1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 30 Sept 1997	3. REPORT TYPE AND DATES COVERED Final Report(15 Aug 93-30 Sept 97)

4. TITLE AND SUBTITLE

Detection and Classification of Synthetic Aperture Radar Targets

5. FUNDING NUMBERS

G#F49620-93-1-0492

AFOSR-TR 97-0615

6. AUTHOR(S)

K. C. Chang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center of Excellence in C3I
George Mason University
Fairfax, VA 22030

8. PERFORMING ORGANIZATION REPORT NUMBER

5-25707-97-1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NM, Sponsoring Scientific Office
Dr. Jon A. Sjogren
110 Duncan Ave., Rm. #B115, Bolling AFB, DC 20332-8030

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

T *(Maximum 200 words)*

This final report described the ASSERT project "Detection and Classification of Synthetic Aperture Radar Targets" associated with the URI Automatic Target Recognition (ATR) project sponsored by DARPA. The main goal of this ASSERT project together with the URI-ATR project is to develop detection and classification algorithms for automatic target recognition. For the ASSERT project, we have focused on the use of Bayesian probabilistic reasoning approach to fuse multiple target feature data for the purpose of target classification. We also developed Bayesian network learning algorithms to automatically construct the Bayesian network model.

In this project, there were two graduate students and one undergraduate students participated in the technical work. Of whom, two of them have received M.S. degrees and one of them is continuing his Ph.D. degree. This project directly or indirectly supported the publications of eight technical papers, two Master thesis, one Ph.D. thesis, and one technical report.

14. SUBJECT TERMS

Synthetic Aperture Radar, Automatic Target Recognition
Bayesian Networks

15. NUMBER OF PAGES
36

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Non-classified	18. SECURITY CLASSIFICATION OF THIS PAGE Non-classified	19. SECURITY CLASSIFICATION OF ABSTRACT Non-classified	20. LIMITATION OF ABSTRACT

Detection and Classification of Synthetic Aperture Radar Targets

Final Report

Grant #F49620-93-1-0492

Principal Investigator

Dr. K. C. Chang

September 30, 1997
George Mason University
Center for Excellence in C3I
Fairfax, Virginia

SUMMARY

This final report described the ASSERT project "Detection and Classification of Synthetic Aperture Radar Targets" associated with the URI Automatic Target Recognition (ATR) project sponsored by DARPA. The main goal of this ASSERT project together with the URI-ATR project is to develop detection and classification algorithms for automatic target recognition. For the ASSERT project, we have focused on the use of Bayesian probabilistic reasoning approach to fuse multiple target feature data for the purpose of target classification. We also developed Bayesian network learning algorithms to automatically construct the Bayesian network model.

In this project, there were two graduate students and one undergraduate students participated in the technical work. Of whom, two of them have received M.S. degrees and one of them is continuing his Ph.D. degree. This project directly or indirectly supported the publications of eight technical papers, two Master thesis, one Ph.D. thesis, and one technical report.

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

1. Introduction

To succeed on the battlefield, it is very important to have an accurate picture of the current tactical picture of the situation. Modern sensor technology such as SAR has vastly improved the quantity and quality of the raw information on which tactical intelligence is based. SAR signatures contain coherent noise and have many unknown parameters such as amplitude, target spatial position and target orientation. The SAR clutter statistics are at least somewhat unknown and the signal is nonstationary. Our research focus is to build on our development of new detection and classification algorithms for SAR target data. The major goal of this research was to develop Bayesian Network algorithms appropriate for SAR ATD/R and to demonstrate the performance of these algorithms.

Problem - A high performance SAR target detection and classification system will make use of multiple features and algorithms. The system requires an approach for combining feature and algorithm output information to make intermediate and final decisions about the presence and identities of targets.

Goal - Develop Bayesian Network and other algorithms appropriate for the SAR ATR problem that can handle non-Gaussian features optimally. Demonstrate this algorithm against the SAR ATR problem.

Approach & Objectives - We have worked on the problem of target discrimination and classification using a Bayesian Network whose input is a collection of target features. These target features include the features of the Lincoln Laboratory discriminator. We are have also examined other applicable approaches such as multi-polarization fusion as well as multi-resolution fusion using wavelet.

2. Technical Approach

The objectives of the current research in ATR are to determine techniques for understanding the nature and special features of a SAR image and use those to develop specific identification techniques for classification. Particularly, we are interested in using the Bayesian network technology to improve ATR performance.

During the past few years, Bayesian networks have received much attention as an efficient way of reasoning under uncertainty. Such networks provide a probabilistic model of the problem by means of graphs. The nodes correspond to the variables of interest, the states in a node can be either discrete-valued or continuous-valued. The arcs, usually given in terms of conditional probabilities, represent the probabilistic relationship between nodes. The networks as a whole can be used to represent various complicated models such as target and sensor models.

Once a Bayesian network has been used to represent the model of a problem, the inference problem is to determine the a posterior probability distribution of the state given the observed evidence. Many techniques have been developed for performing Bayesian network inference. These include methods based on graph-theoretic implementations of Bayes rule and marginalization, methods based on message passing and clustering, and other approximate methods based on simulation.

