas amostras não é vantajosa isoladamente. Nesta base fica claro a importância de utilizar intervalos

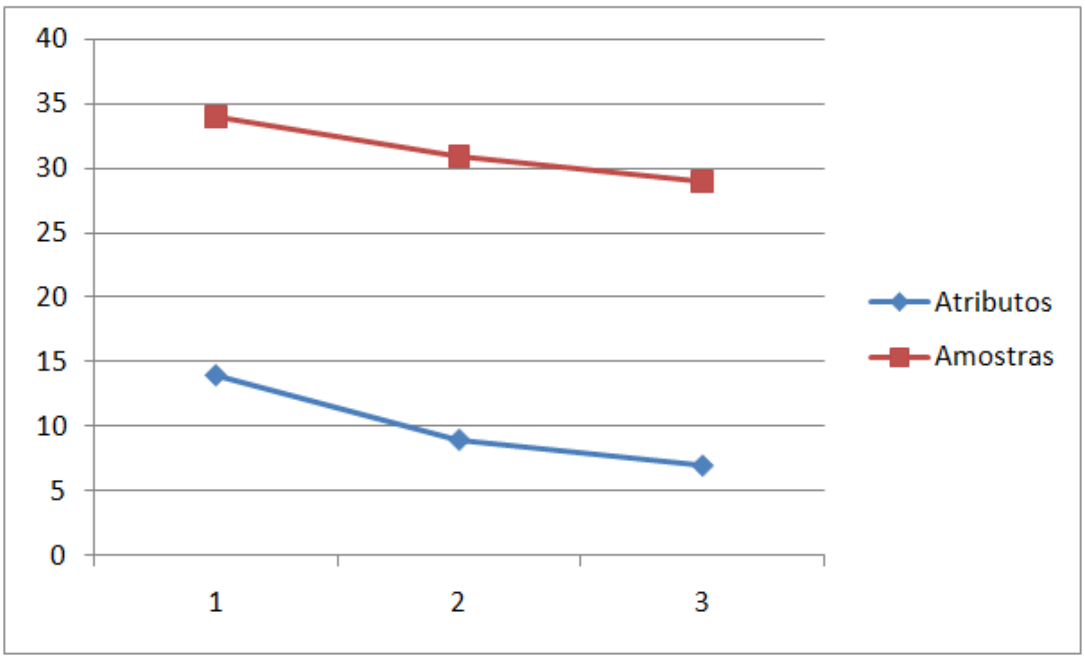
específicos para cada atributo, ao invés de um limiar para toda a base. O que mais interfere, assim como ocorreu na base KKI, é a dificuldade do Bimax em encontrar uma configuração de número de amostras e de atributos que encontrassem biclusters válidos em um tempo aceitável. Das 38 amostras que podem compor a base de treino (excluindo a amostra de teste), a Tabela 15 mostra que o Bimax agrupou menos amostras que o SSBimax, considerando as amostras nos experimentos 1, 2 e 3.

Tabela 15 – Média de amostras rotuladas por biclustering

Métodos	**1**	**2**	**3**
SSBimax	34	31	29
Bimax	20	23	24

As médias de atributos utilizados na construção do bicluster, como podem ser vistas na Figura 11 para os três experimentos, foram de 14, 9 e 7 para o SSBimax, 5, 7 e 5 para o Bimax, enquanto que o PCC selecionou em média mais de 200 atributos como relevantes. Como mostra a Figura 10, mais uma vez, o SSBimax eliminou a maldição de dimensionalidade, agora com quantidades de atributos bem menores que o número de amostras.

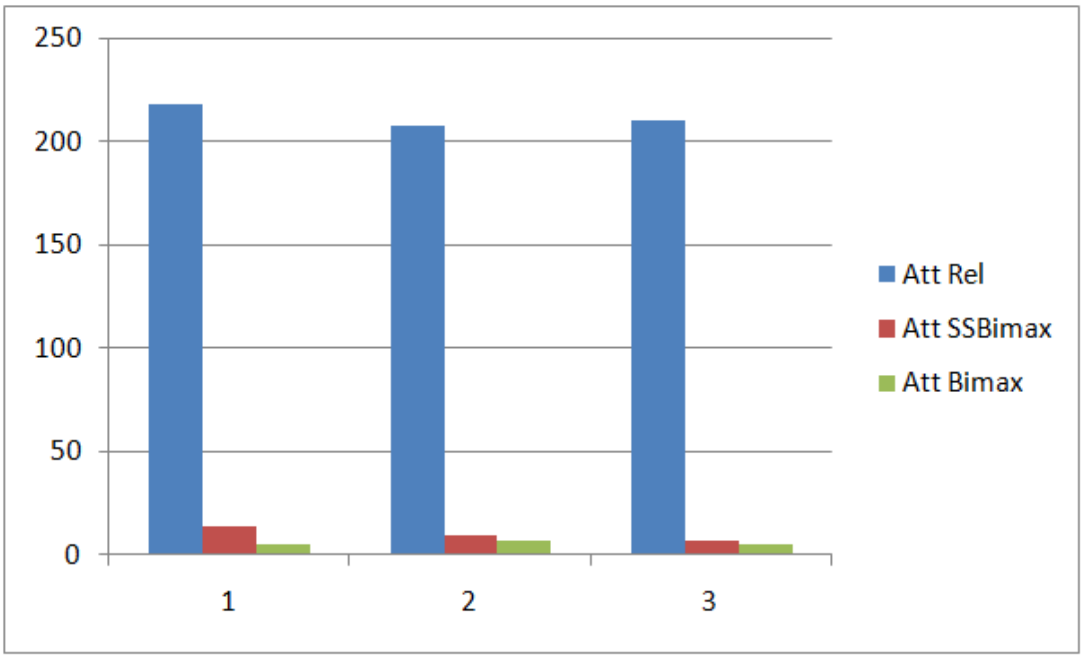
Figura 10 – Relação amostras x atributos no SSBimax



Fonte: Autoria própria

Nas seções a seguir estão, os resultados separados e detalhados de cada classificador para a base NeuroIMAGE nos experimentos realizados.

Figura 11 – Média de atributos por experimento na base NeuroIMAGE



Fonte: Autoria própria

5.2.1 Resultados SVM

Como já mostrado, a média original do classificador SVM para a base NeuroIMAGE foi de 0.6. A Tabela 16 mostra as médias de qualidade quando a base original sofreu redução de rotulação inicial para até 70% e, desde o início, é notável a queda das médias. Além disso, à medida que a quantidade de amostras foi reduzida, gerar subconjuntos de atributos que conseguiam diferenciar classes foi se tornando mais complicado. Como pode ser visto no experimento 1, 90% da base original rotulada já foi o suficiente para reduzir a qualidade de classificação em 0.05, mas o PCC conseguiu melhorar por 0.03. O experimento 2 mostra uma queda maior do PCC, resultando na igualdade de qualidade em 0.54 e, por fim, o experimento 3 já mostra a inferioridade de qualidade do PCC. Apesar do balanceamento, a queda de qualidade ao diminuir a quantidade de amostras que participam da classificação se deve à pequena quantidade de amostras na base.

Apesar das quedas de qualidades, houve métodos semi-supervisionados que conseguiram aumentar a qualidade, de forma a superar também a média original. Foram os casos do S3VM com média 0.62 e o SSBimax com média 0.60, obtidos no experimento 1.

No experimento 2, os mesmos métodos foram os únicos a melhorar a qualidade do PCC, mas não suficiente para superar a média original e o SSBimax foi o melhor desta vez. Inclusive, se fosse considerado apenas esse classificador, o SSBimax não permitiria realizar a redução da base NeuroIMAGE para 70% de amostras rotuladas.

