SEGMENTAÇÃO DO CENSO DEMOGRÁFICO DE MG: UMA PROPOSTA PARA AGRUPAMENTO DE DADOS DE FECUNDIDADE

Jeancarlo Campos Leão[1], Lara Soares Menezes[2]

Resumo: Em razão da transição demográfica, o Brasil contará com uma proporção menor de contribuintes dado o aumento na quantidade de beneficiários, o que pressionará de modo considerável sua despesa e necessidade de financiamento. Assim, o melhor entendimento sobre as características do arranjo familiar, relacionadas com a queda da fecundidade, é relevante para o estudo da sustentabilidade do sistema previdenciário. Apesar da diversidade de casos de uso da mineração de dados, não foram identificados muitos trabalhos com esta abordagem. Com o objetivo de segmentar a base de dados censitários, este trabalho propõe uma nova abordagem de técnica sobre os dados relacionados à fecundidade e ao perfil do arranjo familiar. Foram obtidos grupos deste arranjo com características de maior coesão entre seus membros e separação entre os membros de outros grupos permitindo a otimização de ações sociais com base em perfis da comunidade.

Palavras–chave: Transição demográfica. Dados censitários de Minas Gerais. Mineração de Dados.

Introdução

Grandes mudanças na estrutura demográfica brasileira tem direcionado a atenção para duas das suas principais variáveis relacionadas à transição demográfica: o rápido envelhecimento populacional e a diminuição da população em idade ativa em relação aos aposentados. Em razão da transição demográfica, o Brasil contará com uma proporção menor de contribuintes dado o aumento na quantidade de beneficiários, o que pressionará de modo considerável sua despesa e necessidade de financiamento.

Muito se tem analisado em relação à influência de fatores externos (influência geográfica, de fatores de saúde, urbanização, desenvolvimento tecnológico, de renda e de desenvolvimento econômico). Contudo, não foram identificados estudos sobre a análise das relações entre a fecundidade e das características próprias de cada componente familiar ou do grupo como um todo. Com a possibilidade de encontrar a informação inesperada, a mineração de dados disponibiliza técnicas que, se bem utilizadas juntamente, podem fornecer novas relações ou identificar padrões para construção de conhecimentos.

[1] Docente do IFNMG, Campus Araçuaí. Curso de Análise e Desenvolvimento de Sistemas. Email: jeancarlo.leao@ifnmg.edu.br

[2] Acadêmica do curso de Ciências Atuariais da UFMG, Campus Pampulha (Belo Horizonte) .Email: laras.menezes@gmail.com

O melhor entendimento sobre as características do arranjo familiar, relacionadas com a queda da fecundidade, pode ser considerado relevante para a sustentabilidade do sistema previdenciário. Assim, o objetivo geral deste trabalho é segmentar os dados censitários de 2010 com base nas características relacionadas à fecundidade e os atributos internos da estrutura familiar através de técnicas de mineração de dados.

Material e Métodos

Os dados utilizados nesta pesquisa são constituídos das bases do Censo de 2010 disponibilizados pelo IBGE. Nela, cada entrevista foi feita de forma representativa e foi aplicado a cada entrevista um peso igual ao número de pessoas que ela representa. Dentre os atributos apresentados na especificação de layout (IBGE, 2010), foram considerados relevantes aqueles relacionados à fecundidade e ao perfil do arranjo familiar. Nos experimentos iniciais, foram trabalhados os atributos: V6631,V6632, V6641,V6642,V6660,V6691 e V6692 descritos na documentação de layout de microdados[3] do Censo 2010 (IBGE, 2012) e que, resumidamente, representam a taxa de fecundidade e mortalidade para ambos os sexos separadamente. Para a mineração de dados, o método PCA (ZAKI e MEIRA, 2014) foi utilizado para reduzir um única dimensão, o conjunto destes atributos relacionados à fecundidade, tentando minimizar problemas de alta dimensionalidade na sua mineração e preservando as suas propriedades essenciais.

