CARTAS DE CONTROLE NÃO PARAMÉTRICAS APLICADAS A REGRA K NEAREST NEIGHBOR

MONICA MOURA PACHECO[1]; LEONARDO ROSA ROHDE[2]; ARIANE FERREIRA PORTO ROSA[3]

[1]*Universidade Federal de Pelotas – monicamp3@hotmail.com*
[2] *Universidade Federal de Pelotas – leonardo.rohde@live.com*
[3] *Universidade Federal de Pelotas – afprosa61@gmail.com*

1. INTRODUÇÃO

A detecção de falhas possui grande importância devido à necessidade de incrementar a eficiência do processo produtivo. As falhas podem ser identificadas como derivas que podem ser o deslocamento da média ou o aumento da dispersão de uma ou mais variáveis. O Controle Estatístico de Processos (CEP) possui uma vasta gama de ferramentas voltadas para detecção de falhas, tais como as cartas de controle. Segundo VERDIER e FERREIRA (2011) as cartas multivariadas são as mais utilizadas, sendo a mais conhecida a carta T^2 de Hotteling. Esta só é eficiente quando as variáveis seguem uma distribuição multinormal. Nos dados provenientes de processos industriais a suposição de multinormalidade não é satisfeita, isto dificulta ou impossibilita a correta identificação das derivas.

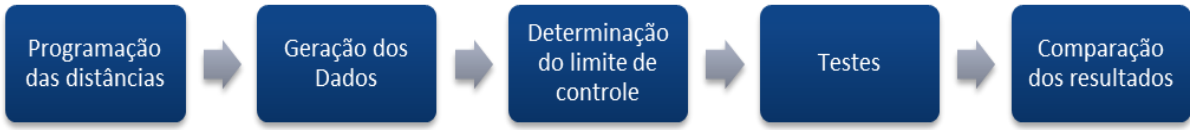
Os métodos estatísticos não paramétricos independem da distribuição de probabilidade dos dados. As cartas não paramétricas são uma opção quando a distribuição dos dados é desconhecida. BAÍLLO e CUEVAS (2003). HE e WANG (2007) desenvolveram uma carta não paramétrica baseada na regra K Nearest Neighbors rule (KNN). VERDIER e FERREIRA (2011) aperfeiçoaram a carta não paramétrica utilizando o KNN aplicando a distância de *Mahalanobis* de forma adaptativa.

A regra k-NND é um método não paramétrico que pode ser usado para a deteção de derivas/falhas. Este método que é baseado somente na amostra de aprendizagem. A principal vantagem deste método é sua simplicidade, assegurando um bom desempenho da detecção.

A maneira mais usual de calcular a distância entre dois pontos x e y no espaço n-dimensional é conhecida como distância euclidiana. O objetivo desse trabalho é comparar a eficiência na detecção de derivas na média de uma amostra multinormal utilizando diferentes distâncias aplicadas à regra KNN. Foram escolhidas as distâncias Manhattan, Canberra, Minkowski e Euclidiana.

2. METODOLOGIA

Esse trabalho é basedado nos projetos desenvolvidos por HE e WANG (2007) e VERDIER e FERREIRA (2011), onde a regra dos k vizinhos mais próximos foi aplicada a um modelo voltado a detecção de falhas. A pesquisa consistiu em aplicar a esse modelo novas distâncias para testar sua robustez na detecção de falhas em comparação com a distância euclidiana quadrada. O trabalho evoluiu conforme mostra o fluxograma abaixo:



O software Scilab foi utilizado para o desenvolvimento dos algoritmos. Utilizamos amostras de dados de duas variáveis da seguinte forma: 1) para cada uma das diferentes métricas é calculada a distância entre os pontos da amostra de treino composta de dados normais; 2) efetua-se a soma das distâncias dos k vizinhos mais próximos; 3) ao inserir uma nova amostra é calculada a distância entre cada ponto desta e da amostra de treino e também a soma das distâncias dos k vizinhos mais próximos de cada ponto; 4) então as soma das distâncias são plotadas num gráfico contra uma linha de controle horizontal, os dados acima dessa linha são classificados como fora de controle.