Como já dito, os métodos S3VM e SSBimax foram os destaques para o classificador

Tabela 16 – Resultados SVM para a base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.55	0.54	0.53
Com PCC	0.58	0.54	0.51
S3VM	0.62	0.56	0.52
COP-KMeans	0.53	0.53	0.48
Grafo	0.51	0.45	0.50
Bimax	0.47	0.49	0.46
SSBimax	0.60	0.59	0.57

SVM. Ambos tiveram sequências de médias decrescentes, mas a queda sofrida pelo S3VM foi muito maior, enquanto o SSBimax sofreu uma pequena queda de 0.03. Assim, o SSBimax é considerado o melhor método semi-supervisionado aqui e o S3VM é o melhor método tradicional. O pior método foi o Bimax.

5.2.2 Resultados KNN

O classificador KNN obteve a média 0.45, a menor média de qualidade para a base NeuroIMAGE. Em contra partida, reduzir a base original só fez aumentar as médias de classificação, como mostra a Tabela 17. Com 90% das amostras rotuladas, a média subiu de 0.45 para 0.46 porém, nos experimentos 2 e 3 essa média subiu para 0.5. Além disso, o PCC foi responsável por aumentar ainda mais essas médias, com 0.51, 0.52 e 0.51 nos experimentos 1, 2 e 3, respectivamente.

Tabela 17 – Resultados KNN para a base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.46	0.50	0.50
Com PCC	0.51	0.52	0.51
S3VM	0.52	0.51	0.50
COP-KMeans	0.49	0.51	0.49
Grafo	0.50	0.50	0.51
Bimax	0.48	0.48	0.50
SSBimax	0.59	0.57	0.57

Com relação aos métodos semi-supervisionados, todos foram superiores à média original em todos os experimentos. Comparando às médias do PCC, no experimento 1 somente o S3VM e o SSBimax conseguiram melhorar as médias e o Bimax foi o pior. No experimento 2, somente o SSBimax foi capaz de aumentar a média de 0.52. Já no experimento 3, com 70% de amostras inicialmente rotuladas, o método baseado em grafo

igualou a média do PCC e mais uma vez o SSBimax foi superior. Dessa forma, se o critério de parada dependesse somente do KNN, o método SSBimax possibilitaria a redução da base NeuroIMAGE para usar 60% da quantidade de amostras originalmente rotuladas.

Ao se tratar dos métodos tradicionais, o S3VM foi mais uma vez o melhor método semi-supervisionado e o Bimax foi eleito o pior. Assim como aconteceu no classificador anterior, a pequena base NeuroIMAGE vai mostrando que o SSBimax tem grande vantagem sobre o Bimax.

5.2.3 Resultados Perceptron

O Perceptron foi outro classificador a obter média de qualidade baixa, 0.47. Assim como aconteceu com o KNN, os experimentos apresentados na Tabela 18 mostram que o Perceptron se comportou bem nas versões reduzidas da base NeuroIMAGE, resultando nas médias 0.49, 0.52 e 0.52. Melhores ainda foram as médias obtidas pelo PCC, que aumentou as qualidades nos três experimentos.

Por outro lado, a alta média obtida pelo PCC no experimento 1 (0.56) não ajudou os métodos semi-supervisionados a melhoras a qualidade através da rotulação automática. A exceção foi o SSBimax, que obteve 0.63 de média e foi muito superior aos outros métodos. No experimento 2, com a queda de qualidade do PCC (0.52), somente o Bimax não foi capaz de rotular amostras e aumentar a taxa de acerto. Por fim, no experimento 3 o PCC voltou a aumentar a média de qualidade e, da mesmo forma que ocorreu no experimento 1, somente o SSBimax foi superior.

Em todas as oportunidades, considerando todos os métodos semi-supervisionados em todos os experimentos, a realização de um experimento 4 seria possível. Principalmente se tratando do SSBimax, que obteve médias altas como 0.63, 0.60 e 0.58, se tornando no melhor método para este classificador. O S3VM, o COP-Kmeans e o método baseado em grafo tiveram comportamentos bem próximos, diferente do Bimax e suas médias inferiores.

Tabela 18 – Resultados Perceptron para a base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.49	0.52	0.52
Com PCC	0.56	0.52	0.55
S3VM	0.50	0.57	0.49
COP-KMeans	0.53	0.53	0.50
Grafo	0.49	0.54	0.49
Bimax	0.51	0.48	0.47
SSBimax	0.63	0.60	0.58

5.2.4 Resultados C4.5

Mais uma vez, o C4.5 obteve a média de qualidade mais alta, com 0.78. Assim como aconteceu na base KKI, com a redução de apenas 10% da quantidade de amostras rotuladas para construção de modelos houve uma queda na qualidade de 0.28. A 19 mostra também que nos experimentos 2 e 3, respectivamente, as médias aumentaram levemente de 0.50 para 0.51 e 0.53. Reduzir também a quantidade de atributos das amostras surtiu uma pequena melhora no experimento 1, quando a média aumentou para 0.56, e no experimento 2, quando aumentou para 0.52.

A alta média obtida pelo C4.5 com o grande efeito negativo provocado pelas reduções da base original, devido a pouca quantidade de amostras, tornou a tarefa de rotular automaticamente muito mais difícil. Ainda na Tabela 19, é possível ver que o SSBimax foi capaz de melhorar a média de qualidade para 0.66. Apesar de ainda muito distante da média original, foi o suficiente para a realização do experimento 2, graças à ação dos demais classificadores. Além do SSBimax, somente o S3VM foi superior ao PCC.

No experimento 2, os métodos que melhoraram a média obtida pelo PCC foram o S3VM (0.53), o método baseado em grafo (0.53) e o SSBimax (0.60). Como eles também foram melhores que a base sem PCC, pode-se dizer que a rotulação de novas amostras foi vantajosa.

No experimento 3, somente o método baseado em grafo foi inferior ao PCC. Porém, o SSBimax foi o único a a fazer da rotulação automática uma característica válida pois, foi o único a ser melhor que a base sem PCC. A média obtida neste experimento (0.61) foi abaixo da média original por 0.27, não sendo o bastante para que o C4.5 juntamente com os outros classificadores pudessem reduzir a base NeuroIMAGE para 60% de amostras rotuladas

Mais uma vez, o método SSBimax foi quem mais chegou perto da média original, sendo o melhor método para este classificador. Já a Minimização de Energia Harmônica (grafo) foi o pior método tradicional.

Tabela 19 – Resultados C4.5 para a base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.50	0.51	0.53
Com PCC	0.56	0.52	0.46
S3VM	0.59	0.53	0.50
COP-KMeans	0.53	0.51	0.50
Grafo	0.49	0.53	0.44
Bimax	0.49	0.50	0.49
SSBimax	0.66	0.60	0.61

5.2.5 Resultados Naive Bayes

O último classificador analisado nesta base é o Naive Bayes, que obteve 0.58 de média para a base NeuroIMAGE original. Diferente do que aconteceu na base KKI, a redução da base nos três experimentos afetou a qualidade de classificação deste classificador. Ainda mais, como mostras a Tabela 20, o PCC não foi eficaz e resultou em médias ainda menores. As médias 0.51, 0.53 e 0.46 obtidas apenas pela redução da base foram baixadas para 0.46, 0.46 e 0.44 quando a base foi aplicada ao PCC nos experimentos 1, 2 e 3, respectivamente.

Nos experimentos 1 e 2, somente o método baseado em grafo não conseguiu aumentar a média de qualidade. Os demais fizeram bom uso das amostras não rotuladas, com destaque maior para o SSBimax, que alcançou as altas médias 0.65 e 0.61.

No último experimento, apesar de todos os métodos semi-supervisionados terem melhorado a qualidade do PCC, o SSBimax sofreu uma queda de média para 0.56, que o fez ficar abaixo da média original e concretizou o fim dos experimentos. Mesmo assim, é possível afirmar que o SSBimax foi o melhor método semi-supervisionado para o Naive Bayes e o método baseado em grafo foi o pior entre o métodos tradicionais.

