Revista Brasileira de Meteorologia, **v.24**, **n.4**, 448 - 460, 2009

VEDALOGIC – UM MÉTODO DE VERIFICAÇÃO DE DADOS CLIMATOLÓGICOS APOIADO EM MODELOS MINERADOS

HENRIQUE GONÇALVES SALVADOR[1], ADILSON MARQUES DA CUNHA[1] E CLEBER SOUZA CORRÊA[2]

[1]ITA – Instituto de Tecnologia de Aeronáutica, São José dos Campos/SP, Brasil
[2] ICEA – Instituto de Controle do Espaço Aéreo, São José dos Campos/SP, Brasil

henriquesalvador@gmail.com, cunha@ita.br e cleber@icea.gov.br

Recebido Janeiro 2008 - Aceito Maio 2009

RESUMO

Neste artigo, apresenta-se um Método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados – VEDALOGIC para o Instituto de Controle do Espaço Aéreo Brasileiro (ICEA). O VEDALOGIC consiste de uma verificação de dados, utilizando-se de modelos criados com algoritmos de Mineração de Dados. O Método utiliza modelos de *clustering*, gerados a partir de uma série histórica, que propiciam a identificação de grupos homogêneos em uma Base de Dados Climatológicos (BDC). A partir desses modelos, torna-se possível a detecção de inconformidades nos dados, denominadas pontos estranhos (*outliers*). Após a detecção de um *outlier*, este é classificado/predito, de acordo com o modelo de árvore de decisão, gerado também a partir de uma série histórica. O valor encontrado com base na árvore de decisão é adotado como sugestão para a correção do *outlier*, contribuindo com a consistência dos dados no BDC. Neste artigo, utilizam-se os seguintes algoritmos: *Expectation-Maximization* (*EM*) e K-means para *clustering*; e REPTree e M5P para classificação/predição. Para a verificação da eficiência do VEDALOGIC, inseriram-se, artificialmente, dados ruidosos em um conjunto de dados, os quais foram todos detectados pelo VEDALOGIC, que sugeriu valores para correção com uma precisão média superior a 98%.
Palavras-Chaves: Mineração de Dados; Banco de Dados Climatológicos; *Clustering*; Verificação de Dados.

ABSTRACT: VEDALOGIC – A METHOD OF CLIMATOLOGIC DATA VERIFICATION BASED ON DATA MINING MODELS

This work presents the VEDALOGIC - Method for Climatologic Data Verification – based on Data Mining Models, to be used by the "Instituto de Controle do Espaço Aéreo Brasileiro" (ICEA). The VEDALOGIC method consists of a data verification using Data Mining algorithm models. The method uses clustering models generated from a historical series that provide the identification of homogeneous groups in the Climatologic Data Base (CDB). This method, based on clustering models, detects unconformities, named outliers. Detected outliers are classified/predicted according to the decision tree *models* which are also built from historic data. The found value based on the decision tree model is used as a suggestion to correct an outlier, contributing to increase the CDB data consistence. In this study, the Expectation-Maximization (EM) and the K-means algorithms were used to generate clustering models, and the REPTree and the M5P algorithms were used to generate decision (classification/prediction) tree models. To verify the efficiency of the proposed method, some noisy data were artificially inserted into CDB. After applying the VEDALOGIC method, all inserted noisy data were detected and the adjustments have an average precision above 98%.

1. INTRODUÇÃO

Esta pesquisa foi motivada pela necessidade do Instituto de Controle de Espaço Aéreo (ICEA) desenvolver um verificador para as inserções de dados, em uma Base de Dados Climatológicos (BDC), que considerasse não apenas verificações por intervalo de valores de cada atributo, mas também as inter-relações entre eles. Ela considera a existência, naquela Instituição, de uma série histórica climatológica contendo dados das últimas cinco décadas, em processo de digitalização gradativa, bem como o armazenamento no BDC, sem uma verificação de consistência, o que impacta na confiabilidade dos dados armazenados.

Neste artigo, será apresentado o Método VEDALOGIC com a geração de modelos de verificação dos dados, utilizando-se das técnicas de mineração de dados. Em um primeiro momento, realizar-se-á o pré-processamento. Em seguida, se aplicará os algoritmos de mineração de dados, adotando-se as técnicas que apresentam resultados com qualidade de classificação e de predição de valores para correção com precisão aceitável e que melhor identifiquem grupos homogêneos no BDC.

2. MINERAÇÃO DE DADOS

Com a evolução computacional, as empresas se especializaram em armazenar informações. Em pouco tempo, essas empresas tiveram sua capacidade de análise de informações superadas, devido ao grande volume armazenado em seus repositórios. A geração desses grandes repositórios foi impulsionada pelos seguintes fatores (Fayyad *et al*., 1996; Goebel e Gruenwald, 1999):

Diminuição dos custos das tecnologias utilizadas para o armazenamento de dados;

Aumento na velocidade e capacidade dos sistemas utilizados;

Melhoria dos Sistemas Gerenciadores de Banco de Dados (SGBDs); e

Popularização dos Armazéns de Dados (*Data Warehouse* - *DW*).

A criação de grandes repositórios gerou a necessidade de um processo que propiciasse a exploração de dados com maior eficiência e agilidade, bem como a extração de informações e conhecimentos novos e implícitos, que auxiliem nas tomadas de decisões (Goebel e Gruenwald, 1999). Com o intuito de gerar conhecimento, a partir desses dados, uniu-se as áreas de inteligência artificial, estatística e banco de dados, resultando em um processo conhecido como Descoberta de Conhecimento em Banco de Dados – DCBD (*Knowledge Discovery in Databases* - *KDD*) (Hand, 1998).

O conhecimento obtido por um processo de transformação de dados em conhecimentos (DCBD) refere-se ao conjunto de etapas, dentre as quais se destaca a etapa de Mineração de Dados (MD) (Berry e Linoff, 2004). A etapa de MD tem recebido maior destaque na literatura, devido a sua importância no processo de DCBD, em alguns casos, adotado como sinônimo do processo como um todo (Imberman, 2001).

A Mineração de Dados utiliza-se de algoritmos para a busca de padrões em grandes volumes de dados, propiciando a criação de modelos. Pode-se usar Mineração de Dados para verificação de hipóteses ou para descobrir novos padrões de comportamento nos dados (Imberman, 2001).

2.1. Descoberta de Conhecimento em Banco de Dados – DCBD

O processo de Descoberta de Conhecimento em Banco de Dados – DCBD tem por objetivo a extração do conhecimento implícito e previamente desconhecido, e a busca da informação potencialmente útil nos dados, por intermédio da intersecção de diferentes áreas, como a aprendizagem de máquina, banco de dados, inteligência artificial, estatística, reconhecimento de padrões, visualização dos dados, entre outras (Fayyad *et al*., 1996; Han e Kamber, 2006).

O conjunto de etapas do processo de DCBD pode ser executado, de forma interativa e iterativa (Fayyad et al. 1996b). Interativa porque envolve a cooperação da pessoa responsável pela análise dos dados com o(s) especialista(s) do domínio do problema, cujo conhecimento sobre o domínio orientará a execução do processo e validará os resultados. Por sua vez, a iteração deve-se ao fato de que, com freqüência, esse processo não se executa de forma seqüencial, envolvendo repetidas seleções de parâmetros e conjuntos de dados, aplicações de técnicas de MD e posterior análise dos resultados obtidos, a fim de refinar o conhecimento extraído.

