Feature Selection through Biclustering to Identify Specific Language Impairment

Marta D. M. Noronha[1], Luis E. Zárate[1]

Pontifícia Universidade Católica de Minas Gerais, Brasil
martadmnoronha@gmail.com, zarate@pucminas.br

Abstract. Failure to express yourself verbally is a condition that affects nearly 7% of children worldwide, known as specific language impairment. The diagnosis is complex, involving specialists such as speech therapists and pediatricians. The dataset used in this work has many attributes and imbalanced data, which can harm knowledge discovery. We used biclustering to identify clusters that characterize children with speech problems and those with typical development. We propose selecting attributes through the significance analysis of biclusters, which enhanced the F-score and accuracy in models generated by using 90% of instances from the training dataset, compared to results from the original data.

CCS Concepts: • **Applied computing** → **Health informatics**; • **Computing methodologies** → **Classification and regression trees**; **Feature selection**.

Keywords: Biclustering, Classification, Data mining, Speech signal processing, Specific language impairment

1. INTRODUÇÃO

A deficiência caracterizada pela incapacidade de se expressar por meio de linguagem natural em crianças sem problemas auditivos, comparado à crianças da mesma faixa etária que não possuem essa deficiência, é uma condição que afeta cerca de 7% das crianças no mundo com idade próxima aos 5 anos. O diagnóstico da deficiência, considerada como um distúrbio do neurodesenvolvimento, é complexo para especialistas como pediatras e fonoaudiólogos, e geralmente envolve o uso de diferentes técnicas como avaliação de morfologia e sintaxe em erros contabilizados quando a criança reproduz uma história, para diagnosticar se essa possui um distúrbio específico da linguagem (*"Specific Language Impairment"* - SLI) ou um desenvolvimento típico (*"Typical Development"* - TD) para sua faixa etária [Gabani et al. 2011; Sharma and Singh 2022; Huang et al. 2022]. Roger Brown propôs segmentar as idades para definir as habilidades de fala que devem estar presentes em cada faixa para diagnosticar o tipo de desenvolvimento de uma criança. Por exemplo, no estágio I (entre 12 e 26 meses) as crianças possuem a habilidade linguística para se expressar com um vocabulário de 50 a 60 palavras, sendo capazes de produzir frases como "Na sala"quando a sua intenção comunicativa seria dizer "Estou na sala", a qual teria sido dito por uma criança com maior maturidade para formar frases completas. A avaliação proposta por Brown mede também o comprimento da sentença falada em uma unidade de medida chamada MLUm's (*mean length of utterance measured in morphemes*), a qual define valores por meio da análise sintática sobre a presença de palavras, sílabas e fonemas, avaliando se a evolução e maturidade da linguagem da criança segue o padrão da normalidade, indicando que a mesma possui

Este trabalho foi realizado por meio de bolsa financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pelo Programa de Apoio à Pós-Graduação em Instituições Comunitárias de Ensino Superior (PROSUC - Programa de Suporte à Pós- Graduação de Instituições Comunitárias de Ensino Superior) / Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil), Código Financeiro 001, Agência Federal de Apoio e Avaliação da Pós-Graduação no âmbito do Ministério da Educação do Brasil.

um TD. Caso contrário, a mesma será diagnosticada como portadora de SLI [Bowen 1998].

Em Hayward et al. [2009], o desenvolvimento foi analisado por meio da avaliação da capacidade de realizar tarefas, como narrar uma história e responder questões relacionadas à mesma, para compreender os efeitos da idade e da complexidade das tarefas em crianças de 4 aos 6 anos que possuem um TD. Em Gillam et al. [1998], a análise por métodos estatísticos compara crianças com SLI e um grupo de controle para validar hipóteses construídas a respeito de características do SLI.