Os dados foram gerados com auxílio da função de dados randômicos do Scilab, seguindo uma distribuição multinormal. O primeiro grupo de dados gerado foi aquele utilizado como amostra de treino e tem como características média $[0 \quad 0]$ e matriz de covariância $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, foram geradas amostras de tamanho 125, 250 e 500 dados.

Segui-se então a definição do limite de controle, ele foi determinado utilizando a função Quantile, ela funciona ordenando o vetor com a soma das distâncias em ordem crescente e determinando um ponto de corte no qual a porcentagem determinada de dados fique abaixo do valor do quantil. Foram determinados dois quantis, o primeiro de 0.9 e o segundo de 0.95. Os testes e a comparação dos resultados obtidos serão discutidos na seção 3 Resultados e Discussão.

3. RESULTADOS E DISCUSSÃO

Os testes foram feitos utilizando amostras de tamanho N=250, 500 e 1000, das quais $N/_2$ eram compostas de dados normais e o restante de dados com derivas. Foram gerados dados com derivas na média para testar a capacidade de detecção de cada distância, os dados foram gerados seguindo as derivas da média contidas na tabela 1 e para cada deriva foram gerados três grupos de dados de tamanhos equivalentes à amostra de treino. O k foi determinado de forma não criteriosa e fixado como 10 e os quantis foram determinados com valores 0.9 e 0.95. Os resultados com amostras de tamanho 1000 foram os que apresentaram os melhores resultados visuais e a grande quantidade de dados facilitou a análise.

Tabela 1

D	Média
0	(0,0)
0,25	(0.176776, 0.176776)
0,5	(0.353553, 0.353553)
0,75	(0.530331, 0.530331)
1	(0.707106, 0.707106)
1,25	(0.883884, 0.883884)
1,5	(1.060661, 1.060661)
1,75	(1.237437, 1.237437)
2,00	(1.414214, 1.414214)
2,25	(1.590991, 1.590991)

2,5	(1.767767, 1.767767)
3,00	(2.121321, 2.121321)
4,00	(2.828428, 2.828428)

Tabela 2

			% Dados Detectados			
Derivas	Quantil	% Alarme Falso	Distância Euclidiana Quadrada	Distância Manhattan	Distância Canberra	Distância Minkowski, q=0,5
0,25	0.9	10%	10%	10%	16,40%	9,80%
	0.95	5%	5,80%	5,20%	9,20%	4,60%
0,5	0.9	10%	14,20%	13,60%	12,00%	13,00%
	0.95	5%	9,40%	9,20%	7,40%	8,60%
0,75	0.9	10%	18,20%	16,80%	10,40%	16,00%
	0.95	5%	11,20%	11,80%	5,60%	11,20%
1	0.9	10%	22,80%	29,40%	12,40%	20,40%
	0.95	5%	14,80%	15,20%	7,20%	14,00%
1,25	0.9	10%	29,20%	27,40%	10,20%	26,60%
	0.95	5%	20,60%	20,00%	5,60%	19,80%
1,5	0.9	10%	37,80%	36,00%	11,60%	35,40%
	0.95	5%	27,80%	28,20%	5,20%	26,80%
1,75	0.9	10%	43,40%	41,40%	12,60%	40,60%
	0.95	5%	34,40%	34,00%	4,80%	33,20%
2	0.9	10%	53,40%	52,20%	15,60%	51,00%
	0.95	5%	41,40%	42,00%	6,00%	40,80%
2,25	0.9	10%	64,40%	63,20%	18,80%	62,00%
	0.95	5%	54%	54,20%	6,60%	53,00%
2,5	0.9	10%	74,20%	73,60%	6,80%	72,20%
	0.95	5%	54%	64,60%	9,00%	63,40%
3	0.9	10%	85,20%	85,20%	40,40%	84,00%
	0.95	5%	79,40%	79,60%	14,40%	79,40%
4	0.9	10%	97,60%	97,60%	74,00%	97,40%
	0.95	5%	96%	96,40%	46,80%	96,60%

A tabela 2 apresenta os resultados dos testes com as distâncias. A Euclidiana foi ineficiente em detectar as derivas de 0,25, 0,5 e 0,75, pois em média 85,8% dos dados ficaram abaixo da linha do quantil de 0.9 e 90,6% do quantil de 0.95, para as derivas 1 até 1,75 o número de pontos plotados abaixo do quantil reduziu para 66,5% e 75,8%. Para derivas de valor 2, 2,25 e 2,5 a porcentagem cai para 35,6% para o quantil de 0.9 e 46% para o quantil de 0.95 o que sugere que o modelo foi capaz de detectar os desvios na média. Os desvios de 3 e 4 foram completamente detectados pelo modelo, para os dados com 3 desvios na média 97,8% dos pontos ficaram acima da linha de controle com quantil de 0.9 e 79,4% ficaram acima da linha do quantil de 0.95. Os dados com 4 desvios da média apresentaram 97,6% dos pontos acima da linha do quantil de 0.9 e 96% acima da linha do quantil 0.95.