Tabela 20 – Resultados Naive Bayes para a base NeuroIMAGE

Métodos	**1**	**2**	**3**
Sem PCC	0.51	0.53	0.46
Com PCC	0.46	0.46	0.44
S3VM	0.52	0.50	0.50
COP-KMeans	0.47	0.50	0.46
Grafo	0.46	0.44	0.49
Bimax	0.50	0.49	0.47
SSBimax	0.65	0.61	0.56

5.3 Base NYU

A NYU é a maior base estudada neste trabalho e as quantidades de suas amostras são balanceadas entre as classes TDAH e Typ. A Tabela 21 mostra as médias de qualidade para cada classificador nessa base. É possível notar a baixa diferença entre elas e, diferente das outras bases analisadas até agora, a baixa média obtida pelo C4.5 (0.39), resultando em uma média de qualidade na base original de 0.48. Esse valor foi baixo o suficiente para que todos os métodos semi-supervisionados conseguissem aumentar a qualidade média usando apenas 10% de amostras previamente rotuladas (Tabela 22).

Tabela 21 – Resultados para base NYU original

SVM	**KNN**	**Perceptron**	**C4.5**	**NB**
0.53	0.52	0.45	0.39	0.49

Como o principal intuito desses experimentos é verificar quanto o método proposto SSBimax é capaz de reduzir uma base fMRI com foco no TDAH, e por motivo de agilizar o processo de obtenção de resultados, o SSBimax utilizou nos experimentos 3-9 as médias de atributos utilizados pelos experimentos 1 e 2.

Por isso, as taxas apresentadas na tabela a seguir não exploram todas as funcionalidades oferecidas pelo método baseado em biclustering. No caso do método clássico Bimax, a obtenção dos resultados foi feita de forma completa, já que só foram encontrados biclusters válidos em tempo aceitável usando baixos números de amostras e atributos.

Tabela 22 – Resultados da base NYU

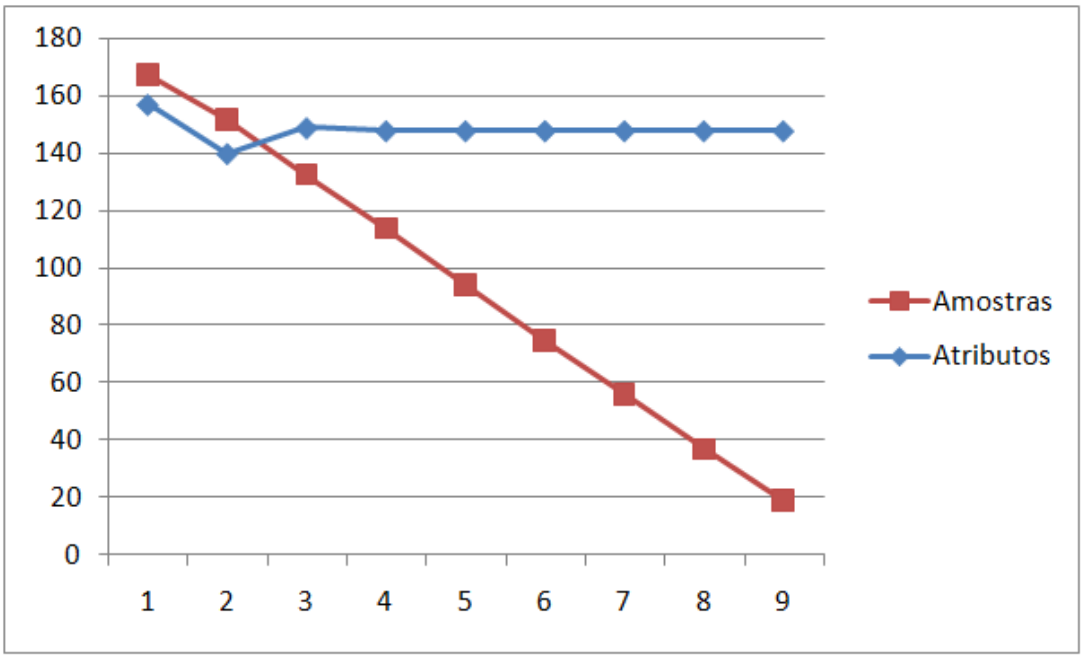
Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.52	0.51	0.51	0.49	0.51	0.51	0.50	0.51	0.50
Com PCC	0.50	0.50	0.52	0.50	0.51	0.51	0.51	0.51	0.51
S3VM	0.49	0.49	0.49	0.50	0.50	0.51	0.50	0.50	0.50
COP-KMeans	0.51	0.50	0.50	0.52	0.51	0.52	0.49	0.50	0.49
Grafo	0.51	0.50	0.51	0.51	0.51	0.51	0.50	0.50	0.51
Bimax	0.51	0.50	0.52	0.51	0.50	0.50	0.51	0.50	0.49
SSBimax	0.52	0.52	0.51	0.51	0.51	0.52	0.50	0.51	0.51

Mesmo assim, a Tabela 22 mostra que o SSBimax fez parte da melhor solução em quase todos os experimentos, estando de fora somente nos experimentos 3, 4 e 7. Já o Bimax esteve entre os melhores métodos em cincos experimentos. Com relação aos demais métodos semi-supervisionados, o método baseado na teoria de grafos foi o melhor em 3 oportunidades, o COP-Kmeans em duas vezes, e por fim, o S3VM não foi o melhor método em nenhum experimento.

Uma característica interessante nesses experimentos foi o ganho de qualidade por parte da base sem PCC, independente da quantidade de amostras previamente rotuladas. A maior média foi obtida no experimento 1, 0.52, e a menor foi obtida no experimento 4, 0.49. Nos demais experimentos houve variação entre 0.50 e 0.51.

Da mesma forma, o PCC conseguiu representar bem todas as bases parcialmente rotuladas, melhorando ou igualando a base sem PCC nos experimentos 3-9, a partir da redução do conjunto de atributos para o nível de significância em 0.01. A partir do experimento 5, a média do PCC se manteve em 0.51.

Figura 12 – Relação amostras x atributos no SSBimax



Fonte: Autoria própria

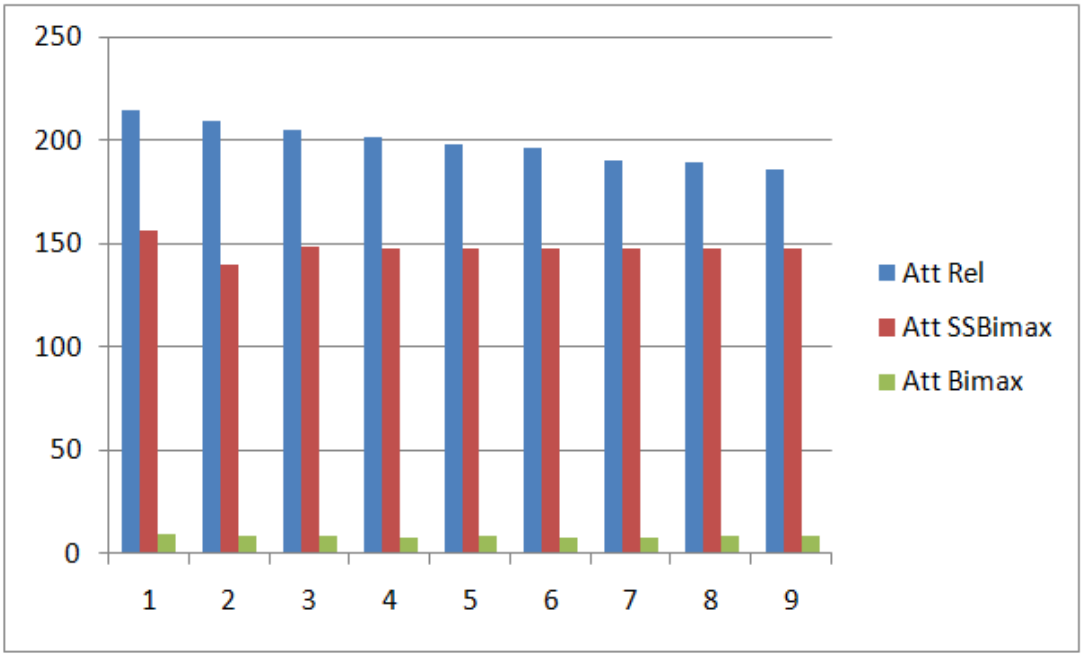
Eliminar a maldição da dimensionalidade também não foi prioridade para essa base. A Figura 12 mostra que nos experimentos 1 e 2, quando o SSBimax usou o parâmetro de escolha mínima de atributos iterativamente, a maldição havia desaparecido. A partir do experimento 3, o mínimo de atributos foi fixado em 150 pois, corresponde à média de atributos nos dois primeiros experimentos. Esse valor obedece o requisito de não ultrapassar a quantidade de atributos considerados relevantes pelo PCC, como visto na Figura 12. Como visto anteriormente, tanto as amostras quanto os atributos utilizados pelo Bimax foram em quantidades muito baixas. A Figura 13 mostra a diferença de atributos usados para o SSBimax e os considerados relevantes pelo PCC.

