Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. XXX, 2006
R. Shaw, F. Hill and D. Bell, eds.

Bayesian Belief Networks for Astronomical Object Recognition and Classification in CTI-II

Mike Ritthaler, George Luger and Robert Young

Computer Science Department, University of New Mexico, Albuquerque, NM, 87131, Email: mike.ritthaler@gmail.com

John McGraw and Pete Zimmer

Physics and Astronomy Department, University of New Mexico, Albuquerque, NM 87131

Abstract. The University of New Mexico (UNM) is currently designing and building the CCD Transit Instrument II (CTI-II)(McGraw et al., 2006), a 1.8m transit survey telescope. The stationary CTI-II uses the time delay and integrate readout mode for a mosaic of CCDs to generate over 100 gigapixels per night which is required to be analyzed within a day of observation. We are attempting to develop robust machine learning techniques that use multiple scientific and engineering data streams to classify both objects within an image frame, and the image frame itself. We propose the use of Bayesian belief nets as both classifiers and as tools to integrate and explore the data streams. This initial report explores the use of Bayesian networks as source/noise separators.

1. Introduction

The goals of the CTI-II project revolve around ground-based millimagnitude photometry and milliarcsecond astrometry sustained under a wide variety of conditions. This involves the use of multiple engineering streams of data from cloud monitors, various optical and structural monitoring instruments, as well as LIDAR and cameras to measure atmospheric extinction(Dawsey et al., 2006). We wish to create systems which can learn and adapt how the scientific data streams are processed or interpreted based on the conditions presented by the engineering data streams. We present models using Bayesian networks as the integration tool to provide for source/noise separation under a wide variety of seeing and sky-brightness conditions.

2. Bayesian networks

Bayesian networks are directed acyclic graph (DAG) representations across distributions of discrete or continuous random variables: $X = \{X_1, X_2, ..., X_N\}$. The resulting graph G is a unique resentation of a joint probability distribution across the set of random variables. The topology of the graph gives the independence relations of the variables according to the Markov condition: any node is conditionally independent of its nondescendents, given its parents. Both the conditional probabilities and the topology of the net can be learned or determined by a domain expert. Given a set of values across some of the random variables, the values of the other variables may be inferred using algorithms such as junction trees, Gibbs sampling, or Pearl's algorithm (Heckerman, 1999).

Bayesian networks provide a natural way of reasoning in uncertainty, or without access to the full set of variables. In a data set of random variables, Bayesian net structure learning can indicate relationships between variables (Pearl & Verma, 1991). Their ability to inference can provide deep insight into how the network processes information.

3. Data sets and processing

We are currently modeling the sky using a source catalog generated by the STUFF(Bertin & Fouqué, 2006) and SKYMAKER(Bertin, 2005) programs, with SKYMAKER generating images using the parameters for the CTI-II telescope. To generate a large dynamic range of seeing and background brightness under controlled conditions, we use a grid of 25 frames where the seeing varies from 0.7 to 1.9 arcsec, and the background brightness varies from magnitude 22.5 to magnitude 19.7.

Since each frame is synthetic, we assume it is the equivalent of a flat-fielded and debiased telescope frame. The mean and standard deviation of the background are calculated from the frame, and the seeing is calculated from the full-width at half-magnitude (FWHM) of the brightest object in the catalog. All groups of contiguous pixels 2.5σ above the background are located and put into a list of possible sources. A size filter is then run across the possible source list, filtering out all single-pixel sources. For the remaining sources, the centroid is calculated and rounded to the nearest pixel. The values of the surrounding 25 pixels (a 5-by-5 window) including the centroid pixel are put in a 25 dimension vector P. From P a normalized vector P_N is calculated: $P_N = \frac{P}{|P|}$. The values in P_N are then binned to discrete values for use in the network. The values for seeing and background are also binned, as is a value for the number of pixels in the source. The network has nodes for each pixel value in P_N as well as nodes for the background and seeing. The final node represents the class of the source, either noise, galaxy or star.

A number of different network topologies were tested, most using the Bayesian network as a naive Bayesian (NB) classifier with the pixel nodes as the attributes of the class node, conditioned on the seeing or the background brightness. Some added in the size node as an additional attribute (see Figure 1, NB-BZP). In naive Bayesian topologies the attributes are considered conditionally independent of each other and no arcs are allowed between them in the graph. Other topologies tested were tree-augmented naive Bayes (TANB), which allows arcs and conditional dependence between the attributes, as well as topologies built using a global Monte Carlo structure search algorithm across the whole data set, where there are no restrictions on the arcs between variables other than the DAG requirement(Cheng & Greiner, 1999; Friedman & Goldszmidt, 1996).

