ANÁLISE DE DADOS EDUCACIONAIS POR MEIO DE REDES NEURAIS ARTIFICIAIS DO TIPO LVQ

Murilo Luiz Freire da Rocha
murilolfrocha@gmail.com
Antônio Manoel Batista
antonio.manoel@uniube.br

RESUMO

O crescimento de cursos de educação à distância e o consequente aumento do número de alunos nessa modalidade têm trazido à tona um grande volume de dados educacionais que são de suma importância às instituições de ensino. Porém, essas informações não são utilizadas pelos gestores de forma adequada, o que demanda a criação de novas ferramentas capazes de tratá-las de maneira sistemática, afim de estabelecer indicadores acadêmicos para esses gestores. Este trabalho demonstra o desenvolvimento de um sistema que cria perfis característicos de alunos a partir de certas informações obtidas da base de dados de uma universidade - Uniube. O sistema utiliza um tipo de inteligência artificial, chamado de Rede Neural Artificial com vetores quantizadores de aprendizagem, para classificar os dados em grupos distintos. Os resultados mostraram uma distribuição coerente dos conjuntos de alunos, além de outros grupos, divergentes e característicos, o que corrobora o juízo de que esse modelo computacional é satisfatoriamente adequado tanto para a mineração de dados educacionais quanto para a descoberta de conhecimento dentre essas mesmas informações.

Palavras-chave: Educação à Distância. Inteligência Artificial. Classificação. Modelo Computacional.

EDUCATIONAL DATA ANALISYS THROUGH LVQ ARTIFICIAL NEURAL NETWORKS

ABSTRACT

The growth of distance education courses and the consequent increase in the number of students in this modality have brought to light a large volume of educational data that is of great importance to educational institutions. However, this information is not used by managers in an adequate way, which demands the creation of new tools capable of treating them in a systematic way, in order to establish academic indicators for these managers. This work demonstrates the development of a software that creates characteristic profiles of students from certain information obtained from the database of a university - Uniube. The system uses a type of artificial intelligence, called Artificial Neural Network with learning vector quantization, to classify the data into different groups. The results showed a coherent distribution of the groups of students, as well as other groups, divergent and characteristic, which corroborates the opinion that this computational model is satisfactory both for the mining of educational data and for the discovery of knowledge among these same information.

Keywords: Distance Education. Artificial Intelligence. Classification. Computational Model.

1. INTRODUÇÃO

As ferramentas de tecnologia da informação e análise de dados se multiplicam de forma exponencial, assim como as técnicas utilizadas para desenvolvê-las. Desde a simples organização de dados até as modernas ferramentas para reconhecimento facial, um método se destaca: a inteligência artificial. Ao procurar emular o raciocínio humano, ela permite a descoberta de diversos tópicos de estudo e aplicação dela derivados.

A inteligência artificial aplicada se subdivide em vários tipos, como algoritmo genético, redes neurais artificiais e raciocínio baseado em caos. As Redes Neurais Artificiais (RNAs) buscam simular o funcionamento do cérebro humano através de neurônios artificiais, os quais solucionam problemas a partir de um treinamento prévio realizado.

Elas possuem diversas arquiteturas distintas, que apresentam diferentes aplicações

e formas de analisar os dados. Uma delas é a LVQ (*Learning Vector Quantization*, quantização vetorial por aprendizagem), que é tipicamente usada em problemas que necessitam a classificação de padrões. Ela produz saídas separadas em classes distintas, as quais são atribuídas a um vetor de pesos e representam *clusters* ou agrupamentos de dados.

Dentre as técnicas de análise de dados, a mineração de dados e a descoberta de conhecimento possuem um grande atrativo. Com elas, podemos delinear os padrões contidos em grandes quantidades de informação, preparando-os de forma adequada para serem analisados. Portanto, unindo as técnicas de mineração de dados às redes neurais artificiais, é possível criar ferramentas capazes de analisar dados e prever comportamentos através de padrões encontrados.

As RNAs são comumente utilizadas na mineração de dados, pois têm a capacidade de reconhecer padrões dentro de uma grande quantidade de informação. Por isso, seu uso é o mais adequado quando deseja-se realizar predições a partir de comportamentos prévios conhecidos.

Escolas e instituições de ensino possuem uma grande quantidade de dados, sendo uns obtidos para o cadastro dos alunos e outros a partir da evolução deles durante o período em que estiveram matriculados nos cursos. Para que uma instituição possa se desenvolver e manter-se no cenário educacional, é necessário que essas informações indiquem aos gestores uma visão do que os alunos representam no contexto institucional e que relacionem seus dados socioeconômicos com indicadores importantes para a tomada de decisões, tais como estatísticas de evasão e desempenho escolar.

Dessa maneira, a partir da utilização de diversos padrões existentes em instituições de ensino, gerados pelos gestores, e tendo em vista a característica da LVQ em classificá-los, formula-se a seguinte questão: uma rede neural artificial do tipo LVQ pode produzir os indicadores necessários para que os gestores de ensino tomem decisões justificadas sobre mudanças a serem implementadas em suas instituições de origem?

