Mineração de regras de associação diversas em dados meteorológicos temporais de múltiplos pontos geográficos via algoritmo genético

Mining of diverse association rules in temporal meteorological data from multiple geographic points via genetic algorithm

DOI:10.34117/bjdv7n12-136

Recebimento dos originais: 12/11/2021
Aceitação para publicação: 01/12/2021

Lucas Ávila Oliveira
Graduando de Ciência da Computação
Departamento de Computação (DC) – Instituto de Biotecnologia (IBiotec) – Universidade Federal de Catalão – (UFCAT)
Endereço: Av. Dr. Lamartine Pinto de Avelar, 1120, Setor Universitário - CEP: 75704-020
E-mail: eu.lucasavila@gmail.com

Matheus Matos Machado
Graduando de Ciência da Computação
Departamento de Computação (DC) – Instituto de Biotecnologia (IBiotec) – Universidade Federal de Catalão – (UFCAT)
Endereço: Av. Dr. Lamartine Pinto de Avelar, 1120, Setor Universitário - CEP: 75704-020
E-mail: m.matos1012.m@gmail.com

Gustavo Evangelista Araújo
Graduando de Ciência da Computação
Departamento de Computação (DC) – Instituto de Biotecnologia (IBiotec) – Universidade Federal de Catalão – (UFCAT)
Endereço: Av. Dr. Lamartine Pinto de Avelar, 1120, Setor Universitário - CEP: 75704-020
E-mail: gustavoevangelistaaraujo@gmail.com

Tércio Alberto dos Santos Filho
Doutor em Engenharia Elétrica, Doutor em Ciência da Computação
Professor Associado do Departamento de Computação (DC) – Instituto de Biotecnologia (IBiotec) – Universidade Federal de Catalão – (UFCAT)
Endereço: Av. Dr. Lamartine Pinto de Avelar, 1120, Setor Universitário - CEP: 75704-020
E-mail: tercioas@ufcat.edu.br

Sérgio Francisco da Silva
Doutor em Engenharia Elétrica, Doutor em Ciência da Computação
Professor Associado do Departamento de Computação (DC) – Instituto de Biotecnologia (IBiotec) – Universidade Federal de Catalão – (UFCAT)
Endereço: Av. Dr. Lamartine Pinto de Avelar, 1120, Setor Universitário - CEP: 75704-020
E-mail: sergio@ufcat.edu.br

RESUMO
O conhecimento das associações de fatores climáticos que influenciam o clima em uma determinada região é importante para análises climáticas e planejamentos de curto a longo prazo. Contudo, os métodos tradicionais existentes na literatura para a descoberta de associações apresentam várias deficiências como alto custo computacional, o que impede sua aplicação até mesmo para conjunto de dados relativamente pequenos, ajuste de vários parâmetros críticos como limiares de suporte e confiança das regras, além de muitas vezes produzirem regras triviais. Tendo em vista esta limitação dos métodos tradicionais da literatura este artigo usa da teoria de Algoritmos Genéticos e suas extensões (memória Tabu e técnica de Nicho, a saber *Clearing*) para desenvolver e experimentar metodologias para mineração de regras de associação de dados temporais quantitativos. Os métodos foram aplicados a dados meteorológicos temporais de múltiplas cidades brasileiras para minerar implicações meteorológicas de um conjunto de cidades na situação meteorológica posterior em um cidade específica. Os experimentos realizados mostram que um dos métodos desenvolvidos que combina memória Tabu e *Clearing*, é bastante promissor, pois minera uma grande quantidade de regras de alta diversidade e não apresenta problema de convergência.

Palavras-chave: algoritmos genéticos, dados temporais quantitativos, mineração de regras de associação .

ABSTRACT
Knowledge of the associations of climatic factors influencing the climate in a given region is important for climate analysis and short to long term planning. However, the traditional methods existing in the literature for discovering associations have several shortcomings such as high computational cost, which prevents their application even for relatively small data sets, adjustment of several critical parameters such as support and confidence thresholds of rules, in addition to often producing trivial rules. Given these limitations of traditional methods in the literature, this article uses the theory of Genetic Algorithms and its extensions (Tabu memory and a Niching technique, namely *Clearing*) to develop and experiment methodologies for association rule mining on quantitative temporal data. The methods were applied to temporal meteorological data from multiple Brazilian cities to mine implications of the weather in a set of cities on the future weather in a specific city. The experiments carried out show that one of the developed methods that combines Tabu memory and *Clearing* is very promising, as it mines a large amount of high diversity rules and does not present the convergence problem.

Keywords: genetic algorithms, quantitative temporal data, mining association rules.

1 INTRODUÇÃO

Muitos fenômenos do mundo real, incluindo atividades, processos e a física da atmosfera, apresentam variáveis correlacionadas. Desta forma, fenômenos reais podem ser melhor compreendidos por descobrir as associações e implicações dos valores de variáveis ao longo do tempo, ou seja, as associações entre certos episódios e suas implicações em episódios futuros. Por exemplo, chuvas na região centro-oeste do Brasil

no período de verão estão normalmente associadas a atuação de massas de ar Tropicais Atlânticas úmidas vindas do Sul e massas de ar Equatorial Continental vindas da Amazônia. Também, as implicações entre episódios são úteis para a geração de previsões (predição). Por exemplo, se há uma forte massa de ar Equatorial Continental atingindo o estado de Mato Grosso e há também uma forte massa de Tropical Atlântica atingindo o estado de São Paulo é altamente provável que em entorno de dois dias formará uma zona de convergência no Estado de Goiás implicando em um alto volume de chuva. Deste modo, no verão, com base nas condições atmosféricas nos estados de São Paulo e Mato Grosso é possível predizer as condições climáticas em Goiás nos próximos dias.

Somada à importância da descoberta de associação em implicações com base em dados temporais quantitativos, fontes deste tipo de dados têm se tornado fartamente disponíveis. Dentre algumas destas fontes pode-se citar: economia, comunicações, astronomia, energia, agronomia, meteorologia e agrometeorologia. Na área de meteorologia, o Brasil, através do Instituto Nacional de Pesquisas Espaciais (INPE), conta com uma grande base de dados, relativa a dados coletados por estações meteorológicas existentes em várias cidades brasileiras. Contudo, os dados produzidos por estas fontes como o INPE têm sido pouco explorados devido a ausência de técnicas para extrair conhecimentos concretos e não-triviais destes. Atualmente, muitos destes dados são analisados por métodos estatísticos e/ou gráficos que têm capacidade limitada para a análise de múltiplas variáveis simultaneamente.