Poucos estudos foram encontrados que usam técnicas de aprendizado de máquina para diagnóstico de SLI em crianças, onde o conjunto de dados geralmente é desbalanceado por haver menor quantidade de instâncias rotuladas como SLI. Em Sharma and Singh [2022] é apresentada uma abordagem que utiliza um modelo de rede neural convolucional de 1 camada (*1D convolutional neural network* - CNN), e outra com um modelo híbrido da CNN com modelo de rede neural recorrente de memória de curto e longo prazo (*Long Short-Term Memory* - LSTM), usados para classificar dados em aúdios provenientes de crianças da República Tcheca. Nestes, foram avaliadas as pronúncias das vogais "a" e "o" em que os modelos alcançaram valores de acurácia e F-Score acima de 90%. Em Huang et al. [2022] foram analisados os efeitos do desbalanceamento dos dados em que o F-Score dos melhores modelos, *Random Forest* com *oversampling* e *back-propagation neural network* (BPNN), ficaram acima de 90%.

A proposta deste trabalho é a utilização de biclusterização para avaliar se as características contidas nos biclusters mais significantes de cada classe, SLI e TD, podem ser selecionadas para construção de modelos de classificação que produzam melhores resultados classificatórios do que um modelo construído usando o conjunto de dados original. O uso da biclusterização para redução do espaço de características diferencia este trabalho daqueles apresentados em Sharma and Singh [2022] e Huang et al. [2022]. A biclusterização descobre subgrupos de instâncias (atributos) que são coerentes, de acordo com algum determinado critério ou medida, com subgrupos de atributos (instâncias). Os dados mais usados em biclusterização são provenientes de dados genéticos, contudo alguns são originados de mineração de texto e outras áreas [Noronha et al. 2022]. O uso de biclusterização para seleção de características é apresentado em Huang et al. [2009] e Busygin et al. [2005], sendo que o primeiro usa uma heurística baseada em clusterização hierárquica e, o segundo, usa biclusterização supervisionada baseada em problemas de otimização inteira não linear para seleção de características, porém o foco dos dois trabalhos não é aplicação em dados de crianças com SLI e TD.

Neste trabalho é usado o conjunto de dados do projeto CHILDES [1], o qual possui muitos atributos e desbalanceamento de classes, sendo de difícil separação de classes por medidas que utilizam todos os atributos para calcular a (dis)similaridade entre as instâncias, como a distância euclidiana. Isto ocorre porque o aumento da quantidade de atributos faz com que a distância entre as instâncias seja o mais próximo da média dos dados, impedindo a separação entre os distintos grupos do conjunto de dados [Hinneburg et al. 2000]. Por este motivo que este trabalho sugere a biclusterização como uma ferramenta para selecionar atributos para reduzir a dimensionalidade dos dados que são utilizados na construção dos modelos de classificação.

Este trabalho se divide em: Seção 2 apresenta os materiais e métodos propostos para a execução do trabalho. A Seção 3 mostra os resultados e apresenta uma breve discussão sobre os mesmos. Por fim, a Seção 4 apresenta as conclusões e propostas para trabalhos futuros.

2. MATERIAIS E METÓDOS

2.1 Descrição do conjunto de dados

O conjunto de dados para diagnóstico de deficiência de linguagens é proveniente do agrupamento de três conjuntos[1]: ENNI que foi coletado de 377 crianças canadenses com idades de 4 a 9 anos;

[1]Link para acesso aos dados CHILDES: https://www.kaggle.com/dgokeeffe/specific-language-impairment

Média do bicluster	Média das instâncias	Média dos atributos
$\mu_{(B_{IJ})} = \frac{1}{I*J}\sum_{i=1}^{I}\sum_{j=1}^{J} b_{ij}$ (1)	$\mu_{(B_{iJ})} = \frac{1}{J}\sum_{j=1}^{J} b_{ij}$ (2)	$\mu_{(B_{Ij})} = \frac{1}{I}\sum_{i=1}^{I} b_{ij}$ (3)

Table I. Médias usadas para cálculo do MSR de um bicluster

Gillam de 668 crianças dos Estados Unidos com idades de 4 a 12 anos; e Conti4 de 118 adolescentes britânicos com idades de 12 a 16 anos. Este conjunto foi disponibilizado pré-processado com remoção de instâncias que continham diversos dados ausentes, junção de atributos semelhantes e inclusão de atributos obtidos pela aplicação de medidas estatísticas sobre atributos desse conjunto de dados. Esse conjunto possui 60 atributos relacionados à idade, aspectos avaliados em narrações de crianças, conjunto de dados de origem, e um atributo de classe (SLI ou TD).

