MÉTODO PARA IDENTIFICAÇÃO DE DADOS METEOROLÓGICOS INCORRETOS[1]

Luciano Pugsley[2], Paulo Henrique Caramori[3], Danilo Augusto Bambini Silva [4]

ABSTRACT – The goal of this work was to evaluate an automated method to verify incorrect meteorological data inserted in a database. The error risk of inserting an incorrect meteorological data was determined by a quadratic Bayesian classifier, using series of data from 32 meteorological stations of Instituto Agronômico do Paraná (IAPAR), in Paraná state, Brazil. The results showed that the method points failures of manual and automatic data insertion. The method with supervision of a specialist makes possible a secure and efficient substitution of incorrect data.

INTRODUÇÃO

No presente trabalho, propõe-se identificar dados meteorológicos incorretos inseridos em uma base de dados agrometeorológica, utilizando-se como estudo de caso os dados do Instituto Agronômico do Paraná. Trabalhos desenvolvidos na área de Agrometeorologia deste Instituto, como o desenvolvimento de banco de dados e métodos de classificação estatísticos para o refinamento e a melhoria da qualidade dos dados meteorológicos, constituem a base para identificação de dados meteorológicos incorretos (Caramori et al. 2003, Cruvinel et al. 2002, Pugsley 2002).

MATERIAL E MÉTODOS

O uso de probabilidade é importante em reconhecimento de padrão de erros de dados meteorológicos, devido à casualidade em que normalmente são geradas as classes de um padrão. É possível fazer uma aproximação de classificação ótima esperando que, em média, aconteça a probabilidade mais baixa de se cometer erros de classificação. A probabilidade de um padrão pertencer a uma classe é dada pela regra de Bayes (Duda, 1973). Assim, supondo-se que um x em particular é o desejo de se realizar uma ação, $p(\mathbf{x}|\ \omega_j)$ representará a função de densidade de probabilidade condicional do padrão e $P(\omega_j)$ a probabilidade a priori do padrão x pertencer ao estado $\boldsymbol{\omega}_j$, conforme mostrado nas equações 1 e 2.

$$P(\omega_j\ |\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})} \qquad p(\mathbf{x}) = \sum_{j=1}^{S} p(\mathbf{x}\mid \omega_j)P(\omega_j)$$

Equação 1 Equação 2

Se o estado verdadeiro da natureza for $\boldsymbol{\omega}_j$, sofre-se a perda $\lambda(\alpha_i|\ \omega_j)$. Considerando que $P(\omega_j|\ \mathbf{x})$ é a probabilidade de que o estado verdadeiro da natureza seja $\boldsymbol{\omega}_j$, a perda esperada associada com a ação $\boldsymbol{\alpha}_i$ será dada por:

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{S} \lambda(\alpha_i|\omega_j)P(\omega_j \mid \mathbf{x})$$

Equação 3

Uma perda pode ser chamada de risco, e $R(\alpha_i|\ \mathbf{x})$ é conhecido como risco condicional. Pode-se minimizar a perda para todo x, selecionando uma ação que minimize o risco condicional. Este procedimento na verdade é uma decisão Bayesiana ótima.

Para minimizar o risco total, basta calcular o risco condicional e selecionar a ação $\boldsymbol{\alpha}_i$ para a qual o risco condicional é mínimo. O risco total mínimo é chamado de risco de Bayes e é o melhor resultado que pode ser alcançado pelo método.

O classificador é visto como uma máquina que calcula c funções discriminantes e seleciona a classe que corresponde ao maior discriminante, como ilustra a figura 1.

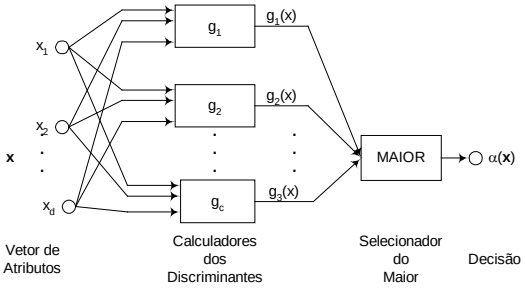

Figura 1. Classificador de padrões.

