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 Categóricos: um Estudo de Caso Baseado em Dados Educacionais

MARIA REGINA RAMALHO[1]
CRISTIANE AKEMI YAGUINUMA[2]
CINTIA MAGNO BRAZOROTTO [3]

1-Graduando em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP Campus Araraquara, em maria.regina@aluno.ifsp.edu.br.
2-Docente do Instituto Federal de São Paulo – Campus Araraquara – cristiane.yaguinuma@ifsp.edu.br (Orientadora).
3-Técnico administrativo do Instituto Federal de São Paulo – Campus Araraquara cbrazorotto@ifsp.edu.br (Colaboradora)

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

RESUMO: Na área de educação, há conjuntos de dados extensos que descrevem diversas características dos estudantes, sendo necessárias técnicas automatizadas que auxiliem a análise de dados. Neste contexto, a mineração de dados educacionais (*Educational Data Mining* - EDM) visa explorar conjuntos de dados coletados em ambientes educacionais de modo a obter padrões relevantes que caracterizem o perfil dos estudantes e do processo de ensino-aprendizagem. Dentre as tarefas de mineração de dados, a tarefa de agrupamento de dados tem como objetivo agrupar objetos de acordo com suas características, contribuindo para tomada de decisões estratégicas. Em dados educacionais, são frequentes os atributos categóricos, ou seja, atributos cujos valores são descritos por estados ou categorias, que podem ou não ter alguma relação de ordem, como gênero (feminino, masculino), escolaridade (fundamental, médio, superior), entre outros. Assim, este projeto realizou agrupamento de dados categóricos aplicados a dados educacionais considerando aspectos socioeconômicos. Os resultados obtidos contribuíram para identificar grupos que descrevem o perfil dos estudantes, principalmente em função do tipo de curso (concomitante ou integrado) e renda.

PALAVRAS-CHAVE: *Mineração de Dados; Agrupamento de Dados; Dados Educacionais.*

INTRODUÇÃO

O processamento manual de grandes conjuntos de dados é inviável, sendo necessárias técnicas para estruturar, organizar e analisar os dados disponíveis a fim de identificar padrões interessantes para tomada de decisão. Neste sentido, a área de Mineração de Dados (do inglês, *Data Mining*, ou DM) pode ser vista como um resultado da evolução natural da tecnologia de informação, permitindo extrair conhecimento a partir do processamento de volumes abundantes de dados brutos (HAN; KAMBER; PEI, 2011).

Com a expansão do suporte computacional às atividades educacionais (ambientes virtuais de aprendizagem, sistemas acadêmicos), diversas pesquisas têm utilizado técnicas de mineração de dados no domínio de educação, dando origem à área de mineração de dados educacionais (do inglês, *Educational Data Mining*, ou EDM). EDM é uma área de pesquisa que visa desenvolver métodos para explorar conjuntos de dados coletados em ambientes educacionais, a fim de compreender os fatores que influenciam o processo de ensino-aprendizagem (DE BAKER; ISOTANI; DE CARVALHO, 2011) (PEÑA-AYALA, 2014).

DE BAKER, ISOTANI e DE CARVALHO (2011) apontam que é crescente o potencial para pesquisa, desenvolvimento e aplicação de EDM considerando o cenário da educação brasileira. Segundo esses autores, há diversos desafios em função da diversidade da população, de fatores econômicos e socioculturais que são intrínsecos à realidade brasileira. Neste contexto, são necessárias pesquisas em EDM para apoiar decisões estratégicas que considerem variáveis como regionalidade, estrutura familiar, faixas de renda, raça, faixa etária, gênero e outras, de modo a aprimorar o processo de ensino-aprendizagem nas instituições e contribuir para o delineamento das políticas para a educação.

No contexto de dados educacionais, é frequente a ocorrência de atributos categóricos ou ordinais. Por exemplo, para o atributo raça, alguns estados possíveis são: branco, preto, pardo, amarelo, indígena, etc. Assim, esta pesquisa visa aplicar mineração de dados, especificamente um método de agrupamento que trate dados não numéricos considerando o domínio educacional. Os resultados obtidos permitem identificar perfil dos estudantes a partir de dados socioeconômicos de candidatos a cursos de nível médio do Instituto Federal de São Paulo. Os grupos obtidos identificam traços peculiares dos estudantes aos gestores da área do ensino.