2.1 Bayesian Networks Representation and Algorithms

The Bayesian network technology is a set of modeling (representation) techniques for encoding large-scale systems with inter-dependent uncertain elements into well-structured probability spaces, coupled with a set of inference techniques to obtain a posterior probabilistic assessments given available data. As a new technology, the Bayesian network has been shown to be both computationally more tractable and more easily understood than its predecessor technology. These advantages are achieved primarily through one technical innovation, the representation of conditional independence relationships. Such relationships limit the information used in an assessment or decision to that which is directly relevant and therefore improve both the efficiency of the representation and the inference process.

Bayesian networks provide a flexible representation for complex models, and efficient inference algorithms. A Bayesian network is an acyclic directed graph in which each node in the graph is a random variable. The nodes in the graph collectively satisfy a certain

Markov Property, i.e., the predecessors of a node "separate" it from the nodes preceding those predecessors. In a Bayesian network, the existence of an arc between two nodes indicates a potential stochastic dependence between the two random variables represented by the two nodes.

In addition to the convenient and flexible representation, a major benefit of using Bayesian networks is the existence of many powerful probabilistic inference algorithms developed in the past few years. In ATR, the goal of inference is to update beliefs in particular target classification in the light of the current state of information and new evidence about a target. The updated beliefs are known as posteriors, while the state of information before the evidence is known as priors. Since the nodes of a network are its most basic unit, it is often desired to know the posterior distribution for each node in the graph. Many algorithms are designed to handle this particular query. They include the distributed algorithm, the influence diagram algorithm, the evidence potential algorithm, and the symbolic probabilistic inference (SPI) algorithm.

2.2 Learning Bayesian Networks from Data

For target recognition, a Bayesian network can be constructed either by expert knowledge or learned from the training database. Bayesian methods for learning networks from data take prior knowledge and combine it with data to produce one or more Bayesian networks. The philosophy of Bayesian learning methods, in principle, is based on a so called score function which is proportional to the posterior probability $p(B_S|D)$ of a network structure B_S given database D. Generally, the score function is derived according to a number of assumptions on the underlined probabilistic model. A Bayesian structure B_S that maximizes the score function is considered as the most possible structure generating the database.

During past years, several methods have been developed for learning Bayesian networks from a given database. Some of these algorithms, due to their inherent nature, are computational intensive, and others which employing a greedy search heuristic can not guarantee to converge to the right network, even if the sample size is sufficiently large. Our research focus partially on developing efficient methods for learning Bayesian networks. A number of attributes about a Bayesian network and learning metric are identified. Based on these properties, we developed new learning algorithms which can be shown to be computationally efficient and guarantee the resulting network converging to a right network given a sufficiently large sample size.

3. Accomplishments and Issues

(A) Major Accomplishments of the Project

- Convert Xpatch multi-polarization, multi-frequency synthetic SAR data for four targets into image chips and wrote MATLAB code to display the images. A total of 3600 images for each target are created.

- Develop multi-polarimetric fusion algorithm for SAR target classification [1,5].

- Conduct research on useful target features to use in our Bayes network algorithm for target discremination and classification [2].

- Develop and implement Bayesian Network inference algorithms for generic networks [7].

- Develop and test a wavelet-based feature identification and fusion algorithm [3].

- Test the Bayesnet algorithms against the real feature data provided by Lincoln Lab. and obtain very good performance [2,6].

- Developed learning algorithms to construct Bayesnet from data automatically [4,8].

(B) Project Issues

Since the parent project (DARPA, URI-ATR) was re-directed and extended (without additional funding) to July 1997, we therefore requested a no-cost extension in June, 1996, for this accompanied ASSERT project. We received an approval for no-cost extension of this project to Feb. 1997, and subsequently to July 31, 1997.

4. Personnel

Faculty: Dr. K. C. Chang, Principal Investigator

Students: Mr. Andrew Hauter, Ph. D. student in CSI, US citizen (1994-1997)
Ms. Ellen Vann , MS student in Systems Engineering Dept., US citizen (1994-1995)
Mr. Luke , BS student in Systems Engineering Dept., US citizen (1996)

Dr. K. C. Chang became the Principal Investigator of the project in Jan., 1995. Mr. Andy Hauter has been performing his research under the direction of Dr. K. C. Chang. Mr. Andy Hauter has developed and evaluated the SAR multi-polarization fusion techniques for discrimination. Ms. Ellen Vann performed her research under the direction of Dr. K. C. Chang. Ms. Vann and Mr. Hauter developed the multi-resolution approach using wavelet as well as Bayesian network approach for SAR target classification. Ms. Vann completed her Master in June, 1996. Mr. Hauter is in his third year of Ph.D. program. Mr. Luke performed his work under the direction of Dr. K. C. Chang. Mr. Luke assisted Dr. Chang and Mr. Hauter in their research.

5. Publications

The following publications are the results of the research supported either directly or indirectly by this ASSERT project. The first four papers are attached in this report.

1. Andrew Hauter, K. C. Chang, and Sherman Karp, "Polarimetric Fusion for SAR Target Classification," *Pattern Recognition*, March, 1997.

