Business Intelligence

Carlos Alberto Poncinelli Filho

Previsão de Microclima de Municípios do Rio de Janeiro Utilizando Técnicas de Mineração de Dados Apoiado em Redes Neurais Artificiais

Monografia de Final de Curso

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence.

Orientador:

Prof. Dra. Karla Figueiredo

Agradeço a Deus pelo dom da vida e Sua semente em minha alma.
À minha esposa e filha, por compreender a ausência durante a realização do curso.

Agradeço a oportunidade que a Petrobras, através da gerência setorial TIC/IDTA/AIM, pelo patrocínio e apoio de forma que pudesse realizar esse curso. Espero que, com os ensinamentos recebidos, poder contribuir profissionalmente para a Companhia.

Agradeço à doutora Luciana Campos pelo código de redes neurais que implementou os algoritmos fundamentais para uso no MatLab e execução dos experimentos.

RESUMO

Esta monografia trata de um modelo de previsão de sensação térmica a partir de técnicas de inteligência computacional utilizando o método de Redes Neurais Artificiais. A partir de uma base original de 30 municípios do estado do Rio de Janeiro foram feitos agrupamentos de acordo com suas similaridades com uso de análise de grupos (*cluster*) e depois sobre esses grupos foram aplicados vários modelos de redes neurais artificiais para verificar qual das redes estava mais adequada a elaborar as previsões de um passo à frente. O trabalho também faz uma análise dos resultados obtidos e sugere investigações para futuros trabalhos correlacionadas. Os anexos da monografia contêm os scripts e detalhes que esclarecem as aplicações utilizadas.

ABSTRACT

This monograph is a model for predicting thermal sensation from computational intelligence techniques using the method of Artificial Neural Networks. From an original base of 30 municipalities in the state of Rio de Janeiro groupings were made according to their similarities with the use of cluster analysis and then on these groups have applied several models of artificial neural networks to determine which network was more appropriate to develop the forecasts one step ahead. In this paper, also analyzes the results and suggests future work related to investigations. The annexes of the monograph contain scripts and details that clarify the applications used.

Sumário

1 INTRODUÇÃO 11

1.1 JUSTIFICATIVAS 11
1.2 OBJETIVO 11
1.3 METODOLOGIA 12

2 MINERAÇÃO DE DADOS – UM BREVE CONTEXTO 13

2.1 O LEGADO COMO FONTE 13
2.2 AVALANCHE DE INFORMAÇÕES 14
2.3 COMBINAÇÃO INTERDISCIPLINAR 14

3 PROPECÇÃO DE CONHECIMENTO E DATA MINING 16

3.1 ETAPAS DO PROCESSO DE KDD 16
3.1.1 Seleção dos Dados 17
3.1.2.Limpeza e Enriquecimento 17
3.1.3 Transformação 18
3.1.4 Data Mining 18
3.1.5 Interpretação 18

4 ANÁLISE DE CLUSTER 19

4.1 DESCREVENDO ANÁLISE DE CLUSTER 19
4.2 MODELO APLICADO NA SOLUÇÃO DO PROBLEMA 20

5 REDES NEURAIS ARTIFICIAIS 21

5.1 DESCREVENDO REDES NEURAIS ARTIFICIAIS 21
5.2 ARQUITETURA DA REDE 22
5.3 APRENDIZADO E VALIDAÇÃO 23
5.4 FUNÇÃO DE ATIVAÇÃO 25
5.5 TAXA DE APRENDIZAGEM 26
5.6 REDE NEURAL APLICADO NA SOLUÇÃO DO PROBLEMA 26

6 APLICANDO TÉCNICAS DE DATA MINING PARA PREVISÃO DE MICRO-CLIMA 28

6.1 LEVANTAMENTO DE DADOS 28
6.2 DEFINIÇÃO DO MÉTODO PARA COLETA DE DADOS 30
6.2.1 Análise dos Dados Coletados 30
6.3 APLICAÇÃO DA ANALISE DE CLUSTER NO AGRUPAMENTO DE MICROCLIMA 33
6.4 MODELAGEM POR REDE NEURAL PARA A PREVISÃO DE SANSAÇÃO TÉRMICA DO MICROCLIMA 37
6.5 ANÁLISE DOS RESULTADOS OBTIDO 38

7 CONCLUSÕES 42

8 SUGESTÃO DE TRABALHOS FUTUROS 43

9 REFERÊNCIAS BIBLIOGRÁFICAS 44

Anexo I
Anexo II

Lista de Tabelas

TABELA 1 – CATEGORIZAÇÃO DAS VARIÁVEIS.....29
TABELA 2 – MUNICÍPIOS DO ESTADO DO RIO DE JANEIRO PARTICIPANTES DA ANÁLISE 31
TABELA 3 – CAMPOS DE CONFIGURAÇÃO DO WEKA PARA EFETUAR A EXECUÇÃO DO CLUSTER 35
TABELA 4 (A) A (G) – CLUSTER PROCESSADOS PELO ALGORITMO DE REDE NEURAL EXECUTADO PELO MATLAB.....40
TABELA 5 – RELAÇÃO DAS MELHORES REDES BASEADA NOS RESULTADOS DE MAPE_VALID ...41

Lista de Figuras

FIGURA 1 – ETAPAS DO PROCESSO DE KDD. 16
FIGURA 2 – DISTÂNCIA EUCLIDIANA ENTRE OBJETOS I E J COM P=2 VARIÁVEIS....19
FIGURA 3– COMPONENTES PARA UM RN....22
FIGURA 4 – MPL TÍPICA COM UMA CAMADA OCULTA 23
FIGURA 5 – REPRESENTAÇÃO DAS FUNÇÕES TANGENTE HIPERBÓLICA E LOGÍSTICA....26
FIGURA 6 – CABEÇALHO DA BASE DE DADOS INICIAL. 30
FIGURA 7 – SENSAÇÃO TÉRMICA MÉDIA PARA OS MESES DO ANO NOS TRINTA MUNICÍPIOS DO RIO DE JANEIRO DE 2000 A 2009 – FONTE: CLIMATEMPO 32
FIGURA 8 – MÁXIMOS E MÍNIMO DE SENSAÇÃO TÉRMICA MÉDIA PARA OS MESES DO ANO NOS TRINTA MUNICÍPIOS DO RIO DE JANEIRO....32
FIGURA 9 – GRÁFICO BOXPLOT PARA SENSAÇÃO TÉRMICA MÉDIA COLETADOS 33
FIGURA 10 – INSERÇÃO DE CABEÇALHO NO ARQUIVO DO WEKA ANTES DA MASSA DE DADOS..34
FIGURA 11 – RESULTADO PARA A ANÁLISE DE CLUSTER QUE MELHOR AGRUPA OS MUNICÍPIOS EM ATÉ 7 CLUSTERS 36
FIGURA 12 – GRÁFICOS PARA OS 7-CLUSTERS DE SENSAÇÃO TÉRMICA MÉDIA 37

1
Introdução

1.1
Justificativas

O estudo do clima e suas previsões com determinada antecedência tem sido sempre objeto de estudo por parte de pesquisas aplicadas.

Nessa direção, estudar os efeitos do clima combinando-se a sensação térmica (definida pela temperatura+umidade+vento) é fundamental para aumentar a eficácia em modelos de previsão, sejam esses para tratar de questões de agronegócio, consumo de energia, construção civil dentre outras. Até a presente data são ainda poucos os trabalhos de sistemas computacionais de suporte à decisão que permitam efetuar previsões correlacionadas ao clima (não apenas temperatura) gerando resultados com precisão e eficácia satisfatórios. Observa-se também que os trabalhos que incluem a previsão de temperatura, efetuam essa previsão por meio de modelos lineares e não consideraram o clima das diversas micro-regiões.

Assim, uma das metas a serem atingidas é reduzir os erros nas estimativas de previsão onde são considerados os micro-climas com aspectos de similaridade.

O trabalho também permite expor o ciclo de vida para um modelo de conhecimento que se inicia na captação dos dados, passando por diversas etapas de pré-processamento até que a submissão dos dados sejam insumos para os algoritmos aplicados pelas redes neurais artificiais.

1.2
Objetivo

Esta monografia tem por objetivo investigar, analisar e desenvolver uma metodologia inovadora que permita relacionar o clima, de forma mais específica, a sensação térmica das diversas regiões ou unidades geográficas a partir de dados históricos em uma região ou unidade geográfica.

Pretende-se que esta metodologia esteja aderente a um processo de maneira que permita seguir ações processuais e sistemáticas, permitindo que o

conhecimento seja explicitado e documentado para uso futuro e adaptado em outras investigações desse tipo.

Como é necessários efetuar diversas etapas relacionadas ao préprocessamentos de dados (tratamentos iniciais dos dados) foi também incluído um capítulo de forma a descrever o processo que sistematizou tais procedimentos nas várias bases de dados utilizadas.

Como fonte de dados foi necessário a aquisição de uma base de dados de clima (temperatura, umidade e vento) mensal para utilização no estudo e proposição do modelo mais adequado ao problema de previsão em um passo à frente.

A contribuição desse sistema de suporte à decisão é, em escala maior, melhorar o planejamento e gerenciamento de soluções estratégias relacionadas à prevenção como: carga de demanda do setor elétrico, agronegócio e outras aplicações, bastando adaptações em determinados dados de entrada.

1.3
Metodologia

O modelo a ser pesquisado e desenvolvido deverá avaliar inicialmente a sensação térmica agrupando-as por grupos de similaridades e depois os submetendo a modelos de rede neurais com arquitetura de múltiplas camadas com função de ativação (não-linear) de modo a ser aplicado em tarefas de previsão. A análise relacionada a regiões geográficas será feita a partir dos resultados obtidos após a análise dessas correlações.