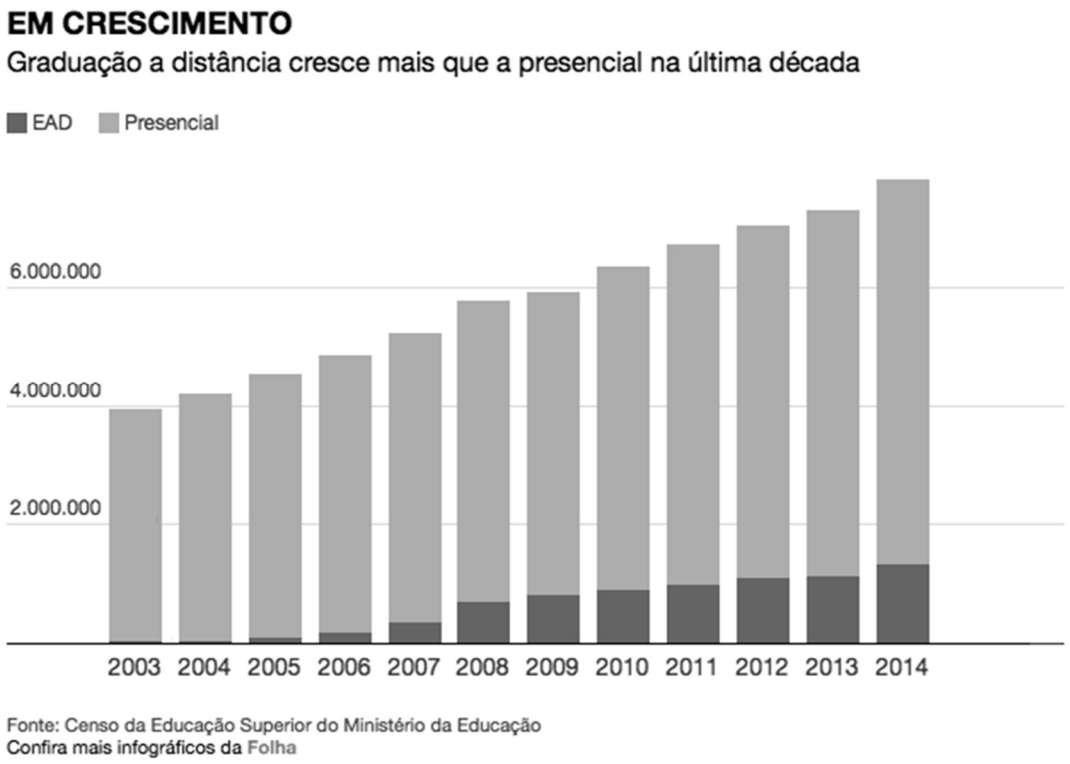
O objetivo do presente trabalho é descrever o desenvolvimento de uma ferramenta para análise de dados educacionais capaz de gerar indicadores acadêmicos a partir dos dados colhidos. Para isso, as metas a serem atingidas são: adquirir e preparar os dados educacionais de alunos para análise, desenvolver a ferramenta e classificar os dados, apresentando-os.

2. O SISTEMA EDUCACIONAL

O sistema educacional sempre foi um grande produtor de dados, sendo eles a respeito de alunos, professores, disciplinas e cursos. Com o passar dos anos e com o advento da internet, essa produção cresceu ainda mais, visto que surgiram modalidades de ensino não presenciais. Elas proporcionaram maior facilidade para os alunos ingressarem nas instituições de ensino, aumentando consideravelmente a quantidade de informação que é coletada e armazenada por elas. A maior parte dessas modalidades de ensino não presenciais fazem o uso dos Ambientes Virtuais de Aprendizagem (AVAs), os quais servem de apoio ao estudante, funcionando como uma plataforma na qual as atividades são realizadas. Essa informatização das atividades estudantis também contribuiu para a ampliação do volume de dados referentes ao ensino.

Conforme pode ser observado na Figura 1, o crescimento da quantidade de alunos na modalidade à distância é maior que o mesmo crescimento na modalidade presencial.

Figura 1 – Aumento da quantidade de alunos por modalidade e ano



Fonte – https://www.educamaisead.com.br/datafiles/uploads/foto_crescimento.png (2016)

Esses dados devem ser aproveitados de maneira adequada. Por meio deles, presume-se, que é possível extrair inúmeras informações benéficas às instituições. Para isso, os dados devem ser analisados de forma metódica, garantindo a qualidade e acurácia das informações, assim como a possibilidade de replicação dessa análise em diversas instituições, com a mesma eficácia.

3. MINERAÇÃO DE DADOS

De acordo com Fayyad (1996), o grande crescimento das bases de dados de empresas, governos e instituições científicas superou a capacidade dos sistemas atuais de digerirem estas informações, criando a necessidade de desenvolvimento de novas ferramentas e técnicas para a análise inteligente das bases de dados. Assim, foi criado o conceito de *Knowledge Discovery in Databases* (KDD), o qual inclui as formas de se encontrar conhecimento e/ou informações relevantes dentro de grandes bancos de dados de instituições.

