International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:5, 2007

Fuzzy Scan Method to Detect Clusters

Laureano Rodríguez, Gladys Casas, Ricardo Grau, and Yailen Martínez

***Abstract*—**The classical temporal scan statistic is often used to identify disease clusters. In recent years, this method has become as a very popular technique and its field of application has been notably increased. Many bioinformatic problems have been solved with this technique. In this paper a new scan fuzzy method is proposed. The behaviors of classic and fuzzy scan techniques are studied with simulated data. ROC curves are calculated, being demonstrated the superiority of the fuzzy scan technique.

***Keywords*—**Scan statistic, fuzzy scan, simulating study
.

I. Introduction

THE temporal scan statistics [1-5] have become a very popular method in disease surveillance for the detection of disease clusters. The standard approach is to look at a single disease or health outcome over the time, such as leukemia incidence [6] or breast cancer mortality [7].

The classical technique uses as input data, the day of diagnosis or the day of onset of the patients. The primary reason for using symptoms rather than diagnosed diseases is that it can take many tests and several days to establish a firm diagnosis [4].

Scan statistics have been traditionally used to scan both time and space for evidence of significant clusters of events. Recently, they have been receiving more attention as a method for genomic analysis [8].

In particular, scan statistics have been used within the field of molecular biology to identify chromosomal regions harboring a greater than expected number of restriction sites or clusters of transcription factor binding sites. Other authors proposed the use of a simple scan statistic for linkage studies to refine the search for new genes. Recent studies have extended the utilities of scan statistics to the analyses of genome-wide gene expression and SNP association by incorporating distance between genomic elements into the identification of significant genomic regions [8].

In 2005, Levin and colleagues have developed a model-based scan statistic that accounts for these aspects of the complex landscape of the human genome in the identification of extreme chromosomal regions of gene expression [9]. This method can be applied to gene expression data regardless of the microarray platform used to generate it. To demonstrate the accuracy and utility of this method, they applied it to a breast cancer gene expression dataset and tested its ability to predict regions containing medium to high level of DNA amplification with good results [9].

This finding strongly suggests that the model-based scan statistic and the expression characteristics of an increased chromosomal region of gene expression can be used to accurately predict chromosomal regions containing amplified genes [9].

In this paper, we present a modification of the classical temporal scan statistic. We obtain a new method: fuzzy scan statistic. We realize an intensive simulating study in order to prove the superiority of the fuzzy technique.

The study is structured as follows: Section 2 shows the main details of a classic scan method whereas Section 3 presents our basic new ideas of a generalized scan technique. The bases of the simulating study are the subject of Section 4. The experiments carried out to prove the behavior of classic scan method is the topic of Section 5. The new technique proposed and the results of the feasibility of our contribution are the topic of Sections 6 and 7. Some graphics are also available in order to summarize our results. Finally, a validation and some conclusions and comments are provided.

II. The Classic Scan Technique

Scan statistics are commonly used to investigate the presence of at least one cluster, inside a larger period of time. Known in the literature as "moving window analysis", the idea is to scan a small fixed window of length t over the data, calculating some local statistic (generally the number of events) for each window. The maximum of these locality statistics is known as the scan statistic, and it will be denoted by hmax.

Under some specified homogeneity null hypothesis H0 on X (Poisson point process) the approach entails specification of a critical value $C\alpha$ such that $P(\eta \geq C_\alpha) = \alpha$. If the observed maximum is larger or equal to $C\alpha$, then we can infer that there is at least one cluster, which is a local region with statistically significant signal [10].

Analysis of the univariate scan process has been considered by many authors, including [5, 11, 12]. For a few simple models, exact p-values are available; but they can not be applied to big databases. Many applications require approximations to the p-value. In 1982 Naus published an

F. A. Universidad Central de Las Villas and Instituto Superior de Ciencias Médicas de Sancti Spiritus. Phone: 53 42 281515; e-mail: corvea@ uclv.edu..