Nas seções a seguir estão detalhados os nove experimentos da base NYU, separados por cada classificador. É mostrado que todos esses experimentos foram possíveis graças à baixa perda de qualidade com a redução das bases de treino; ao baixo valor de qualidade da base original para o classificador C4.5, o que facilitou a melhoria por parte dos métodos semi-supervisionados, incluindo o SSBimax; a grande quantidade de ocasiões em que os classificadores Perceptron e Naive Bayes também foram melhorados e a baixa variação de piora para os classificadores SVM e KNN.

5.3.1 Resultados SVM

A média original do classificador SVM para a base NYU foi de 0.53, a maior desta base. Como apresenta a Tabela 23, a redução da quantidade de amostras rotuladas para

Figura 13 – Média de atributos por experimento na base NYU



Fonte: Autoria própria

90% e 70% não afetou a média de qualidade. Nos demais experimentos, as baixas quedas obtidas foram as que seguem: experimentos 2, 5, 6 e 8 obtiveram 0.52; experimentos 4 e 7 obtiveram 0.51 e, por fim, o experimento 9 foi quando a menor média foi obtidas, com 0.50. Ao aplicar o PCC, as médias de qualidade não foram melhoradas na maioria dos experimentos, sendo as únicas exceções ocorridas no experimento 7, onde as médias foram iguais, e no experimento 9, onde o PCC foi superior. A queda sofrida pelo PCC não foi grande, em que em uma única situação foi uma queda de 0.04 de média no experimento 2. Nos demais experimentos, as maiores quedas foram de 0.02 de média.

Tabela 23 – Resultados SVM para a base NYU

Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.53	0.52	0.53	0.51	0.52	0.52	0.51	0.52	0.50
Com PCC	0.51	0.48	0.52	0.49	0.51	0.50	0.51	0.51	0.51
S3VM	0.48	0.48	0.50	0.50	0.50	0.51	0.51	0.48	0.49
COP-KMeans	0.51	0.49	0.52	0.51	0.51	0.54	0.49	0.49	0.50
Grafo	0.51	0.50	0.50	0.50	0.51	0.51	0.50	0.49	0.52
Bimax	0.50	0.50	0.51	0.51	0.50	0.51	0.51	0.50	0.50
SSBimax	0.50	0.52	0.50	0.49	0.51	0.51	0.51	0.51	0.52

Se fosse considerado apenas este classificador, os testes seriam finalizados no experimento 2, já que a média obtida pelo SSBimax foi 0.50, menor que a média original.

De fato, apenas o COP-Kmeans conseguiu superar a média original, no experimento 6, que também foi maior que a versão reduzida sem PCC. Apesar disso, a queda de qualidade foi tão baixa que, ao reduzir a base para apenas 10% de amostras rotuladas, a média 0.52 obtida pel SSBimax pode ser vista com bons olhos, principalmente se tratando de uma base grande como o NYU.

Além do experimento 6, o COP-Kmeans melhorou as médias do PCC nos experimentos 2 e 4. O método SSL baseado em grafo foi quem foi superior ao PCC em mais experimentos (2, 4, 6 e 9), sendo o experimento 9 também superior à base sem PCC. Da mesma forma, o SSBimax obteve média 0.52 no experimento 9. O SSBimax, que foi o método com a maior média em todos os experimentos, sendo eleito o melhor para este classificador, elevou as médias dos experimentos 2, 6 e 9. O pior método, o S3VM, só melhorou os experimentos 4 e 6.

5.3.2 Resultados KNN

No caso do KNN a média original foi de 0.52, um pouco menor que o classificador anterior. A redução da base NYU para até 80% não foi suficiente para afetar a classificação pelo KNN. Como pode ser vista na Tabela 24, a partir desse experimento houve pequenas quedas de média, como: 0.51 nos experimentos 3 e 6, 0.50 nos experimentos 4 e 7 e, por fim, 0.49 nos dois últimos experimentos e no experimento 5.

Diferente das bases que usaram o conjunto completo de atributos, as bases com amostras reduzidas com PCC tiveram as menores médias nos dois primeiros experimentos. A partir do experimento 3 o PCC igualou ou melhorou as médias do Sem PCC, igualando a média original nos experimentos 3, 6 e 9.

Tabela 24 – Resultados KNN para a base NYU

Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.52	0.52	0.51	0.50	0.49	0.51	0.50	0.49	0.49
Com PCC	0.49	0.49	0.52	0.50	0.51	0.52	0.51	0.51	0.52
S3VM	0.48	0.49	0.50	0.50	0.51	0.53	0.50	0.50	0.51
COP-KMeans	0.50	0.49	0.50	0.51	0.52	0.52	0.48	0.50	0.50
Grafo	0.50	0.50	0.51	0.50	0.51	0.53	0.50	0.49	0.51
Bimax	0.51	0.49	0.52	0.52	0.51	0.52	0.49	0.51	0.50
SSBimax	0.53	0.53	0.52	0.51	0.51	0.53	0.50	0.51	0.51

O SSBimax seria capaz de reduzir a base até o experimento 4 somente avaliando o KNN. Este método superou a média original nos experimentos 1, 2 e 6, foi melhor que o PCC também no experimento 4 e em nenhum experimento foi pior que a versão reduzida sem PCC. Dessa forma, o SSBimax foi o melhor método aqui.

Por outro lado, o método S3VM foi considerado o pior método porque só foi superior que o PCC no experimento 6, única oportunidade maior também que a média original. Empatado com a mesma média em todos os experimentos, o COP-Kmeans também foi o destaque negativo, apesar de ser melhor que o PCC em três oportunidades (1, 4 e 5).

5.3.3 Resultados Perceptron

A média original obtida pelo classificador Perceptron foi 0.45. A Tabela 25 mostra que a menor média encontrada para esse classificador foi 0.47, obtida pelo método S3VM no experimento 5. Dessa forma, todas as médias obtidas foram superiores à média original, incluindo as obtidas pelo SSBimax, identificando o Perceptron como um dos responsáveis pela realização dos experimentos até a base NYU ser reduzida a 10% de amostras rotuladas.

A redução da dimensionalidade não foi certeza de melhora nas qualidades, já que somente foi superior nos experimentos 4 e 7. Por outro lado, quando foi inferior, a perda de qualidade não foi grande, como mostram os experimentos 1 e 2 (maiores diferenças).

O melhor método semi-supervisionado usando o Perceptron foi o SSBimax. Além de obter a maior média para todos os experimentos, ele foi melhor que o PCC nos experimentos 1, 2 e 6, e foi melhor que a versão sem PCC nos experiimentos 2, 4 e 6. Comparando o SSBimax com o Bimax tradicional, esse último só foi melhor no experimento 4. Isso mostra a vantagem que as modificações do SSBimax têm sobre a versão tradicional.