Nos últimos anos, vários métodos de mineração de dados temporais têm sido propostos [George et al. 2021, Cot et al. 2021, Xia et al. 2021, Owadally et al. 2019, Ghosh et al. 2020, Wang et al. 2021, Chen et al. 2020, Wang et al. 2018, Wen et al. 2019a], assim como de mineração de dados quantitativos [Jaramillo et al. 2021, Martín et al. 2018, Moslehi et al. 2020a, Moslehi and Haeri 2020, Medjadba et al. 2020]. Contudo, os métodos existentes na literatura não são efetivamente aplicáveis para extrair múltiplas regras diversas de dados temporais quantitativos, devido a deficiências em satisfazer os seguintes critérios:

- Minerar regras (implicações temporais entre variáveis) e não somente identificar padrões em uma série temporal.
- Operar sem a necessidade do usuário informar parâmetros críticos tais como, limiares de suporte e de confiança das regras pretendidas;
- Minerar várias regras com diversidade e de alta confiança em uma única execução do método;

- Apresentar escalabilidade computacional, de modo a possibilitar sua aplicação para grandes conjuntos de dados.

Métodos da literatura atual não têm a capacidade de minerar regras como:

⟨⟨Campo Verde, Temp. media(oC) = [21.67, 25.79]⟩

AND

⟨Campo Verde, Umidade Rel. Media(%) = [44.99 , 60.24] ⟩

AND

Ibitinga, Temp. media(oC) = [15.16, 19.72]

Horiz.= 1

⇒

⟨Catalao Prec.(mm) = [69.33254704826612 , 93.93254704826612]⟩⟩,

a qual provê conhecimento sobre a precipitação em Catalão no próximo dia (Horiz. = 1) com base nas temperaturas médias em Campo Verde (MT) e Ibitinga (SP) e na umidade relativa do ar em Campo Verde (MT). Este tipo de regra de associação são mineradas pelos métodos baseados em algoritmos genéticos propostos neste artigo. É importante destacar que apesar de ser apresentado neste artigo somente implicações na precipitação para a Cidade de Catalão para o próximo dia, isto é com horizonte de predição igual a um (Horiz. = 1), os métodos permitem definição de quaisquer cidades e quaisquer variáveis para o consequente de regra, assim como usar qualquer horizonte de predição, sendo que este logicamente deve ser maior ou igual a um (1).

Os métodos propostos se baseiam em algoritmo genético de código real, que detectam eventos significativos com base nos valores reais das variáveis. Desta forma, não é feita uma discretização prévia dos dados, que necessitaria de parâmetros adicionais e causaria perda de informações. O algoritmo proposto permite minerar implicações futuras conforme um horizonte de predição especificado pelo usuário. O usuário também pode configurar as variáveis meteorológicas sobre as quais serão extraídas as regras e pontos geográficos do consequente da regra. Nos experimentos deste trabalho foi configurado Catalão como o ponto geográfico do consequente da regra e escolhida a variável precipitação para este ponto. Assim, é possível minerar associações de múltiplas variáveis dos demais pontos geográficos escolhidos que implicam na precipitação em Catalão.

Uma das principais dificuldades do uso de algoritmos genéticos para mineração de regras de associação, a qual tem sido ignorada em várias pesquisas, é a natureza

unimodal destes algoritmos [Petrowski 1996]. Algoritmos genéticos tradicionais são projetados para encontrar uma solução ótima ou sub-ótima. Desta forma, a população (conjunto de soluções candidatas) tende a concentrar em torno da solução ótima no decorrer das gerações (iterações) do algoritmo. Para solucionar este problema, foi usado um mecanismo de preservação de diversidade que permite ao AG explorar simultaneamente várias regiões do espaço de busca e, consequentemente, encontrar várias soluções ótimas e/ou sub-ótimas em uma única execução do método. Vale destacar que o método de preservação de diversidade usado, a saber, *clearing* [Petrowski 1996], é amplamente conhecido pela comunidade de algoritmos genéticos. Contudo, para viabilizar a utilização de *clearing* foi proposta uma medida de distância entre regras, sendo a distância entre duas regras proporcional à diversidade. Adicionalmente, foi usada uma memória Tabu para reter os indivíduos de aptidão máxima ou próxima ao valor máximo encontrados durante a evolução do Algoritmo Genético, descartando estes da população corrente. Para não reinserir um mesmo indivíduo ou indivíduos similares na memória Tabu foi elaborado um filtro com base na distância de um indivíduo aos indivíduos já inseridos no Tabu. Assim, o Tabu retém os indivíduos mais aptos e diversos obtidos pelo processo evolutivo.

O restante deste artigo é organizado da seguinte forma. A Seção 2 apresenta os conceitos base para a definição de regras de associação temporais quantitativas obtidas através de múltiplos pontos geográficos, que são empregados neste trabalho. A Seção 3 sumariza os trabalhos existentes na literatura com focos em mineração de regras de associação temporais e/ou quantitativas. A Seção 4 descreve a base meteorológica de múltiplas cidades brasileiras utilizada, o pré-processamento desta e os métodos baseados em algoritmo genético desenvolvidos para a mineração de regras. Na Seção 5 são reportados e discutidos os resultados obtidos. Por fim, na Seção 6 são destacadas as conclusões.

2 CONCEITOS E DEFINIÇÕES

Este trabalho amplia a noção de regras de associação temporais apresentada em [Silva et al. 2015] para lidar efetivamente com dados meteorológicos. Estes conceitos são base para o desenvolvimento de métodos baseados em algoritmos genéticos para minerar regras que relacionam as condições meteorológicas de múltiplas cidades brasileiras com as condições meteorológicas futuras de Catalão.

De início é importante lembrar o conceito de *itemset* de mineração de regras de associação booleanas. Um *itemset* é basicamente um subconjunto de itens de uma transação. Para abranger o aspecto temporal e quantitativo, um item de um *itemset* passa a ser uma condição intervalar sobre uma variável em um dado intervalo de tempo. Nos experimentos deste trabalho foi considerado o intervalo de tempo como sendo de um dia. Essa definição estendida de item é chamada de episódio. Por exemplo, *temperatura média diária* ∈ [35-39 °C] em *Catalão/Go*. Assim, em substituição ao conceito de *itemset* é definido o conceito de *episodeset* como um conjunto de episódios. Da mesma forma que o elemento base de análise de suporte dos *itemset* são as transações, aqui o elemento base para a análise de suporte dos *episodeset* são os registros temporais das variáveis em múltiplos pontos geográficos (no nosso caso, múltiplas variáveis meteorológicas em múltiplas cidades brasileiras). Contudo, enquanto que a análise de suporte de *itemset* é feita contabilizando a ocorrência dos itens transação por transação, nesta pesquisa conta-se o suporte com base nos registros de um período de tempo; a ideia aqui é prover em mecanismo análogo o de janela deslizante usado para a análise de séries temporais. Estes detalhes são formalizados a seguir.