Nos resultados da tabela 2 para a distância Manhattan vemos que para as derivas de 0,25 a 0,75 em média 84,8% não foram detecctados pelo quantil de 0.9 e 90,8% pelo quantil de 0.95. Os dados com derivas de 1 a 1,75 apresentaram 67,3% dos dados abaixo do quantil de 0.9 e 75,9% do quantil de 0.95, para os

dados com derivas de 2 a 2,5 apenas 36,8% ficaram abaixo do quantil de 0.9 e 45,8% do quantil de 0.95 o que mostra uma detecção de 73,2% e 55,2% dos dados com desvio. A amostra com 3 desvios foi identificada pelo modelo pois 85,2% dos dados com desvio foi identificada pelo quantil de 0.9 e 79,6% pelo quantil de 0.95, já na amostra com 4 desvios 97,6% foram detectadas pelo quantil de 0.9 e 96,4% pelo quantil de 0.95.

A tabela 2 mostra que a distância Canberra foi ineficaz na detecção dos desvios na média. Os dados com desvio ficaram em média 87,5% abaixo do quantil de 0.9 e 93,1% abaixo do quantil 0.95, o modelo detectou menos de 15% dos dados com desvio com o quantil de 0.9 e menos de 10% com o quantil 0.95, mesmo na amostra com 4 desvio foram identificadas apenas 74% dos dados na amostra com quantil 0.9 e 46,8% com quantil 0.95.

Quanto a distância Minkwski é necessário estabelecer o paramêtro q. Foi escolhido q=0,5. Nas amostras com desvios de 0,25 a 1,5 em média apenas 18,2% dos dados com desvio foram detectados pelo quantil 0.9 e 12,65 pelo quantil 0.95. Já nas amostras de 1,75 a 2,5 em média 56,5% foram identficadas pelo quantil 0.9 e 46,9% pelo quantil 0.95. Na amostra com 3 desvios na média 84% o quantil 0.9 detectou 84% dos dados com desvio e 79,4% pelo quantil 0.95, já com 4 desvios cerca de 97,4% foram percebidos pelo quantil 0.9 e 96,6% pelo quantil 0.95.

4. CONCLUSÕES

As comparações dos resultados mostraram que a distância canberra foi a menos eficaz na detecção dos desvios na média e mesmo os maiores desvio utilizados foram fracamente identificados em comparação com as demais distâncias. As distância Euclidiana Quadrada, Manhattan e Minkowski apresentaram um desempenho similar, as amostras com pequenos desvios nas média, de 0,25 a 0,5, não foram detectadas pelas três distâncias. As amostras com desvios de 1,75 a 2,5 foram relativamente identificadas pois para todas essas distâncias cerca de 50% foram detectados como falhas. Já as amostras com 3 e 4 desvios na média foram altamente detectadas pelas três métricas, pois a maioria dos dados com desvios ficaram acima da linha dos quantis.

Seria interessante desenvolver esse trabalho utilizando dados com diferentes distribuições e fazer a otimização do parâmetro k da regra KNN para determinar o k mais robusto na detecção de falhas, essas melhorias serão implementadas na segunda etapa da pesquisa.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BAÍLLO, A.; CUEVAS, A. Parametric versus Nonparametric Tolerance Regions in detection problems. Statistics and Econometrics Series 17, p.3-29, 2003.

HE, Q. P. and WANG, J. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes, IEEE Transactions on Semiconductor Manufacturing, vol. 20 (4), pp. 345–354, Nov. 2007.

VERDIER, G.; ROSA, A.F.P. Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing. IEEE Transactions on Semiconductor Manufacturing, v.24, n.1, p.59-68, 2011.