Como pior método, o S3VM só foi melhor que o PCC no experimento 1 e em nenhuma oportunidade foi melhor que a versão sem PCC, além disso, como já foi descrito, o S3VM foi o método que obteve a pior média, no experimento 5.

Tabela 25 – Resultados Perceptron para a base NYU

Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.52	0.50	0.51	0.49	0.51	0.52	0.50	0.53	0.51
Com PCC	0.49	0.50	0.51	0.50	0.51	0.51	0.51	0.52	0.50
S3VM	0.50	0.50	0.49	0.49	0.47	0.50	0.48	0.51	0.49
COP-KMeans	0.48	0.51	0.49	0.51	0.50	0.51	0.50	0.50	0.49
Grafo	0.50	0.50	0.51	0.51	0.50	0.51	0.50	0.50	0.51
Bimax	0.50	0.49	0.51	0.50	0.49	0.50	0.51	0.50	0.49
SSBimax	0.51	0.51	0.49	0.50	0.50	0.53	0.50	0.52	0.50

5.3.4 Resultados C4.5

Bastante diferente do que aconteceu na bases anteriormente analisadas, a menor média original na base NYU foi 0.39, obtida pelo classificador C4.5. Como resultado

apresentado na Tabela 26, todos os métodos semi-supervisionados foram muito superiores à essa média e o classificador se mostrou ser outro grande responsável pela redução ter chegado até o experimento 1.

Mais uma vez, a redução de dimensionalidade não melhorou as médias do conjunto completo de atributos mas a queda não foi grande, mostrando que a redução também tem vantagem. Com relação aos métodos SSL não houve diferença. O melhor foi SSBimax, com leve vantagem sobre o Bimax provocada no experimento 6, e o pior foi o S3VM (melhor que o PCC somente no experimento 9).

A rotulação automática do SSBimax melhorou as médias do PCC nos experimentos 1, 2, 4 e 6. No experimento 4 ele também foi superior à versão sem PCC. O Bimax tradicional foi quem teve taxas superiores ao PCC por mais vezes (experimentos 1, 2, 3, 4 e 7).

Tabela 26 – Resultados C4.5 para a base NYU

Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.52	0.52	0.51	0.48	0.52	0.52	0.51	0.52	0.50
Com PCC	0.50	0.50	0.51	0.49	0.52	0.50	0.50	0.50	0.49
S3VM	0.48	0.50	0.47	0.47	0.51	0.50	0.49	0.50	0.50
COP-KMeans	0.52	0.49	0.49	0.52	0.50	0.48	0.50	0.51	0.49
Grafo	0.50	0.50	0.50	0.52	0.50	0.49	0.51	0.50	0.50
Bimax	0.52	0.51	0.52	0.52	0.51	0.49	0.51	0.49	0.49
SSBimax	0.52	0.51	0.50	0.51	0.52	0.52	0.50	0.50	0.49

5.3.5 Resultados Naive Bayes

O último classificador analisado para esta base é o Naive Bayes. Sua média original foi 0.49, suficiente para não afetar nas classificações das bases NYU reduzidas. A Tabela 27 mostra que nas reduções, os experimentos 4 e 9 apresentaram as mesmas médias originais e nos demais experimentos a redução corrente foi melhor (0.50 de média). Ainda melhor, o PCC funcionou de forma bastante eficaz ao aumentar as médias em todos os experimentos, mostrando que os subconjuntos de atributos considerados relevantes distinguiam as classes melhor que usando o conjunto completo de atributos.

A eficácia do PCC refletiu nas médias de qualidades obtidas pelos métodos semi-supervisionados. O SSBimax, mais uma vez eleito o melhor método, teve qualidade igual ou superior ao PCC até o experimento 6, quando suas medidas começaram a decrescer. Novamente, este classificador também foi responsável por ter sido possível reduzir a base NYU para apenas 10% de amostras rotuladas. Assim como SSBimax, o COP-Kmeans também teve suas medidas pioradas nos últimos experimentos, quando

obteve no experimento 9 a única média abaixo da original (0.48). Por fim, o pior método foi o S3VM.

Tabela 27 – Resultados Naive Bayes para a base NYU

Métodos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sem PCC	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.50	0.49
Com PCC	0.52	0.51	0.53	0.52	0.51	0.53	0.51	0.52	0.51
S3VM	0.51	0.50	0.50	0.50	0.51	0.52	0.50	0.49	0.51
COP-KMeans	0.53	0.52	0.52	0.53	0.51	0.53	0.49	0.50	0.48
Grafo	0.53	0.51	0.53	0.52	0.52	0.52	0.50	0.50	0.52
Bimax	0.51	0.50	0.52	0.52	0.51	0.52	0.51	0.51	0.51
SSBimax	0.53	0.53	0.53	0.53	0.53	0.53	0.50	0.51	0.51

5.4 Base Peking

Para finalizar os experimentos deste trabalho, a Peking é a segunda maior base analisada, com 150 amostras. Diferente da NYU, base com a maior quantidade de amostras neste trabalho, essa base não é balanceada e a dificuldade na construção de modelos que diferenciam classes é bem maior. Na Tabela 28, a maior medida de qualidade pertenceu ao classificador C4.5 (0.65), como semelhança com as bases KKI e NeuroIMAGE, e a menor medida de qualidade pertenceu ao KNN (0,50).

Com a média original de 0.58, os modelos construídos nesta base foram fracos a ponto de nenhum método semi-supervisionado ter sido capaz de superar a média original. Por conta disso, a base Peking foi reduzida somente usando 90% da quantidade de amostras inicialmente rotuladas e, por falta da necessidade de dividir esta seção em subseções de um experimento só, os resultados obtidos pelos cinco classificadores e a média geral são apresentados na Tabela 29.

Tabela 28 – Resultados para base Peking original

SVM	**KNN**	**Perceptron**	**C4.5**	**NB**
0,58	0,50	0,55	0,65	0,59

A Tabela 29 mostra que os métodos COP-Kmeans e o SSBimax foram os melhores, com 0.03 de diferença entre as suas médias e a média original. Como forma de desempatar, pode-se dizer que o SSBimax ganha essa disputa por dois fatores: utilizar menos atributos, já que usou em média 149 atributos, enquanto o COP-Kmeans utilizou em média 390 atributos, identificados pelo PCC como relevantes; o segundo fator vem da quantidade de

amostras agrupadas, em que o SSBimax agrupou cerca de 134, menos que os 149 agrupados pelo COP-Kmeans.

A ausência do Bimax original nos experimentos desta base se deu pelo fato de que ele não foi capaz de encontrar biclusters válidos para todas as iterações, reforçando a vantagem que o SSBimax levou em todas as bases aqui analisadas.

Apenas a redução da base já foi o suficiente para diminuir as qualidades de classificações. A queda para 0.54 de média teve como maiores responsáveis os classificadores KNN e C4.5 por obterem médias distantes dos demais. Em contra partida, o Naive Bayes conseguiu igualar a média original e o SVM ficou abaixo por apenas 0.01 de média.

Com a aplicação do PCC, a média geral se manteve mas por causa de um acontecimento diferente. O SVM foi o único classificador a ser inferior, com sua qualidade caindo de 0.57 para 0.52 e anulando todas as melhorias obtidas com o PCC. Essa queda influenciou os métodos semi-supervisionados a estacionarem suas médias também em 0.52, impossibilitando o prosseguimento do experimento. Para este classificador, o pior método foi o S3VM, único a obter 0.47 de média.