Após esta pesquisa e análise será desenvolvido um modelo original e inovador de previsão de sensação térmica com um passo à frente, que ajudará a melhorar de forma importante o modelo das previsões de carga em horizonte de médio e longo prazo como sugestões de pesquisas futuras.

As técnicas e teorias, que deverão ser investigadas, terão seus resultados analisados e depois utilizados no processo de previsão da sensação térmica, fundamentalmente Redes Neurais e Modelos Estatísticos. Essas técnicas que buscam obter informações e conhecimento a partir de dados é conhecida por mineração de dados (*data mining*). O capítulo 2 faz uma breve contextualização dessa técnica e sua importância com o advento da inteligência computacional. Esses modelos se adaptam perfeitamente aos objetivos da monografia, uma vez que são capazes de considerar comportamentos não lineares de variáveis climáticas.

2
Mineração de dados – um breve contexto

2.1
O legado como fonte

Nas últimas quatro décadas pode-se ver a evolução, desenvolvimento e implantação de inúmeros produtos na área computacional. Na década de 1960 e 1970 imperaram os equipamentos de grande porte (*mainframes*), aliaram-se a estes os computadores pessoais (*personal computers*), depois equipamentos com capacidades de multi-processamento, e nos últimos tempos a computação de alto desempenho e distribuída.

Em toda esta evolução, cresceram o poder de processamento, os recursos linguagens para implementar algoritmos complexos e o volume de armazenamento de dados. Na última década, o que se viu foram os sistemas de gerenciamento da produção emergirem-se como controladores e executores dos processos operacionais das empresas, a internet como veículo de interligação de repositório de dados e de transações no âmbito mundial e também os processos de automação que invadiram as fábricas.

Porém, a partir da década passada, o papel computacional passou a tomar um rumo diferente daquele processo tradicional e começou aos poucos a ter uma atuação mais sutil para a investigação e solução de questões não lineares, pouco rígidas em parâmetros e que contemplassem mais o mundo real, portanto questões mais complexas. (BARBIERI, 2001).

Neste contexto, uma vez que as grandes massas de dados já estavam depositadas e escondiam muitas preciosidades, vários cientistas e pesquisadores resolveram determinar formas de transformar simplesmente os dados em "conhecimento". Esta foi sem dúvida a palavra que norteou tais pesquisas, e surge agora como a pérola do milênio, dando origem às novas ciências como a biotecnologia, a genética computacional, ao marketing inteligente, dentre outras (WITTEN, 2005).

O ciclo de vida de um sistema de apoio à decisão como este baseado em mineração de dados combina recursos computacionais com a descoberta do conhecimento – Knowledge Discovery – KDD[1], onde princípios científicos subjetivos são transformados em regras objetivas para serem processadas e desta

[1] Extração de novas informações em BD através de vários processos de descoberta do conhecimento.

forma extrair conhecimentos, aprimorar a tomada de decisão, aumentar a segurança, em fim, garimpar dados.

A estrutura deste trabalho procura descrever as técnicas do processo de mineração, o papel das ciências, que de maneira interdisciplinares, estão envolvidas com o conhecimento, além de propor um modelo de processo que permita implementar soluções para apoiar a tomada de decisão. Apenas para se ter uma idéia, o Dr. Arno Penzias, ganhador do prêmio Nobel de Física de 1978, em uma entrevista ao Computer World, definiu mineração de dados (data mining) como "A MUST". Nesta ocasião ele se referia ao assunto como solução para que as empresas conhecessem seus clientes, mercados e toda gama de elementos necessários para manter a sua existência daqui para frente (PONCINELLI, 2002).

2.2
Avalanche de Informações

As empresas passaram estas últimas décadas armazenando dados que agora devem ser extraídos de maneira inteligente e assim serem transformados em conhecimento. O crescimento do acesso aos dados para a tomada de decisão e o mecanismo de sistemas que compõem os bancos de dados, criou o termo "*data warehousing*" - (DW), simbolizandoum enorme armazém de dados. Nos DWs estão depositadas grandes bases de dados contendo dados históricos coletados a partir de transações operacionais.

A mineração dos dados – "*data mining*" – (DM), aparece como elemento complementar para pesquisar o relacionamento e os padrões nesses imensos depósitos e descobrir através de técnicas analíticas um vasto mundo de informações (HAIR, 1998).

2.3
Combinação Interdisciplinar

A mineração de dados (*data mining*) como está diretamente envolvida com a descoberta de conhecimento, tem portanto que ser suportada por um conjunto de disciplinas. As principais, podem ser indicadas como Estatística Multivariada, Ciência Computacional pelo lado da Inteligência Artificial[2] e Redes Neurais e por técnicas não convencionais de resolução de problemas, como Algoritmos Genéticos, Lógica Nebulosa dentre outras que tratam das incertezas naturais e da busca por informações ocultas (HAIR, 1998).

[2] Área da ciência da computação que cria programas para emular o aprendizado do cérebro humano.

Toda essa tecnologia é utilizada para revelar informações estratégicas escondidas nas grandes massas de dados e a atuação combinada de várias disciplinas permite capturar e analisar conjuntos de dados, extraindo significado destes, de modo que se possa descrever características do passado e então predizer tendências futuras (HAIR, 1998).

3
Propecção de conhecimento e Data Mining

Prospecção de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD) é um processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. Trata-se de uma pesquisa de fronteira, que começou a se expandir mais rapidamente nos últimos anos. Sua principal característica é a extração não-trivial de informações a partir de uma base de dados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas e potencialmente úteis.

3.1
Etapas do processo de KDD

O processo para se prospectar conhecimento é composto por um conjunto de atividades contínuas que compartilham resultados descobertos a partir de bases de dados. Esse conjunto é composto de etapas, que são: seleção de dados; pré-processamento e limpeza; transformação; *data mining*; interpretação (análise) e finalmente a obtenção do conhecimento. Como forma de ilustrar a figura 3 a seguir indica essas principais etapas do processo.

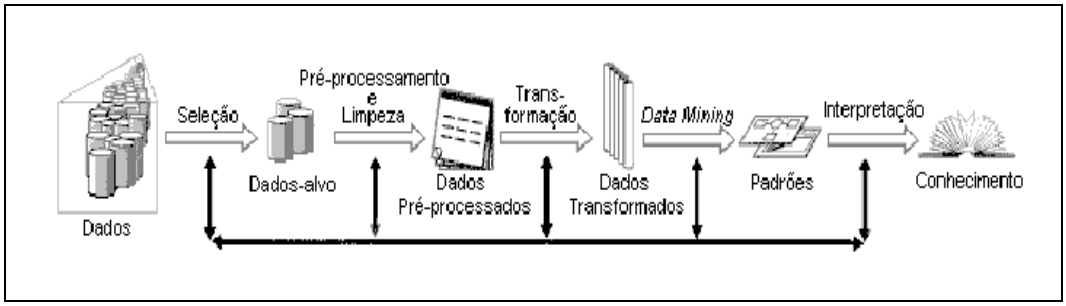


Figura 1 – Etapas do processo de KDD.

O processo de prospecção do conhecimento começa obviamente com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos. Em seguida, é feito um agrupamento organizado de uma massa de dados, alvo da prospecção. A etapa da limpeza dos dados (*data cleaning*) vem a seguir, através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz através da integração de dados heterogêneos, eliminação da incompletude dos dados, repetição de tuplas, problemas de tipagem, etc. De acordo com Pyle (PYLE, 1999), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas. Os dados pré-processados devem ainda passar por uma

transformação que os armazenem adequadamente, visando facilitar o uso das técnicas de *data mining*.

Prosseguindo-se no processo, chega-se à fase que especificamente efetua a mineração. Esta começa com a escolha dos algoritmos a serem aplicados, os quais dependem fundamentalmente do objetivo do processo, tais como, classificação, clusterização, regras associativas, etc. Diversas ferramentas distintas, como redes neurais, indução, árvores de decisão, sistemas baseados em regras e programas estatísticos, usados tanto de maneira isolada quanto em combinação, podem ser então aplicados ao problema. Em geral, o processo de busca é interativo, de forma que os analistas revêem o resultado, formam um novo conjunto de questões para refinar a busca em um dado aspecto das descobertas, e realimentam o sistema com novos parâmetros. Ao final são gerados os resultados que passam então a ser interpretados pelos analistas. Somente após a interpretação das informações obtidas é encontrado o "conhecimento".

3.1.1
Seleção dos Dados

Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias. A maioria das empresas já possui bases de dados. Porém, nem sempre todos os dados necessários estão disponíveis nestas bases o que exige um trabalho de compatibilização (PYLE, 1999).

3.1.2.
Limpeza e Enriquecimento

É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados e onde são estabelecidas as estratégias para resolver os problemas de ausência de dados. As causas que levam à situação de ausência de dados são a não disponibilidade do dado ou a inexistência do mesmo. Uma situação de não disponibilidade ocorre quando da não divulgação do dado, como exemplo, dados de renda da pessoa física em função da obrigatoriedade de sigilo. Ou fator é a inexistência do dado, que ocorre, por exemplo, quando o dado necessário não foi coletado (PYLE, 1999).

3.1.3.
Transformação

Nessa fase, o uso de *data warehouses* se expande consideravelmente, já que nessas estruturas as informações estão alocadas da maneira mais eficiente. Nesses depósitos de dados, os dados são não-voláteis, classificados por assunto, e de natureza histórica, tendendo a se tornarem grandes repositórios de dados. A etapa de transformação permite manipular e adequar os dados a estruturas que sejam mais convenientes para a etapa seguinte (PYLE, 1999).