A descrição do processo apresentada por Fayyad *et al.* (1996), menciona os passos básicos do processo de DCBD, que partindo de dados disponíveis e, normalmente, da definição de um problema, conduzem à descoberta do conhecimento. O processo DCBD, ilustrado na Figura 1, possui as seguintes etapas (Fayyad *et al*., 1996):

Conhecendo o Problema - Representa o ponto de partida do processo de DCBD. Nesta etapa, levantam-se os problemas e/ou objetivos, assim como o reconhecimento dos dados disponíveis e as necessidades de dados complementaes;

Seleção dos Dados - Nesta etapa, cria-se o conjunto de dados que servirá de base para o processo de DCBD, por meio da seleção de conjuntos de origem, de um subconjunto das variáveis, ou ainda, de uma amostra. Este conjunto de dados representa os dados extraídos de um Banco de Dados operacional ou de um DW, criado para servir às diversas necessidades de análise;

Limpeza e Integração - Refere-se ao Pré-Processamento. Nesta etapa, decide-se as estratégias e realiza-se as devidas limpezas dos dados, a fim de remover ruídos e tratar possíveis inconsistências. Em seguida, levanta-se as necessidades do modelo e a disposição dos dados armazenados, os seus tipos e os mapeamentos de valores ausentes e/ou desconhecidos;

Transformação e Redução dos Dados – Como normalmente, os algoritmos de Mineração de Dados (MD) não podem acessar os dados em seu formato nativo, por causa da forma de armazenamento ou normalização adotada na modelagem da base de dados, torna-se necessária a conversão desses dados para um formato mais apropriado, podendo ainda sumarizá-los, a fim de reduzir o número de variáveis consideradas ou criar novos atributos que possam agregar valor à base de dados;

Escolha das Tarefas de Mineração de Dados - Nesta etapa, decide-se qual o tipo de tarefa de MD proverá modelos que melhor se enquadrem no problema que se deseja resolver, podendo-se adotar, por exemplo: classificação, predição, segmentação e/ou análise de dependência;

Escolha dos Algoritmos de Mineração de Dados - Após decidir-se sobre as Tarefas, selecionam-se os algoritmos e parâmetros que propiciem resultados em conformidade com os requisitos apurados no levantamento do problema. Esta escolha dependerá do objetivo da MD, que poderá ser a criação de modelos com alto grau de predição ou apenas a obtenção de um melhor entendimento da base de dados;

Mineração de Dados - Consiste na efetiva aplicação dos algoritmos escolhidos sobre os dados analisados, com o objetivo de localizar os padrões desejados. A qualidade dos resultados desta etapa depende diretamente da correta execução das etapas anteriores; e

Interpretação e Avaliação - Nesta última etapa, analisa-se os resultados gerados pelos algoritmos, verificando-se a necessidade de se retornar aos passos anteriores, para o refinamento do processo, removendo-se os padrões irrelevantes e as redundâncias. Por fim, realiza-se a transformação dos resultados em linguagem acessível ao usuário final.

Em virtude do destaque dado à etapa de Mineração de Dados, na indústria e entre os pesquisadores de conhecimento em Banco de Dados, o termo Mineração de Dados tem sido utilizado para identificar todo o processo de DCBD, popularizando-se como um sinônimo. No entanto, as demais etapas do DCBD possuem tanta ou mais importância que a etapa de Mineração de Dados (Han e Kamber, 2006; Berry e Linoff, 2004).

Assim, por ser uma denominação mais difundida atualmente no meio acadêmico, este artigo utilizará a expressão Mineração de Dados, ao invés de Descoberta de Conhecimento em Banco de Dados (DCDB) ou Knowledge Discovery in Databases (KDD), para denominar o processo como um todo.

2.2. O que é Mineração de Dados?

O termo Mineração de Dados (MD) refere-se ao processo de análise de dados com a finalidade de verificar hipóteses ou de descobrir padrões interessantes, que possam representar informações úteis para a tomada de decisão (Imberman, 2001). Pode-se definir um padrão como uma afirmação S em L, que descreve um subconjunto de fatos F_S dentre um conjunto de

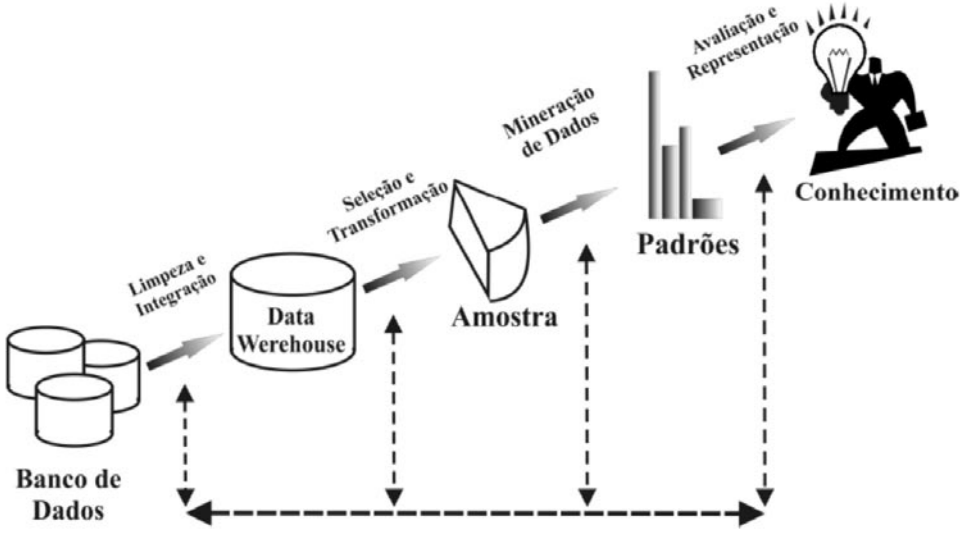

Figura 1 – Processo de DCBD.

fatos F, dado por uma distribuição probabilística, e a afirmação S pode ser expressa, de forma mais simples do que todos os fatos de F_S (Jain e Duin, 2004).

Métodos tradicionais de análise de dados, utilizando planilhas e consultas, tornaram-se inapropriados para tratar grandes volumes de dados. Ao adotar este tipo de abordagem, pode-se até oferecer suporte para a criação de relatórios informativos sobre os dados, mas não se consegue propiciar análises de seu conteúdo, nem tampouco descobertas de conhecimentos importantes para tomada de decisões (Fayyad *et al.,* 1996).

2.3. Tarefas e Técnicas de Mineração de Dados

Torna-se importante identificar a diferença entre as Tarefas e as Técnicas de Mineração de Dados (MD). As Tarefas consistem nas especificações do que se procura nos dados, em que tipo de regularidades ou categoria de padrões tem-se interesse, ou quais tipos de padrões poderiam surpreender. Já as Técnicas de Mineração consistem nas formas pelas quais se executam os métodos para garantir a descoberta de padrões importantes.

Dentre as principais Técnicas utilizadas em MD, tem-se a estatística e a aprendizagem de máquina. Jackson (2002) apresenta, em seu artigo, a combinação entre as Tarefas e as Técnicas ilustradas na Tabela 1. Tem-se nesta tabela: as Tarefas de classificação, predição, análise de dependência e segmentação; e as Técnicas de descrição e visualização, regressão, árvore de decisão, análises de correlação, redes neurais, *clustering* e regras de associação.