2. Jun Liu and K. C. Chang, "Feature-Based Target Recognition with Bayesian Networks," *SPIE Optical Engineering Journal*, Vol. 35, NO. 3, pp. 701-707, March, 1996.

3. Andrew Hauter and K. C. Chang, "A Synthetic Aperture Radar Hybrid ATR System," Proceedings of *SPIE*, Orlando, April, 1996.

4. K. C. Chang and Jun Liu, "Efficient Algorithms for Learning Probabilistic Networks," Proc. of *SMC-96*, Beijing, China, Oct., 1996.

5. Andrew Hauter and K. C. Chang, "Polarimetric Fusion for Synthetic Aperture Radar," Proceedings of *SPIE*, Orlando, April, 1996.

6. Ellen Vann, "Bayesian Networks for Automatic Target Recognition ," Master project, George Mason University, June, 1996.

7. K. C. Chang and Jun Liu, "Bayesian Probabilistic Inference for Target Recognition," Proceedings of SPIE, Orlando, April, 1996.

8. Shulin Yang and K. C. Chang , "On Score Matrics for Joint Probability of Bayesian Network Structure and Database," in Proc. of *SMC-96*, Beijing, China, Oct., 1996.

Pergamon

Pattern Recognition, Vol. 00, No. 0, pp. 000–000, 1996

0031–3203/96 $15.00+.00

0031-3203(96)00099-4

POLARIMETRIC FUSION FOR SYNTHETIC APERTURE RADAR TARGET CLASSIFICATION

ANDREW HAUTER[†], KUO CHU CHANG[‡,*] and SHERMAN KARP[§]
[†] C3I Center, George Mason University, Fairfax, VA 22030-4444, USA
[‡] Systems Engineering Department and C3I Center, George Mason University, Fairfax, VA 22030-4444, USA
[§] Fall Church, VA 22102, USA

(*Received* 2 *November* 1995; *in revised form* 20 *June* 1996; *received for publication* 10 *July* 1996)

Abstract—The problem of target classification using Synthetic Aperture Radar (SAR) polarizations is considered from a Bayesian decision point of view. This problem is analogous to the multi-sensor problem. We investigate the optimum design of a data fusion structure given that each classifier makes a target classification decision for each polarimetric channel. Though the optimal structure is difficult to implement without complete statistical information, we show that significant performance gains can be made even without a perfect model. First, we analyze the problem from an optimal classification point of view using a simple classification example by outlining the relationship between classification and fusion. Then, we demonstrate the performance improvement on real SAR data by fusing the decisions from a Gram Schmidt image classifier for each polarization.

Radar target classification   Automatic target recognition   Synthetic aperture radar
Feature fusion

1. INTRODUCTION

There has been a strong interest in employing multiple sensors for surveillance detection and recognition for a number of years. Some of the motivating factors for this interest are: increased target illumination, increased coverage, and increased information for recognition. For military applications combining several sensors or multiple looks of the same sensor is an effective method for increasing recognition performance while providing strength for resisting environmental effects and countermeasures. The approach we outline offers advantages that can be extended to many applications in multisensor surveillance. We have chosen to demonstrate on an application that is particularly challenging and not fully explored, i.e. the fusing of multi-polarimetric classification decisions from a Synthetic Aperture Radar (SAR) sensor.

In this paper a two-level decision problem is outlined. The first level is the single source classification solution. The second level is where a fusing focal point receives each decision, where each decision is made individually and independent of the others. We analyze this from an optimality criterion and assume that the sources transmit their decision instead of raw data. The optimal decision fusion test is outlined and demonstrated with several examples. Instead of using a binary decision alone in the fusion process, we also utilize the value of the decision statistic for maximum information usage. As a result, we show that the classification performance can significantly improve a single source performance. Examples for the Gaussian case are provided and the implications for more realistic non-Gaussian sensor data are discussed.

The approach we developed offers many advantages for multiple sensor surveillance applications. The main advantage is the algorithm design and implementation, where each sensor classifier can be designed and optimized independent of the others. In fact, we demonstrate our approach with a particularly challenging fully polarimetric SAR example where the polarimetric channels are inherently more correlated than the sources from independent sensors would be.

This work is closely related to the distributed multiple sensor detection problem which has been reported in references (1–4). Tenney and Sandell[(1)] developed a theory for obtaining the distributed Bayesian detection rules. They derived the decision rules for the individual detectors that are coupled with information from the other sensors. This decision process is very difficult to perform when the relationship among sensor data is unknown. Chair and Varshney[(2)] presented an optimum fusion structure given that the detectors were independently designed. The solution was then used for a Neyman–Pearson[(3)] test. Surprisingly, this type of fusion construction has not been applied to the problem of multi-polarimetric channel SAR imagery. The outline of this paper will first cover fusion and single source (single polarization, sensors, etc.) preliminaries; then the fusion algorithm will be demonstrated on the two-class

[*] Author to whom correspondence should be addressed. E-mail: kchang@gmu.edu.

Gaussian classification problem; and finally the fusion paradigm will be demonstrated on fully polarimetric SAR data in conjunction with a Gram Schmidt image feature selection method.

