Análise da Correlação de Circutos Integrados por meio de *Clustering* de Acordo com Volume de Importações

Sandro Leite Furtado
Mestrado em Computação Aplicada
Universidade de Brasília
Brasília DF, 70910-900, Brasil
Email: sandroleitefurtado@gmail.com

Tiago Pereira Vidigal
Mestrado em Computação Aplicada
Universidade de Brasília
Brasília DF, 70910-900, Brasil
Email: tiago.vidigal@aluno.unb.br

William Oliveira Camelo
Mestrado em Computação Aplicada
Universidade de Brasília
Brasília DF, 70910-900, Brasil
Email:

***Resumo*—The abstract goes here.**

1. Introdução

A facilidade de se coletar e armazenar dados gerou a necessidade de analisá-los, de acordo com Han et al. [1]. A mineração de dados é o processo de descoberta de padrões significativos de dados e pode ser usado para essa análise como destaca Witten and Frank [2]. Essa abordagem permite o levantamento de informações relevantes para tomadores de decisão fazerem escolhas baseadas em dados.

Chipus Microeletrônica é uma empresa privada brasileira que atua como *Design House* no setor de Semicondutores e possui o interesse de aumentar o número de projetos de produto, como o desenvolvimento de *Application Specific Integrated Circuits* (ASICs). Esse interesse estratégico visa aumentar o potencial ganho constante que esse tipo de projeto fornece em contraste com projetos de serviço. A definição de quais sistemas possuem uma maior chance de sucesso é desafiadora, mas fornece informações valiosas para atingir os objetivos de negócio.

O aumento de projetos de produto pode ser viabilizado mapeando oportunidades de mercado para a plataforma ICX da Chipus, um conjunto configurável por camada de metal de *Intellectual Properties* (IPs) digitais e analógicos [3]. Um único produto capaz de juntar vários componentes clássicos de forma rápida e barata se torna muito chamativo para empresas importando múltiplos circuitos integrados para seus projetos. No entanto, sua efetividade depende que o ICX contenha os circuitos necessários para o cliente com o mínimo de excesso possível por questões de custo, área e consumo.

O objetivo deste trabalho é mapear grupos de *Integrated Circuits* (ICs) que são comumente utilizados juntos. Isso pode ser estimado pelo volume de importações de cada circuito, informação amplamente disponível para o Brasil e outros países. A geração de *clusters* de acordo com tendências de importação permitirá a definição de diferentes versões da plataforma ICX da Chipus que tenham maior chance de impacto no mercado.

A divisão do trabalho é conforme segue: na seção 2 é apresentada a revisão da literatura sobre o tema apresentando conceitos utilizados na concepção e trabalhos correlatos. Na seção 3 é apresentada a metodologia adotada para o desenvolvimento deste trabalho, a seção 4 apresenta os resultados encontrados e finaliza-se com a seção 5 concluindo os achados.

2. Revisão de Literatura

Essa seção apresentará a revisão de literatura do estado da arte relacionado ao escopo deste trabalho. Primeiro será apresentada uma visão geral sobre *Data Mining*, em seguida sobre a técnica escolhida e por fim sobre o escopo negocial a ser analizado.

2.1. *Data Mining*

Data Mining, ou a mineração de dados em português, pode ser entendida como um processo de descoberta de informações estratégicas para predição de futuros comportamentos e, de acordo com Witten and Frank [2] é a extração implícita, previamente desconhecida, e potencialmente utilizável de informações sobre dados.

Para uma empresa, a aprendizagem dos dados já existentes em bancos de dados, textos ou processos pode ser considerada vantajosa no âmbito da competitividade, de acordo Thuraisingham [4] uma vez que informações estratégicas, para tomada de decisão, sejam geradas.

2.2. Clusterização

A clusterização é uma técnica para classificação de dados que pode ser usada quando se tem um volume grande e pouco conhecimento sobre eles [5]. A técnica de mineração coloca dados relacionados ou homogêneos em grupos sem conhecimento avançado dos grupos, essa abordagem é conhecida como clássica (crisp), uma vez que seus elementos estão contidos em uma única classe. [6].