S. B. Universidad Central de Las Villas. Phone: 53 42 281515; e-mail: gladita@ uclv.edu

T. C. Universidad Central de Las Villas. Phone: 53 42 281515; e-mail: rgrau@ uclv.edu

The last. Universidad Central de Las Villas. Phone: 53 42 281515; e-mail: yailenm@ uclv.edu

approximate formula to calculate p-value significance, based in exact results [13]. The next formulas summarize its fundamental aspects:

Let: t: Window length,

T: Total time considered,

λ: Expected number of points per unit time in the Poisson process,

$w_{y,y+t}$: Number of cases in the [y, y+t) interval

L =T / t

$$p = P^*(w, \lambda L, 1/L) = 1 - Q^*(w, \lambda L, 1/L)$$

where Q^* can be approximately calculated for any $L > 2$

$$Q^*(w,\lambda L,1/L) \approx Q^*(w,2\lambda,1/2)\left[Q^*(w,3\lambda,1/3)/Q^*(w,2\lambda,1/2)\right]^{L-2}$$

For w>2, $p_i = e^{-\lambda}\lambda^i/i!$, $F_w = \sum_{i=0}^{w} p_i$, $\lambda > 0$:

$$Q^*(w,2\lambda,1/2) = F_{w-1}^2 - (w-1)p_w\, p_{w-2} - (w-1-\lambda)p_w\, F_{w-3}$$

$$Q^*(w,3\lambda,1/3) = F_{w-1}^3 - A_1 + A_2 + A_3 - A_4$$

Where:

$$A_1 = 2\,p_w\, F_{w-1}\left((w-1)F_{w-2} - \lambda\, F_{w-3}\right)$$

$$A_2 = 0.5\,p_w^2\left((w-1)(w-2)F_{w-3} - 2(w-2)\lambda\, F_{w-4} + \lambda^2\, F_{w-5}\right)$$

$$A_3 = \sum_{r=1}^{w-1} p_{2w-r}\, F_{r-1}^2$$

$$A_4 = \sum_{r=2}^{w-1} p_{2w-r}\, p_r\left((r-1)F_{r-2} - \lambda\, F_{r-3}\right)$$

Being $F_i = 0$ for all i<0.

As can be seen, all these formulas use probability and cumulative Poisson distribution. We want to notice that Poisson distribution is discrete, that is, it is defined only for integer values, see Table I.

III. Generalizing the Scan Technique

The classic scan technique can be modified in order to apply it to other problems. The main idea consists of transforming the sequence input data into a binary sequence. The number one will represent the interest category and the zero value will correspond to the other categories.

For example, the repeats detection of some substring inside a DNA sequence is a classical problem in bioinformatic. Suppose it is necessary to detect repeats of "gcg". Then, the "gcg" substring will be replaced by "one", and the other letters will be replaced by "zero", see Table I, [14, 15].

TABLE I Binary transformation of a DNA sequence

	Code
Sequence	**...ccccagtctga gcg gcg atg gcg gcg gcg gcagcagca...**
Transformation	...00000000000 1 1 000 1 1 1 000000000

This kind of transformation offers a wide range of applications for the Scan technique, [14, 15].

IV. Bases of the Simulation Study

The purpose of this section is to determine the capacity of the test to detect actual clusters. Scan method capacity for detect clusters depends on the ability of the user to select the correct value for the length of the moving window. A bad choice of this parameter can hide true clusters or show putative ones.

Two different kinds of datasets were generated: one for true clusters and other for the wrong ones. 1000 different sequences were generated for each kind of dataset, in order to be used as input data for both scan techniques. The capacity of compute right responses was calculated using a significance level of 0.05. Graphical results are shown and discussed in Section 5.

A. Generating True Clusters Dataset

Binary sequences were generated according to the following principles:

The first third part of the sequence was generated according to a Bernoulli distribution with 0.2 parameter. The number of ones inside the sequence will be small.

The second part was generated according with another Bernoulli distribution. This time the parameter was higher: 0.8 in order to obtain a significant increment of the number of ones. Doing this, we assure that there is, at least, one cluster.

The third part of the sequence was generated in the same way that the first one.

Besides, in order to characterize the behavior of the test, we generate sequences of different lengths, from 50 to 200 cases.

B. Generating False Clusters Dataset

We generate a binary sequence according to a Bernoulli distribution with 0.3 as parameter value. The number of ones will be dispersed in the whole string without grouping itself.

We also generate sequences of different lengths: 50, 100 and 200 cases.

V. Bases of the Simulation Study

The X-axis of Fig. 1 to 3 represents all possible values of the windows length. In Fig. 1, values change between 1 and 50, in Fig 2, between 1 and 100 and in Fig. 3, between 1 and 200. In all cases, continuous line shows the results of the classic scan method for true clusters and the dashed line shows the same results for false clusters.