3.1.4.
Data Mining

A atividade de descoberta do conhecimento é uma das mais fascinantes, onde são processados os algoritmos de aprendizado de máquina e de reconhecimento de padrões. Na maioria das vezes os métodos de mineração de dados estão baseados em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização, modelos para previsão, dentre outros (PYLE, 1999).

3.1.5.
Interpretação

Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para responder às questões desejadas da solução do problema. Também pode correr que em determinadas situações exista a necessidade de se retornar a qualquer um dos estágios anteriores do processo de descoberta (PYLE, 1999).

4
Análise de Cluster

4.1
Descrevendo Análise de Cluster

A análise de agrupamento é um método que permite usar valores das variáveis para planejar um esquema para formar grupos de objetos em classes de modo que objetos similares estejam na mesma classe. O método usado para efetuar esses agrupamentos precisa ser completamente numérico, e o número de classes não é usualmente conhecido (MANLY, 2008).

O objetivo, portanto desse método, que é de aprendizado não supervisionado visa a formar grupos homogêneos ocorrendo que, dentro de cada grupo se busque obter a máxima similaridade e entre os distintos grupos sejam minimizadas suas similaridades.

De acordo com Manly (MANLY, 2008) muitos algoritmos têm sido propostos, mas uma técnica clássica de obtenção de várias médias dos grupos, denominada *k-means*, tem permitido bons resultados. Essa técnica parte de uma matriz com os dados e forma uma outra matriz de dissimilaridade, a partir do cálculo da distância geométrica entre os objetos. Na figura 2, está ilustrada de forma simples a representação da Distância Euclidiana e na equação (1) é ilustra a representação por meio da Distância Euclidiana entre duas variáveis.

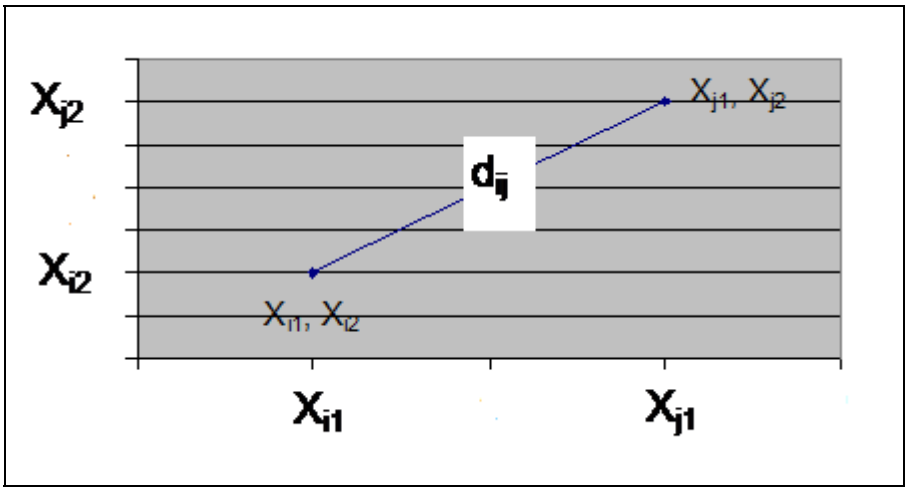


Figura 2 – Distância Euclidiana entre objetos i e j com p=2 variáveis

O algoritmo usualmente utiliza de muitas interações, sendo que a uma dessas iterações vai calculando e envolvendo a distância de centróide de determinado grupo *k*. Nesse algoritmo busca-se minimizar a função V, representada na equação (1).

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2 \quad (1)$$

Nesta equação o parâmetro ***x*** refere-se a cada amostra lida e o valor **μ** é o centróide obtido naquele conjunto de iteração. Quanto mais próximo do centróide o valor está, mais similar o grupo vai se tornando. Com a execução do algoritmo os centróides vão se formando com as amostras ao seu redor, determinando assim grupos como nuvens muito similares (MANLY, 2008).

Witten e Frank [WITTEN, 2005] descreve a implementação do método, que basicamente consiste em escolher o número inicial de clusters e determinar de maneira aleatória seus centróides. A partir desses centróides, vão se associando cada ponto da amostra e verificando-se sua distância, de modo que as menores distâncias são guardadas e consideradas como similares àquele grupamento. Como se trata de um processo iterativo, os centróides vão sendo ajustados, assim como os pontos, de maneira com que uma convergência pela minimização da distância ocorra.

4.2
Modelo aplicado na solução do problema

Em relação ao problema dos microclimas em análise, observou-se que previsões quando se parte de grupos mais similares é possível obter resultados efetivamente mais consistentes. Por essa razão, diversas análises de cluster do tipo *k*-means foram usadas e seus resultados comparados para determinar os grupos com alta similaridade intra-grupos de sensação térmica e baixa similaridade inter-grupos dessa mesma variável. A aplicação do método e sua forma de implementação, assim como o número recomendado de grupos e as respectivas análises de resultados obtidos estão descritos detalhadamente no item 6.3 mais a diante.

5
Redes Neurais Artificiais

5.1
Descrevendo Redes Neurais Artificiais

As redes neurais artificiais são sistemas de processamento paralelo e distribuídos que em algum nível relembram a estrutura do cérebro humano (BRAGA, 2007). Este sistema de processamento paralelamente distribuído é composto de processadores simples, que têm a propensão natural de armazenar conhecimento experimental e torná-lo disponível para o uso. Sua semelhança ao cérebro humano basicamente se dá por dois aspectos, um que o conhecimento é adquirido pela rede a partir de seu ambiente por meio de um processo de aprendizado. O outro aspecto consiste na utilização de pesos para armazenar o conhecimento adquirido (HAYKIN, 1999). A utilização de "algoritmo de aprendizagem" permite instrumentar o processo de aprendizado da rede e sua função é modificar os pesos que atuam nas pontas, denominadas sinapses e assim alcançar a solução para problemas principalmente de entradas não-lineares.

A rede extrai seu poder computacional primeiro por se tratar de uma estrutura paralela distribuída maciça e em segundo que o aprendizado resulta na generalização resultando em saídas adequadas para as entradas que não estavam presentes durante o treinamento a que foi submetida.

Ao longo dos anos, a partir dos trabalhos pioneiros de McCulloch e Pitts em 1943 diversas tentativas de representar e modelar eventos com referência no sistema nervoso. Maiores detalhes históricos e suas evoluções podem ser encontradas em Braga (BRAGA, 2007).

De modo mais amplo, uma rede desse tipo, basicamente é composta de: camada de entrada, o estado de ativação e sua sumarização **"net"**, uma **função de "net"** ou função de ativação que permite a resolução da não-linearidade e por último a camada de neurônios de saída. Também estão presentes os pesos, a estrutura de conexão e os ajustes, "bias". A figura 3 representa esses componentes.

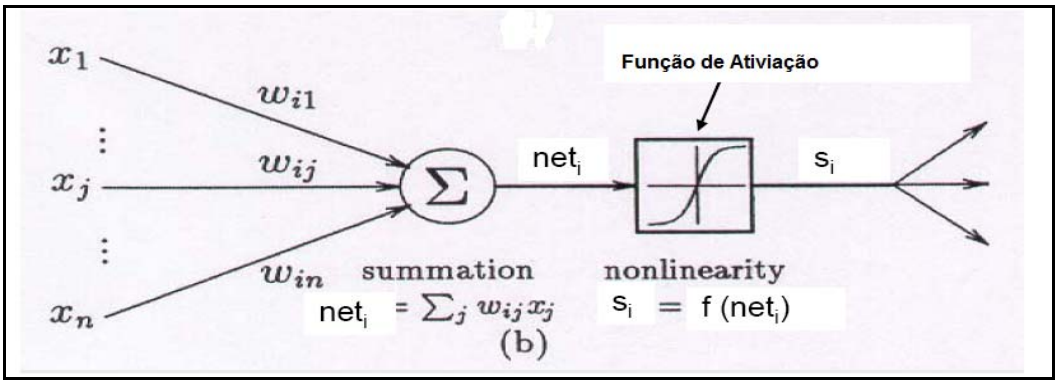


Figura 3– Componentes para um RN

Outro aspecto que muito contribui para a solução de problemas utilizando redes neuronais é o fato dela poder aprimorar-se de modo dinâmico por retroalimentação em sistemas recorrentes. Esse recurso permite, portanto, que os pesos sinápticos efetuem na rede, como um todo, à obtenção da convergência de resultados.

Termos mais específicos e que influencia na estruturação da rede neuronal como: arquitetura, aprendizado, múltiplas camadas com camadas escondidas, épocas, momentum e parâmetros dentre outros serão especificamente abordados nos itens subseqüentes. Tais termos, em função de sua especificidade técnica são imprescindíveis para sua aplicação na solução ao contexto proposto.

5.2
Arquitetura da Rede

Basicamente as redes podem ser configuradas em forma de camada única ou múltiplas camadas. A rede de camada única, onde estão presentes as entradas de seus neurônios fonte e de neurônios de saída contendo nesses neurônios vários dos componentes descritos anteriormente não é objeto de descrição, por suas limitações ou restrições (HAYKIN1999).

No contexto de aplicação com resultados em solução de problemas, principalmente aqueles com tarefas de classificação, categorização ou previsão, a rede de múltiplas camadas é que possui condições para a solução das tarefas mencionadas. Nessa linha, a rede Perceptron de múltiplas camadas (MLP – *Multilayer Perceptron*), com função de ativação (não-lineares) de cada neurônio da rede e da composição da sua estrutura em camadas sucessivas será apropriada, principalmente para tarefas de previsão (BRAGA2007).

