INTER FACES CIENTÍFICAS

EXATAS E TECNOLÓGICAS

ISSN IMPRESSO - 2359-4934

E- ISSN - 2359-4942

DOI - 10.17564/2359-4934.2016v2n2p57-70

MINERAÇÃO DE DADOS METEOROLÓGICOS EMPREGANDO DADOS DE TEMPERATURA: O CASO DE UMA CIDADE GAÚCHA

Morgana Magnus Wagner[1]

Vinicius Gadis Ribeiro[2]

Jorge Zabadal[3]

RESUMO

Neste trabalho implementa-se a mineração de dados meteorológicos da cidade de São Martinho da Serra/RS, com o intuito de analisar as variáveis, buscando encontrar um padrão ou mudança em intervalo de dez anos. Foram empregados dados de temperatura e radiação solar, obtidos pela Plataforma de Coleta de Dados (PCD) do Instituto Nacional de Pesquisas Espaciais (INPE). Os métodos utilizados na mineração de dados mostraram-se viáveis, já que é possível sugerir padrões meteorologicamente coerentes, encorajando a novas pesquisas.

PALAVRAS-CHAVE

Mineração de Dados. Meteorologia. Análise de Dados.

RESUMEN

En este trabajo se implementa la minería de datos meteorológicos de São Martinho da Serra / RS, con el fin de analizar las variables, tratando de encontrar un patrón o un cambio en un intervalo de diez años. Fueron empleados datos de temperatura y radiación solar, obtenidos por la Plataforma de Coleta de Dados (PCD) del Instituto Nacional de Pesquisas Espaciais (INPE). Los métodos utilizados en la minería de datos se han demostrado viables, ya que es posible sugerir normas meteorológicamente consistentes, fomentando nuevas investigaciones.

PALABRAS CLAVE

Minería de datos; Meteorología; Análisis de datos.

ABSTRACT

In this paper it is implemented the meteorological data mining of São Martinho da Serra / RS, in order to analyze the variables, trying to find a pattern or change in ten years. Were used temperature and solar radiation data, obtained by the Data Collection Platform (DCP) of the Instituto Nacional de Pesquisas Espaciais (INPE). The methods used in data mining proved to be viable, since it is possible to suggest meteorologically consistent standards, encouraging further research.

KEYWORDS

Data Mining. Meteorology. Data Analysis.

1 INTRODUÇÃO

A monitoração ou previsão de eventos climáticos são, sem dúvida, essenciais para diversas atividades humanas. Certas áreas de conhecimento, como por exemplo, na agricultura, ou em outros ramos de atividades como na indústria ou no transporte, há a necessidade de se ter previsões confiáveis para seus planejamentos.

Na agricultura a previsão climática é importante, por exemplo, para avaliar a aptidão de um cultivo, a necessidade de irrigação e a melhor época de semeadura, conhecendo-se o clima da região. Mas, analisar dados meteorológicos gerados por estações automáticas não é uma atividade tão simples a qual conseguimos realizar sem uso de ferramentas específicas. Com o uso de técnicas computacionais, conseguimos analisar as correlações, variações, modelos e/ou gerar relatórios dos dados obtidos. A mineração de dados é uma dessas técnicas, na qual auxilia o descobrimento de conhecimento em grandes bases de dados.

O objetivo deste trabalho é verificar a possibilidade de utilização de técnicas de mineração de dados para identificar padrões, tendências ou correlações nos dados meteorológicos. Como prova de conceito, foram obtidos dados de uma pequena localidade do estado do Rio Grande do Sul – não foi escolhida uma grande cidade, pela possibilidade de ocorrência do fenômeno de microclimas diversificados em um mesmo município. Os dados meteorológicos incluem dados de temperatura e radiação solar dos últimos dez anos da cidade de São Martinho da Serra. Os dados foram disponibilizados pelo INPE/CPTEC por meio do Sistema Nacional de Dados Ambientais (SINDA).

2 REFERENCIAL TEÓRICO

O processo de mineração de dados pode ser considerado como uma parte do *Knowledge Discovery in Databases* – Descoberta de Conhecimento em Banco de Dados (KDD), que foi definido por Fayyad (1996, p. 23) como sendo "o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados".

