sid.inpe.br/mtc-m19/2012/01.24.15.41-TDI

USO DE ALGORITMOS DE MINERAÇÃO DE DADOS PARA AGRUPAMENTO E BUSCA DE ERROS EM SÉRIES TEMPORAIS COLETADAS A PARTIR DE GEOSSENSORES: UM ESTUDO DE CASO NA MATA ATLÂNTICA

Gerson da Penha Neto

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Nandamudi Lankalapalli Vijaykumar, e Celso von Randow, aprovada em 17 de fevereiro de 2012.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3B8Q4T5>

INPE
São José dos Campos
2012

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

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELECTUAL DO INPE (RE/DIR-204):
Presidente:
Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)
Membros:
Dr. Antonio Fernando Bertachini de Almeida Prado - Coordenação Engenharia e Tecnologia Espacial (ETE)
Drª Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)
Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)
Dr. Germano de Souza Kienbaum - Centro de Tecnologias Especiais (CTE)
Dr. Manoel Alonso Gan - Centro de Previsão de Tempo e Estudos Climáticos (CPT)
Drª Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação
Dr. Plínio Carlos Alvalá - Centro de Ciência do Sistema Terrestre (CST)
BIBLIOTECA DIGITAL:
Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)
REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:
Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)
Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)
EDITORAÇÃO ELETRÔNICA:
Vivéca Sant´Ana Lemos - Serviço de Informação e Documentação (SID)

sid.inpe.br/mtc-m19/2012/01.24.15.41-TDI

USO DE ALGORITMOS DE MINERAÇÃO DE DADOS PARA AGRUPAMENTO E BUSCA DE ERROS EM SÉRIES TEMPORAIS COLETADAS A PARTIR DE GEOSSENSORES: UM ESTUDO DE CASO NA MATA ATLÂNTICA

Gerson da Penha Neto

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Nandamudi Lankalapalli Vijaykumar, e Celso von Randow, aprovada em 17 de fevereiro de 2012.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3B8Q4T5>

INPE
São José dos Campos
2012

Dados Internacionais de Catalogação na Publicação (CIP)

P376u

Penha Neto, Gerson da.
Uso de algoritmos de mineração de dados para agrupamento e busca de erros em séries temporais coletadas a partir de geossensores: um estudo de caso na Mata Atlântica / Gerson da Penha Neto. – São José dos Campos : INPE, 2012.
xx + 49 p. ; (sid.inpe.br/mtc-m19/2012/01.24.15.41-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2012.
Orientadores : Drs. Nandamudi Lankalapalli Vijaykumar, e Celso von Randow.

1. mineração de dados. 2. séries temporais. 3. sensoriamento remoto. I.Título.

CDU 004.8

Copyright © 2012 do MCT/INPE. Nenhuma parte desta publicação pode ser reproduzida, armazenada em um sistema de recuperação, ou transmitida sob qualquer forma ou por qualquer meio, eletrônico, mecânico, fotográfico, reprográfico, de microfilmagem ou outros, sem a permissão escrita do INPE, com exceção de qualquer material fornecido especificamente com o propósito de ser entrado e executado num sistema computacional, para o uso exclusivo do leitor da obra.

Copyright © 2012 by MCT/INPE. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, microfilming, or otherwise, without written permission from INPE, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use of the reader of the work.

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

Computação Aplicada

Dr. Rafael Duarte Coelho dos Santos

Presidente / INPE / SJCampos - SP

Dra. Lúbia Vinhas

Vice-Presidente / INPE / São José dos Campos - SP

Dr. Nandamudi Lankalapalli Vijaykumar

Orientador(a) / INPE / SJCampos - SP

Dr. Celso Von Randow

Orientador(a) / INPE / Cachoeira Paulista - SP

Dr. Ádamo Lima de Santana

Convidado(a) / UFPA / Belém - PA

Dr. Diego Lisboa Cardoso

Convidado(a) / UFPA / Belém - PA

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): Gerson da Penha Neto

São José dos Campos, 17 de fevereiro de 2012

"A vida será mais complicada se você possuir uma curiosidade ativa, além de aumentarem as chances de você entrar em apuros, mas será mais divertida".

Edward Speyer
em *"Seis Caminhos a Partir de Newton"*, 1994

A meus pais Valmir Paulo e Irani e a minha esposa Luciana

AGRADECIMENTOS

A Deus por toda bondade, por toda inspiração e por cuidar de mim em todos os momentos, por todas as oportunidades e experiências, por todas as pessoas que conheci.

Aos espíritos de luz, paz e bondade que sempre me ajudam.

Muito obrigado aos professores Dr. Celso von Randow e Dr. Rafael Duarte Coelho dos Santos, obrigado pela oportunidade de trabalhar e aprender.

Em especial agradeço ao professor Dr. Nandamudi Lankalapalli Vijaykumar, muito obrigado por ter me guiado de Belém até São José dos Campos, por todo o conhecimento passado e por sempre ter me ajudado quando tinha dificuldades.

Ao Dr. Humberto Rocha, por ter cedido a base de dados necessária para a dissertação.

Ao programa de pós-graduação em Computação Aplicada do INPE e a todos os professores que se dedicaram em passar a mim o máximo de conhecimento.

A todos os amigos que me ajudaram nessa caminhada, obrigador por terem me ajudado e por todos os momentos de descontração após horas de trabalho.

À Fundação de Amparo à Pesquisa do Estado de São Paulo, obrigado por todo o apoio e financiamento.

RESUMO

Séries temporais talvez sejam o mais comum tipo de dado utilizado em vários aspectos pelas pesquisas. São usadas na medicina, em finanças, indústria, em pesquisas científicas e em muitas outras áreas. Instituições de pesquisas como universidades e o próprio INPE estudam fenômenos naturais pertencentes ao escopo da meteorologia, que muitas vezes estão representados em bases de dados como séries temporais. Verificar a qualidade dos dados coletados ajuda a garantir a qualidade de muitos trabalhos e estudos que dependem dos dados. Este trabalho apresenta um estudo de como o uso de ferramentas conhecidas em mineração de dados podem ser utilizadas para agrupar e encontrar erros minuciosos em séries temporais. Utilizou-se um estudo de caso realizado por uma equipe de pesquisa da Universidade de São Paulo na Mata Atlântica. O objetivo é encontrar grupos e erros nas séries temporais que possam indicar possíveis falhas nos equipamentos utilizados durante a coleta de informações na natureza. Foi possível verificar que determinadas série temporais foram consideradas distintas, mesmo sendo obtidas a partir do mesmo equipamento, na mesma posição geografica. Também foi possível encontrar pequenas falhas, ruídos, que provacam variações na classificação e agrupamento das séries.

USE OF DATA MINING ALGORITHMS FOR CLUSTERING AND FIND ERRORS IN TIME SERIES COLLECTED FROM GEOSENSORS: A CASE STUDY IN THE ATLANTIC FOREST

ABSTRACT

Time series are perhaps the most common data type used in various aspects in researches. They are used in medicine, finance, industry, scientific research and in many other areas. Research institutions such as universities and the National Institute for Space Research study natural phenomena pertaining to the scope of meteorology, which are often represented in time series databases. To check the quality of data collected helps to ensure the quality of many papers and studies that rely on data. This work presents a study of how the use of knows data mining tools can be used to group and find errors in time series. We used a case study conducted by a research team from the University of São Paulo in the Atlantic forest. The goal is to find groups in the time series and errors that may indicate possible failures in the equipment used during the collection of information.

LISTA DE FIGURAS

Pág.

2.1 Alinhamento de duas sequências dependentes do tempo para o cálculo de distância. Os pontos casados são indicado pelas retas. 7
2.2 Exemplo da matriz de proximidade e *warping path*. Cada ponto do *warping path* representa o menor valor de distância encontrado. 8
2.3 Classificação de objetos utilizando o kNN com $k = 3$. Cada cor representa uma classe e os círculos marcados com um P representam objetos desconhecidos a serem rotulados. 9
2.4 Classificação de objetos utilizando o k-means com $K = 2$. Em (a) é mostrado um conjunto total de dados ainda não classificados, representados pela cor verde. O centros inicias são representados pelos símbolos em azul e vermelho. O algoritmo termina em (i), quando são formados dois grupos e percebe-se que de (f) para (i) as mudanças na posição dos centros são menos significativas, isso caracteriza o critério de parada. . . 11
2.5 Curvas agrupadas utilizando o algoritmo *Hierarchical Clustering* de forma aglomerativa. 13
2.6 Estrutura da Rede Neural SOM. 14
2.7 Vizinhança. 16
2.8 Exemplo de uma série temporal de medidas de temperatura. Medidas feitas em intervalos de $30s$ durante 1 dia. 17
2.9 Representação de uma série nas sete primeiras componentes de frequência. 18
2.10 Representação de uma série utilizando a função *Haar wavelet*. 20
2.11 Representação de uma série utilizando *Piecewise Aggregate Approximation*. 21
2.12 Valores de probabilidades para $\alpha = 3, \ldots, 10$. 23
2.13 Série temporal normalizada, primeiramente discretizada usando o PAA e então usando a tabela com faixas de probabilidades mapear os coeficientes da representação por PAA em símbolos. Para este caso com, $n = 128, w = 8$ e $\alpha = 3$. A série temporal é mapeada para a cadeia *baabccbc* 23

3.1 Sensores utilizados pelo *Atlantic Rainforest Sensor Networking Project*. . 26
3.2 Topologia da rede de sensores utilizada pelo *Atlantic Rainforest Sensor Networking Project*. 27

3.3 Exemplo do arquivo ASCII de dados brutos. Cada linha contém um *time stemp*, 3 medidas de temperatura e uma medida de umidade relativa respectivamente. 27

4.1 Mapa gerado pela rede SOM para agrupar séries temporais de temperatura. 31
4.2 Temperaturas medidas pelo nó sensor 22 em 24h. 32
4.3 Representação em *Piecewise Aggregate Approximation* para a temperatura média do nó sensor 6. 32
4.4 Agrupamento utilizando o algoritmo *Hierarchical Clustering* os nós sensores 7, 8, 13, 51, 59, 60, 61. 33
4.5 Mapa topológico da rede de sensores. 35
4.6 Séries de temperatura medidas pelo nó sensor 37 em 21 dias. 36
4.7 Representação em SAX da série temporal do sensor 3 do nó sensor 37 . . 37
4.8 Anomalias encontradas no sensor 3 de temperatura do nó sensor 37 . . . 38

5.1 Interface do sistema de processamento e geração de gráficos personalizados desenvolvidos no projeto *Desenvolvimento de Aplicação de Rede de Geossensores para Monitoramento Ambiental*. 42

LISTA DE ABREVIATURAS E SIGLAS

DTW – *Dynamic Time Warping*
kNN – *k-Nearest Neighbors*
HC – *Hierarchical Clustering*
SOM – *Self-Organizing Maps*
DFT – *Discrete Fourier Transform*
DWT – *Discrete Wavelet Transform*

SUMÁRIO

Pág.

1 INTRODUÇÃO 1
1.1 Trabalhos Correlatos 3

2 FUNDAMENTOS TEÓRICOS 5
2.1 Medidas de Distância 5
2.1.1 Distância Euclidiana 5
2.1.2 Matriz de Proximidade 6
2.1.3 DTW - Dynamic Time Warping 6
2.2 Algoritmos de Agrupamento e Classificação 8
2.2.1 k-Nearest Neighbors 8
2.2.2 k-Médias 9
2.2.3 HC - Hierarchical Clustering 12
2.2.4 SOM - Self-Organizing Maps 13
2.3 Séries Temporais 16
2.3.1 Definição 16
2.3.2 Representação de Séries Temporais 18
2.3.2.1 DFT - Discrete Fourier Transform 18
2.3.2.2 DWT - Discrete Wavelet Transform 19
2.3.2.3 PAA - Piecewise Aggregate Approximation 21
2.3.2.4 SAX - Symbolic Aggregate Approximation 22

3 ORIGEM DOS DADOS 25
3.1 *Atlantic Rainforest Sensor Networking Project* 25
3.2 Sensores e Dados Brutos 25

4 EXPERIMENTOS E RESULTADOS 29
4.1 Agrupamentos 29
4.1.1 Agrupamento Utilizando a Rede SOM 29
4.1.2 Agrupamento Utilizando o Hierarchical Clustering (HC) 31
4.2 Detecção de Falhas 34

5 CONCLUSÕES 41

5.1 Contribuições 41
5.2 Trabalhos Futuros 43

REFERÊNCIAS BIBLIOGRÁFICAS 45

1 INTRODUÇÃO

Nos últimos anos a atenção com relação ao clima brasileiro aumentou devido a problemas como secas, enchentes e deslizamentos de terra. O meio ambiente como um todo se tornou foco importante na política, agricultura, em pesquisas científicas, economia e em decisões tomadas com relação a segurança.

Muitos estudos científicos que objetivam compreender e prever os fenômenos do meio ambiente e suas influências no sistema climático, possuem como foco a interação entre a atmosfera e a biosfera terrestre.

Monitorar fenômenos atmosféricos ainda é uma atividade de custo elevado. As estruturas de observação são caras e demandam por manutenção constante. Além disso, as estruturas não podem ser implantadas em qualquer lugar ou de qualquer forma, o que em muitas vezes torna inviável a cobertura adequada da área de estudo.

Existem também dificuldades com relação a resolução temporal com que é feita a coleta de informações; Se ela não for grande o suficiente a compreensão dos fenômenos atmosféricos pode ficar comprometida.

O uso de novas tecnologias ganhou destaque como alternativa para solucionar as dificuldades ao se realizar medições de fenômenos atmosféricos. Em especial, uma tecnologia passou a ser usada por muitas pesquisas científicas: Redes de geossensores.

Geossensores são pequenos dispositivos eletromecânicos que podem medir características ambientais como temperatura, pressão e umidade, por exemplo. Esses dispositivos se comunicam uns com os outros através de uma conexão sem fio e formam uma rede de sensores. Também são capazes de armazenar informações, efetuar pequenos processamentos e distribuí-las.