Os três primeiros componentes principais (PCA) foram selecionados para os experimentos. O primeiro e o segundo componente principal se apresentaram com variação brusca na densidade, concentrando a maior densidade em uma mesma escala. Esta projeção facilmente provocaria o agrupamento de partições com tamanhos muito distintos o que foge ao escopo deste trabalho. O terceiro componente principal apresentou maior homogeneidade e dispersão das entidades e foi selecionado em razão do objetivo de segmentar o conjunto total de dados com um número aproximado de entidades similares em cada grupo.

A medida BetaCV foi utilizada para avaliar e comparar os resultados obtidos de algoritmo distintos em busca daquele que agrupava com melhor coesão e separação dos seus grupos. Esta métrica é definida pela razão entre a distância média do cluster ao significativo distância intercluster (ZAKI e MEIRA, 2014, p. 441). Quanto menor for a razão BetaCV, melhor será o agrupamento uma vez que indica que as distâncias intracluster são, em média, menores do que distâncias intercluster.

[3] A descrição completa e detalhada desta base de dados é disponibilizada na documentação do Censo 2010.

Resultados e Discussão

Além da utilização da técnica de agrupamento usando o algoritmo K-Means, neste trabalho foi proposta a utilização de um novo algoritmo de agrupamento (CTree) que se baseia na combinação de algumas das técnicas apresentadas por Zakie e Meira Jr. (2014). Esta abordagem foi fundamentada nas melhorias obtidas na velocidade do algoritmo; obtenção das relações hierárquicas intracluster; abrangência dos agrupamentos (pontos sem grupo), ser parametrizável e melhores índices BetaCV para grandes quantidades de grupos (Gráfico 2b).

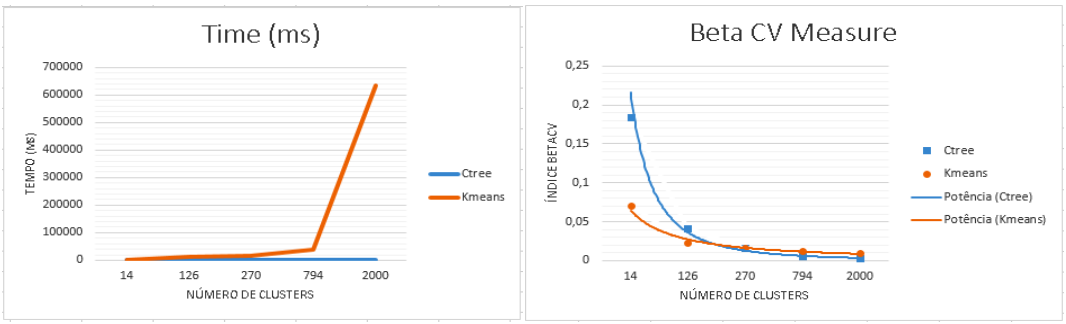

Gráfico 2 – Métricas tempo (a) e BetaCV (b) para diferentes números de clusters sobre mesma amostra de fecundidade.

Conclusões

Técnicas de agrupamento foram utilizadas sobre a base de dados mineira do CENSO de 2010 com o objetivo de segmentar perfis do arranjo familiar. Um novo algoritmo foi concebido objetivando maior velocidade, qualidade do agrupamento e detalhamento de informações interclusters. No contexto deste trabalho, onde há a necessidade de segmentação balanceada, o uso dos parâmetros, permitiu ao algoritmo fornecer grupos mais coesos em suas características de fecundidade. O uso destes dados para melhor direcionamento de investimentos sociais permite atender a perfis de arranjos familiares priorizando os segmentos cujos problemas sociais afetem a taxa de fecundidade. E assim, contribuindo indiretamente para a seguridade da previdência.

Referências

IBGE. Resultados gerais da amostra 2010. p. 239, 2012.

ZAKI, M. J.; MEIRA JR., W. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. 1ª. ed. New York, NY, USA: Cambridge University Press, 2014.

Agradecimentos

Ao IFNMG - Instituto Federal do Norte de Minas Gerais pela condição de bolsista do servidor Jeancarlo Campos Leão através do PBQS - Programa de Bolsas para Qualificação de Servidores. Agradecemos também à empresa que foi receptiva a nós pesquisadores e à ciência neste estudo de caso.