MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2014/11.27.23.46-TDI

PREDIÇÃO DE EVENTOS SEVEROS EM SAÍDAS DE MODELOS METEOROLÓGICOS UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS E METAHEURÍSTICAS PARA REDUÇÃO DE ATRIBUTOS

Alex Sandro Aguiar Pessoa

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Stephan Stephany, aprovada em 14 de novembro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3HFJU3S>

INPE
São José dos Campos
2014

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2014/11.27.23.46-TDI

PREDIÇÃO DE EVENTOS SEVEROS EM SAÍDAS DE MODELOS METEOROLÓGICOS UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS E METAHEURÍSTICAS PARA REDUÇÃO DE ATRIBUTOS

Alex Sandro Aguiar Pessoa

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Stephan Stephany, aprovada em 14 de novembro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3HFJU3S>

INPE
São José dos Campos
2014

Dados Internacionais de Catalogação na Publicação (CIP)

Pessoa, Alex Sandro Aguiar.

P439p Predição de eventos severos em saídas de modelos meteorológicos utilizando a teoria dos conjuntos aproximativos e metaheurísticas para redução de atributos / Alex Sandro Aguiar Pessoa. – São José dos Campos : INPE, 2014.

xx + 126 p. ; (sid.inpe.br/mtc-m21b/2014/11.27.23.46-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Orientador : Dr. Stephan Stephany.

1. Teoria dos conjuntos aproximativos. 2. Metaheurísticas. 3. Eventos severos. I.Título.

CDU 519.22:551.51

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de ***Doutor(a)*** em

Computação Aplicada

Dra. Sandra Aparecida Sandri

Presidente / INPE / SJCampos - SP

Dr. Stephan Stephany

Orientador(a) / INPE / SJCampos - SP

Dr. Luciano Vieira Dutra

Membro da Banca / INPE / SJCampos - SP

Dr. Tércio Ambrizzi

Convidado(a) / IAG/USP / São Paulo - SP

Dr. Carlos Henrique Quartucci Forster

Convidado(a) / ITA / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): ***Alex Sandro Aguiar Pessoa***

São José dos Campos, 14 de Novembro de 2014

RESUMO

A Teoria dos Conjuntos Aproximativos (TCA) é um paradigma para tratamento de informações incertas e imprecisas proposta no início dos anos 80 e vem se difundindo nas últimas duas décadas graças ao aumento das capacidades de processamento e armazenamento de dados. Um ponto central na TCA é a obtenção de conjuntos reduzidos de atributos conhecidos como *reduções*, as quais reduzem a dimensionalidade da classificação. Entretanto, a obtenção de reduções a partir do conjunto completo de atributos possui alta complexidade computacional, recorrendo-se então ao uso de metaheurísticas. Nesta tese, objetiva-se identificar padrões associados à ocorrência de eventos convectivos severos em saídas de modelos numéricos de previsão de tempo utilizando-se TCA. Estes padrões são constituídos por um conjunto selecionado de variáveis meteorológicas e são encontrados a partir de um conjunto de eventos convectivos conhecidos, os quais foram identificados por meio da densidade de ocorrência de descargas elétricas nuvem-solo. A aplicação de metaheurísticas específicas otimiza a identificação desses padrões no escopo da TCA e permite gerar classificadores que possam detectar a possível ocorrência de eventos convectivos em previsões meteorológicas. Isso auxiliaria a previsão operacional de tempo, dada a deficiência que os modelos meteorológicos tem em simular a gênese e evolução de eventos convectivos devida a limitações de resolução espacial e à necessidade de se aprimorar a microfísica correspondente nesses modelos.

PREDICTION OF SEVERE CONVECTIVE EVENTS FROM WEATHER MODEL OUTPUT USING THE ROUGH SET THEORY AND METAHEURISTICS FOR ATTRIBUTE REDUCTION

ABSTRACT

The Rough Set Theory (RST) is a standard proposed to deal with uncertain, incomplete or vague information that was proposed in the early 80s. The use of RST has been spreading over the last two decades thanks to increase of data processing and storage capabilities. A fundamental point of RST is the calculation of reduced sets of attributes known as reducts, which allow to reduce the classification dimensionality. However, the calculation of reducts from the complete set of attributes presents high algorithmic complexity demanding the use of metaheuristics. The aim of this thesis is to identify patterns associated to the occurrence of severe convective events from the output of weather forecast numerical models using RST. These patterns are composed of a selected set of meteorological variables and are found using a set of known convective events, which were identified using the density of occurrence of cloud-to-ground electrical discharges. The application of specific metaheuristics optimizes the identification of such patterns in the scope of RST, and allows to derive classifiers able to detect the possible occurrence of convective events in weather forecasts. This approach would help the operational weather forecasting considering that meteorological models have poor performance to simulate the genesis and evolution of convective events due to spatial resolution limitations and to the need of improving the corresponding microphysics in such models.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE SIGLAS E ABREVIATURAS

BrasilDat – Sistema Brasileiro de Detecção de Descargas Atmosféricas
CPTEC – Centro de Previsão de Tempo e Estudos Climáticos
CRED – Centre for Research on the Epidemiology of Disasters
DCS – *Decrescent Cardinality Search*
EDDA – Estimação de Densidade de Descargas Elétricas Atmosféricas
EM-DAT – Emergency Events Database
GOES – *Geostationary Operational Environmental Satellite*
ILS – *Iterated Local Search*
INPE – Instituto Nacional de Pesquisas Espaciais
KDD – *Knowledge Discovery in Databases*
LAC – Laboratório Associado de Computação e Matemática e Aplicada
MD – Mineração de Dados
NCAR – *National Center for Atmospheric Research*
NCEP – *National Centers for Environmental Prediction*
NDVI – *Normalized Difference Vegetation Index*
NOAA – *National Oceanic and Atmospheric Administration*
NS – nuvem-solo
RINDAT – Rede Integrada Nacional de Detecção de Descargas Atmosféricas
Rosetta – *Rough Set Toolkit for Analysis of Data*
SCM – Sistemas Convectivos de Mesoescala
SLS – *Standard Local Search*
TCA – Teoria dos Conjuntos Aproximativos
TRMM – *Tropical Rainfall Measuring Mission*
VND – *Variable Neighborhood Descent*
VNS – *Variable Neighborhood Search*

SUMÁRIO

Pág.

1 INTRODUÇÃO

Esta tese trata da predição de eventos convectivos severos usando programas baseados na Teoria dos Conjuntos Aproximativos, um paradigma que é aplicado em aprendizado de máquina e em mineração de dados. Objetiva-se identificar padrões associados a eventos convectivos em saídas de modelos numéricos de previsão de tempo. Para isso é assumida como hipótese que áreas com alta taxa de ocorrência de descargas elétricas nuvem-solo (NS), possuam forte atividade convectiva, sujeitas consequentemente a tempestades com forte precipitação. Conforme será exposto adiante, os modelos numéricos de previsão de tempo não conseguem prever com precisão a ocorrência de eventos convectivos.

1.1 Predição de eventos convectivos severos

Eventos convectivos severos são fenômenos meteorológicos que tem o potencial de causar danos de ordem socioeconômica, devido à intensidade e/ou duração dos ventos e chuvas decorrentes, bem como das inundações e deslizamentos de terra que causam. No Brasil, os desastres ambientais registrados na base de dados do EM-DAT: *International Disaster Database* (GUHA-SAPIR et al., 2014), para o período de 1995 a 2014, indicam uma forte predominância de fenômenos meteorológicos entre os dez maiores eventos (número de mortos e de afetados), conforme mostram as Tabelas 1.1(a) e 1.1(b). Esses dados indicam a necessidade de melhores e mais acuradas previsões desses eventos para que se possa mitigar seus efeitos.

Tabela 1.1. Os dez maiores desastres ambientais ocorridos no Brasil de 1995 a 2014, ordenados por: (a) número de mortes e (b) número de afetados.

Desastre	Data	# Mortos
Enchente	Jan/2011	900
Enchente	Abr/2010	256
Enchente	Dez/2003	161
Enchente	Nov/2008	151
Epidêmico	Mar/2008	123
Deslizamento	Fev/1996	96
Enchente	Dez/1995	92
Deslizamento	Mai/1995	86
Deslizamento	Dez/2002	74
Enchente	Dez/2009	74

(a)

Desastre	Data	#Afetados
Seca	Abr/1998	10.000.000
Enchente	Nov/2008	1.500.015
Enchente	Abr/2009	1.150.900
Seca	Jun/2001	1.000.000
Seca	Out/2007	1.000.000
Enchente	Set/2011	1.000.000
Epidêmico	Jan/2011	942.153
Enchente	Nov/2009	680.000
Epidêmico	Jan/2002	317.730
Epidêmico	Mai/1998	213.932

(b)

Fonte: Guha-Sapir et al. (2014).

É de suma importância que a previsão de tais eventos seja realizada em tempo hábil para que se possa tomar as devidas medidas e consequentemente mitigar seus efeitos danosos. Em Meteorologia, dados de diversas fontes, ditos multivariados, tais como dados de satélites, dados observados (gerados por estações de medições em superfície, boias oceânicas, radiossondagens), modelos numéricos de previsão do tempo, dados de ocorrência de descargas elétricas, radar, dentre outros, são usados para realizações de previsões do tempo.

Entretanto, a previsão numérica de eventos severos é prejudicada devido à fato de que no Brasil a malha de dados observados é insuficiente e isso acarreta em erros/imprecisões na previsão, uma vez que os modelos são alimentados com estes dados. A previsão desses eventos também é dificultada pela necessidade de análise deste grande volume de dados na previsão do tempo operacional.

Assim é necessário o desenvolvimento de ferramentas de auxílio à previsão do tempo, para a identificação de eventos severos. No Brasil, alguns estudos foram conduzidos para predição de atividade convectiva e eventos severos. Por exemplo, em Bourscheidt et al. (2002) é sugerida uma abordagem de rastreio de tempestades de descargas elétricas atmosféricas, da rede BrasilDat (Sistema Brasileiro de Detecção de Descargas atmosféricas), usando as

informações provenientes da densidade de descargas. O sistema desenvolvido usa uma função de estimação de núcleo e algoritmos de agrupamentos, tais como DBSCAN e K-Means. Em Dolif e Nobre (2012) foi usada uma rede neural para previsão de ocorrência de precipitação intensa, na cidade do Rio de Janeiro, com base em dados do modelo numérico ETA40 e dados de precipitação de estações meteorológicas. Os resultados obtidos mostraram ser possível de prever 55% dos eventos ligados a chuva intensa, utilizando uma combinação de umidade relativa à 900 hPa e vento meridional à 10m. Mais recentemente, em Garcia et al. (2013) foi desenvolvida uma metodologia para estimar acumulados de precipitação convectiva a partir de dados de descargas NS, usando uma janela deslizante temporal. Os dados utilizados foram de radar de banda-S e dados de descargas elétricas atmosféricas fornecidos pelo RINDAT e processados pelo software EDDA (Estimação de Densidade de Descargas Elétricas Atmosféricas).

Em particular, houve especificamente dois projetos de pesquisa iniciados em 2007 e ligados à presente tese, desenvolvidos pelo Laboratório Associado de Computação e Matemática e Aplicada (LAC/INPE) em colaboração com o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC/INPE). O primeiro foi um projeto FINEP intitulado “ADAPT – Tempestades: desenvolvimento de um sistema dinamicamente adaptativo para produção de alertas para região Sul/Sudeste“ (desenvolvido pelo INPE e por diversas outras instituições), sendo que se menciona aqui a sua meta n^0 2, “Mineração de dados para identificação de condições favoráveis à gênese e evolução de tempestades”. O segundo projeto, do edital Universal do CNPq, processo 479510/2006-7, intitulava-se “Cb-mining - Mineração de dados Associados a Sistemas Convectivos”. Ambos os projetos objetivavam construir classificadores por meio de técnicas de Mineração de Dados (MD) para predição de atividade convectiva, especialmente a severa.

1.2 Eventos convectivos e descargas elétricas atmosféricas

Diversos trabalhos estudam a correlação entre descargas elétricas e atividade convectiva. Em Petersen et al. (1996) foi investigada a ligação entre descargas NS e a convecção oceânica. As análises mostraram que em muitos casos, os picos de frequência das descargas NS aconteciam com forte precipitação. Carey e Rutledge (2000) estudaram o processo de eletrificação das nuvens e a possível relação entre precipitação e descargas NS. Foram utilizados dados de radar polarimétrico na banda C e dados de descargas NS. Concluiu-se que durante a fase madura que as descargas e o campo elétrico da superfície estavam fortemente correlacionados com a precipitação. Zhou et al. (2002) correlacionaram a ocorrência de descargas NS com precipitação convectiva. Kuligowski e Scofield (2005) utilizaram dados de descargas elétricas combinadas com imagens de satélite (*Geostationary Operational Environmental Satellite* ou GOES) para análise e predição da evolução de Sistemas Convectivos de Mesoescala (SCM), de forma a correlacionar a ocorrência de descargas elétricas com a atividade convectiva. No trabalho de Machado et al. (2009) foram estabelecidas relações entre descargas NS e a diferença entre a temperatura de brilho dos canais de vapor de água e o infravermelho (WV-IR). Foram utilizados dados de descargas fornecidos pelo RINDAT (Rede Integrada Nacional de Detecção de Descargas Atmosféricas) e imagens do satélite GOES-10.

Mattos e Machado (2011) analisaram o ciclo de vida de SCMs sobre o estado de São Paulo, usando imagens do canal infravermelho do satélite GOES-10, do canal de micro-ondas dos satélites TRMM (*Tropical Rainfall Measuring Mission*) e NOAA-18 (*National Oceanic and Atmospheric Administration*), além dos dados de descargas elétricas da rede BrasilDat. Constataram, com base em 720 SCMs, que as tempestades elétricas, na média, têm uma duração maior e uma área maior que tempestades comuns. E também constataram que o máximo de densidade de descargas ocorre no início do ciclo do SCM,

enquanto que a taxa de ocorrência de descargas alcança um máximo durante a fase de crescimento, perto da maturação do SCM.

Oliveira e Mattos (2011), estudaram a correlação entre descargas e sistemas convectivos na cidade de São Paulo considerando uma área de 1°x1°, concluindo que durante o verão há descargas NS menos intensas, contrastando com sua maior frequência e associadas a um maior volume de chuva. Por outro lado, no inverno, as descargas NS são mais intensas, porém ocorrem com menor frequência e associadas a um menor volume de chuva. A baixa correlação entre precipitação e descargas é atribuída a uma defasagem temporal entre esses fenômenos. Beneti et al. (2012) demonstraram haver correlação entre dados de radar correspondentes a precipitação e dados de descargas NS para o ciclo diurno de SCMs no Sudeste brasileiro.

1.3 Mineração de dados em Meteorologia

A Mineração de Dados (MD) é parte de um processo maior denominado Descoberta de Conhecimento em Bases de Dados, ou *Knowledge Discovery in Databases* (KDD), cujo objetivo é transformar dados em conhecimento. Esse processo surgiu devido ao grande volume e diversidade de dados produzidos diariamente, inclusive em Meteorologia, inviabilizando sua análise ou interpretação por parte dos especialistas. O processo de KDD possui etapas correspondentes (i) à identificação do problema, (ii) ao pré-processamento dos dados (inclui seleção e transformação), a Mineração de Dados para obtenção de padrões e o pós-processamento desses dados e padrões, que inclui sua visualização e interpretação, para posterior utilização do conhecimento resultante. Estas etapas são realizadas iterativamente, conforme ilustrado na Figura 1.1.

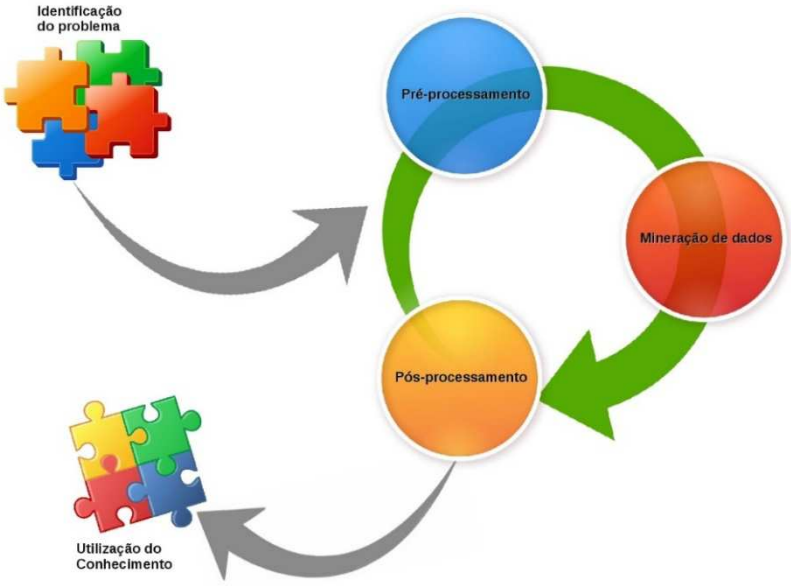

Figura 1.1 – Processo de Descoberta de Conhecimento em Banco de Dados.
Fonte: Rezende (2005)

A MD constitui o cerne do processo de KDD, buscando nos dados padrões desconhecidos ou ocultos e potencialmente úteis, por meio de técnicas computacionais típicas das áreas de estatística, reconhecimento de padrões e inteligência artificial (FAYYAD et al., 1996). Nesta última área, podem-se citar técnicas de aprendizado de máquina. Assim, os padrões são tipicamente obtidos pela aplicação de um algoritmo específico de MD, com uma funcionalidade tal como agrupamento, classificação ou definição de regras de associação. Em particular, este trabalho aborda a classificação, que visa atribuir uma classe a cada instância da base de dados, rotulando-a. A classificação pode ser associada à predição, como será visto adiante.

Várias aplicações de MD foram desenvolvidas em Meteorologia, geralmente relativas à construção de classificadores usando diversas técnicas de MD, conforme exposto a seguir.

Em Peters et al. (2003) foram utilizados dados de radar para classificar células de tempestade relacionadas à ocorrência de granizo, chuva, tornado e vento, para o período de 1997-1999, no Canadá. Assim, foi montada uma base de

dados com 22 variáveis ou atributos condicionais e o atributo de decisão, a qual foi analisada por uma abordagem de TCA. No trabalho de Pessoa et al. (2006) foram usadas redes neurais, do tipo *Perceptron Multi-Camadas*, com o propósito de realizar previsão climática para o Brasil. Os dados utilizados foram de reanálise do NCEP/NCAR (*National Centers for Environmental Prediction*/*National Center for Atmospheric Research*). A TCA foi usada para reduzir os dados de entrada e melhorar o desempenho das redes neurais.

Cano et al. (2004) aplicaram redes Bayesianas em dados diários de chuva e velocidade do vento máxima, gerados por 100 estações meteorológicas na península Ibérica, além de dados de modelo numérico, com o objetivo de gerar padrões que representassem a dependência espacial entre as variáveis observadas e aquelas do modelo. Hruschka Jr et al. (2006), usou redes Bayesianas para construção de um classificador, utilizando dados de estações meteorológicas, para identificar neblina.

No estudo realizado por Sharma (2006) foi estabelecido um relacionamento entre precipitação e o índice NDVI (*Normalized Difference Vegetation Index*) para monitoramento de secas. Dados de precipitação de 1970-2004 foram usados para criar o índice SPI (*Standardized Precipitation Index*) e o índice NDVI da NOAA-AVHRR (*National Oceanic and Atmospheric Administration* - *Advanced Very High Resolution Radiometer*) foi usado para calcular o VCI (*Vegetation Condition Index*) para o período de 1981-2003. As técnicas empregadas nesse trabalho foram regras de associação e Análise de Componentes Independentes (*Independent Component Analysis*, ICA). Em Vestergraard (2011) foi desenvolvido um sistema de *nowcasting*, utilizando dados de radar e de satélite, para previsão de eventos relacionados à ocorrência de precipitação forte na Dinamarca. As técnicas usadas foram Análise de Componentes Principais (*Principal Component Analysis* - PCA), Fator de Máxima Autocorrelação (*Maximum Autocorrelation Factor* - MAF) e Análise de Correlação Canônica (*Canonical Correlation Analysis* - CCA).

Em Cortez e Morais (2007), dados meteorológicos foram usados para predição de incêndios florestais a partir de dados de estações meteorológicas tais como precipitação, temperatura, umidade relativa e direção e velocidade do vento e empregando técnicas de Máquina de Vetor de Suporte (SVM) e de Floresta Aleatória (de árvores de decisão). A previsão de geada ou de deficiência hídrica foi tema central do trabalho de Bucene (2008). Foram usados dados climáticos de temperatura máxima e mínima, de precipitação e também dados específicos de ocorrências de El Niño e La Niña. Uma árvore de decisão permitiu criar o modelo que estabelece a correlação das variáveis de entrada e as geadas ou a deficiência hídrica.

No trabalho de Little et al. (2008) foi utilizado o método de agrupamento Bayesiano para identificar riscos potenciais de enchente, com base em dados de precipitação severa diária no Reino Unido. Kanth et al. (2014) realizaram uma análise do clima na Índia usando o algoritmo de agrupamento K-Means e a árvore de decisão J48. Este estudo indicou também um aumento significativo da temperatura máxima em 112 anos de dados analisados.

Jan et al. (2009) aplicaram o algoritmo de *k*-NN (*k-Nearest Neighbor*) para predição do clima sazonal e interanual. Foram usados dados de precipitação, velocidade do vento, temperatura, temperatura do ponto de orvalho, etc. Kohail e El-Halees (2011) utilizaram técnicas de MD como Naive Bayes, k-NN, árvore de decisão e redes neurais para estabelecer relações entre dados históricos de uma estação meteorológica como médias diárias de umidade relativa, velocidade de vento, direção do vento e rajadas de vento com dados de chuva na faixa de Gaza para o período de 1977-1985. Tsagalidis e Evangelidis (2011) criaram um classificador com as técnicas de árvore de decisão, k-NN e redes neurais, para predizer a ocorrência de precipitação em Micra, Grécia. Foram usadas como variáveis preditoras dados da reanálise ERA-40 fornecido pelo ECMWF (*European Center for Medium-Range Weather Forecast*) e dados de precipitação de uma estação meteorológica em Micra. Os autores destacaram o desbalanceamento da classes comum nesse problema, onde somente 16%

dos casos eram referentes à ocorrência de precipitação. Ping et al. (2012) fizeram um estudo dos agrupamentos dos dados de eventos extremos de tempo e clima ocorridos nos últimos 50 anos na China (1960-2006) utilizando o algoritmo de agrupamento k-NN. Dados de 740 estações meteorológicas distribuídas por toda China compreenderam máximas e mínimas diárias de temperatura e precipitação diária. O estudo mostrou uma diminuição da tendência de ocorrência de eventos extremos de baixas temperaturas e um aumento na tendência de eventos extremos ligados a altas temperaturas nos últimos 50 anos. Mostrou também uma diminuição de chuva torrenciais no mesmo período.

Em Sencan et al. (2011) foram usadas redes dinâmicas para identificar a gênese de furaçõesna região da África Ocidental e prever sua evolução no Atlântico Norte. Essas redes são grafos gerados de acordo com a trajetória efetuada pelo furação, sendo posteriormente correlacionadas com dados meteorológicos. Foi possível prever com 90% de precisão o rastreio de furações com 10-15 dias de antecedência. Os dados utilizados foram referentes ao rastreio de furações efetuado pelo NCDC (*National Climatic Data Center*), dados de reanálise da pressão ao nível do mar e temperatura da superfície do mar, fornecidos pelo NCEP/NCAR. Os algoritmos de DM empregados foram a árvore de decisão C4.5 e SVM.

O estudo realizado por Olaiya e Adeyemo (2012) propõe uma metodologia para previsão do tempo e estudo de mudanças climáticas. Os dados sinóticos são do aeroporto de Ibadan, fornecidos pela Agência Meteorológica da Nigéria, para o período de janeiro de 2000 até dezembro de 2009 (120 meses), constituídos de velocidade do vento, evaporação, formação de nuvens, radiação solar, duração do dia, temperatura máxima e mínima e precipitação. Foram usadas redes neurais e árvore de decisão para construir modelos destinados à predição da temperatura máxima, precipitação, evaporação e velocidade do vento. No trabalho de Babic et al. (2012) foram criados modelos parametrizados e métodos para detecção e predição de padrões associados à

ocorrência de neblina e de cobertura de nuvens baixas, utilizando dados de METAR, SYNOP, radar, satélite e modelos numéricos, com algoritmos de redes neurais e de árvores de decisão.

Dutta e Tahbilder (2014) propuseram a previsão de precipitação utilizando regressão linear múltipla, para dados no período de 2007-2013, fornecidos pelo Departamento Meteorológico de Guwahati, Índia. Esses dados incluem temperaturas máxima e mínima, umidade relativa, velocidade do vento, pressão e precipitação.

Nos dois projetos citados anteriormente (ADAPT e Cb-Mining), algumas abordagens foram adotadas (PESSOA et al., 2012), como a baseada na Teoria dos Conjuntos Aproximativos (TCA) (PESSOA; STEPHANY, 2012a, 2012b) que constitui o escopo desta tese. Outras incluíram similaridade de vetores (LIMA; STEPHANY, 2012, 2013b; LIMA et al., 2010), redes neurais artificiais (LIMA; STEPHANY, 2013a) e árvores de decisão (STRAUSS, 2013; STRAUSS et al., 2012). O objetivo destes trabalhos eram de construir classificadores capazes de identificar, nas saídas do modelo numérico de previsão de tempo, padrões de eventos severos. Foram utilizados dados do modelo numérico de previsão de tempo ETA (MESINGER et al., 1988) com resolução espacial de 20 km e, posteriormente, de 5 km, utilizados operacionalmente no CPTEC/INPE e dados de densidade de ocorrência de descargas NS, gerada pelo software EDDA (STRAUSS et al., 2013; STRAUSS et al., 2010). O EDDA foi desenvolvido no escopo desses projetos e sendo avaliado operacionalmente desde o final de 2012 no Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN) do Ministério da Ciência, Tecnologia e Inovação (MCTI). Os dados de descargas foram gerados pela rede de detecção RINDAT.

1.4 A Teoria dos conjuntos aproximativos

A Teoria dos Conjuntos Aproximativos é um paradigma amplamente usado em MD. A TCA foi proposta no início dos anos 80 (PAWLAK, 1982) para o para tratamento de informações incertas e imprecisas. Seu uso vem se difundindo nas últimas duas décadas graças ao aumento das capacidades de processamento e armazenamento de dados.

Um ponto central na TCA é a obtenção de conjuntos reduzidos de atributos conhecidos como reduções, as quais reduzem a dimensionalidade da classificação, mas preservam o desempenho de classificação. Entretanto, a obtenção de reduções a partir do conjunto completo de atributos possui alta complexidade computacional devido ao número de reduções possíveis combinatoriamente, recorrendo-se então ao uso de metaheurísticas para buscar soluções sub-ótimas no espaço de busca das reduções. A aplicação de metaheurísticas específicas otimiza a identificação desses padrões no escopo da TCA e permite gerar classificadores eficientes.

A TCA foi empregada neste estudo em conjunto com uma técnica chamada de Particionamento Aleatório do Conjunto de Treinamento (PACT), que consiste basicamente em subdividir aleatoriamente o conjunto de treinamento em partições menores, que são mais facilmente tratáveis. Cada partição permite calcular uma redução específica, obtendo-se um conjunto de reduções final, o qual é posteriormente utilizado para induzir um conjunto de regras de associação que corresponde ao classificador TCA e que vai identificar os padrões de ocorrência de atividade convectiva.

Nesta tese a TCA é utilizada para a realização da mineração de dados e consequente construção de classificadores baseados em regras. Os dados utilizados são dados do modelo numérico de previsão do tempo ETA, resolução de 20Km (ETA20), usados como atributos condicionais, e dados de densidade de descargas NS, gerados pelo EDDA, usando dados do RINDAT, usados como atributo de decisão.

Os dados correspondentes às saídas do modelo no formato binário são transformados para o formato texto e dispostos numa forma tabular em que cada linha é um registro correspondente num instante de tempo e num ponto da grade do modelo, sendo cada coluna correspondente a uma das 58 (no caso da abordagem de TCA) variáveis atmosféricas (chamados de atributos condicionais), além de uma coluna extra que contém a densidade de descargas NS (atributo de decisão). Os atributos condicionais são expressos numericamente, enquanto que o atributo de decisão é expresso de forma categórica, tendo sido definidas 3 classes de densidade de descargas, as quais foram assumidas como representativas de atividade convectiva ausente/fraca, moderada e forte.

Para ilustrar as dificuldades da abordagem de TCA proposta, que utiliza 58 atributos condicionais, a Tabela 1.2 mostra o número possível de estados diferentes que um único registro de uma base de dados qualquer pode assumir, considerando valores discretos dos atributos. No caso deste trabalho, os 58 atributos condicionais são também discretizados em 3 intervalos e portanto cada registro pode assumir aproximadamente $5x10^{27}$ valores diferentes.

Tabela 1.2. Número de valores discretos possíveis de um registro de uma base de dados.

Número de valores discretos do atributo	**Número de Atributos Condicionais**		
	10	**20**	**30**
2	1024	1.048.576	$\sim10^{9}$
3	59.049	$\sim10^{9}$	$\sim10^{12}$
4	1.048.576	$\sim10^{12}$	$\sim10^{18}$
5	9.765.625	$\sim10^{14}$	$\sim10^{21}$

Fonte: Düntsch e Gediga (2000)

1.5 Contribuições desta tese

No tocante às contribuições desta tese, é apresentada uma abordagem inovadora para predição da ocorrência de eventos convectivos severos utilizando um classificador derivado da TCA. Padrões compostos por um conjunto de variáveis meteorológicas do modelo de previsão de tempo ETA são analisados pelo classificador indicando a ocorrência ou não de atividade convectiva e se sua provável intensidade (fraca, média ou forte). O classificador foi submetido a treinamento com casos de atividade convectiva pregressos que foram identificados pela densidade de ocorrência de descargas NS.

Esta abordagem começou a ser desenvolvida no escopo de projetos para suporte à previsão meteorológica, especificamente na predição de ocorrência de atividade convectiva, por meio de ferramentas auxiliares que visam suprir deficiências dos modelos de previsão de tempo. Outra contribuição é o aprimoramento do cálculo de reduções em TCA pelo uso inédito de quatro metaheurísticas: a Busca e Descida em Vizinhança Variável (VNS e VNS), Busca em Busca Local Iterativa (ILS) e a Busca em Cardinalidade Decrescente (DCS). Em particular, a DCS é uma metaheurística nova, desenvolvida na pesquisa associada a esta tese, sendo derivada do VNS e tendo como características implementação simples, uso de poucos parâmetros e convergência rápida.

Adicionalmente, também é usada de forma inovadora uma função alternativa de dependência funcional para avaliação das soluções candidatas geradas pelas metaheurísticas, reduzindo drasticamente o tempo de processamento correspondente. Outra contribuição são as implementações computacionais desenvolvidas utilizando softwares gratuitos como MySQL (banco de dados), *shell scripts* e a linguagem *Perl*, alternativamente ao uso do software Rosetta (*Rough Set Toolkit for Analysis of Data*), específico de TCA e também gratuito.

Esta tese tem a seguinte estrutura: o Capitulo 2 trata dos principais conceitos relacionados à Teoria dos Conjuntos Aproximativos. No Capitulo 3 são abordadas ou métodos que usam metaheurísticas para o problema de reduções na TCA. O Capitulo 4 descreve os dados meteorológicos usados nos experimentos desta pesquisa. Os resultados e discussões estão no Capítulo 5. Por fim, o Capítulo 6 apresenta as conclusões e comentários finais.