2. PRELIMINARIES

There are two major options for decision making with multiple sources. The first option provides complete sensor information to a centralized processor. This is sometimes referred to as feature level fusion because all the feature information is transmitted to the fusion center for a combined decision. The second option is to have a decentralized. decision-level fusion. That is, some or all of the signal processing is performed at each individual source and only local decisions are used for fusion.

The second option is more attractive for many applications due to the fact that in the first option the likelihood functions require knowledge of the joint feature distribution among sensors, $p(x \mid w_i, s_j)$, where w_i is the *i*th class, s_j is the *j*th sensor and x represents the feature. This is difficult to obtain even for similar sensors, because the detection thresholds at individual detectors are usually coupled, i.e. they are not independent. Also, the second option is more desirable due to its cost, survivability, and bandwidth considerations. Nevertheless, the trade-off of the decision-level simplicity is a loss in optimality if the data is not or cannot be uncoupled.

In this paper. we consider the problem of making a decision between two hypotheses. A number of N sensors receive observations and independently implement a local[1] test. Let u_j designate the decision of the sensor, having taken into account all the observations available to this sensor at the time of the decision. Every sensor transmits its decision to the fusion center. so that the fusion center has all N decisions available for processing at the time of the decision making.

Before proceeding with the fusion, we first consider the single sensor problem. To minimize the average probability of error[2], we should select the class that maximizes the *a posterior* probability $P(w_i \mid x)$. In other words, for the multi-category case, we minimize the error rate:

$$\text{Decide } w_i \text{ if } P(w_i \mid x) > P(w_j \mid x) \text{ for all } j \neq i. \tag{1}$$

Now, representing the classifier in terms of the discriminant functions $g_i(x)$, any of the following choices can be derived[(4)] giving identical classification results, but some can be simpler to understand or to compute than others:

$$g_i(x) = P(w_i \mid x) \tag{2}$$

$$g_i(x) = \frac{p(x \mid W_i)P(w_i)}{\sum_{j=i}^{c} p(x \mid w_j)P(w_j)} \tag{3}$$

$$g_i(x) = p(x \mid W_i)P(w_i) \tag{4}$$

$$g_i(x) = \log p(x \mid W_i) \log P(w_i) \tag{5}$$

where $P(w_i)$ are the class prior probabilities.

While the two-category case is just a special instance of the discriminant function, it can be written in terms of a single function. The following is particularly convenient and will be used throughout the paper:

$$g_i(x) = \log \frac{p(x \mid W_1)}{p(x \mid w_2)} + \log \frac{P(w_1)}{P(w_2)}. \tag{6}$$

3. Data fusion

The data fusion problem can be viewed as an *m*-hypothesis problem with individual source decisions being the observations. The first fusion rule we will discuss was given by Chair and Varshney[(2)] who showed the Bayesian optimum rule as a weighted sum of local decisions. The weights being functions of local false alarm rates and probabilities of detection. Each detection is assumed to be statistically independent and each detector makes a binary decision where u_i, $i=1,\ldots,n$

$$u_i = \begin{cases} -1, & \text{if } H_0 \text{ is declared} \\ +1, & \text{if } H_1 \text{ is declared} \end{cases}$$

After processing the observations locally the decisions, u_i, are transmitted to the fusion processor. The structure of this fusion rule is derived using the discriminant function discussed earlier equation (2),

$$\begin{aligned} g(u) &= \log \frac{P(H_1 \mid u)}{P(H_0 \mid u)} \\ &= \log \frac{P_1}{P_0} + \sum_{S_+} \frac{P_{D_i}}{P_{F_i}} + \sum_{S_-} \log \frac{1 - P_{D_i}}{1 - P_{F_i}} \end{aligned} \tag{7}$$

where S_+ is the set of all i such that $u_i = +1$, S_- is the set of all i such that $u_i = -1$, P_{F_i} and P_{D_i} are false alarm rates and probabilities of detections of each local sensor, and P_1 and P_0 are the *a priori* probabilities of the two hypotheses. In the case where all the sensors are similar and operate at the same error probability level, i.e. $P_{F_i} = P_F$ and $P_{D_i} = P_D$ for every sensor, equation (6) is particularly easy to analyze because the decision variable has a binomial distribution that is distributed k out of N decisions for favoring a hypothesis. Hence, the fusion center probability of errors, are P_F^f and P_D^f

$$\begin{aligned} P_F^f &= \sum_{i=\lceil T \rceil}^{N} \binom{N}{i} P_F^i (1 - P_F)^{N-i} \\ P_D^f &= \sum_{i=\lceil T \rceil}^{N} \binom{N}{i} P_D^i (1 - P_D)^{N-i} \end{aligned} \tag{8}$$

where $\lceil T \rceil$ indicates the smallest integer exceeding T, the decision threshold of k.

[1] In our examples, local class parameters (feature distributions) are estimated using training data.

[2] Theoretical optimality can only be achieved when all class parameter distributions are known.

