IV Encontro de Iniciação Científica e Tecnológica
IV EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
24 e 25 de outubro de 2019

Agrupamento de Dados para Análise de Satisfação de Estudantes da Educação Profissional e Tecnológica de Nível Superior

JÚLIA MEDEIROS DOMINGUES[1], CRISTIANE AKEMI YAGUINUMA[2], JORGE HENRIQUE DE OLIVEIRA SILVA[3]

[1] Graduando em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP Campus Araraquara, julia.domingues@aluno.ifsp.edu.br.
[2] Docente, IFSP Campus Araraquara, cristiane.yaguinuma@ifsp.edu.br.
[3]Técnico Administrativo, IFSP Campus Araraquara, jorge.henrique@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): Banco de Dados – 1.03.03.03-0

RESUMO: O objetivo geral deste projeto é realizar a tarefa de agrupamento de dados para extrair grupos a partir da análise de dados relativos à Satisfação de Estudantes de Educação Profissional de Nível Superior do Instituto Federal de São Paulo. Desse modo, serão realizadas etapas de pré-processamento dos dados, a tarefa de agrupamento particional com base nos algoritmos *k-means* e *k-modes*, ajustando as medidas de interesse e avaliar, juntamente com os especialistas da área de educação, os grupos extraídos e sua contribuição para analisar a satisfação dos estudantes.

PALAVRAS-CHAVE: mineração de dados; agrupamento particional; *k-means; k-modes*; dados educacionais; satisfação de estudantes.

INTRODUÇÃO

Na área da educação, profissionais como gestores e educadores precisam tomar decisões que dependem de estudos sobre um volumoso conjunto de dados, tarefa inviável manualmente. Neste sentido, a área de mineração de dados pode ser vista como um resultado da evolução natural da tecnologia de informação, permitindo extrair conhecimento a partir do processamento de grandes volumes de dados (HAN; KAMBER; PEI, 2011).

DE BAKER, ISOTANI e DE CARVALHO (2011) apontam que é crescente o potencial para pesquisa, desenvolvimento e aplicação de mineração de dados considerando o cenário da educação brasileira. Segundo esses autores, há diversos desafios na área educacional em função da diversidade da população, de fatores econômicos e socioculturais que são intrínsecos à realidade brasileira. Em especial, há interesse por investigar a satisfação de estudantes com a relação à experiência vivenciada em instituições de ensino, devido a sua importância estratégica para melhorar a qualidade nos serviços educacionais e, consequentemente, gerar maior comprometimento dos alunos.

Neste sentido, este projeto visa estudar uma técnica da mineração de dados chamada agrupamento, aplicando a dados de uma pesquisa sobre satisfação de estudantes matriculados em cursos de nível superior de 28 câmpus do Instituto Federal de São Paulo (IFSP) (SILVA, 2017). Os resultados do projeto devem contribuir para que educadores e gestores possam identificar e analisar grupos associados à satisfação dos estudantes.

FUNDAMENTAÇÃO TEÓRICA

O objetivo de uma técnica de agrupamento é encontrar uma estrutura de *clusters* (grupos) nos dados, onde os objetos pertencentes a cada *cluster* compartilham alguma característica ou propriedade relevante para o domínio do problema em estudo, ou seja, são de alguma maneira similares (FACELI *et al.*, 2011).

Dentre as técnicas de agrupamento, existem os algoritmos particionais baseados no Erro Quadrático. Esses algoritmos otimizam o critério de agrupamento utilizando uma técnica iterativa. O primeiro passo consiste na criação de uma partição inicial, em seguida os objetos são movidos de um *cluster* para outro com o objetivo de melhorar o valor do critério de agrupamento. O critério de agrupamento utilizado por esses algoritmos é o erro quadrático, que garante a propriedade de compactação dos *clusters*. Minimizar o erro quadrático, ou a variação dentro de um *cluster*, é o mesmo que maximizar a variação entre os *clusters* (FACELI *et al.*, 2011).

O processo de agrupamento pode ser dividido em etapas importantes para que se possa garantir que os resultados sejam realmente significativos e úteis (FACELI *et al.*, 2011):

1. Preparação dos dados: Esta etapa engloba vários aspectos relacionados ao pré-processamento, que pode incluir normalizações, conversão de tipos e redução do número de atributos por meio de seleção ou extração de características;

2. Proximidade: Esta etapa consiste em definir as medidas de proximidade apropriadas ao domínio da aplicação e ao tipo de informação que se deseja extrair dos dados. Existem diversas medidas que são apropriadas para calcular a proximidade de objetos, cujos atributos são todos do mesmo tipo.