2 Teoria dos Conjuntos Aproximativos

No mundo real as informações são frequentemente incertas, imprecisas ou incompletas, devido a dificuldades de relatar fenômenos observáveis, fazer medições ou mesmo devido a limitações relativas à resolução espacial e temporal. Muitas teorias foram desenvolvidas para tratar tais imperfeições contidas em base de dados, tais como a teoria dos conjuntos difusos (ZADEH, 1965), teoria da evidência (DEMPSTER, 1967; SHAFER, 1976) e a teoria das possibilidades (ZADEH, 1978). No início da década de 80 surgiu a TCA, ou em inglês, *Rough Sets Theory* (PAWLAK, 1982), caracterizada pelo bom formalismo matemático e simplicidade de aplicação, tornando-a um paradigma interessante no tratamento de informações incertas. Porém somente na década de 90 essa teoria ficou mais difundida, pois a disponibilidade de computadores com mais memória e maior capacidade de processamento viabilizou seu uso em base de dados de maior complexidade. A TCA é uma extensão da teoria dos conjuntos, que enfoca o tratamento de imperfeição da informação por meio de uma relação de indiscernibilidade e que avalia se os elementos de um conjunto são indiscerníveis, ou seja, se possuem as mesmas propriedades. Alguns autores apontam como a principal vantagem da TCA a não necessidade de utilização de informações adicionais, tais como distribuição de probabilidade, grau de pertinência, possibilidade ou atribuição de crença (NICOLETTI; UCHÔA, 1997).

Obviamente, à medida que uma dada metodologia se difunde, mais evidentes se tornam seus pontos fracos. Talvez o ponto fraco mais relevante, da TCA, seja o cálculo das reduções. Reduções são mecanismos de tratamento de dados da TCA, que tentam preservar a mesma quantidade de informação contida no conjunto de dados original, através da eliminação de algumas variáveis (ou atributos condicionais, como será visto adiante). Esse processo de reduzir o número de variáveis/atributos também é conhecido como seleção ou redução de atributos, constituindo uma área de pesquisa bem conhecida e que extrapola o escopo da TCA.

Conforme mencionado no capítulo anterior, o cálculo de reduções possui alta complexidade computacional devido ao grande número de combinações possíveis. Assim, considerando-se o espaço de busca correspondente às reduções, recorre-se ao uso de metaheurísticas para buscar soluções sub-ótimas. Um aspecto relevante desta tese é a implementação do cálculo das reduções em TCA utilizando software desenvolvido pelo autor, sendo que se desconhece outra implementação no país, uma vez que usualmente se recorre a softwares de TCA gratuitos que vêm prontos, porém com limitações em termos da complexidade da base de dados ou das metaheurísticas que podem ser utilizadas.

Nas seções seguintes serão tratadas as definições básicas da TCA, seguidas da definição e cálculo das reduções e, finalmente, o uso de metaheurísticas para esse cálculo.

2.1 Definições básicas

Na TCA, os dados são dispostos de forma tabular, constituindo o chamado *sistema de informação*. Neste tipo de sistema cada linha representa um objeto (um caso, um evento ou paciente) e cada coluna, por sua vez, representa um atributo (uma característica, uma observação ou propriedade). Formalmente, um sistema de informação é um par $S = (U; A)$, onde U é um conjunto finito não-vazio de elementos, chamado de universo e A é um conjunto finito não-vazio de atributos tal que: $a{:}U \rightarrow V_a$ para todo $a \in A$. O conjunto V_a é chamado de *conjuntos de valores* do atributo a. Os elementos do conjunto A são chamados *atributos condicionais* ou simplesmente condições (KOMOROWSKI et al., 1999). Define-se cardinalidade ou dimensionalidade do sistema de informação como sendo o seu número de atributos condicionais, expresso por $|A|$. Abaixo, na Tabela 2.1, apresenta-se um exemplo de sistema de informação definido pelo seu conjunto universo $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, pelos seus

atributos condicionais $A = \{a_1, a_2, a_3\}$ e pelos conjuntos de valores possíveis dos atributos $V_{a1} = \{0,1,2\}$ e $V_{a2} = V_{a3} = \{0,1\}$.

Tabela 2.1. Sistema de Informação.

	a_1	a_2	a_3
x_1	2	0	0
x_2	0	1	0
x_3	1	0	1
x_4	2	0	0
x_5	2	0	1
x_6	0	1	1

Um caso particular do sistema de informação é o sistema de decisão, formado pela união dos atributos condicionais com o chamado atributo de decisão $\{d\}$ que rotula os elementos da base de dados nas classes pré-definidas (KOMOROWSKI et al., 1999). Assim, o sistema de decisão é dado por: $S_d = (U, A \bullet \{d\})$. Na Tabela 2.2 é apresentado um sistema de decisão, que também será usado ao longo deste capítulo para exemplificar os conceitos apresentados da TCA. Essa tabela é obviamente similar à Tabela 2.1, exceto pelo atributo de decisão d, com conjunto de valores $V_d = \{0,1\}$.

Tabela 2.2. Sistema de Decisão.

	a_1	a_2	a_3	d
x_1	2	0	0	0
x_2	0	1	0	1
x_3	1	0	1	0
x_4	2	0	0	1
x_5	2	0	1	1
x_6	0	1	1	1

O atributo de decisão d determina uma partição $CLASSE_A(d) = \{X_1, ..., X_{r(d)}\}$ no universo U, onde $X_k = \{x \in U \mid d(x) = v_k)$ para $1 \leq k \leq r(d)$ e $r(d) = |V_d|$. No caso específico da Tabela 2.2, existem duas classes, conforme mostra o conjunto V_d e a partição é formada por $X_1=\{x_1, x_3\}$ e $X_2=\{x_2, x_4, x_5, x_6\}$, onde X_1 é formado por elementos que pertencem à classe cujo atributo d possui valor "0" e X_2 é formada por elementos com valores de d iguais a "1".

Por questão de conveniência, um sistema de decisão será citado daqui em diante simplesmente como sendo um sistema de informação, uma vez que se trata de um caso particular deste. Outra questão importante é referente à "granularidade" dos valores dos atributos. A TCA utiliza dados discretos e portanto, dados contínuos devem ser discretizados. Em Komorowski et al. (1999) são citadas extensões que possibilitam o uso de dados contínuos, mas não é o caso deste trabalho.

Nas seções seguintes serão abordados a relação de indiscernibilidade e as aproximações de conjuntos. Esses definições são importantes, pois são essenciais para a compreensão dos conjuntos aproximativos, servindo de base para o tratamento e análise de informações na TCA, permitindo, por exemplo, o cálculo de reduções.

2.1.1 Relação de indiscernibilidade

Dado $S = (U; A)$ como sistema de informação, então para qualquer subconjunto de atributos pode-se definir uma relação de equivalência $IND_A(B)$, ou, omitindo o subscrito A, $IND(B)$, com $B \bullet A$:

$$IND(B) = \{(x, x') \in U \mid \forall a \in B, a(x) = a(x')\} \tag{2.1}$$

A expressão acima é chamada de relação de *B-indiscernibilidade*, que particiona A em classes de equivalência, sendo cada classe denotada por $[x]_B$ (KOMOROWSKI et al., 1999). Essas classes de equivalência permitem uma divisão da base de dados em partições definidas de acordo com a certeza dos elementos pertencerem a uma dada classe. Dessa forma um elemento qualquer pertencente ao conjunto U, pode ser classificado como: (i) pertencente somente à uma classe ou (ii) pertencente a mais de uma classe. Se existem elementos que não podem ser definidos como pertencentes a uma

só classe, o conjunto *U* é dito *aproximativo* (PAWLAK, 1982). Caso contrário, diz-se que o conjunto é puro (*crisp*).

Tomando como exemplo o sistema de decisão da Tabela 2.2, a partição induzida pela relação de indiscernibilidade produz as seguintes classes de equivalência:

1) $IND(a_1) = IND\{a_1, a_2\} = \{[x_1, x_4, x_5], [x_2, x_6], [x_3]\}$
2) $IND\{a_2\} = \{[x_1, x_3, x_4, x_5], [x_2, x_6]\}$
3) $IND\{a_3\} = \{[x_1, x_2, x_4], [x_3, x_5, x_6]\}$
4) $IND\{a_2,a_3\} = \{[x_1, x_4], [x_2], [x_3, x_5], [x_6]\}$
5) $IND\{a_1,a_3\} = IND\{A\} = \{[x_1, x_4], [x_2], [x_3], [x_5], [x_6]\}$

2.1.2 Aproximações dos conjuntos

No caso de conjuntos puros, as classes são bem definidas. Na TCA existe o conceito de conjuntos aproximativos, definidos com base nas chamadas aproximações, derivadas a partir das classes e da relação de indiscernibilidade. Assim, pode-se construir uma aproximação de $X \bullet U$, usando somente as informações contidas no conjunto de atributos $B \bullet A$, construindo as aproximações *B*-inferiores e *B*-superiores de *X*, denotados respectivamente por $\underline{B}X$ e $\overline{B}X$, onde:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \tag{2.2}$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \varnothing\} \tag{2.3}$$

Os elementos em $\underline{B}X$ podem ser certamente classificados como membros de *X* na base de conhecimento (conjunto de atributos) *B*, enquanto os elementos em $\overline{B}X$ podem somente serem classificados como possíveis membros de *X* na base de conhecimento *B*.

O conjunto *F,* chamado de *região de fronteira* de *X*, é formado por elementos que não podem ser classificados como pertencentes a *X* na base de conhecimento *B* com absoluta certeza. O conjunto *E* é chamado de *região externa* de *X*, e é formado por elementos que podem ser classificados como não pertencentes a *X*. Assim, um conjunto é dito *aproximativo* se a região de fronteira não é vazia, ou caso contrário, *conjunto puro*. Essas regiões são mostradas na Figura 2.1.

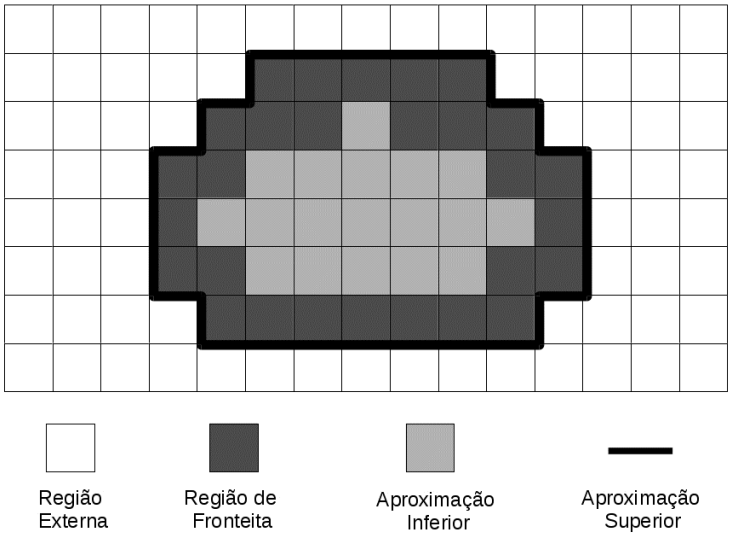

Figura 2.1 - Aproximações de um conjunto na TCA.

Para o exemplo mostrado na Tabela 2.2, as aproximações para as classes $X_1 = \{x_1, x_3\}$ e $X_2 = \{x_2, x_4, x_5, x_6\}$, são:

X_1:

$\underline{B}X_1 = \{[x_3]\}$;

$\overline{B}X_1 = \{[x_1, x_4], [x_3]\}$;

$F(X_1) = \overline{B}X_1 - \underline{B}X_1 = \{[x_1, x_4]\}$;

$E(X_1) = U - \overline{B}X_1 = \{[x_2], [x_5], [x_6]\}$.

X_2:

$\underline{B}X_2 = \{[x_2], [x_5], [x_6]\}$;

$\overline{B}X_2 = \{[x_2], [x_1, x_4], [x_5], [x_6]\}$;

$F(X_2) = \overline{B}X_2 - \underline{B}X_2 = \{[x_1, x_4]\}$;

$E(X_2) = U - \overline{B}X_2 = \{[x_3]\}$.

Como ilustrado no exemplo acima, os elementos x_1 e x_4 pertencem à região de fronteira (*F*), cuja existência torna o conjunto de dados aproximativo.

A TCA define ainda uma região adicional, que é relativa ao lugar onde só existem elementos que certamente pertencem a uma determinada classe. Essa região é chamada de *região positiva*. Formalmente, em TCA, este espaço é definido por:

$$POS_B(d) = \bigcup_X \underline{B}X \tag{2.4}$$

onde $X = \{X_1,...X_{r(d)}\}$ e $r(d) = |V_d|$.

No caso do exemplo apresentado, a região positiva $POS_B(d)$ é dado por:

$$POS_B(d) = \underline{B}X_1 \cup \underline{B}X_2 = \{[e_3] \cup [e_2, e_5, e_6]\}$$

2.2 Reduções

Em um sistema de informação, os dados podem ser desnecessariamente volumosos ao menos de duas formas. Um se refere ao número de elementos duplicados ou iguais na base de dados. Uma modo simples de contornar esse fato é pegar um representante da classe de equivalência, formada pela relação de indiscernibilidade, para representar toda a classe. Outra forma de reduzir ou compactar um conjunto de dados, em TCA, é através da eliminação de atributos condicionais supérfluos. Essa dispensabilidade de atributos é verificada por meio da quantidade de informação que o conjunto de atributos condicionais reduzido preserva, frente ao conjunto completo de atributos condicionais. Por sua vez, essa quantidade de informações pode ser observada com base nas classes de equivalência formadas, por meio da relação de indiscernibilidade. A um conjunto de atributos reduzidos que preserva as mesmas estruturas das classes de equivalência, que o conjunto completo de atributos condicionais é dado o nome de *redução*.

Uma redução pode ser entendida como um subconjunto de atributos condicionais $B \bullet A$, sendo A o conjunto completo de atributos condicionais, tal que $IND(B) = IND(A)$. Em outras palavras, uma redução, $RED(B)$, é o conjunto mínimo de atributos de A que preserva o particionamento do universo realizado pela relação de indiscernibilidade.

O número de reduções de um sistema de informação $S = (U, A \bullet \{d\})$, com $m = |A|$ atributos condicionais, pode ser igual as possíveis combinações dos atributos, dado por (KOMOROWSKI et al., 1999):

$$\binom{m}{\left\lfloor \frac{m}{2} \right\rfloor} \tag{2.5}$$

onde o termo $\lfloor m/2 \rfloor$ denomina o *piso*, ou maior inteiro menor ou igual a $m/2$.

Na Tabela 2.3 é exibido o número de reduções que é possível computar de um conjunto de m atributos condicionais.

Tabela 2.3. Número possível de reduções em função do número de atributos condicionais (m).

m	**Número de reduções**
10	252
20	184.756
30	155.117.520
40	137.846.528.820
50	126.410.606.437.752

Assim, o cálculo das reduções apresenta alta complexidade algorítmica, sendo um problema NP-difícil. Esse problema é contornado pelo uso de heurísticas ou metaheurísticas que fazem uma busca no espaço das reduções gerando sucessivas soluções candidatas (reduções). A Seção 2.3 ilustra a abordagem clássica em TCA para o cálculo das reduções, enquanto que a Seção 2.4, ilustra o cálculo de reduções baseado em dependência de atributos. No escopo deste último, serão propostas no Capítulo 3 as metaheurísticas VNS, VND, ILS

e DCS para o cálculo de reduções, utilizando duas funções de avaliação das soluções/reduções.

2.3 Cálculo de reduções baseado em matriz de discernibilidade

O método proposto por (KOMOROWSKI et al., 1999; PAWLAK, 1982) para cálculo de reduções consiste na simplificação de uma função que é produzida pela matriz de discernibilidade. Dado um sistema de informação $S = (U, A \bullet \{d\})$ com n elementos, define-se a correspondente matriz de discernibilidade como sendo uma matriz simétrica de dimensão n x n, onde cada posição c_{ij} da matriz (para $i \neq j$) é preenchida por um subconjunto de atributos condicionais que diferenciam os elementos x_i e x_j. Claramente, as posições c_{ii} (diagonal da matriz) serão nulas.

$$c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\} \text{ para } i,j = 1, ..., n \text{ e } i \neq j \tag{2.6}$$

Na Tabela 2.4 é mostrada a matriz de discernibilidade para o exemplo mostrado na Tabela 2.2 da Seção 2.1. Na entrada da matriz correspondente aos elementos (x_1, x_2), somente os atributos a_1 e a_2 permitem distingui-los.

Tabela 2.4. Matriz de discernibilidade para o exemplo considerado.

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	-					
x_2	a_1, a_2	-				
x_3	a_1, a_3	a_1, a_2, a_3	-			
x_4	Ø	a_1, a_2	a_1, a_3	-		
x_5	a_3	a_1, a_2, a_3	a_1	a_3	-	
x_6	a_1, a_2, a_3	a_3	a_1, a_2	a_1, a_2, a_3	a_1, a_2	-

A função de discernibilidade f_A para o sistema $S = (U, A \bullet \{d\})$ é uma função Booleana de $m=|A|$ atributos condicionais, correspondentes aos atributos $a_1,..., a_m$, obtida da matriz de discernibilidade, sendo expressa por:

$$f_A(a_1^*,\ldots,a_m^*) = \wedge\left\{\vee c_{ij}^* \mid 1 \le j \le i \le n, c_{ij} \neq \varnothing\right\} \tag{2.7}$$

onde . $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$.

Abaixo é mostrada, como exemplo, a função de discernibilidade obtida da matriz de discernibilidade da Tabela 2.4. Para fins de melhor visualização, cada linha da função de discernibilidade corresponde a uma coluna da matriz de discernibilidade, conforme abaixo.

$$f_A(a_1,a_2,a_3) = [(a_1 \vee a_2) \wedge (a_1 \vee a_3) \wedge (a_3) \wedge (a_1 \vee a_2 \vee a_3)] \wedge$$
$$[(a_1 \vee a_2 \vee a_3) \wedge (a_1 \vee a_2) \wedge (a_1 \vee a_2 \vee a_3) \wedge (a_3)] \wedge$$
$$[(a_1 \vee a_3) \wedge (a1) \wedge (a_1 \vee a_2)] \wedge$$
$$[(a_3) \wedge (a_1 \vee a_2 \vee a_3)] \wedge$$
$$[(a_1 \vee a_2)]$$

O conjunto de atributos condicionais, determinados pela simplificação Booleana de f_A, produz o conjunto de reduções de A. A simplificação Booleana da função de exemplo mostrada acima, gera a seguinte função:

$$f_A(a_1,a_2,a_3) = a_1 \wedge a_3 = a_1 a_3$$

Esta função simplificada corresponde ao seguinte sistema de informação reduzido ilustrado na Tabela 2.5.

Tabela 2.5. Sistema de decisão correspondente à redução a_1a_3

	a_1	a_3	d
x_1	2	0	0
x_2	0	0	1
x_3	1	1	0
x_4	2	0	1
x_5	2	1	1
x_6	0	1	1

No exemplo acima exposto, as reduções foram calculadas considerando todas as colunas da matriz de discernibilidade. Consequentemente a função de discernibilidade sofreu todas as simplificações possíveis, obtendo-se então as chamadas *reduções completas*. Alternativamente, pode-se calcular as reduções parciais a partir de um conjunto reduzido de colunas da matriz de discernibilidade, derivando, portanto de parte da função de discernibilidade completa, isto é não incluindo todas suas cláusulas. No exemplo, uma função de discernibilidade simplificada f_A, construída unicamente a partir da primeira coluna da matriz de discernibilidade (correspondente ao elemento x_1), resultará nas seguintes reduções:

$$f_A(a_1,a_2,a_3) = [(a_1 \vee a_2) \wedge (a_1 \vee a_3) \wedge (a_3) \wedge (a_1 \vee a_2 \vee a_3)] = a_1a_3 \bullet a_2a_3$$

Neste caso o elemento x_1 pode ser diferenciado dos demais elementos de duas formas: (i) usando os atributos condicionais a_1 e a_3 ou (ii) usando os atributos condicionais a_2 e a_3. Essas reduções simplificadas são denominadas *reduções k-relativas* (no caso exemplificado, $1 \leq k \leq |A|$).

2.4 Cálculo de reduções baseado em dependência de atributos

Como mostrado anteriormente, as reduções podem ser calculadas pelo modo originalmente proposto por (KOMOROWSKI et al., 1999; PAWLAK, 1982), que utiliza a simplificação Booleana da matriz de discernibilidade. Entretanto, devido à complexidade deste cálculo para bases de dados complexas, foi

proposto o cálculo de reduções em TCA com base na definição de dependência de atributos, a qual requer uma função específica que avalie a dependência funcional entre o subconjunto de atributos condicionais e o atributo de decisão. Existem diversos métodos alternativos para cálculo de reduções, baseados no aplicação de heurísticas e metaheurísticas para a busca de soluções seja usando a matriz de discernibilidade ou usando a dependência de atributos.

A noção de dependência funcional vem da área de Banco de Dados e se refere a restrições de integridade. Dados dois conjuntos de atributos *A*, *B*, é dito "*B* depende de *A*" ou "$A \bullet B$" se para cada valor de *A* existe um único valor de *B* (KORTH; SILBERSCHATZ, 1993). Essa noção foi incorporada à TCA para que se possa avaliar o grau de dependência entre atributos.

Nesta tese será dada ênfase a dois métodos que usam dependência de atributos, a Redução de Atributos dos Conjuntos Aproximativos (ou RSAR, em inglês *Rough Sets Attribute Reduction*), exposta na Subseção 2.4.1, e a redução de atributos baseada na sua dependência relativa, na Subseção 2.4.2.

Essas novas abordagens são muitos eficazes para encontrar subconjuntos de atributos condicionais que requerem obviamente menos tempo de processamento, mas conservam a informação contida no conjunto original de atributos. Em particular neste estudo, as duas funções de avaliação baseadas em dependência de atributos são utilizadas para avaliar as soluções candidatas (reduções) geradas pelas quatro metaheurísticas (VNS, VND, ILS e DCS), empregadas de forma inédita no cálculo e reduções em TCA (PESSOA; STEPHANY, 2014).

2.4.1 Redução de atributos dos conjuntos aproximativos

Um conceito importante na TCA para a análise de dados é a dependência entre atributos. Uma função chamada de *dependência de atributos* é muito útil para medir a dependência funcional entre os atributos condicionais e o atributo de decisão. Dado dois conjuntos *C* e *D*, se diz que *D* depende totalmente de C, representado por $C \bullet D$, se todos os atributos de *D* são determinados pelos valores de *C*. Em outras palavras: *D* depende de *C* se não existem inconsistência no mapeamento dos elementos de *C* para o elementos contidos em *D*. Assim, se *D* depende totalmente de *C*, existe uma dependência funcional entre os valores de *C* e *D*. Formalmente, a dependência de atributos é dada por (KOMOROWSKI et al., 1999):

$$\gamma_C(D) = \frac{\left|POS_C(D)\right|}{\left|U\right|} \tag{2.8}$$

onde γ_B é chamado de grau de dependência entre atributos *e* $POS_C(D)$ é a região positiva. Se $\gamma_B = 1$ então *D* depende *totalmente* de *C* e se $\gamma_B < 1$, D depende *parcialmente*, em grau γ_B de C. Se $\gamma_B = 0$, então D não depende de B.

Incorporando o conceito de grau de dependência ao cálculo das reduções em TCA, reescreve-se a Equação 2.8 assumindo-se que $C = B$ e $B \subseteq A$ e $D = d$, onde *d* é o atributo de decisão, o que resulta na equação seguinte:

$$\gamma_B(d) = \frac{\left|POS_B(d)\right|}{\left|U\right|} \tag{2.9}$$

O grau de dependência entre os atributos condicionais e o atributo de decisão quantifica a redundância dos primeiros e permite o cálculo de reduções, as quais minimizam as eventuais inconsistências do sistema de informação. Considerando o conjunto completo de atributos condicionais *A*, qualquer subconjunto $B \bullet A$ tal que $\gamma_B(d) \geq \gamma_A(d)$ é uma redução. Isto pode ser provado como mostrado a seguir.

Supondo-se que *B* seja um subconjunto do conjunto completo de atributos condicionais *A*, então $B \subseteq A$ e considerando-se d como atributo de decisão, então o conjunto *R* de todas as possíveis reduções é dado por:

$$R = \{B : B \subseteq A \mid \gamma_B(d) \geq \gamma_A(d)\} \tag{2.10}$$

No conjunto de dados do exemplo da Tabela 2.2, $d = 0$ determina a classe $X_1 = \{x_1, x_3\}$ e $d = 1$ a classe $X_2 = \{x_2, x_4, x_5, x_6\}$. A Equação 2.9 permite calcular o grau de dependência para o conjunto completo de atributos condicionais $A = \{a_1, a_2, a_3\}$:

$$\gamma_A(d) = \frac{|\{x_2\},\{x_3\},\{x_5\},\{x_6\}|}{|U|} = \frac{4}{6}$$

Analogamente, o grau de dependência é calculado para todos os subconjuntos possíveis de *A*:

$$\gamma_{\{a_1\}}(d) = \frac{3}{6}, \gamma_{\{a_2\}}(d) = \frac{2}{6}, \gamma_{\{a_3\}}(d) = 0,$$
$$\gamma_{\{a_1,a_2\}}(d) = \frac{3}{6}, \gamma_{\{a_1,a_3\}}(d) = \frac{4}{6}, \gamma_{\{a_2,a_3\}}(d) = \frac{2}{6}$$

Finalmente, o subconjunto $\{a_1, a_3\}$ pode ser considerado uma redução de *A*, uma vez que possui o mesmo grau de dependência de *A*.

2.4.2 Redução de atributos baseada em dependência relativa

Na Seção 2.4.1 foi apresentado um modo de calcular uma redução usando o grau de dependência de atributos. Uma alternativa mais barata computacionalmente, proposta por Han et al. (2005) é uma métrica chamada de *dependência relativa de atributos*. O grau de dependência relativa de atributos pode ser calculado pela razão entre o número de classes de equivalência da partição do universo de discurso, segundo a relação de indiscernibilidade para um dado subconjunto de atributos condicionais e o número de classes de equivalência da partição criada pela relação de indiscernibilidade de um subconjunto de atributos condicionais unidos com o atributo de decisão. Dado A o conjunto de todos os atributos condicionais, $B \bullet A$ e d o atributo de decisão, a dependência relativa de atributos é dada por:

$$\kappa_B(d) = \frac{|U/IND(B)|}{|U/IND(B \cup d)|} \tag{2.11}$$

O grau de dependência relativa mede a consistência ou certeza de uma base de dados. Quanto menos elementos pertencentes a uma base de dados forem incertos, ou seja, semelhantes segundo os atributos condicionais ($B \bullet A$), porém pertencentes a classes diferentes, maior será o valor do numerador, ou seja, maior será o essa relação.

Empregando a dependência relativa de atributos κ_B ao invés do grau de dependência γ_B na Equação 2.10, resulta no seguinte conjunto de reduções:

$$R = \{B : B \subseteq A \mid \kappa_B(d) \geq \kappa_A(d)\} \tag{2.12}$$

Considerando-se o sistema de informação exemplificado na Tabela 2.2, o atributo de decisão d permite particionar o conjunto universo da seguinte forma:

para d = 0, X_1 = {x_1, x_3} e para d = 1 tem-se a classe X_2 = {x_2, x_4, x_5, x_6}. Considerando o conjunto completo de atributos condicionais A= {a_1, a_2, a_3} é possível calcular $\kappa_A(d)$ pela Equação 2.12:

$$\kappa_A(d) = \frac{|\{x_1, x_4\},\{x_2\},\{x_3\},\{x_5\},\{x_6\}|}{|\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\}|} = \frac{5}{6}$$

A repetição do cálculo de $\kappa_B(d)$ para cada subconjunto $B \subseteq A$ resulta em:

$$\kappa_{\{a_1\}}(d) = \frac{3}{4}, \kappa_{\{a_2\}}(d) = \frac{2}{3}, \kappa_{\{a_3\}}(d) = \frac{2}{4},$$
$$\kappa_{\{a_1,a_2\}}(d) = \frac{3}{4}, \kappa_{\{a_1,a_3\}}(d) = \frac{5}{6}, \kappa_{\{a_2,a_3\}}(d) = \frac{4}{6}$$

Do mesmo modo que na abordagem utilizando a dependência de atributos, somente o subconjunto {a_1, a_3} tem um grau de dependência relativa maior ou igual a κ_A(d) é considera uma redução. Portanto, neste exemplo, o conjunto de reduções é dado por uma única redução R = [{a_1, a_3}].

2.5 Indução de regras de decisão

Cada redução constitui um conjunto reduzido de atributos condicionais, que acoplados ao atributo de decisão constitui um sistema de informação reduzido. Este sistema pode então ser expresso por um conjunto de regras, especificamente chamado de regras de decisão (OHRN, 1999). Assim, a partir do conjunto de reduções pode-se derivar um conjunto de regras de decisão, as

quais serão utilizadas na classificação para rotular/classificar as instâncias da base de dados.

Seja $S = (U, A \bullet d)$ um sistema de informação, onde A é o conjunto de atributos condicionais, U *é* o conjunto universo e d, o atributo de decisão. O padrão α denota uma conjunção de descritores pertencentes aos atributos em A, do tipo $a = a(x)$ com $a \bullet A$ e $x \bullet U$ e o padrão β também denota uma conjunção de descritores, como $d = d(x)$. Uma regra é denotada por $\alpha \bullet \beta$ e lê-se como "se α então β". O padrão α é chamado de *antecedente* e β de *consequente.* Considerando-se o conjunto universo e o atributo de decisão d pode-se derivar uma regra especifica para um instancia x desse sistema, a qual é definida por um produtório que expressa a conjunção de vários condicionais, conforme abaixo.

$$regra(S,x,d) = \prod_{a \in B} a = a(x) \Rightarrow \prod_{d \in D} d = d(x) \tag{2.13}$$

onde $B \bullet A$ e D é um conjunto de atributos de decisão.

Analogamente, pode-se derivar regras para todas as demais instâncias do sistema de informação considerado. O conjunto final de regras de decisão é obtido pela união das diversas regras, eliminando-se as regras repetidas:

$$regras(S,d) = \bigcup_{x \in U} P(S,x,d) \tag{2.14}$$

2.6 Particionamento aleatório do conjunto de treinamento

Como enfatizado anteriormente, o cálculo de reduções em TCA tem alta complexidade algorítmica, frequentemente podendo ser classificado como um problema NP-difícil quando é feito para bases de dados complexas e com

muitas instâncias, como é o caso desta pesquisa. O uso de heurísticas e metaheurísticas, conforme exposto no final da Seção 2.2 e apresentado no Capítulo 3, é uma alternativa bastante empregada. Entretanto, nem assim a mineração de dados meteorológicos proposta aqui se tornou um problema tratável, requerendo o uso de um método adicional, o Particionamento Aleatório do Conjunto de Treinamento (PACT), ou em inglês, *Random Training Set Partitions*, proposto para uso em TCA por Gupta et al. (2006), que consiste em:

1. Geração aleatória de *p* partições iguais (conjuntos de treinamento) da base de dados original;
2. Cálculo de uma redução sub-ótima para cada partição, uma vez que não há como se garantir que se encontrou a solução ótima global para cada partição;
3. Obtenção do conjunto de reduções resultante pela simples união das reduções sub-ótimas das partições.

Este trabalho de Gupta et al. (2006) também demonstra que o PACT pode reduzir o tempo de cálculo das reduções de cada partição por uma razão de *p* vezes em relação ao tempo de cálculo original, que corresponde usar uma partição única. Isso se explica pela complexidade algorítmica do cálculo das reduções, que tem ordem $O(VN^2)$, onde V é a cardinalidade do conjunto de valores possíveis dos atributos condicionais (ou seja, o número de valores diferentes que cada atributo pode assumir) e N é o número de instâncias da base de dados. Assim, o tempo de treinamento tende a variar linearmente com o número de instâncias. Por outro lado, obviamente, o tempo total tende a ser semelhante, pois é preciso calcular *p* partições de forma independente.

Então, qual seria a vantagem da utilização do PACT? Em primeiro lugar, o uso de partições permitiria o cálculo de reduções "especializadas" para as instâncias de cada partição e a união destas constituiria um conjunto abrangente de reduções em relação a todas as partições. Teoricamente, isso possibilitaria a indução de regras de decisão e a obtenção de um correspondente classificador com melhor desempenho de classificação. Em segundo lugar, o cálculo de reduções pode ser feito de maneira independente para cada partição, permitindo facilmente sua paralelização. Entretanto, isso não significa que aumentar indefinidamente o número de partições *p* permita melhorar o desempenho de classificação indefinidamente. A tendência é que, acima de um certo valor, a "especialização" excessiva leve a piorar esse desempenho.

Conforme será exposto no Capítulo 5, o uso do PACT mostrou-se imprescindível para se obter reduções com baixo tempo de processamento graças à paralelização do cálculo das reduções para as partições. Além disso, obteve-se um conjunto de reduções com melhor desempenho de classificação na mineração de dados meteorológicos.