DEFINIÇÃO DE *EPISODESET*

Seja $\mathbf{V} = \{\mathbf{v_1}^{(p)}, \mathbf{v_2}^{(p)}, \ldots, \mathbf{v_m}^{(p)}\}$, $p = 1, \ldots, n$, o conjunto de variáveis observadas para n pontos de observação. Seja $\mathbf{R} = \{ \mathbf{r_1}^{(p)}, \mathbf{r_2}^{(p)}, \ldots, \mathbf{r_m}^{(p)} \}$ os registros das m variáveis nos n pontos, onde $\mathbf{r_i}^{(p)}$ são os registros da variável v_i no ponto p. Um *episodeset* X é um conjunto de episódios, contudo sua análise de ocorrência se limita a $w(\mathbf{R})$, onde w denota o janelamento.

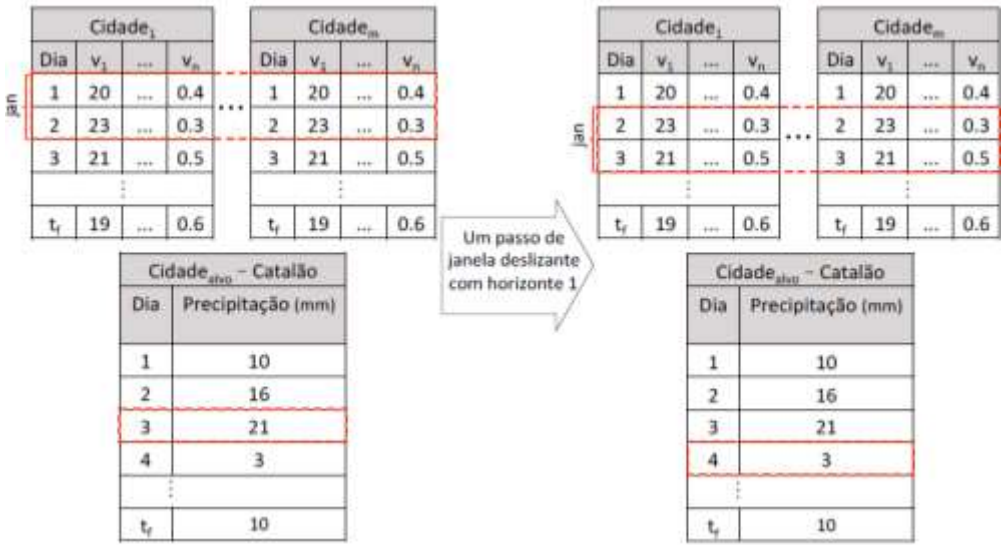
REGRAS TEMPORAIS QUANTITATIVAS

Uma regra de associação temporal quantitativa segue a mesma definição básica de regra de associação, sendo uma implicação da forma $X \Rightarrow Y$ (se X então Y). Contudo, X e Y são conjunções episódicas associadas às variáveis de observação nos pontos geográficos. Como o foco desta pesquisa é minerar regras com implicações de condições meteorológicas dos múltiplos pontos sobre a precipitação na cidade de Catalão/GO, são usadas restrições no Algoritmo Genético através de *flags* para as variáveis, para limitar o consequente a somente episódios relativos a precipitação na cidade de Catalão.

JANELAMENTO E ANÁLISE DE SUPORTE DAS REGRAS

Para a análise de suporte de foi utilizado o tradicional mecanismo de janela deslizante para obter os conjuntos de episódios ao longo do tempo para os cálculos. Para isso, foi utilizado janela temporal de tamanho (*jan*) e o horizonte de predição (*h*). O tamanho da janela temporal (*jan*) corresponde a quantidade de dias sobre todas as cidades e variáveis, que serão utilizadas para checar a ocorrência do antecedente (parte *X*) da regra em análise. Já o horizonte de predição (*h*) é um número que especifica o tempo futuro para o qual será chegado a ocorrência do consequente (parte *Y*) da regra. A Figura 1 ilustra a aplicação de um passo de janela deslizante, com tamanho de janela (*jan*) igual a 2 (dois), e com horizonte de predição (*h*) igual a 1 (um). A parte da esquerda da figura ilustra o posicionamento inicial da janela deslizante e a parte da direita ilustra o posicionamento seguinte. A cada deslizamento da janela são calculadas medidas relacionadas à ocorrência (suporte) do antecedente e consequente da regra.

Figura 1: Ilustração de um passo de janela deslizante sobre os dados temporais de múltiplos pontos geográficos.



3 TRABALHOS CORRELATOS

Nesta seção são descritas as principais pesquisas existentes na literatura que fazem uso de algoritmos genéticos para a mineração de regras de associação quantitativas, temporais e temporais quantitativas. Essas pesquisas são sumarizadas pela Tabela 1 .

As pesquisas de [Martınez-Ballesteros et al. 2016b, Martínez-Ballesteros et al.2014a, Chen et al. 2013] empregam algoritmos genéticos simples para a mineração de regras de associação, sendo que as duas primeiras mineram regras de associação

quantitativas de múltiplas bases de dados fazendo uso de funções de aptidão desenvolvidas pelos autores. Já a última pesquisa minera regras de associação temporais quantitativas de dados financeiros de siderúrgicas, usando densidade e fatores de distorção como medidas de aptidão.

Algoritmos genéticos multiobjetivos têm sido aplicados em vários trabalhos de mineração de regras quantitativas [Martínez-Ballesteros et al. 2016a, Sancho-Asensio et al. 2016, Martínez-Ballesteros et al. 2014b, Matthews et al. 2013] em múltiplas aplicações como análise de dados de microarray e dados sintéticos de compras . Em [Sancho-Asensio et al. 2016, Martínez-Ballesteros et al. 2014b, Martínez-Ballesteros et al. 2011, Matthews et al. 2013] são usados apenas suporte e confiança como funções de aptidão que definem os critérios multiobjetivos. Já em [Martínez-Ballesteros et al.2016a, Martínez-Ballesteros et al. 2011] têm sido aplicado uma quantidade maior de critérios multiobjetivos, incluindo além de suporte e confiança, critérios como interesse, alavancagem, ganho, amplitude e número de atributos na regras. Dentre os trabalhos multiobjetivos listados, o trabalho de [Matthews et al. 2013] tem um diferencial dos demais por minerar regras *fuzzy* quantitativas temporais, enquanto que os demais mineram somente regras quantitativas sem incluir informação temporal.

A pesquisa de [Martín et al. 2018] também lida com algoritmo genéticos multiobjetivos mais o foco é no desenvolvimento de um framework que usa o modelo de programação *map-reduce* para permitir a aplicação desses algoritmos para minerar regras quantitativas de *big-data*.

[Silva et al. 2015] usa um algoritmo genético de nicho para minerar múltiplas regras temporais quantitativas em uma única execução do algoritmo. Para isso é usado como medida de aptidão a confiança relativa que mede a força da implicação. A metodologia é aplicada em múltiplas bases de dados temporais quantitativas.