Existem discussões a respeito da diferenciação entre os termos “mineração de dados” e “KDD”, o que dificulta uma definição concreta. A mineração de dados consiste na separação e classificação automática e sistemática de padrões presentes em bancos de dados, apropriando-se, principalmente, de conhecimentos estatísticos e aprendizado de máquina. O KDD é um conceito mais amplo, o qual abrange a mineração de dados como um todo. Ele se divide em dois grupos: verificação e descoberta de dados. Na verificação, o algoritmo é delineado para separar padrões e mostrar resultados esperados pelo usuário. Já a descoberta de dados procura encontrar novos padrões e identifica-los, produzindo previsões estatísticas desconhecidas. Para o presente trabalho, os termos não serão mais diferenciados.

3.1. Mineração de dados educacionais

A mineração de dados educacionais continua sendo estudada por diversos pesquisadores que utilizam ferramentas e ambientes já consolidados e que fazem uso de vários tipos diferentes de algoritmos. Esses estudos comprovam a eficácia do uso das

Redes Neurais Artificiais para a classificação de dados. Porém, para Manhães (2011), “a acurácia dos classificadores e a taxa de erro são fortemente influenciadas pelos vieses da base de dados”, o que implica na importância da separação correta dos dados e na escolha exata de um algoritmo confiável, já que as diferenças nas bases de dados podem contrariar as expectativas de resultados. Para evitar maiores desvios, faz-se necessária a utilização de ferramentas já existentes como controle para a análise dos resultados. A ferramenta Weka[1], por exemplo, fornece diversos ambientes e algoritmos que permitem uma especificação de resultados controlados para comparação de forma confiável, como pode ser observado em experimentos anteriormente realizados.

Uma das maiores dificuldades na mineração de dados educacionais é a falta de padrões para a catalogação nas bases de dados das instituições. Por isso, as atividades pré-processamento exigem grande esforço por parte dos pesquisadores. De acordo com Costa (2012), deve ser realizada uma tarefa de classificação, a qual procura delimitar um modelo que contenha as classes de dados a serem analisadas. Esse modelo utiliza classes conhecidas para distinguir novas classes que ainda são desconhecidas nas bases de dados, permitindo que a separação seja feita corretamente.

Para realizar esses tratamentos de classificação em padrões, utilizam-se ferramentas analíticas especializadas em mineração de dados. Dentre as técnicas mais utilizadas está a inteligência artificial, devido à forma como ela é aplicada a grandes bases de dados.

4. INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é a ferramenta mais moderna para a análise de dados, desde os numéricos e estatísticos, até os comportamentais. Ela não possui uma definição concreta, sendo descrita por diversos autores, de formas distintas, de acordo com o conceito utilizado para determinar o sucesso do sistema estudado. A Tabela 1 apresenta as diferentes definições da inteligência artificial.

[1] Ferramenta desenvolvida na Universidade de Waikato, Nova Zelândia, que agrega diversos algoritmos de IA dedicados ao aprendizado de máquina.

Tabela 1 – Conceitos de IA de acordo com forma e tipo de sucesso

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal. " (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais. " (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir. " (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas. " (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas. " (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes. " (Poole *et al*., 1998) "AI... está relacionada a um desempenho inteligente de artefatos. " (Nilsson, 1998)

Fonte – Russel (2004)

As definições descritas na figura acima estão dispostas tal que as da parte superior descrevem a IA em relação ao pensamento (raciocínio) e na parte inferior estão aquelas que se relacionam ao comportamento (ação). O lado esquerdo da tabela contém as definições que medem o sucesso baseado no desempenho humano, ou seja, similar ao que uma pessoa faria, e no lado direito estão aquelas que definem o sucesso em relação à racionalidade. Para o presente trabalho, utiliza-se a definição de ação racional, que considera o sistema pelo comportamento e define o sucesso de acordo com o raciocínio.

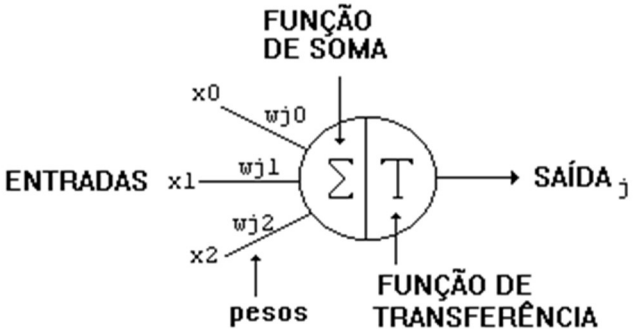
Os sistemas que consideram o sucesso baseado na atuação racional são os que possuem um agente racional, o qual "[...] é aquele que age para alcançar o melhor resultado ou, quando há incerteza, o melhor resultado esperado" (RUSSEL, 2004, p. 6). Ou seja, o sistema é considerado racional quando faz tudo de forma correta (alcança o resultado esperado) ou o mais próximo disso possível.

4.1. Redes Neurais Artificiais

Hoje, a inteligência artificial abrange inúmeras áreas e subcampos, como o algoritmo genético e a Rede Neural Artificial (RNA). Para Silva (2010), as RNAs são modelos que procuram simular a forma com que o sistema nervoso dos seres vivos funciona, adquirindo

conhecimento a partir de informações e são compostas de várias unidades de processamento (neurônios artificiais, representados na Figura 2) interligadas por conexões representadas por vetores e matrizes com pesos distintos. As redes neurais artificiais podem assumir diversas arquiteturas, dependendo da aplicação para a qual será utilizada.