3 HEURÍSTICAS E METAHEURÍSTICAS APLICADAS NA TCA

Conforme exposto no capítulo anterior, as heurísticas e metaheurísticas são usadas na TCA para redução de atributos com a finalidade de diminuir a complexidade algorítmica em relação à abordagem clássica baseada na matriz de discernibilidade, a qual pode consequentemente ser inviável para bases de dados com dimensionalidade muito alta.

Existe também a chamada "maldição da dimensionalidade" (THEODORIDIS; KOUTROUMBAS, 2009), relacionada à degradação da qualidade da informação com o aumento do número de atributos de uma base de dados. Assim, em problemas de classificação, é importante eliminar atributos supérfluos ou que degradem o desempenho do classificador. O tempo de processamento da classificação é proporcional à complexidade da base de dados, ou seja, ao número de registros, ao número de atributos e ao número de classes.

Segundo Gaspar-Cunha et al. (2013), as heurísticas são procedimentos aplicados a problemas de otimização sem garantias teóricas de que uma solução ótima seja obtida, ou seja, que permitem encontrar soluções sub-ótimas, mesmo considerando que uma destas soluções tenha proximidade com a solução ótima. Em outras palavras, são procedimentos simplificados de exploração do espaço de soluções. Heurísticas são tipicamente aplicadas na resolução de alguns tipos de problemas, como, por exemplo, os de complexidade do tipo NP-difícil, para os quais não existem atualmente algoritmos capazes de encontrar uma solução ótima em tempo polinomial.

Nas últimas décadas tem havido um interesse crescente no desenvolvimento de heurísticas inspiradas em fenômenos da natureza, denominadas metaheurísticas e caracterizadas por sua natureza estocástica. Esse é o caso das metaheurísticas bio-inspiradas, características da área de Computação Evolutiva, como é o caso dos algoritmos genéticos. As metaheurísticas são tipicamente aplicáveis a problemas de otimização combinatória e possuem uma

boa capacidade de explorar todo o espaço de busca (ou espaço de soluções), "fugindo" da atração de mínimos locais, tal como ocorre com algoritmos determinísticos (GASPAR-CUNHA et al., 2013).

Nesta tese, em particular, utilizam-se heurísticas e metaheurísticas no cálculo de reduções em TCA, o qual é do tipo NP-difícil (KOMOROWSKI et al., 1999). Desde o surgimento da TCA, pesquisadores da área sempre buscaram métodos capazes de calcular as reduções, especialmente em conjuntos de dados complexos.

Alguns trabalhos merecem destaque no desenvolvimento de métodos para o cálculo de redução. Em Ohrn (1999) é apresentado o software ROSETTA (acrônimo em inglês de *Rough Sets Toolkit for Analysis of Data*), no qual foram implementados um algoritmos genético e o algoritmo de Johnson. Algoritmos genéticos são metaheurísticas bio-inspiradas propostas por John Holland e seus alunos, na Universidade de Michingan, nos anos 1960 (HOLLAND, 1975). O algoritmo de Johnson é uma heurística gulosa, comumente aplicada na TCA, que seleciona os atributos mais frequentes na matriz de discernibilidade, retornando apenas uma redução (JOHNSON, 1973).

Trabalhos mais recentes sobre o cálculo de reduções na TCA, como o de Jensen e Shen (2003) e Jensen e Shen (2005) utilizaram diversas metaheurísticas tais como algoritmos genéticos, Otimização por Colônia de Formigas (em inglês *Ant Colony Optimization* ou ACO) e Recozimento Simulado (em inglês *Simulated Annealing* ou SA). A Otimização por Colônia de Formigas, proposta por Goss et al. (1989) , reproduz o comportamento de uma colônia de formigas na busca por alimentos. A técnica de Recozimento Simulado, proposta por Kirkpatrick et al. (1983) é uma metaheurística inspirada no processo de recozimento de um sólido para obtenção de um estado cristalino mais estável (de menor energia) e que apresente menos defeitos, processo estudado na área de física da matéria condensada. Toda metaheurística necessita de uma forma de avaliação da qualidade da solução

candidata, sendo tipicamente empregado o valor de uma função objetivo para avaliar a solução considerada.

Em Hedar et al. (2008) foi implementada a metaheurística Busca Tabu (do inglês *Tabu Search* ou TS) que foi empregado neste trabalho para o cálculo das reduções e comparado aos resultados encontrados por Jensen e Shen (2003) e Jensen e Shen (2005). A Busca Tabu é uma metaheurística proposta pelos trabalhos independentes de Fred Glover (GLOVER; MCMILLAN, 1986) e Pierre Hensen (HANSEN, 1986) tendo sido concebida como uma técnica para guiar uma heurística de busca local na exploração do espaço de soluções, empregando estruturas de memória específicas, caracterizando-se pelo seu bom desempenho em termos de encontrar soluções sub-ótimas (GASPAR-CUNHA et al., 2013).

Wang et al. (2014) propuseram a metaheurística a chamada Busca por Espalhamento (em inglês *Scatter Search*), também proposta por Fred Glover (GLOVER, 1977). Esta técnica é possivelmente a precursora da Busca Tabu (GASPAR-CUNHA et al., 2013), constituindo uma metaheurística baseada em população na qual as soluções são armazenadas no Conjunto de Referência (RefSet).

No tocante a heurísticas e metaheurísticas aplicadas ao cálculo de reduções em TCA, é preciso ressaltar as limitações típicas devidas à complexidade da base de dados. O tempo de processamento gasto no cálculo das reduções e, posteriormente, na classificação, é proporcional a essa complexidade, a qual depende do número de registros da base de dados, do número de atributos, do número de valores discretos que cada atributo pode assumir e também do número de classes. O cálculo de reduções pela abordagem da matriz de discernibilidade fica limitado a bases de dados não tão complexas devido ao tempo de processamento demandado ou a limitações de memória. As abordagens aqui propostas são mais escaláveis, no sentido de permitirem o uso de bases de dados com maior número de registros.

Neste contexto, foram propostas e implementadas quatro metaheurísticas para o cálculo das reduções na TCA: (i) Busca em Vizinhança Variável (VNS), (ii) Descida em Vizinhança Variável (VND), (iii) Busca Local Iterativa (ILS) e (iv) Busca Decrescente de Cardinalidade (DCS), sendo esta última uma nova heurística derivada do VNS, que tem como característica a busca aleatória de novas soluções de mínima cardinalidade (PESSOA; STEPHANY, 2014).

As soluções candidatas (reduções) geradas pelas metaheurísticas serão avaliadas pelas funções de avaliação apresentadas nas Subseções 2.4.1 e 2.4.2. A notação utilizada para designar cada função de avaliação é $f(\bullet)$, obedecendo à condição $0 \leq f(\bullet) \leq 1$. Eventualmente, estas mesmas funções permitiriam o cálculo de reduções, porém de forma ineficiente, pois requereriam a seleção “manual” dos atributos condicionais. As subseções seguintes abordam a notação adotada para representação das soluções candidatas, o método de busca local padrão e as 4 metaheurísticas propostas.

3.1 Representação das soluções

As soluções candidatas usadas para todas as metaheurísticas deste trabalho são representadas por uma cadeia binária de comprimento $|A|$, onde $|A|$ corresponde à cardinalidade do conjunto de todos os atributos condicionais.

Cada posição desta cadeia corresponde a um atributo condicional. Desta forma a representação “1”, em uma dada posição, indica a presença de um atributo $a \bullet A$, na solução ‘s’, enquanto que a representação “0” indica sua ausência. Por exemplo, considerando-se $A = \{a_1,a_2,a_3,a_4,a_5\}$ e $B \bullet A$, tem-se:

1. $B=\{a_1,a_3\}$:
    - s = 10100
2. $B=\{a_4\}$:
    - s = 00010

3. $B=A=\{a_1,a_2,a_3,a_4,a_5\}$:
    - s = 11111
4. $B=\bullet$:
    - s = 00000

A cardinalidade |s| de uma dada solução *s* representa seu número de atributos condicionais, ou seja, o número de 1's encontrados na cadeia binária, podendo ser representada por:

$$| s |= \sum_{i=1}^{|A|} s_i \tag{3.1}$$

onde s_i é a *i*-ésima posição da solução *s*.

Uma operação comum quando se utiliza este tipo de representação é a permutação do bit que representa o valor de um determinado atributo condicional, implementada aqui pela função “TrocaBit”, a qual alterna o valor de uma dada posição *i* da cadeia ($1 \leq i \leq |s|$) de “1” para “0” ou vice-versa.

A métrica de proximidade adotada aqui entre uma dada solução *s* e as soluções vizinhas é a distância de *Hamming* (THEODORIDIS; KOUTROUMBAS, 2009), calculada como sendo o número de posições com valores diferentes entre as cadeias binárias da solução considerada *s* e da solução vizinha *s'*, ambas com cardinalidade |s|:

$$d_H(s,s^*) = \sum_{i=1}^{|A|} | s_i - s_i^* | \tag{3.2}$$

Seguem abaixo dois exemplos simples para ilustrar o cálculo da distância de *Hamming*:

- s = 1010 e s^* = 1110 • $d_H(s,s^*) = 1$;
- s = 1010 e s^* = 1100 • $d_H(s,s^*) = 2$.

Define-se estrutura de vizinhança $N_k(s)$ de uma solução s ao conjunto de soluções que tem distância de *Hamming* em relação a ela igual a k, as quais constituem a chamada k-ésima vizinhança dessa solução. Assim, por exemplo, a estrutura de vizinhança $N_1(s)$ contém todas as soluções em que apenas o valor binário de um único atributo condicional difere de s (distância de *Hamming* k= 1). No caso da estrutura de vizinhança $N_2(s)$, as soluções diferem de s pelo valor de dois atributos condicionais quaisquer, e assim por diante.

3.2 Busca Local Padrão (SLS)

O algoritmo de Busca Local Padrão (denotado aqui por SLS, do inglês *Standard Local Search*) adotado neste trabalho é uma heurística de busca de um vizinho da solução inicial/corrente (s) que possua um maior (melhor) valor da função de avaliação utilizada. Optou-se por buscar apenas soluções na vizinhança limitada à distância de *Hamming* igual a 1, ou seja, somente soluções com apenas um atributo a mais ou a menos que a solução inicial/corrente (s). Este algoritmo é bastante simples, utilizando a função TrocaBit( ), explicada na seção anterior para gerar as soluções s' da vizinhança de s com k=1. Cada uma destas soluções é avaliada pela função de avaliação $f(\bullet)$ considerada, sendo a melhor solução *s_local* retornada pelo algoritmo.

ALGORITMO 1 – SLS

Considerando-se uma solução inicial/corrente s:

```
início SLS
s_local = s
para i = 1 até i ≤ |A| faça
   s' ← TrocaBit(s,i)
   se f(s') > f(s) então
       s_local ← s'
   fim se
fim para
retorne s_local
fim SLS
```

Nesta tese, a metaheurística VND (*Variable Neighborhood Descent*) foi empregada como alternativa à Busca Local Padrão, que é descrita mais adiante.

3.3 Busca Local Iterativa (ILS)

A Busca Local Iterativa, ou em inglês, *Iterated Local Search* (ILS), foi proposta por vários pesquisadores de forma independente, sendo também conhecida como *Large-step Markov Chains* (MARTIN et al., 1991) ou *Iterated Lin-Kerninghan* (JOHNSON; MCGEOCH, 1997). Lourenço et al. (2002) identificou as semelhanças entre esses dois métodos e fez uma tentativa de unificá-los, adotando o nome ILS. Esta metaheurística tem como principal característica a simplicidade de parametrização, embora tenha bom desempenho em relação a outras metaheurísticas, como, por exemplo, no problema do caixeiro-viajante (GASPAR-CUNHA et al., 2013).

A metaheurística ILS utiliza um algoritmo de busca local qualquer para obter soluções melhores em vizinhanças de uma solução obtida pela perturbação aleatória da solução corrente *s*. Esta perturbação tem como objetivo escapar

da atração de mínimos locais e explorar de forma mais extensa o espaço de busca. Aqui, o esquema de busca local adotado foi o SLS, descrito na seção anterior. O pseudocódigo do ILS é mostrado no Algoritmo 2, que retorna a solução sub-ótima gerada *s**. Nesse pseudocódigo *N* denota um limite de iterações sem que haja melhora da solução.

ALGORITMO 2 – ILS

```
s • Gera()
s' ← BuscaLocal(s)
s* ←s'

início ILS
enquanto (limite de iterações não for alcançado) faça
   s ← Perturba(s')
   s" ← BuscaLocal(s);
   se f(s") > f(s') então
      s* ← s"
   fim se
   se (número de iterações sem melhora > N) então
       s' • Gera()
   fim se
fim enquanto
retorne s*
fim ILS
```

O algoritmo ILS utiliza-se de uma solução inicial *s*, que pode ser gerada aleatoriamente pela função Gera( ) ou então ser dada pelo próprio conjunto completo de atributos condicionais. A seguir, é feita uma busca local padrão para encontrar a melhor solução vizinha *s*', que passa a ser a solução corrente. Iniciam-se assim as iterações e a cada iteração, a melhor solução corrente sofre uma perturbação aleatória efetuada pela função Perturba( ), que consiste na troca aleatória de um certo número, também aleatório, de valores posições (bits) da cadeia que representa a solução. A busca local é então efetuada na vizinhança dessa solução perturbada resultando numa nova solução corrente e passa-se à iteração seguinte. Entretanto, define-se um parâmetro relativo ao limite de iterações (*N*) sem que haja melhora da solução corrente, que quando ultrapassado força a substituição da solução corrente por uma nova solução obtida pela função Gera( ). Assim, o ILS explora sucessivas bacias de atração

para determinar os correspondentes mínimos locais, tentando encontrar uma solução sub-ótima. O algoritmo finaliza ao ser atingido um segundo limite relativo ao número total de iterações.

3.4 Busca e descida em vizinhança variável (VNS e VND)

A Busca em Vizinhança Variável, ou em inglês, *Variable Neighborhood Search* (VNS), é uma metaheurística estocástica que utiliza uma heurística de busca local para exploração de vizinhanças gradativamente maiores (HANSEN; MLADENOVIC, 1997, 2003). Assim como o ILS, tem implementação fácil e apresenta simplicidade de parametrização, chegando a ser considerada um caso particular do ILS (GASPAR-CUNHA et al., 2013). Segundo Hansen e Mladenovic (1997), o VNS baseia-se em três hipóteses:

1. Um ótimo local com relação a uma vizinhança não necessariamente corresponde a um ótimo com relação a outra vizinhança.

2. Um ótimo global corresponde a um ótimo local para todas as estruturas de vizinhança.

3. Em geral ótimos locais para uma determinada vizinhança são relativamente próximos entre si.

O algoritmo VNS utiliza-se de uma solução inicial *s*, que pode ser gerada aleatoriamente, pela função Gera( ) ou então ser dada pelo próprio conjunto completo de atributos condicionais. A cada iteração é realizada uma exploração numa vizinhança gradativamente maior da solução inicial/corrente *s*. A partir desta, é gerada aleatoriamente uma nova solução *s*' pela função GeraVizinho( ), dentro da vizinhança $N_k(s)$, onde $1 \leq k \leq L$, sendo L a máxima distância de *Hamming* adotada. A exploração inicia-se com k=1, sendo efetuada uma busca pela função BuscaLocal( ) para determinar o melhor vizinho *s*''. A geração aleatória de um novo vizinho e a busca local em torno deste são repetidas para

vizinhanças *k* gradativamente maiores até que se encontre uma solução *s*'' melhor avaliada que a solução corrente *s*, substituindo-a. A condição de parada do algoritmo é dada por um critério pré-estabelecido, no caso, um limite do número de iterações. O pseudocódigo do VNS é mostrado abaixo:

ALGORITMO 3 – VNS

```
s • Gera()

início VNS
enquanto(limite de iterações não for alcançado) faça
   k ← 1;
   enquanto (k ≤ L) faça
      s' ← GeraVizinho(N_k(s))
      s'' ← BuscaLocal(s');
      se f(s'') > f(s) então
         s ← s''
         k ← 1
      senão
         k ← k +1
      fim se
   fim enquanto
fim enquanto
retorne s
fim VNS
```

A metaheurística Descida em Vizinhança Variável, ou em inglês *Variable Neighborhood Descent* (VND), é uma variação do VNS na qual uma nova busca local é realizada a cada iteração na vizinhança da solução corrente. Em vez de se utilizar a função GeraVizinho( ) como no VNS e depois fazer uma busca local numa vizinhança *k*=1, o VND efetua uma busca local numa estrutura de vizinhança com cardinalidade *k*>=1, isto é, alternando o valor de um ou mais bits da cadeia binária que representa *s* solução, implementada pela função

k-BuscaLocal( ). A busca local padrão é restrita a *k*=1, conforme exposto anteriormente. O pseudocódigo do VND é mostrado abaixo:

ALGORITMO 4 – VND

```
s • Gera()

início VND
k ← 1;
enquanto (k ≤ L) faça
   s′ ← k-BuscaLocal(s,k)
   se f(s′) > f(s) então
      s = s′
      k = 1
   senão
      k = k +1
   fim se
fim enquanto
retorne s
fim VND
```

O VND também pode ser empregado como um esquema alternativo de busca local (CHAVES et al., 2007; PESSOA; STEPHANY, 2014), como no caso da presente tese.

3.5 Busca Decrescente de Cardinalidade (DCS)

A Busca Decrescente de Cardinalidade, ou em inglês, *Decrescent Cardinality Search* (DCS) é uma heurística nova proposta em PESSOA e STEPHANY (2014), iterativa e estocástica, que é derivada do VNS e que tem como principal característica a convergência rápida na busca de soluções para o problema de redução de atributos na TCA. O algoritmo DCS utiliza-se de uma solução inicial *s*, que pode ser gerada aleatoriamente, pela função Gera( ) ou então ser dada pelo próprio conjunto completo de atributos condicionais. Similarmente ao ILS e ao VNS, o DCS também emprega um esquema de busca local, o SLS, mas a principal característica do DCS é a geração aleatória de vizinhos com cardinalidade necessariamente menor que a da solução corrente.

Assim, a cada iteração, a solução inicial/corrente *s* é utilizada pela função GeraDCS( ) que gera aleatoriamente uma solução *s*' com cardinalidade menor

que a da solução corrente, correspondente a uma redução com menos atributos, ou seja, com mais zeros na sua representação de cadeia binária. Faz-se então uma busca local padrão (BuscaLocal( ) - SLS) na vizinhança de *s*' e a melhor solução vizinha *s*" substitui a solução corrente *s*, caso seja melhor. O algoritmo sofre parada ao ser atingido o limite do número de iterações. O pseudocódigo do DCS é apresentado a seguir, no Algoritmo 5.

ALGORITMO 5 – DCS.

```
s • Gera()

início DCS
enquanto (limite de iterações não for alcançado) faça
   s′ ← GeraDCS(s)
   s″ ← BuscaLocal(s′);
   se f(s″) > f(s) então
      s ← s″
   fim se
fim enquanto
retorne s
fim DCS
```

3.6 Softwares utilizados e implementados

Alguns softwares foram desenvolvidos, sendo em sua maioria gratuitos, para análise de dados utilizando a TCA. Komorowski et al. (1999) listam alguns desses softwares, destacando os seguintes; Grobian, PrimeRose, RSES e Rosetta.

Este trabalho utiliza dois softwares, ambos para o sistema operacional Linux. O primeiro é um software específico para o cálculo de reduções em TCA, desenvolvido pelo autor e que utiliza as metaheurísticas propostas. O segundo é o software gratuito Rosetta, utilizado para gerar o conjunto de regras, as quais são induzidas a partir das reduções. Posteriormente essas regras são aplicadas para a classificação nas bases de dados descritas e o desempenho de classificação é analisado.

O software desenvolvido para cálculo de reduções, com as metaheurísticas VNS, VND, ILS e DCS, foi implementado na linguagem interpretada Perl, com alguns scripts utilizando o interpretador de comandos Bash. Os dados são armazenados no Sistema Gerenciador de Base de Dados (SGBD) MySQL. O software Rosetta (*Rough Set Toolkit for Analysis of Data*) foi desenvolvido por Ohrn (1999) para análise de dados em TCA e possui suporte para todas as etapas do processo de KDD. Possui dois modos de operação, sendo uma por meio de GUI (*Graphical User Interface*) e outra por linha de comando, possibilitando o uso de scripts. Esse software tem aplicação limitada pois somente existem versões para computadores com tamanho da palavra de memória de 32 bits, ou seja, computadores com memória principal máxima de 4 GBytes, o que limita o tamanho da base de dados que pode ser analisado.

A seguir detalha-se o software para cálculo de reduções, que é que é orientado a objetos e cujo diagrama de classes é apresentado na Figura 3.1.

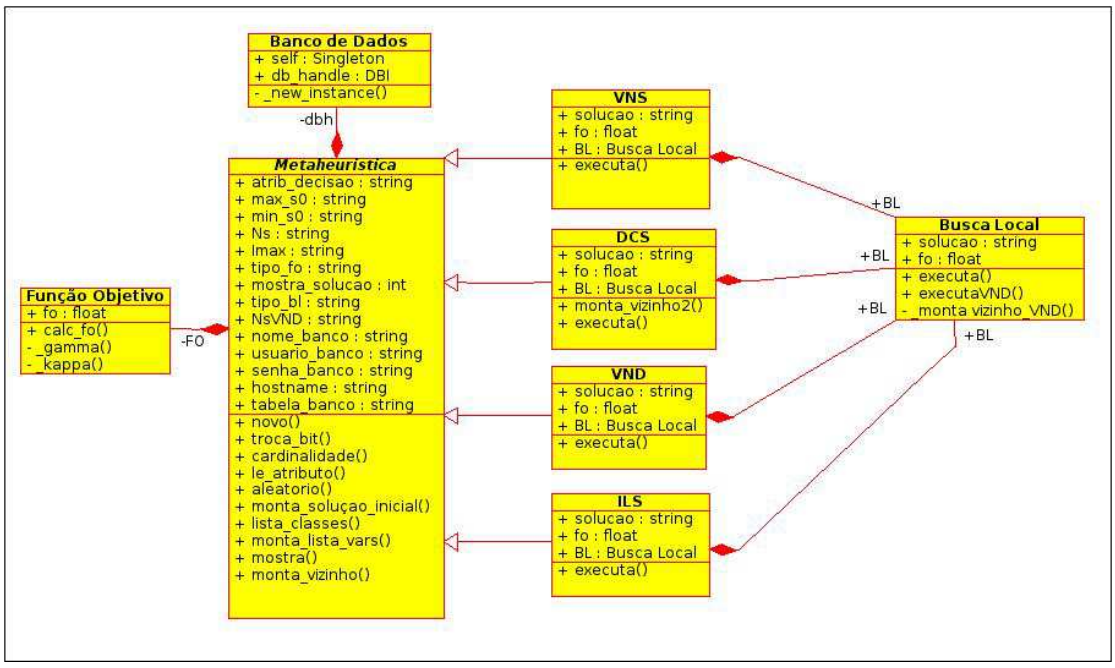

Figura 3.1 – Diagrama de classes utilizado na programação orientada a objetos do cálculo de reduções utilizando as metaheurísticas VNS, VND, ILS e DCS.

A superclasse abstrata “*Metaheurística*” tem 3 subclasses correspondentes e homônimas às metaheurísticas propostas VNS, VND, ILS e DCS, cada qual com o método que a implementa. Associada a estas subclasses há a subclasse "Busca Local" que implementa a busca local padrão (SLS), a metaheurística VND e sua busca local com cardinalidade de estrutura de vizinhança ≥ 1. Essa superclasse possui duas classes associadas, a classe "Função Objetivo", que implementa a função de dependência de atributos γ ou a dependência relativa κ para avaliar soluções candidatas, e a classe "Banco de Dados", implementada segundo o padrão de projeto de software (*design pattern*) conhecido como Singleton, que permite à superclasse o acesso persistente a um objeto dessa classe ("Banco de Dados") evitando a necessidade de estabelecer novas conexões ao banco de dados.

Outros métodos da superclasse “Metaheurística” permitem trocar bits de uma solução candidata (troca_bit), calcular a cardinalidade de uma solução candidata (cardinalidade), exibir a solução candidata na saída padrão (mostra). Os atributos dos objetos incluem usuário, senha, nome do banco de dados (nome_banco), tipo de função objetivo (tipo_fo), tipo de busca local (tipo_bl), etc. Os parâmetros de inicialização dos objetos das subclasses são passados através de um arquivo de configuração, conforme o exemplo a seguir (Figura 3.2).

Dessa forma, uma instância da superclasse refere-se a uma metaheurística com parâmetros específicos: cardinalidade da estrutura de vizinhança, tipo da função objetivo, tipo de busca local. Essa metaheurística é então aplicada a sucessivas instâncias da classe "Banco de Dados".

O capítulo seguinte trata dos dados meteorológicos usados nesta tese.

```
#####################################################################
#Parâmetros do Banco de Dados
hostname : localhost
usuario_banco : saturno
senha_banco : 1234
nome_banco : TESE2
tabela_banco : V_ESEV_RA_P1_48_1
atrib_decisao: classe
#Parâmetros das Metaheurísticas
#Estrutura de Vizinhança (Obrigatório para VNS e VND)
L : 0
#Num. máximo de iterações (Obrigatório para VNS e VND)
Imax : 10
#Metaheurística empregada (VNS, VND, ILS ou DCS). “Case
insensitive”.
algoritmo : dcs
#Busca Local: ls ou vnd
tipo_bl : ls
#Número de estruturas de vizinhança. Usado somente quando o VND for
#usado como busca local
LVND : 2
#Of=g -> gamma (RSAR), Of=k->kappa (relative dependency)
tipo_fo : g
#Exibe resultados
mostra : 1
#Verbosidade 0=nenhuma, 1=nível 1, 2=nível 2 e 3=nível 3
debug : 1
#Número máximo de atributos na solução inicial
max_s0 : max
#Número mínimo de atributos na solução inicial (quando max_s0 for
“max” #esse parâmetro será ignorado)
min_s0 : 10
#Tipo de nova solução (DCS).
#1 -> Snova = |Svelha|-1 ou 2 -> Snova = |Svelha|-rand(|Svelha|)
tipo_nova_sol : 1
#####################################################################
```

Figura 3.2 – Exemplo de um arquivo de configuração de um objeto de uma subclasse, que corresponde a uma determinada variação de uma das metaheurísticas.

4 DADOS METEOROLÓGICOS

Os dados meteorológicos usados nesta tese incluem dados correspondentes a saídas do modelo numérico de previsão do tempo ETA (MESINGER et al., 1988), com resolução de 20Km (ETA20) e a campos de densidade de descargas NS gerados pelo software EDDA (STRAUSS et al., 2013). Os dados do modelo numérico ETA20 são usados como variáveis preditivas dos padrões que identificam eventos convectivos severos, os quais estão correlacionados a altas densidades de ocorrência de descargas NS. Assume-se aqui que esta correlação seja válida para a maioria dos casos.

Na Tabela 4.1 é mostrada uma representação simbólica da base de dados utilizada nesta pesquisa, onde x_{max} é o número de pontos correspondentes às longitudes, y_{max}, é o número de pontos correspondentes às latitudes, t_{max} é o número de horários sinóticos correspondentes às saídas do modelo ETA20, sendo cada registro/instância da base de dados identificado pela tripla (t_i,x_i,y_i) e o número total de registros dado pelo produto p = [t_{max} × x_{max} × y_{max}]. Cada registro tem associado seu valor do atributo de decisão, a densidade de ocorrências de descargas d_i.

Tabela 4.1. Representação simbólica da base de dados.

Registro	**Posição**	**Data**	**Var_1**	**Var_2**	**….**	**Var_n**	**Classe**
e_1	(x_1,y_1)	t_1	$Var_1(e_1)$	$Var_2(e_1)$	...	$Var_n(e_1)$	d_1
e_2	(x_1,y_2)	t_2	$Var_1(e_2)$	$Var_2(e_2)$	...	$Var_n(e_2)$	d_2
...	...	...	...	...	...	...	...
e_p	(x_{max},y_{max})	t_{tmax}	$Var_1(e_p)$	$Var_2(e_p)$	...	$Var_n(e_p)$	d_p

Devido à quantidade excessiva de dados e às dificuldades de se construir classificadores/preditores válidos para toda a área do território brasileiro coberta pela rede RINDAT, optou-se por construí-los para três minirregiões específicas, definidas a seguir.

A. **Pantanal Sul Matogrossense**:
    - Latitudes: 18,4° S a 19,4° S;
    - Longitudes: 56,4° O a 57,4° O.
B. **Alta Sorocabana Paulista**:
    - Latitudes: 21,4° S a 22,4° S;
    - Longitudes: 49,4° O a 50,4° O.
C. **Vale do Paraíba e Litoral Norte Paulista**:
    - Latitudes: 23° S a 24° S;
    - Longitudes: 45° O a 46° O.

Essas três minirregiões são mostradas são mostradas na Figura 4.1.

Figura 4.1 – Minirregiões A, B e C consideradas neste estudo.

Cada uma das minirregiões consideradas possui 42.588 registros na base de dados, referentes a 36 pontos de grade, em uma área quadrada de 1°x1° e a 1183 instantes de tempo, correspondentes às 4 saídas diárias do modelo ETA20 nos horários sinóticos para os meses de janeiro e fevereiro de 2007 a 2011, num total de 296 dias. Apenas os registros referentes à saída das 06 UTC de 25/01/2011 estavam indisponíveis. A distribuição de classes de

atividade convectiva com base na densidade de ocorrência de descargas NS (ausente/fraca, moderada e forte) apresentada na Figura 2.2.

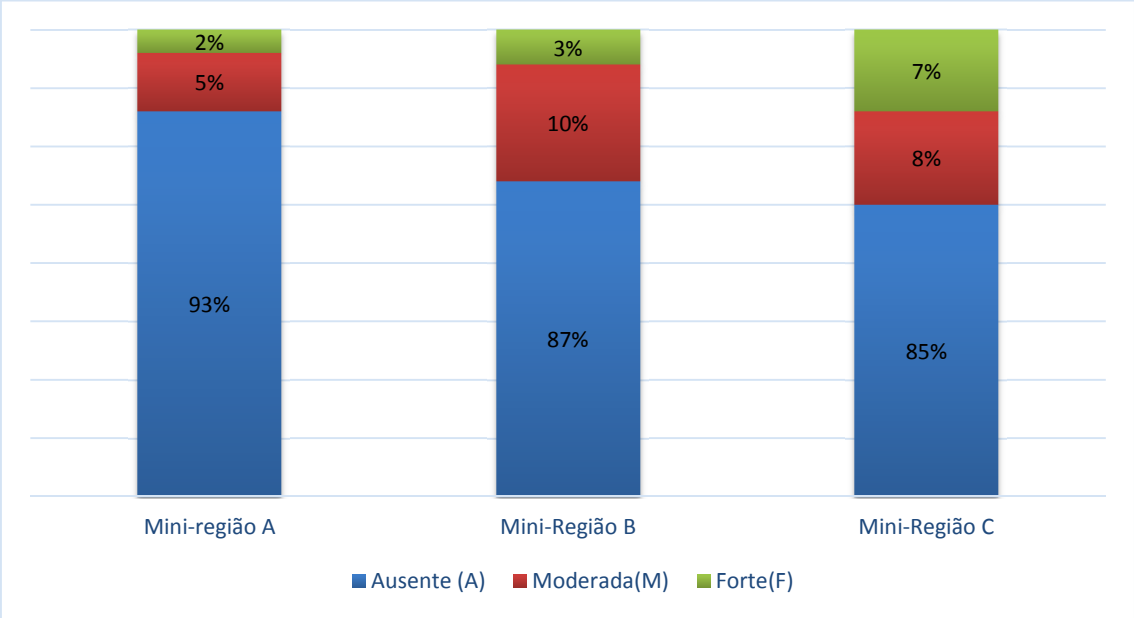

Figura 4.2 - Distribuição de classes de atividade convectiva em porcentagem para as minirregiões A, B e C

Na Tabela 4.2 são mostradas os intervalos de valores da densidade normalizada de descargas NS para cada classe de atividade convectiva. Essa densidade é utilizada como atributo de decisão na base de dados meteorológicos, ou seja, permite atribuir/rotular a classe de cada instância do conjunto de dados de treinamento.

Tabela 4.2. Intervalo de valores considerados da densidade de descargas NS para cada classe de atividade convectiva, sendo a densidade D expressa em valores normalizados no intervalo [0,1].