2.2.1. *Time-Series*. Segundo Antunes and Oliveira [7] uma sequencia temporal ou *time-series*, em inglês, é um tipo de clusterização para uma sequência de elementos contínuos de valor real.

2.2.2. *Fuzzy*. A abordagem fuzzy difere da abordagem clássica de clusterização, uma vez que um elemento pode pertencer a várias classes e assumir valores variados [8].

Segundo Bezdek et al. [9], uma abordagem clássica, por vezes pode ser muito restritiva e inviável, uma vez que pode levar a imprecisão e incompletude dos dados. Essa imprecisão pode ocorrer por diferentes fatores, tais como, erros em instrumentos ou ruídos na amostra de dados podem levar, de forma parcial, a valores não confiáveis de determinados atributos.

Nesse sentido, é adequado o uso da Teoria de Conjuntos Fuzzy (TCF) para representar valores imprecisos. Nessa abordagem utilizam-se variáveis linguísticas e restrições para descrever os valores de atributos, ao contrário de fornecer uma representação numérica exata aos dados com valores incertos dos atributos [8].

2.2.3. *Hierarchical clustering*. Essa abordagem atua agrupando atributos de dados (séries temporais) em uma árvore de clusters. São diferenciados dois tipos de métodos de agrupamento hierárquico, sendo eles: aglomerativo e divisivo.

O método de agrupamento hierárquico aglomerativo atribui a cada objeto um cluster próprio, a partir disso, mescla esses clusters menores em clusters cada vez maiores, de modo a se criar um cluster ainda maior e que contenha a representação de todos os objetos e as condições sejam satisfeitas. Cabe destacar que o método aglomerativo é mais usado que o divisivo em geral.

2.3. Circuitos Integrados

Circuito integrado é um conjunto de componentes ativos e passivos interconectados para implementar um sistema eletrônico em escala micrométrica. O crescimento de complexidade destes sistemas tem crescido de forma acelerada ao longo dos anos, requerendo modularização e integração de blocos de circuito específicos. Isso se assemelha ao que ocorreu com sistemas em software.

A modularização de circuitos proporcionou o desenvolvimento de IPs, sistemas autocontidos que podem ser vendidos para reuso em sistemas distintos. A compra de IPs prontos ao invés do desenvolvimento é uma recorrente tomada de decisão a ser feita em projetos de hardware dado o custo de desenvolvimento e manufatura. Usualmente, múltiplos blocos independentes são integrados para cada produto, potencialmente requerendo lidar com diferentes fornecedores.

2.4. CRISP-DM

A mineração de dados é um processo interativo e iterativo em que muitos passos precisam ser repetidamente refinados, a fim de proporcionar uma solução adequada para o problema de análise de dados, de acordo com Wang [10]. O padrão de mineração utilizado neste trabalho será o CRISP-DM (*Cross Industry Standard Process for Data Mining*). Este modelo define um processo que reflete o ciclo de vida de um projeto de mineração, formado por seis fases, são elas: entendimento do negócio, além do entendimento, preparação e modelagem dos dados, finalizando como sua avaliação e implementação. Kononenko and Kukar [11] identifica que os relacionamentos entre essas fases são ciclicamente iterativas até que algum objetivo desejado seja alcançado.

O entendimento do negócio visa o esclarecimento do contexto, objetivos comerciais da empresa e objetivos da mineração de dados. O entendimento dos dados consiste em uma primeira coleta e avaliação do dado disponível, seguido pela sua preparação que consiste em selecionar, limpar, construir e integrar. Por fim, a modelagem e seu respectivo teste permite a avaliação e revisão dos resultados, embasando o processo de implementação para transformar as descobertas obtidas em ações de melhoria para a empresa [12].