Por meio de análises feitas neste trabalho, foi verificado que os valores de mínimo, máximo, média e desvio padrão de cada atributo numérico são divergentes por origem dos dados, prejudicando a análise sem o auxílio de um especialista. Portanto, na falta do especialista, somente o conjunto Gillam foi selecionado por este possuir a maior quantidade de instâncias, sendo removido o atributo relacionado à origem dos dados. Foram removidos também o atributo "*age_year*" por existir um equivalente expresso em meses e uma instância com sexo biológico não descrito, resultando em 667 instâncias no conjunto de dados. Destas, 70 do sexo feminino e 100 do sexo masculino possuem SLI (25,5% das instâncias), enquanto 254 do sexo feminino e 243 do sexo masculino possuem TD. O atributo referente ao sexo biológico deu origem a dois novos atributos: "*sex_male*" e "*sex_female*". Portanto o conjunto de dados resultante, o qual chamaremos de agora em diante de original, possui 61 atributos (incluso a classe) e 667 instâncias representando crianças com idades entre 60 a 143 meses.

2.2 Biclusterização

Algoritmos de biclusterização são desenvolvidos para encontrar subconjuntos de instâncias que são coerentes em relação a subconjuntos de atributos, e/ou vice-versa, sendo cada subconjunto chamado de bicluster. Em um conjunto de dados A_{MN}, M e N são respectivamente o número de instâncias e de atributos, temos que um bicluster $B_{IJ} \subseteq A_{MN}$, onde $i = 1, 2, .., I$ com $I \subseteq M$ e $j = 1, 2, .., J$ com $J \subseteq N$. Cada valor do par ordenado (instância i x atributo j) de um bicluster é dado por b_{ij} [Madeira and Oliveira 2004].

Diferentes medidas podem ser implementadas nos algoritmos para descobrir biclusters para atender a um ou mais objetivos. Dentre as medidas, uma das mais usadas em biclusterização é a média do quadrado do resíduo (MSR - *Mean Squared Residue*) em que a variância do bicluster é calculada com base em médias avaliadas no bicluster, apresentadas na Tabela I, e o cálculo do MSR dado pela Equação 4 [Noronha et al. 2022; Cheng and Church 2000; Madeira and Oliveira 2004]. O algoritmo δ-bicluster foi escolhido para execução deste trabalho porque os atributos do conjunto de dados são numéricos, favorecendo o uso do MSR para avaliar a coerência interna dos biclusters.

$$MSR(B_{IJ}) = \frac{1}{|I| * |J|)} \sum_{i \in I} \sum_{j \in J} (b_{ij} - \mu_{(B_{Ij})} - \mu_{(B_{iJ})} + \mu_{(B_{IJ})})^2 \quad (4)$$

O teste hipergeométrico p-valor (Equação 5) foi usado para avaliar a significância dos biclusters, medindo a probabilidade de observar $|I_k|$ instâncias da classe C_k no *bicluster* B_{IJ}, com $k = 1, 2, .., K$, sendo que o conjunto de dados possui $|M|$ instâncias das quais $|I|$ pertencem ao bicluster B_{IJ}. O *bicluster* é estatisticamente significante se p-valor $< 5\%$, indicando que este é super-representado e

Parâmetros	Biclusters descobertos	Biclusters rotulados e significantes
$\alpha = 3, \delta = 0,3$	8	6
$\alpha = 4, \delta = 0,4$	7	4
$\alpha = 5, \delta = 0,5$	5	4

Table II. Quantidade de biclusters descobertos e significantes para os parâmetros do algoritmo δ-bicluster

não ocorre por acaso no conjunto de dados [Zhao et al. 2012; Noronha et al. 2022; Eren et al. 2012].

$$p - valor(B_{IJ}) = \frac{\binom{|C_k|}{|I_k|}\binom{|M|-|C_k|}{|I|-|I_k|}}{\binom{|M|}{|I|}} \tag{5}$$