As aplicações das redes de geossensores são muitas, desde detecção de catástrofes naturais, coleta de dados para estudos retrospectivos e o estudo dos ecossistemas (XU, 2002). Também existem aplicações no estudo e preservação da vida animal sensível à presença humana (MAINWARING et al., 2002).

As redes atualmente implantadas ainda são experimentais e possuem algumas dezenas de nós, cuja instalação é feita manualmente. Uma das características mais importantes de redes de geossensores é que seus nós tem limitação de energia, ca-

pacidade computacional e de memória. A limitação de energia é a mais crítica, pois com uma autonomia baixa é necessário descarregar as informações armazenadas nos sensores em períodos curtos.

Embora seja uma nova área de investigação, muito trabalho tem sido feito em redes de geossensores, principalmente com relação ao roteamento dos dados, que é a forma como os dados são transportados ao longo da rede. No entanto, não existe preocupação sobre como analisar a qualidade desses dados.

De modo geral as informações coletadas a partir de uma rede de geossensores são armazenadas em uma base de dados na forma de séries temporais. Isso permite que os dados sejam manipulados e analisados através de algoritmos de mineração de dados.

Neste trabalho é apresentado como o uso de técnicas de mineração de dados, algoritmos de agrupamento, classificação e representação de séries temporais podem avaliar a qualidade de dados coletados a partir de uma rede de geossensores.

Como estudo de caso foram utilizados dados de um projeto desenvolvido por pesquisadores da Universidade de São Paulo - USP. Os dados são referentes a medições realizadas na Mata Atlântica. Os principais objetivos deste trabalho são a busca por grupos e detalhes internos à séries temporais que evidenciem falhas não percebidas por pesquisadores. Isso é importante uma vez que a veracidade dos resultados da pesquisa pode ser influenciada pela qualidade das medições.

Este trabalho está estruturado da seguinte maneira: O Capítulo 2 apresenta os fundamentos téoricos utilizados. É descrito como realizar as principais medidas de dissimilaridade entre séries temporais. Também descreve os algoritmos de agrupamento, classificação e representação de séries temporais utilizados.

O Capítulo 3 apresenta uma descrição sobre o projeto *Atlantic Rainforest Sensor Networking*, utilizado como estudo de caso para este trabalho. Também é apresentada uma breve descrição sobre os sensores utilizados e seus principais componentes.

No Capítulo 4 são apresentados os resultados obtidos após se realizar os agrupamentos e a busca por falhas. São descritos os procedimentos realizados e a aplicação dos algoritmos. Os resultados de falhas são evidenciados com figuras.

Por fim, o Capítulo 5 dedica-se à apresentação das conclusões e considerações finais. São apontadas as contribuições do trabalho e também sugestões para trabalhos futuros.

1.1 Trabalhos Correlatos

Dentre os trabalhos realizados para busca de padrões, agrupamento de séries temporais e busca por falhas podem ser apontados como exemplos os seguintes:

Em Keogh e Pazzani (a) é trabalhada uma forma de representação para séries temporais que consiste em dividir uma série contínua em segmentos lineares. O objetivo da representação é possibilitar a avaliação visual de séries temporais extensas para facilitar possíveis análises visto que em muitas áreas, como a medicina, engenharia e finanças, análise visual é importante.

Além de possibilitar a avaliação visual, outro objetivo importante do trabalho é expor uma forma de representação que seja fiel ao máximo em relação as características da série temporal. Isso garante que resultados numéricos, como medidas de distâncias e autocorrelação sejam mais precisas e corretas.

No trabalho apresentado em Keogh et al. (2000) é introduzida a técnica *Piecewise Aggregate Approximation* (PAA) para redução de dimensionalidade de uma série temporal. Além de apresentar e formalizar o procedimento para o PAA, o trabalho também faz uma comparação com outras formas de representação como, a *Discrete Fourier Transform* (DFT) e a *Discrete Wavelet Transform* (DWT).

A comparação apresentada é de grande importância pois mostrou que o PAA, em conjunto com técnicas para medir distância entre séries temporais, obteve resultados com qualidade igual ou superior em relação as outras representações. Esse fato é importante uma vez que o PAA é de fácil assimilação e implementação se comparado com representações como DFT e DWT.

Em Keogh et al. (2002) é apresentado um algoritmo, denominado de *Tarzan*, para busca de padrões em séries temporais. O algoritmo *Tarzan* usa uma árvore de sufixos e modelos markovianos para encontrar uma frequência relativa entre possíveis padrões internos às séries temporais e assim poder predizê. Neste caso, para que seja possível o uso de árvores de sufixo é feita uma representação da série temporal em uma cadeia de símbolos.

Com relação a representação de séries temporais se destaca o trabalho feito por Lin et al. (2003). No trabalho é apresentada a técnica *Symbolic Aggregate Approximation* (SAX) para representação de séries temporais em cadeias de símbolos.

Os autores consideram SAX como primeira representação simbólica para uma série temporal que permite a redução de dimensionalidade. Essa afirmação acontece porque, para se obter a representação em SAX, primeiramente é preciso realizar o procedimento para o PAA.

Com relação ao uso de geossensores, pode-se destacar os seguintes trabalhos:

Em Xu (2002) é feito um retrato sobre as aplicações possíveis utilizando redes de geossensores. São identificados os principais desafios ao se implantar uma rede de sensores. Também são apresentadas aplicações como monitoramento ambiental, sistemas de observação climáticos e outras aplicações comerciais.

Já em Mainwaring et al. (2002) é fornecido um estudo aprofundado sobre a aplicação de redes de sensores no monitoramento de habitat. É apresentado um conjunto de requisitos e um sistema desenvolvidos para monitorar uma área experimental que serve de morada para aves marinhas. No trabalho é detalhado também as especificações de *hardware* usadas no desenvolvimento dos sensores.

No trabalho desenvolvido por Reis (2005) são descritos conceitos e definições sobre o que são geossensores e uma rede de geossensores. O trabalho apresenta formas de análises estatísticas para avaliar os dados coletados. Também são discutidas técnicas de mineração de dados e análises estatísticas espaciais.

Todos os trabalhos apresentados estão relacionados a pesquisas com sensores e monitoramento ambiental, contudo o foco principal não está em avaliar a qualidade das informações coletadas pelos sensores. Essa é a principal diferença entre o trabalho apresentado nesta dissertação. O objetivo principal é encontrar grupos ou parâmetros que apontam possíveis falhas nos equipamentos e também ajudem na avaliação da qualidade dos dados.

2 FUNDAMENTOS TEÓRICOS

Neste Capítulo são algoritmos de mineração de dados utilizados nesta dissertação. São apresentados os principais conceitos e definições sobre as medidas de distância utilizadas em mineração de dados de séries temporais. O Capitulo finaliza com conceitos sobre representação e redução de dimensionalidade de séries temporais.

2.1 Medidas de Distância

2.1.1 Distância Euclidiana

A distância euclidiana é uma das formas mais utilizadas para se calcular distância entre dois pontos contidos sobre um plano bidimensional (BACKER, 1995). Sua fórmula simples fez com que ela fosse utilizada por diversos trabalhos publicados em áreas variadas (GAUCH, 1982).

Considere dois conjuntos de pontos $P = \{p_1, \ldots, p_n\}$ e $Q = \{q_1, \ldots, q_n\}$. Calcula-se a distância euclidiana entre P e Q via:

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2.1}$$

Deve-se considerar algumas propriedades para distâncias como:

a) **Constância ou Auto-similaridade:** a distância entre um ponto e ele mesmo é zero. Logo

$$d(p, p) = 0$$

b) **Simetria:** a distância entre dois pontos é constante. Portanto

$$d(p, q) = d(q, p)$$

c) **Desigualdade Triangular:** Considerando três pontos p, q e w tem-se

$$d(p, q) \leq d(p, w) + d(q, w)$$

2.1.2 Matriz de Proximidade

A maioria dos algoritmos de agrupamento necessitam estimar um índice de afinidade, associação ou proximidade entre os objetos que se deseja agrupar. Diante disso, uma das formas de se calcular esse parâmetro é utilizar uma matriz de proximidade.

A matriz de proximidade $D(i, j)$ reúne pares de índices de proximidade, onde cada linha e coluna representam um objeto ou padrão a ser agrupado. Os índices contidos na matriz podem ser tanto de similaridade quanto de dissimilaridade e uma das formas de se medir dissimilaridade entre objetos é utilizar a distância euclidiana (ANDERBERG, 1973).

Assim, considerando dois conjuntos de pontos $P = \{p_1, \ldots, p_n\}$ e $Q = \{q_1, \ldots, q_n\}$, pode-se montar uma matriz de proximidade, utilizando a distância euclidiana, como:

$$D(P, Q) = \begin{bmatrix} d(p_1, q_1) & \ldots & d(p_1, q_n) \\ \vdots & \ddots & \vdots \\ d(p_n, q_1) & \ldots & d(p_n, q_n) \end{bmatrix}$$

Neste caso $d(p_j, q_j)$ é a distância euclidiana entre os pontos p_j e q_j e considera-se que $D(P, Q)$ é uma matriz quadrada, portanto P e Q tem a mesma dimensão.

2.1.3 DTW - Dynamic Time Warping

A técnica *Dynamic Time Warping* foi introduzida em meados da década de 80 por Joseph Kruskal and Mark Liberman e ganhou destaque na comunidade que atua em mineração de dados (KRUSKAL; LIBERMAN, 1983). Nos últimos anos tem sido utilizada como métrica em vários trabalhos de classificação e agrupamentos de séries temporais (KEOGH; PAZZANI, b).

DTW permite calcular a similaridade entre duas curvas dependentes do tempo, de maneira mais precisa que a distância euclidiana, distorcendo o eixo do tempo no intuito de encontrar as menores distâncias entre os pontos das curvas, veja Figura 2.1. Em particular DTW é considerado como uma alternativa para debilidades encontradas ao se usar a distância euclidiana (BERNDT; CLIFFORD, 1994).

Devido a flexibilidade do DTW é permitido que duas curvas similares, porém com

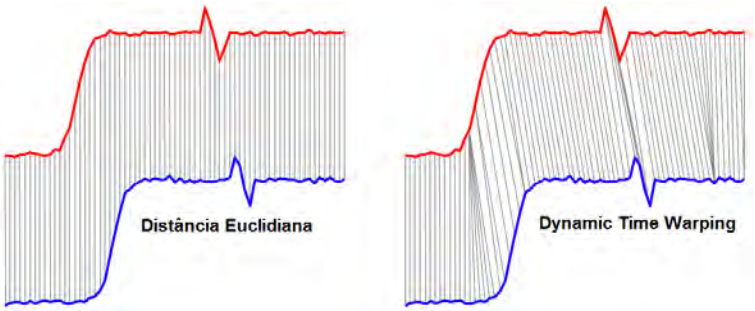

Figura 2.1 - Alinhamento de duas sequências dependentes do tempo para o cálculo de distância. Os pontos casados são indicado pelas retas.

diferentes comprimentos, se alinhem de maneira não linear e apesar de sua complexidade temporal ser $O(n^2)$ é considerada a melhor solução para problemas em mineração de dados relacionados a séries temporais sobre vários domínios como bioinformática (AACH; CHURCH, 2001), medicina (CAIANI et al., 1998), engenharia e entretenimento (ZHU; SHASHA, 2003).

Considerando duas curvas $P = \{p_1, \ldots, p_n\}$, com $n = 1, \ldots, N$ e $Q = \{q_1, \ldots, q_m\}$, com com $m = 1, \ldots, M$ para comparar P e Q usando-se o DTW primeiramente deve-se montar uma matriz de proximidade entre P e Q. Essa matriz de proximidade não será necessariamente quadrada e terá a dimensão MxN.

O objetivo de se montar a matriz de proximidade é encontrar um vetor w_k, com $k = 1, \ldots, K$ que contenha os mínimos valores de distância entre todos os pontos das curvas.

A Figura 2.2 mostra exemplos de como é a matriz de proximidade e o vetor w_k. Denomina-se w_k como *warping path*. Uma vez que w_k é obtido, pode-se então calcular o valor para dissimilaridade entre as curvas como:

$$D_{PQ} = \frac{\sum_{k=1}^{K} w_k}{K} \tag{2.2}$$

Quanto menor D_{PQ} mais alto é a similaridade entre as curvas.

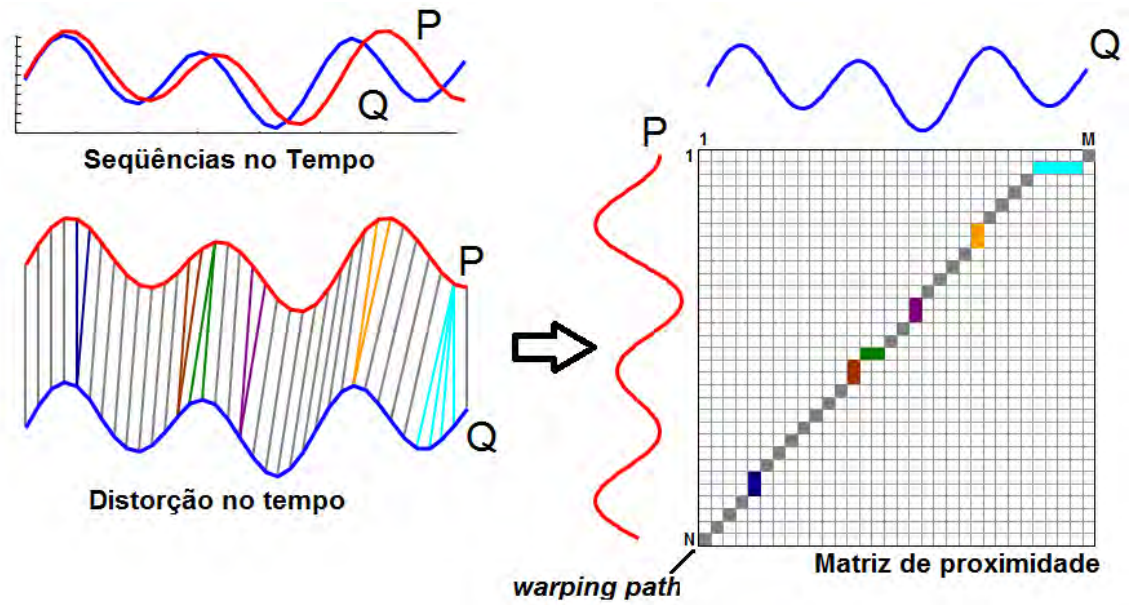

Figura 2.2 - Exemplo da matriz de proximidade e *warping path*. Cada ponto do *warping path* representa o menor valor de distância encontrado.
Fonte: Adaptado de (XI et al., 2006)

2.2 Algoritmos de Agrupamento e Classificação

2.2.1 k-Nearest Neighbors

O algoritmo kNN é um método simples e antigo para classificação de objetos (FIX; HODGES, 1951). O kNN classifica objetos considerando uma semelhança, proximidade ou a distância entre um objeto de teste e um grupo de treinamento (STEINBACH; TAN, 2009).