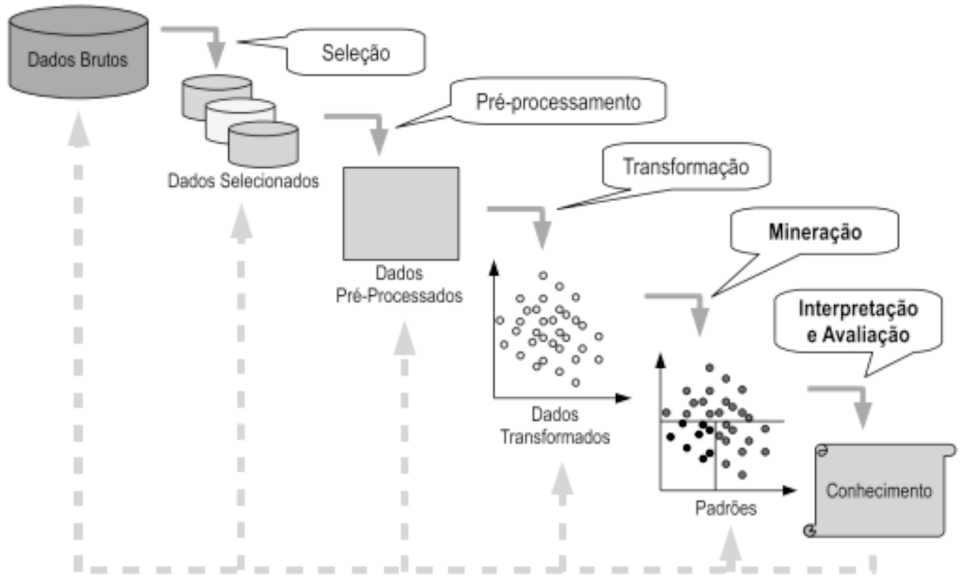
Ele é composto por várias etapas interligadas, que vão desde a definição de domínio, seleção, preparação e transformação dos dados até a etapa de Mineração de Dados, onde se pode analisar os padrões "descobertos" e utilizar técnicas usadas em Estatística e Banco de Dados para extração do conhecimento.

A Figura 1 apresenta – de modo genérico – o processo de Descoberta de Conhecimento em Banco de Dados (KDD) por meio de suas etapas interligadas.

A primeira etapa do processo é a definição e compreensão do domínio, após a definição, é necessário selecionar/criar os dados que serão trabalhados.

Na transformação, trabalhamos com os dados que contém representações ou informações inadequadas para o algoritmo a ser usado, excesso de atributos (redundantes ou desnecessários), atributos insuficientes, excesso de instâncias que podem afetar o tempo de processamento, instâncias insuficientes, instâncias incompletas (sem valores para alguns atributos). Esta etapa do Processo de Descoberta de Conhecimento (KDD) é relevante para se conseguir fazer a mineração de dados – pois, como exemplo, alguns algoritmos de associações só trabalham com valores simbólicos/discretos.

Figura 1 – Etapas do Processo de Descoberta de Conhecimento



Fonte: Fayyad (1996).

Depois dos dados limpos, pré-processados, reduzidos – com o intuito de minimizar ruídos – e transformados de acordo com a proposta, escolhemos a técnica de Mineração de Dados a ser utilizada, assim como seus parâmetros, pois dessa forma os dados serão transformados de acordo com o algoritmo aplicado.

Com o resultado em mãos, já podemos analisar, interpretar ou avaliar o conhecimento descoberto, podendo também repetir alguma etapa, se necessário.

2.1 MINERAÇÃO DE DADOS

Para Hand, Mannila e Smyth (2001, p. 68): "A Mineração de Dados é a análise dos conjuntos de dados observacionais, para encontrar relações insuspeitas e para resumir os dados de maneira compreensíveis e úteis para o proprietário dos mesmos".

Neste contexto, a Mineração de dados – ou Data Mining, é um processo de extração de informações de uma grande base de dados para tomada de decisões e, é aplicada em diversas áreas como empresas, pesquisas e indústrias que utilizam os resultados, como exemplo, para melhoria de processos ou analisar tendências. Este automatiza o processo de transformação e análise dos dados para descrever características do passado ou predizer tendências do futuro. Tem ligação com outras técnicas e ciências, usando muitos conceitos e técnicas de estatística, visualização, reconhecimento de padrões, processamento de alto desempenho, aprendizado por máquina, inteligência artificial etc.

Para extrair o conhecimento, podemos utilizar diversos métodos como: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais, conforme definido por Fayyad (1996).