Trabalhos mais recentes têm desenvolvido metodologia híbridas, envolvendo algoritmos genéticos combinado a outras técnicas para a mineração de regras. Dentre estes, pode-se citar o trabalho de [Moslehi et al. 2020a] que usa um algoritmo genético multiobjetivo combinado a otimização por colônia de partículas (*particle swarm optimization* - PSO) para minerar regras quantitativas de múltiplas bases de dados. Já [Wen et al. 2019a] propuseram um método híbrido de algoritmo genético com o algoritmo de agrupamento *Destiny-based spatial clustering of Application with Noise* (DBSCAN) para encontrar regras temporais quantitativa em dados de tráfego de

trânsito. O algoritmo DBSCAN é utilizado para segmentar as condições de tráfego e o algoritmo genético é aplicado para encontrar implicações temporais quantitativas com base nas condições temporais de tráfego. As regras mineradas são aplicadas para prever níveis de congestionamento de trânsito.

A Tabela 1 sumariza os trabalhos correlatos descritos, sendo estes agrupados conforme o tipo de algoritmo genético empregado, o tipo regra mineradas, a(s) função(ões) de aptidão utilizada(s), os dados sobre os quais os métodos foram aplicados e os critérios usados para avaliar a qualidade das regras mineradas. Com base na tabela pode se notar que uma alta proporção de trabalhos que usam algoritmos genéticos multiobjetivos para mineração de regras quantitativas. Também-se nota uma carência de pesquisas para a mineração de regras temporais quantitativas usando outras modelagens de algoritmos genéticos como o conceito de nicho e o uso de uma memória adicional (chamada de Tabu) para reter os indivíduos ótimos e subótimos da população. Esta pesquisa parte da hipótese que o uso de nicho e memória Tabu auxiliam na descoberta de múltiplas regras diversas em uma única execução do algoritmo genético.

Tabela 1: Tabela de trabalhos correlatos.

Tipo de GA	Tipo de Regra	Função(ões) de aptidão	Dados/aplicação	Critérios de avaliação	Autores
GA simples	quantitativa	propostas pelos autores	diversas bases de repositórios públicos	Influencia dos pesos da medida de aptidão no resultado	[Mart´inez-Ballesteros et al. 2016b]
		propostas pelos autores	diversas bases	comparação de resultados	[Martínez-Ballesteros et al. 2014a]
	temporal quantitativa	densidade e fatores de distorção	dados financeiros de siderúrgicas	Comparação de resultados	[Chen et al. 2013]
GA multiobjetivo	quantitativa	suporte, confiança, alavancagem, precisão, ganho, fator de certeza, amplitude, e o número de atributos da regra	diversas bases de repositórios públicos	suporte da regra, suporte do antecedente, suporte do consequente, confiança, alavancagem, precisão, ganho, fator de certeza, amplitude, e o número de atributos da regra	[Mart´inez-Ballesteros et al. 2016a]
		suporte e confiança	fluxo contínuo de dados não rotulados de regras de associação	suporte e confiança	[Sancho-Asensio et al. 2016]
		suporte e confiança	dados gerados pela aplicação de microarray	comparação de resultados, analisando aspectos como exatidão, precisão, sensibilidade e especificidade	[Mart´inez-Ballesteros et al. 2014b]
		suporte,confiança, precisão, interesse e alavancagem	dados gerado pela aplicação de *microarray*	suporte, confiança, precisão, interesse e alavancagem	[Mart´inez-Ballesteros et al. 2011]

Continua na próxima página

Tipo de GA	Tipo de Regra	Função(ões) de aptidão	Dados/aplicação	Critérios de avaliação	Autores
	temporal quantitativa fuzzy	suporte e confiança	dados sintéticos de cesta de compras de supermercado	comparação e descobrimento de padrões	[Matthews et al. 2013]
GA multiobjetivo com *map-reduce*	quantitativa	suporte e confiança	*big data*	comparação de resultado	[Martin et al. 2018]
GA com nicho	temporal quantitativa	confiança relativa	diversas bases de dados	número de regras e o tempo de execução	[Silva et al. 2015]
Híbrido GA+PSO	quantitativa	confiança, compreensibilidade e interesse	diversas bases de gênero diferentes	numero de regras, confiança e suporte	[Moslehi et al. 2020b]
Híbrido GA + DBSCAN	temporal quantitativa	suporte e confiança	dados de tráfego de trânsito	validação cruzada	[Wen et al. 2019b]

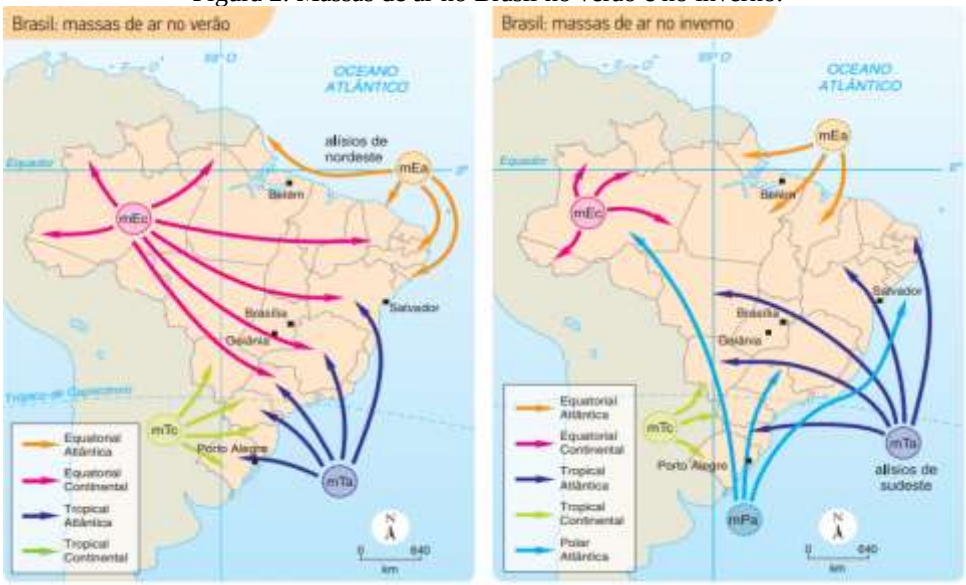
4 MATERIAL E MÉTODOS

Nesta seção é descrito primeiramente a base de dados usada para a mineração de regras e o pré processamento efetuado sobre esta para o tratamento de valores ausentes. Em seguida são descritos os métodos baseados em algoritmos genéticos que foram desenvolvidos.

4.1 BASE DE DADOS E PRÉ PROCESSAMENTO

Os dados meteorológicos foram obtidos através do *site* do Instituto Nacional de Meteorologia (INMET), que possui estações meteorológicas espalhadas por todo o Brasil. Para seleção das estações foram observadas as seguintes condições: Localização de Catalão; Estar sobre influência das mesmas massas de ar, conforme a Figura 2; Possuir disponibilidade de dados de pelo menos 10 anos (de 01/01/2010 até 01/01/2020); Possuir uma quantidade mínima de dados ausentes nesse período de 10 anos, e a estação responsável pela coleta das informações meteorológicas ser automatizada, o que teoricamente é menos suscetível a erros e têm menores quantidade de valores ausentes.