A figura 4 a seguir apresenta uma MLP típica com uma camada intermediária, ou oculta. Esta camada gera uma codificação interna para os padrões de entrada, que é então utilizada pela camada de saída da rede.

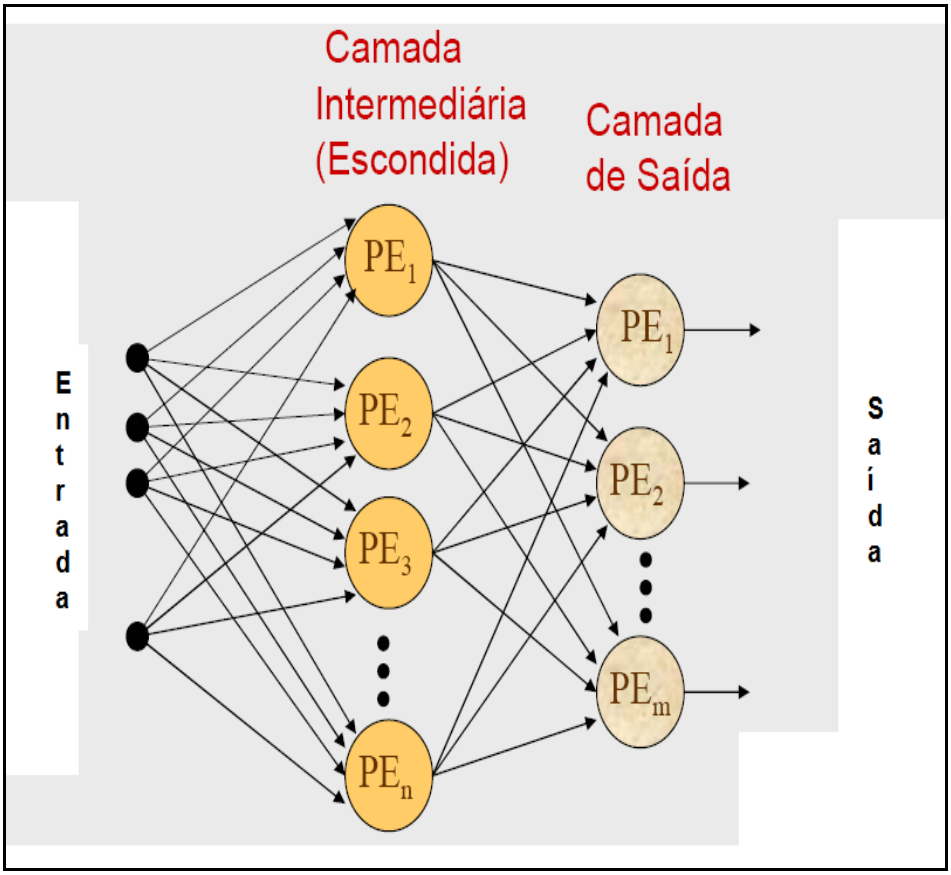


Figura 4 – MPL típica com uma camada oculta

5.3
Aprendizado e Validação

A aprendizagem é um processo pelo qual os parâmetros livres de uma rede neural são adaptados em função de estimulação pelo ambiente no qual a rede está inserida. Vários são esses processos, mas para aplicação em previsões o aprendizado supervisionado, aquele onde são conhecidas saídas e desta maneira é possível a rede efetuar ajustes até encontrar um modelo mais adequado.

O objetivo de um processo de aprendizado é minimizar o erro de todos os processadores da camada de saída para todos os padrões apresentados. Uma forma de efetuar esse controle utiliza a soma dos quadrados dos erros, denominada de "**Esse**", descrita pela equação na figura 5. Nessa equação $\mathbf{t_{pj}}$ é o valor desejado de saída do padrão p para o processador j da camada de saída e $\mathbf{s_{pj}}$ o estado de ativação do processador *j* da camada de saída quando apresentado o padrão *p*.

A minimização do erro quadrático é efetuada pelo método do Gradiente Decrescente, conforme a equação (2). O peso sináptico i do elemento processador j é atualizado proporcionalmente ao negativo da derivada parcial do erro do referido processador em relação ao peso. Quando se está trabalhando com o aprendizado supervisionado, somente se conhece o erro na camada de saída e este é função do potencial interno do processador e depende dos estados de ativação dos processadores da camada anterior e pesos das conexões (RUSSEL, 1999).

$$E_{SSE} = \frac{1}{2} \Sigma_p \Sigma_j (t_{pj} - s_{pj})^2 \quad (2)$$

O algoritmo para o processo de aprendizado "*back propagation*" é executado em duas fases, a primeira as entradas se propagam pela rede desde a entrada até sua saída para a frente (*feed-forward*) e na segunda fase os erros fazem o caminho inverso ao fluxo de dados (*feed-backward*) (BRAGA, 2007).

Algumas vezes torna-se necessária algumas abordagens para resolver problemas de mínimos locais ou ajustes para inércia. Neste caso é incluído o termo *momentum* no ajuste de pesos, podendo assim aumentar a velocidade de convergência em regiões de descida da superfície de erro.

Observa-se que o método de treinamento para redes do tipo "*perceptron*" de múltiplas camadas treinadas com algoritmo "*back propagation*" possui seu erro médio quadrático decrescendo com o número de ciclos (épocas) durante o treinamento (HAYKIN, 1999). Para evitar que o decaimento em direção a um mínimo local na superfície de erro, portanto que a rede acabe sendo excessivamente ajustada aos dados de treinamento. O procedimento que evita a continuidade de treinamentos desnecessários é referido como "parada antecipada". Se comparadas as curvas de aprendizagem de estimação que decresce monotonamente para um número crescente de épocas com a curva de aprendizagem de validação, que decresce para um mínimo, mas depois começa a crescer, conforme o treinamento continue. Portanto este ponto de inflexão é considerado como ponto de parada antecipada, e a partir daqui o que se tem é a presença de ruído contido nos dados de treinamento. O cálculo do erro é obtido de equação (3) a seguir.

$$e_j = (t_j - s_j).\ F'(net_j) \quad (3)$$

Após a etapa de aprendizagem da rede, efetua-se a fase de validação ou teste, para se determinar se o processo de aprendizado foi efetivo e a rede modelada possui realmente a capacidade de aprendizado para todos os padrões em tempo de treinamento. Também é possível verificar se ela está bem generalizada, de modo a não incorrer em especialização o que seria indesejado, uma vez que a modelagem também incluiria os ruídos (RUSSEL, 1999).

Em termos quantitativos o fracionamento da base de dados para atender de um lado ao aprendizado da rede e de outro à etapa de validação, faz-se com proporcionalidade de 80% a 85%. Portanto ficariam para teste os valores restantes entre 15% a 20%, dependendo do tamanho da base (BRAGA, 2007).

De forma a obtermos um resultado comparativo entre as redes e de maneira a selecionar a melhor, emprega-se o cálculo pela porcentagem de erro médio absoluto, representada pelo cálculo do MAPE (Mean Asolute Prcentage Errr). O MAPE é a medida de precisão em um valor de ajuste para a série temporal de dados estatísticos, especificamente para as tendências e geralmente é expressa como uma porcentagem de precisão, tal como definido pela fórmula (4) a seguir.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (4)$$

Onde: $\mathbf{A}_t$ é o valor atual e $\mathbf{F}_t$ o valor de previsão;

A diferença entre os valores $\mathbf{A}_t$ e $\mathbf{F}_t$ é novamente dividida pelo atual valor At . O valor absoluto deste cálculo é somado para cada resultado de ajuste ou previsão, reptindo-se tantas vezes quanto as instâncias e depois dividido pelo numero de destas instâncias. Este cálculo é uma percentagem de erro e permite comparar em nível de ajuste da série. Um ajuste perfeito para o MAPE próximo de zero, o que significa que resultados obtidos nestas proximidades significam melhor ajuste do modelo.

5.4
Função de ativação

De modo a restringir a amplitude da saída de um neurônio é necessário aplicação sobre ele uma função que determine o nível de ativação. Essa função é inserida

após a saída da junção aditiva **(net)** e da *bias*, determinando assim um novo valor do estado de ativação do processador **($s_j = F(net_j)$).**

Alguns tipos destas funções já possuem características matemáticas mais conhecidas o que as tornam mais propícias também à aplicação. Alguns exemplos são a função degrau, tangente hiperbólica, logística e pseudo-linear. Destas citadas, a função logística por ser diferenciavel de modo contínua e sua não-linearidade sigmóide torna sua aplicação em redes *perceptron* de múltiplas camadas. A forma geral é definida pela equação apresentada junto com a função de ativação apresentada na figura 5.