Existem três parâmetros principais que devem ser utilizados pelo algoritmo kNN: Um conjunto de objetos já rotulados, ou seja, um grupo conhecido de objetos, uma métrica de similaridade para se obter a distância entre os objetos e o valor de k, que corresponde ao número de vizinhos próximos considerados para o cálculo da distância e classificação.

Para classificar um objeto desconhecido ou pode-se dizer também não rotulado, calcula-se a distância entre esse objeto e outros objetos já conhecidos pertencentes a dois ou mais grupos. Assim identifica-se os k-vizinhos mais próximos. O rótulo desses k-vizinhos mais próximos é então usado para determinar a qual classe o

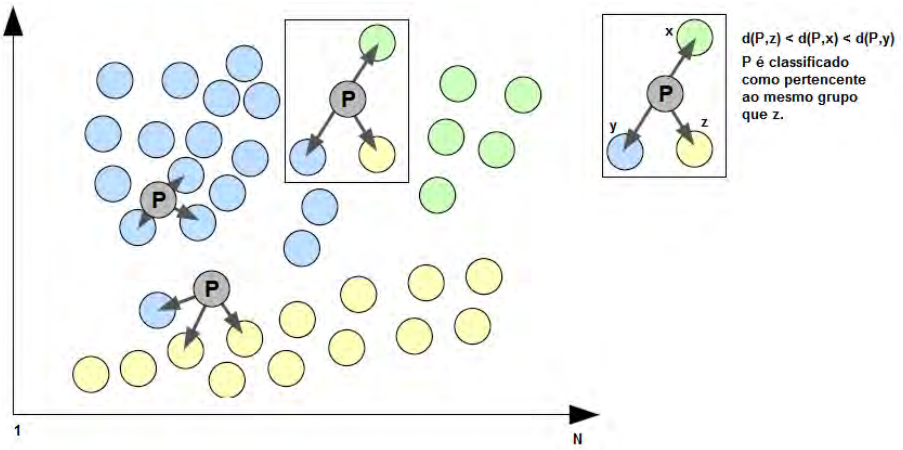

Figura 2.3 - Classificação de objetos utilizando o kNN com $k = 3$. Cada cor representa uma classe e os círculos marcados com um P representam objetos desconhecidos a serem rotulados.
Fonte: Adaptado de (SANTOS, 2010)

objeto desconhecido pertence.

Um ponto importante a ser considerado é o valor escolhido para k. Se o valor de k é pequeno demais, o resultado do algoritmo pode ficar sensível a *outliers*. Por outro lado, se o valor de k for grande demais então a vizinhança pode incluir muitos pontos pertencentes a classes distintas.

Considerando um conjunto de pontos N como sendo o conjunto de treinamento e um objeto de teste p. O algoritmos calcula a distância $d(N, p)$ entre todos os objetos que pertencem a N para determinar uma lista de k-vizinhos mais próximos.

Uma vez determinados os vizinhos mais próximos, o objeto de teste é classificado baseado na maioria das classes dos k-vizinhos mais próximos. A Figura 2.3 ilustra como os k-vizinhos mais próximos são encontrados em um ambiente com três classes e três objetos desconhecidos.

2.2.2 k-Médias

O algoritmo k-médias foi desenvolvido por McQueen (1967) e consiste em um processo para particionar uma população de objetos em k grupos. É fácil de ser implementado e computacionalmente econômico (STEINBACH; TAN, 2009).

Para McQueen (1967) umas das aplicações mais comuns do k-médias é a solução de problemas de agrupamento. Do ponto de vista dos inventores o algoritmo não consiste de uma técnica de agrupamento definitiva ou ótima, contudo é útil no auxílio da compreensão qualitativa e quantitativa de grandes volumes de dados. Por isso, o método deve ser usado em estreita interação entre teoria e intuição.

O algoritmo k-médias se auto-organiza para encontrar grupos. Considerando um conjunto de objetos $P = \{p^{(1)}, \ldots, p^{(N)}\}$ e K como sendo um vetor de médias e contendo a quantidade de grupos a serem formados, $K < N$ e $K = \{z^{(1)}, \ldots, z^{(K)}\}$, cada $z^{(K)}$ é um centro.

O procedimento inicia especificando-se o valor de K, então a cada objeto pertencente a P é atribuido um rótulo em relação a média mais próxima, ou seja, em relação ao centro $z^{(K)}$ mais próximo. A partir desse ponto têm-se os primeiros grupos formados, a quantidade de grupos é igual a quantidade de elementos em K.

Calculam-se novas médias, ou seja, novos centros para cada grupo formado, considerando as posições de cada elemento pertencente a cada grupo. O procedimento finaliza quando a variação das médias é pequena.

Para verificar se a variação das médias é pequena utiliza-se um critério de parada. Um valor que especifica a distância necessária entre as novas e anteriores médias para que o algoritmo pare. A Figura 2.4 mostra um exemplo da aplicação do algoritmo k-médias em um conjunto de dados, considerando $K = 2$.

Uma das formas de se estimar a eficiência do agrupamento é determinar o erro médio quadrático para o particionamento encontrado. Considerando o conjunto P particionado em K grupos $\{C_1, C_2 \ldots, C_K\}$, sendo que cada grupo C_K tem um determinado número p_k de objetos associados a ele, então

$$\sum_{k=1}^{K} p_k = P$$

O vetor médio $m^{(K)}$, ou centro, do grupo C_K é definido como o *centróide* do grupo

$$m^{(K)} = \frac{\sum_{i=1}^{p_k} {p_i}^k}{p_k} \tag{2.3}$$

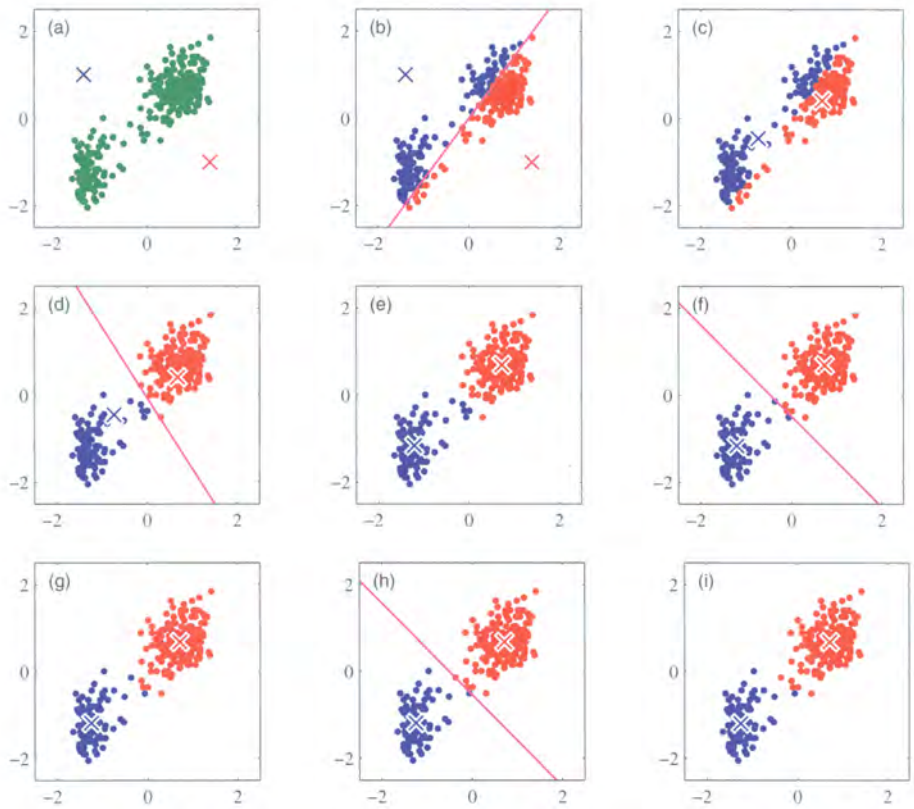

Figura 2.4 - Classificação de objetos utilizando o k-means com $K = 2$. Em (a) é mostrado um conjunto total de dados ainda não classificados, representados pela cor verde. O centros inicias são representados pelos símbolos em azul e vermelho. O algoritmo termina em (i), quando são formados dois grupos e percebe-se que de (f) para (i) as mudanças na posição dos centros são menos significativas, isso caracteriza o critério de parada.
Fonte: Adaptado de (ALPAYDIN, 2004)

onde ${p_i}^k$ é o i-ésimo objeto pertencente ao grupo C_K. O erro quadrático ${e_k}^2$ para o grupo C_K é a soma das distâncias entre cada objeto em C_K e o centro $m^{(K)}$

$$ {e_k}^2 = \sum_{i=1}^{p_k} ({p_i}^k - m^{(K)})^T ({p_i}^k - m^{(K)}) \tag{2.4} $$

$$ {E_K}^2 = \sum_{k=1}^{K} {e_k}^2 \tag{2.5} $$

o erro quadrático total ${E_K}^2$ para o agrupamento é a soma de todos os ${e_k}^2$.

2.2.3 HC - Hierarchical Clustering

Hierarchical Clustering consiste em uma sequência de grupos ou partições em que cada grupo está aninhado dentro de outro grupo mais próximo(BACKER, 1995) .

Também pode ser entendido como o processo de transformar a matriz de proximidade em uma sequência de partições aninhadas. O HC foi desenvolvido por King (1967) na década de 60 e popularizado por Johnson (1967), quem formalizou seu procedimento.

É um método de apredizagem não supervisionada e pode ser aplicado para mineração de dados, reconhecimento de padrões. Uma das vantagens de se usar o HC é que ele oferece uma excelente visualização dos grupos formados (KEOGH; PAZZANI, a).

Outra vantagem desse algoritmo é sua generalidade, dado que para o algoritmo não é necessário a entrada de parâmetros ou informar a quantidade de grupos a serem formados.

A Figura 2.5 mostra um dendograma, uma das formas mais usadas para visualizar o resultado de um agrupamento feito pelo HC. Usando-se um dendograma pode-se facilmente verificar a quantidade de grupos e sub-grupos formados e outras caracteristicas, como por exemplo, quais elementos foram unidos primeiro.

A semelhança entre dois objetos em um dendograma é representada como a altura do menor nó interno que eles partilham.

Existem três modos para se escolher a distância entre os objetos no algoritmo HC, o *single linkage*, o *complete linkage* e o *average linkage*.

No *single linkage*, inicialmente calcula-se a distância entre os objetos a serem agrupados e monta-se a matriz de proximidade. Baseado na matriz de proximidade entre os objetos o algoritmo começa unir objetos para formar novos grupos. Neste caso, os objetos escolhidos, por primeiro, para serem agrupados são os que apresentam a menor distância entre si.

Cada união forma um novo objeto, ou seja, um novo grupo a ser comparado com os demais. O procedimento finaliza quando todos os objetos estiverem agrupados.

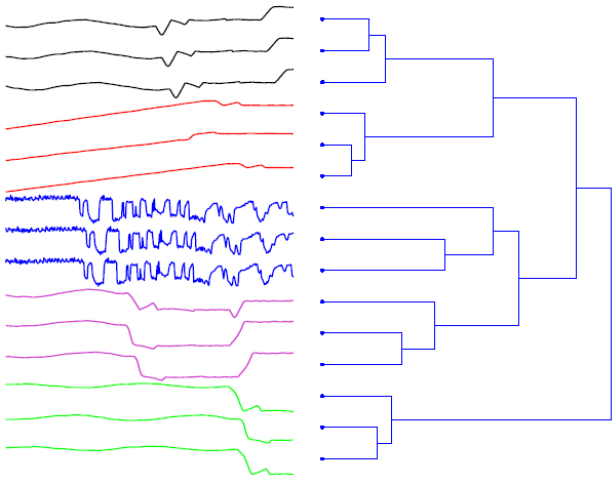

Figura 2.5 - Curvas agrupadas utilizando o algoritmo *Hierarchical Clustering* de forma aglomerativa.
Fonte: Adaptado de (LIN et al., 2004)

No *complete linkage* calcula-se a distância e monta-se a matriz de proximidade como no *single linkage*. Neste caso os objetos agrupados por primeiro são os que apresentam a maior distância entre si. Os novos grupos formados também serão agrupados de maneira recursiva.

No *average linkage* também é feito o mesmo procedimento para calculo de distâncias e montagem da matriz de proximidade. Neste caso os objetos escolhidos para serem agrupados são aqueles que apresentam a menor distância média. Este procedimento é feito até que todos os objetos estejam agrupados, assim como no *single linkage* e *complete linkage*.

2.2.4 SOM - Self-Organizing Maps

Também conhecido como Mapa de Kohonen, SOM consiste em uma rede neural para agrupar e classificar dados desenvolvida por Teuvo Kohonen na década de 80.