Figura 2 – Estrutura de um neurônio artificial



Fonte – http://www.cerebromente.org.br/n05/tecnologia/neuronio_artificial.gif (2016)

Para o objetivo do presente trabalho, utiliza-se a LVQ, devido a sua característica de agrupamento e classificação de grandes quantidades de dados. O processo de quantização vetorial é realizado atribuindo-se um vetor (quantizador) para cada classe de amostras, as quais deseja-se classificar, e aproximando-se os valores destas amostras, através de operações vetoriais, ao vetor atribuído às respectivas classes. Dessa forma, as amostras são classificadas em classes distintas, a partir de sua aproximação do vetor quantizador.

O treinamento de uma rede neural LVQ começa inicializando-se os vetores quantizadores com valores aleatórios. Então, para cada vetor de entrada, encontra-se o vetor quantizador mais próximo e o atribui a essa entrada, determinando sua classe. O cálculo da distância euclidiana *d* entre o vetor quantizador *q* e o vetor de entrada *p* é realizado de acordo com a quantidade de neurônios *n*, pela seguinte equação:

$$\boldsymbol{d} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{p_i} - \boldsymbol{q_i})^2}$$

Após encontrar a menor distância entre os vetores, calcula-se um valor $\propto$ (alpha) baseado nessa distância, na taxa de aprendizado *t* utilizada e na quantidade de iterações *n*, onde $\propto = (\boldsymbol{d} * \boldsymbol{t})/(\boldsymbol{n} * \boldsymbol{10})$. A equação apresentada foi determinada após diversos testes

para encontrar o cálculo satisfatório do $\propto$, já que não existe uma regra definida para isso. Finalmente, soma-se o $\propto$ ao vetor quantizador mais próximo e subtrai-o dos demais. Assim, obtém-se o conjunto de *clusters* para os dados analisados.

Os *clusters* definidos no treinamento podem, então, ser utilizados para classificar dados relacionados àqueles do treinamento de forma semelhante: atribui-se ao dado de entrada a classe que possui a menor distância euclidiana em relação a ela.

5. ANÁLISE DE DADOS EDUCACIONAIS

Os dados gerados a partir de instituições educacionais, quaisquer sejam, vêm crescendo diariamente, e são de suma importância para o bom funcionamento dessas instituições. Eles trazem informações socioeconômicas que identificam dados como: a distância do local de estudo, faixa de renda familiar, faixa etária, etc. Além disso, possuem informações em relação a histórico escolar, desempenho e assiduidade dos alunos. A partir desses dados, os gestores tomam decisões informadas a respeito de mudanças a serem realizadas no currículo dos cursos, na estrutura escolar e no corpo docente. Porém, as informações são tratadas pelos próprios gestores ou por especialistas em planilhas informativas, o que pode ocasionar erros na classificação e separação dos dados. Para incluir as instituições de ensino na atual era da informação, é preciso desenvolver ferramentas que proporcionem a criação de perfis estabelecidos dos alunos e que possam prever a situação futura dos mesmos, baseando-se em informações passadas e presentes. Auxiliando, assim, os tomadores de decisões com indicadores precisos acerca do comportamento dos estudantes.

6. DELINEAMENTO DO TRABALHO

O presente trabalho consiste de três fases principais que são detalhadas nas seções subsequentes, sendo elas:

- Aquisição e preparação dos dados;
- Desenvolvimento da ferramenta de análise;

• Análise dos resultados com o auxílio da ferramenta desenvolvida.

6.1. Aquisição e preparação dos dados

Os dados necessários para a realização da análise final foram obtidos da Universidade de Uberaba, curso de Direito. Eles são referentes a alunos matriculados no segundo semestre de 2016 e suas informações geradas até o fim desse período. Contém os dados de cadastro dos alunos e do desempenho escolar de cada um. Em relação ao cadastro, apresentam informações como a data de nascimento, sexo, cidade de origem, dados do ensino médio, percentual de bolsa e financiamento, se é portador de diploma de ensino superior e matrícula. O desempenho escolar é quantificado através da pontuação final e do percentual de presença computados ao final de cada período do curso do aluno.

Para serem analisados pela ferramenta desenvolvida, os dados são preparados de modo que sejam representados por números reais. A classificação é realizada a partir das informações presentes na Tabela 2.

Tabela 2 – Informações obtidas dos alunos e descrição da preparação

Informação	**Descrição**
Idade	Obtida a partir da data de nascimento
Sexo	Representado por 0 (Masculino) ou 1 (Feminino)
Distância de origem	Calculada medindo-se a distância da cidade de origem até a cidade da instituição
Tempo de conclusão do ensino médio	Quantidade de anos decorridos desde a conclusão até a data de realização da análise
Natureza do ensino médio	Representado por 0 (instituição pública) ou 1 (particular)
Desempenho no período	Não é necessária a preparação, já que esses dados estão no formato desejado
Percentual de presença	Não é necessária a preparação, já que esses dados estão no formato desejado
Percentual de Bolsa	Não é necessária a preparação, já que esses dados estão no formato desejado
Percentual de Financiamento	Não é necessária a preparação, já que esses dados estão no formato desejado
Portador de diploma	0 (sim) ou 1 (não)
Evasão	Representado por 0 (aluno está matriculado) ou 1 (aluno não está matriculado)

Os valores obtidos devem ser convertidos em números de ponto flutuante variando de 0 a 1. Tal conversão é realizada obtendo os valores mínimo e máximo de cada classe de dados informada e realizando a interpolação do valor dentro do intervalo definido. Por exemplo, se as idades dos alunos possuem o mínimo de 20 anos e o máximo de 40 anos, a idade de um aluno com 30 anos de vida seria representada por 0,5, já que o intervalo definido é de 20 anos e ele encontra-se na metade do mesmo.