Densidade de descargas NS (D)	Classe
0,0025 ≤ D	Ausente/Fraca (A)
0,0025< D ≤ 0,01	Moderada (M)
0,01 < D	Forte (F)

Na Seção 4.1 são apresentados os dados gerados pelo modelo numérico de previsão do tempo ETA20, fornecidos pelo CPTEC/INPE, enquanto que na Seção 4.2, os dados de descargas elétricas atmosféricas NS, gerados pela rede RINDAT e utilizados pelo software EDDA para geração dos campos de densidade de ocorrências.

4.1 Modelo de mesoescala ETA

Os modelos numéricos de previsão do tempo são implementações computacionais que resolvem numericamente as equações físico-matemáticas que descrevem o comportamento da atmosfera. Em cada instante de tempo, o estado da atmosfera é representado por um conjunto de variáveis para os pontos de uma grade tridimensional. Modelos regionais, como no caso do empregado nesta tese, utilizam condições iniciais fornecidas por um modelo global e por dados observacionais incorporados por meio de assimilação de dados, processo no qual o estado corrente da atmosfera resultante de previsões anteriores é atualizado com base nesses dados. Modelos regionais usualmente são modelos utilizados na Meteorologia de mesoescala, que abrangem dimensões de 5 km até centenas de quilômetros, em oposição à escala sinótica, que abrange dimensões acima de 1000 km, ou à microescala, abaixo de 1 km.

O modelo de previsão do tempo utilizado nesta tese é o ETA20, com resolução espacial de 20 Km. O modelo ETA foi inspirado no seu antecessor HIBU, desenvolvido no Instituto Hidrometeorológico e Universidade de Belgrado, na Iugoslávia, durante a década de 1970. Nos anos 1980 seu código foi atualizado sendo uma das modificações a implementação das coordenadas verticais η (letra grega “eta”), resultando no modelo ETA, que se tornou operacional em 1993 no NCEP norte-americano (BLACK, 1994; MESINGER et al., 1988). Desde então, muitos países passaram a utilizar este modelo, inclusive o CPTEC/INPE, que implantou este modelo operacionalmente em 1996.

O modelo ETA20 é executado duas vezes ao dia no CPTEC/INPE, nos horários de 00 e 12 UTC, que correspondem às chamadas "análises" pois incorporam dados observacionais e dados de um modelo global por meio de assimilação de dados. As saídas do ETA20 são de 6 em 6 horas, correspondendo aos horários sinóticos das 00, 06, 12 e 18 horas (UTC). Neste estudo foram utilizadas previsões de 24, 48 e 72 horas. Para um determinado dia D, as saídas referentes às previsões de 24hs (00, 06, 12 e 18 UTC) correspondem à execução do modelo a 00 UTC do dia anterior (D-1), enquanto que as saídas para previsões de 48hs (00, 06, 12 e 18 UTC), à execução do modelo a 00 UTC do dia (D-2). E as saídas para previsões de 72hs (00, 06, 12 e 18 UTC), à execução do modelo a 00 UTC do dia (D-3).

No escopo dos projetos supracitados (Adapt e Cb-mining), meteorologistas selecionaram 15 variáveis do modelo ETA20, sendo 9 de superfície e outras 7 variáveis distribuídas em 7 níveis específicos (1000, 925, 850, 700, 500,300 e 250 hPa), num total de 58 variáveis.

Estas variáveis são utilizadas como atributos condicionais de um sistema de informação, cujo atributo de decisão é a densidade de descargas. Assim, espera-se associar valores específicos de um conjunto de variáveis do modelo ETA20 para classes distintas de densidade de ocorrência de descargas NS. Estas classes são representativas de atividade convectiva ausente/fraca, moderada e forte, segundo a hipótese de que essa densidade esteja correlacionadas à atividade convectiva.

O conjunto de variáveis do modelo assume valores característicos para cada classe de atividade convectiva, compondo os padrões indicativos de cada classe. É importante enfatizar que o conjunto de variáveis utilizado deriva dos dados de previsão do modelo ETA20. Assim, espera-se que estes padrões permitam um “aprendizado” do viés do modelo.

A Tabela 4.3 traz uma breve descrição das variáveis selecionadas do ETA20, com suas unidades e seus níveis (superfície ou níveis de pressão em hPa), usadas neste estudo.

Tabela 4.3. Variáveis selecionadas do Modelo ETA20.

Variável	Descrição	Nível
PSLM	Pressão Média ao Nível do Mar [hPa]	Superfície
PSLC	Pressão na Superfície [hPa]	Superfície
TP2m	Temperatura	2 metros
DP2m	Temperatura do Ponto de Orvalho [K]	2 metros
CAPE	Energia Potencial Convectiva Disponível [m^2/s^2]	Superfície
CINE	Energia Inibitória Convectiva [m^2/s^2]	Superfície
BLI	Best Lifted Index [K]	Superfície
FZHT	Altura de congelamento [m]	Superfície
AGPL	Água Precipitável Instantânea [kg/m^2]	Superfície
U	Vento Zonal [m/s]	1000, 925, 850, 700, 500,300 e 250 [hPa]
V	Vento Meridional [m/s]	1000, 925, 850, 700, 500,300 e 250 [hPa]
Z	Altura Geopotencial [gpm]	1000, 925, 850, 700, 500,300 e 250 [hPa]
TABS	Temperatura Absoluta [K]	1000, 925, 850, 700, 500,300 e 250 [hPa]
OMEGA	Omega [Pa/s]	1000, 925, 850, 700, 500,300 e 250 [hPa]
UREL	Unidade Relativa [ ]	1000, 925, 850, 700, 500,300 e 250 [hPa]
UMES	Umidade Especifica [kg/kg]	1000, 925, 850, 700, 500,300 e 250 [hPa]

4.3 Dados de descargas elétricas atmosféricas

A descarga elétrica atmosférica é um fenômeno atmosférico complexo caracterizado por um intenso fluxo de corrente de curta duração, que ocorre na atmosfera e em alguns casos, atinge a superfície da Terra. A principal causa das descargas são as nuvens de tempestade originadas de atividade convectiva intensa, as quais são conhecidas como *Cumulus Nimbus* (Cb). Essas nuvens são formadas devido à convecção caracterizada por fortes correntes ascendentes de ar quente e úmido, ou então, por frentes frias que se deslocam abrangendo grandes altitudes, com intensos fluxos descendentes. Os Cb's podem ter de 1 a 20 km de diâmetro, e suas alturas de base de nuvem variam de 3 km, para as mais próximas do equador, a 1 km, para as mais distantes. Seus topos podem alcançar até 20 Km de altitude (FILHO, 2005).

As descargas podem ser tipificadas de três modos: (i) segundo o percurso da descargas, (ii) pela direção de propagação da descarga e (iii) pelo sinal da descarga.

Segundo seu percurso as descargas atmosféricas podem ser de três tipos:

- IntraNuvem (IN): ocorrem internamente às nuvens, sendo o tipo mais frequente.
- Nuvem-Nuvem (NN): ocorrem entre centros de cargas negativas e positivas localizados em nuvens diferentes.
- Nuvem-Solo (NS): ocorrem entre a nuvem e o solo, representando cerca de 25% das descargas totais.

Quanto à direção, as descargas NS podem ser classificadas como:

- Descendentes: são descargas que o fluxo de elétrons se origina na nuvem e é descarregado para o solo, constituindo a maior parte das descargas NS.

- Ascendentes: são descargas que o fluxo de elétrons se origina no solo e é descarregado para a nuvem, sendo menos frequentes.

Por fim, as descargas NS também podem ser classificadas de acordo com o sinal da carga ou polaridade que é transferida para o solo (embora essa classificação possa ser estendida aos demais tipos de descargas):

- Positivas: quando a nuvem está carregada positivamente, sendo neutralizada por uma descarga em que há um fluxo ascendente de elétrons.

- Negativas: quando a nuvem está carregada negativamente, sendo neutralizada por uma descarga em que há um fluxo descendente de elétrons.

Nesta tese são usados dados brutos de descargas fornecidos pela rede RINDAT e posteriormente processados pelo EDDA, para geração dos campos de densidade de ocorrência de descargas elétricas NS. Embora a rede RINDAT permita a detecção de descargas do tipo NS e NN, são apenas consideradas neste estudo as descargas NS, que melhor se correlacionam com eventos convectivos.

A rede RINDAT surgiu em 2001 graças a uma cooperação entre o SIMEPAR (Sistema Meteorológico do Paraná), CEMIG (Companhia Energética de Minas Gerais) e Furnas. Em 2003, o INPE passou a integrar essa rede, incorporando seus próprios sensores. A RINDAT conta com dois tipos de sensores. O (i) sensor LPATS (*Lightning Position and Tracking System*), que utiliza a diferença dos tempos de chegada do pulso eletromagnético causado pela descarga aos sensores (mínimo de três sensores) para estimar a posição da descarga, na técnica conhecida em inglês como *Time of Arrival* (ToA). E o (ii) sensor IMPACT (*Improved Performance Combined Technology*), que além da técnica ToA utiliza a técnica de Indicação de Direção Magnética (em inglês, *Magnetic Direction Finder* ou MDF), que mede os ângulos das componentes norte-sul e

leste-oeste do campo magnético gerado pela descarga utilizando duas antenas circulares (em inglês, *loop antennas*). Sensores IMPACT possibilitam uma melhor detecção de descargas devido à melhor precisão de localização e maior probabilidade de detecção (FILHO, 2005). A Figura 4.3 mostra a distribuição geográfica dos sensores da rede RINDAT:

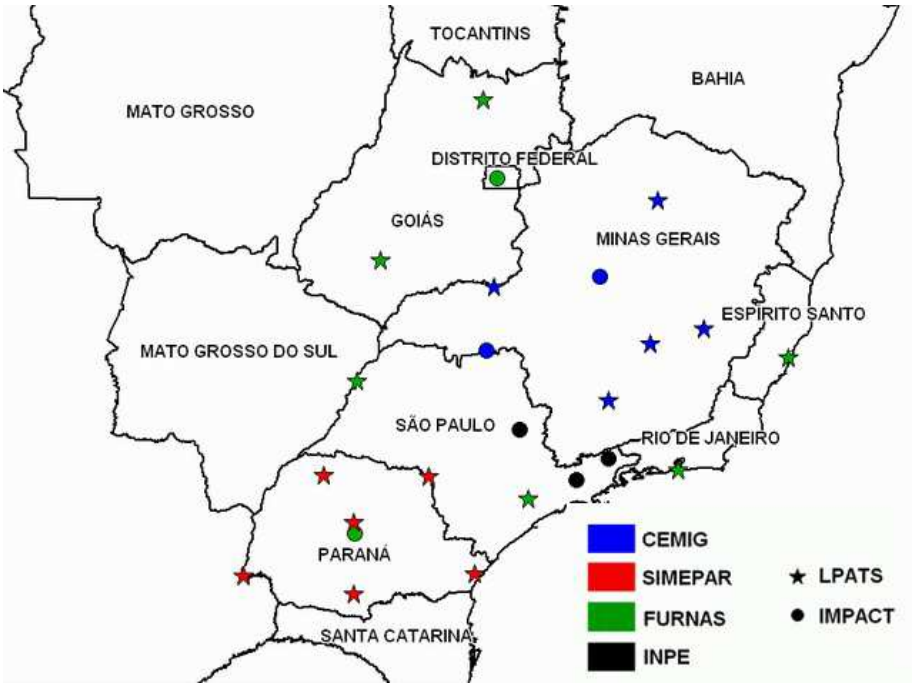

Figura 4.3 – Localização dos sensores da rede RINDAT.
Fonte: RINDAT (2014)

Os dados brutos são fornecidos pela rede RINDAT no formato ASCII denominado UALF (*Universal ASCII Lightning Format*). Os correspondentes dados de descargas NS são processados por uma ferramenta de estimação de densidade de ocorrência de descargas, o software EDDA, desenvolvido por Strauss et al. (2013). Esse software utiliza estimação de núcleo (*kernel estimator*), em cada ponto de grade (x, y), para integrar as ocorrências de descargas NS para a área e intervalo de tempo selecionados, de forma a gerar o correspondente campo de densidade.

Os dados da rede RINDAT possuem uma resolução temporal da ordem de milissegundos, sendo fornecidos em arquivos contendo 5 minutos de ocorrências, aproximadamente. Neste trabalho o software EDDA integra esses dados para períodos de 1 hora, antecedentes aos horários das correspondentes saídas do modelo ETA20.

A densidade de ocorrências de descargas NS é utilizada como atributo de decisão indicativo de atividade convectiva na classificação/predição em que os atributos condicionais são as variáveis selecionadas do modelo ETA20 para as previsões e análises cujas saídas ocorrem de 6 em 6 horas, correspondentes aos horários sinóticos de 00, 06, 12 e 18 UTC. Assim, para cada hora sinótica, associa-se as descargas NS ocorridas na hora anterior, ou seja, nos intervalos 23-00, 05-06, 11-12 e 17-18 UTC.

No próximo capitulo serão mostrados os resultados obtidos pela aplicação das metaheurísticas propostas, para o cálculo de reduções em TCA. Primeiramente, serão abordados os resultados pertinentes à aplicação das metaheurísticas em base de dados de uso geral. Em seguida, serão apresentados os resultados referentes à aplicação da TCA nos dados meteorológicos, no escopo de predição da atividade convectiva.

5 RESULTADOS

Neste capítulo são exibidos os resultados referentes à aplicação das metaheurísticas apresentadas no cálculo de reduções em TCA para fins de classificação.

Foram realizados testes com bases de dados de uso geral (PESSOA; STEPHANY, 2014), apresentados na Seção 5.1, e testes relativos à base de dados meteorológicos, sendo relativos à predição de eventos convectivos, objeto principal desta tese e apresentados na Seção 5.2.

As metaheurísticas propostas foram VNS, VND, ILS e DCS, esta última desenvolvida neste trabalho. A metaheurística VND não deu bons resultados, sendo usada unicamente como algoritmo alternativo de busca local.

Exceto pela VND, que utiliza um esquema próprio de busca local, as demais três metaheurísticas utilizam um esquema de Busca Local Padrão (SLS, Seção 3.2), denotada por s. Conforme mencionado, as mesmas foram também testadas utilizando como alternativa de busca local a própria metaheurística VND, denotada por v.

Além disso, para cada metaheurística, deve-se especificar qual a função objetivo empregada (função de dependência de atributos γ ou então a dependência relativa κ). Assim, uma dada metaheurística é denotada por seu nome, a função objetivo e o esquema de busca local. Como exemplo, a identificação VNS(γ)-s é referente ao uso da metaheurística VNS com a função de dependência de atributos γ e o algoritmo de Busca Local Padrão. Na Tabela 5.1 são mostradas todas as possíveis variações de algoritmos usados neste trabalho.

Tabela 5.1. Descrição das 16 variações possíveis para as metaheurísticas empregadas segundo opções de cardinalidade de estrutura de vizinhança (L), função objetivo e de esquema de busca local.

Metaheurística	*L*	Função objetivo	Busca Local
VNS	4	γ	s
		γ	v
		κ	s
		κ	v
	8	γ	s
		γ	v
		κ	s
		κ	v
ILS	-	γ	s
		γ	v
		κ	s
		κ	v
DCS	-	γ	s
		γ	v
		κ	s
		κ	v

No caso dos dados meteorológicos (Seção 5.2), foi adicionalmente empregado o Particionamento Aleatório do Conjunto de Treinamento (PACT), sendo adotadas 1, 8, 16 e 32 partições para os testes realizados.

A avaliação das metaheurísticas propostas para cálculo de reduções em TCA pode ser feita verificando-se o desempenho de classificação dos correspondentes classificadores para algumas bases de dados ou, então, contabilizando-se a frequência e a cardinalidade das reduções encontradas. No primeiro caso, utilizam-se duas métricas bem conhecidas na literatura de classificação: a acurácia (*Ac*) e o índice Kappa de Cohen (K). No segundo caso,

uitliza-se uma métrica proposta em Pessoa e Stephany (2014), denominada *skill score* (*S*), baseada no número de atributos de cada redução obtida.

A acurácia e o índice Kappa de Cohen são métricas calculadas a partir da matriz de confusão, amplamente empregada em classificação. Dada a matriz de confusão $M = [A(i,j)]$, com $i=j=1..|V_d|$, cada elemento $A(i,j)$ representa o número de instâncias que pertencem a classe *i* e foram classificadas como pertencentes à classe *j* (THEODORIDIS; KOUTROUMBAS, 2009). Assim, os elementos da diagonal da matriz correspondem a instâncias corretamente classificadas, enquanto que os elementos fora da diagonal correspondem a falsos positivos e falsos negativos. A matriz de confusão é obviamente uma matriz quadrada. Considerando-se a matriz de confusão *A* e a classificação de *N* instâncias do conjunto de testes, a acurácia é definida por:

$$Ac = \frac{1}{N}\sum_{i=1}^{|V_d|} A(i,i) \tag{5.1}$$

Pode-se definir também a acurácia do produtor para uma classe *i*, denominando-se TL_i como sendo o total de elementos da linha *i*:

$$Ac_i = \frac{A(i,i)}{TL_i} \tag{5.2}$$

O índice de concordância Kappa de Cohen pode ser definido em função da matriz de confusão *A* para *N* instâncias classificadas como sendo (COHEN, 1960; LIMA; STEPHANY, 2013b):

$$\mathrm{K} = \frac{N\sum_{i=1}^{|Vd|} A(i,i) - \sum_{i=1}^{|Vd|}(TC_i \times TL_i)}{N^2 - \sum_{i=1}^{|Vd|}(TC_i - TL_i)} \tag{5.3}$$

Na definição acima, TC_i é a soma da *i*-ésima coluna e TL_i é a soma da *i*-ésima linha da matriz de confusão *A*, enquanto que $|V_d|$ é o número de classes possíveis do atributo de decisão, que corresponde à dimensão da matriz. Valores do índice Kappa entre 0,60 e 0.79 indicariam uma classificação razoável, enquanto que valores acima dessa faixa, uma boa classificação (LANDIS; KOCH, 1977).

Em Pessoa e Stephany (2014), foi proposta uma nova métrica denominada *skill score* (*S*) para avaliar as reduções obtidas pelas diversas metaheurísticas. Essa métrica atribui um valor maior às reduções obtidas com menor cardinalidade tomando como referência a menor cardinalidade conhecida. Esse valor também depende da frequência com que as reduções são obtidas, considerando múltiplas execuções com uma mesma heurística, privilegiando as repetidas vezes com que uma redução de baixa cardinalidade é obtida. Dada uma base de dados, o *skill score S* para *T* execuções de uma metaheurística específica, é definido em função das *M* reduções diferentes obtidas como:

$$S = \frac{C_{\min}}{T} \sum_{i=1}^{M} \frac{Q_i}{C_i} \tag{5.4}$$

onde C_{min} é a cardinalidade da melhor redução conhecida obtida por qualquer metaheurística para a base de dados considerada, C_i denota a cardinalidade da *i*-ésima redução obtida pela metaheurística considerada e Q_i denota o número de vezes que essa cardinalidade foi obtida. Como exemplo, se 8 for a melhor cardinalidade obtida por qualquer metaheurística, e se a metaheurística considerada for executada 20 vezes obtendo 9 reduções de cardinalidade 8 ($8^{(9)}$) e 11 reduções de cardinalidade 9 ($9^{(11)}$), denotados por $8^{(9)}9^{(11)}$, o *skill score* será calculado como (8/20)×[(9/8)+(11/9)] ou 0,9389 (note que neste caso *M*=2).

O cálculo de reduções foi executado 10 (base de dados meteorológicos) ou 20 vezes (base de dados gerais) para cada variação de metaheurística. O número

máximo de iterações, aplicável para as metaheurísticas ILS, VNS e DCS, foi fixado em 10. No caso da metaheurística VND, utilizada apenas para busca local, adotou-se uma estrutura de vizinhança com cardinalidade máxima de 2. A Tabela 5.2 apresenta as características das máquinas utilizadas nesta tese, sendo que os testes com bases de dados gerais (Seção 5.1) foram executados 20 vezes para cada caso, apenas nas máquinas "Jupiter" e "Matrix" do LAC/INPE, destinadas aos projetos supracitados e que têm desempenhos computacionais bastante similares. Os testes com a base de dados meteorológicos demandaram muito mais processamento, sendo executados 10 vezes por caso em todas as máquinas descritas na tabela.

Tabela 5.2. Descrição das máquina utilizadas.

Nome	Processador	Freq. (GHz)	Núcleos/ Pipelines	Memória (GB)	Sistema Operacional
c3.4xlarge *	Intel Xeon E5-1670 v2	3.60	4/16	32	Amazon Linux AMI 2014.09 (kernel 3.14.19-17.43)
c3.8xlarge *	Intel Xeon E5-1670 v2	3.60	16/32	64	Amazon Linux AMI 2014.09 (kernel 3.14.19-17.43)
c3.8xlarge *	Intel Xeon E5-1670 v2	3.60	16/32	64	Amazon Linux AMI 2014.09 (kernel 3.14.19-17.43)
Matrix	Intel Xeon X5670	2.93	12/24	32	Ubuntu 12.04.4 LTS (kernel 3.2.0-52)
Jupiter	Intel Xeon E5530	2.40	8/16	24	Ubuntu 12.04.5 LTS (kernel 3.2.0-48)
Saturno	AMD Phenom II X4 965	3.4	4	16	openSUSE 13.1 (kernel 3.11.10-7)
Rigel	Intel Core i7 4500U	1.8	2/4	16	openSUSE 13.1 (kernel 3.11.10-7)

*- Amazon Elastic Compute Cloud (EC2)

As seções seguintes apresentam os resultados obtidos pela aplicação das metaheurísticas propostas e suas variações, conforme descrito na Tabela 5.1, para cálculo de reduções com as bases de dados gerais (Seção 5.1) e com a base de dados meteorológicos (Seção 5.2).

5.1 Resultados – bases de dados de uso geral

Nesta seção são apresentados os resultados relativos à aplicação das metaheurísticas VNS, ILS e DCS, para o cálculo de reduções em TCA para bases de dados de uso geral originarias de diversas áreas tais como medicina, biologia, economia, eletrônica e política. Os resultados foram publicados em Pessoa e Stephany (2014). Essas bases incluem dados reais e sintéticos, sendo usadas em trabalhos para validar metodologias de análise de dados. Estes dados estão disponíveis na UCI *Machile Learning Repository* (BACHE; LICHMAN, 2013; HEDAR et al., 2008; JENSEN; SHEN, 2003). A Tabela 5.3 ilustra detalhes das 13 bases de dados utilizadas, tais como a cardinalidade $|A|$ (número de atributos), o número de instâncias/elementos $|U|$ e o número de classes $|V_d|$.

Tabela 5.3. Descrição das 13 bases de dados de uso geral utilizadas.

Conjuntos de dados	$\|A\|$	$\|U\|$	$\|V_d\|$
M-of-N	13	1000	2
Exactly	13	1000	2
Exactly2	13	1000	2
Heart	13	294	2
Vote	16	300	2
Credit	20	1000	2
Mushroom	22	8124	2
LED	24	2000	10
Letters	25	26	26
Derm	34	366	6
Derm2	34	358	6
WQ	38	521	13
Lung	56	32	3

Segue-se uma breve descrição de cada base de dados.

A. M-of-N: constituída de dados sintéticos, possui uma distribuição uniforme e seus atributos condicionais e de decisão, tem valores binários.

B. Exactly e Exactly2: constituída de dados sintéticos, caracteriza-se pela dificuldade de treinamento ao serem utilizados algoritmos de aprendizado de máquina, devido a suas classes não serem linearmente separáveis.

C. Heart: constituída de dados referentes ao diagnóstico de doenças cardíacas, sendo os dados coletados em quatro hospitais:

1. Cleveland Clinic Foundation, Estados Unidos;
2. Hungarian Institute of Cardiology, Budapeste, Hungria;
3. V.A. Medical Center, Long Beach, CA; Estados Unidos;
4. University Hospital, Zurique, Suíça.

D. Vote: muito empregada, seus dados correspondem aos votos de congressistas americanos para o ano de 1984; seus 16 atributos condicionais correspondem a 16 posições-chave ligadas a interesses específicos listadas pelo CQA (*Congressional Quarterly Almanac*), enquanto que o atributo de decisão é o partido do congressista (republicano ou democrata).

E. Credit: destinada a avaliação do cliente que solicita de crédito, possui alguns atributos numéricos e outros categóricos.

F. Mushroom: referente a 23 espécies de cogumelos que devem ser identificados como da família *Agaricus* ou *Lepiota*.

G. LED: base de dados sintética com domínio relativo a um mostrador digital; embora a base tenha 20 atributos, somente 7 deles são relevantes, pois correspondem aos diodos emissores de luz.

H. Letters: base de dados sintética destinada a identificar as 25 letras do alfabeto.

I. Derm e Derm2: ambas são bases de dados reais relativas ao diagnóstico de doenças Eritêmato-Escamosas.

J. WQ: base de dados reais relativa ao diagnóstico da qualidade da água de uma estação de tratamento.

L. Lung: base de dados reais relativa ao diagnóstico de tipos de câncer de pulmão.

Os testes relativos ao cálculo de reduções para essas 13 bases de dados utilizaram as 3 metaheurísticas (VNS/ILS/DCS), cada uma com duas funções de avaliação (função de dependência e função de dependência relativa), dois esquemas de busca local (Busca Local Padrão e VND), sendo cada variação executada 20 vezes, repetição desejável por se tratarem de algoritmos estocásticos. Adicionalmente, as variações da metaheurística VNS foram testadas com distância de *Hamming* L=4 ou L=8, como ilustrado adiante.

Os resultados aqui apresentados foram comparados por resultados previamente obtidos por outros autores, também no escopo de redução de atributos em TCA para as 13 bases de dados citadas. Esses resultados, apresentados na Tabela 5.4 usados como referência nesta seção, foram obtidos por Hedar et al. (2008), que propôs o uso da metaheurística Busca Tabu (em inglês, *Tabu Search* ou TS) como alternativa a metaheurísticas anteriormente propostas por Jensen e Shen (2005), que incluíam Otimização por Colônia de Formigas (ACO), Recozimento Simulado (SA) e Algoritmo Genético (AG). Os resultados foram obtidos com 20 execuções, exceto no SA, para os conjuntos de dados Heart, Vote e Derm2, que executaram 30, 30 e 10 vezes, respectivamente, e para o AG no conjunto de dados M-of-N, que foi executado 18 vezes.

Na Tabela 5.4, bem como nas seguintes, $|A|$ denota a cardinalidade original de cada base de dados, enquanto que os demais números representam as cardinalidades das reduções obtidas, sendo que os números sobrescritos entre parênteses expressam o número de vezes em que cada cardinalidade foi obtida nas 20 execuções de cada experimento. Quando esse número sobrescrito é omitido, significa que a cardinalidade foi a mesma nas 20 execuções.

Tabela 5.4. Cardinalidade das reduções obtidas para as 13 bases de dados consideradas pelas diversas metaheurísticas propostas anteriormente para cálculo de reduções em TCA, ou seja, ACO, AS, GA e TS.

Datasets	|A|	ACO	SA	GA	TS
M-of-N	13	6	6	$6^{(6)}$ $7^{(12)}$	6
Exactly	13	6	6	$6^{(10)}$ $7^{(10)}$	6
Exactly2	13	10	10	$10^{(9)}$ $11^{(11)}$	10
Heart	13	$6^{(18)}$ $7^{(2)}$	$6^{(29)}$ $7^{(1)}$	$6^{(18)}$ $7^{(2)}$	6
Vote	16	8	$8^{(15)}$ $9^{(15)}$	$8^{(2)}$ $9^{(18)}$	8
Credit	20	$8^{(12)}$ $9^{(4)}$ $10^{(4)}$	$8^{(18)}$ $9^{(1)}$ $11^{(1)}$	$10^{(6)}$ $11^{(14)}$	$8^{(13)}$ $9^{(5)}$ $10^{(2)}$
Mushroom	22	4	4	$5^{(1)}$ $6^{(5)}$ $7^{(14)}$	$4^{(17)}$ $5^{(3)}$
LED	24	$5^{(12)}$ $6^{(4)}$ $7^{(3)}$	5	$6^{(1)}$ $7^{(3)}$ $8^{(16)}$	5
Letters	25	8	8	$8^{(8)}$ $9^{(12)}$	$8^{(17)}$ $9^{(3)}$
Derm	34	$6^{(17)}$ $7^{(3)}$	$6^{(12)}$ $7^{(8)}$	$10^{(6)}$ $11^{(14)}$	$6^{(14)}$ $7^{(6)}$
Derm2	34	$8^{(3)}$ $9^{(17)}$	$8^{(3)}$ $9^{(7)}$	$10^{(4)}$ $11^{(16)}$	$8^{(2)}$ $9^{(14)}$ $10^{(4)}$
WQ	38	$12^{(2)}$ $13^{(7)}$ $14^{(11)}$	$13^{(16)}$ $14^{(4)}$	16	$12^{(1)}$ $13^{(13)}$ $14^{(6)}$
Lung	56	4	$4^{(7)}$ $5^{(12)}$ $6^{(1)}$	$6^{(8)}$ $7^{(12)}$	$4^{(6)}$ $5^{(13)}$ $6^{(1)}$

Fonte: Hedar et al. (2008).

Na Tabela 5.5 são mostradas as cardinalidades das reduções obtidas para as variações da metaheurística VNS com estrutura de vizinhança L=4, ou seja, vizinhos que apresentam distância de *Hamming* correspondente à troca de até 4 bits em relação à solução corrente considerada.

Tabela 5.5. Cardinalidade das reduções obtidas para as 13 bases de dados consideradas pelas variações da metaheurística VNS com distância de *Hamming* L = 4.

Datasets	**|A|**	**VNS(γ)-s**	**VNS(κ)-s**	**VNS(γ)-v**	**VNS(κ)-v**
M-of-N	13	6	6	6	6
Exactly	13	6	6	6	6
Exactly2	13	10	10	10	10
Heart	13	$6^{(19)}7^{(1)}$	6	6	6
Vote	16	8	8	8	8
Credit	20	$8^{(15)}9^{(4)}10^{(1)}$	$8^{(16)}9^{(4)}$	8	8
Mushroom	22	3	3	3	3
LED	24	5	5	5	5
Letters	25	$8^{(13)}9^{(7)}$	$8^{(16)}9^{(4)}$	$8^{(19)}9^{(1)}$	8
Derm	34	$6^{(16)}7^{(4)}$	$6^{(13)}7^{(7)}$	$6^{(17)}7^{(3)}$	$6^{(18)}7^{(2)}$
Derm2	34	$9^{(14)}10^{(6)}$	$9^{(19)}10^{(1)}$	$8^{(1)}9^{(18)}10^{(1)}$	$8^{(2)}9^{(18)}$
WQ	38	$12^{(5)}13^{(15)}$	$12^{(6)}13^{(14)}$	$12^{(9)}13^{(11)}$	$12^{(15)}13^{(5)}$
Lung	56	$3^{(9)}4^{(10)}5^{(1)}$	$3^{(6)}4^{(12)}5^{(2)}$	$3^{(14)}4^{(6)}$	$3^{(12)}4^{(8)}$

As cardinalidades das reduções obtidas para as variações da metaheurística VNS com estrutura de vizinhança L=8 são mostrados na Tabela 5.6.

Tabela 5.6. Cardinalidade das reduções obtidas para as 13 bases de dados consideradas pelas variações da metaheurística VNS com distância de *Hamming* L=8.

Datasets	**|A|**	**VNS(γ)-s**	**VNS(κ)-s**	**VNS(γ)-v**	**VNS(κ)-v**
M-of-N	13	6	6	6	6
Exactly	13	6	6	6	6
Exactly2	13	10	10	10	10
Heart	13	6	6	6	6
Vote	16	8	8	8	8
Credit	20	8	8	8	8
Mushroom	22	3	3	3	3
LED	24	5	5	5	5
Letters	25	8	8	8	8
Derm	34	$6^{(18)}7^{(2)}$	$6^{(19)}7^{(1)}$	6	6
Derm2	34	$8^{(1)}9^{(19)}$	$8^{(1)}9^{(18)}10^{(1)}$	$8^{(1)}9^{(19)}$	$8^{(4)}9^{(16)}$
WQ	38	$12^{(13)}13^{(7)}$	$12^{(12)}13^{(8)}$	$12^{(19)}13^{(1)}$	$12^{(18)}13^{(2)}$
Lung	56	$3^{(14)}4^{(5)}5^{(1)}$	$3^{(13)}4^{(7)}$	$3^{(18)}4^{(2)}$	$3^{(18)}4^{(2)}$

Na Tabela 5.7 são exibidas as cardinalidades das reduções obtidas para as variações da metaheurística ILS (note-se que o parâmetro L, correspondente à distância de *Hamming*, não se aplica).