3. Agrupamento: Nesta etapa em que um ou mais algoritmos de agrupamento são aplicados aos dados para a identificação das possíveis estruturas de *clusters* existentes nos dados.

4. Validação: Esta é a etapa que avalia o resultado de um agrupamento e deve, de forma objetiva, determinar se a resolução é representativa para o conjunto de dados analisados.

5. Interpretação: é o processo de examinar cada *cluster* com relação a seus objetos para rotulá-los, descrevendo a sua natureza.

O principal representante da categoria de algoritmos particionais baseados no Erro Quadrático é o k-médias (*k-means*), que particiona o conjunto de dados em k *clusters*, onde k é um valor fornecido pelo usuário. Os *clusters* são formados de acordo com alguma medida de similaridade. O algoritmo em questão utiliza uma técnica de realocação iterativa, minimizando a distância entre cada objeto e o centroide do *cluster* ao qual ele pertence. Ele é sensível à escolha inicial dos centroides e da sua forma de atualização.

O algoritmo k-médias segue cinco passos, inicialmente são definidos os k centroides de forma aleatória. Em seguida é calculada a distância entre os elementos e cada centroide, gerando assim uma matriz de distâncias. O terceiro passo é alocar os elementos de acordo com a sua distância do centroide de cada classe, a classificação ocorre da seguinte forma, o elemento vai pertencer ao *cluster* representado pelo centroide mais próximo. No quarto passo, são calculados novos centroides para cada *cluster*, os valores das coordenadas dos centroides são refinados. Por fim, o algoritmo repete até a convergência, ou seja, os elementos permanecem no mesmo *cluster*.

Para utilizarmos esse algoritmo trabalhamos com a ferramenta Weka (HALL et al., 2009), um programa de origem da Nova Zelândia, que possibilita de uma maneira mais rápida o processo de mineração de dados.

METODOLOGIA

Foi utilizado um conjunto com 505 dados, referentes a satisfação de estudantes matriculados em cursos de nível superior de 28 câmpus do Instituto Federal de São Paulo (IFSP) (SILVA, 2017). Esse conjunto totalizava 28,51% em dados do curso de Análise e Desenvolvimento de Sistemas, mais que ¼, e 71,49% em dados dos outros 19 cursos entre eles licenciaturas, bacharelados e tecnólogos.

Seguindo as etapas do processo de agrupamento, primeiramente na preparação dos dados foram selecionados dois conjuntos menores de atributos, um referente a satisfação dos estudantes sobre a instituição, nomeados com prefixo SAT, e outro referente a empregabilidade, com prefixo EMP. Os atributos correspondem às seguintes afirmações:

- SAT1: Esta é uma instituição de ensino que pode ser considerada próxima do ideal;
- SAT2: Estou satisfeito com competências profissionais adquiridas ao longo do curso;
- SAT3: De modo geral, minha família está satisfeita com a minha escolha de estudar nesta instituição;
- EMP1: Desconsiderando o efeito do desempenho/crise econômico do país, em geral, a instituição proporciona uma formação que garante o meu ingresso no mercado de trabalho após a graduação;

- EMP2: De modo geral, esta instituição é uma instituição de ensino que os empregadores valorizam;
- EMP3: De modo geral, a realização desta graduação permite uma perspectiva de melhoria profissional;
- EMP4: De modo geral, a instituição confere uma formação profissional esperada pelo mercado de trabalho;
- EMP5: De modo geral, a instituição estimula minha formação continuada (realização de cursos de pós-graduação ou outros tipos de cursos).

Para cada afirmação, os dados consistem em respostas dos estudantes, seguindo o seguinte nível de concordância: 1 para "Sem posição"; 2, "Discordo plenamente"; 3, "Discordo"; 4, "Discordo parcialmente"; 5, "Concordo parcialmente"; 6, "Concordo", e 7, "Concordo plenamente". Em seguida foi realizado o primeiro teste de agrupamento na ferramenta Weka (HALL *et al.*, 2009), utilizando o algoritmo K-médias com valor inicial dos *clusters* igual a dois e a medida de proximidade Euclidiana.

RESULTADOS PRELIMINARES

O resultado obtido foi o seguinte: *Cluster 0* – 212 dados (42%) *e Cluster 1* - 293 dados (58%). A Tabela 1 apresenta os centroides de cada *cluster*.

TABELA 1. Resultado de centroides para dois grupos.
Fonte: próprio autor.