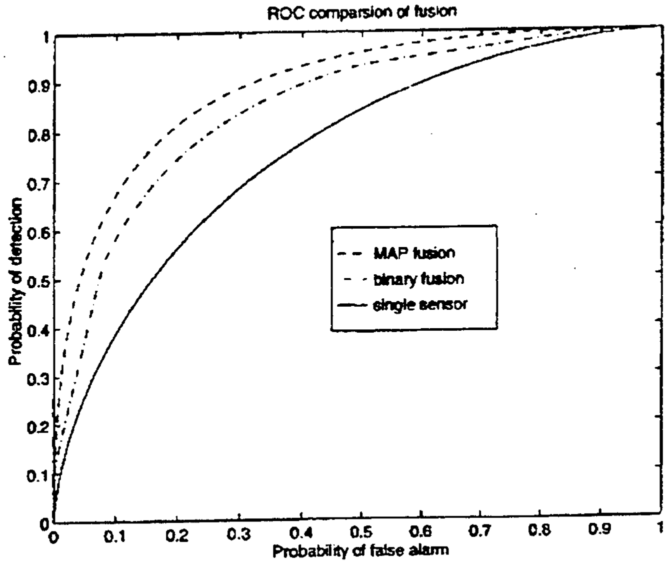

Fig. 1. Information fusion approaches compared.

Thomopoulos *et al.*[(3)] obtained similar results using the *N-P* criterion in addition to showing improved performance by transmitting a single confidence bit along with the hard decision. We follow basically the latter approach, but do not restrict the fusion rule to a single confidence bit. Lee and Chao[(5)] observed that by using a 3-bit fusion paradigm the detection performance of the system was nearly optimal (i.e. the performance of an optimum centralized system). The binary decision process certainly implies an information loss, but is used for distributed sensor problems where the communication channel is an issue. For our application we are not hindered by the communication channel. Specifically, we assume that the decision statistic as well as the decision threshold of each local detector are given. We investigate two fusion algorithms based on this assumption. The first method is the *MAP* rule introduced in equation (1) as a classifier, which applies equally well as a fusion algorithm, i.e.:

$$\text{Decide } w_i \text{ if } P(w_i \mid u) > P(w_i \mid u) \text{ for all } j \neq i. \quad (9)$$

Naturally, equation (9) is the desired implementation because it provides an optimal[3] solution. However, in general, we may not be able to accurately estimate the feature distribution functions. Therefore, the second method is a heuristic approach that directly extends the optimal binary approach introduced by Varshney [equation (7)]. The fusion algorithm is the same but the decision region is expanded to include the full threshold range:

$$g(u) = \log\frac{P_1}{P_0} + \sum_{\Lambda_1} \log\frac{P_i^d}{P_i^f} + \sum_{\Lambda_0} \log\frac{1-P_i^d}{1-P_i^f} \quad (10)$$

where Λ_1 is the set of all i such that $\{g_i(x) \geq T_i\}$ and with T_i being the individual source threshold for partitioning the decision regions, and are the probabilities defined by the Cumulated Probability Functions (CDFs) for the *each* decision statistic, e.g. $P_i^d = P(u_i > g_1(x) \mid w_1)$. This is simply an extension of equation (6) to include more than binary decision information. In practice, the CDFs will be quantized and estimated from training on the individual sensor's classifier error probabilities, which is the same for equation (9), but the advantage here is if the estimates are not ideal, then the added information from passing the decision partitions can provide some strength against environmental uncertainties. In a distributed scenario the weighting can be computed at each sensor and transmitted to the fusion center where they will be summed and compared to the decision threshold.

To demonstrate a simple example we consider a system of three sensors, N=3, where the observation of each sensor is distributed normally as $N(0,1)$ for H_0 and $N(1,1)$ for H_1. We compare the performance with the best centralized scheme, which utilizes raw data, not decisions, from the different sensors. Figure 1 displays the single sensor performance, the optimal binary fusion [equation (7)], and the *MAP* fusion.

Up to this point we have looked at the design of a decision-level fusion scheme. The present analysis can be extended in many directions. Equation (7) is certainly not difficult to implement, since it only processes a single hard bit decision. Let us examine the extra difficulty in implementing equations (9) and (10) using the full range of decision statistics from training. This is not the typical difficulty associated with the classical methods that attempt to assimilate the likelihood ratio solution because the probability distributions for each hypothesis must be defined. In the next section we apply the decentralized scheme to fuse the returns of a fully polarimetric SAR classifier. With three linear polarizations which are known to have low correlation, the gain

[3] Again, optimality can only be achieved when all feature distributions are known.

is shown to be considerable. First, we introduce the classification procedure that will be used for each polarization and then we discuss the fusion methods that operate on each classified decision.

3.1. Description of the classifier and fusion implementation

A simplified block diagram of a complete multi-stage ATR system is shown in Fig. 2. The first stage locates regions of interest that contain "target-like" features. The goal of this stage is to rapidly sift through the sensor imagery to make quick and computationally efficient decisions while still keeping the false alarm rate at a reasonable level. Although the first stage is still an active research problem, this paper focuses on the latter two stages (target classification and fusion) where the burden of the final decision making process lies. For the classifier we use a signal subspace technique popularized in communications to get target features where each signal is defined by a separate basis function in an orthonormal set. If this set can be described, then the optimum classifier easily follows, hence the minimum probability of error. One method of determining a set of orthonormal basis functions from a signal set is by using the Gram–Schmidt orthonormalization procedure. This is similar to several orthogonal methods such as the Karhunen–Loeve expansion. The Gram–Schmidt basis functions offer a reduced yet complete signal set. For this preliminary study, we use a target set having an angular coverage of approximately 68°. The target images are comprised of 32×32 pixel Lincoln Laboratory's millimeter wave SAR sensor data in the spotlight mode. The images are separated by one degree azimuth increments; thus, we have 64 images of each target type. We use a subset for designing the classifier. The classifier is designed by processing a target set through an orthonormalization Gram–Schmidt analysis to determine a subset that best accounts for the characteristics of all the images within the target space. Let the design images be denoted by vectors X_1, $X_2, \ldots$, X_N. The residuals of each image are combined into a matrix formed from the vectors as