2.5. Trabalhos Correlatos

D'Urso et al. [13] abordou a classificação de séries temporais utilizando fuzzy, no contexto financeiro. Neste trabalho foram propostos dois modelos de agrupamento fuzzy baseados em modelos GARCH. Na adequação das medidas de distância para o primeiro modelo D'Urso et al. [13] utilizou a métrica autorregressiva clássica. O segundo utilizou a distância de Caiado, uma distância do tipo Mahalanobis, baseada em parâmetros GARCH estimados e covariâncias que levaram em consideração as informações sobre a estrutura de volatilidade das séries temporais. O trabalho apresenta uma aplicação ao problema de classificação de 29 séries temporais de taxas de câmbio do euro em relação às moedas internacionais e identificou três clusters de taxas de câmbio do Euro, caracterizados por diferentes níveis de estabilidade em termos de flutuações da volatilidade condicional. O trabalhou concluiu que existe uma sugestão de que o uso da distância de Caiado ajuda a descobrir a imprecisão da estrutura do grupo. Os resultados do trabalho mostraram que os clusterings propostos são capazes de revelar características importantes das séries temporais analisadas e detectaram padrões ocultos em grandes amostras de séries temporais financeiras.

3. Metodologia

A metodologia seguida neste trabalho é baseada no *CRISM-DM* [12]. O entendimento de negócios resumido foi consolidado na seção 1, enquanto as demais fases intermediárias são descritas nesta seção. A fase final de implementação foge do escopo deste artigo.

3.1. Entendimento dos Dados

Os dados de importação iniciais estão disponíveis na base de dados da Receita Federal do Brasil. Os campos dos dados obtidos são listados e uma primeira exploração deles levantar as primeiras impressões. O mapeando problemas de formatação e inconsistências a serem tratadas deve ser realizado e registrado para guiar a preparação dos dados.

Importações de outros países podem fornecer informações interessantes, tornando valioso o entendimento desses dados. Demais países do Mercosul são candidatos naturais dada as similaridades, porém integrantes da União Europeia (EU) e dos Estados Unidos (USA) também podem ser usados. A diversificação de países auxilia na redução de particularidades nacionais e aumenta a quantidade de dados que podem ser usados na mineração.

3.2. Preparação dos Dados

Circuitos integrados em trocas internacionais são registrados por códigos de prefixo 8542 como definido pelo padrão de nomenclatura internacional *Harmonized System* (HS) [14]. Esse código pode ser usado para selecionar apenas os registros de interesse dos dados de importação. Essa filtragem pode ser repetida para todos os países sendo analisados.

Os dados selecionados são limpos para resolver quaisquer problemas de qualidade levantados e nenhum dado novo precisará ser construído uma vez que apenas o volume de importações será analisado. A integração dos dados será necessária caso múltiplos países sejam analisados, requerendo padronizar o nome dos campos e a formatação dos dados nos registros.

Os países, apesar de seguirem o padrão HS, geralmente possuem uma nomenclatura derivada com subcategorias para maior especificidade dos produtos. Países do Mercosul, como o Brasil, seguem a Nomenclatura Comum do Mercosul (NCM) [15], países da EU pelo padrão *Combined Nomenclature* (CN) [16] e USA pelo padrão *Schedule B* [17]. Uma vez que as subcategorias entre esses vários padrões não é correspondente, a modelagem deve ser feita por conjuntos de países com nomenclaturas idênticas.

3.3. Modelagem

Os grupos de circuitos integrados mais utilizados juntos podem ser determinados com as tendências de importação, que é uma série temporal de dados de mesmo tamanho. Estes componentes não precisam ser exclusivos de apenas um grupo, uma vez que é comum circuitos clássicos estarem presentes em várias aplicações distintas. Com isso, *fuzzy time-series clustering* é uma modelagem compatível com a análise proposta [5].