FUNDAMENTAÇÃO TEÓRICA

No âmbito da mineração de dados, o processo de descoberta de conhecimento a partir de dados (do inglês, *Knowledge Discovery from Data*, ou KDD) é composto por uma sequência iterativa das seguintes etapas (HAN; KAMBER; PEI, 2011):

1. Limpeza de dados, para remoção de ruídos e dados inconsistentes;
2. Integração de dados, na qual múltiplas fontes de dados podem ser combinadas;
3. Seleção de dados, na qual dados relevantes para a análise são selecionados do conjunto total de dados;
4. Transformação de dados, onde dados são transformados e consolidados em formatos apropriados para mineração por meio de operações de sumarização ou agregação;
5. Mineração de dados, sendo o processo essencial em que algoritmos são aplicados para extrair padrões de dados;
6. Avaliação de padrões, para identificar padrões realmente interessantes que representem conhecimento, com base em medidas de interesse; e
7. Apresentação do conhecimento, onde técnicas de visualização e representação do conhecimento são utilizadas para apresentar os resultados da mineração aos usuários interessados.

Dentre os métodos de mineração de dados, o agrupamento de dados é bastante utilizado na literatura por permitir a descoberta de grupos de objetos, de modo a maximizar a similaridade entre objetos de um mesmo grupo e minimizar a similaridade entre objetos pertencentes a grupos distintos. Dentre os métodos de agrupamento existentes, os métodos de agrupamento particional visam obter todos os grupos simultaneamente para definir uma partição dos dados. Um dos métodos de agrupamento particional mais conhecidos na literatura é o algoritmo *k-means* e suas variações (HAN; KAMBER; PEI, 2011).

No contexto de dados educacionais, é frequente a ocorrência de atributos categóricos ou ordinais, cujos valores podem variar entre dois ou mais estados. Por exemplo, para o atributo raça, alguns estados possíveis são: branco, preto, pardo, amarelo, indígena, etc. Quando há uma relação de ordem, os atributos são considerados ordinais – por exemplo o atributo renda familiar (menos de um salário mínimo; um salário mínimo; entre 2 e 5 salários mínimos; entre 5 e 10 salários mínimos; acima de 10 salários mínimos). Para realizar o agrupamento considerando tais tipos de atributo, é necessário utilizar funções de cálculo de distância que representem a dissimilaridade entre os estados possíveis. Existem diversas abordagens na literatura para o agrupamento de dados que estendem o algoritmo de particionamento clássico *k-means* considerando atributos categóricos, como *k-modes* (HUANG, 1998), *Fuzzy c-modes* (HUANG; NG, 1999), ROCK (GUHA et al., 2000) e a extensão proposta por Ahmad (2007).

O algoritmo *k-modes* é definido da seguinte forma (HUANG, 1998):

1. Selecione k centroides iniciais, um para cada *cluster* (*cluster* refere-se a um grupo);
2. Alocar uma instância para o *cluster* cujo centroide é mais próximo a ela, utilizando a medida de dissimilaridade de correspondência simples (*simple matching*). Ou seja, considerando dois valores de

atributos categóricos, se eles forem diferentes, a distância é o maior valor de dissimilaridade (valor 1); se eles forem iguais, a distância é o menor valor de dissimilaridade (valor 0);

3. Depois que todas as instâncias tiverem sido alocadas para *clusters*, atualizar os centroides utilizando a moda, ou seja, considerando o valor mais frequente para cada atributo das instâncias de um *cluster*.
4. Repita 2 e 3 até que nenhuma instância tenha alterado de *cluster* após um ciclo completo de execução considerando todo o conjunto de dados.

Embora o algoritmo *k-modes* seja de grande importância para solucionar o agrupamento de dados categóricos na mineração de dados, há poucas opções de plataformas que o implementam. Verificou-se que a plataforma RStudio (R-PROJECT, 2018) apresenta na sua biblioteca de funções o algoritmo *k-modes*. Uma outra opção identificada foi a ferramenta Weka (HALL, 2009), que disponibiliza a implementação *SimpleKmeans*, que também trata dados categóricos, mas não faz menção ao algoritmo *k-modes*. Portanto, como a implementação do algoritmo *k-modes* está disponível na plataforma RStudio, este foi o algoritmo considerado para o desenvolvimento deste projeto.

METODOLOGIA

Para o desenvolvimento do projeto, foram realizadas atividades segundo o processo de KDD.

As etapas de limpeza, seleção e transformação dos dados foram realizadas a partir de uma planilha contendo 23957 linhas e 78 colunas, onde cada linha representa uma inscrição de candidato a algum curso de nível médio do IFSP e as colunas contêm informações socioeconômicas dos candidatos, tais como: dados cadastrais básicos, formação escolar, formação escolar dos pais, escolha do curso, se possui equipamentos eletrônicos, aspectos de deficiências, entre outros. Foram padronizadas as nomenclaturas dos dados e posteriormente realizada transformação para formatos compatíveis com as ferramentas de mineração de dados (formato CSV). Na sequência, foram selecionados atributos para iniciar os estudos, com a supervisão de especialista no domínio. Os atributos selecionados foram: gênero, curso, tipo de curso (Concomitante ou Integrado), raça e renda.