$$\bar{X} = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N]. \tag{11}$$

The Gram–Schmidt decomposition process requires that we choose a subset of images that provides the best representation of the target set. We refer to this subset of images as the Gram–Schmidt (GS) set. The GS algorithm is described in many linear algebra textbooks[6]. In our case an orthonormal set Q is produced by the algorithm from $\bar{X}$ to make up the components of the newly reconstructed design set Y, such that:

$$Y_i = \sum_{i=1}^{N} Q_i \tag{12}$$

where N is the rank of Q.

A subset consisting of N images per target is selected for the classifier design. For the 68 target chips, a design chip was selected every other degree. Figure 3 shows a transformation example for the third GS-element image, the corresponding basis image Q_3 (third element of a GS set of eight), and $Y_3 = Q_1+Q_2+Q_3$. Visually, we can see how the combined Y_3 is nearly a perfect reconstruction of the original image X_3.

The classifier operates on the unknown data using the GS set as follows. A test image X_t is projected onto the N dimensional space spanned by the GS set. The components of the projection are calculated as follows:

$$\Theta_i = (X_t - \bar{X}) \bullet Y_i^{\mathrm{T}} \tag{13}$$

where $\bar{X}_t$ is the residual test image. This test will determine if there is enough energy contained in the test image to fall within one of the target classes. The scalar components are placed into an N dimensional feature vector

$$\Theta_t = [\Theta_1, \Theta_2, \ldots, \Theta_N] \tag{14}$$

The max(Θ_i) for i=1,...,N is the distance decision measure corresponding to the best projection onto the GS-set. Hence, if the GS-set completely characterizes the target set then all of the target images should project onto one of the N vectors completely. Also, this procedure was tested against a "cultural" clutter chip set in the same manner. Referring to Fig. 2, these chips were passed through a stage 1 prescreener developed by Lincoln laboratory.[7] These consist of 100 challenging false alarms. Fig. 4 shows each single channel HH, HV, and VV performance of the classifier operating against all the 100 false alarm chips and the 68 target chips. Larger GS-set's were tested with improved performance, but the goal was to retain a reasonably small filter set. Also, other researchers have reported algorithm performance on this data set[7], e.g. an eigen-image approach, a quadratic distance correlation classifier, and a shift invariant 2D pattern matcher.

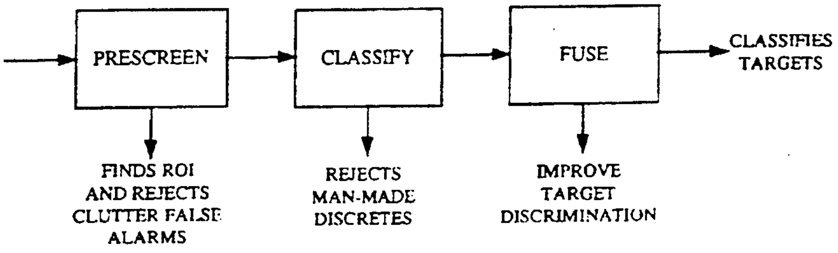
PRESCREEN
CLASSIFY
FUSE
CLASSIFIES TARGETS
FINDS ROI AND REJECTS CLUTTER FALSE ALARMS
REJECTS MAN-MADE DISCRETES
IMPROVE TARGET DISCRIMINATION

Fig. 2. Block diagram of Automatic Target Recognition (ATR) system.