Figura 2: Massas de ar no Brasil no verão e no inverno.



Fonte: SENE e MOREIRA (2013, 153).

Após pesquisa seguindo as condições supracitadas, foram escolhidas 11 cidades para extração dos dados, sendo Campo Verde-MT, Casa Branca-SP, Catalão-GO, Conceição das Alagoas-MG, Florestal-MG, Guiratinga-MT, Ibitinga-SP, Itapaci-GO, Juína-MT, Morrinhos-GO e Silvânia-GO.

Os dados meteorológicos usados, contêm informações sobre os seguintes atributos: código da estação; data da coleta; precipitação total (diária); pressão atmosférica (média diária); temperatura do ponto de orvalho (média diária); temperatura máxima, média e mínima (diária); umidade relativa do ar média e mínima (diária); vento rajada máxima e velocidade média (diária).

4.2 TRATAMENTO DE VALORES AUSENTES - IMPUTAÇÃO

Os dados obtidos na etapa de captação dos dados possuem o problema de valores ausentes, causados por algum tipo de falha na medição ou armazenamento dos valores. Para tratar esse problema visando a aplicação posterior de métodos de predição de valores futuros de séries temporais, foi aplicado o método de imputação denominado regressão pelo k-vizinhos mais próximos (do inglês, *k-nearest neighbors* (kNN))[Zhang 2012], que é um dos métodos mais simples e tradicional da literatura para o preenchimento de valores ausentes. A regressão kNN encontra os registros completos mais próximos do registro ausente e preenche-o pela média da respectiva variável para cada valor ausente encontrado. Mesmo sendo um método simples, há registros na literatura de que a imputação por regressão kNN produz resultados superiores que as de outros métodos.

4.3 ALGORITMOS GENÉTICOS PARA MINERAÇÃO DE REGRAS

Nesta seção são apresentados os algoritmos genéticos desenvolvidos para a mineração de regras de associação temporais quantitativas. Primeiramente é descrito o Algoritmo Genético base, o qual é chamado de Algoritmo Genético Simples, que é o método base para a mineração de regras. Em seguida são apresentadas extensões sobre este método base, as quais são denominadas: Algoritmo Genético com *Niching* - que usa a técnica de *clearing* para manter a diversidade da população, evitando convergência prematura; Algoritmo Genético com Tabu, que usa uma memória para armazenar os melhores indivíduos, os quais são retirados da população do AG; e Algoritmo Genético Completo que usa *Niching* e Tabu.

4.4 ALGORITMO GENÉTICO SIMPLES

Seja V o conjunto de variáveis e P o conjunto de pontos geográficos (as cidades da base de dados). Na codificação de cromossomo para cada elemento de $V \times P$ será associado um gene. Subsequente a estes genes, terá mais u genes associados à cidade de Catalão, onde u corresponde a quantidade de variáveis consideradas para Catalão. Nos experimentos deste trabalho usamos $u=1$, sendo selecionada somente a variável precipitação. Isso significa que a representação de uma regra pode conter episódios no antecedente (parte X da regra) de qualquer variável em qualquer ponto (cidade). Já no caso dos experimentos, os episódios do consequente estão associados somente à precipitação em Catalão/GO. Essa codificação permite mineração de regras que tenham ocorrência de episódios relacionados a Catalão tanto no antecedente quanto no consequente. Contudo, vale destacar que o tempo relacionado aos episódios do consequente é posterior ao tempo dos episódios do antecedente, pois deseja-se minerar implicações futuras.

A representação de cada gene é adaptada de [Silva et al. 2015], não sendo considerado o tempo na representação, uma vez que esta informação é analisada pelo mecanismo de janela deslizante. Assim um gene é codificado pelos seguintes valores (w, v_0, v_1) , onde w é um peso que é comparado a um limiar para indicar se o episódio temporal representado pelo gene fará ou não parte da regra, e v_0 e v_1 são os limites inferior e superior, respectivamente, do intervalo da variável relacionada ao episódio**.**

População inicial : Gerar uma população de cromossomos (C) aleatoriamente de acordo com a codificação de cromossomo para as regras**;**

Avaliação de aptidão : Avaliar cada cromossomo *C* da população conforme a função de aptidão. Como não se planeja-se fazer uma otimização multiobjetivo é preciso definir uma medida única de avaliação das regras. Uma medida usada neste trabalho é chamada de confiança relativa [Yan et al. 2009], dada por:

$$ConfiançaRelativa(X \Rightarrow Y) = \frac{Suport(X \cup Y) - Suporte(X)Suporte(Y)}{Suporte(X)(1 - Suporte(Y))} \quad (1)$$

A confiança relativa mede o grau de relação entre *X* e *Y*. Ela retorna valores máximos quando *X* e *Y* ocorrem em conjunto.

Quanto aos operadores genéticos, foram aplicados operadores tradicionais, sendo eles a seleção por ordenação, cruzamento uniforme, e uma mutação simples. Foi usado seleção por ordenação ao invés de seleção por roleta, devido a seleção por ordenação tem menor pressão seletiva, evitando assim a convergência prematura**.**

4.5 ALGORITMO GENÉTICO COM *NICHING*

Para permitir uma otimização multimodal na mineração de regras, de forma a possibilitar encontrar diversas regras distintas com qualidade máxima em uma única execução do algoritmo, é aplicado um operador de *niching*. O algoritmo de *niching* utilizado, denominado *clearing* e esboçado no algoritmo da Figura 3 . Neste algoritmo o raio de *clearing* ($\boldsymbol{\sigma}$) e a capacidade de nicho ($\boldsymbol{\varkappa}$) deve ser definido pelo usuário. Estes parâmetros têm influência direta na exploração de diversidade de regras no espaço de atributos da base de dados.

Figura 3: algoritmo de *niching* denominado *clearing*.

Input: σ (raio de *clearing*), κ (capacidade de cada nicho).
Output: *Clearing* – atribuição dos recursos de um nicho ao indivíduo mais apto.