As can be seen, classic scan method is not quite good for detecting true clusters when the sequence length is either too small or too large. In a small population (size 50), when the windows size is between 9 and 14, better results were obtained. In these cases the capacity of detection is superior to 50%, Fig. 1 shows the results.

False databases clusters were also presented as input sequences to classic and fuzzy techniques. In all cases, the wrong classification percent was 0. Therefore, we can assure that, for the methods' accuracy classification it is only important the capacity of detect true clusters.

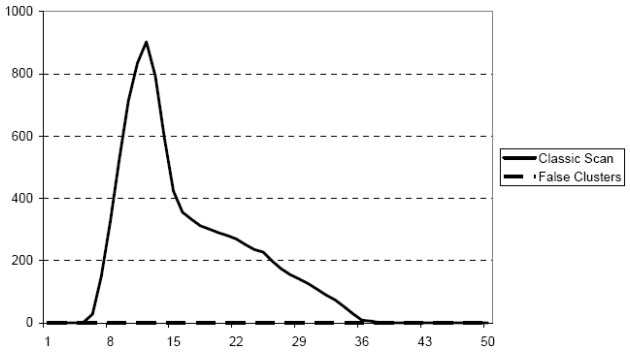

Fig. 1 Results of classic scan with 50 cases length sequences and True Clusters

When the sequence's length is increased, true cluster's detection capacity of both methods are also increased. Fig. 2 and 3 show this graphically.

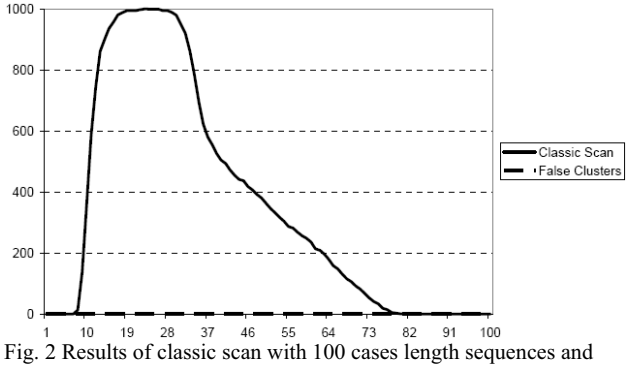

Fig. 2 Results of classic scan with 100 cases length sequences and True Clusters

Fig. 3 Results of classic scan with 200 cases length sequences and True Clusters

VI. The Fuzzy Scan Technique

In this section we propose a modification to the classic scan method in order to obtain better results in the cluster detection process. The idea is to change the fixed length of the moving window by a fuzzy length with a certain membership function in both extremes of the interval.

The membership function is a graphical representation of the magnitude of participation of each value. It associates a weighting with each of the inputs that are processed in the extremes of the moving window. Fuzzy scan technique uses the membership values as weighting factors to determine their influence on the fuzzy output number of cases detected in the interval. A triangular membership function was used in both extremes of the moving window.

The mathematical formulation of the test is essentially the same: the method scans the same data using a fuzzy moving window. Being the scanning window now fuzzy, the maximum number of cases reported in a window, (that is the scan statistic hmax) will be a real number, not an integer like in classic method. Then, we have a problem with the compute of the p-value, because Poisson distribution is only defined for integer values.

An interpolating function solved our problem. In general terms, interpolation is a method of constructing new data points from a discrete set of known data points. In our case, the set of known data is the probability Poisson distribution data or the cumulative Poisson distribution data, see Table I. With these datasets we build other two functions: interpolating probability Poisson distribution and interpolating cumulative Poisson distribution see Fig. 4.

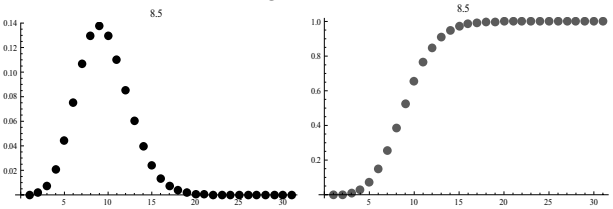

Fig. 4 Probability and cumulative Poisson distribution

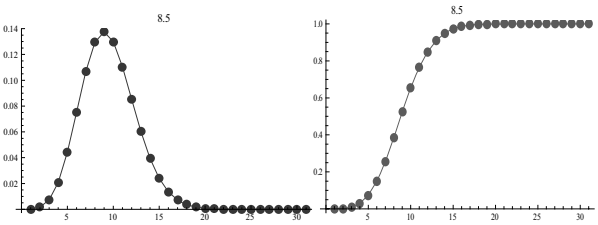

Fig. 5 Interpolating function for probability and cumulative

Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points. Then, when we evaluated the interpolating functions in integer points we obtained exactly the same value as in the original Poisson functions. We used interpolating polynomial of four degree, see Fig. 5.

Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points. Then, when we evaluated the interpolating functions in integer points we obtained exactly the same value as in the original Poisson functions. We used interpolating polynomial of four degree.

VII. Bases of the Simulation Study

Show graphically the results of classic and fuzzy scan techniques with true clusters datasets. In all cases, continuous line corresponds to the classic scan method and the dashed line shows the results of the fuzzy scan statistics.

As can be seen, the results of the fuzzy technique are better than those obtained by classic technique for small window length values, see Fig. 6 to 8. Therefore, it increases the range

in which the method capacity of detection is superior to 60%.

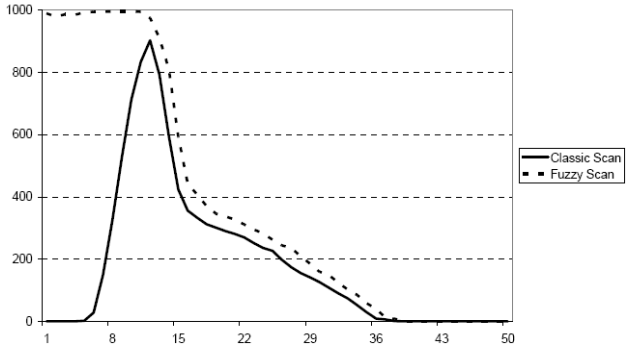

Fig. 6 Results of classic and fuzzy scan with 50 cases length sequences and True Clusters

When the sequence's length is increased, true cluster's detection capacity of both methods are also increased, see Fig. 7 and 8.

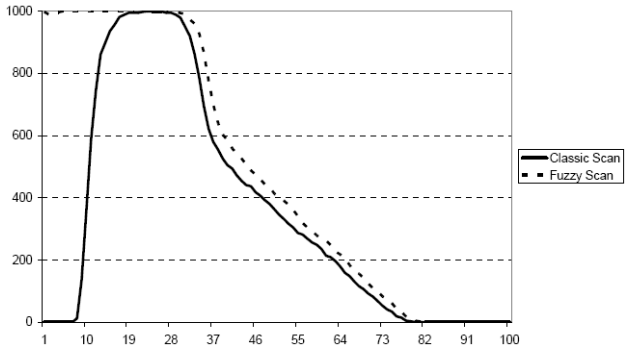

Fig. 7 Results of classic and fuzzy scan with 100 cases length sequences and True Clusters

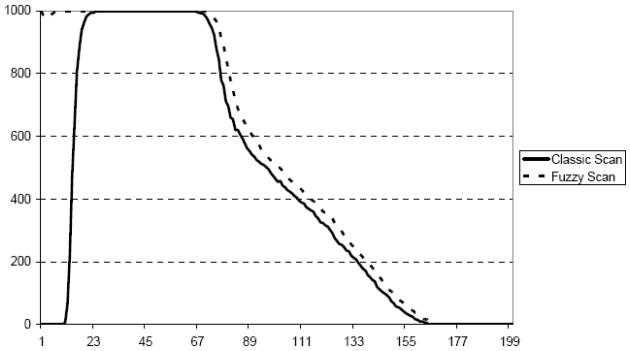

Fig. 8 Results of classic and fuzzy scan with 200 cases length sequences and True Clusters

False databases clusters were also presented as input sequences to classic and fuzzy techniques. In all cases, the wrong classification percent was zero. Therefore, we can assure that, for the methods' accuracy classification it is only important the capacity of detect true clusters.

VIII. Validation and Discussion

The detection of clusters using scan technique can be seen as a classification problem. Given a binary sequence of length n, it is necessary to determine if there is at least a cluster of 1 or not. Two different classifiers are shown: the classic scan technique and the fuzzy one.

The ROC Curve procedure provides a useful way to evaluate the performance of classification schemes that categorize cases into one of two groups. The area under the curve represents the probability that the cluster detection scan result for a randomly chosen positive case will exceed the result for a randomly chosen negative case.

The area under the ROC curve was calculated for both scan techniques: classic and fuzzy. Results are shown in Table II.