Para o caso geral, considera-se $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$, de forma que o maior resultado das funções discriminantes corresponderá ao menor risco condicional. Para o caso MAP, pode–se considerar $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$, sendo que o maior resultado das funções discriminantes corresponderá à maior probabilidade a posteriori.

Densidades multivariadas gaussianas têm sido freqüentemente utilizadas para caracterizar classes. A forma geral dessa densidade é dada por:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Equação 4

onde $\mathbf{x}$ é o vetor (d x 1) de atributos, $\boldsymbol{\mu}$ é o vetor (d x 1) de médias, $\boldsymbol{\Sigma}$ é a matriz (d x d) de covariância, $(\mathbf{x} - \boldsymbol{\mu})^t$ é a transposição de $\mathbf{x} - \boldsymbol{\mu}$, $\boldsymbol{\Sigma}^{-1}$ é a inversa de $\boldsymbol{\Sigma}$, e $|\boldsymbol{\Sigma}|$ é o determinante de $\boldsymbol{\Sigma}$.

A densidade multivariada normal é especificada pelos parâmetros $d = d\ (d + 1) / 2$, os elementos do vetor de médias $\boldsymbol{\mu}$ e os elementos independentes da matriz de covariância $\boldsymbol{\Sigma}$. Amostras tiradas de uma população normal tendem a se juntar em um único agrupamento.

A equação 5 mostra as funções discriminantes para o caso gaussiano multivariado

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

Equação 5

[1] Trabalho parcialmente financiado pela AGROCONSULT – Ministério da Agricultura, RJ, Brasil.
[2] AGROCONSULT, Rio de Janeiro, RJ, Brasil. Consultor pesquisador colaborador na área de Agrometeorologia–IAPAR.
[3] Pesquisador do IAPAR, Londrina, PR, Brasil. PhD. Agrometeorologia.
[4] Depto. de Computação, Universidade Estadual de Londrina (UEL), Londrina, PR, Brasil. Bolsista CNPQ/PIBIC.

RESULTADOS E DISCUSSÃO

A condição de se aceitar um dado inválido ou não e a interpretação dos resultados foram analisados conforme o risco total mínimo apresentado pelos dados históricos de cada estação meteorológica. Foram utilizados para o vetor de padrão os dados de temperatura máxima, temperatura média e temperatura mínima e posteriormente os dados de umidade relativa, evaporação e insolação.

O método apresentou várias situações de dados inválidos, aumentando a possibilidade de correção da base de dados pelo gerente da base ou por um sistema automatizado, que poderá realizar as alterações nos dados conforme critérios definidos pelo especialista.

O item 1 da Tabela 1 ilustra a ocorrência de um rico de 11%, representado por um valor negativo, pois o erro representa um risco. Quanto menor o valor do risco bayesiano maior a probabilidade deste valor ser inválido. Este caso mostra a temperatura máxima abaixo da temperatura mínima e média. Outro caso ilustrado no item 2 da Tabela 1 com esta mesma percentagem ocorrendo quando a temperatura máxima de 12 ^{o}C foi inserida como 22 ^{o}C e o item 3 da Tabela 1 quando foi inserido 3 ^{o}C, ao invés de 13 ^{o}C. Os itens 5 a 9 mostram casos semelhantes. O item 4, mesmo com 11% de risco de alguma temperatura estar incorreta, não apresentou nenhum dado incorreto, conforme julgamento do especialista. O item 10 da Tabela 1 revelou um problema de inserção automática de dados, onde deveria ter sido inserido um valor nulo (NULL) ou um código de erro para valores de dados inexistentes, foi inserido o valor 0, que para o sistema representa um valor válido.

Tabela 1. Risco de aceitar um valor de temperatura incorreto entre 0 - 100 %. Valores entre parênteses e em negrito representam as temperaturas incorretas.