As tarefas de MD dividem-se em tarefas com Aprendizagem Supervisionada e tarefas com Aprendizagem Não-Supervisionada, definidas a seguir (Harrison, 1998) (Silva, 2004):

Tarefas com Aprendizagem Supervisionada - As tarefas deste gênero possuem como característica básica a predição, onde se determina, na base de treinamento, quais conjuntos de dados pertencentes a cada uma das classes. As tarefas propiciam a inferência de padrões, com o intuito de prever ou indicar tendências de informações ausentes ou desconhecidas, a partir da base de treinamento, como por exemplo, a determinação da classe a qual pertenceria um conjunto de dados específicos e ainda não classificados. Fazem parte deste gênero, as seguintes Tarefas (JACKSON, 2002):

- **Classificação** – Propicia a realização do mapeamento dos dados para o atributo-meta ou classe correspondente (categórica, por exemplo, "Classe A" ou "Classe B"), conforme aprendido com a base de treinamento, e
- **Predição** – Possui a mesma funcionalidade da classificação, diferenciando-se pelos tipos das classes, que ao invés de categóricas, são numéricas.

Tarefas com Aprendizagem Não-Supervisionada - Neste tipo de aprendizagem, não se fornece classes ou pré-determinações de como se formarão os modelos gerados pelas tarefas; no máximo, define-se a quantidade de grupos nos quais os dados deverão ser divididos. Na Aprendizagem Não-Supervisionada, os modelos resultantes apresentam uma descrição concisa dos dados da base de treinamento, fornecendo características e propriedades dos dados minerados. Dentre as tarefas que adotam a Aprendizagem Não-Supervisionada tem-se:

- **Segmentação** – Define-se como uma tarefa que propicia a realização do agrupamento natural dos dados. Diferentemente da classificação, onde se pré-determinam as classes, a segmentação encontra as classes baseando-se nas maximizações da similaridade dos itens intra-classe e nas diferenças entre as classes, e
- **Análise de Dependência** – Propicia a descrição da dependência entre os atributos, podendo ocorrer em dois níveis: estrutural, onde se determina quais variáveis são dependentes localmente (expressas graficamente); e quantitativa, que expressa, numericamente, qual o grau de dependência existe entre os atributos.

Dentre as Técnicas de Mineração de Dados, destacam-se (Han e Kamber, 2006):

Tabela 1 - Tarefas e Técnicas de Análise de Dados. Adaptado de Jackson (2002).

TAREFAS / TÉCNICAS	Supervisionada		Não-Supervisionada	
	Classificação	Predição	Dependência	Segmentação
Descrição e Visualização			X	X
Regressão		X	X	
Árvore de Decisão	X	X		
Análise de Correlação			X	
Redes Neurais	X	X		X
Clustering				X
Regras de Associação			X	

Árvore de Decisão - Apresenta-se como um fluxograma, disposto sob forma de uma árvore, conforme ilustrado na Figura 2a, no qual cada nó contém o nome do atributo que terá o seu valor testado. Nos galhos, encontram-se as saídas dos testes; e nas folhas, as classes resultantes da classificação;

Árvore de Regressão – Possui uma estrutura semelhante à árvore de decisão. No entanto, ao invés de valores categóricos, nas folhas da árvore, apresentam-se valores numéricos; e

Clustering – Esta técnica procura agrupar os dados de tal forma que maximize a similaridade dos objetos de um mesmo grupo e a diferença entre grupos distintos, conforme ilustrado na Figura 2b. Pode-se utilizar a Técnica de *clustering* para detectar os *outliers*. Os *outliers* representam objetos não pertencentes a nenhum dos grupos encontrados pelo algoritmo de *clustering* e devem ser analisados caso a caso, pois podem representar algum tipo de fraude, um novo nicho de mercado ou erro na inserção de dados.

3. MATERIAIS E MÉTODOS

3.1. VEGALOGIC

Visando utilizar o processo de MD para a geração de modelos que propiciassem a verificação dos dados armazenados em um sistema real, firmou-se uma parceria com o Instituto de Controle do Espaço Aéreo Brasileiro (ICEA). A partir desta parceria, iniciou-se a concepção e o desenvolvimento de um Método de Verificação de Dados Climatológicos (VEDALOGIC), apoiado em modelos gerados por algoritmos de MD, a partir de uma série história armazenada em uma Base de Dados Climatológicos (BDC).

Para gerar os modelos de *clustering* e de árvore de decisão, utilizados pelo método de verificação dos dados, adotou-se o conjunto de ferramentas *WEKA* (*Waikato Environment for Knowledge Analysis*) (Han e Kamber, 2006), ferramenta livre e de código aberto. O WEKA é escrito em Java, portanto, portável para a maioria dos sistemas operacionais, bastando que o usuário possua a Maquina Virtual Java (JVM) instalada.

Dentre os vários algoritmos disponíveis no *WEKA*, os que apresentaram modelos mais precisos, de acordo com os testes preliminares, foram:

Para a geração de modelos de *clustering*:
- *Expectation-Maximization* (Dempster et al 1977), e
- *K-means* (Macqueen, 1967); e

Para a geração de modelos de árvore de decisão:
- *REPTree* (Quinlan, 1986), e
- M5P (Wang e Witten, 1997).

3.1.1. Geração dos Modelos de Verificação de Dados

Inicialmente, criaram-se os modelos de *clustering* e de árvore de decisão. Os modelos de *clustering* foram utilizados para a detecção de *outliers*. Já os modelos de árvores de decisão foram utilizados para predição de valores de correções para dados suspeitos e possíveis ruídos. Para esta etapa, tem-se ilustrado na Figura 3, o fluxo de criação destes modelos, onde se adota um único modelo de *clustering*, que servirá para identificar dados suspeitos, e vários modelos de árvore de decisão, uma para cada combinação de atributos com dados suspeitos.

Para evitar uma explosão combinatorial e para que o VEDALOGIC gerasse sugestões com um grau de confiabilidade aceitável, realizaram-se estudos preliminares, considerando-se os dados do BDC, que indicaram um número limite de até três atributos contendo dados suspeitos. Para cada modelo de árvore de decisão, até o limite de três atributos suspeitos, um será o atributo predito e os demais atributos serão retirados para a

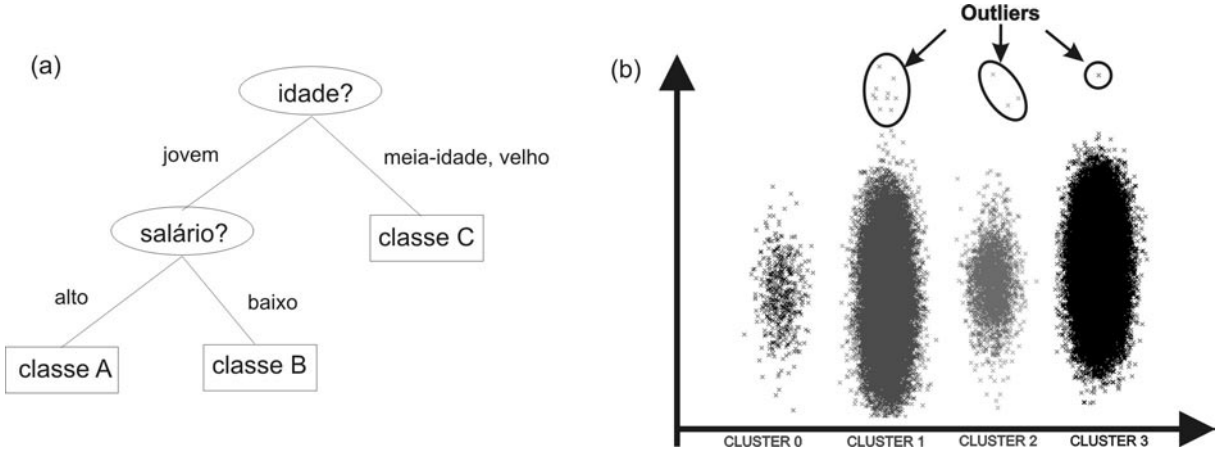

Figura 2 - Exemplos Técnicas de Mineração de Dados: a) árvore de decisão; e b) *clustering*

Figura 3 - Sequência de criação dos modelos utilizados pelo VEDALOGIC

geração dos modelos de árvore de decisão.