Table II Classification results

Method	Length sequence	Area under ROC curve	Asymptotic signification
Classic	50	0.840	0.00
	100	0.870	0.00
	200	0.885	0.00
Fuzzy	50	0.880	0.00
	100	0.915	0.00
	200	0.923	0.00

As can be seen, the area under ROC curves is bigger in fuzzy scan techniques compared to the equivalent area in classic scan technique.

The asymptotic significance is less than 0.05 in all cases, which means that using the scan method for clustering detection is better than guessing.

Besides, in both methods there is an obvious relationship between the sequence length and the window length to obtain correct results for true clusters.

For small sequences (length 50), Fig. 6 shows that the correct values of the windows length for classic scan technique must be between 10 and 13 in order to guarantee a positive detection rate superior to 60%. In the fuzzy scan method, the value of the parameter must be between 1 (the lowest possible value) and 14, see also Fig. 6.

The same analysis can be done for the other simulated sequences. The range of the windows length values detecting true clusters is wider in the fuzzy scan technique in comparison with the range of the equivalent classic method. Therefore, fuzzy scan technique produces better results than the classic one.

IX. Conclusion

Fuzzy method increases the capacity of classic scan technique to detect true clusters. Fuzzy technique solves the problem of the small values for the moving window, and increases the detection capacity of larger ones.

However, in general terms, the values of the window length must be smaller than the half of the length of the sequence in order to achieve correct results. Generally, bioinformatic applications are based for instance, in the analysis of big DNA sequences, then, this constraint will not be a real problem.

REFERENCES

[1] J. Glaz, and N. Balakrishnan, *Scan Statistics and Applications*, 1999.
[2] J. Glaz, J. Naus, and S. Wallenstein, "Scan Statistics," Springer Verlag, 2001, pp. 370.
[3] M. Kulldorff, "A spatial scan statistic. Communications in Statistics," *Theory and Methods*, vol. 26, pp. 1481–1496. 1997.
[4] M. Kulldorff, F. Mostashari, L. Duczmal, K. Yih, K. Kleinman, et al., "Multivariate scan statistics for disease surveillance," *Statistics in Medicine*, vol. vol. 26, No. 8. 2007.
[5] J. Naus, "Clustering of random points in two dimensions.," *Biometrika*, vol. 52, pp. 263–267. 1965.
[6] U. Hjalmars, M. Kulldorff, G. Gustafsson, and N. Nagarwalla, "Childhood Leukaemia in Sweden: Using GIS and a Spatial Scan Statistic for Cluster Detection," *Statistics in Medicine*, vol. 15 pp. 707 - 715. 1998.
[7] C.E. Hsu, H. Jacobson, and F. Soto, "Evaluating the disparity of female breast cancer mortality among racial groups - a spatiotemporal analysis," *International Journal of Health Geographics*, vol. 3 (4), pp. 1-11. 2004.
[8] Y.V. Sun, D.M. Jacobsen, and S.L. Kardia, "ChromoScan: a scan statistic application for identifying chromosomal regions in genomic studies.," *Bioinformatics*, vol. 22. (23), pp. 2945-2947. 2006.
[9] A.M. Levin, D. Ghosh, K.R. Cho, and S.L. Kardia, "A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors," *Bioinformatics*, vol. 21, pp. 2867-2874. 2005.
[10] C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, "Scan Statistics on Enron Graphs," *Book Scan Statistics on Enron Graphs*, Series Scan Statistics on Enron Graphs, ed., Editor ed.^eds., 2005, pp.
[11] N. Cressie, "On Some Properties of the Scan Statistic on the Circle and the Line " *Journal of Applied Probability*, vol. 14, pp. 272-283. 1997.
[12] C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, "Scan Statistics on Enron Graphs " *Computational & Mathematical Organization Theory*, vol. 11, pp. 229-247. 2005.
[13] J.I. Naus, "Approximations for distributions of Scan statistics," *Journal of the American Statistical Association*, vol. 77, pp. 177-183, Mar. 1982.
[14] L. Rodríguez, G. Casas, R. Grau, G. Cardoso, S. Ortega, et al., "Scan Statistics. Bioinformatics Applications," in *First International Workshop on Bioinformatics Cuba-Flanders*, City.
[15] L. Rodríguez, G. Casas, R. Grau, and M. Pupo, "Generalización de dos métodos de detección de conglomerados. Aplicaciones en Bioinformática.," *Revista de Matemática: Teoría y Aplicaciones.*, vol. 15 (1), pp. 27 - 40. 2008.