7º EnPE
Encontro de Pesquisa e Extensão

COMUNICAÇÃO ORAL

Análise de algoritmos de aprendizagem de máquina para auxílio no diagnóstico do transtorno do espectro autista em crianças

Raquel da Silva Dornelas[1], Danielli Araújo Lima[2]
Instituto Federal do Triângulo Mineiro (IFTM) Campus Patrocínio
Laboratório de Inteligência Computacional e Robótica (LICRo)
Pró-Reitoria de Pesquisa, Pós-graduação e Inovação do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro (PROPI)

Resumo: Atualmente o transtorno do espectro autista é visto como um dos principais distúrbios do neurodesenvolvimento. Sua origem ocorre nos primeiros anos de vida, e em algumas crianças é possível perceber os sintomas rapidamente após o nascimento. Entretanto na maior parte dos casos os sintomas só são constatados entre 1 a 2 anos de idade. O tempo de espera pelo diagnóstico deste transtorno normalmente é lento. Este é um sério problema, uma vez que o diagnóstico precoce possibilita aos profissionais de saúde iniciarem um processo de intervenção individual, desenvolvendo um tratamento que trará ganhos significativos, por ocorrer em um período do desenvolvimento em que o cérebro é altamente maleável. O objetivo deste trabalho é analisar por meio de uma base de dados pública algumas características individuais de crianças que passaram por um processo de triagem médica com a finalidade de encontrar padrões do espectro autista que sejam capazes de auxiliar no diagnóstico deste transtorno. A mineração de dados foi realizada por meio de oito algoritmos de classificação. Quanto maior a acurácia encontrada, maior é o grau de previsibilidade da base de dados que permite através de um conjunto de regras compreender o transtorno do espectro autista.

Palavras-chave: Mineração de Dados. Autismo. Diagnóstico automático. Transtorno do espectro autista. Ciência de Dados.

1 Introdução

O Transtorno do Espectro Autista (TEA) não é uma doença, mas sim uma condição neurológica caracterizada por alterações notáveis no desenvolvimento da linguagem e da interação social. Há também a existência de comportamentos estereotipados e repetitivos, alterações sensoriais e interesses limitados (VIEIRA; BALDIN, 2017). Por se tratar de um espectro o TEA está classificado em três níveis, onde cada um deles se encontra ligado ao grau de apoio exigido pelo paciente: leve (necessita apoio), moderado (necessita de apoio substancial) e severo (necessita de muito apoio substancial) (APA, 2014).

Técnicas de aprendizado de máquina vêm sendo adotadas por pesquisadores com a intenção de aprimorar o método de diagnóstico do TEA otimizando o tempo de análise e precisão

[1] Aluna do Curso de Tecnólogo em Análise e Desenvolvimento de Sistemas do IFTM Campus Patrocínio, <rakels.dornelas@gmail.com>
[2] Professora Efetiva do IFTM Campus Patrocínio, Doutora em Ciência da Computação, <danielli@iftm.edu.br>.

dos sintomas para conceder aos pacientes rápido acesso ao devido tratamento do transtorno (THABTAH; KAMALOV; RAJAB, 2018). Tendo em vista de que a classificação é um dos trabalhos mais importantes realizado no processo de diagnóstico do autismo, este trabalho tem como propósito realizar análises através do uso de algoritmos de aprendizagem de máquina com o intuito de melhorar a eficiência e precisão preditiva em um conjunto de dados ligados à triagens realizadas em crianças para diagnóstico do autismo, contendo características individuais capazes de auxiliar nos casos de Transtorno do Espectro Autista. Além disso, é importante ressaltar que essa análise tem como objetivo apenas proporcionar assistência na previsão ao diagnóstico, pois somente os médicos estão habilitados a emitir laudos atestando que a criança possua o TEA. Sendo assim, caso o resultado apresente uma forte probabilidade em possuir algum grau de autismo é necessário a busca por um especialista para obter o laudo final.

2 Fundamentação Teórica

De acordo com (APA, 2014), o transtorno do espectro autista caracteriza-se por déficits persistentes na comunicação social e na interação social em múltiplos contextos, incluindo déficits na reciprocidade social, em comportamentos não verbais de comunicação usados para interação social e em habilidades para desenvolver, manter e compreender relacionamentos. Uma informação importante abrangendo tanto as amostras clínicas quanto as epidemiológicas foi o de que há uma maior incidência de autismo em meninos do que em meninas, com proporções médias relatadas de cerca de 3,5 a 4,0 meninos para cada menina (KLIN, 2006).