Em geral contém uma única camada de neurônios e uma camada de entrada ramificada e seu treinamento se dá de forma não-supervisionada (KOHONEN, 2001).

Para Kono (2007) Kohonen introduziu uma idéia de vizinhança topológica, sendo que a área da vizinhança e seus limites ou fronteiras são alterados dinamicamente por influência das modificações dos pesos dos nós vencedores.

De acordo com Beale e Jackson (1990) os estados internos da rede SOM, ou seja os pesos da rede, se modificam para modelar as características encontradas nos dados de treinamento.

De maneira geral, a implementação de um SOM é um arranjo de neurônios em duas dimensões (largura e altura) compondo um mapa (KONO, 2007). Os neurônios são arranjados ou organizados em um único plano.

Considerando uma configuração genérica para uma rede SOM, existem M neurônios em uma camada única e cada um tem um vetor de pesos $v^{(m)}$ com dimensão N, que é a mesma dimensão do vetor de entrada $x^{(q)} = ({x_1}^{(q)} + \ldots + {x_N}^{(q)})$, onde $q = 1, \ldots, Q$.

O objetivo da rede SOM é representar um padrão de entrada, com dimensão variada, em um mapa discreto unidimensional ou bidimensional de forma adaptável e com topologia ordenada (LOONEY, 1997).

A Figura 2.6 apresenta uma estrutura genérica para uma rede neural do tipo SOM. Percebe-se que cada valor do vetor de entrada $x^{(q)}$ está ligado a cada neurônio da rede e cada neurônio corresponde a uma saída.

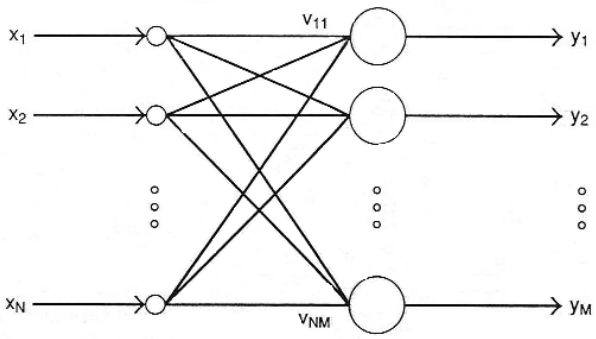

Figura 2.6 - Estrutura da Rede Neural SOM.
Fonte: Adaptado de (LOONEY, 1997)

A rede SOM busca por neurônios que se relacionam ao máximo com as entradas apresentadas, aumentando o grau de similaridade deles (KOHONEN, 2001).

Para esse processo existe uma fase de competição onde um exemplar dos vetores de entrada é selecionado e inserido na rede SOM. Então calcula-se as distâncias D_{qm}

entre $x^{(q)}$ e cada $v^{(m)}$, com $m = 1, \ldots, M$.

Uma das formas de se obter esta distância é utilizando a fórmula da distância euclidiana. Assim pode-se calcular D_{qm} como:

$$D_{qm} = D_{qm}(x^{(q)}, v^{(m)}) = \sum_{n=1}^{N} \left({x_n}^{(q)} - {v_n}^{(m)}\right)^2 \tag{2.6}$$

Após isso determina-se então a distância mínima D_{qm*} que corresponde ao neurônio m_*, que é o vencedor. Como m_* é o vencedor atualiza-se os pesos deste neurônio da seguinte forma:

$$v^{m*} = v^{m*} + \eta(x^{(q)} - v^{(m*)}) \tag{2.7}$$

Onde η é a taxa de aprendizagem. Os demais neurônios mantêm os valores dos seus pesos.

O resultado do algoritmo é uma organização de neurônios em vizinhanças próximas que funcionam como classificadores ou agrupadores dos dados de entrada gerando um mapa topográfico. A organização aleatória dos neurônios na grade apontam a existência de áreas representativas no mapa auto-organizado.

Uma outra forma para atualizar os pesos dos neurônios é utilizar uma estratégia conhecida por inibição lateral ou *lateral inhibition*. A estratégia de inibição lateral consiste em atulaizar positivamente todos os demais neurônios que estão próximos ao neurônio vencedor e atualizar negativamente todos os neurônios afastados.

Neste caso, uma vizinhança quadrada é formada em relação ao neurônio vencedor e todos os demais neurônios contidos nesta area também têm seus pesos atualizados. A Figura 2.7 mostra um exemplo do mapa e da vizinhança formados para o treinamento com inibição lateral.

No treinamento por inibição lateral, os ajustes dos pesos também são realizados de acordo com uma taxa de adaptação ou de aprendizagem. Essa taxa de adaptação, assim como o raio de vizinhança são reduzidos a cada ciclo, seguindo alguma função de decremento.

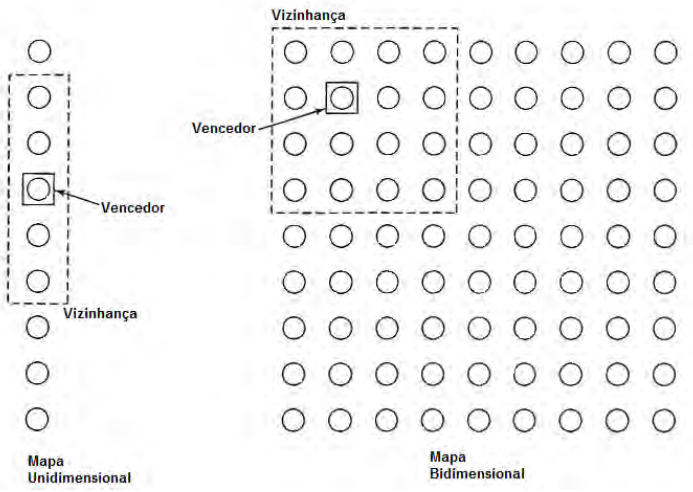

Figura 2.7 - Vizinhança.
Fonte: Adaptado de (LOONEY, 1997)

Decide-se o fim do treinamento quando algum parâmetro de finalização previamente estabelecido é alcançado. Este parâmetro pode ser, por exemplo, o número de ciclos de processamento, o tempo de execução, o tamanho do raio de vizinhança ou o valor da taxa de aprendizagem.

2.3 Séries Temporais

2.3.1 Definição

Séries temporais são usadas em muitas aplicações e em diversas atividades, como economia, previsão de vendas, processos de controle de qualidade, observações de fenômenos naturais, em engenharia e medicina.

Uma série temporal consiste de uma sequência de valores ou eventos obtidos em medições repetidas no tempo (HAN; KAMBER, 2006).

Para Kono (2007), uma série temporal é uma sequência de observações ao longo do tempo, que pode ser espaçada de maneira uniforme ou não. Comumente, os dados mensurados das observações são obtidos de forma contínua e rotineira à intervalos de tempo constantes.

Keogh e Pazzani (a) definem série temporal como porções significativas de dados

armazenados em uma base de dados, que podem ser geradas por aplicações em áreas variadas.

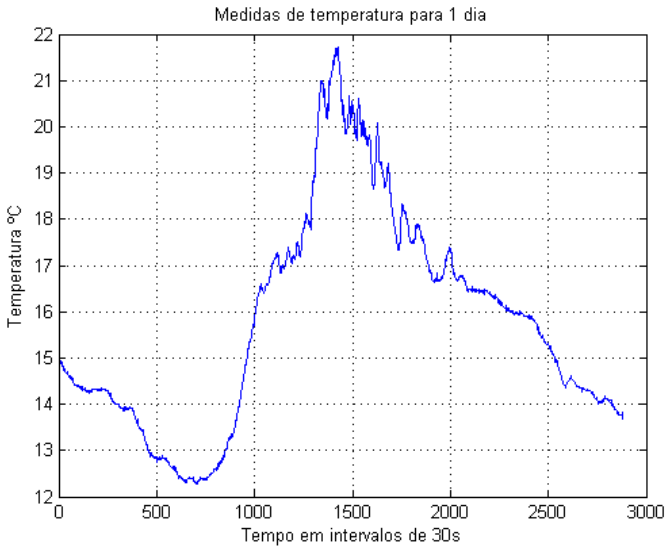

Figura 2.8 - Exemplo de uma série temporal de medidas de temperatura. Medidas feitas em intervalos de 30s durante 1 dia.

Pode-se aplicar testes estatísticos sobre séries temporais, como os que determinam coeficientes de auto-correlação, tendências de medidas lineares e outras métricas. A utilização desses testes varia de acordo com a capacidade e a necessidade de visualizar os padrões embutidos nesse tipo de dado.

Séries temporais podem ser diferenciadas entre contínuas e discretas. As séries consideradas contínuas possuem medidas realizadas ao longo do tempo de forma contínua, como por exemplo séries oriundas de um termógrafo ou um sismógrafo.

Séries consideradas discretas podem ter as medidas realizadas em tempos específicos e igualmente espaçados, como por exemplo séries de temperatura média, oriundas de sensores térmicos ou ainda séries de índices pluviométricos.

Um exemplo de série temporal é apresentado na Figura 2.8. No exemplo é apresentada uma série temporal de medições de temperatura coletadas a partir de um termistor. As medições foram realizadas em intervalos de 30s durante 1 dia.

2.3.2 Representação de Séries Temporais

2.3.2.1 DFT - Discrete Fourier Transform

No inicio do século IXX, Jean Baptiste Fourier desenvolveu um trabalho sobre propagação de calor em corpos sólidos no qual concluiu que uma função contínua pode ser expressada como uma soma de séries trigonométricas.

Após isso, vários outros métodos foram desenvolvidos para se usar a decomposição descoberta por Fourier. Um dos métodos usados para se trabalhar com séries discretas é a *Discrete Fourier Transform*.

A DFT de uma série $x(t)$, $t = 0, 1, \ldots, N-1$, é um vetor de números complexos $X(f)$:

$$X(\frac{f_k}{N}) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x(t) e^{\frac{-j2\pi kt}{N}} \tag{2.8}$$

Com $k = 0, 1, \ldots, N-1$ e $j \in$. A DFT representa uma série por meio de uma combinação linear de senoides complexas.

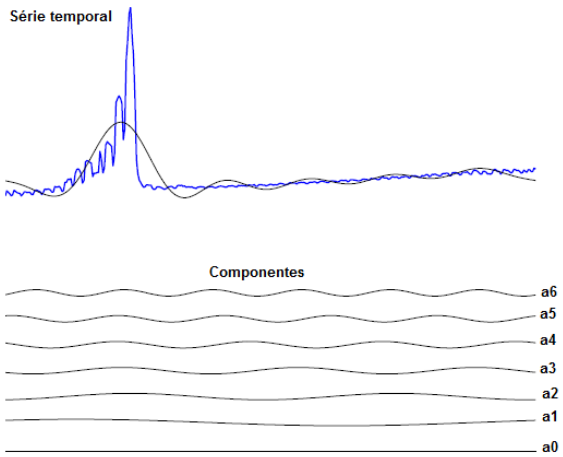

Figura 2.9 - Representação de uma série nas sete primeiras componentes de frequência. Fonte: Adaptado de (VLACHOS, 2004)

De posse dos coeficientes da decomposição, também chamados de componentes de frequência, pode-se recuperar a série original usando a DFT inversa:

$$x(t) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} X(\frac{f_k}{N}) e^{\frac{-j2\pi kt}{N}} \tag{2.9}$$

Muitas vezes, utilizando-se apenas poucos coeficientes é possível reconstruir a série original de maneira a não se perder informações importantes. Isso é um fator forte para se usar DFT na representação de séries.

Desconsiderando a maioria dos coeficientes pode-se essencialmente comprimir a série original, e assim reduzir sua dimensionalidade sem perdas significativas. A Figura 2.9 mostra uma série decomposta, utilizando-se a DFT, em sete componentes de frequência.

2.3.2.2 DWT - Discrete Wavelet Transform

Wavelets é uma outra técnica para representação de séries temporais. É usado com sucesso em várias áreas como, processamento de imagens (WINK; ROERDINK, 2004), predição de tráfego (SUN et al., 2004) e na biologia (RIVA et al., 2004).

Assim como a transformada de Fourier decompõe uma série em ondas senoidais, a DWT descreve uma série como uma versão deslocada em escalas de alguma *wavelet* mãe ψ.

O coeficiente c *wavelet* de uma série $x(t)$, em uma determinada posição k e escala j, pode ser descrito como:

$$c_{j,k} = \sum_j \sum_k x(t)\psi_{j,k}(t) \tag{2.10}$$

Onde $j = 0, 1, \ldots, N$, $k = 0, 1, \ldots, 2^{j-1}$. De maneira geral ψ pode ser descrita como:

$$\psi_{j,k} = \sqrt{2^j}\psi(2^j t - k) \tag{2.11}$$

A DWT é ortogonal e a série original pode ser completamente reconstruida dos

coeficientes $c_{j,k}$ por:

$$x(t) = \sum_j \sum_k c_{j,k} \psi_{j,k}(t) \tag{2.12}$$

A DWT também pode ser usada para redução de dimensionalidade. Isso é possível porque uma grande quandidade de coeficientes $c_{j,k}$ são zero, portanto podem ser descartados.

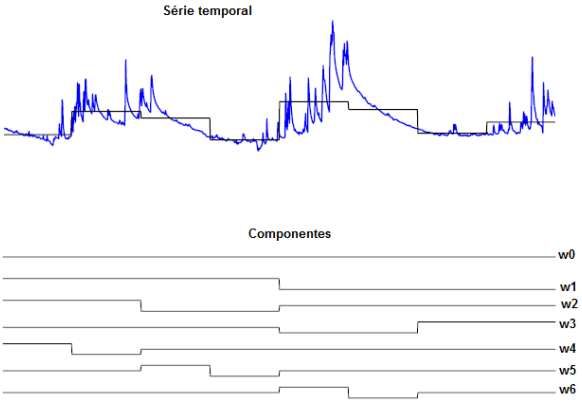

Figura 2.10 - Representação de uma série utilizando a função *Haar wavelet*.
Fonte: Adaptado de (VLACHOS, 2004)

Existem várias funções *wavelet* mãe, também chamadas de funções bases. Contudo uma das funções *wavelet* mais utilizadas em mineração de dados de séries temporais é a *Haar wavelet* (KEOGH; PAZZANI, a), veja Figura 2.10.