Os resultados são obtidos por 3 variações nos parâmetros do algoritmo δ-bicluster [2] devido este algoritmo produzir biclusters diferentes pela variação de seus parâmetros. Estes parâmetros são α, usado para exclusão de nós (linhas e colunas), e δ, relacionado ao valor limiar (chamado de *threshold*) que não deve ser ultrapassado na descoberta dos biclusters. Ou seja, $MSR(B_{IJ}) \leq \delta$. Os parâmetros foram selecionados por meio de experimentos empíricos, sendo eles: $\alpha = 3, \delta = 0,3$; $\alpha = 4, \delta = 0,4$; e, $\alpha = 5, \delta = 0,5$. Nos experimentos quanto maior o valor de α e/ou δ, uma quantidade menor de biclusters foram descobertos ou os biclusters continham uma quantidade maior de instâncias.

Cada bicluster recebeu o rótulo, SLI ou TD, de acordo com a classe majoritária que ele contém, sendo considerada acima de 50% de instâncias de uma determinada classe.

2.3 Árvore de decisão

A árvore de decisão usada neste trabalho é disponibilizada no pacote *rpart*[3] da linguagem R. O pacote oferece implementação de particionamento recursivo para classificação e árvores de regressão. Como as classes são categóricas, o valor do método para classificação foi selecionado como "*class*" e a divisão da árvore foi feito utilizando o índice "*gini*". Os demais parâmetros foram mantidos como "*default*", sendo que neste caso o vetor de probabilidades à priori das classes são proporcionais à contagem das mesmas e as perdas da matriz são padronizadas para 1.

Para comparação de resultados de classificação com o conjunto de dados original, foram gerados diferentes conjuntos de dados a partir dos atributos contidos em cada bicluster estatisticamente significante. Cada um dos conjuntos de treino possuem proporções entre 60% a 90% de instâncias do conjunto original. Logo os conjuntos de teste, com instâncias não vistas pelo modelo de classificação, possuem respectivamente de 40% a 10% das instâncias restantes.

3. RESULTADOS E DISCUSSÃO

A Tabela II apresenta a quantidade de biclusters descobertos e, dentre estes, quantos são significantes (p-valor $< 5\%$) em cada execução do δ-bicluster. Para a execução com parâmetros $\alpha = 4$ e $\delta = 0,4$, um bicluster não rotulado foi removido por possuir exatamente 50% de instâncias de cada classe.

As Tabelas III, V e VII apresentam, em suas respectivas seções, a classe atribuída ao bicluster significante, a quantidade de instâncias e atributos, o percentual de instâncias da classe majoritária e a significância do bicluster. Também são mostrados os valores do F-Score para cada classe, SLI e TD, além da acurácia do modelo criado sobre o conjunto de dados que possui somente os atributos do bicluster. As medidas de F-Score e acurácia do conjunto de dados original também são apresentadas para fins comparativos. Nestas Tabelas, os valores de acurácia e F-Score correspondem aos resultados da classificação onde o treino do modelo utilizou 90% do conjunto de dados e o teste utilizou os 10%

[2]Disponível no pacote *biclust* em: https://cran.r-project.org/web/packages/biclust/biclust.pdf
[3]Disponível em: https://cran.r-project.org/web/packages/rpart/rpart.pdf

bicluster	classe	instâncias	atributos	majoritária	significância	F-Sc.SLI	F-Sc.TD	Acurácia
cc-3	TD	208	23	84,13%	**3,45E-05**	73,33%	92,31%	88,06%
cc-1		131	26	87,02%	6,11E-05	**75,86**%	**93,33**%	**89,55**%
cc-6		39	14	56,41%	5,40E-03	27,27%	85,71%	76,12%
cc-5		50	15	60,00%	7,80E-03	59,26%	89,72%	83,58%
cc-4	SLI	51	15	54,90%	**2,17E-06**	64,29%	90,57%	85,07%
cc-8		10	9	70,00%	3,26E-03	**70,97**%	**91,26**%	**86,57**%
Original	-	-	-	-	-	**62,5**%	**88,24**%	**82,09**%

Table III. δ-biclusters com parâmetros δ= 0.3 e α = 3, e F-Score (F-Sc.) e acurácia do conjunto de teste no modelo gerado sobre 90% das instâncias de treino

restantes. As Tabelas IV, VI e VIII mostram os valores da acurácia obtidas no conjunto de teste, onde os modelos foram gerados em conjuntos de dados contendo diferentes proporções de instância no conjunto de treino (60% a 90%).