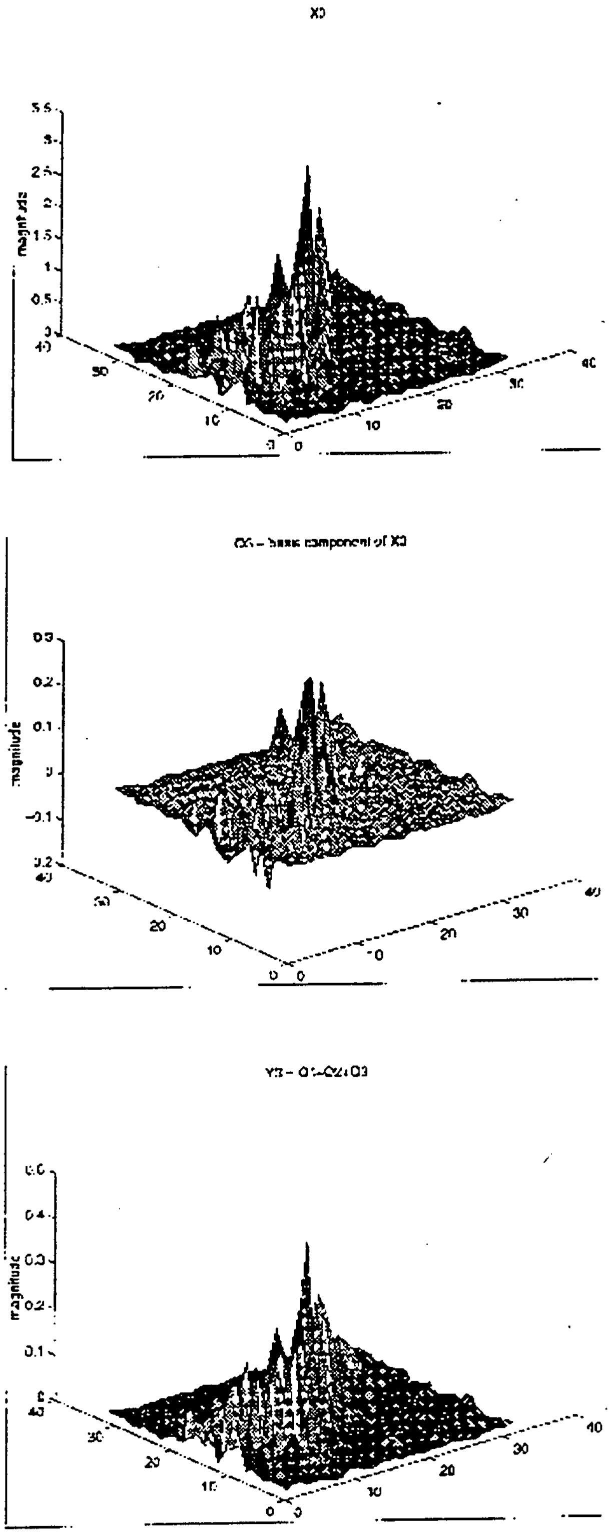

Fig. 3. GS orthogonal composition example.

Our algorithm reported very similar performance for a single channel classifier.[4]

Nevertheless, by combining the classification decisions using the rules such as equations (9) and (10), further refining of classification performance can be obtained. In our next experiment, we demonstrate this by combining the three classifier decisions from each polarization into the heuristic fusion rule. A previous

[4] All four algorithms reported similar performance within one or two false alarms. There was not enough target chips to draw conclusive confidence.

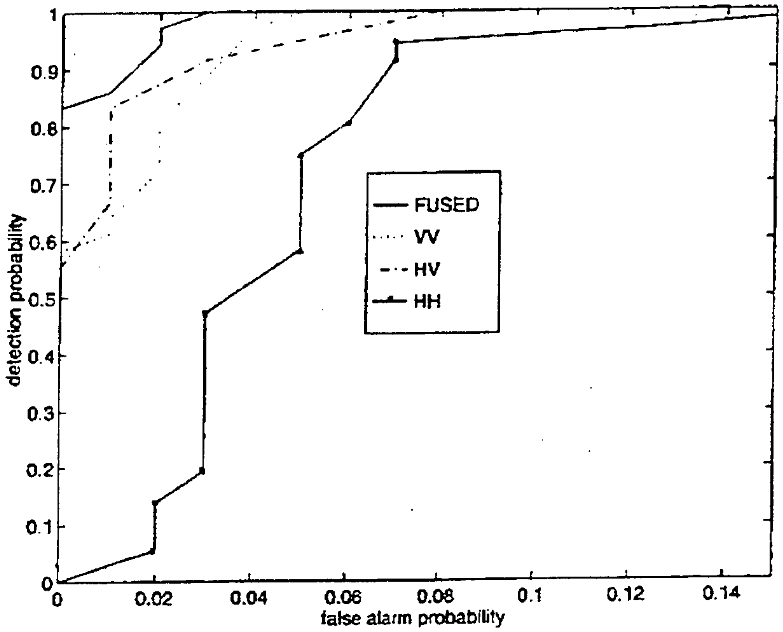

Fig. 4. Heuristic fusion using Gaussian estimate

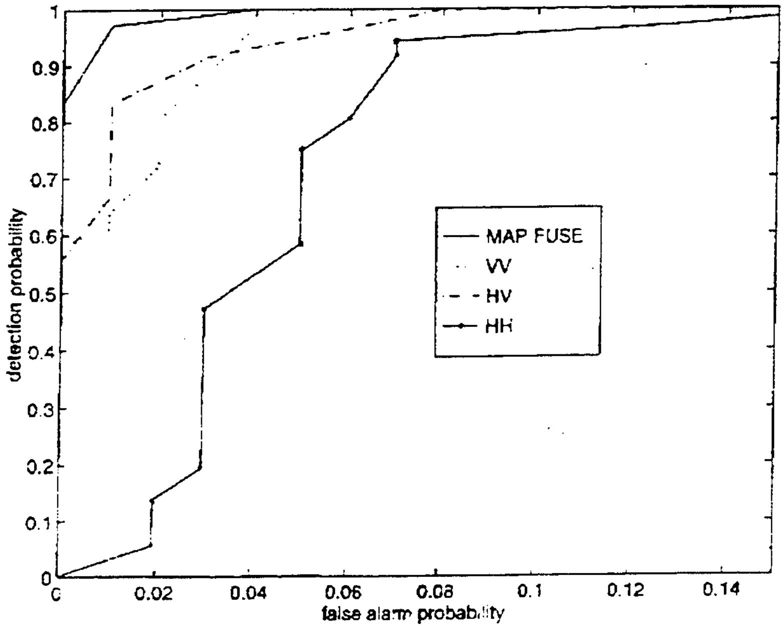

Fig. 5. MAP fusion using empirical estimate.