Os métodos de Data Mining analisados neste trabalho para escolha na implementação foram:

2.1.1 CLASSIFICAÇÃO

Aprendizado de uma função a ser usada para mapear dados em uma de várias classes discretas definidas previamente.

Segundo Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, tambem, a natureza da diferença, de forma que, com base em um conjunto de variáveis independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas. Dentre os métodos de Classificação podemos citar:

2.1.1.1 MÍNIMA DISTÂNCIA EUCLIDIANA

A distância Euclidiana é a distância entre dois pontos, que pode ser provada pela aplicação do teorema de Pitágoras. No método usa-se o protótipo de uma classe como assinatura e comparam-se atributos de uma instância com os protótipos. O protótipo mais próximo, considerando a distância Euclideana indica a classe.

2.1.1.2 ÁRVORES DE DECISÃO

É uma estrutura de árvore, que faz representações simples do conhecimento, tem a função de particionar recursivamente um conjunto de treinamento, até que casa subconjunto contenha casos de uma única classe.

Segundo Quinlan (1993), os resultados obtidos, após a construção de uma árvore de decisão, são dados organizados de maneira compacta, com a árvore podendo ser utilizada para classificar novos casos.

Contém nodos que representam os atributos, arcos que correspondem ao valor de um atributo e nodos folha que designam uma classificação. A árvore pode ser lida a partir do teste encontrado na parte superior da mesma, normalmente chamado nó raiz da árvore. Como exemplo de algoritmos que implementam a árvore de decisão temos o ID3, C4.5 e C5.0.

2.1.1.3 VIZINHOS MAIS PRÓXIMOS

Nesta técnica, o conjunto de dados mais comum é mantido na memória para que os novos dados sejam comparados com estes.

Se uma instância de classe desconhecida estiver perto de uma classe conhecida, a que foi mantida para comparação, as classes devem ser as mesmas. Neste método, não criamos protótipos ou assinaturas, usamos as próprias instâncias.

2.1.1.4 REDES NEURAIS

Uma Rede Neural Artificial (RNA) é uma técnica de construção de um modelo matemático, originalmente concebida com base no estudo do cérebro humano. Tem capacidade para aprendizado, generalização, associação e abstração. As redes neurais apresentam uma estrutura altamente interconectada e trabalhando em paralelo, assim como no cérebro humano. É composta por processadores simples (neurônios artificiais) conectados entre si.

2.1.2 REGRESSÃO OU PREDIÇÃO

Aprendizado de uma função usada para mapear os valores associados aos dados. Observa-se, conforme Gujarati (2000), que o método dos mínimos quadrados ordinários, tem propriedades estatísticas relevantes e apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão.

2.1.3 AGRUPAMENTO OU CLUSTERIZAÇÃO

Identicação de grupos de dados onde os dados têm características semelhantes aos do mesmo grupo e onde os grupos tenham características diferentes entre si. Neste tipo de análise, segundo Pereira (1999), o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas

(variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles. Na sequência, os agrupamentos por proximidade geométrica são efetuados, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados. O algoritmo K-Médias pode ser destacado neste método. Ele minimiza o erro quadrático calculado entre as instâncias e os centróides dos grupos.

2.1.4 SUMARIZAÇÃO

Descrição compacta do que caracteriza um conjunto de dados (ex. conjunto de regras que descreve o comportamento e relação entre os valores dos dados de meteorologia). As medidas de posição e variabilidade são exemplos simples de sumarização.

2.1.5 REGRAS DE ASSOCIAÇÃO

Identicação de grupos de dados que apresentam co-ocorrência entre si (ex. cesta de compras). A ideia é a derivação de correlações variadas que permitam subsidiar a tomada de decisões.

2.1.6 SÉRIES TEMPORAIS

Um conjunto de observações tomadas em tempos determinados, comumente em intervalos iguais (MURRAY, 1993). Entendemos que é uma sequência de observações sobre a variável de interesse, que é observada em pontos temporais discretos. A descrição do processo ou fenômeno se dá pela análise do comportamento.

2.1.7 DETECÇÃO DE DESVIOS OU *OUTLIERS*

Todas as técnicas de detecção de *outliers* fazem a seleção/identificação de dados que deveriam seguir um padrão ou comportamento esperado, mas não o fazem. Detecção de *outliers*, detecção de anomalias, detecção de ruído, detecção de desvio e mineração de exceções são outras nomenclaturas para o método. (HODGE; AUSTIN 2004)

3 ESTADO DA ARTE

Para a atividade de Mineração de Dados, foi escolhido o software Tanagra, por dispor de uma coleção de algoritmos para diversas tarefas de mineração de dados, pelo seu uso ser livre e gratuito e pela facilidade de uso.