O KNN foi melhorado com o PCC e o SSBimax elevou ainda mais a sua média, para 0.55. Os demais métodos tiveram médias 0.53, como o S3VM e o COP-Kmeans, e 0.52 para o baseado em grafo, pior para este classificador. Para o Perceptron, o único método a realizar rotulação bem sucedida foi o COP-Kmeans, que chegou aos 0.55. SSBimax e o de grafo ficaram com 0.53 e o S3VM, pior novamente, ficou com 0.49. A média do S3VM se manteve como a pior para o C4.5, em que o baseado em grafo igualou a média do PCC, o COP-Kmeans aumentou para 0.55 e SSBimax foi o maior, com 0.56.

Por fim, o classificador Naive Bayes proporcionou as melhores médias, inclusive maior que a média original. Com a média 0.60 do PCC, tanto o SSBimax e a Minimização de Energia Harmônica obtiveram a mesma média, possibilitando a redução da base para 80% da base original de amostras rotuladas. Mais uma vez, o S3VM foi o pior com 0.50 de média.

Tabela 29 – Resultados da base Peking

Métodos	**SVM**	**KNN**	**Perceptron**	**C4.5**	**NB**	**Média**
Sem PCC	0.57	0.50	0.54	0.49	0.58	0.54
Com PCC	0.52	0.53	0.54	0.53	0.60	0.54
S3VM	0.47	0.53	0.49	0.49	0.50	0.50
COP-KMeans	0.52	0.53	0.55	0.55	0.59	0.55
Grafo	0.52	0.52	0.53	0.53	0.60	0.54
SSBimax	0.52	0.55	0.53	0.56	0.60	0.55

Capítulo 6
Considerações Finais

O objetivo principal deste trabalho foi de apresentar o SSBimax, um novo método de aprendizado semi-supervisionado baseado em biclustering, aplicado em matrizes de conectividade com foco no TDAH e obtidas pelas técnicas de fMRI.

Apesar de ser o transtorno psiquiátrico mais comumente diagnosticado entre crianças no mundo, estudos relacionando o TDAH não é tão popular quanto outras doenças psiquiátricas, como o Autismo ou Mal de Alzheimer. A partir deste fato vem a motivação de auxiliar em análises de ativação e diagnósticos desse transtorno, que afeta tanto o desenvolvimento cognitivo quanto o desenvolvimento social.

Como já é de conhecimento, muitos transtornos psiquiátricos são caracterizados pelo funcionamento anormal de algumas regiões cerebrais, e as técnicas de fMRI têm sido ferramentas importantes para um diagnóstico mais preciso dessas doenças.

Por conta do alto custo para rotular manualmente imagens funcionais de Ressonância Magnética do cérebro, se faz favorável o uso de técnicas computacionais baseadas no aprendizado semi-supervisionado. Além disso, poucas regiões cerebrais são afetadas pelo TDAH, o que torna essencial uma análise local dentro da grande quantidade de informação contida em uma base fMRI. Por fim, a presença de ruídos causados pela movimentação da cabeça podem comprometer a qualidade das imagens, tornando essa amostra não representativa para a base fMRI. A busca por agrupamentos de um subconjunto de regiões usando apenas as amostras representativas na base de dados sugere o uso de técnicas biclustering.

A fim de atender a todas as especificações anteriores, este trabalho apresentou o SSBimax e aplicou em quatro bases de matrizes de conectividade com foco no TDAH obtidas por técnicas fMRI (KKI, NeuroIMAGE, NYU e Peking). Os experimentos serviram para mostrar o poder de rotulação automática do SSBimax, que conseguiu reduzir a quantidade de amostras previamente rotuladas em três das quatro bases, mas ainda mantendo o modelo em um nível qualidade que equivaleu ou superou a qualidade de classificação da base original, onde somente uma amostra não era rotulada para compor a base de teste no LOOCV.

O maior destaque dos experimentos foi a base NYU. Por se tratar de uma base grande e balanceada, a construção de modelos confiáveis se tornou mais fácil. Como resultado, os métodos semi-supervisionados conseguiram manter a média de qualidade superior à qualidade original nos modelos construídos usando apenas 10% de amostras

rotuladas.

O poder de rotulação automática foi uma demonstração de que a estratégia de construir biclusters pode levar vantagem sobre outras estratégias usadas por métodos semi-supervisionados populares, como o S3VM (bastante usado em estudos envolvendo bases de neuroimagens), Minimização de Energia Harmônica e o COP-Kmeans. Essa vantagem teve como responsáveis a rotulação apenas de amostras representativas e o seu parâmetro iterativo de menor quantidade de atributos em um bicluster, que permitiu eliminar, na maioria das bases, a maldição da dimensionalidade. Vale lembrar que o Bimax busca por biclusters constantes, que é a forma mais simples, e que não são exaustivos, ou seja, nem todas as amostras devem ser agrupadas.

Considerando apenas esses métodos populares, o S3VM foi considerado o melhor método semi-supervisionado em duas bases (KKI e NeuroIMAGE), o que justifica o seu uso em estudos de neuroimagens. Nas demais, o COP-Kmeans foi melhor na Peking e quase empatou na NYU, onde o melhor foi o método baseado em grafo. Assim foi mostrado que não houve um método semi-supervisionado tradicional unânime nas bases de dados. Nos experimentos onde a base rotulada foi reduzida, o PCC se mostrou uma boa alternativa de redução de dimensionalidade. Nos exemplos em que não conseguiu melhorar a qualidade dos classificadores quando comparados à base usando o conjunto completo de atributos, ao menos a perda de qualidade não foi grande. Mesmo assim, os métodos semi-supervisionados ficaram dependentes da sequência de relevância estabelecida pelo PCC. Por mais que o uso do PCC tenha sido estimulado com base em estudos usando as mesmas bases fMRI, ele não tinha sido testado em bases reduzidas. Com isso, como proposta de um trabalho futuro, será realizado o estudo de diferentes métodos de seleção de atributos para avaliação de bases fMRI reduzidas. Por outro lado, pelo SSBimax utilizar apenas um subconjunto desses atributos, não importando o seu ranking ordenado pela relevância, foi capaz de contornar mais esse problema graças ao uso dos parâmetros de quantidade mínima e máxima de atributos.

Além da construção de biclusters ter levado vantagem sobre as estratégias dos métodos semi-supervisionados tradicionais, experimentos como os da base NeuroIMAGE mostram também a superioridade da utilização de faixa de limiares sobre o limiar único global utilizado pelo Bimax na construção de biclusters melhores. Um fato que demonstrou essa vantagem foi a dificuldade do Bimax em gerar biclusters pois, sempre foi requerido pequenas quantidades de amostras, o que diminui a confiabilidade dos biclusters, e de atributos, pelo motivo de que conjuntos maiores de atributos não compartilharem comportamentos com mais amostras ao mesmo tempo.

O prosseguimento dos experimentos dependia da média geral do método SSBimax. Na base KKI, a relação do método semi-supervisionado com os classificadores se deu da seguinte forma: SVM e Naive Bayes foram favoráveis à redução da base para 60%

de amostras rotuladas; KNN e Perceptron obtiveram médias menores mas nada que afetasse muito rendimento geral; o grande responsável foi o C4.5 por média 0.09 inferior ao original. Da mesma forma ocorreu na base NeuroIMAGE, o classificador C4.5 foi o responsável pela descontinuidade dos experimento por causa de média 0.17 menor que o original; desta vez, o KNN e o Perceptron foram superiores e o SVM e Naive Bayes foram inferiores com pequena diferença de qualidade. Na base NYU, apesar de ter sido realizados os nove experimentos, os classificadores SVM e KNN foram os que mais dificultaram a realização dos experimentos pelas suas médias abaixo do original. No único experimento da base Peking, os classificadores KNN e Naive Bayes foram favoráveis à continuidade do experimento, mas foi impedido principalmente pelo SVM e C4.5.