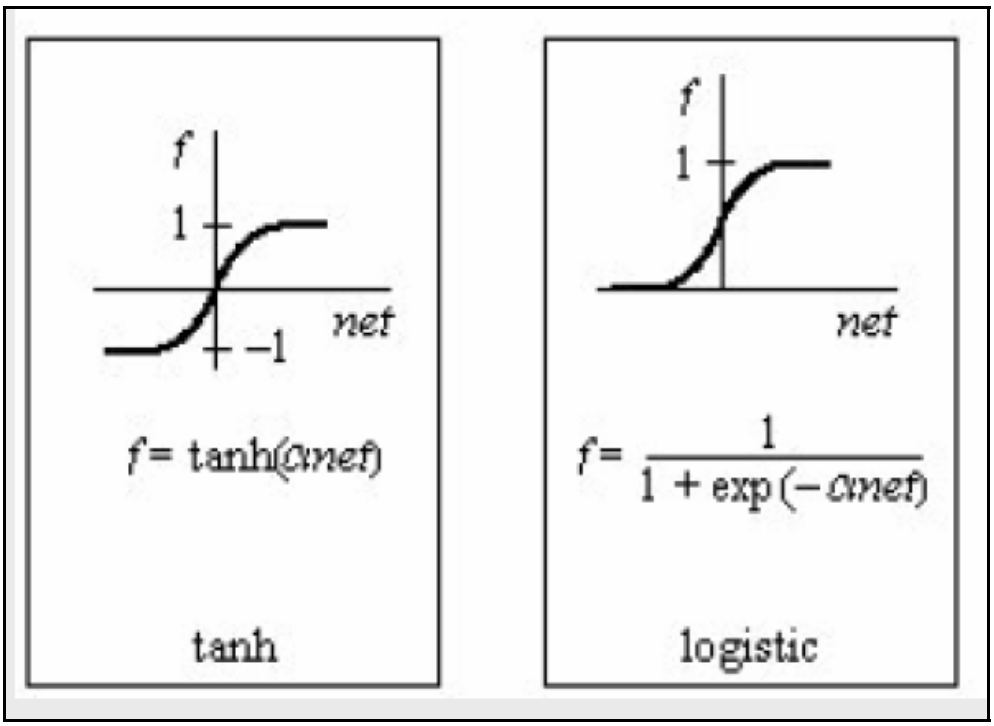


Figura 5 – Representação das funções tangente hiperbólica e logística

5.5
Taxa de aprendizagem

O algoritmo de retropropagação fornece uma "aproximação" para o espaço de pesos que é calculada pelo método de descida mais íngreme. Nesse parâmetro de aprendizagem, quanto menor for a taxa **η**, menor serão as variações de pesos sinápticos da rede, de uma iteração para a outra e mais suave será a trajetória no espaço de pesos. Em contrapartida uma taxa **η** grande para acelerar a taxa de aprendizagem pode tornar a rede instável. Para melhor equilibrar o fator de estabilidade deve conjugar a taxa de aprendizagem com a taxa de momentum como comentado no item 5.3 anteriormente.

5.6
Rede neural aplicado na solução do problema

Em relação ao problema dos microclimas em análise, observou-se que as previsões são efetuadas para períodos à frente. No caso da aplicação com MLP considerou-se o cenário de microclima para cada um dos clusters determinados e a previsão seria de um passo à frente. A consideração de todas as características de arquitetura de rede, algoritmos, escolha de função de ativação, demais parâmetros descritos seriam especificados no momento de execução da rede. Como variáveis de entrada foram determinados: os quatro últimos meses, o décimo segundo mês, assim como as médias móveis dos dois e três últimos meses. Em termos de aprendizado foi determinado em 88% e para validação e teste em 12%.

A aplicação do método e sua forma de implementação, assim como o número recomendado de grupos e as respectivas análises de resultados obtidos estão descritos com mais detalhes no item 6.3 mais a diante.

6

Aplicando Técnicas de Data Mining para Previsão de Micro-clima

6.1

Levantamento de Dados

A base de dados a ser analisada e sobre a qual será elaborado modelos para previsão dos referidos micro-climas é composta de:

- Série histórica mensal da sensação térmica, temperatura mínima e máxima, umidade por microclima de trinta municípios do estado do Rio de Janeiro desde janeiro de 1998 até dezembro 2009 (Fonte: Climatempo);

  Base de dados original: planilha em MS-Excel (DADOS DE CLIMA-ORIGINAL.XLS)

  Metadados: Arquivo formato XLS, versão 2003, 789 Kb, contendo o resumo mensal dos municípios do estado do Rio de Janeiro. Os dados que foram manipulados encontram-se em uma aba (planilha), contendo no cabeçalho: "ano"; "mês"; "município"; "bairro"; "tmax"; "tmin"; "st"; "umax"; "intensidade". Abaixo deste os dados coletados.

A seguir a tabela 1 de investigação e categorização dessas variáveis.

Nome	papel	Nível de mensura-ção				Descrição e observação
Ano	Input	Intervalar	Qualitativa	Nominal	discreta	Ano varia de 1998 a 2009
mês	Input	Intervalar	Qualitativa	Nominal	discreta	Ano varia de 1 a 12 (são os meses de janeiro a dezembro)
municipio	Input	Nominal	Qualitativa	Nominal		Contém a UF do RJ
bairro/município	Input	Nominal	Qualitativa	Nominal		Contém todos os municípios do estado do Rio de Janeiro
Tmax	Input	Intervalar	Quantitativa		contínua	Temperatura maxima atingida no mês em análise
St	Input	Intervalar	Quantitativa		contínua	Sensação térmica calculada (*)
Umax	Input	Intervalar	Quantitativa		contínua	Umidade maxima medida
intensidade	Input	Intervalar	Quantitativa		contínua	intensidade do vento (**)

Tabela 1 – Categorização das variáveis

(*) Modelo de Köppen - **Classificação climática de Köppen-Geiger**, mais conhecida por **classificação climática de Köppen**, é o sistema de classificação global dos tipos climáticos mais utilizados em geografia, climatologia e ecologia. A classificação foi proposta em 1900 pelo climatologista alemão Wladimir Köppen, tendo sido por ele aperfeiçoada em1918, 1927 e 1936. A classificação é baseada no pressuposto, com origem na fitossociologia e na ecologia, de que a vegetação natural de cada grande região da Terra é essencialmente uma expressão do clima nela prevalecente. Assim, as fronteiras entre regiões climáticas foram seleccionadas para corresponder, tanto quanto possível, às áreas de predominância de cada tipo de vegetação, razão pela qual a distribuição global dos tipos climáticos e a distribuição dos biomas apresenta elevada correlação. Na determinação dos tipos climáticos de Köppen-Geiger são considerados a sazonalidade e os valores médios anuais e mensais da temperatura do ar e da precipitação.

(**) informações como vento, temperatura e umidade da tabela são utilizadas para se determiner a sensação térmica.

Esta planilha contém os dados brutos que serão analisados e tratados na etapa do pré-processamento. Procedimentos e detalhes de prepraração de dados para *data mining* podem ser encontrados em Pyle (PYLE, 1999).

Problemas encontrados nas bases de Dados Enviadas

- Inconsistência do nome da coluna ”município”, esta coluna contém a UF referente ao estado “RJ”. Provavelmente não será utilizada por se tratar de informações que se repetem para todos os municípios.

- Coluna “bairro/município”, essa coluna contém apenas os municípios e portanto seu nome correto deve ser apenas “município”.

6.2
Definição do Método para Coleta de Dados

Os dados foram obtidos da base fonte: CLIMATEMPO e estão descritas na tabela1 anteriormente apresentada.

6.2.1
Análise dos Dados Coletados

Os dados originais foram anteriormente tratados pelos técnicos do ICA (PUC-RJ). Nesse tratamento pudemos observar que os dados já haviam sido limpos e substituídos quando fosse o caso. Também notamos que a faixa dos anos coletados anteriormente de 1998 a 2009 foi reduzida para uma faixa de 2000 a 2009.

Nessas análises foram verificadas para os dados referentes às temperaturas: máxima e mínima e também para a sensação térmica.
Para cada município foram geradas médias dos dez anos coletados (2000 a 2009) para os meses de janeiro a dezembro, das temperaturas máxima e mínima e também da sensação térmica, referente ao mesmo período.
Esta preparação inicial ocorreu porque se pretende efetuar uma análise de cluster, uma vez que este tipo de análise é recomendável para situações de agrupamento que ainda não são conhecidos. Assim o objetivo é poder determinar quais seriam os municípios com climas semelhantes e agrupá-los em conjuntos de características semelhantes. No arquivo “Base_Processamento”, a aba da planilha contém os dados de entrada e seus respectivos meta dados.
A etapa de tratamento dos dados a serem processados pelo software WEKA de forma a gerarem cluster é executada da seguinte forma:

Os dados foram inicialmente preparados em planilhas. Cada planilha com sua respectiva aba, congregando os nomes e dados para Temperatura Máxima, Temperatura Mínima e Sensação Térmica.

Desses dados, foram escolhidos com base para a previsão aqueles dados relativos à sensação térmica (temperatura, umidade e vento), por ser mais significativa do que apenas as temperaturas máximas e mínimas.
Na figura 6 esta representada a estrutura de cabeçalho.

Numero	Município	jans	fevs	mars	abrs	mais	juns	juls	agos	sets	outs	novs	dezs

Figura 6 – cabeçalho da base de dados inicial.

O campo município é composto por trinta municípios do estado do Rio de Janeiro e são apresentados na tabela 2.

Numero	Município
1	Barra do Pirai
2	Barra Mansa
3	Belford Roxo
4	Carmo
5	Comendador Levy Gasparian
6	Duque de Caxias
7	Engenheiro Paulo de Frontin
8	Itaguai
9	Japeri
10	Mendes
11	Mesquita
12	Miguel Pereira
13	Nilopolis
14	Nova Iguacu
15	Paracambi
16	Paraiba do Sul
17	Paty do Alferes
18	Pinheiral
19	Pirai
20	Quatis
21	Queimados
22	Rio Claro
23	Rio das Flores
24	Sao Joao de Meriti
25	Sapucaia
26	Seropedica
27	Tres Rios
28	Valenca
29	Vassouras
30	Volta Redonda

Tabela 2 – Municípios do estado do Rio de Janeiro participantes da análise

Em cada planilha havia as 30 cidades com os dados das médias relacionados em séries. Por exemplo. A cidade Barra do Piraí, contendo para os meses de janeiro a dezembro as médias desses meses para todos os anos (de 2000 a 2009).