3.1 Classificação usando atributos de biclusters gerados com parâmetros $\alpha = 3$ e $\delta = 0,3$

Os resultados mostrados na Tabela III apontam que, considerando a classe TD, o bicluster "cc-3" possui a maior significância e valores de F-Score e acurácia pouco menores do que "cc-1", o qual possui o segundo melhor valor de significância, a melhor acurácia e F-Score. Situação semelhante ocorre ao se observar a significância, F-Score de cada classe e acurácia do modelo nos biclusters "cc-4" e "cc-8", ambos SLI, sendo o primeiro mais significante enquanto o segundo possui melhor resultado classificatório sobre o conjunto de testes no modelo gerado com 90% de instâncias no treinamento.

Na Tabela IV, pode-se observar que o modelo "cc-3", classe TD, possui acurácia maior ou igual do que "cc-1" em partições de 60% a 75% do conjunto de dados em treino (4 de 7 partições), permanecendo com valor de no máximo 1,51% abaixo de "cc-1" nas demais partições. Na classe SLI, o modelo "cc-4" possui os melhores valores de acurácia com exceção do modelo criado com a partição de 90% dos dados para treino, sendo este 1,50% menor do que "cc-8". Os melhores valores de acurácia obtido por biclusters de cada classe são destacados em negrito na Tabela IV. Os modelos de classificação usando os atributos dos biclusters mais significantes de cada classe obtiveram resultados (medidos no conjunto de teste) superiores ao conjunto original em no mínimo 6% com 85% de treino e 2,98% com 90% de treino.

3.2 Classificação usando atributos de biclusters gerados com parâmetros $\alpha = 4$ e $\delta = 0,4$

Os biclusters mais significantes de cada classe, apresentados na Tabela V, também possuem melhores resultados de classificação que o conjunto original. Porém devido ao empate no F-Score e acurácia dos biclusters da classe TD, "cc-1" e "cc-2", a acurácia foi medida sobre os respectivos conjuntos de teste de cada modelo, onde os modelos foram gerados com diferentes proporções de instâncias no conjunto de treino (Tabela VI).

Pode-se observar que o modelo "cc-1" obteve resultados inferiores ao "cc-2" em 2 partições (65% e 70%), onde a maior diferença é de apenas 0,5%. O modelo gerado sobre "cc-2" teve pior desempenho

		Proporção de instâncias usadas no conjunto de treino						
Bicluster	Classe	60%	65%	70%	75%	80%	85%	90%
cc-3	TD	81,65%	81,62%	82,00%	79,52%	81,20%	85,00%	88,06%
cc-1		81,65%	79,91%	80,50%	78,31%	82,71%	86,00%	**89,55**%
cc-4	SLI	79,78%	79,49%	80,00%	79,52%	82,71%	86,00%	85,07%
cc-8		78,28%	77,35%	74,50%	74,70%	78,20%	84,00%	**86,57**%
Original		80,90%	80,77%	**82,50**%	81,33%	81,20%	78,00%	82,09%

Table IV. Acurácia obtida nos conjuntos de testes com modelos usando *delta*-bicluster com δ = 0,3 e *alpha* = 3

bicluster	classe	instâncias	#atributos	majoritária	significância	F-Sc. SLI	F-Sc.TD	Acurácia
cc-2	TD	232	24	81,47%	**7,32E-04**	**75,86%**	**93,33%**	**89,55%**
cc-1		283	26	79,86%	1,73E-03	**75,86%**	**93,33%**	**89,55%**
cc-5	SLI	24	17	91,67%	**4,88E-12**	**70,97%**	**91,26%**	**86,57%**
cc-3		57	17	54,39%	7,51E-07	64,29%	90,57%	85,07%
Original	-	-	-	-	-	**62,5%**	**88,24%**	**82,09%**

Table V. Informações δ-biclusters sob parâmetro δ= 0.4 e $\alpha = 4$, e F-Score (F-Sc.) e acurácia do conjunto de teste no modelo gerado sobre 90% das instâncias de treino