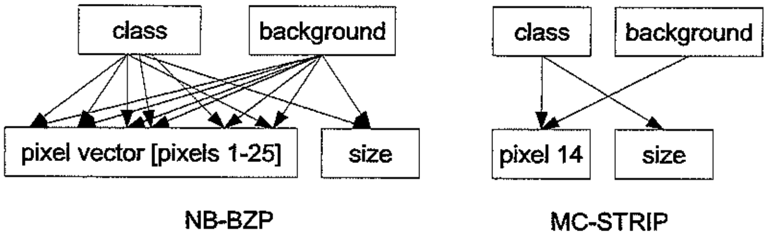

Figure 1. The best performing network, NB-BZP and the intriguing MC-STRIP network, which uses only four variables.

4. Results

Table 1. Results from various network topologies.

Network	False negative	False positive	Unknown	Total error score
NB-BZP	17.1%	2.3%	0.5%	19.9%
NB-BP	17.8%	3.2%	0.5%	21.5%
MC-STRIP	15.8%	6.3%	0.0%	22.1%
MC-BSZP	15.4%	6.0%	3.1%	24.5%
NB-BSP	13.9%	7.0%	4.0%	24.9%
NB-SP	18.2%	7.0%	0.4%	25.6%
MC-BZP	19.3%	4.3%	2.5%	26.1%
TANB-BP	5.7%	6.0%	39.5%	51.2%

[a] Topologies: MC=Monte Carlo, NB=Naive Bayes, TANB=Tree Augmented Naive Bayes, STRIP=4 node classifier

[b] Nodes: B=background, S=seeing, Z=size, P=Pixels

[c] All scores are a percentage of the entire test set, the unknown category represents cases where the network couldn't determine a class due to lack of information

5. Discussion

There are a number of suprising results from this research. Intially, we thought that the seeing node would give good information to the network about point spread function change. However for our synthetic data set, such information was far less helpful than the measure of the background brightness. Since the pixel vector is normalized to unit length, perhaps the brightness information is needed to supplement the absolute measure lost in the normalization. If the normalization was not done, perhaps the brightness variable information would not be needed. An even more interesting result is the score of the network MC-STRIP in Table 1. This network only contains 4 nodes: class, brightness, size and a node representing the value of a pixel next to the centroid pixel (see Figure 1). This network is within 2% of the performance of the two best peforming networks NB-BZP(see Figure 1) and NB-BP, using only 4 variables compared to their 29 and 28 variables respectively. It is important to point out that the structure of MC-STRIP was learned, while NB-BZP and NB-BP are both naive Bayes structures. As shown by the results of the TANB-BP (see Table 1), the tree augmented naive Bayes structures suffer from a combinatorial explosion, where some nodes require too many training cases to be effective, and thus the network has a high number of 'unknown' errors in testing.

6. Future work

Our next step is increasing the size of our test data sets and moving from synthetic to real data to get a wider variety of objects and conditions. We are also exploring the use of self-organizing machines to replace simple histogramming in discretizing the scientific and engineering data.(Young, et al. 2006). Since CTI-II will operate in multiple bands and multiple nights, this raises the possibility of networks with nodes representing data from previous nights and multiple wavelengths. Finally, we wish to model the integration of more information from the engineering data of the telescope.

Acknowledgments. This work has been kindly supported by UNM Press, the UNM Physics and Astronomy and Computer Science departments, as well as a grant from the UNM Graduate and Professional Student Association.

References

Bertin, E., 2005, STUFF V1.11, ftp://ftp.iap.fr/pub/from_users/bertin/stuff

Bertin, E. & Fouqué, P., 2006, SKYMAKER V3.03, http://terapix.iap.fr/

Cheng, J. & Greiner, R. 1999, in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 101-108

Dawsey, M., Gimmestad, G., Roberts, D., McGraw, J., Zimmer, P. & Fitch, J. 2006, Proceedings of SPIE, 6270, 62701F

Friedman, N. & Goldszmidt, M. 1996, in Proceedings AAAI-96 Thirteenth National Conference on Artificial Intelligence, 1277-1289

Heckerman, D. 1999, in Learning in Graphical Models, (Cambridge, Massachusetts: MIT Press), 310-354

McGraw, J. T., Ackermann, M. R., Gerstle, W. H., Williams, T., & Zimmer, P.C. 2006, Proceedings of SPIE, 6267, 62673T

Pearl, J. & Verma T. 1991 in Principles of Knowledge Representation and Reasoning Proceedings of the Second International Conference, ed. J. F. Allen, R. Fikes & E. Sandewall, 441-452

Young, R., Caudell, T., Ritthaler, M., McGraw, J. & Zimmer, P. 2006 in ASP Conf. Ser., Vol. 125, ADASS VI, ed. G. Hunt & H. E. Payne (San Francisco: ASP), [P1.03]