A *Haar wavelet* foi desenvolvida pelo matemático húngaro Alfréd Haar em 1909 e considerada a primeira função *wavelet* de base apresentada. A equação genérica para função *Haar wavelet* é defina como:

$$\psi_{j,k} = \begin{cases} 1 & \text{se} \quad t \in [0, \frac{1}{2}] \\ -1 & \text{se} \quad t \in [\frac{1}{2}, 1] \\ 0 & \text{caso contrário} \end{cases} \tag{2.13}$$

2.3.2.3 PAA - Piecewise Aggregate Approximation

A técnica *Piecewise Aggregate Approximation* é utilizada na representação de séries temporais para facilitar a classificação e agrupamento, identificação de correlações, outliers e a visualização das séries em diferentes escalas.

Desenvolvido por Keogh et al. (2000), o PAA é uma forma de representação de séries temporais recomendada quando se deseja diminuir a dimensionalidade de uma série extensa. É de fácil entendimento e implementação, também é aplicável a medidas de distância mais sofisticadas como DTW.

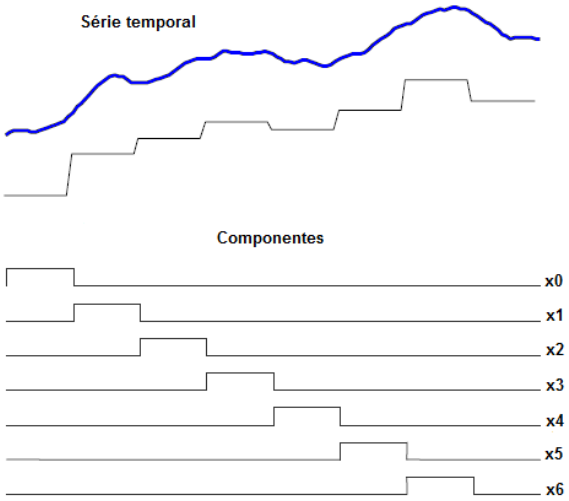

Figura 2.11 - Representação de uma série utilizando *Piecewise Aggregate Approximation*. Fonte: Adaptado de (KEOGH et al., 2000)

Considerando uma série temporal $x(t) = x_1, \ldots, x_n$ e seja N a dimensão da representação desejada para $x(t)$. A série temporal $x(t)$ pode ser representada no espaço N, por um vetor $\overline{x_i}$, calculado como:

$$\overline{x_i} = \frac{N}{n} \sum_{t=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_t \tag{2.14}$$

A série $x(t)$ é dividida em N janelas igualmente espaçadas. Para cada janela, é

calculado um valor médio e o vetor desses valores é a representação da série $x(t)$ em PPA.

O PAA pode ser entendido como uma média móvel uma vez que, se $N = n$, $\overline{x_i} = x(t)$ e se $N = 1$, $\overline{x_i}$ é a média da série original. A Figura 2.11 ilustra um exemplo de uma série representada pelo PAA, para efeito de comparação entre as representações por DFT e DWT utilizou-se uma figura semelhante as Figuras 2.9 e 2.10.

2.3.2.4 SAX - Symbolic Aggregate Approximation

Symbolic Aggregate Approximation é uma técnica desenvolvida por Lin et al. (2003) para representar uma série temporal de comprimento arbitrário n em uma cadeia de símbolos com comprimento w, sendo $w < n$.

De acordo com Lin et al. (2003) SAX é a primeira técnica desenvolvida para representação simbólica de séries temporais e para a maioria dos problemas de mineração de dados em séries temporais obtém resultados tão bons quanto representações como DFT e DWT.

Para se obter a representação em SAX, primeiramente transforma-se a série temporal utilizando a técnica PAA, com a intenção de reduzir dimensionalidade e numerosidade. É necessário que a série temporal seja normalizada pela média e desvio padrão, ficando assim com média igual a 0 e variância igual a 1.

Uma vez obtida a representação por PAA pode-se então aplicar outra transformação para representar a série em forma de uma cadeia de símbolos, sendo desejável que se tenha uma distribuição de símbolos com eqüiprobabilidade (APOSTOLICO et al., 2002).

De acordo com Larsen e Marx (1986), a maioria das séries temporais oriundas de fenômenos naturais ou não naturais possuem uma distribuição gaussiana. Isso é melhor evidenciado em Lin et al. (2003) onde se extraiu subsequências de oito diferentes séries temporais normalizadas com 128 pontos de comprimento. Ao se observar a distribuição acumulativa dos dados chegou-se a conclusão que esta era uma distribuição gaussiana.

Dado que a distruibuição de probabilidade de uma série temporal é uma gaussiana, pode-se então atribuir faixas de probabilidades, $B = \beta_1, \ldots, \beta_{\alpha-1}$, que irão produzir

β_i \ a	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Figura 2.12 - Valores de probabilidades para $\alpha = 3, \ldots, 10$.
Fonte: Adaptado de (LIN et al., 2003)

os símbolos, igualmente espaçados sobre a curva de uma distribuição gaussiana. Para isso usa-se uma tabela de probabilidades que contém as faixas que dividem uma distribuição gaussiana em um número de regiões eqüiprováveis, de acordo com o tamanho de α. A Figura 2.12 mostra a tabela que contém os valores para as faixas de probabilidade que serão usadas para distribuir os símbolos.

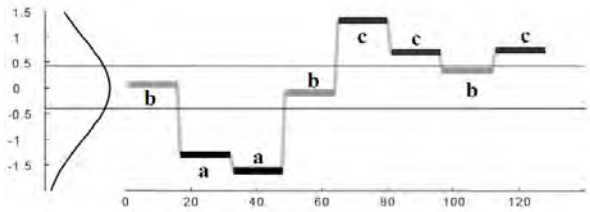

Figura 2.13 - Série temporal normalizada, primeiramente discretizada usando o PAA e então usando a tabela com faixas de probabilidades mapear os coeficientes da representação por PAA em símbolos. Para este caso com, $n = 128, w = 8$ e $\alpha = 3$. A série temporal é mapeada para a cadeia *baabccbc*
Fonte: Adaptado de (LIN et al., 2003)

Uma vez obtidos a tabela com as faixas de probabilidades e a representação da série normalizada em PAA, pode-se então mapear os valores reais da série temporal original em respectivos símbolos B de acordo com a tabela de símbolos e o alfabeto α. A Figura 2.13 ilustra o procedimento de como fazer o mapeamento de uma série temporal representada por PAA em símbolos.

3 ORIGEM DOS DADOS

Neste capítulo é abordada a origem dos dados utilizados. É apresentado um resumo do projeto de pesquisa Atlantic Rainforest Sensor Networking Project e dos equipamentos utilizados na coleta dos dados.

3.1 *Atlantic Rainforest Sensor Networking Project*

As florestas tropicais brasileiras são de imensa importância para o mundo. Eles ajudam a regular a temperatura do clima, absorvem parte da chuva, e são habitat para animais e plantas.

O projeto de pesquisa denominado *Atlantic Rainforest Sensor Networking Project*, implementado na costa sudeste do Brasil, está monitorando uma pequena parte da Mata Atlântica. Utilizando técnicas de micrometeorologia, fundamentais para a compreensão de como diferentes ecossistemas afetam o clima do planeta (CHANG, 2010).

A meta para os cientistas ambientais sobre o projeto foi para avançar significativamente na sua compreensão de como as funções da floresta se comportam, através da obtenção de dados mais detalhados sobre o que acontece dentro da atmosfera em uma escala pequena contida na copa da floresta.

Eram necessárias medidas reduzidas para eventos dentro de uma área pequena, como por exemplo a encosta de uma colina. A escala de tempo também teve de ser comprimida, com eventos que estão sendo medidos minuto a minuto.

O projeto é uma parceria entre a Universidade de São Paulo, Fundação de Amparo à Pesquisa de São Paulo (FAPESP), Instituto Nacional de Pesquisas Espaciais (INPE), *Johns Hopkins University* e do grupo *Microsoft Research* de pesquisa esterna.

3.2 Sensores e Dados Brutos

O desafio técnico foi a construção de um sistema de sensores que poderia sobreviver as condições apresentadas por uma floresta tropical, sem considerar possíveis interferências por causa da vida animal, e transmitir dados de forma confiável.

Cada sensor é protegido por uma embalagem de PVC em forma de copo, que é preenchido com epóxi. Pequenos orifícios nas extremidades das embalagens permitem

Figura 3.1 - Sensores utilizados pelo *Atlantic Rainforest Sensor Networking Project*. Fonte: Adaptado de (CHANG, 2010)

alta sensibilidade para as medições necessárias.

Os sensores são agrupados em quatro e cada conjunto de sensores é anexado a um equipamento que será um nó na rede de sensores. Cada nó sensor possui uma placa *wireless* e uma memória *flash*.

Os nós sensores coletam os dados a partir das medições feitas por cada sensor e transmitem através de uma conexão *wireless*. Também armazenam os dados na memória *flash*, uma precaução necessária para o caso da comunicação *wireless* falhar, assim cada nó mantém um conjunto completo dos dados coletados. A Figura3.1 mostra como é o sensor utilizado no projeto, percebe-se que o equipamento nó também possui uma embalagem especial para proteção.

A topologia da rede de sensores consiste de uma torre central com 60 metros de altura cercada por cinco torres menores. As torres são ligadas por cabos pendurados 2 metros acima da copa das árvores.

Nós sensores foram colocados nas torres e nos cabos que são irradiados pela torre central. Também existem outros cabos, ligando as torres, pendurados a cerca de 1 metro do solo, nesses cabos também foram colocados nós sensores. A Figura 3.2 ilustra como é a topologia da rede de sensores utilizada no projeto.

A equipe brasileira periodicamente entra na área com equipamentos, como com-

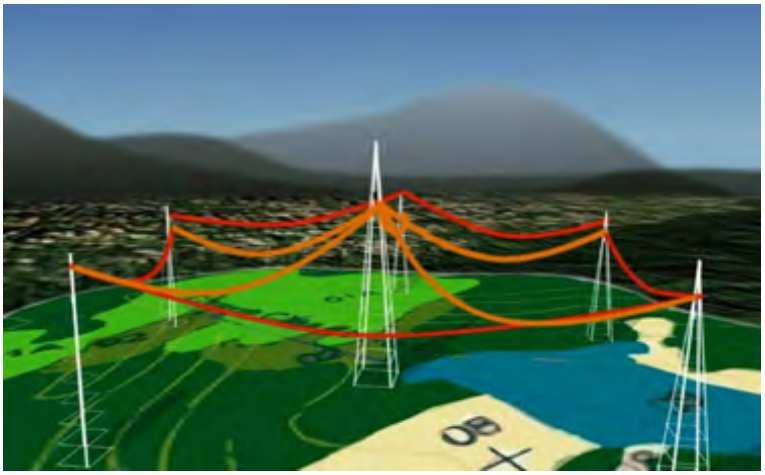

Figura 3.2 - Topologia da rede de sensores utilizada pelo *Atlantic Rainforest Sensor Networking Project*.
Fonte: Adaptado de (CHANG, 2010)

putadores portatéis, para fazer a coleta dos dados. Após o retorno da equipe para laboratório, os dados são enviados para a *Johns Hopkins University*, onde os arquivos de dados são processados para um formato adequado para análise e são enviados de volta. Os dados processados devolvidos são denominados de dados brutos.

```
17-11-2009 23:58:38,13.72732,13.71718,14.03399,114.7875
17-11-2009 23:59:08,13.72732,13.66496,13.82586,114.7875
17-11-2009 23:59:38,13.80557,13.69107,13.79978,114.7711
18-11-2009 0:00:08,13.64896,13.61268,13.82586,114.7703
18-11-2009 0:00:38,13.70121,13.69107,13.82586,114.7481
18-11-2009 0:01:08,13.67509,13.61268,13.82586,114.6749
18-11-2009 0:01:38,13.64896,13.61268,13.79978,114.7703
18-11-2009 0:02:08,13.64896,13.58652,13.79978,114.7703
18-11-2009 0:02:38,13.64896,13.56034,13.7476,114.4669
```

Figura 3.3 - Exemplo do arquivo ASCII de dados brutos. Cada linha contém um *time stemp*, 3 medidas de temperatura e uma medida de umidade relativa respectivamente.

Os dados processados pela *Johns Hopkins University* são devolvidos em formato *ASCII*. São arquivos de texto com extensão *.csv* contendo em cada linha um registro das temperaturas obtidas dos sensores.

Na Figura3.1 percebe-se que cada nó sensor possui acoplado a ele 4 sensores: 3 sensores são para medir temperatura em Celsius e 1 para medir a umidade relativa do ar.

Cada arquivo armazena em média 2.880 registros por dia. Além dos registros de temperatura e umidade relativa, cada nó sensor acrescenta uma medida de tempo em cada registro, um *timestemp*, que também fica registrado no arquivo de dados brutos. As medidas de tempo obtidas por cada nó sensor são feitas a cada 30 segundos.

Neste trabalho foram utilizadas séries temporais de 53 nós sensores. Isso totaliza a quantidade de 610.560 registros diários, somando a quantidade de medições feitas pelos sensores de temperatura e umidade relativa. A Figura 3.3 mostra um exemplo de como é o arquivo de dados brutos entregue após o processamento realizado pela *Johns Hopkins University*.

4 EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados os algoritmos utilizados para busca de grupos e erros nas séries temporais coletadas. São apresentados os principais resultados sobre os agrupamentos realizados e busca por falhas e ruídos que causam baixa qualidade nas medições e que podem revelar falhas nos sensores.