Tabela 5.7. Cardinalidade das reduções obtidas para as 13 bases de dados consideradas pelas variações da metaheurística ILS.

Datasets	**|A|**	**ILS(γ)-s**	**ILS(κ)-s**	**ILS(γ)-v**	**ILS(κ)-v**
M-of-N	13	6	6	6	6
Exactly	13	6	6	6	6
Exactly2	13	10	10	10	10
Heart	13	$6^{(12)}7^{(8)}$	$6^{(12)}7^{(8)}$	6	6
Vote	16	8	8	8	8
Credit	20	$8^{(7)}9^{(2)}10^{(8)}11^{(3)}$	$8^{(7)}9^{(2)}10^{(8)}11^{(3)}$	$8^{(17)}9^{(2)}10^{(1)}$	$8^{(14)}9^{(3)}10^{(3)}$
Mushroom	22	$3^{(19)}4^{(1)}$	$3^{(19)}4^{(1)}$	3	3
LED	24	$5^{(5)}7^{(4)}8^{(11)}$	$5^{(7)}7^{(4)}8^{(9)}$	5	5
Letters	25	$8^{(4)}9^{(16)}$	$8^{(4)}9^{(16}$	$8^{(15)}9^{(4)}$	$8^{(19)}9^{(1)}$
Derm	34	$6^{(1)}7^{(19)}$	$7^{(16)}8^{(4)}$	$6^{(15)}7^{(5)}$	$6^{(15)}7^{(5)}$
Derm2	34	$9^{(1)}10^{(12)}11^{(3)}12^{(4)}$	$9^{(1)}10^{(13)}11^{(4)}12^{(2)}$	$8^{(2)}9^{(14)}10^{(4)}$	$9^{(19)}10^{(1)}$
WQ	38	$12^{(1)}13^{(17)}14^{(2)}$	$13^{(17)}14^{(3)}$	$12^{(3)}13^{(16)}14^{(1)}$	$12^{(8)}13^{(12)}$
Lung	56	$4^{(4)}5^{(16)}$	$4^{(3)}5^{(17)}$	$3^{(11)}4^{(6)}5^{(3)}$	$3^{(6)}4^{(8)}5^{(6)}$

As cardinalidades das reduções obtidas para as variações da metaheurística DCS são mostrados na Tabela 5.8 (note-se que o parâmetro L, correspondente à distância de *Hamming*, não se aplica).

Tabela 5.8. Cardinalidade das reduções obtidas para as 13 bases de dados consideradas pelas variações da metaheurística DCS.

Datasets	**|A|**	**DCS(γ)-s**	**DCS(κ)-s**	**DCS(γ)-v**	**DCS(κ)-v**
M-of-N	13	6	6	6	6
Exactly	13	6	6	6	6
Exactly2	13	10	10	10	10
Heart	13	6	6	6	6
Vote	16	8	8	8	8
Credit	20	$8^{(17)}9^{(3)}$	$8^{(17)}9^{(2)}10^{(1)}$	$8^{(19)}9^{(1)}$	8
Mushroom	22	3	3	3	3
LED	24	5	5	5	5
Letters	25	$8^{(6)}9^{(13)}10^{(1)}$	8	$8^{(12)}9^{(7)}10^{(1)}$	8
Derm	34	6	6	6	6
Derm2	34	$9^{(19)}10^{(1)}$	$9^{(4)}10^{(16)}$	$8^{(4)}9^{(16)}$	$8^{(3)}9^{(16)}10^{(1)}$
WQ	38	$12^{(6)}13^{(12)}$ $14^{(2)}$	$12^{(3)}13^{(17)}$	$12^{(6)}13^{(13)}$ $14^{(1)}$	$12^{(7)}13^{(13)}$
Lung	56	3	$3^{(19)}4^{(1)}$	3	$3^{(19)}4^{(1)}$

As Tabelas 5.5 a 5.8 mostram que as reduções obtidas são mais robustas em relação aos resultados anteriormente obtidos em Hedar et al. (2008). Vê-se que, de maneira geral, as cardinalidades das reduções obtidas aqui são de menor cardinalidade em relação àquelas apresentadas na Tabela 5.4 e, frequentemente, reduções com cardinalidades menores foram obtidas com frequência maior nas 20 execuções de cada variação de cada metaheurística. Vale ressaltar que nem sempre reduções que tem a mesma cardinalidade sejam iguais, pois podem ser constituídas de atributos condicionais diferentes.

A Figura 5.1 apresenta o número médio de iterações utilizadas no cálculo de reduções para as 20 execuções de cada variação das metaheurísticas ILS, DCS e VNS para cada uma das 13 bases de dados consideradas. No caso do VNS, conforme exposto no Algoritmo 3, cada iteração inclui buscas locais com estruturas de vizinhança de cardinalidades crescentes, ou seja, distâncias de *Hamming* crescentes de 1 a 4, ou de 1 a 8. Assim, uma iteração do VNS é mais complexa que as iterações do ILS ou DCS, incluindo, por exemplo 4 sub-iterações correspondentes a essa buscas locais. Consequentemente, para efeito de comparação, nessa figura considerou-se o número de iterações do VNS como sendo o número de suas correspondentes sub-iterações. Assim, o gráfico apresentado na Figura 5.1 permite a comparação de iterações de complexidade algorítmica similar para as diversas metaheurísticas. Os mesmos resultados são sumarizados na Figura 5.2, no qual aparece a média das iterações utilizadas no cálculo de reduções de todas as bases de dados para cada variação de metaheurística, sempre considerando as 20 execuções.

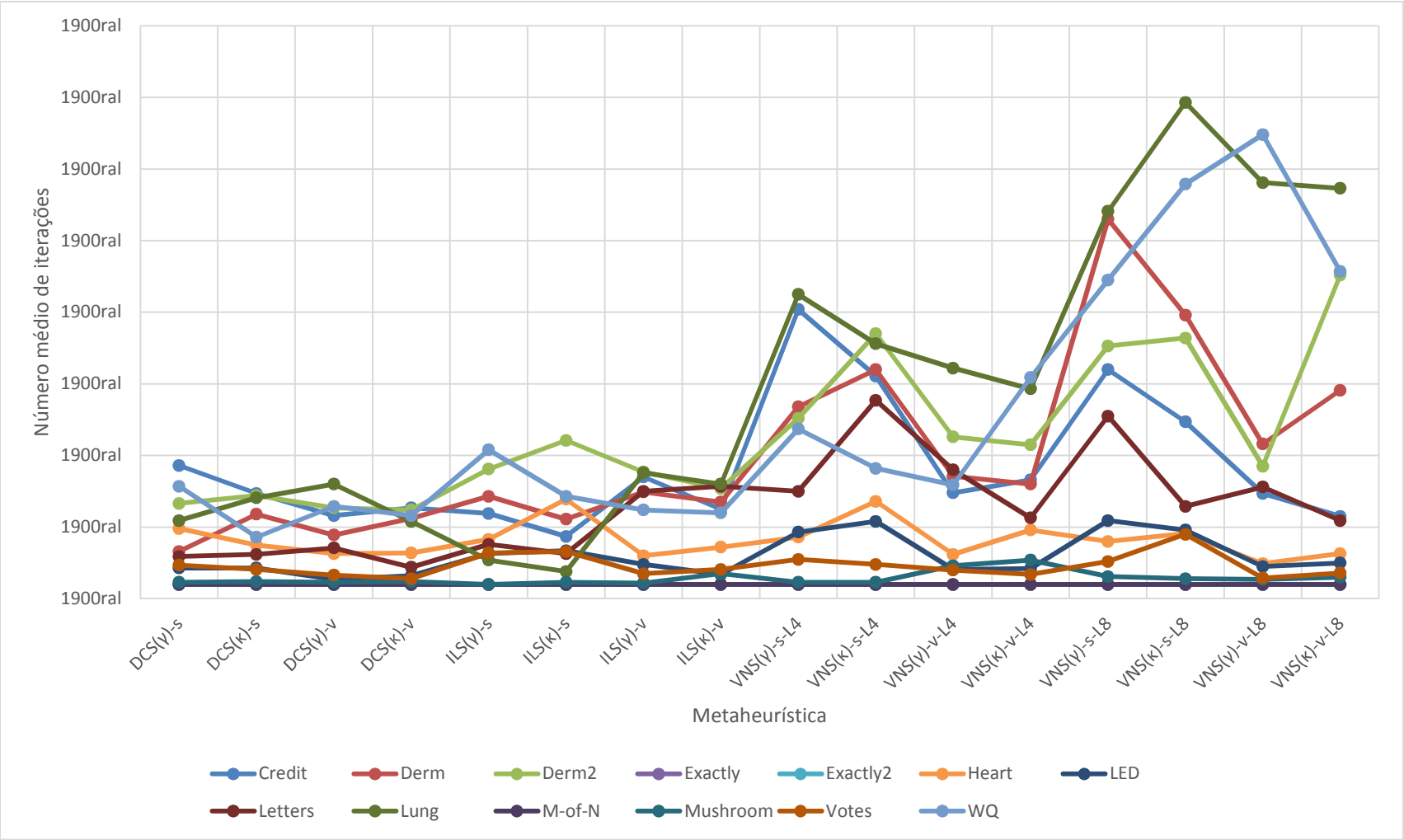

Figura 5.1 – Número médio de iterações para as 20 execuções de cada variação de metaheurística no cálculo de reduções para as 13 bases de dados consideradas.

As Figuras 5.1 e 5.2 mostram que as variações da metaheurística DCS demandaram menos iterações que a metaheurística ILS e esta demandou menos iterações que a metaheurística VNS, como seria de se esperar, uma vez que, a cada iteração, o VNS gera um vizinho da solução corrente e realiza buscas locais em estruturas de vizinhança com cardinalidade crescente. Por outro lado, a cada iteração, o ILS perturba a solução corrente e faz uma busca local em torno desta e eventualmente gera uma nova solução inicial reiniciando as iterações. Finalmente, o DCS, a cada iteração, gera uma nova solução candidata com cardinalidade menor, realizando uma busca mais agressiva no espaço de soluções/reduções. Entretanto, embora útil para dimensionar futuros testes, a comparação do número de iterações pode ser enganosa, pois em se tratando de algoritmos estocásticos, a convergência para uma solução melhor, ou seja, uma redução com menos atributos, pode variar.

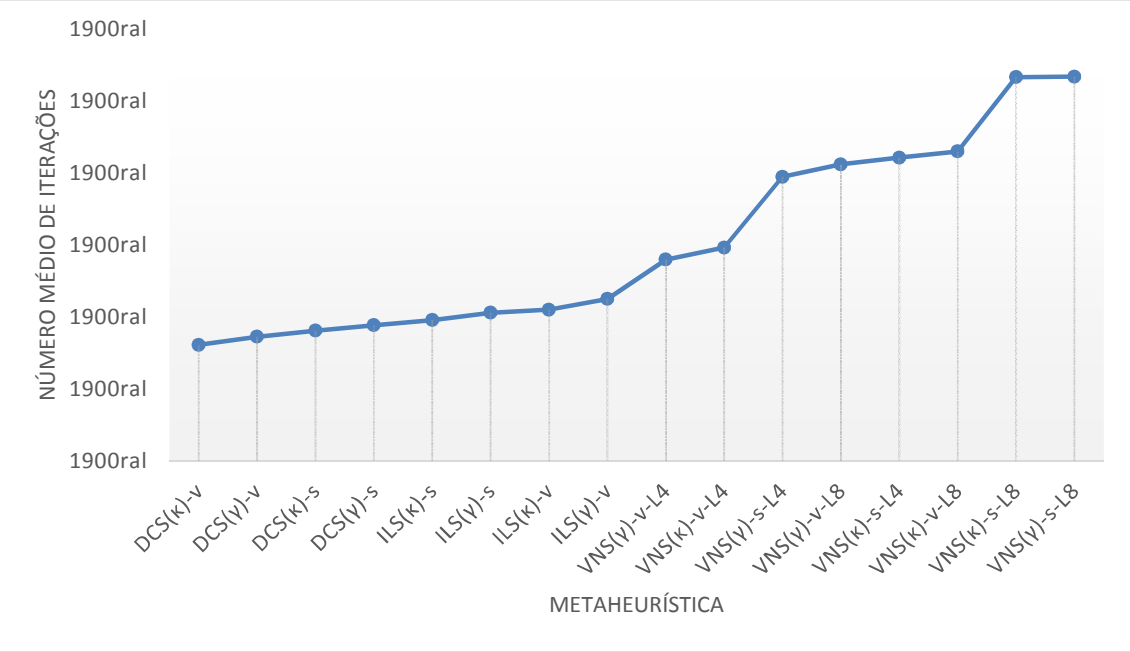

Figura 5.2 – Número médio de iterações demandado para o cálculo de reduções para as 20 execuções e para as 13 bases de dados consideradas de cada variação de metaheurística.

Assim, é interessante comparar os *skill scores* médios (Equação 5.4) e os tempos médios de processamento de cada variação de cada metaheurística, considerando-se as 20 execuções para todas as 13 bases de dados consideradas, conforme aparecem na Figura 5.3, na qual as variações das metaheurísticas estão ordenadas conforme tempos de processamento crescentes. Nota-se que, em média, as variações que utilizam como função objetivo a função de dependência de atributos (γ) demandaram muito mais tempo de processamento que as correspondentes variações que utilizaram a dependência relativa (κ), embora com *skill scores* médios similares. Nota-se, também em média, que as variações que utilizaram o VND como busca local (v) demandaram mais tempo de processamento, porém resultando em *skill scores* médios melhores que as correspondentes variações que utilizaram a busca local padrão (s). Isso se explica pela simplicidade da busca local padrão em relação à busca efetuada com VND.

Também de maneira geral, as variações do VNS obtiveram os melhores *skill scores* de todas as metaheurísticas, mas a custa de mais tempo de processamento. Obviamente, as variações do VNS com L=8 obtiveram resultados melhores que aquelas com L=4, mas também à custa de mais tempo de processamento. Apenas as variações do ILS com busca local padrão (s) obtiveram *skill scores* médios abaixo de 0.96, mostrando ser essa metaheurística sensível ao esquema de busca local. Exceto nesses casos, todas as metaheurísticas conseguiram *skill scores* médios iguais ou acima de 0.96, mas as variações do DCS com função objetivo dada pela dependência relativa (κ) obtiveram bons *skill scores* médios com tempos de processamento muito baixos. Embora algumas variações do VNS se mostrem muito competitivas, como a VNS(κ)-v, é preciso lembrar que esses são resultados médios.

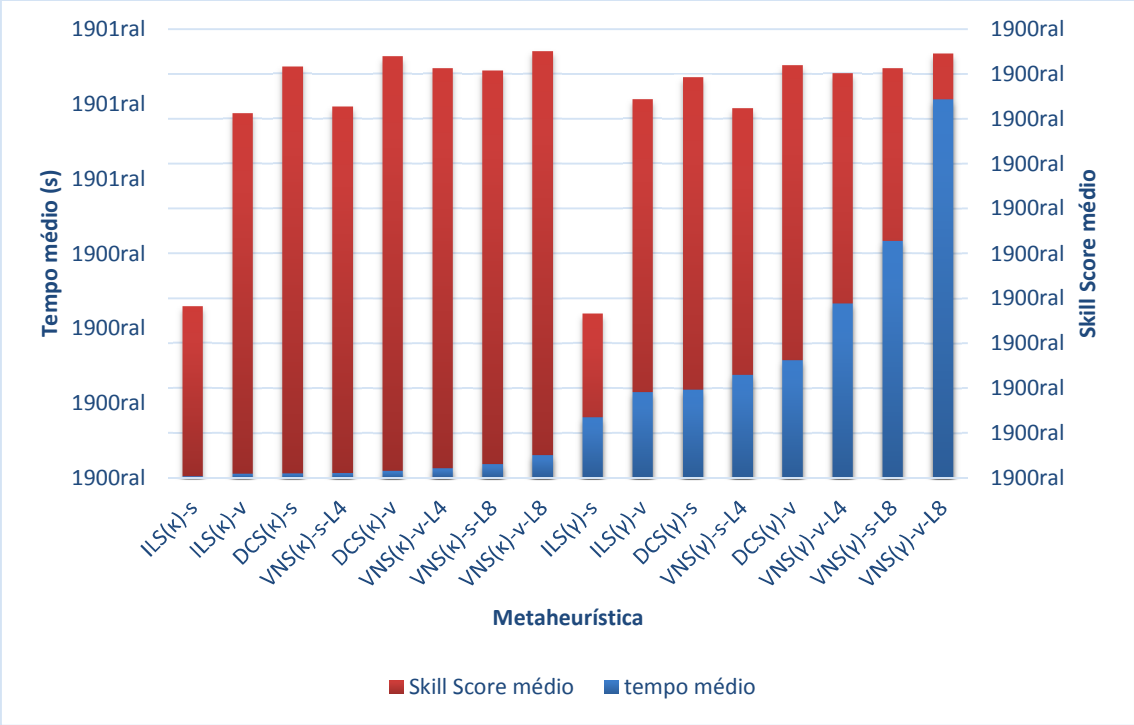

Figura 5.3 – Tempos médios de processamento e *skill scores* médios para todas as variações da metaheurísticas para todas as bases de dados consideradas (20 execuções) ordenadas segundo tempos de processamento crescentes.

Uma análise mais detalhada para cada base de dados é apresentada a seguir.

A Figura 5.4 ilustra os *skill scores* médios obtidos para 20 execuções de cada variação de cada metaheurística para cada uma das 13 bases de dados consideradas. Entretanto, são também incluídos os *skill scores* médios obtidos pelas metaheurísticas anteriormente propostas por outros autores, ACO, SA, GA e TS. Nota-se que estas últimas obtiveram os piores *skill scores* médios para a maioria das bases de dados, juntamente com as variações do ILS que utilizam busca local padrão. Os *skill scores* médios obtidos para as bases de dados Vote e Derm2 foram ligeiramente inferiores para as variações do ILS com busca local VND e para algumas variações do VNS. Entretanto, as variações do VNS tiveram bons *skill scores* médios para as demais bases de dados. Finalmente, a nova metaheurística DCS, proposta no escopo desta tese, obteve *skill scores* médios bons para todas as bases de dados. Em

particular, o DCS teve duas variações, DCS(κ)-s e DCS(κ)-v, que obtiveram *skill scores* médios entre os 5 melhores, mas que demandaram tempos de processamento muito baixos (Figuras 5.3 e 5.5), uma vez que as variações do DCS são as que demandam menos iterações (Figura 5.2). É importante notar que, conforme o detalhamento das reduções apresentado nas tabelas desta seção, as metaheurísticas aqui propostas obtiveram resultados melhores que as anteriormente propostas (ACO, SA, GA e TS), especialmente para a base de dados Mushroom, para a qual nenhuma destas últimas conseguiu obter reduções de cardinalidade 3. É interessante fazer a ressalva de que as metaheurísticas ACO, SA, GA e TS poderiam ser eventualmente melhoradas, mas isso extrapola o escopo deste trabalho.

A Figura 5.5 é similar à Figura 5.3, comparando os *skill scores* médios (Equação 5.4) e os tempos médios de processamento de cada variação de cada metaheurística, considerando-se as 20 execuções para todas as 13 bases de dados consideradas, mas ordenando as variações das metaheurísticas segundo *skill scores* médios crescentes.

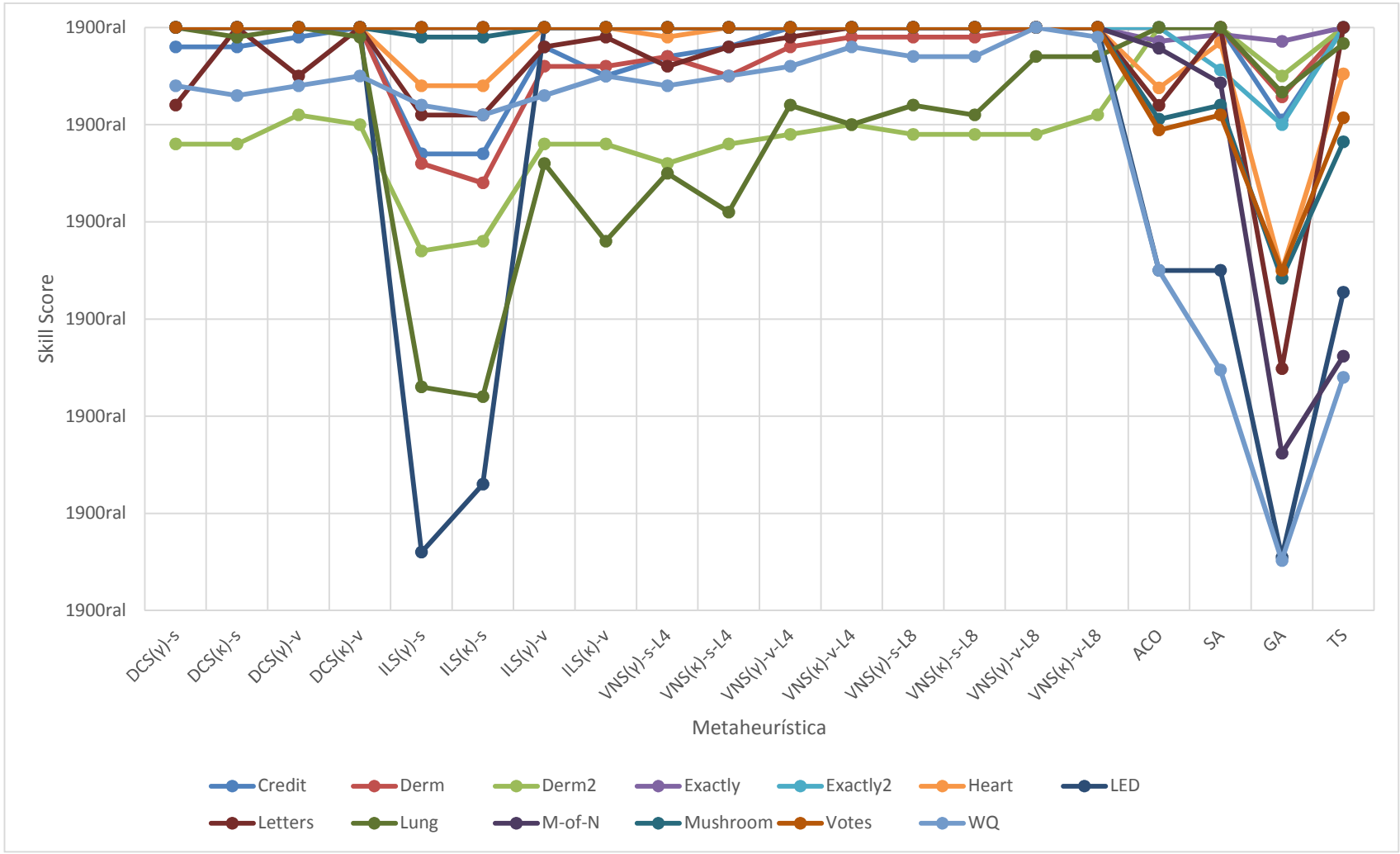

Figura 5.4 – *Skill scores* médios obtidos para 20 execuções de cada variação de cada metaheurística proposta para cada uma das 13 bases de dados consideradas e também para as metaheurísticas propostas anteriormente.

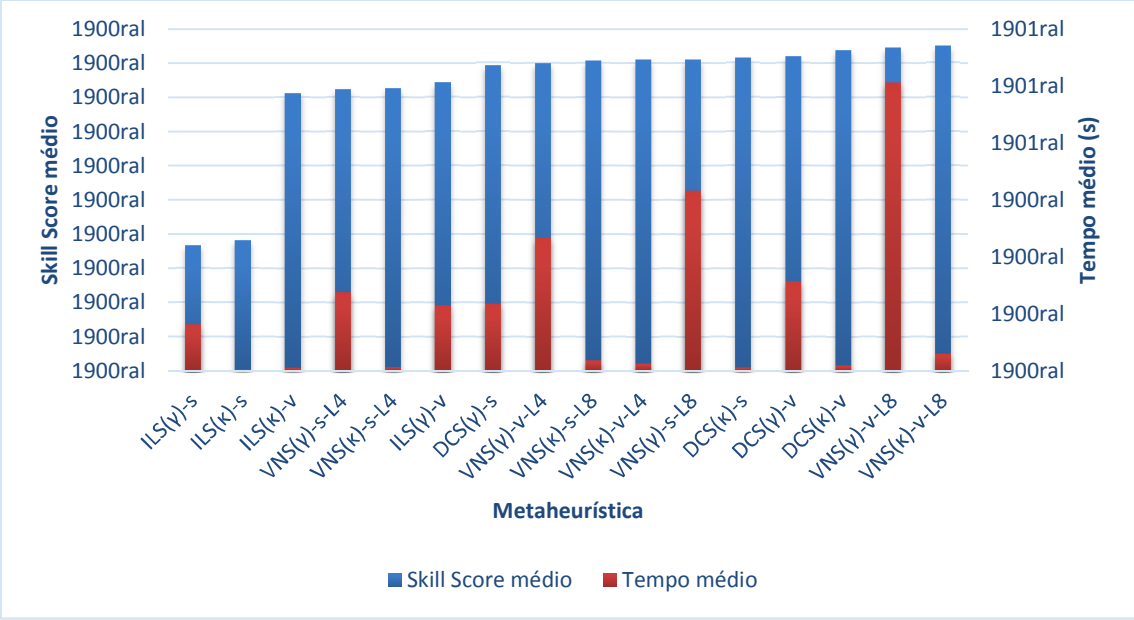

Figura 5.5 – Tempos médios de processamento e *skill scores* médios para todas as variações da metaheurísticas para todas as bases de dados consideradas (20 execuções) ordenadas segundo *skill scores* médios crescentes.

5.2 Resultados – base de dados meteorológicos

Nesta seção são apresentados e discutidos os resultados referentes à aplicação das metaheurísticas VNS, ILS e DCS no cálculo de reduções em TCA, para construção de classificadores com a base de dados meteorológicos descrita no Capítulo 4 com o objetivo de fazer a predição de ocorrência de atividade convectiva. Na seção anterior, que abrangeu o cálculo de reduções para bases de dados gerais, as reduções obtidas pelas diversas variações das metaheurísticas foram avaliadas unicamente pela sua cardinalidade utilizando a métrica *skill score* aqui proposta.

Nesta seção, além da avaliação das reduções por essa métrica, avaliam-se também os classificadores derivados das reduções obtidas quanto ao seu desempenho de classificação na predição de atividade convectiva para as 3 minirregiões definidas A, B e C (Figura 4.1). Assim, cada saída/instância do conjunto de teste, formado por previsões do modelo numérico ETA20, é

classificada como pertencente à classe ausente/fraca, moderada ou forte de atividade convectiva. Obtém-se assim, para cada conjunto de teste, uma matriz de confusão que sumariza as instâncias correta e incorretamente classificadas de cada classe, e que permite calcular os índices de desempenho de classificação aqui adotados, a acurácia e o índice Kappa de Cohen.

Obtiveram-se classificadores, baseados em regras, para cada uma das 16 variações das metaheurísticas propostas, apresentadas na Tabela 5.1. Entretanto, considerando-se as 3 minirregiões, há 3 conjuntos de treinamento que resultam em 48 opções de classificadores, um para cada variação de metaheurística e para cada minirregião. Adicionalmente, para cada uma dessas opções foram utilizadas previsões de 24, 48 ou 72 horas do modelo ETA20 e os dados de treinamento foram aleatoriamente divididos em 1, 8, 16 ou 32 partições (método PACT, Seção 2.6), com os casos denominados 1P, 8P, 16P e 32P, respectivamente. Assim, foram construídos 576 classificadores diferentes. Nos conjuntos de treinamento são utilizados, além das saídas do modelo ETA20, dados de densidade de ocorrência de descargas elétricas NS, usadas aqui como indicativas de atividade convectiva, conforme descrito no Capítulo 4 (Figura 4.2). Como o treinamento é executado 10 vezes para cada caso, de forma a se obter valores médios de acurácia e do índice Kappa, efetuaram-se 5,760 treinamentos para cada conjunto de partições. Se levado em conta que o treinamento é realizado por partição, então o número de treinamentos, então têm-se [1+8+16+32] ou 57 partições para cada caso, e o total real de treinamentos ascende a [5760/4] × 57 ou 82,080. Entretanto, ao se empregar partições, as 10 execuções não são independentes, pois, para cada partição, executam-se 10 treinamentos e seleciona-se a melhor redução obtida, sendo assim o conjunto final de reduções formado pela melhor redução de cada partição. Este conjunto será então empregado para inferência das regras de decisão que compõem o classificador.

Um ponto importante foi a construção dos conjuntos de treinamento e de teste, para a qual foi adotado o esquema de *Holdout*, em que selecionam-se

aleatoriamente 80% dos dados para treinamento e 20% para teste. Os dados de treinamento são então novamente divididos, também aleatoriamente, de forma a compor as partições de treinamento.

A seguir, expõem-se os resultados médios obtidos (acurácia e índice Kappa) para as 16 variações de metaheurísticas, discriminados por minirregião (A, B e C) e por previsão (24, 48 e 72 UTC) e pelo número de partições (P1, P8, P16 e P32). Estes resultados aparecem nas Figuras 5.6, 5.7 e 5.8 para a minirregião A (respectivamente para previsões de 24, 48 e 72 UTC), nas Figuras 5.9, 5.10 e 5.11 para a minirregião B (respectivamente para previsões de 24, 48 e 72 UTC) e nas Figuras 5.12, 5.13 e 5.14 para a minirregião C (respectivamente para previsões de 24, 48 e 72 UTC). Nestas figuras são mostradas as médias das 16 variações para os valores máximos, médios e mínimos do índice Kappa (diagrama de barras) e as médias para a acurácia (linha com pontos discretos), em função do número de partições.

Nas mesmas figuras, pode-se observar que o desempenho dos classificadores com partição única (P1) foi muito pobre, enquanto que o uso de 32 partições (P32) resultou em desempenho pior do que 8 (P8) ou 16 (P16), sugerindo que o número ideal de partições seja um destes. Apesar de que as figuras apresentam valores médios de acurácia e de índice Kappa para todas as 16 variações de metaheurísticas, ressalva-se que os resultados apresentados adiante corroboram que os melhores desempenhos de classificação foram obtidos com P8 ou P16. Entretanto, considerando valores máximos e mínimos absolutos, vê-se que ocorrem casos pontuais em que uma única metaheurística com P8 obteve resultados piores, como na previsão de 24hs para a minirregião A (Figura 5.6), em que o Kappa da variação VNS(κ)-s-L8 (P8) ficou abaixo dos Kappas médios para P1 e P32, ou então na previsão de 24hs para a minirregião B (Figura 5.9), em que também o Kappa da variação VNS(κ)-s-L4 (8) ficou abaixo dos Kappas médios para P1 e P32. Entretanto, em ambos casos, as correspondentes acurácias não seguiram esse comportamento.

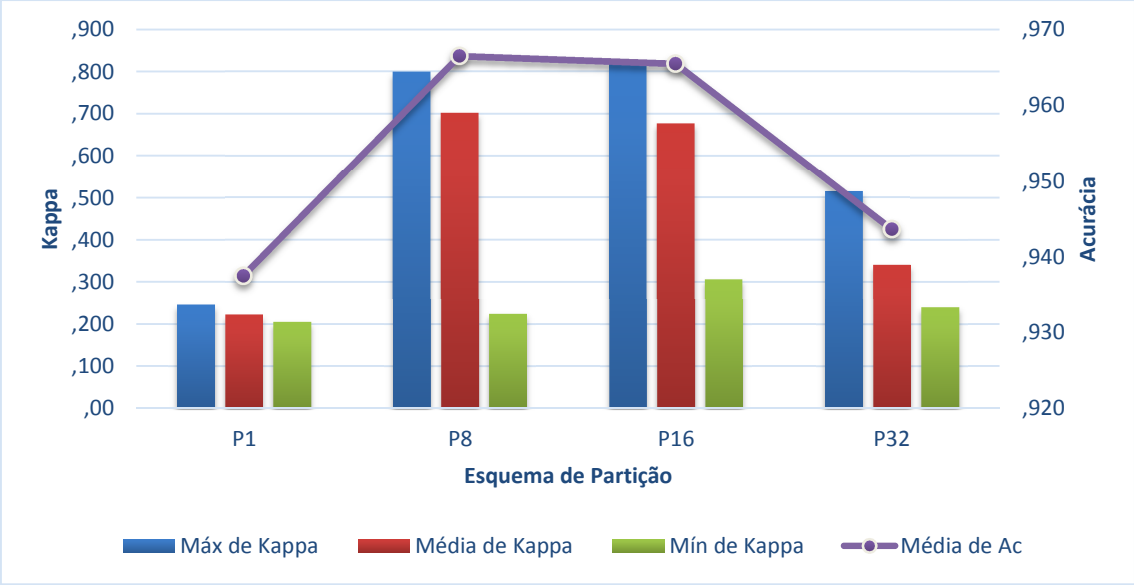

Figura 5.6 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião A com uma base de dados de previsões de 24hs.