3.1 TANAGRA

O Tanagra é um software livre de mineração de dados desenvolvido em Delphi por pesquisadores da Universidade de Lyon. É utilizado para fins acadêmicos e de pesquisa. Como é um open source (código aberto), os usuários podem acessar seus códigos e adicionar seus próprios algoritmos.

Este projeto é o sucessor do SIPINA que implementa vários algoritmos de aprendizado supervisionado, especialmente uma construção interativa e visual de árvores de decisão. O Tanagra propõe diversos métodos de mineração de dados, análise exploratória e classificação estatística. Conta com processos de classificação supervisionada e não-supervisionada, tais como clusterização, análise fatorial, estatísticas parametrizadas e não parametrizadas e regras de associação. Este é um sistema integrado para análises estatísticas e de Mineração de Dados (RAKOTOMALALA, 2005).

4 DESCRIÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS APLICADO AO PROBLEMA

4.1 IDENTIFICAÇÃO DO PROBLEMA

O volume de informações meteorológicas não permite que sua análise seja feita pelos métodos tradicionais (planilhas, gráficos etc), já que com esses métodos podemos gerar relatórios, mas não a extração do conhecimento. Dessa forma, as técnicas de mineração de dados serão aplicadas nas variáveis da

cidade de São Martinho da Serra no Rio Grande do Sul que são: radiação solar e temperatura.

4.2 PRÉ-PROCESSAMENTO

Os dados foram coletados pelo Sistema Nacional de Dados Ambientais (SINDA), que é alimentado pelas Plataformas de Coleta de Dados (PCDs).

Foram selecionados os dados de radiação solar acumulada, temperatura do ar, temperatura máxima e temperatura mínima, juntamente com a data do fenômeno, do mês de janeiro de 2000 a dezembro de 2014. Para extração dos dados, as opções possíveis de formato eram xls, xml e csv.

Como há integração entre as ferramentas Tanagra e Excel, os dados foram extraídos em formato xls, sendo as tabelas já separadas por tabulações, o que permitiu ser importada pelo Tanagra sem alteração nenhuma na estrutura da planilha. Os registros a qual algum campo não tivesse valor foram filtrados e apagados da planilha, não comprometendo o resultado final, já que para cada dia do ano temos um registro a cada 3 horas. A Figura 2 apresenta um exemplo da estrutura da planilha como é extraída.

Figura 2 – Dados exportados para Excel

	A	B	C	D	E	F
1	DataHora	RadSolAcu	TempAr	TempMax	TempMin	UmidRel
2	2003-04-30 21:00:00.0	0.2	18.0	20.5	16.5	99.0
3	2003-04-30 18:00:00.0	1.0	20.0	20.5	16.5	99.0
4	2003-04-30 15:00:00.0	0.9	20.0	20.0	16.5	99.0
5	2003-04-30 12:00:00.0	0.1	18.5	19.0	16.5	99.0
6	2003-04-30 09:00:00.0	0.0	17.0	18.0	16.5	99.0
7	2003-04-30 06:00:00.0	0.0	17.5	18.0	16.5	99.0
8	2003-04-30 03:00:00.0	0.0	17.5	18.5	16.5	99.0
9	2003-04-30 00:00:00.0	0.0	16.5	18.5	16.5	99.0
10	2003-04-29 21:00:00.0	0.0	17.0	19.0	16.5	99.0
11	2003-04-29 18:00:00.0	0.1	17.5	19.0	16.5	99.0
12	2003-04-29 15:00:00.0	0.1	17.5	19.0	16.5	99.0
13	2003-04-29 12:00:00.0	0.1	17.5	19.0	16.5	99.0
14	2003-04-29 09:00:00.0	0.0	17.0	19.0	16.0	99.0
15	2003-04-29 06:00:00.0	0.0	17.0	19.0	16.0	99.0
16	2003-04-29 03:00:00.0	0.0	17.5	19.0	16.0	99.0
17	2003-04-29 00:00:00.0	0.0	18.5	19.0	16.0	99.0
18	2003-04-28 21:00:00.0	0.2	18.0	18.5	16.0	99.0

Fonte: Dados da pesquisa.