As bases fMRI analisadas neste trabalho têm revelado vários padrões com relação ao TDAH, com exceção da base Peking. Em (GARCIA; PARAISO; NIEVOLA, 2017), a base Peking foi a única que os métodos de seleção de atributos não conseguiram aumentar o aproveitamento dos classificadores para predição da doença. Neste trabalho, apesar do SSBimax ter conseguido taxas de qualidade melhores do que as obtidas nos trabalhos supracitados, a base Peking foi a única em que a redução de informação foi impossível para reconhecimento de algum padrão, para todos os métodos semi-supervisionados. Assim, outras técnicas de reconhecimento de padrão serão abordadas para a análise desta base a fim de desmistificar essa base que tem se mostrado tão difícil de ser trabalhada.

Diferentes métodos semi-supervisionados baseados em outras estratégias biclustering (LAS e SAMBA) tiveram alguns experimentos prévios porém, os melhores biclusters considerados por essas estratégias não foram capazes de sobressair que nem o SSBimax. Como possível problema para esses métodos e mais uma proposta de trabalho futuro, serão estudadas formas de combinar múltiplos biclusters específicos de classes diferentes a fim de construir modelos mais completos e exatos, resultando também em versões evoluídas do SSBimax.

Por fim, o método semi-supervisionado baseado em biclustering pode ser considerado como bem sucedido ao ser aplicado em bases fMRI, de forma que possa vir a ser uma solução alternativa no auxílio do diagnóstico do TDAH. Por outro lado, para esse cenário se concretizar, um longo caminho deverá ser percorrido até que se torne confiável ao ponto de diagnosticar e de fornecer conhecimento mais profundo sobre o funcionamento desta disfunção que atinge tanta criança no mundo.

Referências

AKIL, Huda; MARTONE, Maryann E; ESSEN, David C Van. Challenges and opportunities in mining neuroscience data. *science*, American Association for the Advancement of Science, v. 331, n. 6018, p. 708–712, 2011. Citado na página 13.

ASHBURNER, Michael; BALL, Catherine A; BLAKE, Judith A; BOTSTEIN, David; BUTLER, Heather; CHERRY, J Michael; DAVIS, Allan P; DOLINSKI, Kara; DWIGHT, Selina S; EPPIG, Janan T et al. Gene ontology: tool for the unification of biology. *Nature genetics*, Nature Publishing Group, v. 25, n. 1, p. 25, 2000. Citado na página 31.

BANASCHEWSKI, Tobias; ZUDDAS, Alessandro; ASHERSON, Philip; COGHILL, David; BUITELAAR, Jan; DANCKAERTS, Marina; DÖPFNER, Manfred; SONUGA-BARKE, Edmund. *ADHD and hyperkinetic disorder*. [S.l.]: Oxford Psychiatry Library, 2015. Citado na página 13.

BANSAL, Ravi; STAIB, Lawrence H; LAINE, Andrew F; HAO, Xuejun; XU, Dongrong; LIU, Jun; WEISSMAN, Myrna; PETERSON, Bradley S. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. *PloS one*, Public Library of Science, v. 7, n. 12, p. e50698, 2012. Citado na página 16.

BARKOW, Simon; BLEULER, Stefan; PRELIĆ, Amela; ZIMMERMANN, Philip; ZITZLER, Eckart. Bicat: a biclustering analysis toolbox. *Bioinformatics*, Oxford University Press, v. 22, n. 10, p. 1282–1283, 2006. Citado na página 31.

BASU, Sugato; BANERJEE, Arindam; MOONEY, Raymond. Semi-supervised clustering by seeding. In: CITESEER. *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. [S.l.], 2002. Citado na página 30.

BEHRENS, Timothy EJ; SPORNS, Olaf. Human connectomics. *Current opinion in neurobiology*, Elsevier, v. 22, n. 1, p. 144–153, 2012. Citado 2 vezes nas páginas 17 e 18.

BROWN, Jesse A; RUDIE, Jeffrey D; BANDROWSKI, Anita; HORN, John D Van; BOOKHEIMER, Susan Y. The ucla multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. *Frontiers in neuroinformatics*, Frontiers, v. 6, p. 28, 2012. Citado 2 vezes nas páginas 18 e 41.

BUSYGIN, Stanislav; BOYKO, Nikita; PARDALOS, Panos M; BEWERNITZ, Michael; GHACIBEH, Georges. Biclustering eeg data from epileptic patients treated with vagus nerve stimulation. In: AIP. *AIP Conference Proceedings*. [S.l.], 2007. v. 953, n. 1, p. 220–231. Citado na página 16.

CALIFANO, Andrea; STOLOVITZKY, Gustavo; TU, Yuhai et al. Analysis of gene expression microarrays for phenotype classification. In: *Ismb*. [S.l.: s.n.], 2000. v. 8, p. 75–85. Citado na página 35.

CHAPELLE, Olivier; SCHOLKOPF, Bernhard; ZIEN, Alexander. Introduction to semi-supervised learning. *Semi-supervised learning (chapelle, o. et al., eds.; 2006)*, The MIT Press, p. 1–8, 2006. Citado 4 vezes nas páginas 14, 22, 24 e 25.

CHENG, Yizong; CHURCH, George M. Biclustering of expression data. In: *Ismb*. [S.l.: s.n.], 2000. v. 8, n. 2000, p. 93–103. Citado na página 31.

DU, Yuhui; SUI, Jing; YU, Qingbao; HE, Hao; CALHOUN, Vince D. Semi-supervised learning of brain functional networks. In: IEEE. *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*. [S.l.], 2014. p. 1–4. Citado na página 16.

DY, Jennifer G. Unsupervised feature selection. *Computational methods of feature selection*, CRC Press Boca Raton, FL, USA, p. 19–39, 2008. Citado na página 23.

EREN, Kemal; DEVECI, Mehmet; KÜÇÜKTUNÇ, Onur; ÇATALYÜREK, Ümit V. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics*, Oxford University Press, v. 14, n. 3, p. 279–292, 2012. Citado na página 31.

ESSEN, David C Van; UGURBIL, Kamil. The future of the human connectome. *Neuroimage*, Elsevier, v. 62, n. 2, p. 1299–1310, 2012. Citado 2 vezes nas páginas 17 e 19.

ESSEN, David C Van; UGURBIL, Kamil; AUERBACH, E; BARCH, D; BEHRENS, TEJ; BUCHOLZ, R; CHANG, Acer; CHEN, Liyong; CORBETTA, Maurizio; CURTISS, Sandra W et al. The human connectome project: a data acquisition perspective. *Neuroimage*, Elsevier, v. 62, n. 4, p. 2222–2231, 2012. Citado na página 18.

FATEHI, Kavan; BOZORGI, ARASTOO; ZAHEDI, MOHAMMAD SADEGH; ASGARIAN, EHSAN. Improving semi-supervised constrained k-means clustering method using user feedback. JOURNAL OF COMPUTING AND SECURITY, 2014. Citado 2 vezes nas páginas 29 e 30.

FILIPOVYCH, Roman; DAVATZIKOS, Christos; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (mci). *NeuroImage*, Elsevier, v. 55, n. 3, p. 1109–1119, 2011. Citado 2 vezes nas páginas 16 e 25.

FREITAS, Adelaide; AYADI, Wassim; ELLOUMI, Mourad; OLIVEIRA, Joséluis; HAO, Jin-Kao. Survey on biclustering of gene expression data. *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, John Wiley & Sons, Inc., p. 591–608, 2013. Citado na página 32.

GARCIA, Rodolfo; NIEVOLA, Julio Cesar; PARAISO, Emerson Cabrera. Estudo comparativo de métodos de seleção de atributos na predição de matrizes de conectividades em tdah obtidas pela técnica de resting-state fmri. In: *XIII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.: s.n.], 2016. p. 637–647. Citado 2 vezes nas páginas 44 e 47.