		Proporção de instâncias usadas no conjunto de treino						
Bicluster	Classe	60%	65%	70%	75%	80%	85%	90%
cc-2	TD	81,65%	80,34%	81,00%	78,31%	82,71%	86,00%	**89,55%**
cc-1		81,65%	79,91%	80,50%	78,31%	82,71%	86,00%	**89,55%**
cc-5	SLI	79,78%	80,77%	82,50%	83,73%	81,20%	86,00%	**86,57%**
Original	—	80,90%	80,77%	**82,50%**	81,33%	81,20%	78,00%	82,09%

Table VI. Acurácia obtida nos conjuntos de testes com modelos usando *delta*-bicluster com δ = 0,4 e *alpha* = 4

nas partições de 65% a 75% (3 de 4 partições) comparado com os modelos "cc-5" e original, sendo a maior diferença observada na partição de 75% (5,42% entre "cc-2" e "cc-5"). Porém nos modelos criados a partir de 80% de instâncias, os atributos dos biclusters mais significantes de cada classe possuem melhores resultados classificatórios do que o conjunto original. Além disso, nos modelos com 90% de instâncias de treino, o modelo "cc-2" (TD) supera o modelo "cc-5" (SLI).

3.3 Classificação usando atributos de biclusters gerados com parâmetros $\alpha = 5$ e $\delta = 0,5$

A Tabela VII mostra que os biclusters mais significantes de cada classe obtiveram os melhores F-Score e acurácia do que os dados originais, com exceção do F-Score do bicluster "cc-3", SLI, o qual ficou 0,43% menor do que o obtido no conjunto original. Na Tabela VIII, foi observado que dentre os 7 modelos de "cc-1" e "cc-3", respectivamente, 2 e 3 modelos possuem menor acurácia do que o original sendo a maior diferença, de 3,02%, registrada na partição de 75% para treino com "cc-1". Porém, os atributos dos biclusters mais significantes possuem resultados classificatórios superiores ao conjunto original em partições a partir de 85%.

3.4 Discussão

A seleção de atributos por meio da escolha de biclusters mais significantes mostrou ser uma boa ferramenta para melhorar os resultados de classificação neste conjunto de dados, sobretudo quando o modelo é gerado de um conjunto de treino contendo a partir de 85% de instâncias para treino. Porém embora a significância dos biclusters cujos atributos geraram os melhores modelos seja alta, nota-se que o modelo "cc-5" (SLI) é mais significante e possui menor acurácia do que "cc-2" (TD) (ver Tabela V). Embora a significância tenha melhorado os resultados da classificação, é necessário avaliar outras características do bicluster que podem influenciar nos resultados classificatórios, como a relação entre o tamanho do bicluster e a proporção da classe que este representa. Busca-se desta forma verificar se

bicluster	classe	instâncias	#atributos	majoritária	significância	F-Sc. SLI	F-Sc.TD	Acurácia
cc-1	TD	505	27	78,42%	**2,91E-05**	**75,86%**	**93,33%**	**89,55%**
cc-4		25	14	60,00%	4,46E-02	50,00%	86,79%	79,10%
cc-3	SLI	49	20	51,02%	**4,23E-05**	**62,07%**	**89,52%**	**83,58%**
cc-5		9	11	66,67%	9,14E-03	37,04%	84,11%	74,63%
Original	-	-	-	-	-	**62,5%**	**88,24%**	**82,09%**

Table VII. Informações δ-biclusters sob parâmetro δ= 0.5 e $\alpha = 5$, e F-Score (F-Sc.) e acurácia do conjunto de teste no modelo gerado sobre 90% das instâncias de treino

		Proporção de instâncias usadas no conjunto de treino						
Bicluster	Classe	60%	65%	70%	75%	80%	85%	90%
cc-1	TD	81,27%	81,20%	81,50%	78,31%	82,71%	86,00%	**89,55**%
cc-3	SLI	81,65%	79,91%	80,50%	79,52%	81,20%	**85,00**%	83,58%
Original	—	80,90%	80,77%	**82,50**%	81,33%	81,20%	78,00%	82,09%

Table VIII. Acurácia obtida nos conjuntos de testes com modelos usando *delta*-bicluster com $\delta = 0{,}5$ e $alpha = 5$

é possível selecionar somente um único bicluster, seja o mais significativo da classe minoritária ou da majoritária, para construir um modelo de classificação em que a performance seja melhor do que a obtida pelo modelo original.