4.3 TRANSFORMAÇÃO

Nesta fase de preparação dos dados alguns campos serão codificados ou transformados de forma a tornar viável ou facilitar a extração de padrões.

A seguir, estão listadas as ações efetuadas para transformação dos dados.

• A representação de data e hora do registro que estava na forma DD/MM/AAAA e HH:MM:SS foi

convertido para data, no formato MM/AA, com o objetivo de facilitar o processo de mineração de dados, diminuindo a quantidade de variáveis.

Como os dados foram extraídos em planilhas separadas por mês do ano, foram agrupados os dados referentes a cada mês em uma planilha separada por estações.

Na Figura 3 é apresentado um exemplo sobre como foram importados os dados. Essa transformação nos dados foi feita para aplicação da árvore de decisão sobre as variáveis. Para aplicação das regras de sumarização, foram aplicadas as seguintes transformações:

• Todos os campos relativos aos dados numéricos eram do tipo discreto na base original. Como as técnicas de mineração de dados aplicadas ao trabalho, trabalham com variáveis do tipo contínuo, para geração das regras, todos os campos foram convertidos de discreto para contínuo, apenas substituindo os pontos por vírgulas;

• A transformação, do campo data e hora, foi realizada da mesma forma que para a árvore de decisão;

• O arquivo de dados está separado por ano.

Figura 3 – Dados Formatados

	A	B	C	D	E	F	G
1	DataHora	RadSolAcu	TempAr	TempMax	TempMin	UmidRel	UmidInt
2	2000-07	5,5	12	15	9,5	89	35
3	2000-07	5,5	11,5	15	9,5	91	35
4	2000-07	0	10	15	9,5	97	35
5	2000-07	0	10	15	9,5	95	35
6	2000-07	0,25	13	15	10	90	35
7	2000-07	2,5	18,5	15	10	87	35
8	2000-07	3,75	19,5	15	10	93	0
9	2000-07	4	20	15	10	97	0
10	2000-07	4	19,5	21	10	97	0
11	2000-07	4	19	21	10	95	0
12	2000-07	0	19,5	21	10	90	0
13	2000-07	0	20			82	
14	2000-07	0	11,5	21	11,5	97	0
15	2000-07	0,25	7	21	11,5	97	35
16	2000-07	2,25	10	21	11,5	81	35

Plan1 Plan2 Plan3

Pronto

Fonte: Dados da pesquisa.

4.4 TÉCNICAS DE MINERAÇÃO ESCOLHIDA

A partir do estudo das técnicas de mineração de dados, as que mais se adaptaram às informações que se pretende obter do banco foram às técnicas de árvore de decisão e sumarização.

A seguir, são descritas as funcionalidades de cada técnica em relação ao banco de dados utilizado:

• Sumarização – Esta técnica permite a descrição do que caracteriza um conjunto de dados. Por exemplo, agrupar os dados de temperatura e

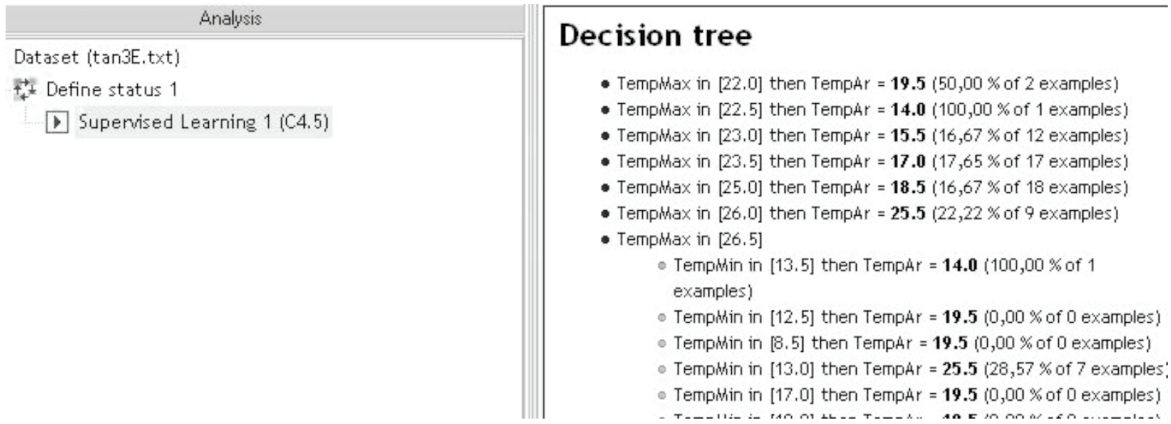
radiação solar por estações ou separados por ano para verificar suas estatísticas;

• Árvores de Decisão – Por meio de árvores de decisão são geradas regras que podem ser úteis na consulta do comportamento das variáveis, tais como verificar a comparação da temperatura real com a amplitude térmica.