Após a geração dos modelos de detecção de *outliers* e de predição de valores, armazenam-se os modelos e as respectivas avaliações de qualidade de cada um dos modelos gerados. Essas avaliações servirão para a determinação de quais modelos, dentre os modelos gerados, possuem maior precisão para a detecção de *outliers*, bem como para a sugestão de valores de correção.

O Método VEDALOGIC propicia a detecção e correção de valores ruidosos, executando os seguintes passos descritos a seguir, conforme ilustrado na Figura 4:

Primeiramente, o usuário entrará com os dados coletados junto a uma Plataforma de Coleta de Dados (PCD) e anotados em uma planilha. Em seguida, ao solicitar a inserção desses dados na base de dados, esses passarão pelo módulo de verificação;

Na verificação dos dados, define-se o respectivo perfil, enquadrando-se o conjunto de dados (tupla) em um determinado grupo (*cluster*), o qual fornecerá um valor médio e um desvio padrão para cada um dos atributos, utilizados para determinar se o dado é ou não um dado suspeito;

Após a definição do *cluster* ao qual a tupla pertence, verifica-se o intervalo de valores aceitos para cada um dos atributos. Caso o valor de um determinado atributo se encontre fora do intervalo estabelecido, um *outlier*, esta tupla seguirá para um modelo de árvore de decisão (predição);

A partir do modelo de árvore de decisão, sugere-se um valor alternativo para o atributo suspeito, de acordo com a série histórica armazenada na base de dados;

Compara-se, então, o valor alternativo com o valor suspeito, caso estes sejam diferentes, considera-se os dados suspeitos como ruídos;

Uma vez detectado o ruído, o usuário receberá uma mensagem informando os atributos que contenham possíveis erros, assim como os respectivos valores de sugestão para a correção e o intervalo de valores aceitáveis para estes atributos; e

Caberá ao usuário adotar ou não as sugestões emitidas pelo VEDALOGIC. Caso ele opte por realizar as alterações sugeridas, os dados passarão por uma nova verificação. Em caso contrário, inserem-se os dados na base dados.

Com a adoção do VEDALOGIC, busca-se diminuir a quantidade de ruídos na Base de Dados, adotando-se como referência a série histórica armazenada e o conhecimento do responsável pela inserção dos dados.

3.2. Estudo de Caso

Para a verificação da eficiência do método proposto, utilizou-se a Base de Dados Climatológicos (BDC) do Instituto

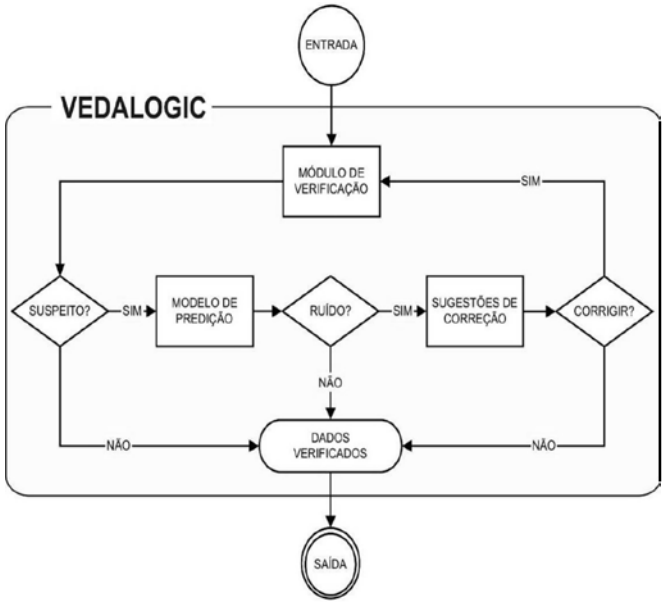

Figura 4 - Fluxograma do método de verificação VEDALOGIC

de Controle do Espaço Aéreo (ICEA), subordinado ao Comando da Aeronáutica. Adotou-se, especificamente, para este estudo de caso, os dados coletados entre os anos de 1974 e 1990, na cidade de Bagé, no estado do Rio Grande do Sul – Brasil. Adotaram-se esses dados, por representarem um conjunto de dados com a maior série histórica completamente digitalizada no BDC.

Os dados utilizados, neste estudo de caso, foram coletados em uma Estação Meteorológica de Superfície (EMS), dentre mais de 130 espalhadas nos aeródromos de todo o território brasileiro, e anotados em uma planilha denominada Impresso Especial de Proteção ao Vôo – IEPV, modelo 105-25. Os dados do IEPV 105-25 encontram-se armazenados na tabela TF10525 do BDC. Os atributos, as respectivas descrições, as distribuições dos dados (médias e desvios padrão ou modas), os intervalos com o piso e teto para cada atributo da série histórica e o número de tuplas com dados nulos, contidos na tabela TF10525 do SBD, encontram-se na Tabela 2.

A verificação dos dados climatológicos se faz necessária por conta do método adotado para a inserção dos dados no BDC, inserção manual, na qual o observador meteorológico faz a leitura dos dados em uma Plataforma de Coleta de Dados – PCD e os anota em uma planilha. Em um segundo momento, realiza-se a inserção desses dados em um sistema.

Para a digitalização dos dados climatológicos, adotam-se os sistemas Climat I e Climat II. Estes sistemas realizam verificações por intervalo de valores, em cada um dos atributos,

Tabela 2 – Atributos da tabela TF10525, com a descrição, a média e o desvio padrão ou moda (estatística), o intervalo de valores [piso; teto] da série histórica e o número de tuplas com dados ausentes.