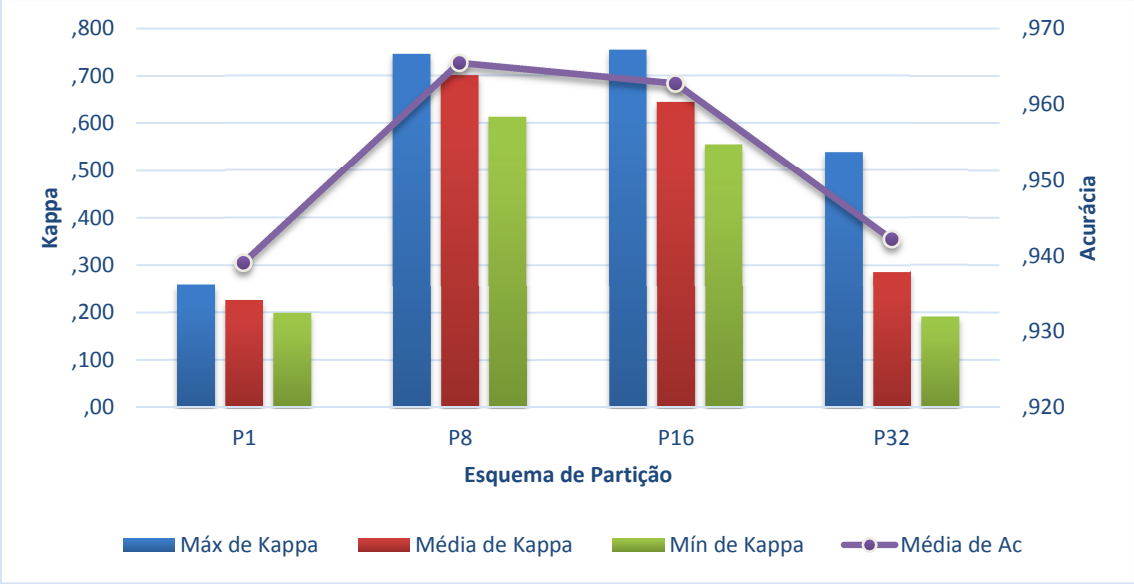

Figura 5.7 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião A com uma base de dados de previsões de 48hs.

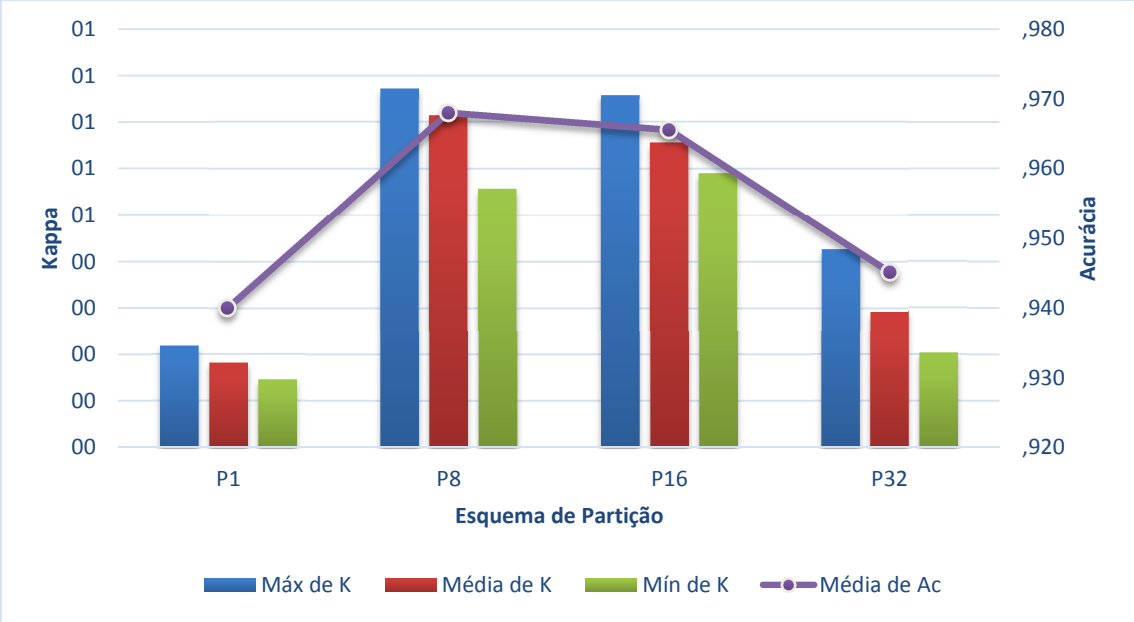

Figura 5.8 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião A com uma base de dados de previsões de 72hs.

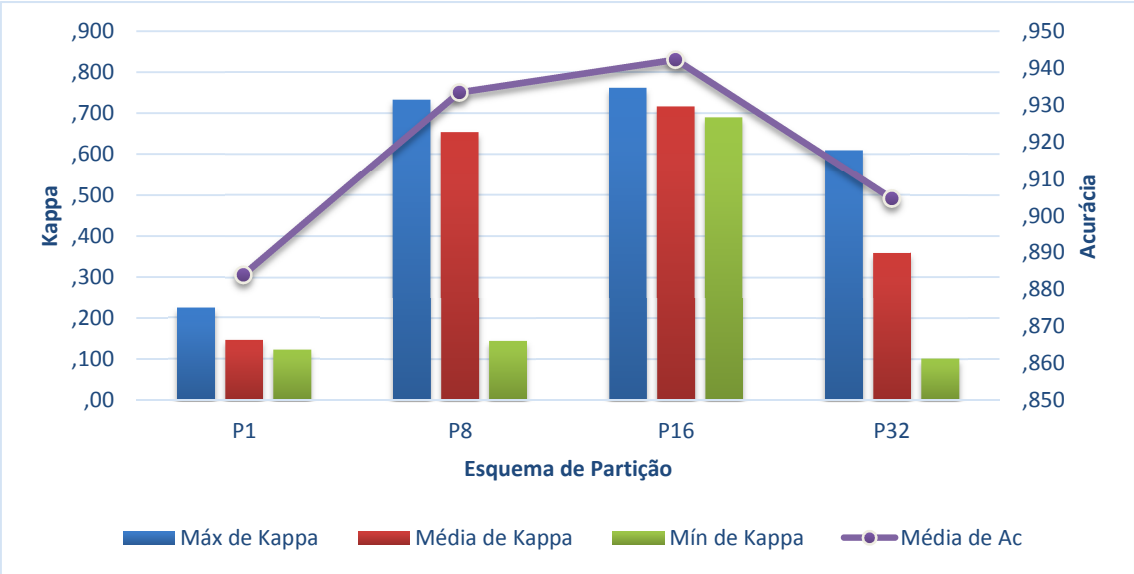

Figura 5.9 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião B com uma base de dados de previsões de 24hs.

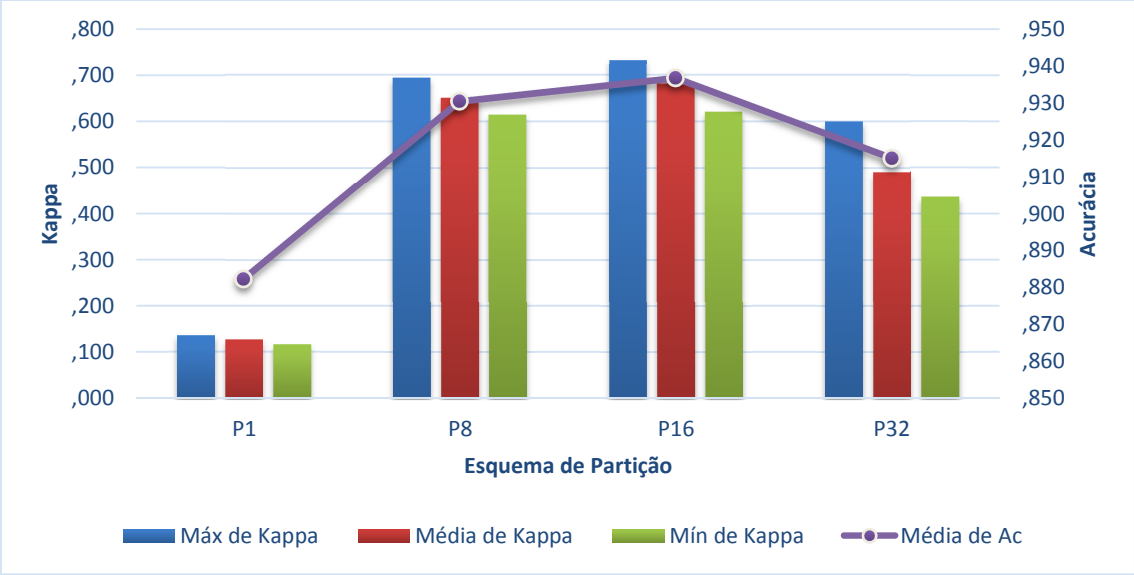

Figura 5.10 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião B com uma base de dados de previsões de 48hs.

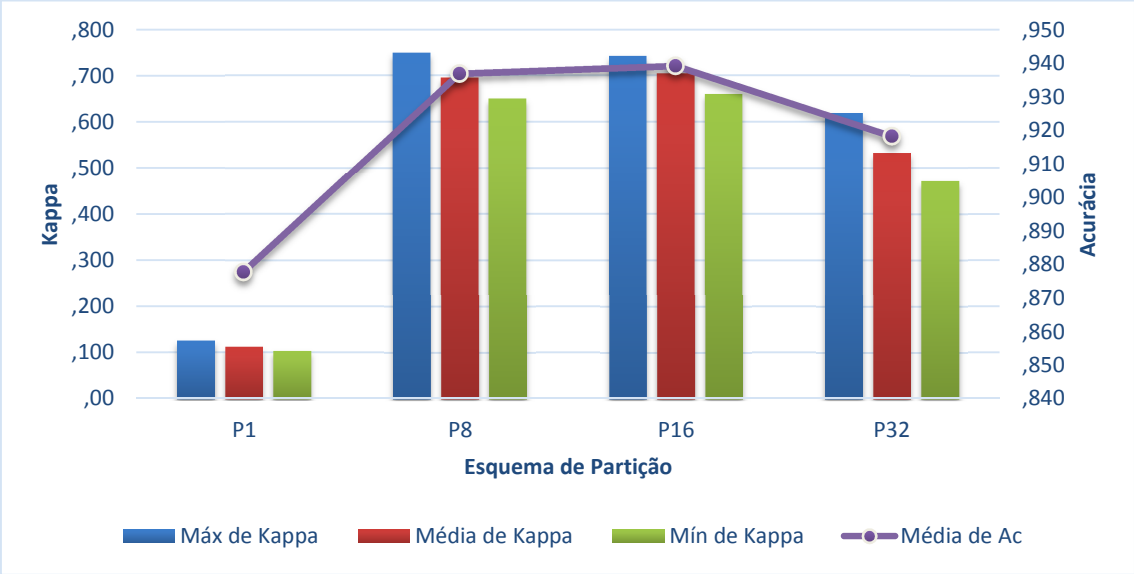

Figura 5.11 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião B com uma base de dados de previsões de 72hs.

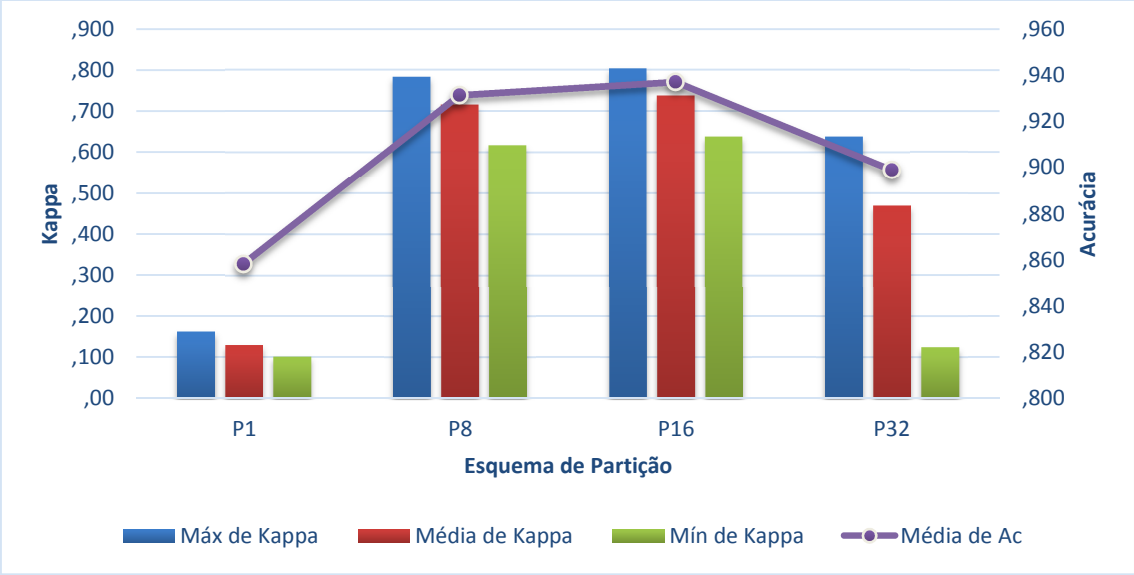

Figura 5.12 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião C com uma base de dados de previsões de 24hs.

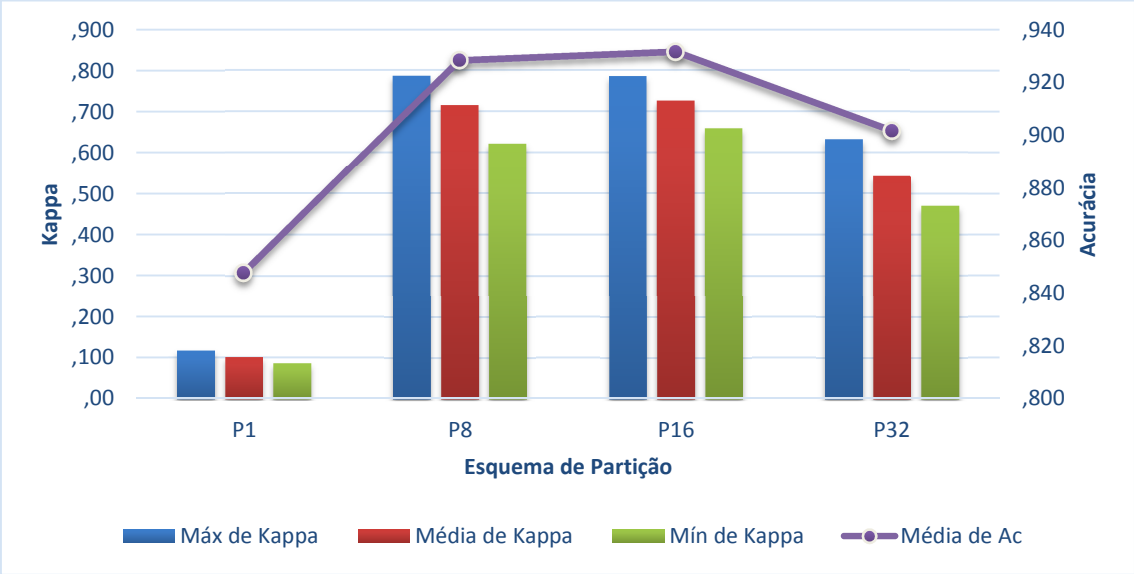

Figura 5.13 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião C com uma base de dados de previsões de 48hs.

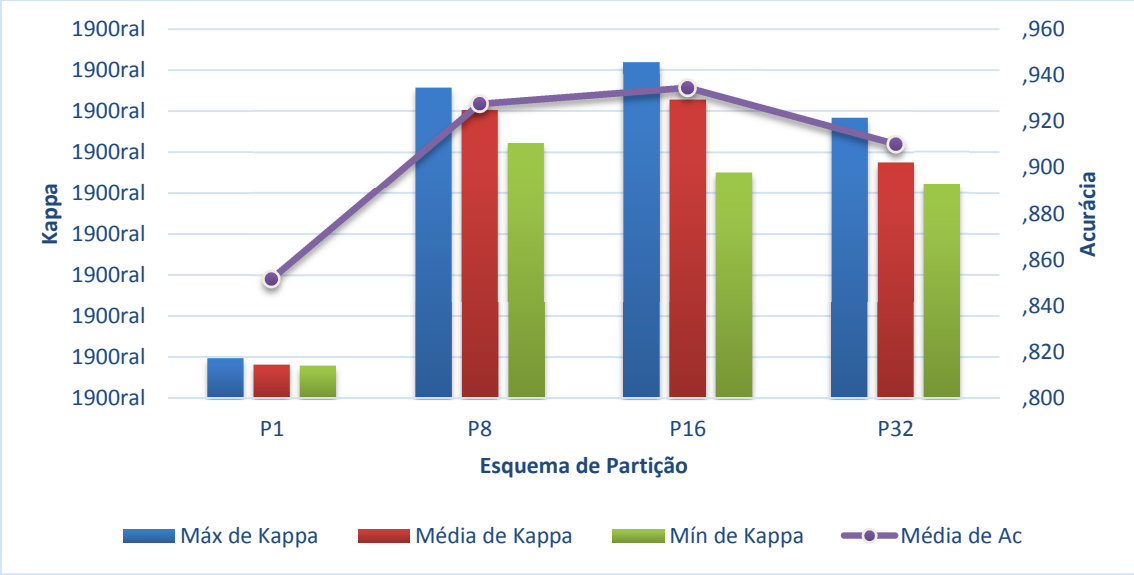

Figura 5.14 – Valores médios, máximos e mínimos absolutos do índice Kappa para as 16 variações de metaheurísticas (diagrama de barras) e valores médios de acurácia (linha com pontos discretos), em função do número de partições na classificação de eventos convectivos para a minirregião C com uma base de dados de previsões de 72hs.

Assim, nos resultados seguintes são apenas levados em consideração apenas aqueles obtidos pelos esquemas de particionamento (PACT) de 8 e 16 partições (P8 e P16). Descartadas as partições P1 e P32 devido ao seu pior desempenho de classificação, objetiva-se a seguir comparar unicamente os esquemas P8 e P16. Enfatiza-se que, apesar disso, o cálculo de reduções, e as correspondentes induções de regras de decisão e classificações foram efetuados para todas as 16 variações de metaheurísticas e com todos os esquemas de partição para cada minirregião e horário de previsão.

A título de exemplo, os tempos de processamento médios por esquema de partição, considerando-se as 16 variações de metaheurística, as 10 execuções e os casos de teste possíveis (3 minirregiões e 3 horários de previsão) para a base de dados meteorológicos foram, em segundos, para execução sequencial (P1) ou paralela (8P, 16P e 32P), ambas na máquina "Matrix": 27,866 (P1), 4,196 (P8), 1,660 (P16) e 687 (P32).

Na Figura 5.15 é mostrada a acurácia média das variações de metaheurística nos esquemas P8 e P16, para as 3 minirregiões e os 3 horários de previsão. A acurácia média para P16 foi quase sempre superior do que para P8, exceto para a variação ILS(*γ*)-v. As variações do DCS obtiveram acurácias médias maiores, enquanto que aquelas do ILS, menores. As melhores variações do VNS foram VNS(*κ*)-s-L8 e VNS(*κ*)-v-L8, ambas para P8.

Analogamente, a Figura 5.16 mostra o índice Kappa médio das variações de metaheurística nos esquemas P8 e P16, para as 3 minirregiões e os 3 horários de previsão. Novamente, a variação ILS(*γ*)-v foi a exceção mais notável, embora tenham aparecido mais casos em que variações com P8 tenham obtido índices Kappa superiores às correspondentes com P16, mas em nenhum destes casos (todos referentes a variações VNS), a diferença excedeu o valor de 0.02, ou seja, menos de 3%.

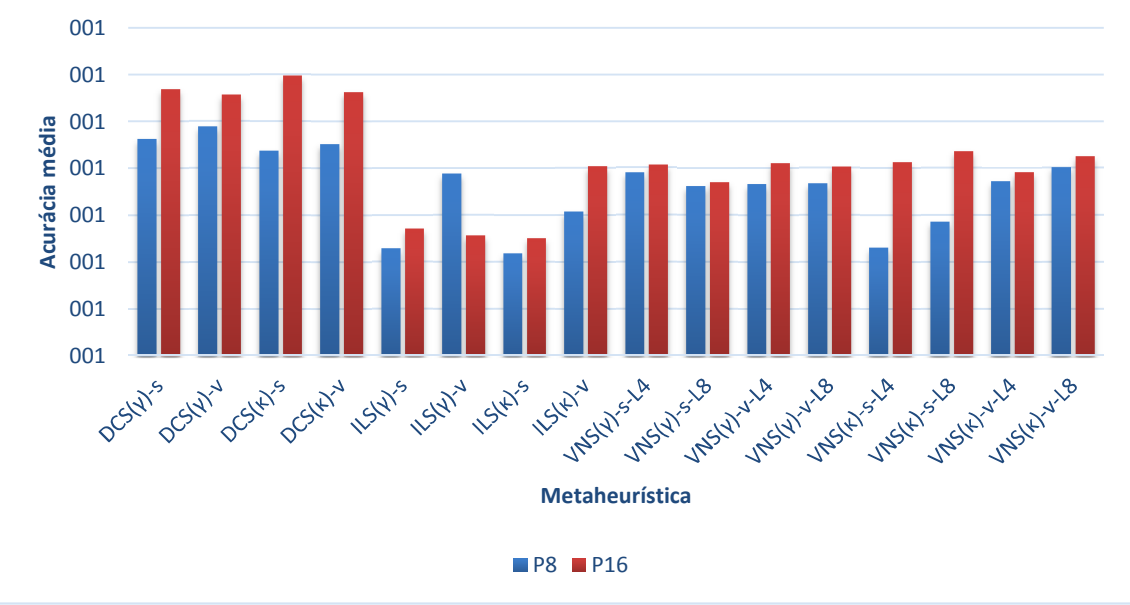

Figura 5.15 – Acurácia média obtida pelas 16 variações das metaheurísticas para as 3 minirregiões e os 3 horários de previsão utilizando os esquemas P8 e P16.

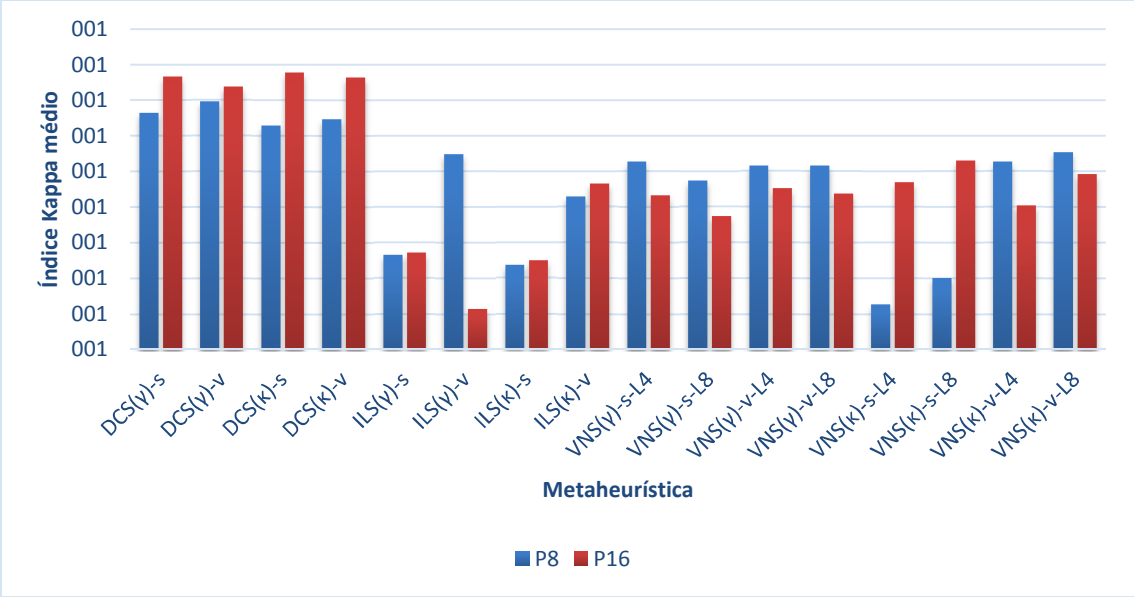

Figura 5.16 – Índice Kappa médio obtido pelas 16 variações das metaheurísticas para as 3 minirregiões e os 3 horários de previsão utilizando os esquemas P8 e P16.

Finalmente, são comparados os tempos de processamento médios das variações de metaheurística nos esquemas P8 e P16, para as 3 minirregiões e os 3 horários de previsão, conforme ilustrado na Figura 5.17. O cálculo de reduções com P8 sempre demandou mais tempo de processamento do que com P16 para cada variação considerada. As variações do VNS com estrutura de vizinhança de cardinalidade L=8 foram as que mais consumiram tempo, seja com P8 ou com P16, como seria de se esperar devido à maior complexidade algorítmica do VNS. A Figura 5.18 mostra a razão entre os tempos de processamento das versões P8 e P16 para cada variação mostrada na figura anterior. Seria de se esperar uma razão de 2:1 entre P8 e P16, mas essa razão oscilou entre 3,57 e 1,42 conforme a variação considerada.

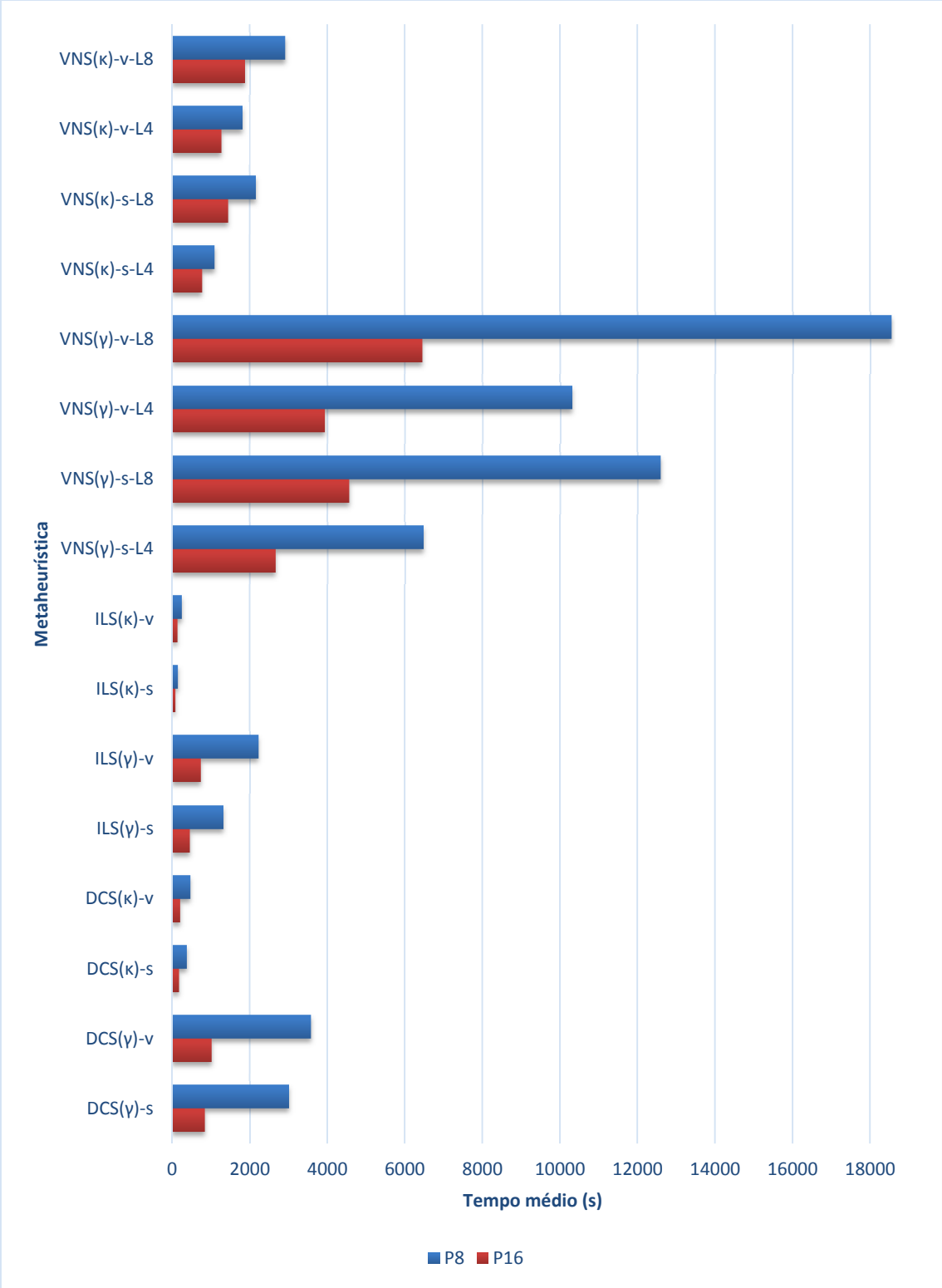

Figura 5.17 – Tempos de processamento médios demandados pelas 16 variações das metaheurísticas para as 3 minirregiões e os 3 horários de previsão utilizando os esquemas P8 e P16.

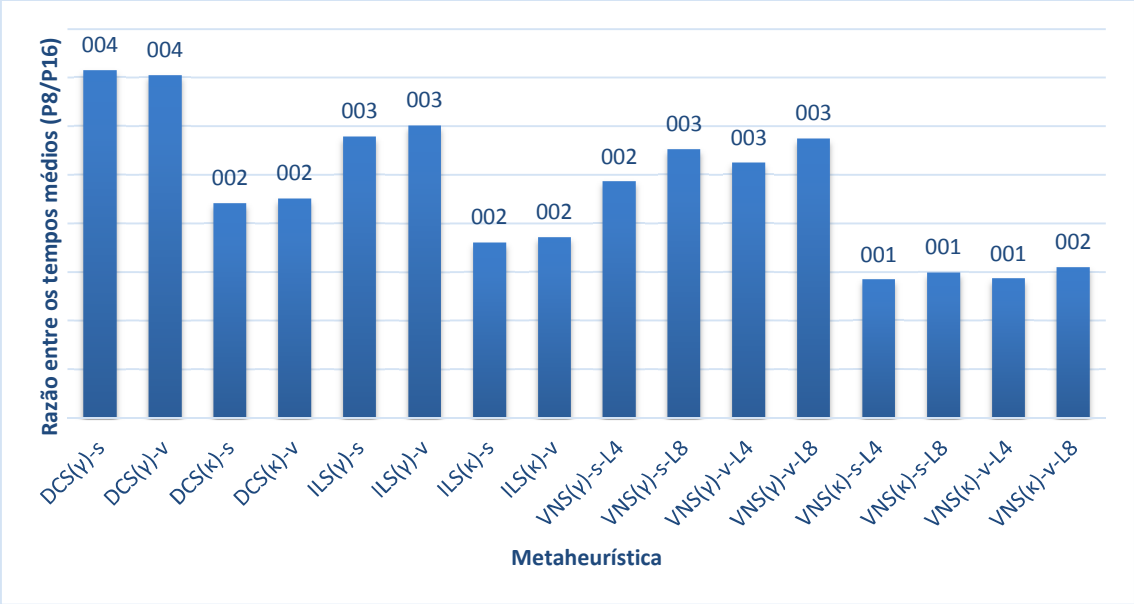

Figura 5.18 – Razão entre os tempos de processamento médios demandados entre os esquemas P8 e P16 para as 16 variações das metaheurísticas e para as 3 minirregiões e os 3 horários de previsão utilizando os esquemas P8 e P16.