Para mineração de dados, utilizou-se a técnica de árvores de decisão com o algoritmo C4.5 – desenvolvido por Quinlan (2001), que permite trabalhar com valores contínuos, indisponíveis, podar árvores de decisão e derivar regras. Assim, foi gerada uma árvore a partir dos parâmetros de temperatura, temperatura mínima e temperatura máxima, a fim de analisar se a previsão de amplitude térmica corresponde à temperatura registrada no dia, e verificar se a temperatura registrada segue algum padrão de aproximação da amplitude.

Na Figura 4, pode-se observar o exemplo da árvore de decisão criada a partir dos dados da estação verão de 2005.

Figura 4 – Árvore de Decisão

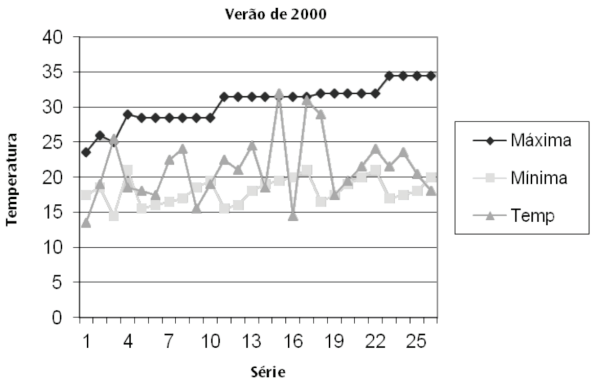


Fonte: Dados da pesquisa.

Depois de extraída a árvore de decisão, os dados mais relevantes apontados pela árvore – topo – foram passados para uma planilha Microsoft Excel, onde foi gerado o gráfico, comparando a amplitude térmica com a temperatura real para melhor visualização.

Nas Figuras 5, 6 e 7, visualizam-se os gráficos das estações de verão, para avaliar os resultados e comparar as evoluções.

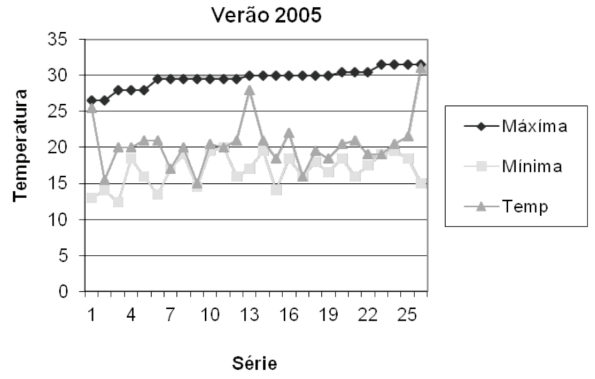
Figura 5 – Dados do verão de 2000



Fonte: Dados da pesquisa.

Pode-se observar na Figura 5 que a temperatura real tende a se aproximar com a temperatura mínima nos dados da estação verão de 2005.

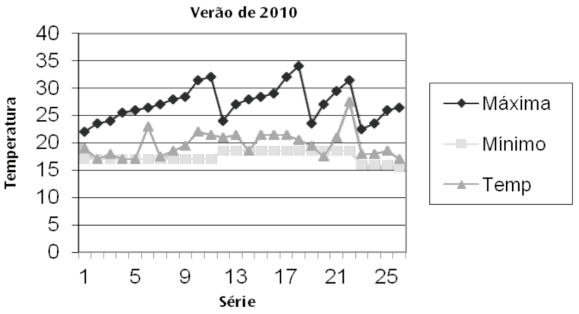
Figura 6 – Dados do verão 2005



Fonte: Dados da pesquisa.

Nos dados referentes ao verão de 2005, verifica-se que a temperatura real continua apresentando proximidade com os dados de temperatura mínima, além da temperatura máxima, não demonstrar variância significativa.

Figura 7 – Dados do verão 2010



Fonte: Dados da pesquisa.