4.1 Agrupamentos

O agrupamento de objetos é uma das principais tarefas realizadas em mineração de dados. Através do agrupamento pode-se, de maneira rápida, encontrar e agrupar padrões e *outliers*.

Pode-se também evidenciar caracteristicas classificatórias, úteis se o objetivo da mineração for a classificação de objetos. Portanto, seguindo os objetivos propostos no início, definiu-se como primeira fase de experimentos o agrupamento das séries temporais dos sensores.

Não foram aplicados todos os algoritmos de agrupamentos descritos no capítulo 2 devido a semelhanças nos procedimentos dos algoritmos, alguns resultados poderiam ser iguais ou a visualização dos resultados seria inviável. Foram selecionados a rede SOM e o *Hierarchical Clustering* baseando-se em resultados apresentados em outros trabalhos, como por exemplos em Reis et al. (2007), Oliveira et al. (2009) e Keogh et al. (2002).

4.1.1 Agrupamento Utilizando a Rede SOM

Como descrito no capítulo 3, cada nó sensor possui quatro sensores acoplados: três para medir a temperatura e um para medir a umidade relativa do ar. O objetivo de se medir a temperatura com três sensores é ter uma garantia contra erros ou possíveis interferências, de animais ou outros eventos, que possam prejudicar a qualidade e qualidade das medições.

Com o propósito de validar o funcionamento dos sensores de temperatura e a veracidade das medições, utilizou-se uma rede neural SOM para agrupar as séries temporais de sensores pertencentes ao mesmo nó sensor.

A escolha da rede SOM adveio da necessidade de formar grupos de objetos que possuem múltiplas dimensões, como é o caso das séries temporais utilizadas neste

trabalho.

A rede SOM pode mapear as séries temporais dos sensores de temparatura em um mapa bi-dimensional. Desse modo, séries temporais próximas no espaço de atributos original tendem a permancer próximas no mapa formado pela rede SOM. Isso permitiu evidenciar, através de inspeção visual, a formação de grupos e falhas nas medições de temperatura.

Com o objetivo de evitar que uma dimensão se sobreponha em relação às outras e que o aprendizado fique estagnado, antes de iniciar o treinamento da rede SOM, normalizou-se os dados utilizados como entrada para o treinamento da rede.

Para isso utilizou-se a técnica de normalização conhecida como *Max-Min*, descrita na equação 4.1. Dessa forma os dados de entrada ficaram normalizados linearmente no intervalo $[0, 1]$.

$$x_n = \frac{x - min(x)}{max(x) - min(x)} \tag{4.1}$$

Foi selecionada uma parte dos dados para treinamento da rede SOM. A cada época de treinamento um neurônio da rede SOM é escolhido como vencedor. A escolha é feita baseada na menor distância euclidiana, descrita pela equação 2.6, entre os pesos do neurônio e os dados de entrada.

Alguns nós sensores apresentavam falhas em suas séreis temporais, como falta de medições. Portanto, somente nós sensores que apresentavam dados completos foram selecionados para esta fase de agrupamento com a rede SOM. O período escolhido para as medições foi de 3 dias.

O resultado obtido foi um mapa bi-dimensional, no qual percebeu-se a formação de grupos e o distanciamento de séries que deveriam estar próximas. A topologia da rede SOM é formada por 40 neurônios, o que gerou um mapa de 20 por 20 pontos. A Figura 4.1 mostra o mapa gerado pela rede SOM para o agrupamento das séries temporais de temperatura.

Percebe-se no agrupamento, apresentado pela Figura 4.1, que determinados sensores possuem séries muito distantes, mesmo estando geograficamente próximas, pois pertencem ao mesmo nó sensore. O nó sensor 22 é um dos que apresenta a maior dis-

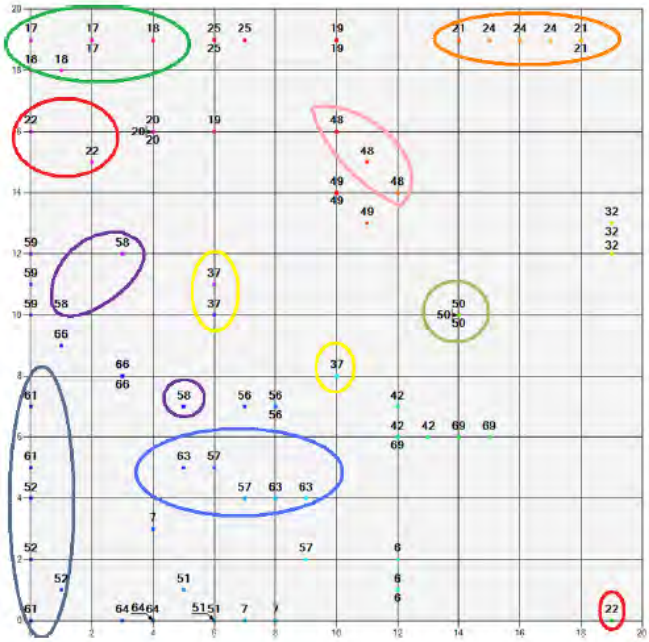

Figura 4.1 - Mapa gerado pela rede SOM para agrupar séries temporais de temperatura. Fonte: Adaptado de (SANTANA et al., 2011)

tância no mapa topológico. Isso evidencia que uma das duas séries de temperatura está muito distante das outras.

Também apresentam distâncias significativas os nós sensores 37 e 58. A Figura 4.2 mostra as séries temporais de temperatura obtidas pelo nó sensor 22, no intervalo das primeiras 24 horas. É evidente que uma das séries temporais está em discordância das demais.

4.1.2 Agrupamento Utilizando o Hierarchical Clustering (HC)

O algoritmo *Hierarchical Clustering* foi outra forma escolhida para agrupar as séries temporais. O objetivo dessa fase também foi gerar grupos e descobrir informações como as encontradas no agrupamento utilizando a rede SOM. Contudo este agrupamento foi realizado entre todos os nós sensores e não entre séries de apenas um nó sensor.

Considerando novamente as medidas de temperatura, como sabe-se que cada nó sensor tem 3 sensores de temperatura acoplados, precisou-se representar cada nó sensor por apenas uma única série temporal.

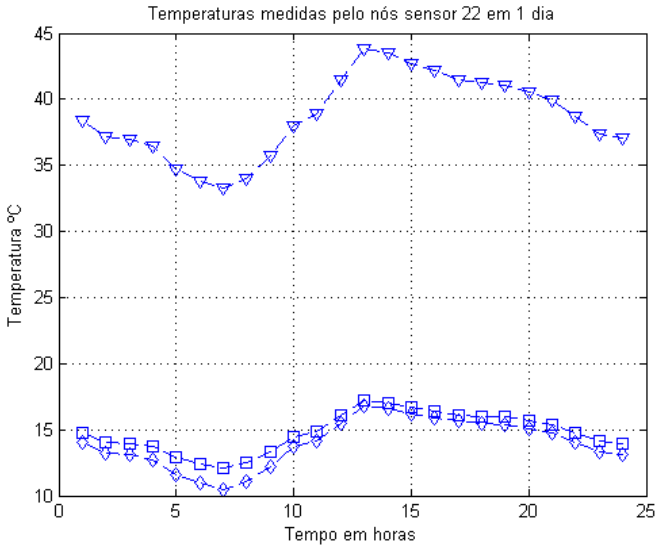

Figura 4.2 - Temperaturas medidas pelo nó sensor 22 em 24h.

Desse modo fez-se a média entre as séries temporais de temperatura dos 3 sensores de cada nó sensor. A partir disso, cada nó sensor passou a ser representado por apenas uma única série temporal de temperatura.

No capítulo 3 foi descrita a origem dos dados e explicado que cada nó sensor armazena, diáriamente, 2880 pontos para cada série temporal. Ao todo, cada série temporal tem em média 90500 colhidos ao longo de 21 dias. Portanto, a base de dados que contém as informações do nós sensores foi considerada extensa.

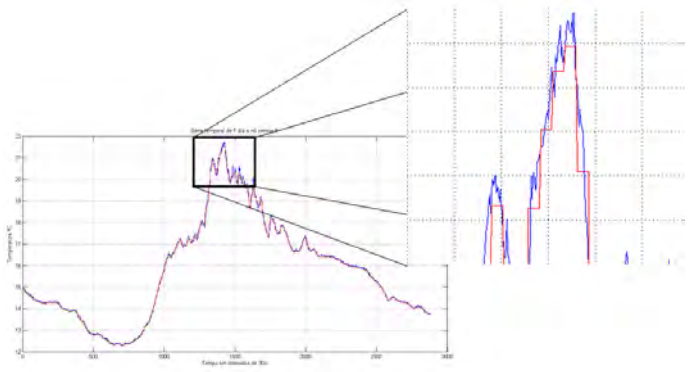

Figura 4.3 - Representação em *Piecewise Aggregate Approximation* para a temperatura média do nó sensor 6.

Muitos algoritmos de cálculo de distância, agrupamentos e classificação tornam-se lentos quando aplicados a base de dados extensas (KEOGH; PAZZANI, b). Por esse motivo é necessário usar de artifícios que possam reduzir a quantidade da base de dados.

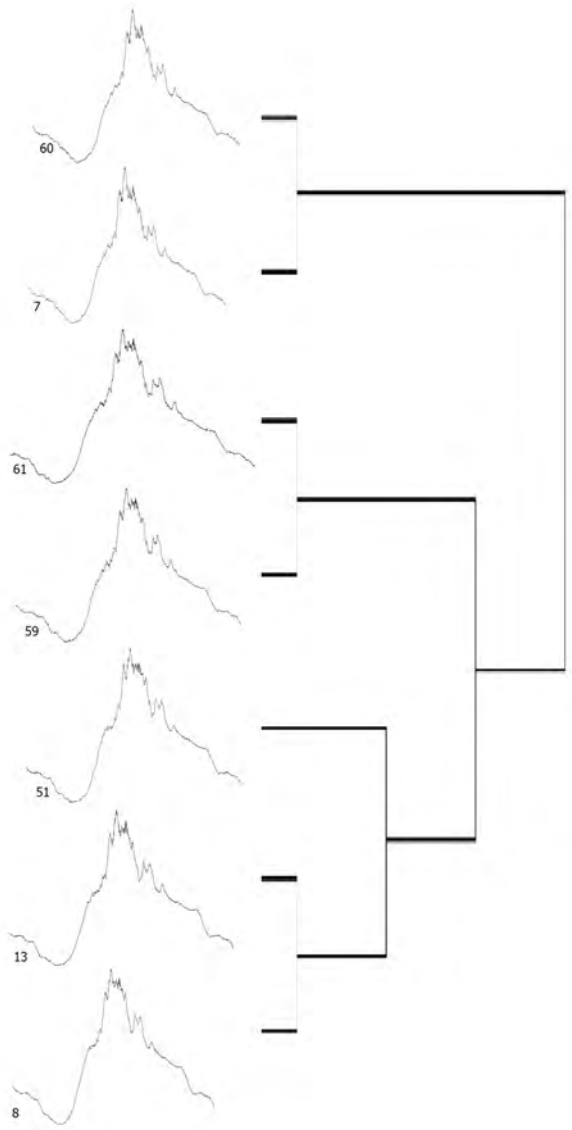

Figura 4.4 - Agrupamento utilizando o algoritmo *Hierarchical Clustering* os nós sensores 7, 8, 13, 51, 59, 60, 61.

De maneira geral, quando se pretende reduzir uma base de dados de séries temporais a redução ocorre na dimensionalidade das séries (KEOGH; PAZZANI, a). É desejável

uma redução sem prejudicar a qualidade, ou seja, sem perdas de informações significativas.

Utilizou-se a representação *Piecewise Aggregate Approximation* para reduzir a dimensionalidade das séries de temperatura, pertencentes a cada nó sensor. A Figura 4.3 mostra a série de temperatura do nó sensor 6. Considerou-se apenas as medições para um dia.

De acordo com Keogh et al. (2000), existe uma restrição no tamanho da dimensão da série representada pelo PAA. Para que o algoritmo funcione, o tamanho da representação em PAA deve ser divisor do tamanho da série original.

Portanto, para lidar com essa restrição, o tamanho da representação escolhida neste caso foi o máximo divisor comum de 2880, que é 1440. Isso representa uma redução pela metade na dimensionalidade das séries temporais.

O HC foi realizado na forma aglomerativa e utilizou-se a técnica *Dynamic Time Warping* como medida de distância, visto que o DTW apresenta maior precisão quando se mede dissimilaridade entre séries temporais.

A Figura 4.4 mostra um dendograma gerado após o agrupamento das séries de temperatura dos nós sensores 7, 8, 13, 51, 59, 60, 61. Inspecionando visualmente o dendograma da Figura 4.4 observa-se a formação dos grupos.

Entretanto, percebe-se que, visualmente, é difícil encontrar onde exatamente estão as diferenças, entre as séries, que permitiram o agrupamento. Por esse motivo a técnica DTW foi fundamental para este agrupamento.

Para verificar possíveis falhas, ou *outliers*, a partir do agrupamento feito pelo HC, montou-se um mapa com a topologia da rede de sensores e as posições dos nós sensores utilizados no agrupamento. Este mapa é apresentado pela Figura 4.5. Pôde-se ver que as séries dos nós sensores 8, 13 e 51 não estão próximos geograficamente como as demais séries agrupadas.