ATRIBUTO	DESCRIÇÃO	ESTATÍSTICA	INTERVALO	NULOS
NSINOTICO	Código da Localização da coleta	avg = 83981 +/- 0	[83.981; 83.981]	1,0
DIA	Dia dos mês	avg =15,690 +/- 8,776	[1,000 ; 31,000]	1,0
MES	Mês dos Ano	avg = 6,517 +/- 3,446	[1,000 ;12,000]	1,0
ANO	Ano da coleta do dado	avg =1.981,61 +/- 4,87	[1.974,000 ;1.990,000]	1,0
HORA	Hora do dia	avg =12,952 +/- 4,393	[0,000 ;23,000]	1,0
TOTNUVEM	Total de Nuvens	avg = 4,500 +/- 3,049	[0,000 ; 9,000]	2,0
DIRVENTO	Direção do Vento	avg =13.205+/-10,587	[0,000 ; 36,000]	6,0
VELVENTO	Velocidade do Vento	avg = 6,493 +/- 4,476	[0,000 ; 50,000]	3,0
VISIB	Visibilidade	avg=1.687,91+/-523,29	[0,000 ; 8.000,000]	1,0
CGT	Condição Geral do Tempo	avg =1,475 +/-2,226	[0,000 ; 9,000]	7,0
QTDNUVEM1	Quantidade de nuvens na1º camada	avg = 3,962 +/-2,329	[1,000 ; 9,000]	16470,0
TIPONUVEM1	Tipo das nuvens na1º camada	avg = 5,538 +/-2,539	[0,000 ; 9,000]	16697,0
DIRNUVEM1	Direção do deslocamento na1º camada	avg = 4,534 +/-2,055	[0,000 ; 9,000]	16730,0
ALTNUVEM1	Altura das nuvens na1º camada	avg =180,45+/-211,59	[0,000 ; 900,000]	17131,0
QTDNUVEM2	Quantidade de nuvens na2º camada	avg = 4,918 +/-2,389	[0,000 ; 9,000]	55749,0
TIPONUVEM2	Tipo das nuvens na2º camada	avg = 3,391 +/-2,322	[0,000 ; 9,000]	55749,0
DIRNUVEM2	Direção do deslocamento na2º camada	avg = 5,410 +/-1,582	[0,000 ; 9,000]	55888,0
ALTNUVEM2	Altura das nuvens na2º camada	avg = 345,35+/-228,14	[0,000 ; 900,000]	55749,0
QTDNUVEM3	Quantidade de nuvens na 3º camada	avg = 5,723 +/-2,370	[1,000 ; 8,000]	82004,0
TIPONUVEM3	Tipo das nuvens na 3º camada	avg =2,836 +/-2,100	[0,000 ; 9,000]	82004,0
DIRNUVEM3	Direção do deslocamento na 3º camada	avg = 5,825 +/-1,267	[0,000 ; 9,000]	82053,0
ALTNUVEM3	Altura das nuvens na 3º camada	avg = 421,34+/-200,85	[0,000 ; 900,000]	82004,0
QTDNUVEM4	Quantidade de nuvens na 4º camada	avg = 6,385 +/-2,149	[1,000 ; 8,000]	91217,0
TIPONUVEM4	Tipo das nuvens na 4º camada	avg =2,641 +/-1,520	[1,000 ; 9,000]	91217,0
DIRNUVEM4	Direção do deslocamento na 4º camada	avg = 6,044 +/-1,126	[1,000 ; 8,000]	91232,0
ALTNUVEM4	Altura das nuvens na 4º camada	avg = 437,77+/-190,25	[30,000 ; 900,000]	91216,0
QNH	Ajuste do Altímetro	avg =1.014,12+/-8,62	[930,000 ;1.054,800]	4,0
PO	Ponto de Orvalho	avg =13,367 +/- 5,295	[-6,000 ; 99,000]	42,0
RAJADA	Velocidade da Rajada de Vendo	avg =24,832 +/- 9,747	[0,000 ; 85,000]	92163,0
QFF	Pressão ao nível médio do mar	avg =1.013,59+/-9,14	[930,000 ;1.059,900]	16,0
QFE	Pressão da estação ao nível da Pista	avg = 990,61+/-15,17	[900,000 ;1.024,000]	152,0
TENDPRESSAO	Tendência da Pressão atmosférica	avg = 4,371 +/-2,613	[0,000 ; 9,000]	58593,0
DIFPRESSAO	Diferença de Pressão	avg =1,074 +/- 0,853	[0,000 ; 9,900]	58592,0
BSECO	Temperatura do Ar	avg =19,448 +/- 6,929	[-3,400 ; 55,600]	26,0
BUMIDO	Temperatura de Orvalho	avg =15,670 +/- 5,117	[-3,400 ; 50,000]	69,0
PRECIP	Precipitação em milímetros (mm)	avg = 0,284 +/- 8,027	[0,000 ; 900,000]	1176,0
UR	Umidade Relativa do Ar	avg = 72,271+/-20,03	[1,000 ;100,000]	58430,0
FASECONSIST	Controle da consistência no BDC	mode = c (82542)	c(82542),T(3308),a(6640)	1,0
ORIGEMDADO	Origem dos dados	mode = A (82542)	A(82542),T(3308),I(6640)	1,0
FLAGIAE	Flag de Controle	mode = E (92490)	E (92490)	1,0
FLAGEXP	Flag de Controle	mode = E (89182)	E (89182)	3309,0

realizando também algumas poucas comparações entre os dados de atributos diferentes, ou seja, algumas regras inferidas por um especialista. Após a digitação dos formulários nos sistemas Climat I e Climat II, os dados passam por uma verificação de fidelidade por amostragem, a partir do CLIVER. A verificação, por fidelidade, realizada pelo CLIVER consiste em sortear uma quantidade de 300 tuplas, de um total de 4800, coletadas no período de um ano, para que o técnico responsável pela verificação compare os dados digitalizados com os dados anotados na planilha. Caso ocorram erros em menos de 5% das tuplas, insere-se os dados no BDC. Caso contrário, faz-se a redigitação dos dados referente ao ano reprovado. Esta tarefa consome, aproximadamente, 240 h/homem.

Apesar da realização da verificação por intervalo de valores e por fidelidade, o processo de inserção de dados adotado pelo ICEA possui suscetibilidade a erros, pois não há uma verificação por consistência, considerando-se a inter-relação de todos os atributos de uma mesma tupla. Esta suscetibilidade a erros poderá impactar na confiabilidade dos dados contidos no BDC e, conseqüentemente, nas pesquisas e decisões tomadas com base nesses dados.

3.2.1. Pré-Processamento

Para a geração de modelos de MD confiáveis, deve-se seguir todas as etapas do processo de DCBD, uma das etapas de grande importância para a geração de modelos confiáveis é o pré-processamento. Os dados se encontram em um BDC e na tabela TF10525, totalizando 92.491 tuplas e 42 atributos. Dentre os 42 atributos, foram excluídos 9 com baixa representatividade para a geração dos modelos de árvore de decisão e de *clustering*. Os atributos excluídos foram: *flags* de controle interno; chave primária; e demais atributos com altos índices de dados ausentes.

Os atributos referentes à *flags* de controle interno e chave primária ou estrangeira são descartados, nos primeiros passos da etapa de pré-processamento. Após esta exclusão de atributos, pode-se aplicar, combinadamente, avaliadores com os métodos de busca, para auxiliarem na escolha dos atributos para aplicação do processo de Mineração de Dados. Para maiores informações sobre avaliadores e métodos de busca, consulte Witten & Frank (2005). Ao final desta etapa de pré-processamento, restaram apenas os 33 atributos, hachurados na Tabela 2, e 80.989 tuplas. Estes atributos serão considerados para a geração dos modelos de verificação e sugestão de valores.

3.2.2. Geração dos Modelos

Para a geração dos modelos de verificação e de sugestão de valores de correção do VEDALOGIC, utilizaram-se os dados da série histórica entre os anos de 1974 e 1989, reservando-se o ano de 1990 para a simulação de um novo ano de leituras. Este conjunto de dados, referente ao ano de 1990, será utilizado para a verificação da eficiência do VEDALOGIC ao detectar e sugerir valores alternativos para dados ruidosos.