A etapa de mineração de dados, dentro do processo KDD, consiste na aplicação de algoritmo para extrair padrões de dados. No projeto, foi utilizado o software RStudio e especificamente o algoritmo *k-modes*, conforme explicado na seção anterior. A definição da quantidade de *clusters* (k) pode ser realizada por uma técnica chamada silhueta (ROUSSEEUW, 1987) que considera a medida da semelhança de um objeto comparado a outros clusters. Esta métrica é calculada por medida de distância, que pode ser Euclidiana ou de Manhattan. Ao realizar testes com esta técnica não foram obtidos resultados, uma vez que todos os dados são categóricos e a medida de distância não se torna compatível. Como não foi possível definir o valor de k através da silhueta, foram realizadas execuções repetidas para buscar o valor de k que direcionasse à identificação de convergência para grupos similares.

Para verificar se haveria uma reincidência de centroides em diferentes execuções, definiu-se o primeiro experimento para 100 execuções do algoritmo *k-modes* com os seguintes parâmetros: K = 2, 100 interações. Determinou-se o número de 100 execuções para apresentar uma amostragem satisfatória e k = 2 por ser o valor mínimo inicial para realizar agrupamento de dados.

Um segundo experimento foi realizado envolvendo 100 execuções com k=3 e um terceiro experimento com 100 execuções com k=4. Quando analisados os resultados observou-se grande combinação divergente de centroides, não havendo uma reincidência para análise. Portanto, esses resultados não foram considerados, pois cada execução obteve centroides distintos das demais, não indicando uma convergência no perfil dos grupos. Para os resultados gerados das 100 execuções com 2 *clusters* observou-se realmente um padrão de centroides reincidentes, podendo assim direcionar a um perfil do agrupamento dos dados.

RESULTADOS E DISCUSSÃO

Dentre os três experimentos realizados, o experimento com 100 execuções do *k-modes* e 2 *clusters* permitiu a identificação de um padrão de repetição no centroide de alguns *clusters*, consolidando um perfil de grupos nos dados. Na Tabela 1, são apresentados os centroides reincidentes e os respectivos percentuais de repetição, considerando 100 execuções com k=2. Centroides com repetições menores que 4% não são exibidos na Tabela 1, por serem considerados resultados pouco frequentes.

Para melhor visualização dos resultados e observação do comportamento dos *clusters*, foram geradas tabelas (Tabelas 2 a 6) para descrever as características dos *clusters*, considerando os grupos obtidos com os centroides que possuem a maior frequência de repetição (16%). Nessas tabelas, as células em cor cinza destacam os valores mais frequentes em cada *cluster*, que representam os valores dos centroides obtidos por meio da moda com base no algoritmo *k-modes*.

TABELA 1. Centroides reincidentes e respectivos percentuais de repetição.
Fonte: Próprio autor.

	Centroides				
Percentual de repetição	**Gênero**	**Curso**	**Tipo**	**Raça**	**Renda**
16%	MASCULINO	EDIFICAÇÃO	CONCOMITANTE	BRANCA	1 a 2 Salário Mínimo
	MASCULINO	INFORMATICA	INTEGRADO	BRANCA	1 a 2 Salário Mínimo
7%	FEMININO	INFORMATICA	INTEGRADO	BRANCA	1 a 2 Salário Mínimo
	MASCULINO	AUTOMAÇÃO	CONCOMITANTE	BRANCA	1 a 2 Salário Mínimo
6%	MASCULINO	INFORMATICA	CONCOMITANTE	BRANCA	1 a 2 Salário Mínimo
	FEMININO	INFORMATICA	INTEGRADO	BRANCA	1 a 2 Salário Mínimo
4%	MASCULINO	INFORMATICA	INTEGRADO	BRANCA	1 a 2 Salário Mínimo
	MASCULINO	EDIFICAÇÃO	CONCOMITANTE	BRANCA	2 a 3 Salário Mínimo

TABELA 2. Distribuição de tipo de curso para cada cluster.
Fonte: Próprio autor.

	CLUSTER 1	***CLUSTER 2***
Concomitante	0	12004
Integrado	11953	0

TABELA 3. Distribuição dos gêneros para cada cluster.
Fonte: Próprio autor.

	CLUSTER 1	***CLUSTER 2***
Feminino	5703	4667
Masculino	6250	7337

TABELA 4. Distribuição dos cursos para cada cluster.
Fonte: Próprio autor.

	CLUSTER 1	***CLUSTER 2***
Administração	347	1739
Agroindústria	106	79
Automação	1087	1762
Edificações	75	2138

Eletrotécnica	234	658
Informática	5207	419
Mecânica	773	984
Mecatrônica	626	894
Química	892	38

TABELA 5. Distribuição de raças para cada cluster.
Fonte: Próprio autor.