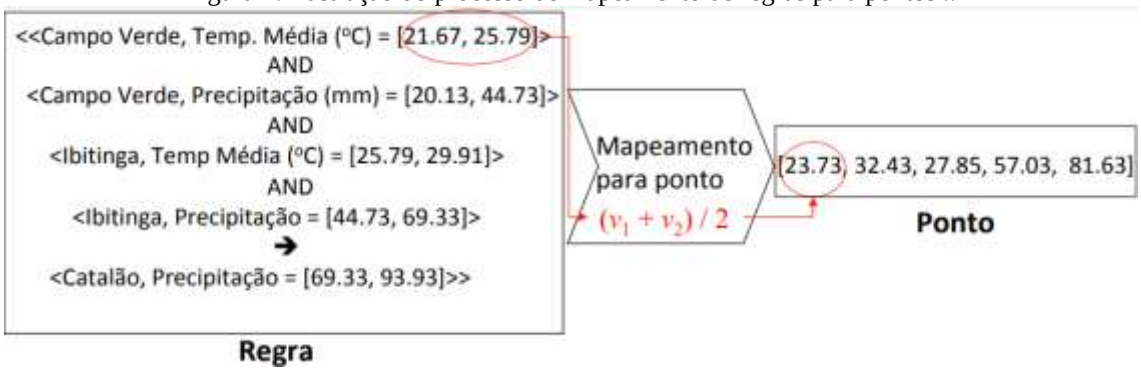
```
1:  OrdenarFitness(C);
2:  para i = 0 até n_C-1
3:      se Fitness(C[i]) > 0
4:          nbWinners = 1;
5:          para j = i + 1 até n_C − 1
6:              se Fitness(C[j]) > 0 AND Distância(f(C[i]), f(C[j])) < σ
7:                  se nbWinners < κ
8:                      nbWinners = nbWinners + 1;
9:                  senão Fitness(C[j]) = 0;
10:                     Fitness(C[j]) = 0;
```

Fonte: Adaptada de [Silva et al 2015]

As funções usadas pelo algoritmo de *clearing* são:

- *OrdenarFitness*(**C**): ordena a população de cromossomos em ordem decrescente de acordo com a aptidão.
- *Fitness*(**C**[*i*]): retorna a aptidão do *i*-ésimo cromossomo da população **C**.
- *Distância*(*f* (**C**[*i*]), *f* (**C**[*j*])): retorna a distância Euclidiana entre dois pontos *f* (**C**[*i*]) e *f* (**C**[*j*]) onde *f* (**C**) é um ponto correspondente ao cromossomo **C**;
- *f* (**C**): é uma função que mapeia o cromossomo **C** para um ponto em espaço dado por $|V x P| + u$ dimensões, através do cálculo da médias dos valores de v_0 e v_1 de cada gene. A Figura 4 ilustra o processo de mapeamento de cromossomo para ponto no espaço de busca. Por simplicidade da ilustração assume-se que *V* tenha duas variáveis, que *P* contenha dois pontos e que *u* seja igual a um, o que corresponde a uma variável relativa a um ponto no consequente da regra.

Figura 4: Ilustração do processo de mapeamento de regras para pontos ..



4.6 ALGORITMO GENÉTICO COM TABU

Algoritmos genéticos, na grande maioria dos casos, são aplicados para buscar uma solução ótima global. Assim, para buscar múltiplos pontos ótimos locais utiliza-se outros métodos/elementos em conjunto com o algoritmo genético tradicional [Kurahashi and Terano 2000]. Um desses elementos usados para permitir encontrar múltiplas soluções ótimas é o uso de uma lista chamada de Tabu, para reter os indivíduos ótimos, retirando-os da população corrente do algoritmo genético. Essa ideia ajuda a evitar convergência prematura pois os indivíduos de alta aptidão são retirados da população, evitando assim que os demais indivíduos tendam convergir para pontos próximos a este indivíduo. Além disso, permite encontrar múltiplas soluções ótimas em uma única execução do algoritmo genético.

Neste trabalho, para garantir que seja minerado um conjunto diverso de soluções ótimas ou sub-ótimas é aplicado um filtro aos indivíduos a serem inseridos no Tabu. Primeiramente, calcula-se a menor distância de um indivíduo candidato a ser inserido no

Tabu (no nosso caso, com aptidão ≥ 0.99) para todos os indivíduos que já se encontram no Tabu. Se essa distância for maior que um dado limiar α, insere-se o indivíduo no Tabu, retirando-o da população corrente. Em seguida é gerado um novo indivíduo aleatório e calculado sua aptidão, o qual é inserido na população corrente, em substituição ao indivíduo removido para o Tabu. O limiar α é calculado por:

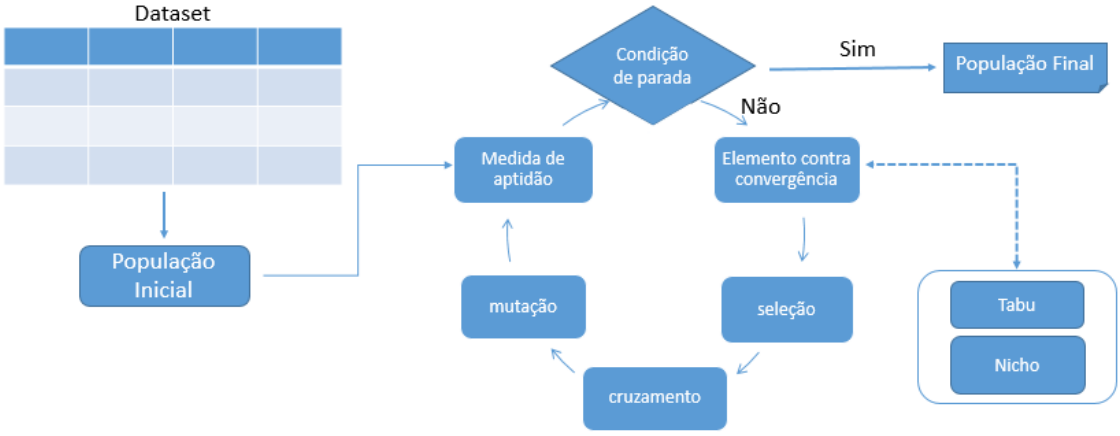
$$\alpha = c * maxDist(f(PopInicial)) \quad (2)$$

onde $f(PopInicial)$ corresponde ao mapeamento de todos os indivíduos da população inicial do algoritmo genético para pontos no espaço $|V \times P| + u$, $maxDist(f$ (PopInicial)) calcula a distância máxima entre os pontos correspondentes aos indivíduos da população inicial, e c é uma constante no intervalo real [0,1] que representa a porcentagem da distância máxima. No experimento foi usado c = 0,2. Isso significa que para um indivíduo ser inserido no Tabu, este deve diferir em 20% ou mais da distância máxima de indivíduos da população inicial para todos os indivíduos já inseridos no Tabu.

4.7 ALGORITMO GENÉTICO COMPLETO

O algoritmo genético denominado como completo é dado pelo algoritmo genético simples acrescido do mecanismo de *Niching* e da memória Tabu. A Figura 5 ilustra os passos do algoritmo genético.

Figura 5: Fluxo de métodos



5 RESULTADOS E DISCUSSÃO

Para a análise de desempenho das diferentes metodologias de Algoritmo Genético desenvolvidas para o problema, foram experimentados três diferentes tamanhos de população (50, 100 e 200 indivíduos) e três diferentes quantidade de gerações (50, 100 e 200 gerações). Como o algoritmo genético conta com elementos

probabilísticos como o operador de seleção, na mutação e no cruzamento, para cada experimento foram efetuadas 10 (dez) execuções usando sementes aleatórias para o gerador de números randômicos. Assim, para cada metodologia de Algoritmo Genético tem-se 9 execuções de experimento, sendo que para cada experimento é reportado a média de 10 execuções. Para cada experimento foi mensurado o número de regras mineradas com aptidão máxima (com valor de confiança relativa, dada pela Equação. 1, igual a 1) e o desvio padrão entre as regras. Para o cálculo de desvio padrão entre as regras faz-se primeiramente o mapeamento das regras para pontos no espaço de regras (operação descrita na Figura 4) e calcula o desvio padrão entre pontos, conforme o posicionamento destes.