Para a criação dos modelos de verificação (*clustering*) utilizou-se os algoritmos *Expectation-Maximization* (EM) e o K-means (implementado no WEKA com o nome *SimpleKmeans*). A aplicação do algoritmo EM encontra o número adequado de *cluster* (grupos) a serem gerados pelo modelo. Neste caso, foram

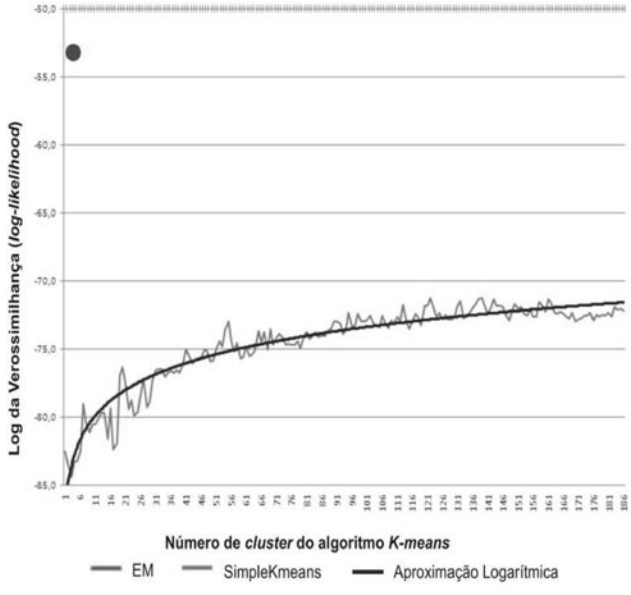

Figura 5 – Gráfico comparativo da avaliação dos modelos gerados pelos algoritmmos EM e K-means (SimpleKmeans).

encontrados apenas dois *clusters*. A aplicação do algoritmo K-means necessita da determinação de qual número de *clusters* serão gerados pelo algoritmo. Por este motivo, gerou-se um conjunto de modelos variando-se o número de *clusters* de 2 até 190. Para a avaliação da qualidade dos modelos gerados, utilizou-se o log da verossimilhança (*log-likelihood*). Para mais detalhes sobre o cálculo desta função, consulte (Dempster et al. 1977).

Na Figura 5 tem-se o gráfico comparativo da qualidade dos modelos obtidos com a utilização dos algoritmos EM e K-means. Neste gráfico, observa-se que o resultado da avaliação do modelo gerado pelo algoritmo EM, encontra-se representado por um ponto no canto superior direito, atingindo o valor de -53.10. Já o algoritmo K-means, encontra-se representado por uma linha que indica a evolução da precisão, a medida que se aumenta o número de *clusters*, acompanhando uma função logarítmica (Aproximação logarítmica) e tendendo a uma estabilidade.

O algoritmo K-means, mesmo com um número elevado de *cluster*, propiciou a geração de modelos com qualidade inferior ao modelo gerado pelo EM. Além disso, com o aumento demasiado do número de *cluster*, surge o problema de super-especialização (*overfiting*) do modelo gerado, diminuindo a sua capacidade de generalização. Com os resultados obtidos, adotou-se o modelo gerado pelo algoritmo EM para o módulo de verificação do VEDALOGIC.

Para a geração dos modelos de sugestão de valores, utilizou-se os algoritmos M5P e REPtree, ambos propiciaram a criação de modelos que atingiram precisões semelhantes, com uma média de 99%. Com isso, decidiu-se armazenar os modelos gerados para a utilização no módulo de predição, adotando-se, para a sugestão de valores, o modelo com maior precisão, conforme a combinação de atributos ruidosos.

Na Figura 6, tem-se um gráfico de pontos, contendo a distribuição dos dados do BDC com o atributo BSECO, no eixo *y,* e o atributo MÊS, no eixo *x*. Em (a), tem-se as linhas que delimitam o piso e o teto da verificação realizada pelo Climat I e Climat II, por intervalo de valores. Em (b), tem-se a delimitação da área sem cobertura da verificação realizada pelo Climat I e II. No entanto, utilizando-se o VEDALOGIC, torna-se possível identificar os valores ruidosos que estiverem compreendidos na área assinalada, assim como os demais que extrapolarem as linhas indicadas em (a).

3.3. Aplicação do VEDALOGIC nos Dados do BDC

Após a geração dos modelos de detecção e verificação de ruídos do VEDALOGIC, criaram-se três cenários para a verificação da eficiência do método, ao detectar valores suspeitos e sugerir valores alternativos para a correção desses dados. Para esta verificação, alteraram-se os dados previamente

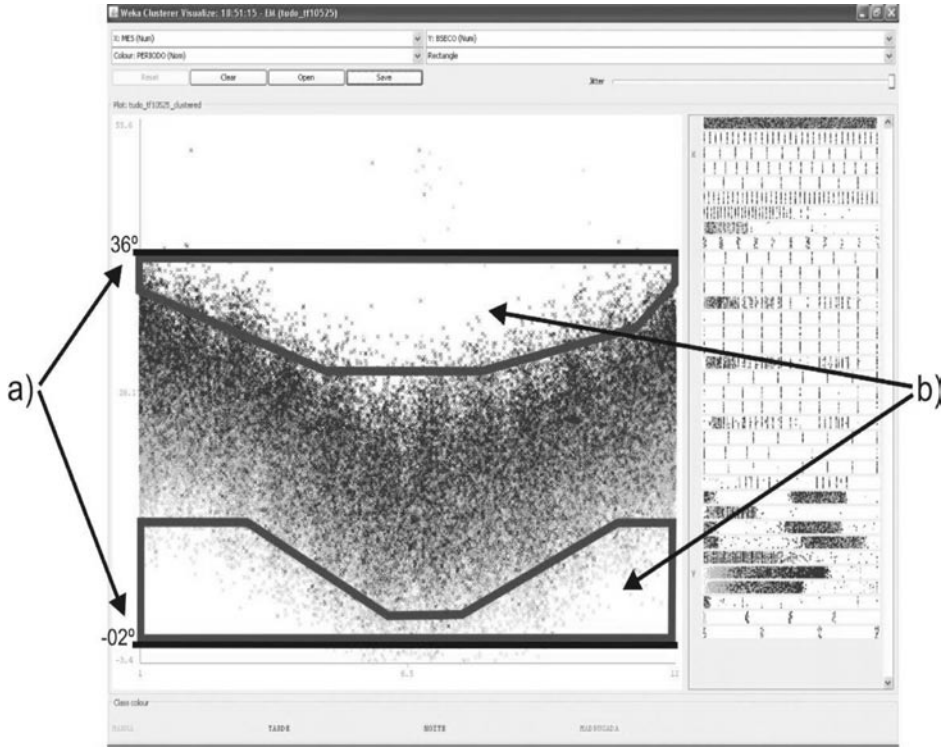

Figura 6 – Distribuição dos dados do BDC com BSECO, no eixo y, e MÊS, no eixo x: a) Linhas de piso e teto da verificação por intervalo de valores; b) Área sem cobertura da verificação por intervalo de valores e bordada pela verificação do VEDALOGIC.

selecionados, referente ao ano de 1990, simulando-se um novo ano de leituras. Os dados referentes ao ano de 1990, não foram considerados para geração dos modelos. Os valores sugeridos pelo VEDALOGIC serão comparados com os valores calculados, a partir dos princípios da termodinâmica. Para maiores detalhes sobre as fórmulas utilizadas, consulte (SWIDT, 1995).

No primeiro cenário, inseriram-se ruídos no BSECO, alterando-se os valores das temperaturas de 23,00ºC para 35,00ºC, em 73 ocorrências. Neste cenário, o VEDALOGIC identificou todos os ruídos inseridos, bem como sugeriu valores com erro médio de 0,47ºC, ou seja, 99,21% de precisão média, ilustrados na Figura 7 – BSECO. Os valores calculados, com base nos princípios da termodinâmica, atingiram erro médio de 0,53ºC, ou seja, 99,11% de precisão média.