A visualização através de gráficos mostra que as temperaturas estão maiores nos meses das pontas e menores nos meses internos (abril a setembro).

O gráfico apresentado na figura 7 ilustra essas variações. Os dados encontram-se na aba “sensação térmica”.

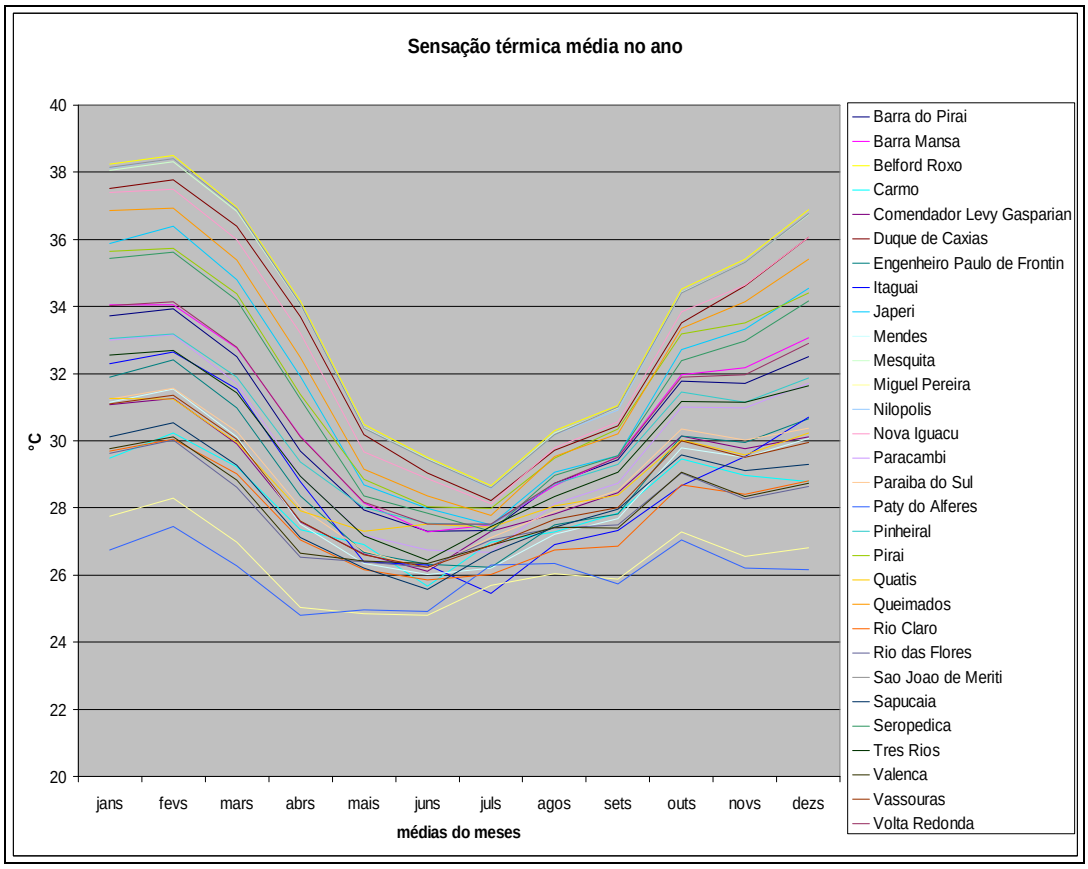


Figura 7 – Sensação térmica média para os meses do ano nos trinta municípios do Rio de Janeiro de 2000 a 2009 – Fonte: Climatempo

Junto com a preparação das médias de sensação térmica é importante uma análise estatística de perfil. Nessa análise são apresentados os valores máximos, mínimos e é possível verificar se existem pontos fora de perfil (*outliers*).

A figura 8 mostra os valores máximos, mínimos e a quantidade de valores para as respectivas sensações térmicas médias coletadas.

	jans	fevs	mars	abrs	mais	juns	juls	agos	sets	outs	novs	dezs
Máximo	38,24387	38,49446	36,95742	34,157	30,49581	29,501	28,66903	30,31548	31,04933	34,52968	35,415	36,88194
Mínimo	26,7529	27,44831	26,27774	24,79667	24,85613	24,79	25,45871	26,04613	25,74033	27,05129	26,20333	26,14839
Quantidad	30	30	30	30	30	30	30	30	30	30	30	30

Figura 8 – Máximos e mínimo de sensação térmica média para os meses do ano nos trinta municípios do Rio de Janeiro

A figura 9 mostra o gráfico *boxplot*, gerado no software MatLab para os dados de sensação térmica média coletados.

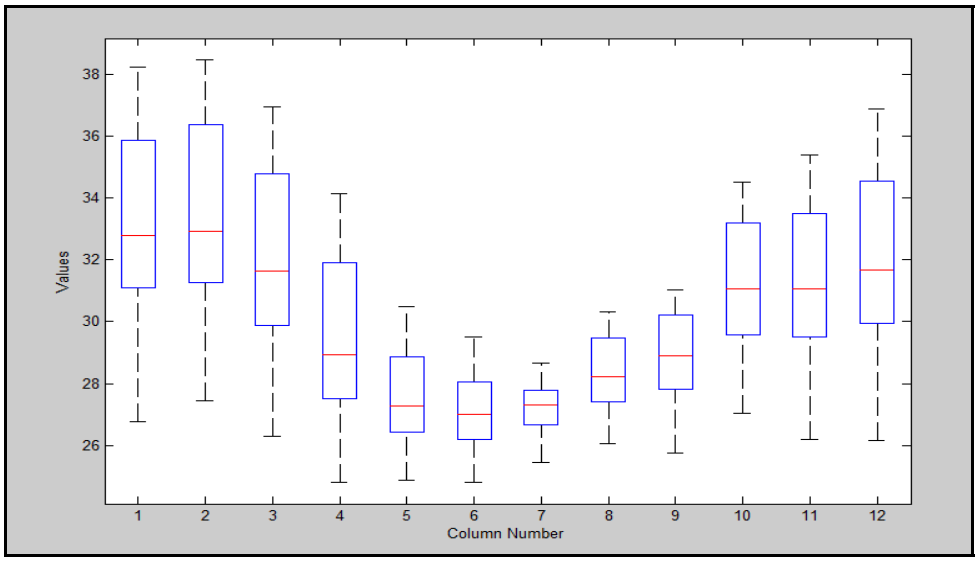


Figura 9 – Gráfico boxplot para sensação térmica média coletados

Este gráfico é importante pois facilita o reconhecimento de *outliers*, o que de fato não ocorreu com esses dados de entrada. No eixo horizontal, estão relacionados os valores médios para os doze meses do ano.

Após a preparação das planilhas pelas médias, a próxima etapa era preparar os dados para se aplicar a clusterização a partir do uso do software WEKA.

6.3
Aplicação da analise de cluster no agrupamento de microclima

Por se tratar de uma técnica que investiga as amostras, classificando-as sem conhecer previamente a qual grupo elas pertencem. Isto significa que a classificação em informações de acordo com a comparação dos valores numéricos entre os dados. Desta forma, o algoritmo vai fornecer uma classificação sem nenhuma supervisão humana, isto é, sem nenhuma pré-classificação. Esta característica é considerada como *não* supervisionado. Detalhes podem ser encontrados em Witten, (WITTEN, 2005).

No Anexo 1 estão descritos os procedimentos de manipulação do software WEKA na obtenção dos resultados quando da submissão da base de dados. A análise de cluster utilizada foi SIMPLE K-MEANS, com uso da Distância Euclidiana (EuclideanDistance).

O objetivo é gerar vários resultados para as análises de cluster, observando como as cidades estão agrupadas em termos de sensação térmica média, verificando a coerência desses resultados. A análise de cluster levará em consideração no caso, grupamentos de 5, 6 e 7 clusters.

A definição para a quantidade de cluster mais adequada será estabelecida para aqueles clusters onde a ocorrência da variação se dê em até um limite de 1 grau para mais ou para menos na variação da temperatura.

Como o software para desenvolvimento da análise de cluster foi utilizado o WEKA, os dados devem ser preparados para serem inseridos no software.

Preparação para o WEKA

Deve-se gear um cabeçalho para que o software consiga interpretar os dados. Esse cabeçalho deve ser adicionado à base de dados. O formato é um arquivo texto, com extensão "TXT". Este arquivo deve ser convertido para o formato de separação dos campos por vírgula e a extensão no caso é "CSV". Somente após essas conversões é possível gerar o arquivo que poderá ser lido pelo WEKA, formato de extensão "ARFF".

O procedimento de conversão dos dados da planilha XLS para formato de dados TXT é obtido salvando esta planilha com a opção de arquivo para CSV. Depois esse arquivo deve ser gerado com a separação decimal utilizando ponto ".", e os valores numéricos com separação por ",".

A inserção de um cabeçalho especial para o reconhecimento do WEKA é apresentado na figura 10 a seguir.

Abrir no Word Pad, colar o arquivo de cabeçalho:

```
@relation clima_muncipios_estado_RJ_sensterm

@attribute jans REAL
@attribute fevs REAL
@attribute mars REAL
@attribute abrs REAL
@attribute mais REAL
@attribute juns REAL
@attribute juls REAL
@attribute agos REAL
@attribute sets REAL
@attribute outs REAL
@attribute novs REAL
@attribute dezs REAL
@data
```

Salvar esse arquivo com extensão ARFF
Processá-lo no WEKA

Figura 10 – Inserção de cabeçalho no arquivo do WEKA antes da massa de dados

A definição para a quantidade de cluster mais adequada será estabelecida para aqueles clusters onde a ocorrência da variação se dê até um limite de 1 grau para mais ou para menos na variação da temperatura.

