Búsquedas de fuentes puntuales con el algoritmo EM en el telescopio de neutrinos ANTARES

J.A. Aguilar[1], E. Carmona[1], J.J. Hernández[1], F. Salesa[1], J. Zúñiga[1]

[1] IFIC.- Instituto de Física Corpuscular. CSIC – U. de Valencia. Apdo. 22085, E-46071 Valencia

I. INTRODUCCIÓN

La colaboración ANTARES ha empezado la construcción de un telescopio de neutrinos en el fondo del mar Mediterráneo. El detector completo estará totalmente desplegado y operativo para finales del 2006. Uno de los principales objetivos del experimento es la búsqueda de fuentes de neutrinos de alta energía. Entre los posibles emisores de neutrinos podemos encontrar Núcleos Galácticos Activos (AGNs), explosiones de Rayos Gamma (GRBs), Remanentes de Supernovas (SNRs), etc.

La detección de fuentes puntuales se basa en la identificación de agrupamientos (*clusters*) de sucesos sobre un fondo formado principalmente por los neutrinos atmosféricos y neutrinos provenientes de todas las fuentes distribuidas en el espacio (flujo difuso). En este marco, hemos desarrollados algoritmos basados en métodos generales de *clustering* o lo que es lo mismo, la identificación de grupos de observaciones cohesivas y diferenciadas de otros grupos. El análisis de *clustering* es un problema recurrente e interdisciplinar que se presenta en muy diversas áreas, como por ejemplo dentro de la industria textil en la búsqueda de imperfecciones en el tejido[1]. Así pues, diversas técnicas han sido desarrolladas durante décadas para este tipo de análisis. En el caso de ANTARES hemos adaptado dos de los métodos más usados a las peculiaridades de la búsqueda de fuentes puntuales; el algoritmo Nearest-Neighbour (NN) y el Expectation-Maximization (EM).

II. ALGORITMO NEAREST-NEIGHBOUR

El algoritmo Nearest-Neighbour (NN) es una de las técnicas llamadas aglomerativas. Una de las desventajas de este tipo de técnicas es que carecen de una estadística asociada y no pueden proporcionar un criterio de selección. El NN es uno de los más antiguos y se define fácilmente de la siguiente manera; dos objetos **a** y **b** pertenecen a un mismo cluster **a,** a un nivel d, si existe una cadena de objetos intermedios $i_1, \ldots, i_m$ uniéndolos de manera que las distancias entre objetos cumplen:

$$d_{i_k, i_{k+1}} \leq d \qquad \text{para } k = 0, \ldots, m-1 \qquad [1]$$

donde $i_0 = \mathbf{a}$ y $i_m = \mathbf{b}$.

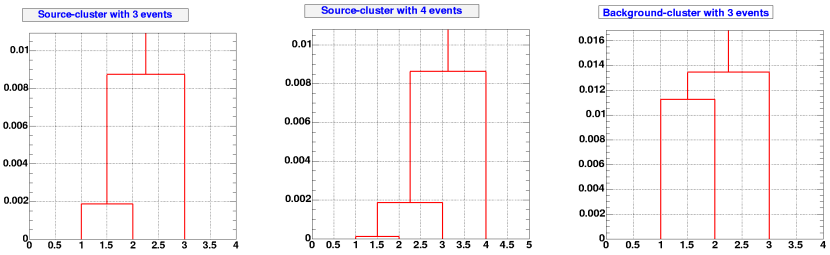

Figura 1. Izquierda: Dendrograma de un *cluster* de una fuente de 3 sucesos. Centro: Dendrograma para una fuente de 4 sucesos. Derecha: Dendrograma para un *cluster* de 3 sucesos debido al fondo

El NN permite obtener como primer resultado lo que se conoce como *dendrogramas*. Estos dendrogramas son árboles anidados donde se representa a qué niveles los elementos de un *cluster* se han unido. En la figura 1 se pueden ver tres distintos dendrogramas para tres distintos tipos de *clusters*. El NN es utilizado en nuestro caso para proporcionar las parámetros iniciales al algoritmo EM, el cual comenzará un bucle de maximización usando como primera estimación las coordenadas asignadas por el NN.

III. ALGORITMO EXPECTATION-MAXIMIZATION

El algoritmo EM es uno de los más usados dentro del análisis de *clustering*[2]. Es un método para maximizar analíticamente la verosimilitud de distribuciones de probabilidad cuyo único enfoque puede ser la maximización numérica. El procedimiento general se basa en suponer que en realidad la muestra observada es una versión *incompleta* de otra muestra más general, donde una coordenada extra indica a qué grupo pertenece un determinado suceso. Una vez construida una nueva distribución de probabilidad para el conjunto *completo* de datos, podemos en un primer paso **estimar** el valor esperado de la verosimilitud de dicho conjunto *completo* condicionado por los datos observados (conjunto *incompleto*). Una vez calculado el valor esperado, el siguiente paso es **maximizar** dicho valor, el método garantiza que los parámetros que maximizan el valor esperado son los mismos que maximizan la verosimilitud de la muestra original (*incompleta*).
Como criterio para confirmar o descartar la existencia de uno o varios agrupamientos (fuentes puntuales) se usa el factor de Bayes aproximado por el *Bayesian Information Criterion* o BIC.

IV. RESULTADOS

Los resultados obtenidos son para un estudio Monte Carlo de un año de toma de datos, y en ellos se ha calculado la probabilidad de obtener una fuente en dos declinaciones distintas para un nivel de confianza de 3σ. La probabilidad se presenta en función del número de sucesos N_{source} emitidos por la fuente en dos posiciones distintas. Se puede comprobar que para declinaciones medias, una fuente emitiendo 5 sucesos por año, tiene una probabilidad de ser detectada mayor del 60 %. Otros resultados para distintas declinaciones y mayor tiempo de exposición están también siendo estudiados así como la sensibilidad a distintos índices espectrales de la fuente emisora.

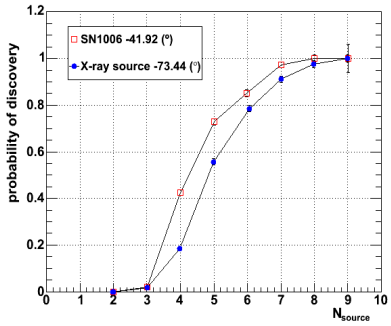

Figura 2. Probabilidad para una fuente emitiendo N_{source} sucesos en dos declinaciones distintas

Referencias

[1] J.G. Campbell et al, *Linear flaw detection in woven textiles using model-based clustering*, Pattern Recognition Letters, 1997. 18:1539-1548.

[2] A.P. Dempster et al, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society Series B, 1977. 39:1-18