4.2 Detecção de Falhas

O processo de agrupamento foi importante no sentido de se encontrar grupos entre as séries temporais de sensores e nós sensores. Também pôde-se visualizar algumas das diferenças encontradas, como no caso do nó sensor 22 que apresenta uma das

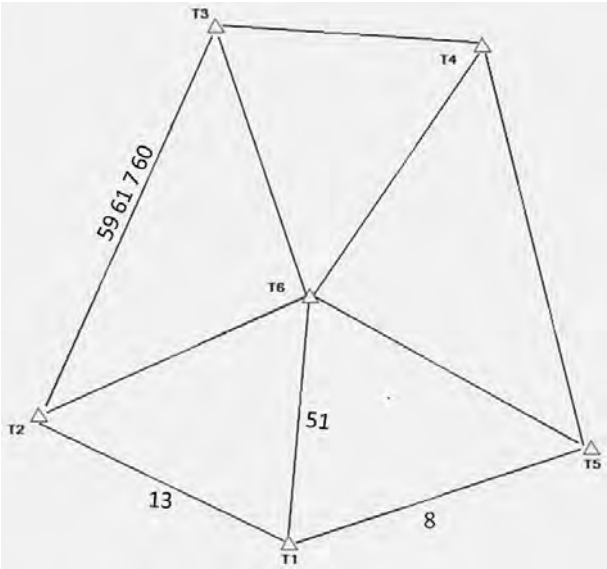

Figura 4.5 - Mapa topológico da rede de sensores.

séries de temperatura distante das demais.

O agrupamento utilizando a rede SOM conseguiu evidenciar que alguns nós sensores apresentavam falhas em suas medições. Contudo essas falhas nem sempre são tão evidentes.

Na Figura 4.1, o nó sensor 37 apresentou uma das séries de temperatura distante das outras. Assim como no nó sensor 22, isso significa que a série temporal distante tem medições diferentes das outras. Mas o que provocou essa distância, no mapa gerado pela rede SOM, entre as séries foi um pequeno ruído, uma falha inserida durante o processo de medição. Essa falha é, em muitos casos, sutil, portanto difícil de ser descoberta apenas de maneira visual.

A Figura4.6 mostra as 3 séries de temperatuas medidas pelo nó sensor 37 ao longo de 21 dias. A primeira vista, é difícil perceber diferenças irrefutáveis. Ao se ampliar a imagem fica claro que uma das medidas, a do sensor 3, está 1 °C acima das outras.

O objetivo da detecção de falhas é poder encontrar onde medidas diferentes das demais ocorrem. Para isso é necessário utilizar técnicas que permitam a busca de falhas ao longo de uma série temporal. Neste caso, utilizou-se a técnica de representação de séries temporais *Symbolic Aggregate Approximation*.

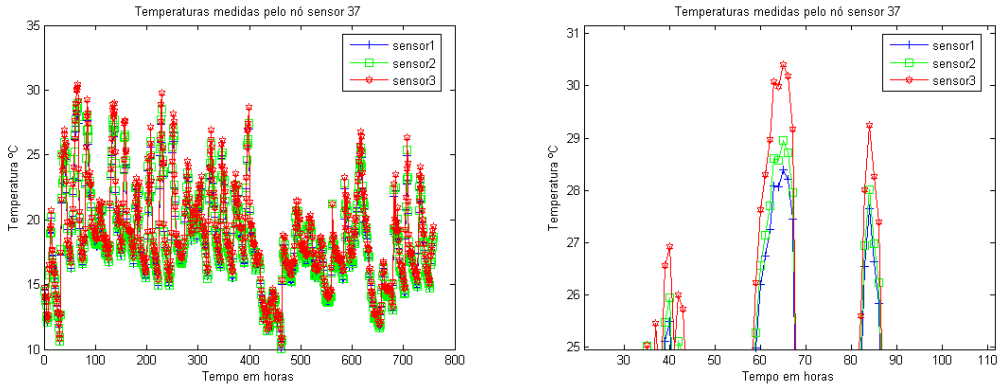

Figura 4.6 - Séries de temperatura medidas pelo nó sensor 37 em 21 dias.

O primeiro experimento foi realizado com as medições de temperatura do nó sensor 37. Considerou-se as medidas dos sensores 1 e 2 como medidas normais e a medida do sensor 3 como contendo anomalias.

Para transformar uma série temporal em sua representação por SAX, primeiramente é nescessário se fazer a representação em PAA (LIN et al., 2003). Antes de se aplicar o algoritmo PAA, normalizou-se as medições do nó sensor 37 utilizando a técnica conhecida como *Z-score*. Descrita na equação 4.2.

$$x_n = \frac{x - \mu_x}{\sigma_x} \tag{4.2}$$

A técnica *Z-score* normaliza os dados através da média e desvio padrão. Após a normalização as séries temporais ficaram com média 0 e desvio padrão 1.

O nó sensor 37 tem, ao todo, 90956 medições feitas ao longo de 21 dias. Obedecendo a restrição de que para o PAA funcionar é preciso que o tamanho da representação seja divisor do tamanho da série original, o tamanho escolhido para representação em PAA foi de 45478.

Após concluída a representação por PAA pôde-se então mapear os novos valores para símbolos. Utilizou-se uma distribuição gaussiana para fazer o mapeamento dos símbolos, como descrito na secção 2.3.2.4.

O tamanho do alfabeto foi escolhido arbitrariamente, uma vez que não se tem in-

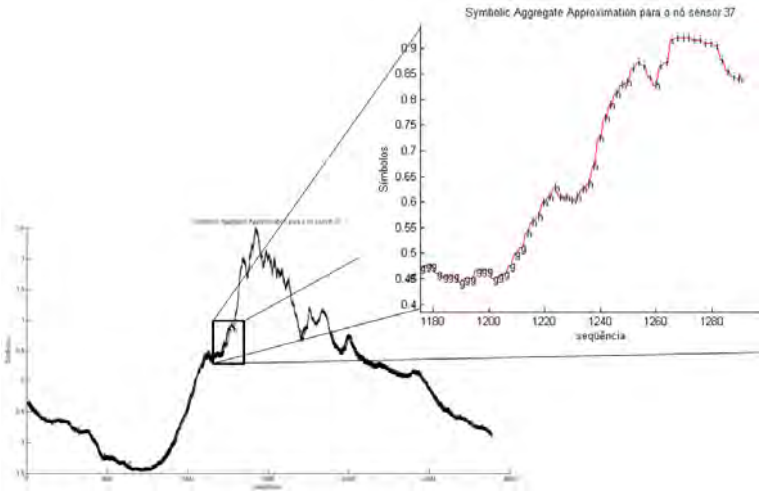

Figura 4.7 - Representação em SAX da série temporal do sensor 3 do nó sensor 37

dicação do melhor tamanho (LIN et al., 2003). Escolheu-se como símbolos as 10 primeiras letras do alfabeto português. A Figura 4.7 mostra um parte da representação em SAX da série temporal do sensor 3.

Após concluída a representação por SAX pôde-se então realizar a busca por falhas. O processo de busca consiste em comparar símbolo a símbolo e reportar o local e quantidade das diferenças encontratas. Onde houver diferenças, existem então falhas.

Como o processo de representação em SAX transforma uma série temporal em uma cadeia de símbolos, pode-se então usar algoritmos específicos para busca de símbolos em *strings* (GUSFIELD, 1997).

A maneira mais simples para se buscar uma falha é sobrepor as cadeias de símbolos e comparar cada par de símbolos, até que se esgote os pares símbolos a serem comparados. Esse método pode se tornar exaustivo uma vez que uma falha pode ser representada por uma seqüência.

Portanto, utilizou-se técnicas de busca de símbolos mais eficientes, capazes de diminuir a quantidade de tempo e número de comparações para se encontrar uma falha por completo.

Para a busca de falhas utilizou-se o algoritmo de Boyer-Moore. O algoritmo utiliza duas táticas denominadas de *Right-to-left scan* e *Bad character rule*, capazes de agilizar o processo de busca de símbolos (GUSFIELD, 1997) (HUME; SUNDAY, 1991).

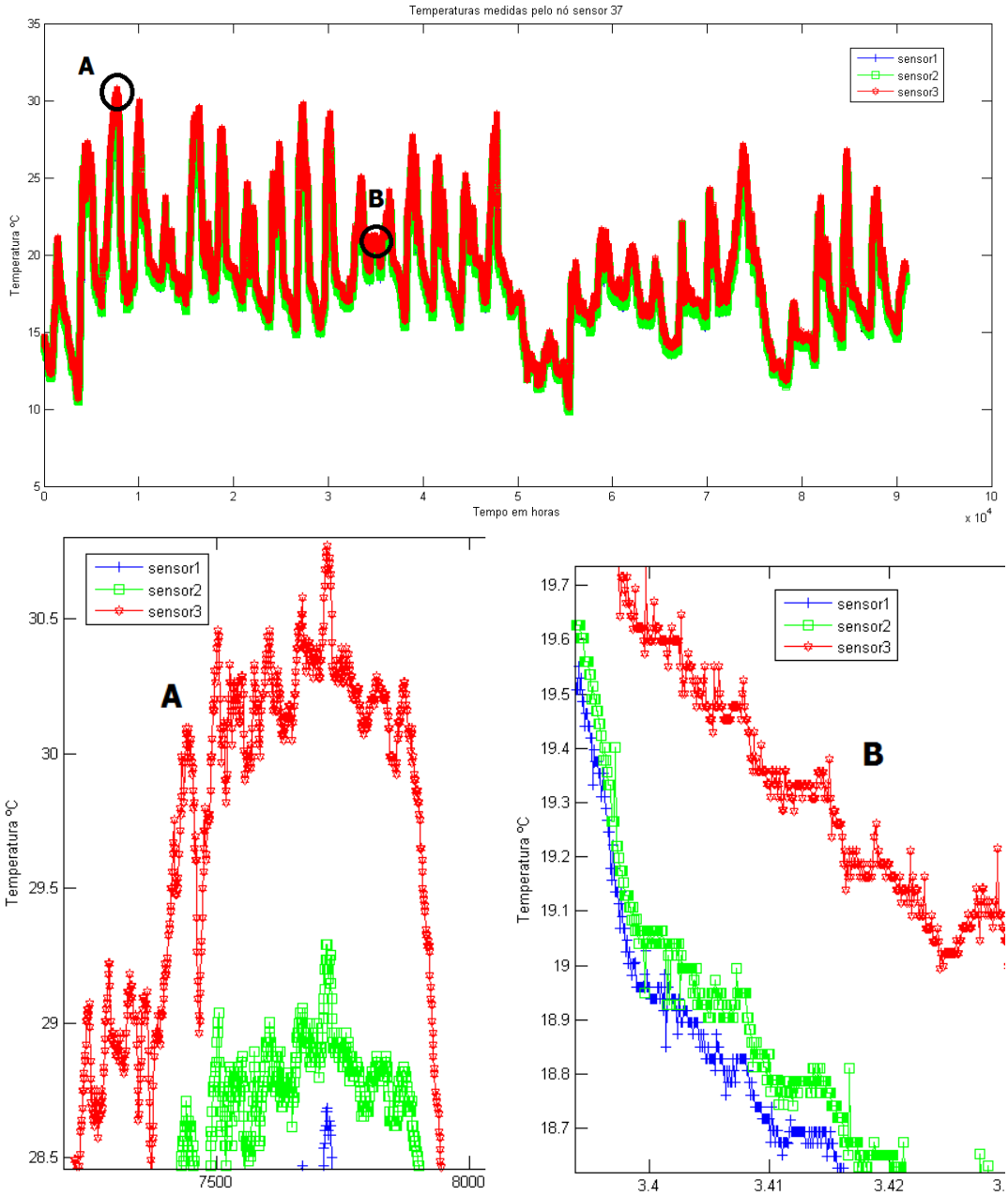

Figura 4.8 - Anomalias encontradas no sensor 3 de temperatura do nó sensor 37

Neste caso, para que a busca por falhas fosse mais precisa, as comparações foram realizadas em subcadeias que representavam as medições feitas em intervalos de 1 hora, ou seja, buscou-se falhas a cada 1 hora de medições.

Como resultado são apresentados na Figura 4.8, 2 falhas encontras durante o processo de busca. Percebe-se que as 2 falhas confirmam o fato de que em determinados

pontos a temperatura de um dos sensores está acima dos outros dois.

5 CONCLUSÕES

Este trabalho apresentou conceitos e características de algoritmos de mineração de dados, com o objetivo de encontrar informações e buscar agrupamentos e falhas em séries temporais. Em especial o trabalho realizou experimentos em séries temporais de fenômenos naturais coletadas por uma rede de sensores instalada na Mata Atlântica.

Os algoritmos de agrupamento utilizados se mostraram eficientes em relação a formar grupos de séries temporais. Foi possível verificar que determinadas séries foram selecionadas para grupos que estavam geograficamente distantes. Por este fato, conclui-se que as medições apresentavam falhas. Contudo, apenas com a aplicação dos algoritmos de agrupamento não foi possível identificar onde as falhas ocorriam.

Os algoritmos de representação de séries temporais também se mostraram eficientes. Com a aplicação do PAA a base de dados foi reduzida, o que gerou um ganho computacional na transformação das séries em cadeias de símbolos. A representação símbolica permitiu a busca por medições erradas de maneira minuciosa. Isso possibilitou encontrar em detalhes onde as falhas ocorriam e como elas afetavam a forma da série temporal. Pode-se verificar variações de até 1 °C para sensores muitos próximos.

Não é possível afirmar o que provocou as falhas. Elas podem ser geradas por influência de animais, falta de habilidade ao se instalar os sensores ou ainda por defeitos inseridos durante a fabricação ou calibração dos mesmos. O importante é que uma vez revelada a existência de falhas, fica a critério dos pesquisadores a execução de providencias para corrigir o problema e garantir os resultados de suas pesquisas. Por fim, espera-se que as metodologias e os resultados apresentados neste trabalho ajude demais pesquisadores a validar os dados de suas pesquisas para que estes não percam tempo buscando informação em fontes corrompidas ou acabem publicando informações equivocadas.

5.1 Contribuições

Foram publicados dois artigos no XV Simpósio Brasileiro de Sensoriamento Remoto realizado em Curitiba-PR no período de 30 de Abril a 5 de Maio de 2011, sendo eles:

- SANTANA, A. F.; SANTOS, R.; NETO, G. P.; VIJAYKUMAR, N. L.

Aplicação de mapas auto-organizáveis para análise de dados de geosensores: Estudo de caso na mata atlântica. XV SBSR 2011.