Ao se executar o processamento por meio do WEKA é possível validar qual a variação da temperatura. Foi estabelecido com base em outros estudos e situações que para essa quantidade de amostras (30 municípios), utilizaria cinco

clusters, aumentando para seis, sete, até alcançar a variação desejada. O parâmetro de análise ideal seria manter a variação de temperatura abaixo de 1°C.

Base de entrada para o WEKA

A partir da planilha [BASE_PROCESSAMENTO – SENSAÇAO_TERMICA.XLS] contendo a média de "sensação térmica" para os 30 municípios, converteu-se esses dados no arquivo em formato TXT, com separadores por ";".

Este arquivo gerado é que será lido pelo software WEKA de modo a ser processado e analisado em cluster.

Execução dos experimentos utilizando Análise de Cluster para a Sensação Térmica Média

Como o software WEKA permite a configuração de parâmetros antes de efetivamente executar os resultados. A tabela 3 contém os seguintes procedimentos de configurações e parâmetros devem ser efetuados:

Aba: "Cluster"	
Choose: "K-Means"	
Especificar os parâmetros:	
DisplayStDev: True	
DistanceFunction: euclidian	
MaxIterations: 500	
Cluster: 5	Partindo-se de que em 30 cidades, um grupo interessante de cluster deva variar de 5 a 9.
Seed	Fornece o número de sementes, inicialmente 10

Tabela 3 – campos de configuração do WEKA para efetuar a execução do cluster

Após a execução feita a partir do botão "start", os resultados gerados são obtidos clicando com o botão da direita sobre o resultado do quadro de lista de resultados. A opção "visualize cluster assignments";

Salvar o resultado, que é um novo arquivo "ARFF", o qual permite observar como foram agrupados os dados por clusters.

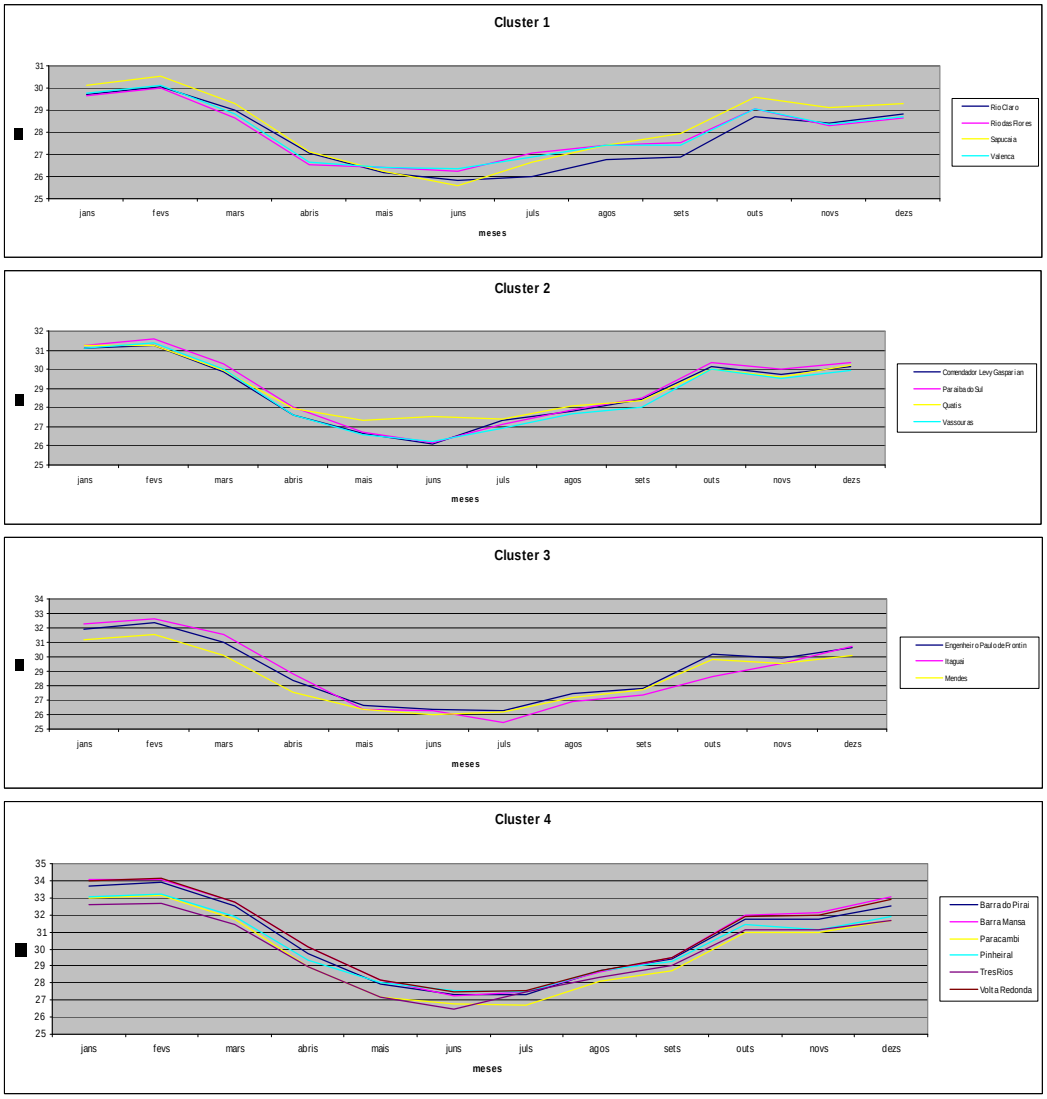
Diversos experimentos foram executados e o escolhido para representar grupo de 7-cluster estão apresentados na figura 11 a seguir.

Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6		Cluster7	
Cidade	SensaTerm	Cidade	SensaTerm	Cidade	SensaTerm	Cidade	SensaTerm	Cidade	SensaTerm	Cidade	SensaTerm	Cidade	SensaTerm
Carmo	28,18152267	Comendador	28,8473788	Engenheiro	29,07702525	Barra do Pir	30,55116833	Miguel Pereira	26,33043792	Belford Roxo	33,72583458	Japeri	31,8608174
Rio Claro	27,78196742	Paraiba do S	29,0222635	Itaguai	28,87988808	Barra Mansa	30,77770217	Paty do Alfere	26,0791775	Duque de Caxias	33,09797042	Pirai	31,9139553
Rio das Flores	27,94437642	Quatis	29,0796907	Mendes	28,59688742	Paracambi	29,83135567	**Média**	**26,20480771**	Mesquita	33,6175805	Queimados	32,4663349
Sapucaia	28,23986242	Vassouras	28,746425	**Média**	**28,85126692**	Pinheiral	30,253771	**Desv-pad**	**0,17667944**	Nilopolis	33,62411642	Seropedica	31,5038877
Valenca	27,99375342	**Média**	**28,983111**	**Desv-pad**	**0,13397027**	Tres Rios	29,84011458	**Max-Min**	**0,251260417**	Nova Iguacu	32,97309967	**Média**	**31,9362488**
Média	**28,03693223**	**Desv-pad**	**0,12100371**	**Max-Min**	**0,480137833**	Volta Redon	30,774696			Sao Joao de Meri	33,63888067	**Desv-pad**	**0,39754007**
Desv-pad	**0,212650488**	**Max-Min**	**0,33326567**			**Média**	**30,33813463**			**Média**	**33,44624704**	**Max-Min**	**0,96244725**
Max-Min	**0,457895**					**Desv-pad**	**0,43855947**			**Desv-pad**	**0,32293486**		
						Max-Min	**0,9463465**			**Max-Min**	**0,75273491**		

Figura 11 – Resultado para a análise de cluster que melhor agrupa os municípios em até 7 clusters

A representação gráfica das curvas de cada um dos 7 clusters das sensações térmicas médias encontra-se na figura 12.

Observa-se que os municípios escolhidos possuem semelhanças de similaridade por temperatura, satisfazendo à condição de diferenças máximas e mínimas menores ou iguais a 1°C.