Dentre os resultados apresentados, os biclusters da classe TD, "cc-1" na Seção 3.1, "cc-2" na Seção 3.2 e "cc-1" na Seção 3.3, obtiveram simultaneamente o melhor valor de acurácia e F-Score tanto nos modelos criados com 90% do conjunto de dados para treino em relação ao conjunto original quanto naqueles da classe SLI. Os 3 modelos representam a mesma árvore de classificação, a qual contém 5 atributos. Destes, 3 foram gerados por teste estatístico z-Score, o qual mede o quanto um valor difere da média em termos de desvio padrão, sendo calculado sobre pontuações obtidas nos testes pelos participantes. As regras extraídas da árvore são apresentadas na Tabela IX.

Logo, por este modelo é possível classificar as instâncias com 89,55% de acurácia utilizando as regras apresentadas na Tabela IX. Estas regras são formadas por atributos calculados sobre o z-Score medido sobre a quantidade de palavras erradas, o comprimento médio da fala, e o número de expressões verbais que uma criança portadora de SLI pode pronunciar. Além disso, a média de sílabas e a pontuação no teste Flesch-Kincaid podem auxiliar no diagnóstico, conforme apontado pelas regras 4 a 7.

Seguindo a metodologia proposta neste trabalho foi possível obter 93,33% de F-Score com o melhor modelo, sendo semelhante àqueles obtidos nos trabalhos relacionados que investigam o SLI por meio de aprendizado de máquina porém estes não abordam a redução de dimensionalidade.

4. CONCLUSÃO

Neste trabalho foi mostrado que os atributos de biclusters mais significantes podem ser usados para melhorar a classificação em um conjunto de dados desbalanceado e com muitos atributos. Para isto, foi usado o algoritmo δ-bicluster para descobrir diferentes biclusters por meio da variação de seus parâmetros. Foi observado que o modelo gerado usando os atributos selecionados do bicluster mais significante, representante da classe majoritária, em cada uma destas variações possuía os mesmos valores de acurácia e F-Score para as classes SLI e TD por representarem a mesma árvore (respecti-

	Regra	**Classe**	**Perc_Acerto**	**iFolha**
1	zwes≥-0.37 E zms< 0.063	SLI	88,52%	61
2	zwes≥-0.37 E zms≥ 0.063 E zwes≥0.041	SLI	76,19%	21
3	zwes≥-0.37 E zms≥ 0.063 E zwes<0.041	TD	73,8%	42
4	zwes<-0.37 E zms< 0.12 E avs< 1.1	SLI	61,11%	18
5	zwes<-0.37 E zms< 0.12 E avs≥1.1 E zus≥-0.51 E f_k≥1.54	SLI	87,50%	8
6	zwes<-0.37 E zms< 0.12 E avs≥1.1 E zus≥-0.51 E f_k<1.54	TD	58,82%	17
7	zwes<-0.37 E zms< 0.12 E avs≥1.1 E zus<-0.51	TD	80%	50
8	zwes<-0.37 E zms≥ 0.12	TD	90,34%	383

z_word_errors_sli (zwes): z-Score baseado no número de erros de palavras; z_mlu_sli (zms): z-Score com base no comprimento médio da fala do grupo SLI; average_syl (avs): número médio de sílabas por palavra; z_utts_sli (zus): z-Score baseado no número de expressões verbais do grupo SLI; Flesch-Kincaid *Score* (f_k): pontuação no teste que avalia o quão difícil é a compreensão da leitura de uma passagem de texto em inglês; iFolha: quantidade de instâncias de treino por folha; Perc_Acerto: o percentual de acerto da classificação no conjunto de treino.

Table IX. Regras da árvore obtida simultaneamente pelos melhores modelos, com iFolha representando o número de instâncias na folha

vamente 89,55%, 75,86% e 93,33%). Logo estes produziram as mesmas regras.