Os dados referentes ao verão de 2010 mantiveram o mesmo comportamento: temperatura real, tendendo à temperatura mínima. Mas pode-se observar que a temperatua máxima teve maior variância.

Com as regras de sumarização, utilizando o algoritmo *Statistics – Group Characterization*, pode-se verificar também, algumas tendências pela Figura 8.

Figura 8 – *Group Characterization*

DataHora=2001-09			
Examples			[8,8 %] 139
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
UmidRel	0,78	87,15 (15,37)	86,17 (15,56)
RadSolAcum	-0,42	1,59 (2,32)	1,67 (2,41)
TempMin	-4,99	11,83 (3,99)	13,97 (5,27)
TempAr	-5,11	15,37 (4,75)	17,88 (6,07)
TempMax	-5,68	20,87 (4,64)	23,63 (6,00)
Discrete attributes : [Recall] Accuracy			

DataHora=2001-04			
Examples			[10,9 %] 171
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
UmidRel	4,23	90,93 (12,74)	86,17 (15,56)
TempMin	3,54	15,32 (2,14)	13,97 (5,27)
TempMax	2,81	24,85 (3,43)	23,63 (6,00)
TempAr	2,36	18,92 (3,31)	17,88 (6,07)
RadSolAcum	-1,63	1,39 (2,19)	1,67 (2,41)
Discrete attributes : [Recall] Accuracy			

DataHora=2001-08			
Examples			[12,3 %] 193
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
TempMax	0,19	23,71 (4,25)	23,63 (6,00)
RadSolAcum	-0,78	1,54 (2,26)	1,67 (2,41)
TempAr	-1,74	17,17 (4,87)	17,88 (6,07)
TempMin	-4,01	12,54 (3,65)	13,97 (5,27)
UmidRel	-5,61	80,28 (18,89)	86,17 (15,56)
Discrete attributes : [Recall] Accuracy			

Fonte: Dados da pesquisa.

O algoritmo *Group Characterization* permite confrontar diversos subgrupos, comparando as estatísticas descritivas sobre os dados. Realizada a caracterização do subconjunto de dados de alguns meses aleatórios referente ao ano de 2001 pode-se fazer a comparação com dados do ano de 2011 na Figura 9.

Figura 9 – *Group Characterization*

DataHora=2011-10			
Examples			[8,9 %] 205
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
RadSolAcum	1,21	2,64 (3,25)	2,39 (3,11)
UmidRel	-2,13	87,70 (19,07)	90,09 (16,79)
TempMax	-2,65	23,64 (3,82)	24,55 (5,18)
TempAr	-3,15	17,36 (4,67)	18,51 (5,48)
TempMin	-5,45	12,26 (2,74)	13,73 (4,04)
Discrete attributes : [Recall] Accuracy			

DataHora=2011-11			
Examples			[9,4 %] 217
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
TempMax	7,66	27,12 (4,38)	24,55 (5,18)
TempAr	5,08	20,31 (5,42)	18,51 (5,48)
RadSolAcum	3,83	3,16 (3,60)	2,39 (3,11)
TempMin	1,25	14,05 (3,07)	13,73 (4,04)
UmidRel	-7,36	82,10 (20,31)	90,09 (16,79)
Discrete attributes : [Recall] Accuracy			

DataHora=2011-04			
Examples			[9,1 %] 209
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
TempMax	0,83	24,83 (2,77)	24,55 (5,18)
TempMin	-0,03	13,72 (2,41)	13,73 (4,04)
TempAr	-0,40	18,37 (3,94)	18,51 (5,48)
UmidRel	-1,51	88,41 (17,65)	90,09 (16,79)
RadSolAcum	-2,91	1,79 (2,55)	2,39 (3,11)
Discrete attributes : [Recall] Accuracy			

Fonte: Dados da pesquisa.

A coluna *Test Value* (Valor de Teste) mostra a resistência da diferença. Quanto maior é o valor absoluto desse indicador, maior é a diferença entre a média calculada no subgrupo e a média calculada sobre o conjunto de dados inteiro.

Dessa forma, percebe-se que, por exemplo, o mês de abril apresentou variância muito maior de temperatura dentro do ano de 2001 do que em 2011. Com o algoritmo *Univariate Continuous Stat*, consegue-se verificar as estatísticas básicas de média de todo período do ano de 2001 e 2011, nas Figuras 10 e 11.