Após serem preparados, os dados são inseridos em instâncias da classe desenvolvida, que representa a abstração do aluno no desenvolvimento orientado a objeto, e dispostos em uma lista encadeada, formando o grupo de alunos a serem analisados. O tratamento descrito faz-se necessário, devido à forma com que a rede neural LVQ analisa informações.

6.2. Desenvolvimento da ferramenta

A ferramenta de análise dos dados foi desenvolvida como uma aplicação desktop, ou seja, funciona apenas em computadores comuns e não utiliza recursos de internet, e apenas em ambientes que utilizam o sistema operacional Windows.

Para o desenvolvimento da aplicação foi utilizada a interface de programação *Visual Studio 2017*. O tipo de projeto é o *Windows Forms Application*, o qual possui a função de criar as janelas do programa e permite a criação dos códigos necessários para o desenvolvimento da ferramenta. A linguagem de programação utilizada foi o C#, dentro do framework *.NET,* versão 4.5.2.

O projeto é composto de três classes simples (contém apenas propriedades), as quais abstraem os componentes utilizados no programa, uma classe de negócio, que realiza os processos de treinamento e classificação, e de um formulário principal com a função de exibir os resultados.

A classe Aluno contém as propriedades representativas dos dados dos alunos (idade, sexo, distância de origem, tempo de conclusão do ensino médio, natureza do ensino médio, posse de diploma de curso superior, desempenho, percentual de frequência, percentual de bolsas, percentual de financiamento e evasão), a classe Neurônio possui como propriedades o código de identificação e a lista de grupos nos quais os alunos serão classificados e a classe Grupo representa o agrupamento dos alunos em cada cluster,

contendo as propriedades identificação, quantidade total de alunos, média de idades, quantidade de homens, quantidade de mulheres, média das distâncias de origem, média do tempo decorrido desde o término do ensino médio, quantidade de alunos de escolas públicas, quantidade de alunos de escolas privadas, quantidade de alunos com diploma de curso superior, médias de desempenho, percentual de frequência, percentual de bolsas e percentual de financiamento e a quantidade de alunos não matriculados. A estrutura dessas classes pode ser observada nas Figuras 3, 4 e 5.

Figura 3 – Classe Aluno e suas propriedades

```
public class Aluno
{
    public double Idade { get; set; }
    public double Sexo { get; set; }
    public double DistOrigem { get; set; }
    public double AnosEnsinoMedio { get; set; }
    public double TipoEscola { get; set; }
    public double Diploma { get; set; }
    public double Desempenho { get; set; }
    public double PercFrequencia { get; set; }
    public double PercBolsa { get; set; }
    public double PercFinanciamento { get; set; }
    public double Evadiu { get; set; }
    public int Classe { get; set; }
}
```

Figura 4 – Classe Neurônio e suas propriedades

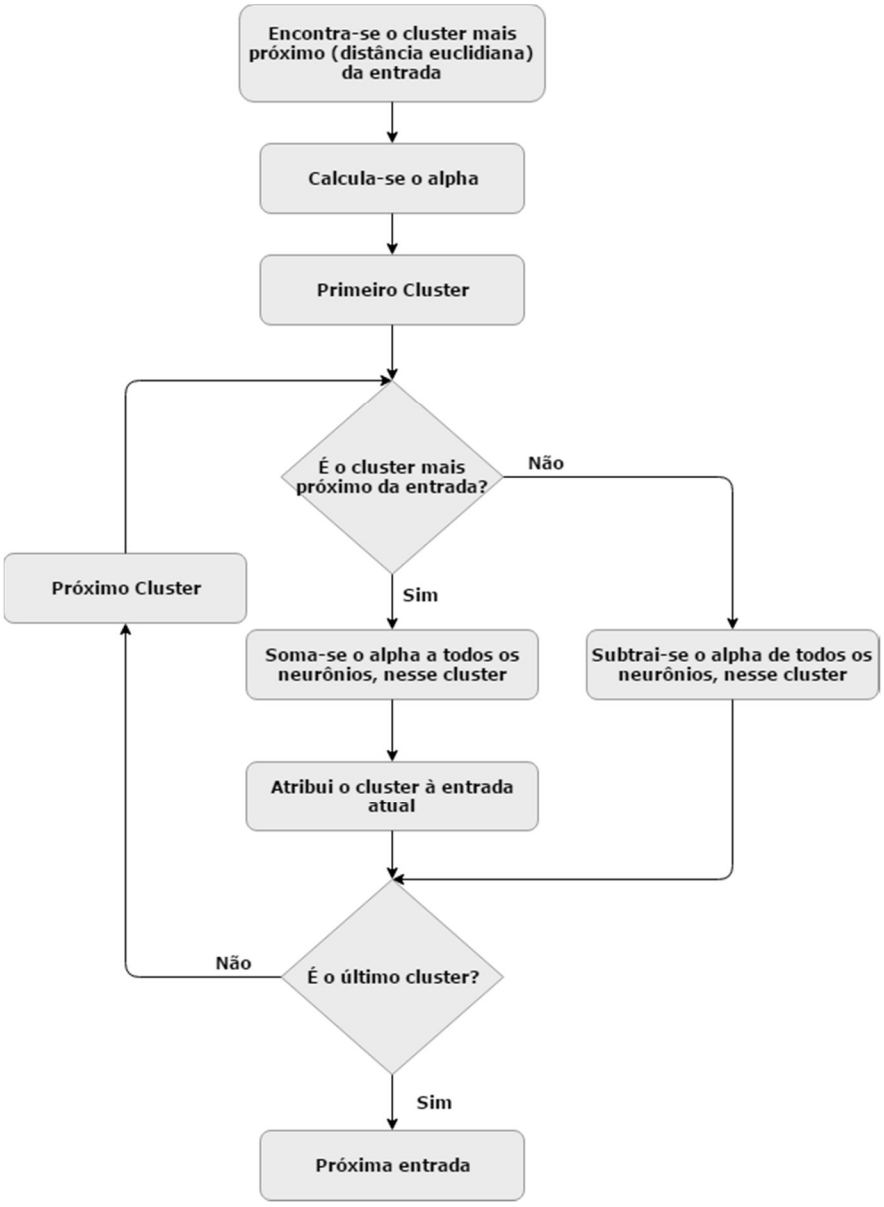
```
public class Neuronio
{
    public int NeuronioId { get; set; }
    public List<double> Clusters { get; set; }
}
```