Como vantagem da metodologia usada nesse trabalho, têm-se a identificação das características por meio da análise dos biclusters, onde a classificação é melhorada por já ter sido assegurada a coerência (dada pelo MSR) e a significância estatística do bicluster, permitindo assim selecionar atributos representativos que permitem gerar um bom modelo de classificação.

Como limitações aponta-se a) a ausência de comparação com outros algoritmos de biclusterização ou algoritmos específicos para seleção de características baseados em filtros ou wrappers; b) a ausência do especialista para auxiliar no tratamento e análise dos dados provenientes das demais fontes (ENNI e Conti4) dificulta a análise da generalização do modelo gerado; e c) na biclusterização, devido à existência de poucas instâncias no conjunto Gillam, todos os registros do conjunto de dados foram usados com o objetivo de maximizar a diversidade de informação para extrair as características que podem ser úteis para classificar os perfis das crianças que possuem SLI ou TD. Mas as instâncias dos conjuntos de teste não foram vistos na geração de modelos.

Em trabalhos futuros a) pretende-se investigar como a significância de um bicluster é afetada pelo tamanho e pela classe que o mesmo representa (majoritária e minoritária); b) avaliar a significância estatística da performance de modelos de classificação, comparando modelos criados com dados originais com aqueles obtidos pela redução por biclusterização, e; c) mitigar as limitações deste trabalho para assegurar adequadamente a classificação usando a metodologia proposta por este trabalho.

REFERENCES

Bowen, C. Brown's stages of syntactic and morphological development. https://www.speech-language-therapy.com/ speech/index.php?option=com_content&view=article&id=33:brown&catid=2:uncategorised&Itemid=117, 1998.

Busygin, S., Prokopyev, O., and Pardalos, P. Feature selection for consistent biclustering via fractional 0-1 programming. *Journal of combinatorial optimization* 10 (1): 7–21, 2005.

Cheng, Y. and Church, G. M. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, California, USA, pp. 93–103, 2000.

Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, Ü. V. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics* 14 (3): 279–292, 2012.

Gabani, K., Solorio, T., Liu, Y., Hassanali, K.-n., and Dollaghan, C. A. Exploring a corpus-based approach for detecting language impairment in monolingual english-speaking children. *Artificial Intelligence in Medicine* 53 (3): 161–170, 2011.

Gillam, R. B., Cowan, N., and Marler, J. A. Information processing by school-age children with specific language impairment: Evidence from a modality effect paradigm. *Journal of Speech, Language, and Hearing Research* 41 (4): 913–926, 1998.

Hayward, D., Schneider, P., and Gillam, R. B. Age and task-related effects on young children's understanding of a complex picture story. *Alberta Journal of Educational Research* 55 (1): 54–72, 2009.

Hinneburg, A., Aggarwal, C. C., and Keim, D. A. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th International Conference on Very Large Data Bases*. VLDB '00. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 506–515, 2000.

Huang, G., Cheng, A., and Gao, Y. Machine learning improvements to the accuracy of predicting specific language impairment. In *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, Xi'an, China, pp. 553–566, 2022.

Huang, Q., Jin, L., and Tao, D. An unsupervised feature ranking scheme by discovering biclusters. In *2009 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, San Antonio, TX, USA, pp. 4970–4975, 2009.

Madeira, S. C. and Oliveira, A. L. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1 (1): 24–45, Jan., 2004.

Noronha, M. D., Henriques, R., Madeira, S. C., and Zárate, L. E. Impact of metrics on biclustering solution and quality: A review. *Pattern Recognition* vol. 127, pp. 108612, 2022.

Sharma, Y. and Singh, B. K. One-dimensional convolutional neural network and hybrid deep-learning paradigm for classification of specific language impaired children using their speech. *Computer Methods and Programs in Biomedicine* vol. 213, pp. 106487, 2022.

Zhao, H., Liew, A., Wang, D., and Yan, H. Biclustering analysis for pattern discovery: Current techniques, comparative studies and applications. *Current Bioinformatics* 7 (1): 43–55, 3, 2012.