Nas Tabelas 2,3,4,5 são apresentados os resultados acerca das regras mineradas pelo Algoritmo Genético Simples, Algoritmo Genético com Clearing, Algoritmo Genético com Tabu e Algoritmo Genético Completo. Nas tabelas, Tpop denota o tamanho do população, nGen é o número de gerações, "n° regras" corresponde ao número de regras mineradas e "desvio padrão (pontos)" corresponde ao desvio padrão das regras mineradas calculado conforme as distâncias entre seus respectivos pontos no espaço de busca, e "desvio padrão (execução) corresponde ao desvio padrão da quantidade de regras mineradas em cada execução" . Desde modo, Tpop e nGen são parâmetros de configuração dos algoritmos genéticos e "n° regras" juntamente com o desvio padrão das regras denota a qualidade dos resultados produzidos pela mineração.

Começando a análise pelo Algoritmo Genético Simples (Tabela 2), nota-se que esse minera uma quantidade pequena de regras em média a cada execução, e que com um tamanho de população (Tpop) fixo ele converge em 50 gerações, não minerando mais regras novas quando é aumentado o número de gerações. Além disso, não há um aumento significativo no número de regras mineradas quando aumenta o tamanho da população, conforme pode ser notado com o aumento da população de 50 indivíduos para 200 indivíduos, onde o número de regras médio minerado aumentou somente de 3,3 para 3,8.

Tabela 2: Resultados do algoritmo Genético Simples

Tpop	nGen	n° regras	desvio padrão (pontos)	desvio padrão (n° regras)
50	50	3,3	5,89	1,42
50	100	3,3	5,89	1,42
50	200	3,3	5,89	1,42
100	50	3,4	7,58	1,43
100	100	3,4	7,58	1,43
100	200	3,4	7,58	1,43
200	50	3,8	8,42	1,40
200	100	3,8	8,42	1,40
200	200	3,8	8,42	1,40

Analisando os resultados do Algoritmo Genético com *Clearing* (Tabela 3) pode-se notar que este minera aproximadamente o dobro de regras em comparação ao GA Simples, e que o algoritmo não converge prematuramente, o que pode ser observado pela quantidade de regras mineradas com um tamanho de população fixo e variado a quantidade de gerações (nGen). Contudo, nota-se que o aumento do tamanho da população não aumenta significativamente a quantidade de regras mineradas.

Tabela 3: Resultados do algoritmo Genético com *Clearing*

Tpop	nGen	n°regras	desvio padrão (pontos)	desvio padrão (n° regras)
50	50	5,2	7,74	2,09
50	100	7,5	7,56	3,04
50	200	7,52	7,58	9,48
100	50	6,9	8,71	2,39
100	100	8,3	8,82	1,85
100	200	8,1	8,81	1,70
200	50	5,5	9,3	2,42
200	100	6,8	9,31	1,94
200	200	8,98	6,8	2,48

Pode-se notar que o Algoritmo Genético com Tabu (Tabela 4) minera uma quantidade maior de regras que os anteriores, contudo a quantidade de regras não aumenta mantendo o tamanho de população fixo e aumentando o número de gerações (nGen) de 50 para 200. Isso acontece devido ao algoritmo convergir com 50 gerações, não acrescentado, dessa forma, regras novas no tabu - devido a essa convergência o desvio padrão do número de regras mineradas a cada execução é o mesmo para um tamanho de população fixo, sendo que o mesmo acontece com o desvio padrão dos pontos no espaço de busca. Também pode se notar que o número de regras mineradas

não aumenta de forma estritamente proporcional ao aumento da população. Nota-se que quando aumenta o tamanho da população (Tpop) de 50 para 200, o que corresponde a um aumento de 400%, o número de regras mineradas aumenta de 10,3 para 11,9, o que corresponde somente a um aumento de somente 15,5%.

Tabela 4: Resultados do algoritmo Genético com Tabu

Tpop	nGen	n°regras	desvio padrão (pontos)	desvio padrão (n° regras)
50	50	10,3	6,56	3,38
50	100	10,3	6,56	3,38
50	200	10,3	6,56	3,38
100	50	9,8	7,89	4,98
100	100	9,8	7,89	4,98
100	200	9,8	7,89	4,98
200	50	11,9	8,46	4,23
200	100	11,9	8,46	4,23
200	200	11,9	8,46	4,23

Analisando por fim os resultados do Algoritmo Genético Completo, que usa *clearing* e tabu (Tabela 5), pode se notar que este minera uma quantidade significativamente maior de regras que todas as versões anteriores, e que este não converge quando se mantém fixo o tamanho da população. Para todos os tamanhos de população (Tpop) experimentados, pode se notar que o número de regras aumenta de forma proporcional ao aumento da quantidade de gerações. Por exemplo, aumentando nGen de 50 para 200, com Tpop = 50, o número de regras mineradas foi de 28,5 para 163,5; com Tpop = 100, o número de regras mineradas foi de 25,2 para 145,2; e com Tpop = 200 o número de regras mineradas foi de 31,2 para 166. Assim, se nota que o uso de *clearing* e tabu na composição do Algoritmo Genético, permitiu a elaboração de uma metodologia bastante promissora pois não converge e com isso, pode-se aumentar a quantidade de regras mineradas aumentando o número de gerações. Além disso, o desvio padrão dos pontos correspondentes as regras mineradas é, em média, maior que nas versões de algoritmos genéticos anteriores. Isso mostra que esta versão permite minerar uma quantidade grande de regras com diversidade.