Figura 5 – Classe Grupo e suas propriedades

```
public class Grupo
{
    public int ClasseId { get; set; }
    public int QtdTotal { get; set; }
    public double Idade { get; set; }
    public int QtdHomens { get; set; }
    public int QtdMulheres { get; set; }
    public double DistOrigem { get; set; }
    public double AnosEnsinoMedio { get; set; }
    public int QtdPublica { get; set; }
    public int QtdPrivada { get; set; }
    public int QtdDiploma { get; set; }
    public double Desempenho { get; set; }
    public double PercFrequencia { get; set; }
    public double PercBolsa { get; set; }
    public double PercFinanciamento { get; set; }
    public double Evasao { get; set; }
}
```

O código que realiza o tratamento dos dados encontra-se na classe RedeNeural, a qual utiliza métodos assíncronos para o processamento dos dados. Esta classe, cujo código-fonte poder ser encontrado no Anexo I deste trabalho, possui o método ClassificarEntradasAsync, que inicializa a lista de Neurônios a serem utilizados pelo processo e realiza as iterações e cálculos necessários para caracterizar os Neurônios e classificar os Alunos. A quantidade de neurônios instanciados é limitada à quantidade de propriedades da classe Aluno e a quantidade de iterações realizadas é igual a dez vezes a quantidade de alunos analisados. Para cada aluno (entrada), o algoritmo do fluxograma da Figura 6 é aplicado.

Figura 6 – Fluxograma do ajuste de clusters por entrada



O formulário principal possui os métodos responsáveis por importar os dados (PreencherAlunos), ativar o processo de treinamento e classificação (TreinarAsync) e apresentar o resultado final da classificação (PreencherDetalhes). O método PreencherAlunos busca os dados dos alunos de uma planilha, instanciando-os em uma lista encadeada, em que cada linha da planilha representa uma instância de Aluno. Para isso, o tratamento dos dados descrito na fase de preparação é aplicado. TreinarAsync é executado em uma Thread diferente daquela utilizada pela aplicação, impedindo o travamento do programa durante o treinamento da rede neural. Ele aguarda a execução do método responsável pelas iterações de classificação e, então, exibe os resultados no formulário principal, ao chamar o método PreencherDetalhes. Este agrupa os alunos pertencentes a cada classe, calculando as médias e totais para cada atributo analisado.

7. RESULTADOS

Após a preparação dos dados, eles são carregados na ferramenta desenvolvida para análise. Foram realizadas 2 classificações distintas afim de examinar o comportamento da RNA ao agrupar os dados. Na primeira execução, foi determinado à ferramenta que gerasse 3 grupos de classificação e, na segunda, 5. Essa disposição foi considerada para estudar a distribuição dos alunos em quantidades distintas de grupos, criando perfis mais, ou menos, específicos.

Os grupos são representados pela classe Grupo, demonstrada na seção anterior. Ela possui como atributos a quantidade total de alunos, a média de idade, a quantidade de homens, a quantidade de mulheres, a média da distância entre a cidade de origem e a cidade da instituição de ensino, a média de quantos anos se passaram desde a conclusão do ensino médio, a quantidade de alunos provenientes de escolas públicas e privadas, a quantidade de alunos portadores de diploma, a média de desempenho escolar e frequência no período, a média da porcentagem de gratuidade e de financiamento que os alunos possuem e a quantidade de evasões ao final do período. As tabelas 3 e 4 apresentam os dados obtidos.