Em relação aos tempos de processamento, foram apresentados os tempos relativos ao cálculo de reduções para 8 e 16 partições. Esses cálculos foram implementados no software específico desenvolvido no escopo desta tese. Como esses cálculos são independentes, foram paralelizados usando um pacote de paralelização da linguagem Perl (ProcQueue) que provê a execução paralela de múltiplos processos explorando a arquitetura de máquinas multiprocessadas e processadores multinúcleo, como no caso das servidoras "Jupiter" e "Matrix". Esse pacote gera novos processos para executar o cálculo de reduções de maneira independente em cada partição, sendo esses processos escalonados pelo sistema operacional Linux para execução nos vários cores. Isso viabilizou a execução dos inúmeros casos de teste. Os casos referentes a partições únicas (P1) demandaram tempos de processamento muito longos

Na abordagem de TCA, além do cálculo de reduções é preciso induzir as regras de decisão a partir dessas reduções. Essas regras constituem o

classificador derivado desse conjunto de reduções. A indução de regras e a classificação de um conjunto de dados da base de dados considerada foram realizados pelo software Rosetta. Não foi possível distinguir esses 2 tempos, mas pode-se dizer que o tempo de classificação é muito menor que o da indução de regras. No caso do PACT, como o software Rosetta não é paralelo, quanto mais partições, mais demorada será a indução de regras, que é feita para cada partição. Os tempos médios para todas as variações de metaheurísticas e para todos os casos de teste (3 minirregiões e 3 horários de previsão), nas máquinas supracitadas, foram aproximadamente de:

- P1 = 120 s
- P8 = 530 s
- P16 = 760 s
- P32 = 855 s

Conforme o acima exposto, a análise do desempenho de classificação de dados meteorológicos foi referente ao PACT, ou seja, ao número de partições utilizado, considerando todas as variações de metaheurísticas. Isso se justifica pois o uso de partições viabilizou a classificação com dados meteorológicos. Entretanto, pode-se retomar as métricas utilizadas na Seção 5.1 para as bases de dados gerais, no caso, a cardinalidade das reduções obtidas e o próprio *Skill Score*. A Figura 5.19 exibe a cardinalidade média das reduções obtidas por cada metaheurística para as 3 minirregiões e os 3 horários de previsão, comparando os esquemas P8 e P16.

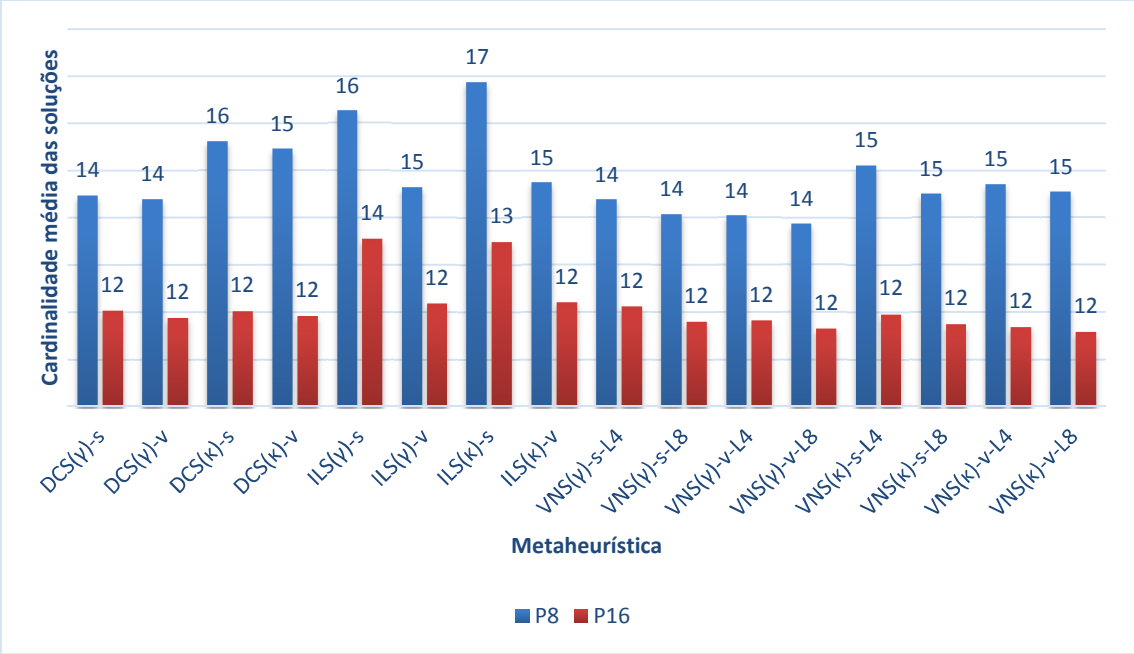

Figura 5.19 – Cardinalidade média das reduções obtidas pelas 16 variações das metaheurísticas, para as 3 minirregiões e para os 3 horários de previsão, considerando os esquemas P8 e P16.

Normalmente, partições menores, que são conjuntos de treinamento com menores instâncias, tendem a gerar reduções de menor cardinalidade devido à "especialização" de cada partição. Assim, observa-se na Figura 5.19 que a cardinalidade média para P8 foi 15, enquanto que, para P16, foi de 12. A título de informação, a cardinalidade média para P1 foi aproximadamente 24 e a de P32, aproximadamente 8. Conforme já mencionado, uma "especialização" excessiva degrada o desempenho de classificação, o que se constatou no esquema P32, porém viu-se que o desempenho de classificação do esquema sem partição (ou partição única) foi o pior de todos.

Uma vez analisada a cardinalidade, passa-se à análise do *Skill Score* médio para as 16 variações das metaheurísticas, para as 3 mini regiões e para os 3 horários de previsão, unicamente no esquema P16. Na Seção 5.1, essa métrica permitia uma análise mais direta, sem necessidade de verificar o desempenho de classificação, pois dispunha-se das cardinalidade das

reduções obtidas em trabalhos anteriores de outros autores, os quais serviam como valores de referência e que também permitiam o cálculo do *Skill Score*. Conforme definida no início deste capítulo, o *Skill Score* avalia a obtenção frequente de reduções com menor cardinalidade, i.e. com menos atributos condicionais, assumindo-se que quanto menor a cardinalidade, melhor a redução, o que nem sempre é verdadeiro.

As metaheurísticas calculam as reduções efetuando uma busca no espaço de reduções/soluções e avaliando cada solução candidata por uma função de avaliação. Considerando-se que, as metaheurísticas propostas sempre partem de uma solução inicial que corresponde à cardinalidade máxima, ou seja, uma redução contendo todos os atributos condicionais, esta solução inicial tem seu valor da função de avaliação próximo de 1, ou seja, próximo do máximo. Ao longo das iterações de qualquer uma dessas metaheurísticas, novas soluções somente substituem a solução corrente se forem melhores. Assim, os valores assumidos pela função de avaliação tende a permanecer próxima de 1, de forma que pode-se considerar todas as reduções como "quase-ótimas" e, nesse caso, as reduções obtidas seriam distinguidas pelo *Skill Score*, que conforme mencionado, privilegia reduções com cardinalidade baixa. Isso pode ser observado na Figura 5.20, que apresenta os valores médios da função de avaliação e do *Skill Score* das 16 variações das metaheurísticas, para as 3 minirregiões e para os 3 horários de previsão, considerando apenas o esquema P16.

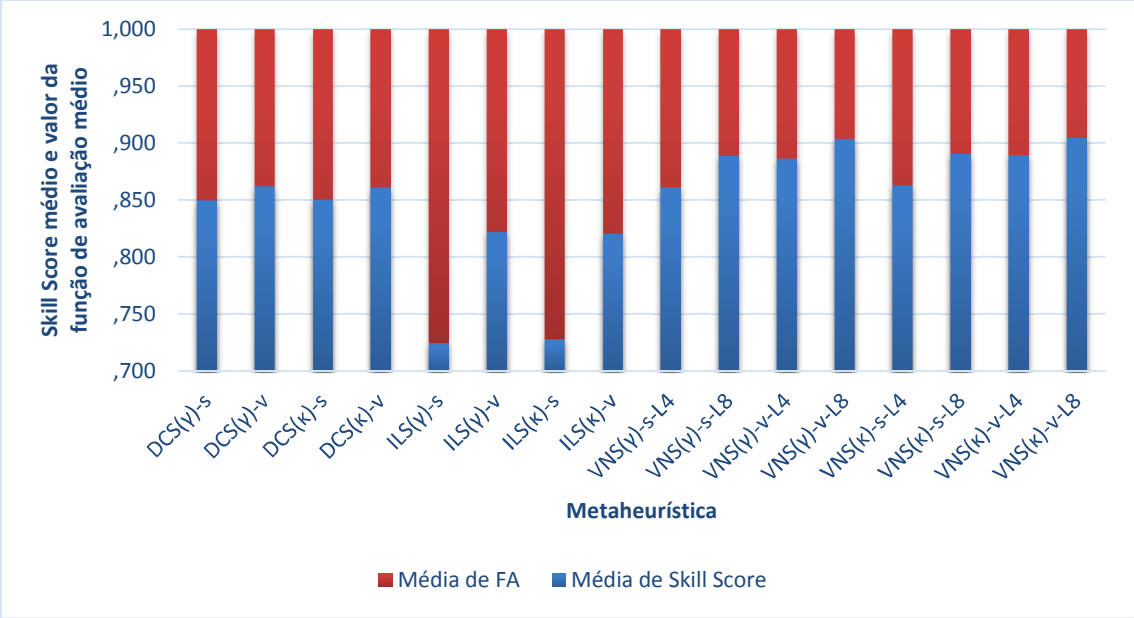

Figura 5.20 – Valores médios da função de avaliação e do *Skill Score* das reduções obtidas pelas 16 variações das metaheurísticas, para as 3 minirregiões e para os 3 horários de previsão (esquema P16).

Nesta figura pode-se observar que o valor da função de avaliação é muito próximo da unidade (aproximadamente 0.98) para todas as 16 variações das metaheurísticas, enquanto que os *Skill Scores* médios tem baixíssima variabilidade para as variações de DCS e VNS. Aqui, como em todos os resultados apresentados nesta seção, consideram-se valores médios para 10 execuções de cada variação de metaheurística, o que corresponde, no esquema P16, a 10 execuções por partição, ou um total de 160 execuções. À semelhança dos *Skill Scores* médios obtidos na Seção 5.1 (Figuras 5.3, 5.4 e 5.5), as variações das metaheurísticas ILS obtiveram baixos valores, e aquelas propostas anteriormente por outros autores também. A análise dos *Skill Scores* médios para os dados meteorológicos na Figuras 5.20 permite inferir que as variações do DCS diferiram por menos de 0,01 entre si, enquanto que as variações do VNS, por menos de 0,05 aproximadamente.

Se na seção anterior (bases de dados gerais) os *Skill Scores* médios permitiam uma cômoda comparação com as demais metaheurísticas de outros autores, nesta seção (base de dados meteorológicos) pode-se apenas dizer que as variações do DCS e VNS são melhores. Foi constatado que, em média, para cada conjunto de 10 execuções para uma partição, as 10 reduções obtidas tendem a apresentar apenas duas cardinalidades, tanto para as variações do DCS como do VNS. No caso do ILS, esse valor dobra, mostrando uma instabilidade na busca por reduções/soluções sub-ótimas.

Assim, pode-se concluir que as variações do DCS e VNS são melhores, sempre no esquema P16, com a vantagem de que as primeiras demandam menos tempo de processamento. Essa conclusão é baseada no desempenho médio de classificação e também no *Skill Score*, tendo sido já definido o PACT "ótimo" como P16. Entretanto, considerando-se a complexidade da base de dados meteorológicos, torna-se mais conveniente elencar os melhores e piores casos em termos de desempenho de classificação para essas metaheurísticas, discriminando-se por minirregião e por horário de previsão, como exposto a seguir.

As tabelas seguintes (Tabelas 5.9 a 5.14) apresentam as melhores e piores classificações obtidas para cada uma das três minirregiões (A, B e C) e para cada um dos horários de previsão (24, 48 e 72 h), especificando qual a variação de metaheurística que as obteve, apresentando a matriz de confusão resultante e os valores correspondentes de acurácia e do índice Kappa. As 3 classes correspondem a atividade convectiva fraca/ausente, moderada e forte.

Tabela 5.9. Melhores resultados de classificação para a minirregião A para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião A		Predito								
		24h			48h			72h		
		DCS(*κ*)-v			DCS(*γ*)-s			DCS(*γ*)-v		
		A	M	F	A	M	F	A	M	F
Atual	A	7737	16	1	7618	22	1	7519	27	0
	M	131	320	8	163	273	9	159	252	8
	F	19	9	112	26	9	94	20	1	99
		Kappa: 0,8179 **Acurácia:** 0,9780 **Ac_F** 0,8000			**Kappa:** 0,7541 **Acurácia:** 0,9720 **Ac_F** 0,7287			**Kappa:** 0,7568 **Acurácia:** 0,9734 **Ac_F** 0,8250		

Tabela 5.10. Piores resultados de classificação para a minirregião A para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião A		Predito								
		24h			48h			72h		
		VNS(*γ*)-s-L8			VNS(*κ*)-v-L4			VNS(*γ*)-s-L4		
		A	M	F	A	M	F	A	M	F
Atual	A	7732	19	3	7622	16	3	7526	18	2
	M	228	226	5	254	186	5	236	180	3
	F	60	7	73	67	3	59	56	3	61
		Kappa: 0,6363 **Acurácia:** 0,9617 **Ac_F** 0,5214			**Kappa:** 0,5729 **Acurácia:** 0,9576 **Ac_F** 0,4574			**Kappa:** 0,5906 **Acurácia:** 0,9607 **Ac_F** 0,5083		

Tabela 5.11. Melhores resultados de classificação para a minirregião B para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião B		Predito								
		24h			48h			72h		
		DCS(*γ*)-s			DCS(*κ*)-s			DCS(*κ*)-v		
		A	M	F	A	M	F	A	M	F
Atual	A	7218	77	3	7080	91	2	6951	68	3
	M	233	561	27	244	514	26	248	545	23
	F	27	46	155	49	39	158	65	28	142
		Kappa: 0,7622 **Acurácia:** 0,9505 **Ac_F** 0,6798			**Kappa:** 0,7329 **Acurácia:** 0,9450 **Ac_F** 0,6423			**Kappa:** 0,7418 **Acurácia:** 0,9461 **Ac_F** 0,6043		

Tabela 5.12. Piores resultados de classificação para a minirregião B para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião B		Predito								
		24h			48h			72h		
		VNS(γ)-s-L4			VNS(κ)-s-L4			VNS(γ)-s-L8		
		A	M	F	A	M	F	A	M	F
Atual	A	7201	85	12	7080	90	3	6926	92	4
	M	303	488	30	332	425	27	326	465	25
	F	53	36	139	71	39	136	83	24	128
		Kappa: 0,6901			**Kappa:** 0,6491			**Kappa:** 0,6600		
		Acurácia: 0,9378			**Acurácia:** 0,9315			**Acurácia:** 0,9314		
		Ac_F 0,6096			**Ac_F** 0,5528			**Ac_F** 0,5447		

Tabela 5.13. Melhores resultados de classificação para a minirregião C para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião C		Predito								
		24h			48h			72h		
		DCS(γ)-s			DCS(κ)-s			DCS(κ)-v		
		A	M	F	A	M	F	A	M	F
Atual	A	7021	37	15	6849	39	9	6756	48	13
	M	201	465	34	208	447	35	215	411	36
	F	88	35	456	122	40	472	109	36	441
		Kappa: 0,8030			**Kappa:** 0,7857			**Kappa:** 0,7857		
		Acurácia: 0,9509			**Acurácia:** 0,9449			**Acurácia:** 0,9449		
		Ac_F 0,7876			**Ac_F** 0,7445			**Ac_F** 0,7526		

Tabela 5.14. Piores resultados de classificação para a minirregião C para cada horário de previsão, expressos pela matriz de confusão, acurácia e índice Kappa.

Minirregião C		Predito								
		24h			48h			72h		
		VNS(γ)-s-L8			VNS(γ)-v-L8			VNS(γ)-s-L8		
		A	M	F	A	M	F	A	M	F
Atual	A	7012	41	20	6841	43	13	6748	44	25
	M	279	392	29	280	370	40	252	378	32
	F	174	26	379	173	41	420	149	51	386
		Kappa: 0,7100			**Kappa:** 0,7089			**Kappa:** 0,7166		
		Acurácia: 0,9319			**Acurácia:** 0,9282			**Acurácia:** 0,9314		
		Ac_F 0,6546			**Ac_F** 0,6625			**Ac_F** 0,6587		

As matrizes de confusão apresentadas nas Tabelas 5.9 - 5.14 mostram os erros e acertos de classificação, sendo que as acurácias refletem os elementos da diagonal (i.e. instâncias corretamente classificadas) e o índice Kappa, os elementos fora da diagonal (instâncias incorretamente classificadas). Uma vez que o desempenho de classificação teve pequena variação, considerando-se apenas as variações das metaheurísticas VNS e DCS, ambas com P16 e 10 execuções, optou-se por mostrar os melhores e piores casos. Em geral, as acurácias acima de 95% nos melhores casos, mas sempre acima de 90%, mesmo nos piores casos, enquanto que os Kappas, próximos ou acima de 70% nos melhores casos e próximos ou acima de 60% nos piores.

As variações do DCS obtiveram sempre os melhores resultados, mas considerando-se a pouca diferença entre melhores e piores casos, pode-se dizer que o desempenho de classificação de todas as variações do DCS e do VNS, ambas propostas neste trabalho, seja aceitável, considerando-se as dificuldades de se efetuar a classificação numa bases de dados tão complexa. Entretanto, levando em conta o tempo de processamento, a melhor opção seria a variação DCS(κ)-s, que é a que demanda menos tempo de processamento e que em vários casos foi a que obteve melhor resultado. De maneira geral, as variações do DCS com a função de avaliação κ são mais rápidas que as demais, sendo aquelas com a busca local padrão (s) ligeiramente mais rápidas do que com a busca local VND (v).

Embora os desempenhos de classificação para a minirregião A tenham sido ligeiramente melhores do que para as minirregiões B e C, não é possível distinguir diferenças significativas para os correspondentes horários de previsão. Caso contrário, seria possível constatar que o modelo ETA20 teria previsões mais confiáveis para um determinado horário de previsão (24, 48 ou 72 h), o que acarretaria melhores predições de atividade convectiva, expressos nas matrizes de confusão. Obviamente, o maior interesse é pela predição de ocorrências de atividade convectiva forte (classe F), correspondentes às terceiras linhas das matrizes de confusão. Nos piores casos, apenas cerca de

50% das instâncias F foram classificadas corretamente, enquanto que nos melhores casos, mais de 70%. As instâncias mais numerosas foram as correspondentes à classe ausente/fraca (A), sendo classificadas corretamente em sua grande maioria. Isso pode ser explicado pelo maior número de instâncias desta classe nos conjuntos de treinamento, pois a ocorrência de atividade convectiva é relativamente rara, considerando-se a grade espacial e temporal. Por outro lado, as instâncias pior classificadas foram as moderadas (classe M), sendo a maioria das incorretamente classificadas atribuída à classe A.

É preciso observar que as instâncias de treinamento e teste se referem a anos diferentes (2007-2011), sendo que há uma certa variabilidade na razão entre instâncias de diferentes classes, conforme pode ser observado na Tabela 5.15 extraída de (LIMA; STEPHANY, 2013a), que discrimina os totais de instâncias para esses anos, porém adotando apenas duas classes: SCA para atividade convectiva forte e NSCA, para atividade convectiva moderada, fraca ou ausente. Obviamente, existe uma sazonalidade que não foi levada em conta, dadas as restrições do volume de dados, mesmo considerando-se que se referem a apenas 2 meses de cada um desses 5 anos. Talvez o ideal seria ter dados de vários meses de anos em que ocorreram fenômenos meteorológicos significativos como El Ñino ou La Ñina.

Tabela 5.15. Número de instâncias para as classes NCSA (A+M nesta tese) e SCA (F) para cada minirregião para os meses de janeiro e fevereiro do período 2007-2011.

Ano	Minirregião A		Minirregião B		Minirregião C	
	NSCA	SCA	NSCA	SCA	NSCA	SCA
2007	8362	134	8168	328	8319	177
2008	8489	7	8357	139	7805	691
2009	8456	40	8430	66	8189	307
2010	8323	173	8434	62	7869	627
2011	8450	10	8026	434	7871	589

Fonte: (LIMA; STEPHANY, 2013a)

Nas Tabelas 5.16, 5.17 e 5.18 são mostradas as variáveis condicionais de maior ocorrência nas reduções encontradas pelas variações da metaheurística DCS, correspondentes as Tabelas 5.9, 5.11 e 5.13. Foram consideradas de maior ocorrência apenas as variáveis condicionais encontradas em mais de 40% das reduções. Nessas tabelas, esses porcentagens de ocorrência aparecem entre parênteses, considerando-se o total de 16 reduções (uma para cada partição).

Tabela 5.16. Variáveis condicionais de maior ocorrência nas reduções encontradas pelas variações do DCS para cada horário de previsão e para a minirregião A.

DCS(*κ*)-v	**DCS(*γ*)-s**	**DCS(*γ*)-v**
24h	**48h**	**72h**
omega250 (43,75 %) omega850 (56,25 %) pslc (62,5 %)	fzht (43,75 %) v1000 (43,75 %) v250 (43,75 %) omega1000 (50,0 %) pslc (68,75 %)	v1000 (43,75 %) omega925 (43,75 %) pslc (62,5 %) omega1000 (68,75 %)

Tabela 5.17. Variáveis condicionais de maior ocorrência nas reduções encontradas pelas variações do DCS para cada horário de previsão e para a minirregião B.

DCS(*γ*)-s	**DCS(*κ*)-s**	**DCS(*κ*)-v**
24h	**48h**	**72h**
cape (43,75 %) omega250 (43,75 %) v925 (50,0 %) tabs1000 (50,0 %) omega1000 (56,25 %) omega700 (56,25 %) pslc (87,5 %)	tp2m (43,75 %) v250 (43,75 %) tabs1000 (43,75 %) omega850 (50,0 %) omega1000 (56,25 %) omega250 (56,25 %) pslc (81,25 %)	v925 (43,75 %) v700 (43,75 %) omega1000 (62,5 %) omega850 (75,0 %) pslc (87,5 %)

Tabela 5.18. Variáveis condicionais de maior ocorrência nas reduções encontradas pelas variações do DCS para cada horário de previsão e para a minirregião C.

DCS(*γ*)-s	**DCS(*κ*)-s**	**DCS(*κ*)-v**
24h	**48h**	**72h**
bli (43,75 %) v850 (43,75 %) v500 (43,75 %) omega1000 (62,5 %)	u700 (43,75 %) omega1000 (50,0 %) tp2m (56,25 %) omega700 (56,25 %)	u925 (43,75%) u500 (43,75 %) z500 (43,75 %) urel250 (43,75 %) pslc (50,0 %) omega700 (50,0 %) omega1000 (62,5 %)

Note-se que variáveis como “plsc” e “omega1000” são bastante frequentes nos resultados, mostrando sua importância na composição de padrões associados à ocorrência de eventos severos.

Esses resultados de classificação referem-se ao esquema de *holdout* aleatório, em que o conjunto de instâncias de todos os anos (2007-2011) é dividido aleatoriamente num conjunto de treinamento (que depois é novamente dividido aleatoriamente em 16 partições), e num conjunto de teste, na proporção de 80% e 20%, respectivamente. Entretanto, uma predição no âmbito operacional exigiria que se obtivessem bons resultados de classificação com um "*holdout* cronológico", no qual o conjunto de instâncias seria particionado por meio da escolha de uma data/horário, sendo os dados precedentes a essa data utilizados para treinamento e os demais, para teste. Esse esquema de *holdout* simularia a operação real de um classificador para predição de eventos convectivos. Infelizmente, vários testes realizados com o esquema de "*holdout* cronológico" não foram bem sucedidos, i.e. obtiveram baixo desempenho de classificação, não sendo apresentados aqui. Esses testes incluíram a amostragem da base de dados, de forma a obter uma razão de aproximadamente 2:1 entre instâncias da classe A e da classe F. Incluíram também o uso de instâncias de um único ano, em vez do total de instâncias dos meses de janeiro e fevereiro dos 5 anos (2007-2011). Outros testes realizados foram para o esquema "*leave-one-out*", em que se utilizam todas as instâncias para treinamento, exceto uma única, que constitui o conjunto de teste, também sem sucesso. Essas falhas podem ser atribuídas à complexidade dos dados e também ao número baixo de instâncias de treinamento, referentes a apenas 10 meses para cada minirregião.

6 CONSIDERAÇÕES FINAIS

Esta tese trata de classificadores desenvolvidos com base na Teoria dos Conjuntos Aproximativos (TCA). Aborda mais especificamente metaheurísticas que permitem reduzir o número de atributos condicionais utilizados nas fases de treinamento e classificação por cada classificador. Em TCA, os conjuntos reduzidos de atributos condicionais são denominados reduções. O cálculo de reduções é um problema NP-difícil para bases de dados complexas, sendo contornado pelo uso de metaheurísticas que efetuam buscas no espaço de reduções/soluções. É preciso destacar que o cálculo de reduções em TCA constitui um importante mecanismo para tratamento de incertezas e imprecisões em bases de dados, além de reduzir o custo computacional demandado, como em qualquer abordagem de seleção de atributos. Esta tese propõe o uso inovador de algumas metaheurísticas (DCS, ILS e VNS) para o cálculo de reduções em TCA, sendo uma destas nova (DCS), implementado num software desenvolvido pelo autor. O desempenho dessas metaheurísticas foi comparado com o de outras metaheurísticas propostas anteriormente por outros autores para bases de dados de uso geral, com bons resultados.

Outro foco importante desta tese foi propor a mineração de dados meteorológicos com o objetivo de predição de atividade convectiva, no qual objetiva-se identificar padrões compostos por um conjunto de variáveis selecionadas do modelo ETA20 de previsão numérica do tempo, padrões indicativos de atividade convectiva ausente/fraca, moderada ou forte. No conjunto de treinamento, as instâncias são rotuladas com base na densidade de ocorrência de descargas NS. Estes padrões seriam então empregados na classificação das saídas do modelo. A mesma abordagem de TCA com o cálculo de reduções pelas metaheurísticas propostas foi utilizada para esses dados meteorológicos visando obter classificadores para 3 minirregiões diferentes do território brasileiro e para 3 horários diferentes de previsão. O desempenho de classificação foi promissor no esquema de *holdout* aleatório, mas ruim no esquema de *holdout* cronológico, o qual simularia uma predição

real. Isso pode ser atribuído ao relativamente pequeno número de instâncias da base de dados meteorológicos e à sazonalidade de certos fenômenos meteorológicos.

É importante ressalvar que as metaheurísticas propostas obtiveram bom desempenho de classificação com a base de dados meteorológicos no esquema *holdout* aleatório, especialmente a metaheurística nova DCS. Outro fator determinante desse bom desempenho foi o uso de partições aleatórias (PACT), o qual também permitiu a paralelização de forma a se efetuar o cálculo das partições concorrentemente.

O trabalho aqui exposto enquadra-se na mineração de dados meteorológicos, tema de pesquisa corrente em Meteorologia, dado o crescente volume de dados multidimensionais cuja análise demanda ferramentas semiautomáticas de auxílio ao meteorologista na previsão do tempo. Soma-se a isto a imprecisão de modelos numéricos na previsão de atividade convectiva, deficiência que motivou esta pesquisa. Mesmo com o aprimoramento destes modelos, no tocante ao uso de modelagem mais eficiente e adequada à região da microfísica, a qual compreende a parte de convecção, e no tocante a resoluções espaciais e temporais melhores, viabilizadas pela crescente capacidade de processamento disponível, a abordagem proposta ainda é válida, permitindo a predição da atividade convectiva e/ou até mesmo servindo de feedback para o aperfeiçoamento dos modelos. Obviamente, ainda falta muito para que a abordagem proposta possa ser utilizada operacionalmente, conforme foi exposto no capítulo de resultados. Porém, os resultados são promissores e podem-se vislumbrar possíveis trabalhos futuros para dar continuidade a esta pesquisa.

Como trabalhos futuros, pretende-se migrar para outro modelo numérico mais conveniente, pois a versão utilizada do ETA20 não é mais executada no CPTEC/INPE desde que houve a migração de um supercomputador vetorial (NEC SX-6) para um supercomputador massivamente paralelo (Tupã). A

metodologia proposta pode ser facilmente estendida para utilizar dados de outros modelos, o que possivelmente gerará resultados melhores, uma vez que novos modelos tem maior resolução espacial (o próprio ETA tem uma versão com resolução de 5 km) e saídas mais frequentes que as de periodicidade 6 h que foram empregadas aqui. Outra possibilidade seria utilizar um ensemble de previsões do mesmo modelo, em vez de uma única previsão como foi utilizado aqui. Adicionalmente, pretende-se explorar novos esquemas para geração de regras de decisão, por exemplo, considerando-se não apenas a melhor redução de cada partição, mas um pequeno conjunto destas. Há também a opção de, utilizando essas mesmas reduções, gerar um conjunto de classificadores que utilizam conjuntos mutuamente exclusivos de reduções para se obter resultados num esquema de votação (*polling*).

REFERÊNCIAS BIBLIOGRÁFICAS

BABIC, F. et al. Meteorological phenomena forecast using data mining prediction methods. **Computational Collective Intelligence. Technologies and Applications**, v. 6922, p. 458–467, 2012.

BACHE, K.; LICHMAN, M. **UCI Machine Learning Repository**. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 29 set. 2014.

BENETI, C. et al. Weather radar and lightning observations of mesoscale systems in the south of Brazil. In: WMO/WWRP International Symposium on Nowcasting and Very Short Range Forecasting. 3., Rio de Janeiro**. Anais**... WMO/WWRP. Rio de Janeiro: 2012. Disponível em: <http://www.labhidro.iag.usp.br/wsn12/papers/wea3.pdf>

BLACK, T. L. The New NMC Mesoscale Eta Model: Description and Forecast Examples. **Weather and forecasting Forecasting**, v. 9, n. 2, p. 265–278, 1994.

BOURSCHEIDT, V.; PINTO JR, O.; NACCARATO, K. Tracking thunderstorm cells using lightning density information. American Meteorological Society Annual Metting, 93., . **Proceedings**...Austin, Texas - EUA: 2002. Disponível em: <https://ams.confex.com/ams/93Annual/webprogram/Manuscript/Paper213984/manuscript_after_conf.pdf>. Acesso em: 29 set. 2014

BUCENE, L. C. **Mineração de dados climáticos para previsão local de geada e deficiência hídrica**. Tese: (Doutorado em Engenharia Agrícola) - Universidade Estadual de Campinas, Campinas, 2008.

CANO, R.; SORDO, C.; GUTIÉRREZ, J. M. Applications of Bayesian networks in meteorology. **Advances in Bayesian networks**, p. 309–327, 2004.

CAREY, L. D.; RUTLEDGE, S. A. The Relationship between Precipitation and Lightning in Tropical Island Convection: A C-Band Polarimetric Radar Study. **Monthly Weather Review**, v. 128, p. 2687–2710, 2000.

CHAVES, A. A. et al. Metaheurísticas híbridas para resolução do problema do caixeiro viajante com coleta de prêmios. **Revista Produção**, v. 17, n. 2, p. 263–272, 2007.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1 abr. 1960.

CORTEZ, P.; MORAIS, A. **A Data Mining Approach to Predict Forest Fires using Meteorological Data**13th Portuguese Conference on Artificial Intelligenc. **Anais**...Guimarães, Portugal: 2007. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/8039>. Acesso em: 29 set. 2014

DEMPSTER, A. P. Upper and lower probabilities induced by a multivalued mapping**.** **Annals of Mathematical Statistics** v. 38, n. 2, 1967, p.325-339.

DOLIF, G.; NOBRE, C. Improving extreme precipitation forecasts in Rio de Janeiro, Brazil: are synoptic patterns efficient for distinguishing ordinary from heavy rainfall episodes? **Atmospheric Science Letters**, v. 13, n. 3, p. 216–222, 2012.