No segundo cenário, substituiu-se o valor do atributo PO de 23,0ºC por 28,0ºC, em 76 tuplas. O VEDALOGIC identificou todos os 76 ruídos inseridos, e sugeriu valores para correção com erro médio de 0,64ºC (98,23% de precisão), contra um erro médio de 0,30ºC (99,17% de precisão) dos valores calculados. Os resultados encontram-se ilustrados na Figura 7 – PO.

No terceiro e último cenário, inseriu-se ruídos no atributo BUMIDO, no qual se alterou o valor de 22,0ºC para 30,0 ºC, em 79 tuplas. Todos os ruídos foram identificados pelo VEDALOGIC, que sugeriu valores de correções com erro médio de 0,24ºC, atingindo uma precisão média de 99,55%. Já os valores calculados obtiveram erro médio de 0,17ºC e precisão de 99,68%. Os resultados encontram-se ilustrados na Figura 7 - BUMIDO.

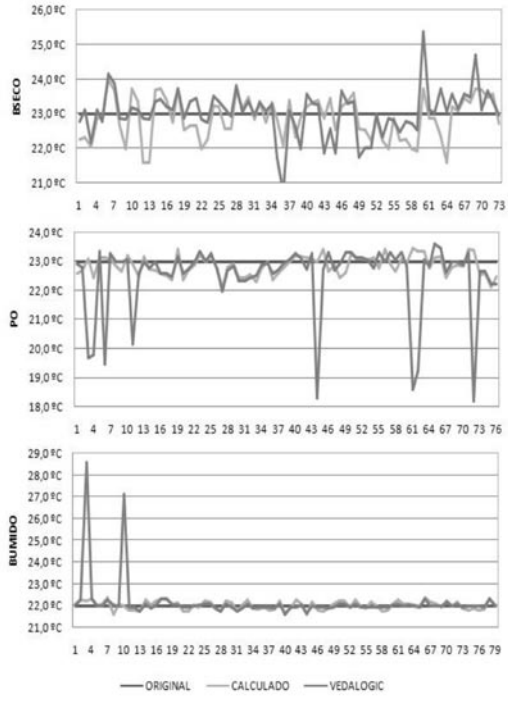

Figura 7 – Gráfico comparativo, em ºC, entre os valores originais, os calculados e os sugeridos pelo VEDALOGIC: BSECO - para o atributo BSECO; PO - para o atributo PO; e BUMIDO - para o atributo BUMIDO.

Os valores ruidosos inseridos encontram-se na área de verificação sem cobertura pelo método de verificação por intervalo de valores do Climat I e II. Logo, esses dados ruidosos provavelmente seriam inseridos no BDC, impactando na confiabilidade dos seus dados.

4. DISCUSSÃO DOS RESULTADOS

Para a análise dos principais resultados, adotou-se como base para a comparação, a precisão dos valores sugeridos pelo VEDALOGIC e os valores calculados com base nos princípios da termodinâmica, em relação ao valor original dos dados (valores antes da inserção dos ruídos). A precisão do valor sugerido e do valor calculado representa a grandeza de proximidade destes com o seu original. Para a medida de precisão em percentual, consideraram-se os valores de piso e teto da série histórica do atributo analisado.

Ao analisar os principais resultados obtidos nos três cenários, observou-se que os valores sugeridos pelo VEDALOGIC, em mais da metade dos casos, foram mais precisos em relação aos valores calculados pela termodinâmica, conforme o gráfico da Figura 8a. Como se pode observar, para o atributo BSECO, em 61,64% dos casos, o VEDALOGIC encontrou valores mais precisos do que os valores calculados pela termodinâmica, em relação aos dados originais. Para o atributo PO, o VEDALOGIC obteve precisão superior aos valores calculados, em 51,32% dos casos. Já para o atributo BUMIDO, o VEDALOGIC apresentou maior precisão média dentre os cenário considerados, obtendo, em 83,54% dos casos, um resultado mais preciso do que os valores encontrados com base nos cálculos da termodinâmica.

Na Figura 8b, tem-se o erro médio, em graus Celsius (ºC), atingido pelos valores sugeridos pelo VEDALOGIC e pelos valores calculados com base na termodinâmica. Observa-se que o erro médio das sugestões de valores de correção fornecidas pelo VEDALOGIC, para o atributo BSECO, ficou em 0,47ºC, enquanto o valor calculado obteve um erro médio de 0,53ºC, portanto, superior ao erro médio alcançado pelo VEDALOGIC. Para o atributo PO, o VEDALOGIC atingiu um erro de 0,64ºC e os valores calculados com 0,30ºC de erro médio. A sugestão de valores para o atributo BUMIDO atingiu o menor erro dentre os cenários estudados, com erro médio de apenas 0,24ºC para as sugestões do VEDALOGIC e 0,17ºC de erro médio para os valores calculados.

Os resultados alcançados pelo VEDALOGIC indicam o potencial de sua utilização, uma vez que os algoritmos de

Mineração de Dados propiciaram a criação de modelos que chegaram bem próximos aos valores calculados com base na termodinâmica, que já foram desenvolvidos, aprimorados e consolidados ao longo do tempo. Com base na série histórica, os algoritmos propiciaram a criação de modelos, que se aproximam da realidade dos dados climatológicos e que, na maioria dos casos, propiciaram a sugestão de valores mais precisos dos que os calculados pela termodinâmica, bem como a identificação dos dados suspeitos de conterem ruídos.

5. CONCLUSÃO

Este artigo teve por objetivo apresentar um método para a verificação de dados e informações existentes na Base de Dados Climatológicos (BDC) do Instituto de Controle do Espaço Aéreo (ICEA), visando reduzir ruídos e erros, aumentar a qualidade e a confiabilidade dos dados e melhorar a eficiência das pesquisas realizadas com base nos dados do ICEA. O método desenvolvido foi denominado VEDALOGIC.

O desenvolvimento do VEDALOGIC foi iniciado, realizando-se, um estudo bibliográfico, abordando os principais conceitos sobre o processo de Descoberta de Conhecimento em Banco de Dados (DCBD) e de Mineração de Dados (MD) e seus principais algoritmos.

Para a MD, estudaram-se algoritmos que propiciassem a realização apropriada das tarefas de segmentação e predição, utilizando-se, especificamente, algoritmos de *clustering* e de árvore de decisão, de tal forma a fundamentar e situar o desenvolvimento desta pesquisa e fornecer subsídios necessários ao seu entendimento.

Apesar de não abordado neste artigo, para o desenvolvimento desta pesquisa, estudou-se também as principais ferramentas de MD existentes e de código aberto, que propiciassem a reutilização de código-fonte. Após a avaliação dos ambientes de ferramentas existentes, optou-se pelo ambiente WEKA, por possuir código aberto, ser desenvolvido em linguagem Java e, conseqüentemente, multiplataforma.

Em seguida, adotou-se os algoritmos *Expectation-Maximization* (EM) e K-means para a produção de modelos de *clustering* e os algoritmos REPTree e M5P para a geração dos modelos de árvores de decisão, por apresentarem modelos mais precisos dentre os algoritmos estudados.

O método de verificação VEDALOGIC baseou-se em modelos minerados do BDC para a criação dos modelos de verificação e de sugestão de valores para a correção de valores ruidosos.

A abordagem adotada pelo VEDALOGIC não esgota o assunto referente à verificação de dados em toda a sua complexidade, mas por ter sido fundamentada em uma pesquisa, fornece uma alternativa ao processo de verificação de dados existente no ICEA.