- NETO, G. P.; SANTANA, A. F.; VIJAYKUMAR, N. L. SANTOS, R. Mineração e Visualização Remota de Dados Coletados de uma Rede de Sensores Ambientais. XV SBSR 2011

O trabalho realizado também é parte de um projeto de pesquisa denominado: *Desenvolvimento de Aplicação de Rede de Geossensores para Monitoramento Ambiental*. O projeto é financidao pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, uma das principais agências de fomento à pesquisa científica e tecnológica do país, e pela Microsof Research, divisão que se dedica a realização de pesquisa básica e aplicada em ciência da computação e engenharia de software.

O projeto propõe atividades de pesquisa científica e tecnológica cooperativa em resposta à chamada de propostas do Instituto Virtual de Pesquisas FAPESP Microsoft Research. Os principais responsáveis pelo projeto são: Dr. Celso von Randow, Dr. Rafael Duarte Coelho dos Santos e Dra. Maria do Carmo de Andrade Nono.

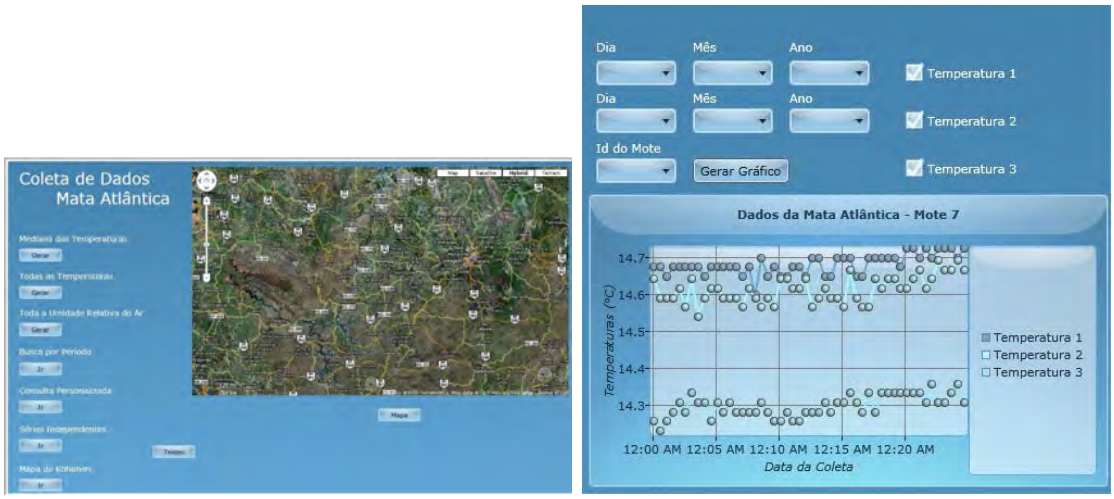

Figura 5.1 - Interface do sistema de processamento e geração de gráficos personalizados desenvolvidos no projeto *Desenvolvimento de Aplicação de Rede de Geossensores para Monitoramento Ambiental*.
Fonte: Adaptado de (NETO et al., 2011)

Um dos objetivos do projeto é o desenvolvimento de ferramentas de software para gerenciamento e integração de dados coletados a partir de geossensores. Os dados coletados pela rede de geossensores devem ser centralizados, processados, armazenados e distribuídos, de maneira que também possam ser usados como fonte de dados para outras aplicações.

Portanto outra contribuição do trabalho foi o auxílio no desenvolvimento de um sistema web para gerenciamento destes dados. As principais funcionalidades do sistema são:

- Entrada de dados obtidos dos sensores, armazenamento no banco de dados e criação dos metadados sobre a entrada.
- Relatórios simples, visualização simples dos dados coletados.
- Web services para acesso ao banco de dados por programas externos ao sistema.

Os resultados e ferramentas desenvolvidas no projeto foram apresentados no XI Workshop de Computação Aplicada, realizado no INPE de São José dos Campos, no período de 08 a 10 de Novembro de 2011 e no XV Simpósio Brasileiro de Sensoriamento Remoto. A Figura 5.1 mostra duas das interfaces desenvolvidas, uma para o sistema de processamento de dados e outra para o sistema de geração de gráficos.

5.2 Trabalhos Futuros

Como trabalhos futuros pretende-se melhorar os algoritmos apresentados para que os mesmo fiquem mais rápidos e possam extrair mais informações sobre séries temporais. Para que esse objetivo seja alcançado será necessário uma maior aproximação dos pesquisadores da área de computação e meteorologia. A troca de idéias, experiências e opiniões deverá trazer grandes benefícios as futuras pesquisas.

Outro ponto a ser estudado é a integração de novos algoritmos à pesquisa em séries temporais. Considerando o algoritmo SAX e que ele transforma uma série temporal em uma sequência de símbolos pode-se então utilizar outras técnicas em conjunto para busca de padrões e falhas. Como exemplo são as técnicas já utilizadas na área da biologia para comparação de DNA e outras características hereditárias.

REFERÊNCIAS BIBLIOGRÁFICAS

AACH, J.; CHURCH, G. Aligning gene expression time series with time warping algorithms. **Bioinformatics**, v. 17, p. 495–508, 2001. 7

ALPAYDIN, E. **Introduction to machine learning**. 1. ed. [S.l.]: MIT Press, 2004. 11

ANDERBERG, M. **Cluster analysis for applications**. New York: Academic Press, 1973. 6

APOSTOLICO, A.; BOCK, M. E.; LONARDI, S. Monotony of surprise and large-scale quest for unusual words. In: INTERNATIONAL CONFERENCE ON RESEARCH IN COMPUTATIONAL MOLECULAR BIOLOGY, 6., 2002. **Proceedings...** [S.l.], 2002. p. 22–31. 22

BACKER, E. **Computer-assisted reasoning in cluster analysis**. 1. ed. Hemel Hempstead: Prentice Hall, 1995. 5, 12

BEALE, R.; JACKSON, T. **Neural computing**. Bristol: IOP Publishing Ltd, 1990. 14

BERNDT, D.; CLIFFORD, J. An algorithm for the machine calculation of complex fourier series. **AAAI Workshop on Knowledge Discovery in DataBases**, p. 229–248, 1994. 6

CAIANI, E.; PORTA, A.; BASELLI, G.; TURIEL, M.; MUZZUPAPPA, S.; PIERUZZI, F.; CREMA, C.; MALLIANI, A.; CERUTTI, S. Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. **IEEE Computers in Cardiology**, p. 73–76, 1998. 7

CHANG, J. **Atlantic Rainforest Sensor Networking Project**. 2010. Disponível em: <http://research.microsoft.com/apps/video/default.aspx?id=138806>. Acesso em: 30 Maio 2011. 25, 26, 27

FIX, E.; HODGES, J. Discriminatory analysis, nonparametric discrimination. **USAF School of Aviation Medicine**, Randolph Field, 1951. 8

GAUCH, H. G. **Multivariate analysis in community ecology**. Cambridge: Cambridge University Press, 1982. 5

GUSFIELD, D. **Algorithms on Strings, Trees and Sequences**: Computer science and computational biology. Cambridge: Cambridge University Press, 1997. 37

HAN, J.; KAMBER, M. **Data Mining**: Concepts and techniques. 2. ed. [S.l.]: Elsevier, 2006. 16

HUME, A.; SUNDAY, D. Fast string searching. In: SOFTWARE - PRACTICE AND EXPERIENCE, 1991. **Proceedings...** [S.l.], 1991. p. 1221–1248. 37

JOHNSON, L. Hierarchical clustering schemes. **Psychometrika**, v. 32, p. 241–254, 1967. 12

KEOGH, E.; CHAKRABARTI, K.; PAZZANI, M.; MEHROTRA, S. Dimensionality reduction for fast similarity search in large time series databases. **Knowledge and Information Systems**, v. 3, n. 3, p. 263–286, 2000. 3, 21, 34

KEOGH, E.; LONARDI, S.; CHIU, W. Finding surprising patterns in a time series database in linear time and space. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 8., 2000. **Proceedings...** [S.l.], 2002. 3, 29

KEOGH, E.; PAZZANI, M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-98), 4., 1998. **Proceedings...** New York, USA: AAAI Press. p. 239–241. ISBN 1-57735-070-7. 3, 12, 16, 20, 33

_____. Scaling up dynamic time warping to massive datasets. Prague, Czech Republic., p. 1–11. 6, 33

KING, B. Step-wise clustering procedures. **Journal of the American Statistical Association**, v. 69, p. 86–101, 1967. 12

KOHONEN, T. **Self-organizing maps**. 3. ed. Berlin: Springer, 2001. 13, 14

KONO, Y. **Utilização de Rede Neural LVQ para previsão do nivel do rio Paraguai**. 2007. 152 p. (INPE-15191-TDI/1300). Dissertação de Mestrado do Curso de Pos-Graduação em Computação Aplicada — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2007. 13, 14, 16

KRUSKAL, J.; LIBERMAN, M. **The Symmetric Time-Warping Problem: From Continuous to Discrete**: Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. Massachusetts: Addison-Wesley Publishing Co., 1983. 125-161 p. 6

LARSEN, R. J.; MARX, M. L. **An Introduction to Mathematical Statistics and Its Applications**. 2. ed. [S.l.]: Prentice Hall, 1986. 22

LIN, J.; KEOGH, E.; LONARDI, S.; CHIU, B. A symbolic representation of time series, with implications for streaming algorithms. In: ACM SIGMOD WORKSHOP ON RESEARCH ISSUES IN DATA MINING AND KNOWLEDGE DISCOVERY, 8., 2003. **Proceedings...** [S.l.], 2003. 4, 22, 23, 36, 37

LIN, J.; VLACHOS, M.; KEOGH, E.; GUNOPULOS, D. Iterative incremental clustering of time series. **IX Conference on Extending Database Technology**, 2004. 13

LOONEY, C. **Pattern Recognition Using Neural Networks**: Theory and algorithms for engineers and scientists. 1. ed. New York: Oxford University Press, 1997. 14, 16

MAINWARING, A.; POLASTRE, J.; SZEWCZYK, R.; CULLER, D.; ANDERSON, J. Wireless sensor networks for habitat monitoring. In: **ACM International Workshop on Wireless Sensor Networks and Applications**. [S.l.: s.n.], 2002. p. 8897. 1, 4

MCQUEEN, J. Some methods for classification and analysis of multivariate observations. **V Berkeley Symposium on Mathematical Statistics and Probability**, Randolph Field, p. 281–297, 1967. 9, 10

NETO, G. P.; SANTANA, A. F.; VIJAYKUMAR, N. L.; SANTOS, R. Mineração e visualização remota de dados coletados de uma rede de sensores ambientais. In: EPIPHANIO, J. C. N.; GALVÃO, L. S. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2011. p. 5479–5485. ISBN 978-85-17-00056-0 (Internet) and 978-85-17-00057-7 (DVD). Disponível em: <http://urlib.net/dpi.inpe.br/marte/2011/07.15.12.03>. Acesso em: 28 mar. 2012. 42

OLIVEIRA, D. N.; SOUZA, B. F. S.; SILVA, F. A. T. F.; TEIXEIRA, A. S. Aplicaçâo de mapas de kohonen em imagem de satélite do semi-árido e

comparaçâo com o método da máxima verossimilhança. In: XIV SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO - SBSR, 14., 2009. **Anais...** [S.l.], 2009. p. 7877–7884. 29

REIS, I. A. Alternatives for geosensors networks data analysis. In: **VII Simpósio Brasileiro de Geoinformática**. [S.l.: s.n.], 2005. p. 94–104. 4

REIS, I. A.; CÂMARA, G.; ASSUNçãO, R.; MONTEIRO, A. M. V. Data-aware clustering for geosensor networks data collection. In: XIII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO - SBSR, 13., 2007. **Anais...** [S.l.], 2007. p. 6059–6066. 29

RIVA, L.; MENOZZI, G.; SIRONI, M.; CERUTTI, S.; POZZOLI, U. A wavelet based method to predict nucleosome positions. **Bioinformatics Italian Society Meeting**, 2004. 19

SANTANA, A. F.; SANTOS, R.; NETO, G. P.; VIJAYKUMAR, N. L. Aplicação de mapas auto-organizáveis para análise de dados de geosensores: Estudo de caso na mata atlântica. In: XV SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO - SBSR, 15., 2011. **Anais...** [S.l.], 2011. p. 8931–8937. 31

SANTOS, R. **Principios e Aplicações de Mineração de Dados**. São José dos Campos: INPE, 2010. Disponível em: <http://www.lac.inpe.br/~rafael.santos/cap359.jsp>. Acesso em: 3 Julho 2010. 9

STEINBACH, M.; TAN, P. **The Top Ten Algorithms in Data Mining**. 1. ed. University of Vermont: CRC Press, 2009. 8, 9

SUN, H.; YANG, F.; RAN, B. Wavelet preprocessing for local linear traffic prediction. In: TRANSPORTATION RESEARCH BOARD ANNUAL MEETING PREPRINT CD-ROM, 83., 2004. **Proceedings...** [S.l.], 2004. 19

VLACHOS, M. **Similarity and Indexing in Multidimensional Spaces**. 2004. 152 p. /textitA Dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science — University of California Riverside, Riverside, 2004. 18, 20

WINK, A.; ROERDINK, J. Denoising functional mr images: a comparison of wavelet denoising and gaussian smoothing. **IEEE Transactions on Medical Imaging**, v. 23, p. 374–387, 2004. 19

XI, X.; KEOGH, E.; SHELTON, C.; WEI, L.; RATANAMAHATANA, C. A. Fast time series classification using numerosity reduction. **23nd International Conference on Machine Learning**, Pittsburg, 2006. 8

XU, N. A survey of sensor network applications. **IEEE Communications Magazine**, v. 40, p. 102114, 2002. 1, 4

ZHU, Y.; SHASHA, D. Warping indexes with envelope transforms for query by humming. **SIGMOD**, p. 181–192, 2003. 7

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.