Em algumas crianças os sintomas são percebidos logo após o nascimento. Na maioria dos casos os sintomas só são identificados entre os 12 e 24 meses. Contudo, o diagnóstico do TEA é realizado em média dos 4 aos 5 anos de idade (LOUREIRO; AUZIER, 2019). As causas biológicas do autismo ainda não foram determinadas. Assim, tratamentos curativos por enquanto não são possíveis. Os diagnósticos precoces, ajudam profissionais de saúde a desenvolverem tratamentos paliativos e programas de prevenção para essas crianças, que, com frequência, reduzem a gravidade do transtorno. Se iniciados cedo, esforços de prevenção, às vezes, podem alterar significativamente a trajetória de desenvolvimento de uma criança autista (WHITMAN, 2019).

A classificação de padrões em pacientes é uma das formas que existe para descoberta do transtorno do espectro autista. A computação pode contribuir com essa descoberta, auxiliando através da automatização onde os dados dos pacientes são armazenados em banco de dados e ao atingir uma quantidade suficiente, essa base pode ser manipulada (ALVES; LIMA, 2018). Ás técnicas de mineração de dados podem realizar essa manipulação (LIMA; ZATI; SILVA, 2017), através de algoritmos de aprendizado supervisionado ou não-supervisionado.

3 Metodologia

Para a estruturação e análise dos dados deste trabalho serão adotadas técnicas de mineração de dados capazes de realizar a classificação do Transtorno do Espectro Autista. Usaremos 8 diferentes tipos de algoritmos para o aprendizado de máquina: a árvore de decisão (AD), naive bayes (NB), support vector machine (SVM), multi layer perceptron (MLP), probabilistic neural networks (PNN), random forest, tree ensemble e o gradient boosted. Será usado em cada uma das técnicas de aprendizado supervisionado um conjunto de dados de treinamento constituído por entradas, compostas pelos dados dos pacientes e, saídas, que equivalem às classes do TEA.

O conjunto de dados usados nesse trabalho foi encontrado no UCI Machine Learning Repository. A base de dados possui 292 registros e 21 atributos[3]. De acordo com (DORNELAS; LIMA, 2019), ao realizar a mineração de dados desta mesma base com os 10 atributos

[3] Dados extraídos da UCI Machine Learning Center for Machine Learning and Intelligent Systems <https://archive.ics.uci.edu/ml/about.html>.

constituídos por características comportamentais foram gerados ótimos desempenhos de modelo de classificação obtendo 100% de precisão quanto às taxas de verdadeiro e falso positivo, e consequentemente 0% de erro na classificação a partir da árvore de decisão. Esse resultado foi obtido para confirmar o que foi revisado na literatura, em que bastava que 7 atributos do questionário clínico fosse respondido positivamente, para que uma possível ajuda médica tivesse que ser procurada. Em face dos dados apresentados, neste trabalho foram selecionados os 10 atributos compostos por características individuais coletados por meio de triagens realizadas em crianças de 4 a 11 anos, sendo alguns dados convertidos para base numérica para simplificar no processo de mineração. O objetivo é identificar qual dos algoritmos de classificação possui como resultado a melhor acurácia na predição e, a mais baixa taxa de erro.

4 Descrição da Proposta

Nesta seção será descrito o modelo utilizado para construção do fluxograma contendo a mineração de dados na ferramenta Knime. Neste caso, apresentaremos a classificação do transtorno do espectro autista em crianças usando a ferramenta KNIME Analytics Platform. Ou seja, a mineração dos dados foi realizada através da plataforma de análise KNIME. Foram usados os mesmos nodes(nós) dentro do modelo do fluxo de trabalho em cada um dos algoritmos (ver Figura 1). Os dados estavam armazenados em arquivos do tipo `ARFF`.