DÜNTSCH, I.; GEDIGA, G. **Rough set data analysis: A road to non-invasive knowledge discovery**. Ontario, Canada: Methodos Primers, 2000. p. 108

DUTTA, P. S.; TAHBILDER, H. Prediction of rainfall using datamining technique over assam. **Indian Journal of Computer Science and Engineering**, v. 5, n. 2, p. 85–90, 2014.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–54, 1996.

GARCIA, J. V. C.; STEPHANY, S.; D'OLIVEIRA, A. B. Estimation of convective precipitation mass from lightning data using a temporal sliding-window for a series of thunderstorms in Southeastern Brazil. **Atmospheric Science Letters**, v. 14, n. 4, p. 281–286, 9 out. 2013.

GASPAR-CUNHA, A.; TAKAHASHI, R.; ANTUNES, C. H. **Manual de Computação Evolutiva e Metaheurística**. 1. ed. Belo Horizonte: Editora UFMG e Imprensa da Universidade de Coimbra, 2013. p. 453

GLOVER, F. Heuristics for Integer Programming Using Surrogate Constraints. **Decision Sciences**, v. 8, n. 1, p. 156–166, jan. 1977.

GLOVER, F.; MCMILLAN, C. The general employee scheduling problem. An integration of MS and AI**. Computers & Operations Research**, v. 13, n. 5, p. 563-573, 1986.

GOSS, S.; DENEUBOURG, J. L.; PASTEELS, J. M. Self-organized shortcuts in the argentine ant. **Naturwissenschaften**, v. 514, n. 1959, p. 579–581, 1989.

GUHA-SAPIR, D.; BELOW, R.; HOYOIS, P. **EM-DAT**: The International Disaster Database**.** Disponível em: <www.emdat.be>. Acesso em: 12 out. 2014.

GUPTA, K. M.; AHA, D. W.; MOORE, P. R**ough set feature selection algorithms for textual case-based classification.** European Conference on Case-Based Reasoning, 8. **Proceedings**...Olundeniz, Turquia: 2006

HAN, J.; SANCHEZ, R.; HU, X. Feature selection based on relative attribute dependency: An experimental study. **Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing**, v. 3641, n. 1, p. 214–223, 2005.

HANSEN, P. **The steepest ascent mildest descent heuristic for combinatorial programming.** Congress on Numerical Methods in Combinatorial Optimization. **Anais**...Capi, Italia: 1986. Disponível em:

<https://www.researchgate.net/publication/243633287_The_steepest_ascent_mildest_descent_heuristic_for_combinatorial_programming>

HANSEN, P.; MLADENOVIC, N. Variable neighborhood search. **Computers and Operations Research**, v. 24, p. 1097–1100, 1997.

HANSEN, P.; MLADENOVIC, N. **A tutorial on variable neighborhood search**. Montreal: [s.n.]. Disponível em: <http://yalma.fime.uanl.mx/~roger/work/teaching/mecbs5122/5-VNS/VNS-tutorial-G-2003-46.pdf>. Acesso em: 28 maio. 2012.

HEDAR, A.-R.; WANG, J.; FUKUSHIMA, M. Tabu search for attribute reduction in rough set theory. **Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications**, v. 12, n. 9, p. 909–918, 2008.

HOLLAND, J. H. **Adaptation in Natural and Artificial Systems**. [s.l.] University of Michigan Press, 1975. v. Ann Arborp. 1–200.

HRUSCHKA JR, E. R.; HRUSCHKA, E. R.; EBECKEN, N. F. F. Applying Bayesian networks for meteorological data mining. **Applications and Innovations in Intelligent Systems XIII**, v. 1, n. 1, p. 122–133, 2006.

JAN, Z. et al. Seasonal to inter-annual climate prediction using data mining KNN technique. **Wireless Networks, Information Processing and Systems**, v. 20, p. 40–51, 2009.

JENSEN, R.; SHEN, Q. Finding rough set reducts with ant colony optimization. **Proceedings of the 2003 UK workshop on Computational Intelligence**, 2003.

JENSEN, R.; SHEN, Q. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. **Knowledge and Data Engineering, IEEE**, v. 17, n. 1, p. 1–15, 2005.

JOHNSON, D. S. **Approximation algorithms for combinatorial problems**Proceedings of the fifth annual ACM symposium on Theory of computing - STOC '73. **Anais**...New York, New York, USA: ACM Press, 1973. Disponível em: <http://portal.acm.org/citation.cfm?doid=800125.804034>. Acesso em: 22 ago. 2012

JOHNSON, D. S.; MCGEOCH, L. A. The Traveling Salesman Problem: A Case Study in Local Optimization. In: AARTS, E. H. L.; LENSTRA, J. K. (Eds.). **Local Search in Combinatorial Optimization**. Chichester, Reino Unido: John Wiley & Sons, 1997. p. 215–310.

KANTH, T. V. R.; BALARAM, V. V. S. S. S.; RAJASEKHAR, N. **ANALYSIS OF INDIAN WEATHER DATA SETS USING DATA MINING TECHNIQUES.** (D. Nagamalai, S. Vaidyanathan, Eds.)Fourth International Conference on Advances in Computing and Information Technology. **Anais**...Delhi, India: 2014. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Analysis+of+Indian+Weather+Data+Sets+Using+Data+Mining+Techniques#0>. Acesso em: 29 set. 2014

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science**, v. 220, n. 4598, p. 671–680, 1983.

KOHAIL, S. N.; EL-HALEES, A. M. Implementation of Data Mining Techniques for Meteorological Data Analysis. **International Journal of Information and Communication Technology Research**, v. 1, n. 3, p. 96–100, 2011.

KOMOROWSKI, J. et al. **Rough sets: A tutorial**. Warsaw, Poland: [s.n.]. Disponível em: <http://secs.ceas.uc.edu/~mazlack/dbm.w2011/Komorowski.RoughSets.tutor.pdf>. Acesso em: 20 ago. 2012.

KORTH, H. F.; SILBERSCHATZ, A. **Sistema de Banco de Dados**. 2. ed. São Paulo: Ed. Makron Books, 1993. p. 748

KULIGOWSKI, R. J.; SCOFIELD, R. A. Moving toward multi-spectral, multi-platform operational satellite precipitation estimates at nesdis. **II Workshop of the International Precipitation Working Group**, p. 1–9, 2005.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, p. 159–174, 1977.

LIMA, G. R. T. et al. **Mineração de dados meteorológicos para previsão de eventos severos pela abordagem de similaridade de vetores.** XXXIII Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2010

LIMA, G. R. T.; STEPHANY, S. **Comparação de abordagens de classificação para um problema de mineração de dados meteorológicos.** XXXIV Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2012

LIMA, G. R. T.; STEPHANY, S. Training a neural network to detect patterns associated with severe weather. **Learning and Nonlinear Models**, v. 11, n. 2, p. 123–152, 2013a.

LIMA, G. R. T.; STEPHANY, S. A new classification approach for detecting severe weather patterns. **Computers & Geosciences**, v. 57, p. 158–165, ago. 2013b.

LITTLE, M. A.; RODDA, H. J. E.; MCSHARRY, P. E. Bayesian objective classification of extreme UK daily rainfall for flood risk applications. **Hydrology and Earth System Sciences**, v. 5, p. 3033–3060, 2008.

LOURENÇO, H. R.; MARTIN, O.; STÜTZLE, T. Iterated Local Search. In: GLOVER, F.; KOCHENBERGER, G. A. (Eds.). **Handbook of Metaheuristics**. Norwell, Estados Unidos: Kluwer Academic Publishers, 2002. p. 321–353.

MACGORMAN, D. R.; RUST, W. D. **The Electrical Nature of Storms**. 1. ed. Nova Iorque, EUA: Oxford University Press, 1998. p. 432

MACHADO, L. A. T. et al. Relationship between cloud-to-ground discharge and penetrative clouds: A multi-channel satellite application. **Atmospheric Research**, v. 93, n. 1-3, p. 304–309, jul. 2009.

MARTIN, O.; OTTO, S. W.; FELTEN, E. W. Large-step Markov chains for the traveling salesman problem. **Complex Systems**, v. 5, p. 299–326, 1991.

MATTOS, E. V.; MACHADO, L. A. T. Cloud-to-ground lightning and Mesoscale Convective Systems. **Atmospheric Research**, v. 99, n. 3-4, p. 377–390, mar. 2011.

MESINGER, F. et al. The step-mountain coordinate: model description and performance for cases of alpine lee cyclogenesis and for a case of an appalachian redevelopment. **Monthly Weather Review**, v. 116, n. 7, p. 1493–1518, 1988.

NICOLETTI, M. DO C.; UCHÔA, J. Q. **Conjuntos aproximados sob a perspectiva de função de pertinência.** Simpósio Brasileiro de Automação Inteligente, 3. **Anais**...Vitória - ES, Brazil: 1997. Disponível em: <http://www.ginux.ufla.br/~joukim/publicacoes/artigo3sbai.ps.gz>. Acesso em: 20 ago. 2012

OHRN, A. **Discernibility and rough sets in medicine:** tools and applications. [s.l.] Norwegian University of Science and Technology, 1999.

OLAIYA, F.; ADEYEMO, A. B. Application of data mining techniques in weather prediction and climate change studies. **International Journal of Information Engineering and Electronic Business**, v. 4, n. 1, p. 51–59, 27 fev. 2012.

OLIVEIRA, R. A. J.; MATTOS, E. V. **The spatial-temporal relationship between cloud-to- ground lightning and precipitation distributions in the State of São Paulo.** XIV International Conference on Atmospheric Electricity. **Anais**...Rio de Janeiro: 2011. Disponível em: <http://mtc-m19.sid.inpe.br/col/sid.inpe.br/mtc-m19/2011/07.07.16.51/doc/Mattos_ the spatial_temporal.pdf?languagebutton=en>

PAWLAK, Z. Rough sets. **International Journal of Computing and Information Sciences**, v. 11, n. 5, p. 341–356, 1982.

PESSOA, A. S. A. et al. Mineração de dados meteorológicos para previsão de eventos severos. **Revista Brasileira de Meteorologia**, v. 27, n. 1, p. 287–294, 2012.

PESSOA, A. S. A.; SILVA, J. D. S. DA; JUNIOR, C. H. **Redução de dados meteorológicos aplicados a previsão climática por redes neurais.** XIV Congresso Brasileiro de Meteorologia. **Anais**...Florianópolis: 2006. Disponível em: <http://www.cbmet.com/cbm-files/14-14147710b7b4d0fd7f69720ab6a03e32.pdf>

PESSOA, A. S. A.; STEPHANY, S. **Seleção de atributos com novas metaheurísticas na teoria de conjuntos aproximativos.** XXXIV Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2012a.

PESSOA, A. S. A.; STEPHANY, S. **Mineração de dados meteorológicos pela teoria dos conjuntos aproximativos utilizando algoritmo de johnson e particionamento aleatório.** XXXIV Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2012b.

PESSOA, A. S. A.; STEPHANY, S. An innovative approach for attribute reduction in rough set theory. **Intelligent Information Management**, v. 06, n. 05, p. 223–239, 2014.

PETERS, J. F. et al. Classification of meteorological volumetric radar data using rough set methods. **Pattern Recognition Letters**, v. 24, n. 6, p. 911–920, mar. 2003.

PETERSEN, W. A.; RUTLEDGE, S. A.; ORVILLE, R. E. Cloud-to-ground lightning observations from TOGA COARE: selected results and lightning location algorithms. **Monthly Weather Review**, v. 124, p. 602–620, 1996.

PING, Y.; WEI, H.; GUO-LIN, F. The characteristics of clusters of weather and extreme climate events in China during the past 50 years. **Chinese Physics B**, v. 21, n. 1, p. 019201, 26 jan. 2012.

RAMÍREZ, M. C. V.; FERREIRA, N. J.; VELHO, H. F. DE C. Linear and nonlinear statistical downscaling for rainfall forecasting over southeastern Brazil. **Weather and Forecasting,** v. 21, n. 6, p. 969–989, 2006.

Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT). Localização dos sensores da rede RINDAT**.** Disponível em: <www.inpe.br/webelat/rindat/imagens/Rede_RINDAT_24ss_2008.jpg>. Acesso em: 2 out. 2014.

REZENDE, S. O. **Sistema Inteligentes: Fundamentos e Aplicações**. 1. ed. Barueri: Ed. Manole, 2005. p. 525

SENCAN, H. et al. **Classification of emerging extreme event tracks in multivariate spatio-temporal physical systems using dynamic network structures:** Application to hurricane track prediction**.** XXII International Joint Conference on Artificial Intelligence. **Anais**...Barcelona, Espanha: 2011. Disponível em: <http://ijcai.org/papers11/Papers/IJCAI11-249.pdf>

SHAFER, G. **A mathematical theory of evidence**. Estados Unidos: Princeton University Press, 1976. p. 376

SHARMA, A. **Spatial data mining for drought monitoring:** an approach using temporal NDVI and rainfall relationship. [s.l.] International Institute for Geo-Information Science and Earth Observation, 2006.

STRAUSS, C. et al. **Análise quantitativa das regras da ferramenta objetiva de previsão de tempo do CPTEC.** XVII Congresso Brasileiro de Meteorologia. **Anais**...2012

STRAUSS, C. **Monitoramento e previsão de atividade convectiva usando abordagens de mineração de dados**. São José dos Campos. Instituto Nacional de Pesquisas Espaciais, 2013.

STRAUSS, C.; ROSA, M. B.; STEPHANY, S. Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events. **Atmospheric Research**, v. 134, p. 87–99, dez. 2013.

STRAUSS, C.; STEPHANY, S.; CAETANO, M. **A ferramenta EDDA de geração de campos de densidade de descargas atmosféricas para mineração de dados meteorológicos.** XXXIII Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2010

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 4. ed. Canada: Academic Press, 2009. p. 961

TSAGALIDIS, E.; EVANGELIDIS, G. **Meteorological data mining:** exploiting domain expertise in the class imbalance problem**.** 7th International Conference on Advanced Data Mining and Applications (ADMA). **Anais**...Pequim, China: 2011. Disponível em: <http://arnetminer.org/PDF/adma2011/session3D/ADMA_conf_100.pdf>. Acesso em: 29 set. 2014

VESTERGRAARD, J. S. **Improved nowcasting of heavy precipitation using satellite and weather radar data**. [s.l.] Technical University of Denmark, 2011.

VISACRO FILHO, S. **Descargas atmosféricas:** uma abordagem de engenharia. 1. ed. São Paulo: Artliber, 2005. p. 268

WALLACE, J. M.; HOBBS, P. V. **Atmospheric Science, Second Edition:** An Introductory Survey. 2. ed. Canada: Academic Press, 2006. p. 504

WANG, J. et al. A rough set approach to feature selection based on scatter search metaheuristic. **Journal of Systems Science and Complexity**, v. 27, n. 1, p. 157–168, 2 fev. 2014.

ZADEH, L. A. Fuzzy Sets. **Information and Control**, v. 8, p. 338–353, 1965.

ZADEH, L. A. Fuzzy sets as a basis for a theory of possibility. **Fuzzy Sets and Systems**, v. 1, p. 3–28, 1978.

ZHOU, Y.; QIE, X.; SOULA, S. A study of the relationship between cloud-to-ground lightning and precipitation in the convective weather system in China. **Annales Geophysicae**, p. 107–113, 2002.

ANEXO A – ARTIGOS PUBLICADOS RELACIONADOS À TESE

PESSOA, A. S. A.; STEPHANY, S. An Innovative Approach for Attribute Reduction in Rough Set Theory. **Intelligent Information Management**, v. 06, n. 05, p. 223–239, 2014.

PESSOA, A. S. A. et al. Mineração de Dados Meteorológicos para Previsão de eventos Severos. **Revista Brasileira de Meteorologia**, v. 27, n. 1, p. 287–294, 2012.

PESSOA, A. S. A.; STEPHANY, S. **Seleção de Atributos com Novas Metaheurísticas na Teoria de Conjuntos Aproximativos.** XXXIV Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2012a.

PESSOA, A. S. A.; STEPHANY, S. **Mineração de Dados Meteorológicos pela Teoria dos Conjuntos Aproximativos Utilizando Algoritmo de Johnson e Particionamento Aleatório.** XXXIV Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2012b.

PESSOA, A. S. A.; STEPHANY, S. An Innovative Approach for Attribute Reduction in Rough Set Theory. **Intelligent Information Management**, v. 06, n. 05, p. 223–239, 2014.

PESSOA, A. S. A. et al. **Mineração de dados meteorológicos associada a eventos severos no Pantanal Sul Matogrossense**XXXIII Congresso Nacional de Matemática Aplicada e Computacional. **Anais**...Águas de Lindóia: 2010. Disponível em: <http://bibdigital.sid.inpe.br/dpi.inpe.br/plutao/2010/11.11.16.31.33>. Acesso em: 18 set. 2014.

Intelligent Information Management, 2014, 6, 223-239
Published Online September 2014 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2014.65022

Scientific Research

An Innovative Approach for Attribute Reduction in Rough Set Theory

Alex Sandro Aguiar Pessoa, Stephan Stephany

National Institute for Space Research (INPE), Sao Jose dos Campos, Brazil
Email: asapessoa@gmail.com, stephan.stephany@lac.inpe.br

Received 4 July 2014; revised 3 August 2014; accepted 25 August 2014

Abstract

The Rough Sets Theory is used in data mining with emphasis on the treatment of uncertain or vague information. In the case of classification, this theory implicitly calculates reducts of the full set of attributes, eliminating those that are redundant or meaningless. Such reducts may even serve as input to other classifiers other than Rough Sets. The typical high dimensionality of current databases precludes the use of greedy methods to find optimal or suboptimal reducts in the search space and requires the use of stochastic methods. In this context, the calculation of reducts is typically performed by a genetic algorithm, but other metaheuristics have been proposed with better performance. This work proposes the innovative use of two known metaheuristics for this calculation, the Variable Neighborhood Search, the Variable Neighborhood Descent, besides a third heuristic called Decrescent Cardinality Search. The last one is a new heuristic specifically proposed for reduct calculation. Considering some databases commonly found in the literature of the area, the reducts that have been obtained present lower cardinality, *i.e.*, a lower number of attributes.

Keywords

Rough Set Theory, Reducts, Attribute Reduction, Metaheuristics

1. Introduction

Large amounts of data are generated everyday and the ability to analyze them is normally a challenge. Experts need efficient data mining methods to extract useful information and to perform the analysis of the data. This is the case of the Rough Sets Theory (RST), emerging in the early 80s [1] to deal with uncertain, incomplete or vague information. In addition, RST has a good mathematical formalism, and it is easy to use, since it does not require additional information, such as the probability distribution, *a priori* probability or pertinence degree.

How to cite this paper: Pessoa, A.S.A. and Stephany, S. (2014) An Innovative Approach for Attribute Reduction in Rough Set Theory. *Intelligent Information Management*, **6**, 223-239. http://dx.doi.org/10.4236/iim.2014.65022

Revista Brasileira de Meteorologia, **v.27**, **n.1**, 61 - 74, 2012

MINERAÇÃO DE DADOS METEOROLÓGICOS PARA PREVISÃO DE EVENTOS SEVEROS

ALEX SANDRO AGUIAR PESSOA[1], GLAUSTON ROBERTO TEIXEIRA DE LIMA[1], JOSÉ DEMÍSIO SIMÕES DA SILVA[1], STEPHAN STEPHANY[1], CESAR STRAUSS[1], MIRIAN CAETANO[2] E NELSON JESUS FERREIRA[2]

[1]Instituto Nacional de Pesquisas Espaciais, Laboratório Associado de Computação e Matemática Aplicada (INPE/ LAC)
[2]Centro de Previsão de Tempo e Estudos Climáticos (INPE/CPTEC), São José dos Campos, SP, Brasil

asapessoa@gmail.com, glau11@gmail.com, demisio@lac.inpe.br, stephan@lac.inpe.br, cstrauss@cea.inpe.br, mirian.caetano@cptec.inpe.br, nelson.ferreira@cptec.inpe.br

Recebido Agosto de 2010 – Aceito Julho de 2011

RESUMO

O objetivo do trabalho proposto é detectar antecipadamente possíveis ocorrências de eventos convectivos severos, por meio do monitoramento das saídas do modelo de previsão numérica de tempo Eta, para cada intervalo de previsão e para um conjunto de variáveis selecionadas. O período de estudo estende-se de janeiro a fevereiro de 2007. Classificadores foram desenvolvidos pela abordagem de similaridade de vetores e de conjuntos aproximativos, de forma a identificar saídas do modelo Eta que possam ser associados a esses eventos. Assumiu-se como premissa que os eventos convectivos severos possam ser correlacionados com grande número de ocorrências de descargas elétricas atmosféricas. Os classificadores agruparam as saídas do modelo Eta, compostas por essas variáveis, com base na densidade de ocorrência de descargas elétricas atmosféricas nuvem-solo. Ambos os classificadores apresentaram bom desempenho para os testes realizados para um período de dois meses escolhido para três mini-regiões selecionadas do território brasileiro.
Palavras-Chave: mineração de dados, previsão meteorológica, eventos convectivos.

ABSTRACT: METEOROLOGICAL DATA MINING FOR THE PREDICTION OF SEVERE CONVECTIVE EVENTS

This work aims the early detection of possible occurrences of severe convective events in Central and Southeast Brazil by means of monitoring the output of the Eta numerical weather prediction model for each forecasted time interval and for a selected set of variables. The studied period ranges from January to February 2007. Classifiers were developed by two approaches, vector similarity and rough sets, in order to identify Eta outputs that can be associated to such events. It was assumed that severe convective events can be correlated to a large number of atmospheric electric discharges. The classifiers grouped the Eta meteorological model outputs for these selected variables based on the density of occurrences of cloud-to-ground atmospheric electrical discharges. Both classifiers show good performance for the chosen 2-month period at the three selected mini-regions of the Brazilian territory.
Keywords: data mining, weather forecast, convective events.

1. INTRODUÇÃO

A previsão de eventos convectivos severos de forma semi-automática e com antecedência desejável é um tema atual de pesquisa em Meteorologia. A necessidade de análise da crescente quantidade de dados meteorológicos e imagens, gerados por sensores ou por modelos de previsão numérica de tempo, demanda técnicas computacionais avançadas. Nesse escopo, um dos objetivos da mineração de dados é descobrir correlações potencialmente úteis entre os diversos dados ou encontrar regras quantitativas associadas aos mesmos (Fayyad et al., 1996).

No caso do presente trabalho, tenta-se inferir a possibilidade de ocorrência de eventos convectivos severos a partir das saídas do modelo de previsão numérica de tempo Eta,

CNMAC2012
XXXIV CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL
17 a 21 de setembro de 2012 - Águas de Lindóia/SP

ISSN 1984-8218

Seleção de Atributos com Novas Metaheurísticas na Teoria de Conjuntos Aproximativos

Alex Sandro Aguiar Pessoa,
INPE - Programa de Pós-graduação em Computação Aplicada (CAP)
12.227-010, São José dos Campos, SP
E-mail: asapessoa@gmail.com

Stephan Stephany
INPE – Laboratório Associado de Computação e Matemática Aplicada (LAC)
12.227-010, São José dos Campos, SP
E-mail: stephan@lac.inpe.br

***Resumo**: Técnicas de seleção de atributos são aplicadas em algoritmos de aprendizagem de máquina para reduzir a dimensionalidade de uma base de dados com o objetivo de melhorar a qualidade dos resultados e o desempenho computacional. A Teoria dos Conjuntos Aproximativos é empregada em mineração de dados com ênfase no tratamento de informações incertas e imprecisas. No caso da classificação, esta teoria implicitamente calcula reduções de atributos, eliminando aqueles que são supérfluos. Neste contexto, o cálculo das reduções é tipicamente efetuado por um algoritmo genético, sendo que o presente trabalho propõe o uso inovador de algumas metaheurísticas para este cálculo. Os resultados apresentados mostram que foi possível obter reduções com menor cardinalidade, ou seja, com menor número de atributos, para algumas bases de dados comuns na literatura da área.*

Palavras-chave: *seleção de atributos, teoria dos conjuntos aproximativos, metaheurísticas.*

1. Introdução

A Teoria dos Conjuntos Aproximativos (TCA), do termo em inglês *Rough Set Theory*, foi proposta por Pawlak [1], vem se mostrando uma ferramenta muito eficiente e eficaz no tratamento de incerteza em bases de dados, que surgem com inexatidão, ruídos ou informações incompletas. A TCA vem se difundindo como técnica de mineração de dados, em particular na classificação [2]. Suas principais características são o bom formalismo matemático, facilidade de uso, não requerer informações adicionais, tais como grau de pertinência ou probabilidade *a priori* e compactação de bases de dados. Esta última característica é decorrente de dois conceitos que são inerentes à TCA: a relação de indiscernibilidade e a redução de atributos. Enquanto a indiscernibilidade diminui o número de objetos ou elementos de uma base de dados, a redução provê a diminuição do número de atributos ou variáveis redundantes, em perda de informações.

A indiscernibilidade é uma relação de equivalência que particiona o universo (conjunto finito contendo todos os elementos de uma base de dados) de acordo com classes. Estas classes são rótulos atribuídos a elementos que possuem determinadas características em comum, expressas por meio de um atributo de decisão.

Para cada classe e para o conjunto de atributos considerado, particiona-se o universo em 3 regiões: uma região interna composta por objetos que certamente pertencem à classe, uma região externa, de elementos que certamente não pertencem a ela e uma região de borda ou fronteira, de elementos que, embora indiscerníveis em relação a esse conjunto de atributos, estão rotulados ou não como pertencentes à classe considerada. Considerando-se a região interna, todos os elementos pertencem à classe considerada para o conjunto de atributos definido, mas pode-se subdividi-los em classes de equivalência em função dos valores de seus atributos. Assim, um único elemento pode representar todos os demais de sua classe de equivalência, reduzindo significativamente o número de objetos da base de dados.

CNMAC2012
XXXIV CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL
17 a 21 de setembro de 2012 - Águas de Lindóia/SP

ISSN 1984-8218

Mineração de Dados Meteorológicos pela Teoria dos Conjuntos Aproximativos Utilizando Algoritmo de Johnson e Particionamento Aleatório

Alex Sandro Aguiar Pessoa,
INPE - Programa de Pós-graduação em Computação Aplicada
12.227-010 São José dos Campos, SP
E-mail: asapessoa@gmail.com

Stephan Stephany
INPE – Laboratório Associado de Computação e Matemática Aplicada (LAC)
12.227-010 São José dos Campos, SP
E-mail: stephan@lac.inpe.br

***Resumo**: Este trabalho busca a detecção de padrões associados à ocorrência de eventos convectivos severos em dados meteorológicos. Utiliza-se a Teoria dos Conjuntos Aproximativos conjuntamente com o algoritmo de Johnson, que é uma heurística que simplifica o cálculo das reduções. Adota-se também um esquema de particionamento dos dados de treinamento de forma a viabilizar a mineração de grandes volumes de dados.*

Palavras-chave: *mineração de dados, eventos convectivos, previsão meteorológica, Teoria dos Conjuntos Aproximativos.*

1. Introdução

A grande quantidade e multiplicidade de dados e imagens meteorológicas gerados por modelos numéricos e sensores, embarcados em satélites ou não, torna mais complexo o trabalho do meteorologista na previsão do tempo. Assim, ferramentas auxiliares tornam-se desejáveis e ultimamente a aplicação de técnicas de mineração de dados vem se expandindo. Um dos objetivos é a detecção automática nos dados de padrões associados a determinados fenômenos meteorológicos. Em particular, a detecção de eventos convectivos severos tem importância devido a seu impacto sócio-econômico, sendo objeto do presente trabalho.

Eventos convectivos severos meteorológicos são fenômenos associados a chuvas e ventos fortes de curta duração, ou então de intensidade média, porém de duração prolongada que, em geral, causam sérios danos, tornando sua previsão altamente desejável. Eventos deste tipo, devido à sua baixa freqüência são de difícil previsão por meteorologistas, embora existam esquemas de alerta antecipado. Neste contexto, este trabalho aborda a mineração de dados meteorológicos com o objetivo de detectar padrões associados à atividade convectiva severa nos dados do modelo numérico de previsão de tempo Eta [1] por meio de um classificador baseado na Teoria dos Conjuntos Aproximativos (TCA).

Se o modelo numérico fosse suficientemente preciso, poderia-se detectar atividade convectiva severa analisando-se certas variáveis nas saídas do próprio modelo. Obviamente isso não se verifica, embora o modelo Eta, em particular, possa simular com razoável precisão variáveis tais como pressão, temperatura, conteúdo de umidade ou ventos em vários níveis de pressão. Optou-se então pelo uso de dados de descargas elétricas atmosféricas, assumindo-se que uma grande quantidade destas possa ser indicativa de atividade convectiva severa.

Os dados de descargas são tratados e agrupados espaço-temporalmente por meio de uma técnica de análise espacial aplicada de maneira inovadora a esse tipo de dados [2][3]. Esse agrupamento gera um campo de densidade de ocorrências de descargas que permite identificar regiões mais densas como sendo núcleos de atividade elétrica (NAEs). O campo de densidade

ISSN 1984-820X
Anais do CNMAC
v.3
PUBLICAÇÃO DA SBMAC
CNMAC2010
XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

Mineração de Dados Meteorológicos Associada a Eventos Severos no Pantanal Sul Matogrossense

Alex Sandro Aguiar Pessoa,
INPE - Programa de Pós-graduação em Computação Aplicada
12.227-010 São José dos Campos, SP
E-mail: asapessoa@gmail.com

José Demísio Simões da Silva, **Stephan Stephany,**
INPE – Laboratório Associado de Computação e Matemática Aplicada (LAC)
12.227-010 São José dos Campos, SP
E-mail: demisio@lac.inpe.br, stephan@lac.inpe.br

César Strauss,
INPE – Coordenação de Ciências Espaciais e Atmosféricas (CEA)
12.227-010 São José dos Campos, SP
E-mail: cstrauss@cea.inpe.br

Mirian Caetano, **Nelson Jesus Ferreira**
INPE – Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)
12.630-000 Cachoeira Paulista, SP
E-mail: mirian.caetano@cptec.inpe.br, nelson.ferreira@cptec.inpe.br

Resumo: *O objetivo do trabalho proposto é detetar possíveis ocorrências de eventos convectivos severos por meio do monitoramento das saídas do modelo meteorológico Eta para cada timestep simulado e para um conjunto de variáveis selecionadas. Um classificador foi desenvolvido pela abordagem de conjuntos aproximativos de forma a identificar saídas referentes a timesteps simulados do modelo que possam ser associados a esses eventos. Assumiu-se como premissa que os mesmos possam ser correlacionados com grande número de ocorrências de descargas elétricas atmosféricas. O classificador foi treinado agrupando-se saídas do modelo Eta compostas por essas variáveis com base na densidade de ocorrência de descargas elétricas atmosféricas nuvem-solo. O classificador apresentou ótimo desempenho para os testes realizados para o Pantanal Sul Matogrossense.*

Palavras-chave: mineração de dados, previsão meteorológica, eventos convectivos, Teoria dos Conjuntos Aproximativos

1. Introdução

A previsão de eventos convectivos severos de forma semi-automática e com antecedência desejável é um tem atual de pesquisa em Meteorologia. A necessidade de análise da crescente quantidade de dados e imagens meteorológicos, gerados por sensores ou por simulações, demanda técnicas computacionais avançadas. Nesse escopo, um dos objetivos da mineração de dados é descobrir correlações potencialmente úteis entre os diversos dados ou encontrar regras quantitativas associadas aos mesmos.

No caso do presente trabalho, tenta-se inferir a possibilidade de ocorrência de eventos convectivos severos a partir das saídas do modelo meteorológico regional Eta, as quais fornecem o valor simulado de muitas dezenas de variáveis meteorológicas para um tempo de simulação futuro. Um classificador é o programa que atribui uma classe para o conjunto de valores das variáveis meteorológicas de cada timestep gerado pelo modelo meteorológico. As classes compreendem, por exemplo, evento convectivo severo, ou de média ou fraca intensidade, ou ainda ausência de atividade convectiva. O classificador incorpora conceitos de aprendizagem de máquina, os quais possibilitam que o mesmo seja "treinado" a partir de um conjunto de instâncias conhecidas. No caso, as instâncias são o conjunto de saídas do modelo Eta para os quais a intensidade da atividade convectiva é conhecida de forma indireta por meio