Para a verificação da eficiência do VEDALOGIC, implementou-se um protótipo, no qual foi utilizado um modelo de *clustering* para o módulo de verificação dos dados, com a finalidade de melhor identificar valores suspeitos. Utilizou-se também um conjunto de modelos de árvore de decisão, para propiciar a sugestão de valores alternativos para a correção de valores ruidosos.

Ao se inserir ruídos no conjunto de dados de verificação, mais especificamente nos atributos referentes às temperaturas do ponto de orvalho, de bulbo seco e de bulbo úmido, o

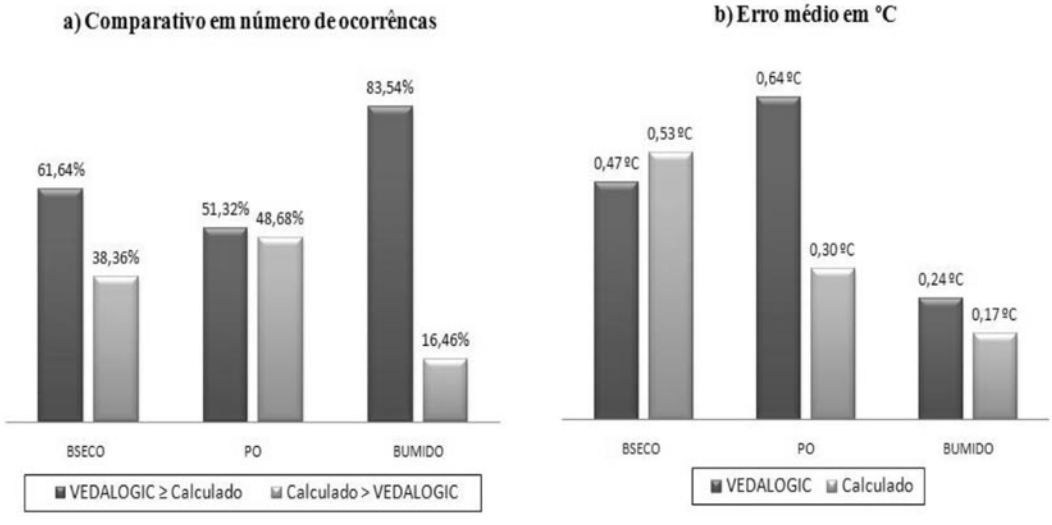

Figura 8 – a) Gráfico comparativo de precisão pelo número de ocorrências em que o VEDALOGIC sugeriu valores mais precisos dos que os valores calculados. b) Gráfico comparativo do erro médio, em graus Celsius (ºC), atingido pelos valores sugeridos pelo VEDALOGIC e os erros médios dos valores calculados.

VEDALOGIC propiciou a detecção de todos os ruídos inseridos. As sugestões de ajustes oferecidas obtiveram precisão média acima de 98%. Em mais da metade dos casos analisados, os valores sugeridos pelo VEDALOGIC ficaram tão ou mais próximos dos dados originais, do que os valores calculados pelos princípios da termodinâmica.

Assim concluiu-se, que a partir da inserção do VEDALOGIC no processo de verificação dos dados no ICEA, pode-se aumentar consideravelmente, a qualidade dos dados no BDC, fornecendo-se informações auxiliares para o ajuste de dados ruidosos. As informações complementares darão suporte a ajustes coerentes e precisos.

5.1. Recomendações

Recomenda-se que seja adotado o VEDALOGIC para minimizar os erros e inconsistências no BDC do ICEA, uma vez que os dados nele contidos são utilizados para pesquisas e tomadas de decisões nos aeroportos de todo o território brasileiro e até mesmo consultados por organismos internacionais.

Recomenda-se ainda que seja criado um conjunto de modelos para cada aeroporto, tendo em vista que, para cada região, existem diferentes particularidades nas oscilações climatológicas. Os padrões encontrados, por exemplo, na cidade de Bagé, no Rio Grande do Sul, localizada na região sul, podem não ser aplicáveis à cidade de Cuiabá, em Mato Grosso, localizada na região centro-oeste do Brasil.

6. AGRADECIMENTOS

Os autores deste artigo agradecem: ao ICEA, pela disposição e colaboração da sua equipe e pela disponibilização dos dados utilizados no estudo de caso; ao Pesquisador Ieso de Miranda, do Instituto de Aeronáutica e Espaço – IAE, pela sua colaboração no desenvolvimento deste trabalho de pesquisa; ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pelo apoio financeiro; e ao Instituto Tecnológico de Aeronáutica – ITA, pelo apoio de infra-estrutura para o desenvolvimento com sucesso desta pesquisa.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BERRY, Michael J. A., LINOFF, Gordon S. **Data Mining Techniques For Marketing, Sales, and Customer Relationship Management.** Indianápolis: Wiley, 2004.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. *Maximum likelihood from incomplete data via the em algorithm*. ***Journal of the Royal Statistical Society. Series B (Methodological), Blackwell Publishing for the Royal Statistical Society,*** v. 39, n.1, p.1-38,1977. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2984875>. Acesso em: 20 out 2008.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. ***Advances in Knowledge Discovery and Data Mining.*** [S.l.]: AAAI/MIT Press,1996.

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

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *The kdd process for extracting useful knowledge from valumes of data.* **COMMUNICATIONS OF THE ACM**, v. 39, p.27-34, NOV1996b.

HAN, J.; KAMBER, M. ***Data Mining: Concepts and Techniques***. San Francisco, LA: Morgan Kaufmann, 2006.

HAND, David J. Statistics *and Data Mining: Interesting Disciplines*. **Association for Computers and Machinery - SIGKDD Explorations**, v.1 – Issue 1, p.16-19, June, 1999.

HARRISON, Thomas H. **Intranet Data Warehouse: Ferramentas e Técnicas para Utilização do Data Warehouse na Intranet**. São Paulo: Berkeley Brasil, 1998.

IMBERMAN, S. P. **Effective use of the kdd process and data mining for computer performance professionals**. In: Int. CMG Conference. 2001. p. 611-620.

JACKSON, Joyce. **Data mining: a conceptual overview**. *Communications of the Association for Information Systems*. v. 8, p.267-296, Mar, 2002.

JAIN, A.; DUIN, R. ***Pattern recognition.*** In: GREGORY, R. (Ed.). The Oxford Companion to the Mind. Second edition. Oxford, UK: Oxford University Press, 2004. p. 698-703.

MACQUEEN, J. B. *Some methods of classication and analysis of multivariate observations.* In: **Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability.**1967. p.281-297.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. Morgan Kaufmann Publishers,1993.

QUINLAN, J. R. ***Induction of decision trees. Machine Learning***, Kluwer Academic Publishers, Hingham, MA, USA, v.1, n.1, p. 81-106, March1986. ISSN 0885-6125.

REZENDE, Solange Oliveira et al. **Sistemas inteligentes: fundamentos e aplicações**. São Paulo: Malone, 2003, p. 307-333.

SILVA, M. P. S. **Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka**. 2004.

SWIDT, J. **TDFTSTW - Cálculo da temperatura do ponto de orvalho em função da temperatura de bulbo seco e da temperatura de bulbo úmido**. IAE - Instituto de Aeronáutica e Espaço. 1995.

WANG, Y.; WITTEN, I**.** *Induction of model trees for predicting continuous classes*. In: **Proc European Conference on Machine Learning Poster Papers.** Prague, Czech Republic:1997. p.128-137.

WITTEN, Ian H.; FRANK, Eibe. **Data mining: practical machine learning tools**. 2. ed. San Francisco, CA: Morgan Kaufmann, 2005.