Primeiramente foi realizada a leitura dos dados por um no de leitura do KNIME. Em

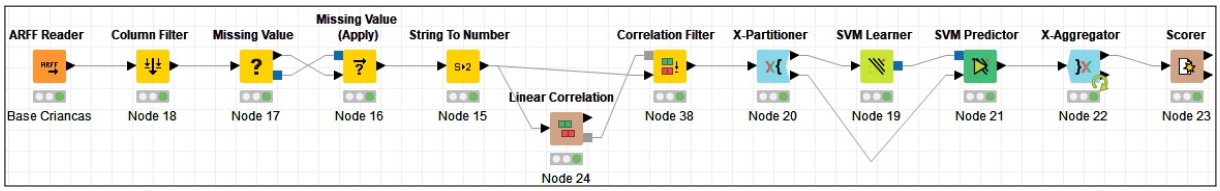

Figura 1 – Interface do usuário do KNIME Analytics Platform apresentando o modelo do fluxo de trabalho, neste exemplo, é apresentado o algoritmo Support Vector Machine (SVM).

seguida, as colunas que representam o questionário das 10 perguntas clínicas $\{A_1, A_2 \cdots A_{10}\}$ já estudadas em (DORNELAS; LIMA, 2019) foram retiradas em virtude de já compreendermos o resultado, em que basta que um somatório (≥ 7) seja obtido para que a família da criança procure ajuda médica. Foi realizada também uma filtragem dos valores faltantes na base de dados (*missing values*), para que esses não atrapalhassem na descoberta do conhecimento pelas técnicas de mineração de dados. Alguns valores passaram de cadeia de caracteres (*string*) para número, para que os algoritmos pudessem manipular os dados de forma adequada. O nó de correlação é responsável por definir as colunas redundantes, para que seja realizado então a filtragem. Para cada coluna no modelo de correlação, a contagem de colunas correlacionadas é determinada com base em um valor limite (*threshold*) para o coeficiente de correlação.

A primeira mineração executada foi sem a utilização de filtro ($= 1$), contendo então todas as características individuais das crianças. A segunda foi realizada contendo como valor limite ($= 0.2$), excluindo apenas a coluna referente ao continente. Já o terceiro filtro com o valor limite de ($= 0.1$) considerou como dados redundantes os atributos de etnia, ictericia e o resultado da Classificação/TEA. Por fim, no filtro com valor limite de ($= 0.05$), foram considerados o gênero, o continente e a Classificação/TEA como as características mais redundantes.

5 Resultados

Ao final do fluxo do trabalho o Scorer gera a matriz de confusão. Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Geralmente, essa tabela de contingência 2×2 especial é também chamada de matriz de erro.

Para o nosso trabalho é possível observar todos os resultados mostrando qual a porcentagem de acurácia e erro de cada um dos algoritmos testados. A Tabela 1 foi montada com base nesses resultados, sendo que à esquerda temos os nomes dos algoritmos usados e à direita temos os resultados obtidos após a aplicação de cada algoritmo alterando-se os filtros através do nó Correlation Filter, uma filtragem presente no KNIME Analytics Platform.

Tabela 1 – Tabela de comparação para análise dos resultados com as diferentes técnicas de aprendizado supervisionado.

Algoritmo	**Resultados**					
Naive Bayes	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	48,973%	51,027%	-0,013	50.000%	50.000%	0,009
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	48,63%	51,37%	-0,028	53,425%	46,575%	0,057
Decision Tree	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	46,223%	53,767%	-0,078	47,26%	52,74%	-0,056
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	58,219%	41,781%	0,163	55,137%	44,863%	0,099
SVM	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	50.000%	50.000%	-0,01	47,26%	54,74%	-0,059
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	46,575%	53,425%	-0,078	52,397%	47,603%	0,032
MLP	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	44,863%	55,137%	-0,104	50,342%	49,658%	0,005
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	58,904%	41,096%	0,177	53,082%	46,918%	0,057
Random Forest	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	53,767%	46,233%	0,068	52,74%	47,26%	0,048
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	58,219%	41,781%	0,162	55,822%	44,178%	0,11
Tree Ensemble	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	52,397%	47,603%	0,039	52,397%	47,603%	0,04
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	55,479%	44,521%	0,109	57,534%	42,466%	0,145
PNN	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	46,918%	53,082%	-0,067	50,685%	49,315%	0,009
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	60,274%	39,726%	0,205	57,534%	42,466%	0,145
Gradient Boosted	**Sem filtro**			**Filtro = 0,2**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	51,37%	48,63%	0,027	51,37%	48,63%	0,027
	Filtro = 0,1			**Filtro = 0,05**		
	Acurácia	Erro	Kappa	Acurácia	Erro	Kappa
	58,904%	41,096%	0,117	58,219%	41,781%	0,162