	CLUSTER 1	***CLUSTER 2***
Amarelo	409	180
Branca	7034	6220
Indígena	54	67
Parda	3560	4150
Preta	776	1226
Prefiro não declarar	120	161

TABELA 6. Distribuição das rendas para cada cluster.
Fonte: Próprio autor.

	CLUSTER 1	***CLUSTER 2***
Acima de 20 salários mínimo	96	11
10 a 20 salários mínimo	552	92
5 a 10 salários mínimo	2122	816
3 a 5 salários mínimo	2592	1929
2 a 3 salários mínimo	2700	2843
1 a 2 Salários Mínimo	2996	4586
1 salário mínimo	757	1463
Meio salário mínimo	138	264

Nos experimentos realizados, observou-se um comportamento padronizado da separação dos clusters e seus centroides em tipo do curso integrado e concomitante. Tal configuração dos clusters confirma a divisão de comportamento dos alunos também de acordo com a renda, pois analisando os gráficos para o tipo de curso Concomitante os valores de renda mais frequentes são de 1 salário mínimo junto com 1 a 2 salário mínimo. Para o tipo de curso integrado, os valores mais frequentes de renda são de 1 a 2 salário mínimo até 3 a 5 salário mínimo, ou seja, renda um pouco mais elevada. Em conhecimento preliminar de especialista, há uma percepção de que os alunos com renda mais baixa escolhem os cursos noturnos de qualificação, que são da modalidade Concomitante. A modalidade Integrado, sendo um curso diurno, a renda já se apresenta um pouco mais alta. Pode-se concluir que os candidatos a cursos do tipo Concomitante possuem uma renda ligeiramente mais baixa que os candidatos dos cursos do tipo Integrado.

Na análise de gênero e raça, a divisão se mantém proporcional ao número de inscritos. É nítida a incidência dos cursos preferenciais de informática e edificação em cada modalidade disponível, confirmando assim traços como o gênero masculino ter uma frequência mais acentuada nos cursos de informática e edificações, além da maioria de alunos serem da raça branca.

CONCLUSÕES

Esta pesquisa buscou a identificação de perfil dos alunos inscritos nos cursos técnicos do IFSP, através de dados obtidos de questionários socioeconômicos. Este questionário gerou um conjunto de dados quase na sua totalidade de dados categóricos. A partir destes dados, aplicou-se o método *k-modes*, uma das técnicas de mineração de dados para agrupamento de dados categóricos.

A metodologia definida para os experimentos considerou repetição de execuções do método de agrupamento, com diferentes quantidades de grupos, a fim de obter uma convergência de centroides, indicando uma tendência de perfil nos grupos. Foram analisados os resultados que mostraram reincidência de perfil dos grupos e, como resultado, foi obtido um perfil real dos alunos inscritos, que confirma hipóteses de especialistas da área de educação.

Os perfis dos alunos obtidos com agrupamento revelam as seguintes características entre os tipos de cursos:

- Concomitante – Gênero masculino, raça branca, renda de 1 a 2 salário mínimos e curso técnico profissionalizante em edificações.
- Integrado – Gênero masculino, raça branca, renda de 1 a 3 salários mínimos e curso técnico em informática.

Com perfis tão específicos, traz a possibilidade para os gestores terem embasamento para novas ações educacionais.

Para trabalhos futuros será possível continuar e aprofundar experimentos de outros perfis de aluno e comparar estes resultados gerado desta pesquisa, com outros anos de inscrição, verificando se este perfil permanece e se para os alunos matriculados também se mantem este quadro.

AGRADECIMENTOS

Agradecimentos ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP) - Edital nº 021/2017 - por financiar esta pesquisa.

REFERÊNCIAS

AHMAD, A.; DEY, l. A k-mean clustering algorithm for mixed numeric and categorical data. **Data & Knowledge Engineering**, v. 63, n. 2, p. 503-527, 2007.

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.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: A Robust Clustering Algorithm for Categorical Attributes. **Information Systems**, v. 25, n. 5, p. 345-366, 2000.

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

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

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.

HUANG, Z.; NG, M. K. A fuzzy k-modes algorithm for clustering categorical data. **IEEE Transactions on Fuzzy Systems**, v. 7, p. 446–452, 1999.

PEÑA-AYALA, A. Educational data mining : A survey and a data mining-based analysis of recent works. **Expert Systems with Applications**, v. 41, n. 4, p. 1432–1462, 2014.

ROUSSEEUW, P. J.; Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Computational and Applied Mathematics**, v. 20, p. 53-65, 1987.

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