Tabela 3 – Distribuição dos 3 grupos por atributo

	Grupo 1	Grupo 2	Grupo 3
Total	455	667	587
Homens	452	0	241
Mulheres	3	667	346
Média Idade	27,74	23,43	23,24
Média Distância	28,47	29,93	48,69
Fim Ensino Médio	5,51	4,95	5,15
Pública	81	6	473
Privada	374	661	114
Portador de Diploma	4	17	34
Média Desempenho	69,6	73,08	77,62
Média Frequência	89,17	90,72	93,46
Média Bolsas	33,19	45,83	28,37
Média Financiamento	0,6	4,13	8,18
Evasão	54	48	44

Tabela 4 – Distribuição dos 5 grupos por atributo

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Total	742	39	1	484	443
Homens	55	38	1	186	413
Mulheres	687	1	0	298	30
Média Idade	23,58	21,82	26	22,67	25,24
Média Distância	39,41	91,23	1078	32,04	27,36
Fim Ensino Médio	5,15	3,92	8	4,62	5,89
Pública	35	39	1	432	53
Privada	707	0	0	52	390
Portador de Diploma	40	0	1	14	0
Média Desempenho	75,27	67,69	78,28	78,37	66,55
Média Frequência	92,3	87,04	79,58	93,1	87,88
Média Bolsas	53,58	34,1	0	12,43	34,36
Média Financiamento	1,36	0	0	13,33	0,83
Evasão	38	13	0	56	39

8. DISCUSSÃO

A primeira análise gerou 3 grupos distintos. O grupo 1 possui majoritariamente homens e o grupo 2, mulheres. O terceiro grupo é um grupo mais equilibrado em relação ao sexo dos alunos e pode ser considerado como controle. Dentre os dois primeiros, é possível observar uma maior presença de alunos provenientes de escolas privadas, ao

contrário do grupo 3, onde encontramos superioridade nos alunos de escolas públicas. As médias de desempenho e frequência possuem comportamentos semelhantes através dos grupos, demonstrando a proporção linear entre os dois atributos, que aumentam de acordo com o grupo (menor entre os homens, aumenta com as mulheres e maior no grupo controle). A média do percentual de financiamento, devido a seu baixo índice, não parece influenciar na formação dos grupos. Além disso, a média do percentual de gratuidade (bolsas) é maior no grupo majoritariamente do sexo feminino. Da mesma forma, o tempo decorrido desde o fim do ensino médio e média da distância da cidade de origem não divergem significativamente entre os grupos por causa de sua homogeneidade entre os alunos e do número limitado de grupos. Em relação à evasão, observa-se uma predominância nos grupos com alunos provenientes de escolas privadas e que, proporcionalmente, originam-se de cidades mais próximas à instituição.

A segunda classificação realizada formou 5 grupos de classificação. Assim como na análise anterior, a RNA separou dois grupos distintos pelo sexo (grupos 1 e 5) e um marcado pelo equilíbrio dessa característica (grupo 4). Em relação a esses grupos, os padrões encontrados anteriormente se repetem. Os grupos 4 e 5 apresentam, em sua maioria, alunos provenientes de escolas privadas e o grupo 4, de escolas públicas. A proporção linear entre as médias de frequência e desempenho também pode ser observada nessa fase, com a mesma variação dos grupos anteriores. A média do percentual de gratuidade continua maior no grupo majoritariamente do sexo feminino e, da mesma forma, o percentual de financiamento não parece influenciar na formação dos grupos. Para esses grupos principais, as distâncias da cidade de origem e o tempo desde a conclusão do ensino médio, como visto anteriormente, possuem uma distribuição homogênea. A evasão mantém seu comportamento em relação à distribuição entre os grupos.

Destaca-se, nessa segunda análise, a criação de 2 grupos divergentes dos principais. O grupo 3 possui apenas 1 aluno, e contém algumas características que, provavelmente, foram responsáveis por isolá-lo dos demais. A distância da cidade de origem é consideravelmente maior e a média do percentual de frequência é significativamente menor que a encontrada no restante das classificações. Já o grupo 2 traça um perfil de aluno um pouco diferente e marcante. É formado principalmente por homens, com baixa média de idade, originários de escolas públicas e de cidades da mesma região (distância menor que 100 Km de distância) da instituição de ensino. Esses alunos

têm percentual de frequência dentro dos padrões dos demais grupos, porém com desempenho abaixo dos 70%. O percentual de bolsas acompanha o grupo 5, também formado principalmente por homens. A taxa de evasão é maior que 30% nesse grupo, sendo mais alta do que nos outros, nos quais varia entre 5% e 11%.

9. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho descreveu o desenvolvimento de uma ferramenta para análise dos dados educacionais que atuou sobre as informações de alunos do curso de Direito da Uniube, matriculados no segundo semestre de 2016. Neles, foi aplicado o funcionamento de uma Rede Neural Artificial do tipo LVQ, que os classificou em grupos de maneira satisfatória, visto que foi possível desenvolver a ferramenta e analisar os dados obtidos.

Os perfis criados em ambas as análises realizadas mostram um panorama dos estudantes, onde pode ser observado o comportamento majoritário do conjunto de alunos, além de identificarem grupos que divergem da maioria. A repetição do padrão de comportamento encontrado nos 3 grupos principais comprova a eficácia e coerência da ferramenta ao classificar os mesmos dados em distribuições distintas. Ademais, a criação do perfil distinto na segunda análise demonstra a capacidade de descoberta de conhecimento que o algoritmo selecionado possui.