O teste dos resultados do modelo serão feitos com um subconjunto dos dados tratados. O critério de excelência é definido pela taxa de erro do modelo usando métodos de medida de similaridade para geração de métricas simples e estáveis [5]. A distancia euclidiana é uma solução clássica e pode ser usada diretamente, mas outro métodos podem ser associados para maior precisão, como correlação cruzada [18].

3.4. Avaliação

As iterações com diferentes configurações do modelo e algoritmos cria diferentes conjuntos de soluções, cada um com um número distinto de *clusters* e composição de circuitos. Essas informações levantadas podem gerar descobertas que devem ser registradas.

As soluções e descobertas devem ser avaliadas considerando os objetivos de negócios junto a um especialista. Uma revisão dessas informações evidencia êxitos e levanta pontos a serem melhorados, podendo inclusive gerar uma nova iteração da mineração a fim de refinar dados e/ou modelo. A mineração pode ser concluída apenas ao atingir os objetivos de negócio.

4. Resultados

To be done.

5. Conclusão

To be done.

Referências

[1] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, Jun. 2011, google-Books-ID: pQws07tdpjoC.

[2] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, Jul. 2005, google-Books-ID: QTnOcZJzlUoC.

[3]

[4] B. Thuraisingham, *Data Mining: Technologies, Techniques, Tools, and Trends*. CRC Press, Dec. 1998, google-Books-ID: UX9yMMpbLFkC.

[5] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - a decade review," *Inf. Syst.*, vol. 53, no. C, p. 16–38, Oct. 2015. [Online]. Available: https://doi.org/10.1016/j.is.2015.04.007

[6] P. Rai and S. Singh, "Article:a survey of clustering techniques," *International Journal of Computer Applications*, vol. 7, no. 12, pp. 1–5, October 2010, published By Foundation of Computer Science.

[7] C. Antunes and A. L. Oliveira, "Temporal data mining: an overview," in *KDD Workshop on Temporal Data Mining*, 2001.

[8] F. S. Yonamine, L. Specia, V. O. Carvalho, and M. C. Nicoletti, "Aprendizado não supervisionado em domínios fuzzy – algoritmo fuzzy c-means," Ph.D. dissertation, Universidade Federal de São Carlos - Centro de Ciências Exatas e de Tecnologia, 2002.

[9] J. Bezdek, E.-K. Tsao, and N. Pal, *Fuzzy Kohonen clustering networks*. IEEE Press, 1992.

[10] J. Wang, *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*. IGI Global, May 2008, google-Books-ID: 1bpEifVEi2MC.

[11] I. Kononenko and M. Kukar, *Machine Learning and Data Mining*. Elsevier, Apr. 2007, google-Books-ID: NUikAgAAQBAJ.

[12] *Guia do IBM SPSS Modeler CRISP-DM*, IBM, 2015. [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/br_po/ModelerCRISPDM.pdf

[13] P. D'Urso, C. Cappelli, D. Di Lallo, and R. Massari, "Clustering of financial time series," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 9, pp. 2114–2129, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437113000617

[14] "Harmonized system nomenclature, section xvi, chapter 85," World Customs Organization, 2017. [Online]. Available: http://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/instruments-and-tools/hs-nomenclature-2017/2017/1685_2017e.pdf?la=en

[15] "Tarifa externa comum - brasil," Secretaria da Fazenda. [Online]. Available: http://www5.sefaz.mt.gov.br/documents/6071037/6401784/Tabela+NCM+-+MDIC+atualizada.pdf/bc780e4b-fd2f-4312-879c-65d5fd1ff49d

[16] "Combined nomenclature," Official Journal of the European Union, 2021. [Online]. Available: http://www5.sefaz.mt.gov.br/documents/6071037/6401784/Tabela+NCM+-+MDIC+atualizada.pdf/bc780e4b-fd2f-4312-879c-65d5fd1ff49d

[17] "Schedule b," US Census Bureau, 2021. [Online]. Available: https://www.census.gov/foreign-trade/schedules/b/2021/index.html

[18] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320305001305