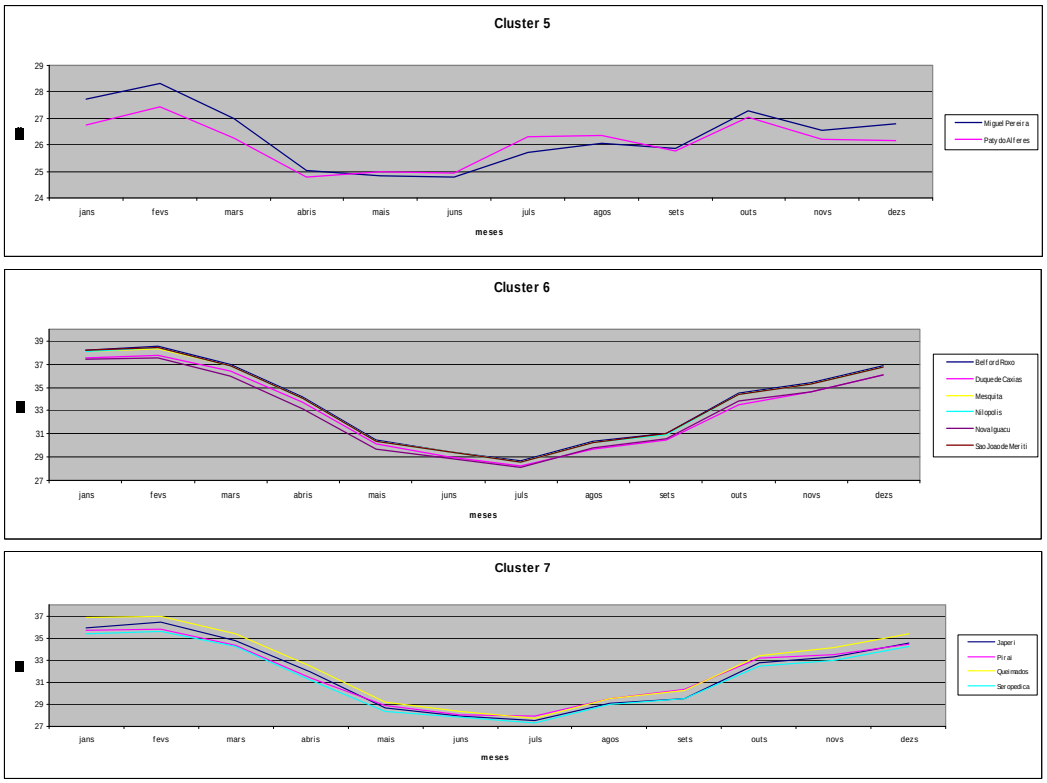


Figura 12 – Gráficos para os 7-clusters de sensação térmica média

6.4

Modelagem por rede neural para a previsão de sansação térmica do microclima

A partir da aplicação do algoritmo em uma rede Perceptron de múltiplas camadas (MLP – *Multilayer Perceptron*), com função de ativação (não-lineares) de cada neurônio desta rede e da composição da sua estrutura em camadas sucessivas apropriada. O entrego da rede respeitou a seguinte arquitetura:

Uma camada escondida, configurada com função: 'tansig' e 'logsig'

9 entradas de neurônios, contendo:

Mês, M-12, M-4, M-3, M-2, M-1, MM-2, MM-3 e o objetivo de previsão (target)

1 saída de previsão imediatamente à frente o algoritmo, implementado no software Matalab, permitiu que a rede fosse executada para cada cluster, totalizando os sete grupos (*clusters*) e validando pelo MAPE (Mean Absolute Percentagem Error) percentual, conforme explicado no item 5.3.

Os dados de entrada, transformados em uma matriz 9X96, foram submetidos individualmente ao algoritmo (Anexo 2) e os resultados compilados em planilha e apresentados para análise no item 6.5.

6.5
Análise dos resultados obtido

Os resultados obtidos após executar o algoritmo com dez inicializações de rede neural para cada cluster executado.
As tabelas 4, com letras de "a" a "g" a seguir apresentam os resultados referentes a cada uma dessas dez inicializações, separadas por cada um dos sete *cluster* analisados. Para facilitar a análise tais resultados foram ordenados do menor para o maior percentual de erro de validação, *Mape_valid*.

Cluster 1		
Mape_train	**Mape_valid**	**Mape_test**
4.32	3.50	6.52
4.15	3.56	6.72
4.25	3.86	6.34
4.06	4.02	6.49
4.22	4.02	6.16
4.38	4.21	5.55
4.48	4.28	6.61
4.23	4.32	6.28
6.12	4.37	6.36
4.11	4.64	6.20

(a)

Cluster 2		
Mape_train	**Mape_valid**	**Mape_test**
5.36	3.76	6.35
4.47	3.89	6.45
4.00	3.94	5.83
4.30	4.11	6.05
4.02	4.22	4.97
4.57	4.27	6.66
5.37	4.36	7.63
4.15	4.41	5.55
4.81	4.44	7.65
4.15	4.68	5.98

(b)

Cluster 3		
Mape_train	**Mape_valid**	**Mape_test**
5.22	4.70	8.20
5.27	4.81	7.82
5.43	4.99	8.33
5.44	5.02	8.11
6.20	5.06	8.40
5.13	5.13	8.28
5.66	5.15	7.71
6.17	5.26	8.51
5.40	5.33	8.67
5.06	5.56	7.58

(c)

Cluster 4		
Mape_train	**Mape_valid**	**Mape_test**
4.48	4.26	6.10
4.52	4.32	6.00
4.52	4.38	5.85
4.61	4.38	6.44
4.56	4.39	5.92
4.64	4.46	6.61
4.60	4.55	5.99
4.79	4.64	6.49
5.75	4.70	8.29
4.52	4.85	6.04

(d)

Cluster 5		
Mape_train	**Mape_valid**	**Mape_test**
4.59	4.54	5.31
4.88	4.55	6.61
3.99	4.62	5.11
4.59	4.63	5.61
4.93	4.70	6.01
4.15	4.85	4.88
4.22	4.85	5.43
4.22	5.00	5.65
3.98	5.08	5.84
4.13	5.26	5.29

(e)

Cluster 6		
Mape_train	**Mape_valid**	**Mape_test**
4.11	3.78	6.43
4.72	3.81	7.44
4.09	3.84	6.49
3.90	3.84	6.71
3.91	3.94	6.46
3.93	3.97	6.64
3.76	4.03	6.62
3.84	4.03	6.64
4.77	4.46	8.50
8.61	8.54	10.23

(f)

Cluster 7		
Mape_train	**Mape_valid**	**Mape_test**
7.69	4.71	7.16
9.48	5.09	9.04
7.35	5.19	6.71
6.93	5.21	7.14
8.37	5.83	6.81
4.47	5.94	7.35
8.25	6.04	10.28
4.37	6.05	7.08
6.23	6.26	6.49
4.13	6.76	7.48

(g)

Tabela 4 (a) até (g) – Cluster processados pelo algoritmo de rede neural executado pelo MatLab

Portanto, pode-se observar que os percentuais obtidos em cada *cluster* variam de um valor mínimo de 3,50% ao máximo de 8,54% para o MAPE de validação, pode-se considerar resultados de validação expressivos, refletindo-se na rede de treinamento. Este resultado sinaliza, portanto, que a rede possui um nível de generalização bom, consequentemente baixa especialização.

A tabela 5 a seguir apresenta a relação das melhores redes baseadas em seus MAPE de validação (*Mape_valid*) referentes a cada *cluster*, alem de informar os valores da média geral de todos os valores de sensação térmica (ST), de desvio padrão destas, de diferenças entres as sensações médias máximas e mínimas e também das quantidades de municípios por *cluster*.

Numero do Cluster	Qtde Municípios	Média ST	Variação Maxima-Minima ST	Desvio	MAPE_Valid
Cluster 1	5	28,03	0,46	0,21	3,5
Cluster 2	4	28,92	0,33	0,12	3,76
Cluster 3	3	28,85	0,48	0,14	4,7
Cluster 4	6	30,17	0,95	0,43	4,26
Cluster 5	2	26,2	0,25	0,18	4,54
Cluster 6	6	33,45	0,75	0,32	3,78
Cluster 7	4	31,94	0,96	0,4	4,71

Tabela 5 – Relação das melhores redes baseada nos resultados de MAPE_valid

Infelizmente não se pode inferir uma análise que associe qualquer uma dessas variações ao melhor desempenho da rede, uma vez que os resultados parecem não seguir nenhum comportamento desse tipo.

7
Conclusões

A execução do trabalho baseado em experimentos foi muito importante para permitir executar todo o ciclo de vida de uma aplicação de obtenção de conhecimento por técnicas diversas, as quais combinaram desde a coleta da informação, inclusão de metadados, investigação de qualidade de dados e pré-processamento, até busca do conhecimento por técnicas de agrupamento e similaridades e modelagem de previsão por meio de redes neurais do tipo Perceptron de múltiplas camadas (MLP – *Multilayer Perceptron*), com função de ativação (não-lineares) de cada neurônio desta rede.

Também foi possível verificar que análises desse tipo não se concentram em um único software, mas sim, está baseada em conjuntos de métodos que suportem todo um ciclo de vida.

Os resultados oriundos dos algoritmos aplicados devem servir de base para se efetuar trabalhos investigativos em outras condições de microclimas, podendo assim validar e aperfeiçoar tais métodos. A aplicação do método de agrupamentos (analise de cluster) mostrou-se como forte aliada para conhecimento de grupos em situações de aprendizado não supervisionadas.

8
Sugestão de Trabalhos Futuros

No escopo de trabalhos futuros, esse estudo pode descortinar uma série de sugestões que vão desde sua aplicação em investigações de sensação térmica de microclimas de outras regiões, passando por inclusão de variáveis como altitude e fenômenos como El Ninõ e La Ninã como parâmetros de entrada das redes neurais utilizadas. Também pode-se sugestionar alterações das funções de ativação e camadas ocultas de neurônios como forma de comparação aos resultados aqui obtidos e analisados.

9
Referências Bibliográficas

BARBIERI, CARLOS, **BI – Business Intelligent – Modelagem & Tecnologia**. Rio de Janeiro: Axcel Books, 2001, 350p.

HAIR Jr, J. F., et al, **Multi Variate data Analysis.** New Jersey **:** Prentice Hall, 4a. ed.,1998, 600 p.

HAYKIN, S., **Redes Neurais – Princípios e Práticas.** São Paulo: Bookman, 2a Ed., 1999.

MANLY, B. J. F., **Métodos Estatísticos Multivariados – Uma Introdução.** Porto Alegre: Artmed – Bookman, 3a ed., 2008.

PONCINELLI, C. A. et al, **Data Mining e Suas Técnicas**. Campinas: Mestrado PUC Campinas, 2002, 47p.

PYLE, D., **Data Preparation for Data Mining.** San Francisco: Morgan Kaufmann, 1999, 539p.

RUSSEL, D. R, et al., **Neural Smithing.** Cambridge : MIT Press, 1999.

WITTEN, I. H., **Data Mining – Pratical Machine Learning Tools and Techniqes**. San Francisco: Morgan Kaufmann, 2005, 505p.