Figura 10 – Univariate Continuous Stat 2001

Attributes : 5
Examples : 2297

Results

Attribute	Min	Max	Average	Std-dev	Std-dev/avg
RadSolAcum	0	12,7	2,3860	3,1051	1,3014
TempAr	5	36	18,5089	5,4750	0,2958
TempMax	10,5	36,5	24,5527	5,1840	0,2111
TempMin	5	22,5	13,7255	4,0407	0,2944
UmidRel	13	100	90,0858	16,7860	0,1863

Computation time : 0 ms.

Fonte: Dados da pesquisa.

Ao analisar a variável temperatura, temos uma média de 18,5 com o desvio padrão de 3,1.

Figura 11 – Univariate Continuous Stat 2011

Attributes : 5
Examples : 1571

Results

Attribute	Min	Max	Average	Std-dev	Std-dev/avg
RadSolAcum	0	10,9	1,6705	2,4076	1,4413
TempAr	0,5	33,5	17,8835	6,0663	0,3392
TempMax	7	34	23,6314	6,0041	0,2541
TempMin	0,5	22,5	13,9663	5,2730	0,3776
UmidRel	33	100	86,1738	15,5550	0,1805

Computation time : 0 ms.

Fonte: Dados da pesquisa.

Já na Figura 11 tem-se uma média de 17,88 com desvio padrão de 6,0 – uma média de temperatura ainda mais baixa que a de 2001. A partir dos resultados obtidos pode-se sugerir, ao contrário do que se esperava – que não se observa diferença significativa de temperatura nos últimos dez anos na cidade analisada, assim como a variação da radiação solar não foi consideravelmente alta.

5 CONSIDERAÇÕES FINAIS

Acredita-se que os objetivos propostos para o presente trabalho foram alcançados, tendo-se em vista que todas as etapas previstas foram realizadas: estudo do banco de dados; estudo da ferramenta utilizada; estudo e definição das técnicas de mineração a serem adotadas; pré-processamento dos dados; aplicação das técnicas de mineração por meio da ferramenta Tanagra; conclusão sobre os dados minerados.

Algumas dificuldades foram encontradas, principalmente no que diz respeito à documentação da ferramenta Tanagra. Devido às dificuldades de se trabalhar com os métodos da ferramenta por falta de documentação, foi dispendido tempo apenas para a escolha do tipo de mineração a ser realizado, já que os testes foram realizados com a maioria dos algoritmos disponíveis na ferramenta, até se encontrar um resultado satisfatório.

A metodologia proposta de mineração de dados possibilitou encontrar de forma automática e quantitativa alguns padrões de conhecimento geral sobre climatologia, e buscou encontrar padrões que possam ser úteis para a meteorologia e áreas semelhantes.

REFERÊNCIAS

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards

a Unifying Framework. **Proceedings of Second International Conference on Knowledge Discovery and Data Mining**. Portland: Oregon, 1996.

GUJARATI, D.N. **Econometria Básica**. São Paulo: Makron Books, 2000.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge, Massachusetts: The MIT Press, 2001.

HODGE, V.J. A **Survey of outlier detection methodologies**. Rotterdan: Kluwer Academic Publishers, 2004 .

MATTAR, F.N. **Pesquisa de marketing**. São Paulo: Atlas, 1998.

MURRAY, R.S. **Estatística**. São Paulo: Makron Books, 1993.

PEREIRA, J.C.R. **Análise de dados qualitativos**. São Paulo: Edusp/Fapesp, 1999.

QUINLAN, J.C. **C4.5**: Programs for Machine Learning. San Mateo: Morgan Kaufmann, 1993.

RAKOTOMALALA, R., TANAGRA: a free software for research and academic purposes. **Proceedings of EGC'2005, RNTI-E-3**, France, 2005.

TUBELIS, A.N; Nascimento, F.J.L. **Meteorologia descritiva**: fundamentos e aplicações brasileiras. São Paulo: Nobel, 1992.

Recebido em: 22 de Março de 2016
Avaliado em: 25 de Março de 2016
Aceito em: 7 de Abril de 2016

1. Bacharel em Sistemas de Informação, Centro Universitário Ritter dos Reis. E-mail: morgana.wagner@gmail.com
2. Doutor em Ciência da Computação, Centro Universitário Ritter dos Reis. E-mail: vinicius@uniritter.edu.br
3. Doutor em Engenharia Mecânica, Universidade Federal do Rio Grande do Sul. E-mail: jorge.zabadal@ufrgs.br