Para trabalhos futuros, considera-se a ampliação da quantidade de atributos e alunos analisados. Além disso, deve ser verificado o comportamento da Rede Neural Artificial ao classificar os dados em mais grupos, ou seja, criando grupos mais específicos, o que só é possível com um número maior de informações. Devem ser experimentadas, também, taxas de aprendizado distintas e comparadas entre si, aperfeiçoando ainda mais a acurácia e coerência das classificações.

REFERÊNCIAS

COSTA, Evandro et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. In: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO, 2012, Rio de Janeiro, **Anais da Jornada de Atualização em Informática na Educação**, Rio de Janeiro: UFRJ, 2012. Disponível em: < http://br-ie.org/pub/index.php/pie/article/view/2341/2096 >. Acesso em: 30 out. 2016.

FAYYAD, Usama M. **Advances in knowledge discovery and data mining.** Menlo Park: AAAI :MIT, c1996. xiv, 611 p.

KUGLER, Mauricio; TORTATO JÚNIOR, Jorge; LOPES, Heitor S.. Desenvolvimento de uma Rede Neural LVQ em Linguagem VHDL para Aplicações em Tempo-Real. **Proceedings Of The Vi Brazilian Conference On Neural Networks: VI Congresso Brasileiro de Redes Neurais**, São Paulo, Sp, v. 1, n. 1, p.103-108, 5 jun. 2003. Disponível em: <http://tupi.elcom.nitech.ac.jp/publications/cbrn2003_mauricio_kugler.pdf>. Acesso em: 27 maio 2017.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, Porto Alegre, Rs, v. 14, n. 1, p.43-67, jul. 2007. Semestral. Disponível em: <http://www.seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67/3543>. Acesso em: 27 maio 2017.

MANHÃES, Lacy Mary Barbosa et al. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 22., 2011, Aracaju. **Anais do SBIE 2011** Aracaju: SBIE, 2011. Disponível em: < http://www.br-ie.org/pub/index.php/sbie/article/view/1585/1350 >. Acesso em: 30 out. 2016.

RUSSELL, Stuart J**. Inteligência artificial.** Rio de Janeiro (RJ): Elsevier : Campus, c2004. 1021 p.

SILVA, Ivan Nunes da; SPATTI, Danilo Hernane; FLAUZINO, Rogério Andrade. **Redes neurais artificiais:** para engenharia e ciências aplicadas : curso prático. São Paulo (SP): Artliber, 2010. 399 p.

ANEXO I

...ual studio 2017\Projects\Murilo.TCC\Murilo.TCC.RedeNeural\Classes\Aluno.cs 1

```
1  public static Task<List<Aluno>> ClassificarEntradasAsync(List<Aluno> entradasTreinamento,
2                                                           DataGridView dgvAlunos,
3                                                           ProgressBar pgbTreinamento,
4                                                           int qtdClusters)
5  {
6      return Task<List<Aluno>>.Run(() =>
7      {
8          double taxaAprendizado = 0.001;
9          double alpha = 0;
10         int idMenorDistancia = 0;
11         int cont = 0;
12         int qtdTreinamento = entradasTreinamento.Count * 10;
13
14         while (cont < qtdTreinamento)
15         {
16             foreach (var entrada in entradasTreinamento)
17             {
18                 double menorDistancia = 10000000000000000.0;
19                 for (int i = 0; i < qtdClusters; i++)
20                 {
21                     Neuronio n = new Neuronio();
22                     List<double> clusters = new List<double>();
23                     double distancia = 0;
24
25                     for (int j = 0; j < pesos.Count; j++)
26                     {
27                         double valorEntrada = Convert.ToDouble(entrada.GetType()
28                                                                       .GetProperties()[j]
29                                                                       .GetValue(entrada));
30                         n = pesos.Single(neuronio => neuronio.NeuronioId == j);
31                         clusters = (List<double>)n.GetType().GetProperty("Clusters").GetValue(n);
32                         double valorCluster = Convert.ToDouble(clusters[i]);
33
34                         distancia += Math.Pow(valorEntrada - valorCluster, 2);
35                     }
36                     distancia = Math.Sqrt(distancia);
37                     if (distancia < menorDistancia)
38                     {
39                         menorDistancia = distancia;
40                         idMenorDistancia = i;
41                     }
42                 }
43                 alpha = (taxaAprendizado * menorDistancia) / (qtdTreinamento * 10);
44                 for (int i = 0; i < qtdClusters; i++)
45                 {
46                     foreach (Neuronio peso in pesos)
47                     {
48                         if (i == idMenorDistancia)
49                         {
50                             entrada.Classe = i;
51
52                             peso.Clusters[i] += alpha;
53                         }
54                         else
55                             peso.Clusters[i] -= alpha;
56                     }
57                 }
58             }
59             cont++;
60             dgvAlunos.Invoke((MethodInvoker)delegate
61             {
62                 dgvAlunos.Refresh();
63                 pgbTreinamento.PerformStep();
64             });
65         }
66         return entradasTreinamento;
67     });
68 }
69
```