Tabela 5: Resultados do algoritmo Genético Completo

Tpop	nGen	n°regras	desvio padrão (pontos)	desvio padrão (n° regras)
50	50	28,5	8,26	11,75
50	100	73,5	8,68	25,26
50	200	163,5	9,26	54,66
100	50	25,2	9,08	12,69
100	100	65,2	9,38	32,33
100	200	145,2	9,68	72,20
200	50	31,2	9,46	10,34
200	100	76,2	9,61	25,18
200	200	166	9,68	55,05

6 CONCLUSÃO

Partindo de conceitos existentes na literatura sobre Algoritmos Genéticos (GAs) e extensões, este trabalho fez uma análise de desempenho de quatro diferentes metodologias de GAs para a mineração de regras de associações temporais quantitativas, sendo um GA Simples, um GA com Clearing, um GA com Tabu e um GA com Clearing e Tabu. Vale destacar que os GAs propostos não necessitam de discretização prévia dos dados em intervalos, assim como, da especificação de parâmetros críticos específicos de bases de dados como os limiares de suporte e confiança, necessários em algoritmos de mineração de regras de associação clássicos. Notou-se que o GA Simples tem pior desempenho, não conseguindo minerar uma grande quantidade de regras por convergir prematuramente, o que se deve a sua natureza de otimização unimodal. Já o GA com *clearing* consegue minerar um número maior de regras em relação ao GA simples e não apresenta o problema de convergência com o aumento do número de gerações. O GA com Tabu minera uma quantidade de regras superior ao GA Simples e também ao GA com Clearing, contudo volta a manifestar o problema de convergência em 50 gerações. Analisando por último o GA com Clearing e Tabu (denominado de GA Completo), notou-se que este minera uma quantidade de regras bastante superior aos demais e não apresenta o problema de convergência, o que é um aspecto muito positivo pois tem-se um maior número de regras mineradas por simplesmente aumentar a quantidade de gerações. Também é importante destacar que os GAs desenvolvidos para mineração de regras de associação temporais quantitativas podem ser aplicados em quaisquer área de conhecimento que produzem dados temporais quantitativos e que carecem de análise de correlações (implicações) temporais.

REFERÊNCIAS

Chen, C.-H., Chou, H., Hong, T.-P., and Nojima, Y. (2020). **Cluster-based membership function acquisition approaches for mining fuzzy temporal association rules.** IEEE Access, 8:123996–124006.

Chen, C.-H., Tseng, V. S., Yu, H.-H., and Hong, T.-P. (2013). **Time series pattern disco- very by a pip-based evolutionary approach.** **Soft Computing**, 17(9):1699–1710.

Cot, C., Cacciapaglia, G., and Sannino, F. (2021). **Mining google and apple mobility data: temporal anatomy for covid-19 social distancing**. Scientific Reports, 11(1):4150.

George, Y., Karunasekera, S., Harwood, A., and Lim, K. H. (2021). **Real-time spatio-temporal event detection on geotagged social media**. Journal of Big Data, 8(1):91.

Ghosh, S., Ghosh, S. K., and Buyya, R. (2020). **Mario: A spatio-temporal data mining fra- mework on google cloud to explore mobility dynamics from taxi trajectories**. Journal of Network and Computer Applications, 164:102692.

Jaramillo, I. F., Garza´s, J., and Redchuk, A. (2021). **Numerical association rule mining from a defined schema using the vmo algorithm**. Applied Sciences, 11(13):1–21.

Kurahashi, S. and Terano, T. (2000). **A genetic algorithm with tabu search for multi-modal and multiobjective function optimization**. In Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation, pages 291–298.

Martín, D., Martínez-Ballesteros, M., García-Gil, D., Alcalá-Fdez, J., Herrera, F., and Riquelme-Santos, J. (2018). **Mrqar: A generic mapreduce framework to discover quan- titative association rules in big data problems**. Knowledge-Based Systems, 153:176– 192.

Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., and Riquelme, J. C. (2014a). **Selecting the best measures to discover quantitative association rules**. Neurocomputing, 126:3–14.

Martínez-Ballesteros, M., Nepomuceno-Chamorro, I., and Riquelme, J. C. (2011). **In-ferring gene-gene associations from quantitative association rules**. In 2011 11th International Conference on Intelligent Systems Design and Applications, pages 1241–1246. IEEE.

Martínez-Ballesteros, M., Nepomuceno-Chamorro, I. A., and Riquelme, J. C. (2014b). **Discovering gene association networks by multi-objective evolutionary quantitative association rules**. Journal of Computer and System Sciences, 80(1):118–136.

Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., and Riquelme, J. C. (2016a). **Improving a multi-objective evolutionary algorithm to discover quantitative association rules**. Knowledge and Information Systems, 49(2):481–509.

Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., and Riquelme, J. C. (2016b). **Obtaining optimal quality measures for quantitative association rules**. Neurocomputing, 176:36–47.

Martín, D., Martínez-Ballesteros, M.,García-Gil, D., Alcala´-Fdez, J., Herrera, F., and Riquelme-Santos, J. (2018). **Mrqar: A generic mapreduce framework to discover quan- titative association rules in big data problems**. Knowledge-Based Systems, 153:176– 192.

Matthews, S. G., Gongora, M. A., and Hopgood, A. A. (2013). **Evolutionary algorithms and fuzzy sets for discovering temporal rules**. International journal of applied mathematics and computer science, 23(4):855–868.

Medjadba, Y., Hu, D., Liu, W., and Yu, X. (2020). **Combining graph clustering and quantitative association rules for knowledge discovery in geochemical data problem**. IEEE Access, 8:40453–40473.

Moslehi, F. and Haeri, A. (2020). **A genetic algorithm-based framework for mining quan- titative association rules without specifying minimum support and minimum confi- dence**. Scientia Iranica, 27(3):1316–1332.

Moslehi, F., Haeri, A., and Martínez-Álvarez, F. (2020a). **A novel hybrid ga–pso fra- mework for mining quantitative association rules**. Soft Computing, 24(6):4645–4666.

Owadally, I., Zhou, F., Otunba, R., Lin, J., and Wright, D. (2019). **An agent-based system with temporal data mining for monitoring financial stability on insurance markets**. Expert Systems with Applications, 123:270–282.

Petrowski, A. (1996). **A clearing procedure as a niching method for genetic algorithms.** In Proceedings of IEEE International Conference on Evolutionary Computation, pages 798–803.

Sancho-Asensio, A., Orriols-Puig, A., and Casillas, J. (2016). **Evolving association stre- ams**. Information Sciences, 334:250–272.

Silva, S., Batista, M., and Traina, A. (2015). **Mineração de regras de associação temporais quantitativas por meio de algoritmo genético**. In Anais do XI Simpósio Brasileiro de Sistemas de Informação, pages 195–202. SBC.

Wang, L., Gui, L., and Zhu, H. (2021). **Incremental fuzzy temporal association rule mining using fuzzy grid table**. Applied Intelligence.

Wang, L., Meng, J., Xu, P., and Peng, K. (2018). **Mining temporal association rules with frequent itemsets tree**. Applied Soft Computing, 62:817–829.

Wen, F., Zhang, G., Sun, L., Wang, X., and Xu, X. (2019a). **A hybrid temporal association rules mining method for traffic congestion prediction**. Computers Industrial Engineering, 130:779–787.

Xia, D., Jiang, S., Yang, N., Hu, Y., Li, Y., Li, H., and Wang, L. (2021). **Discovering spa- tiotemporal characteristics of passenger travel with mobile trajectory big data**. Physica A: Statistical Mechanics and its Applications, 578:126056.

Yan, X., Zhang, C., and Zhang, S. (2009). **Genetic algorithm-based strategy for identifying association rules without specifying actual minimum suppor**t. Expert Systems with Applications, 36:3066–3076.

Zhang, S. (2012). **Nearest neighbor selection for iteratively knn imputation**. Journal of Systems and Software, 85(11):2541–2552.