Dentre as várias técnicas de classificação implementadas neste conjunto de dados para avaliação do desempenho, foi possível observar (ver Tabela 1) que o algoritmo Probabilistic Neural Networks (PNN) com o filtro de correlação em 0.1 (considerando a etnia e a incidência de icterícia) mostrou melhor desempenho comparado aos demais, obtendo o valor de 60.274% de acurácia nos resultados. Adicionalmente, este algoritmo foi também o que apresentou o melhor

resultado em relação ao coeficiente de Cohen kappa ($k = 0.205$). O coeficiente Cohen kappa é uma estatística usada para medir a confiabilidade interexaminadores para itens qualitativos (MCHUGH, 2012). Considera-se geralmente uma medida mais robusta do que o simples cálculo percentual de concordância, pois k leva em consideração a possibilidade de a concordância ocorrer por acaso. Por outro lado, a pior classificação que obtivemos foi com o algoritmo MLP, com acurácia de 44.863 e $k = -0.104$, quando não é usado filtro de correlação ($= 1$).

6 Considerações Finais

O transtorno de autismo afeta o sistema nervoso central, e portanto, a gravidade dos sintomas podem variar amplamente. Dentre os sintomas, podemos citar, dificuldade com interações sociais, interesses obsessivos e até mesmo repetição de comportamentos. Dessa forma, se o reconhecimento for precoce, bem como as terapias é possível reduzir os sintomas melhorando a aprendizagem. Neste artigo foram apresentados diferentes resultados obtidos através de análises dos dados gerados pelos algoritmos de aprendizado de máquina. O melhor resultado obtido foi através da técnica PNN e o pior com a técnica MLP. Em trabalhos futuros, tentaremos entender como a icterícia pode influenciar no diagnóstico do espectro autista.

Referências

ALVES, F. B.; LIMA, D. A. Uso de la clasificación para el análisis y la minería de datos en la herramienta de enseñanza-aprendizaje google classroom. p. 178–189, 2018. Citado na página 2.

APA. **Manual Diagnóstico e Estatístico de Transtornos Mentais - DSM-5**. [S.l.]: ARTMED EDITORA LTDA, 2014. v. 5. Citado 2 vezes nas páginas 1 e 2.

DORNELAS, R. da S.; LIMA, D. A. Ciência de dados aplicada à classificação do comportamento no transtorno do espectro autista. **EnPE**, v. 6, n. 1, 2019. Citado 2 vezes nas páginas 2 e 3.

KLIN, A. Autismo e síndrome de asperger: uma visão geral. **Brazilian Journal of Psychiatry**, SciELO Brasil, v. 28, p. s3–s11, 2006. Citado na página 2.

LIMA, D. A.; ZATI, A. F.; SILVA, E. C. Análise de dados no google classroom para auxiliar na diminuição do distanciamento transacional nas disciplinas da área de lnformática. In: **TISE Conferência Internacional sobre Informática na Educação**. [S.l.: s.n.], 2017. Citado na página 2.

LOUREIRO; AUZIER, A. Transtorno do espectro do autismo. 2019. Citado na página 2.

MCHUGH, M. L. Interrater reliability: the kappa statistic. **Biochemia medica: Biochemia medica**, Medicinska naklada, v. 22, n. 3, p. 276–282, 2012. Citado na página 5.

THABTAH, F.; KAMALOV, F.; RAJAB, K. A new computational intelligence approach to detect autistic features for autism screening. **International journal of medical informatics**, Elsevier, v. 117, p. 112–124, 2018. Citado na página 2.

VIEIRA, N. M.; BALDIN, S. R. Diagnóstico e intervenção de indivíduos com transtorno do espectro autista. **Encontro Internacional de Formação de Professores e Fórum Permanente de Inovação Educacional**, v. 10, n. 1, 2017. Citado na página 1.

WHITMAN, T. L. **O desenvolvimento do autismo**. [S.l.]: M. Books Editora, 2019. Citado na página 2.