GARCIA, Rodolfo; PARAISO, Emerson Cabrera; NIEVOLA, Julio Cesar. Comparative study of dimensionality reduction methods using reliable features for multiple datasets obtained by rs-fmri in adhd prediction. In: SPRINGER. *Canadian Conference on Artificial Intelligence*. [S.l.], 2017. p. 97–102. Citado 4 vezes nas páginas 15, 44, 47 e 74.

HARTIGAN, John A. Direct clustering of a data matrix. *Journal of the american statistical association*, Taylor & Francis Group, v. 67, n. 337, p. 123–129, 1972. Citado na página 31.

HEUVEL, Martijn P Van Den; POL, Hilleke E Hulshoff. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, Elsevier, v. 20, n. 8, p. 519–534, 2010. Citado 2 vezes nas páginas 18 e 20.

HORTA, Danilo; CAMPELLO, Ricardo JGB. Similarity measures for comparing biclusterings. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 11, n. 5, p. 942–954, 2014. Citado na página 47.

JOACHIMS, Thorsten. Transductive inference for text classification using support vector machines. In: *ICML*. [S.l.: s.n.], 1999. v. 99, p. 200–209. Citado na página 45.

JOHNSTON, Blair A; MWANGI, Benson; MATTHEWS, Keith; COGHILL, David; KONRAD, Kerstin; STEELE, J Douglas. Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. *Human brain mapping*, Wiley Online Library, v. 35, n. 10, p. 5179–5189, 2014. Citado na página 31.

LAZAR, Cosmin; TAMINAU, Jonatan; MEGANCK, Stijn; STEENHOFF, David; COLETTA, Alain; MOLTER, Colin; SCHAETZEN, Virginie de; DUQUE, Robin; BERSINI, Hugues; NOWE, Ann. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 9, n. 4, p. 1106–1119, 2012. Citado na página 45.

LIANG, Sheng-Fu; HSIEH, Tsung-Hao; CHEN, Pin-Tzu; WU, Ming-Long; KUNG, Chun-Chia; LIN, Chun-Yu; SHAW, Fu-Zen. Differentiation between resting-state fmri data from adhd and normal subjects: based on functional connectivity and machine learning. In: IEEE. *Fuzzy Theory and it's Applications (iFUZZY), 2012 International Conference on*. [S.l.], 2012. p. 294–298. Citado na página 15.

LIM, Lena; MARQUAND, Andre; CUBILLO, Ana A; SMITH, Anna B; CHANTILUKE, Kaylita; SIMMONS, Andrew; MEHTA, Mitul; RUBIA, Katya. Disorder-specific predictive classification of adolescents with attention deficit hyperactivity disorder (adhd) relative to autism using structural magnetic resonance imaging. *PLoS One*, Public Library of Science, v. 8, n. 5, p. e63660, 2013. Citado na página 15.

LIU, Xiaowen; WANG, Lusheng. Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics*, Oxford University Press, v. 23, n. 1, p. 50–56, 2006. Citado na página 47.

MADEIRA, Sara C; OLIVEIRA, Arlindo L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 1, n. 1, p. 24–45, 2004. Citado 5 vezes nas páginas 14, 31, 32, 33 e 37.

MILHAM, Michael P; FAIR, Damien; MENNES, Maarten; MOSTOFSKY, Stewart HMD et al. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, Frontiers, v. 6, p. 62, 2012. Citado 2 vezes nas páginas 15 e 41.

MORADI, Elaheh; PEPE, Antonietta; GASER, Christian; HUTTUNEN, Heikki; TOHKA, Jussi; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Machine learning framework

for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, Elsevier, v. 104, p. 398–412, 2015. Citado 2 vezes nas páginas 16 e 25.

MWANGI, Benson; TIAN, Tian Siva; SOARES, Jair C. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, Springer, v. 12, n. 2, p. 229–244, 2014. Citado 2 vezes nas páginas 15 e 31.

OGHABIAN, Ali; KILPINEN, Sami; HAUTANIEMI, Sampsa; CZEIZLER, Elena. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one*, Public Library of Science, v. 9, n. 3, p. e90801, 2014. Citado na página 31.

PADILHA, Victor A; CAMPELLO, Ricardo JGB. A systematic comparative evaluation of biclustering techniques. *BMC bioinformatics*, BioMed Central, v. 18, n. 1, p. 55, 2017. Citado 3 vezes nas páginas 31, 33 e 36.

PEETERS, René. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, Elsevier, v. 131, n. 3, p. 651–654, 2003. Citado na página 33.

PRELIĆ, Amela; BLEULER, Stefan; ZIMMERMANN, Philip; WILLE, Anja; BÜHLMANN, Peter; GRUISSEM, Wilhelm; HENNIG, Lars; THIELE, Lothar; ZITZLER, Eckart. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, Oxford University Press, v. 22, n. 9, p. 1122–1129, 2006. Citado 3 vezes nas páginas 36, 38 e 46.

SATO, João Ricardo; HOEXTER, Marcelo Queiroz; FUJITA, André; ROHDE, Luis Augusto. Evaluation of pattern recognition and feature extraction methods in adhd prediction. *Frontiers in systems neuroscience*, Frontiers, v. 6, p. 68, 2012. Citado na página 15.

SEEGER, Matthias. *A taxonomy for semi-supervised learning methods*. [S.l.], 2006. Citado na página 24.

SPORNS, Olaf. The human connectome: a complex network. *Annals of the New York Academy of Sciences*, Wiley Online Library, v. 1224, n. 1, p. 109–125, 2011. Citado 2 vezes nas páginas 18 e 19.

SPORNS, Olaf; TONONI, Giulio; KÖTTER, Rolf. The human connectome: a structural description of the human brain. *PLoS computational biology*, Public Library of Science, v. 1, n. 4, p. e42, 2005. Citado na página 17.

VAPNIK, Vladimir. *Statistical learning theory. 1998*. [S.l.]: Wiley, New York, 1998. Citado na página 25.

WAGSTAFF, Kiri; CARDIE, Claire; ROGERS, Seth; SCHRÖDL, Stefan et al. Constrained k-means clustering with background knowledge. In: *ICML*. [S.l.: s.n.], 2001. v. 1, p. 577–584. Citado 2 vezes nas páginas 25 e 29.

WESTBROOK, Catherine; ROTH, Carolyn Kaut. *MRI in Practice*. [S.l.]: John Wiley & Sons, 2011. Citado na página 18.

WOLFERS, Thomas; BUITELAAR, Jan K; BECKMANN, Christian F; FRANKE, Barbara; MARQUAND, Andre F. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics.

Neuroscience & Biobehavioral Reviews, Elsevier, v. 57, p. 328–349, 2015. Citado 3 vezes nas páginas 13, 15 e 16.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005. Citado na página 23.

ZENG, Ling-Li; SHEN, Hui; LIU, Li; WANG, Lubin; LI, Baojuan; FANG, Peng; ZHOU, Zongtan; LI, Yaming; HU, Dewen. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, Oxford University Press, v. 135, n. 5, p. 1498–1507, 2012. Citado na página 15.

ZHU, Chao-Zhe; ZANG, Yu-Feng; CAO, Qing-Jiu; YAN, Chao-Gan; HE, Yong; JIANG, Tian-Zi; SUI, Man-Qiu; WANG, Yu-Feng. Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *Neuroimage*, Elsevier, v. 40, n. 1, p. 110–120, 2008. Citado na página 15.

ZHU, Xiaojin; GHAHRAMANI, Zoubin; LAFFERTY, John D. Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*. [S.l.: s.n.], 2003. p. 912–919. Citado na página 27.

ZHU, Xiaojin; GOLDBERG, Andrew B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado 3 vezes nas páginas 25, 26 e 27.

ZHU, Xiaofeng; SUK, Heung-Il; WANG, Li; LEE, Seong-Whan; SHEN, Dinggang. A novel relational regularization feature selection method for joint regression and classification in ad diagnosis. *Medical image analysis*, Elsevier, v. 38, p. 205–214, 2017. Citado na página 15.