Atributo	Cluster 0	Cluster 1
CURSO	Análise e Desenvolvimento de Sistemas	Análise e Desenvolvimento de Sistemas
SAT1	2.9528	5.2014
SAT2	2.9906	5.9556
SAT3	4.0472	6.3276
EMP1	3.3726	5.9352
EMP2	3.4434	5.7065
EMP3	4.3443	6.3925
EMP4	3.1651	6.0273
EMP5	3.3208	5.7133

A partir desses resultados pode-se concluir, primeiramente a clara divisão em que o *Cluster* 0 possui os valores inferiores a 4.5, ou seja, respostas de discordância, e o *Cluster* 1 possui as respostas superiores a 5, que são as respostas de concordância. Já os dados que contém os cursos, por serem elementos categóricos, recebem a medida de proximidade diferente da Euclidiana, levando em consideração a moda, ou seja, dados que aparecem com mais frequência, sendo assim, em ambos os *clusters*, o resultado é igual a Análise e Desenvolvimento de Sistemas, sendo o elemento maior densidade no conjunto de dados. Outro resultado que pode ser ressaltado, é que em ambos os grupos a categoria com respostas de predominância discordando foi a SAT1, relativa à afirmação sobre a instituição ser considerada próxima do ideal.

Posteriormente, no *cluster* 0 pode-se observar que os alunos também discordam da afirmação SAT2 sobre a satisfação com as competências profissionais adquiridas ao longo do curso. Além disso, esse centroide mostra que os alunos em geral discordam com EMP4 (a instituição confere uma formação profissional esperada pelo mercado de trabalho). Concluindo, assim, que a principal insatisfação dos estudantes que pertencem a esse *cluster*, está ligada às competências profissionais.

Já no *cluster* 1, por outro lado, pode-se observar que os alunos concordam com a SAT3 mostrando que, de modo geral, as famílias estão satisfeitas com as suas escolhas de estudar nesta instituição. Além disso,

concordam com a EMP3 afirmando que a realização da graduação permite uma perspectiva de melhoria profissional. Portanto, o *cluster* 1 representa o grupo dos alunos satisfeitos em relação às competências profissionais, além de concordarem com a satisfação dos familiares referente à instituição escolhida por eles.

CONCLUSÕES

Os resultados alcançados, até esse momento da pesquisa, foram obtidos a partir de um único experimento com k = 2, o que nos conduziu ao questionamento de qual seria a maneira mais adequada de definir um valor para k confiável. A partir de pesquisas no material de fundamentação teórica, conclui-se que há a necessidade de estudar medidas de avaliação que possibilitem a análise e comparação de resultados dos agrupamentos, pois não há uma resposta esperada com a qual comparar os resultados obtidos pelos algoritmos, e, muitas vezes, não existe uma resposta única.

Na literatura, uma das medidas de avaliação utilizadas é a silhueta, que se baseia na proximidade entre os objetos de um *cluster* e na distância dos objetos de um c*luster* ao *cluster* mais próximo. A silhueta pode ser utilizada para avaliar uma partição ou a adequação de cada objeto ao seu *cluster* e consequentemente a qualidade de cada *cluster* individualmente (FACELI *et al.,* 2011). Por essa razão e considerando que a ferramenta Weka, utilizada até este momento, não possui as medidas de avaliação necessárias, optamos por avançar os estudos utilizando as ferramentas da linguagem R (R-PROJECT, 2018). Além disso, há a necessidade de aprofundar os estudos do algoritmo *K-Modes* que possibilita agrupamento de dados categóricos, ou seja, não numéricos, que são frequentes em dados educacionais.

AGRADECIMENTOS

Agradecimentos ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP) - Edital nº 018/2018 - por financiar esta pesquisa.

REFERÊNCIAS

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

DE BAKER, R. S. J.; ISOTANI, S.; DE CARVALHO, A. M. J. B. Mineração de dados educacionais: oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, 2011.

FACELI, K. et al. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

HALL, M. et al. The WEKA Data Mining Software: An Update. **SIGKDD Explor. Newsl.**, v. 11, n. 1, p. 10–18, 2009.

HUANG, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. **Data Mining and Knowledge Discovery**, v.2, p.283-304, 1998.

R-PROJECT. **R:** A Language and Environment for Statistical Computing. Disponível em: <http://www.R-project.org/>. Acesso em: 27 abr. 2018.

SILVA, J. H. O. **Modelo de satisfação de estudantes na Educação Profissional: integrando qualidade em serviços, resultados da aprendizagem, empregabilidade, imagem, valor e lealdade.** 2017. 130f. Dissertação (Mestrado em Gestão de Organizações e Sistemas Públicos). Universidade Federal de São Carlos, 2017.