	DATA	ESTAÇÃO	Risco	TMax	TMed	TMin
1	21/05/1978	2251027	-11	**(10)** 20	16	18
2	21/07/2000	2349030	-11	**(22)** 12	10	-2
3	05/07/1978	2350018	-11	27	17	**(3)** 13
4	02/06/1978	2551010	-11	20	5	-4
5	10/10/2001	2551010	-11	26	15	**(0,8)** 8
6	15/05/1967	2548038	-12	28	**(36)** 23	16
7	18/04/1967	2548038	-14	31	**(39)** 21	16
8	07/11/1973	2653012	-14	30	**(22)** 21	**(3)** 13
9	27/04/1978	2452050	-15	33	19	2
10	16/05/1967	2548038	-21	28	28	**(0)** -99

Risco - Risco de se aceitar um valor de temperatura icorreto;
TMed – Temperatura Máxima
TE – Temperatura Média
TI – Temperatura Mínima

Na Tabela 2 pode-se observar que o resultado do método extrapola a porcentagem de 100% do risco de erro, quando os valores analisados pelo método já possuem a indicação de um código de valor inexistente. Como apresentam os itens 1 e 2, com risco de 307%.

O item 2 da Tabela 2 revelou o que ocorre quando um valor está muito fora dos valores dos dados da estação. Neste caso o valor de temperatura média deveria ser de 18^{o}C e foi inserido o valor de 44^{o}C. No item 4 pode-se verificar um erro de inversão de sinal apontado pelo método. O item 5 indica a temperatura média maior que a temperatura máxima. Os itens de 6 a 10 mostram percentuais de erro muito altos, devido em grande parte aos valores de códigos de erro inseridos indicando valores nulos (-99). Isto apenas indicou um registro com falhas sem a possibilidade de estimativa destes valores.

Tabela 2. Risco de aceitar um valor de temperatura incorreto acima de 100 %. Valores entre parênteses e em negrito representam as temperaturas incorretas.

	DATA	ESTAÇÃO	Risco	TMax	TMed	TMin
1	09/07/1965	2351008	-307	20	-99	13
2	07/10/1993	2351008	-662	-99	-99	9,2
3	14/8/1975	2350018	-636	27	**(44)** 18	-99
4	05/08/1982	2554026	-382	23	**(-26)**26	-99
5	23/11/1964	2550025	-583	23	37	-99
6	23/03/1973	2453023	-141	-99	20	-99
7	01/08/1982	2554026	-164	-99	28	-99
8	29/08/1989	2548070	-264	18	-99	-99
9	22/08/1965	2351008	-292	18	-99	5,8
10	31/08/1989	2548070	-300	27	-99	-99

Risco - Risco de se aceitar um valor de temperatura icorreto;
TMed – Temperatura Máxima
TE – Temperatura Média
TI – Temperatura Mínima

Conclui-se que o sistema desenvolvido possibilitou a correção eficiente e confiável do Banco de Dados Agrometeorológicos do IAPAR.

REFERÊNCIAS

Caramori, P. H., Gonçalves, S. L., Pugsley, L., Faria, R. T. Sistema de Redução de Riscos Climáticos para a Cultura de Trigo. In: XIII Congresso Brasileiro de Agrometeorologia, 2003, Santa Maria. Anais do XIII Congresso Brasileiro de Agrometeorologia. , 2003.

Cruvinel, P. E., Pugsley, L., Caramori, P. H. Modelagem para otimização de zonas de risco em sistemas agrícolas. In: Reunião Brasileira de Manejo e Conservação do solo e da Água, 2002, Cuiabá. Anais do Reunião Brasileira de Manejo e Conservação do solo e da Água. , 2002.

Duda, R. O. & Hart, P.E.. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.

Pugsley, L. Sistema para a tomada de decisão sobre zonas de risco agroclimático com técnicas do processamento de imagens digitais. São Carlos : Universidade Federal de São Carlos - Departamento de Computação, 2002 p.233.