Table 1. Minimum average probability of errors.

	Bayes error
HH	6.28%
HV	4.00%
VV	2.50%
Fusion	1.50%

analysis[8] determined the polarizations to be marginally correlated, which indicated the possibility for a decision level fusion success. Each polarization channel was classified by the GS classifier to determine: first, a minimum probability of error threshold and secondly, to estimate the CDFs by using a Gaussian approximation from the classifier decisions, as are required by the weighting functions [equation (10)]. Figure 4 displays the improved performance resulting from fusing the polarization decisions, and Table 1 shows their minimum average probability of errors[5]. By choosing a larger GS set the performance improved, but we desired

[5] Average of missed detection probability and false alarm probability

a less than perfect classifier to fully demonstrate the fusion benefit.

In the second experiment, instead of using the Gaussian approximation for the distribution, we formed an empirical distribution from the histogram. In this case the MAP fusion performed significantly better as indicated in Fig. 5.

4. CONCLUSION

We have studied a two-level decision system in which the local decision statistics and their performance characteristics are given. Instead of using a binary decision alone in the fusion process we utilize the value of the decision statistic for maximum information usage. As a result, we show by example that the classification performance can significantly improve a single source performance. Also, we developed practical implementations to improve the binary fusion strategy. The approach we developed offers many advantages for multiple sensor surveillance applications. The main advantage is the algorithm design and implementation, where each local classifier can be designed and optimized independent of the others. We demonstrated our approach with a particularly challenging fully polarimetric SAR example where the polarimetric channels are inherently more correlated than the sources from independent sensors.

REFERENCES

1. R. R. Tenney and N. R. Sandell, Jr., Detection with distributed sensors, *IEEE Trans. Aerospace and Electronic Systems* **AES-17**, 501–510 (1981).
2. Z. Chair and P. K. Varshney, Optimal data fusion in multiple sensor detection systems , *IEEE Trans. Aerospace and Electronic Systems* **AES-22**(1), 98–101 (1986).
3. S. C. Thompoulos, R. Viswanathan and D. Bougoulias Optimal decision fusion in multiple sensor systems, *IEEE Trans. Aerospace and Electronic Systems* **AES-23**(5), 644–652 (1987).
4. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, p. 32, Wiely, New York (1973).
5. C. C. Lee and J. J. Chao, Optimum local decision space paritioning for distributed detection, *IEEE Trans. Aerospace and Electronic Systems*, AES25 **4**, 536-544 (1989).
6. J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*, pp. 266–272, Prospects Heights, Ill., Waveland, (1965).
7. L. M. Novak and G. J. Owrika, Radar target identification using an eigen-image approach, *IEEE National Radar Conference*, Atlanta, GA, March (1994).
8. A. Hauter, A williams, G. Orsak, and V. Diehl, Benchmarking ATR test data, *ATR Systems and Technology Proceedings*, November(1994).

About the Author — ANDREW HAUTER received the M.S. Degree in Electrical and Computer Engineering from George Manson University (GMU) in 1995 and is currently a Ph.D. student in Computational Sciences and Informatics (SCI) in GMU. His research interest include multiresolution approaches for modeling multi-hypothesis target classes; decentralized fusion for combining multi-feature and multi-sensor classifier decisions; and non-parametric classification strategies.

About the Author — KUO-CHU CHANG received the B.S. degree in Communication Engineering from the National Chiao-Tung University, Taiwan, in 1979 and the M.S. and Ph.D. degrees both in Electrical Engineering from the University of Connecticut in 1983 and 1986 respectively. From 1983 to 1992, he was a senior research scientist in Advanced Decision Systems (ADS) division, Booz-Allen and Hamilton, Mountain View, California. He jointed the Systems Engineering department, George Mason University in 1992 as an associate professor. His research interests include estimation theory, optimization, signal processing, and data fusion. He is particularly interested in applying unconventional techniques in the conventional decision and controls systems. He has published more than fifty papers in the areas of multitarget probabilistic tracking, distributed sensor fusion, and Bayesian probabilistic inference. He is currently an editor on navigation/tracking systems for IEEE Transactions on Aerospace and Electronic Systems. Dr. Chang is also a member of Etta kappa Nu and Tau Beta Pi

About the Author — SHERMAN KARP received the BSEE and MSEE from MIT, in 1960 and 1962, and Ph.D. from University of Southern California in 1967. From 1978 to 1981 he was a principle scientist in the Strategic technologies Office in DARPA where, among other things, he inverted the blue/green laser satellite to submarine optical communications program and was principal for the development of the mini-GPS system. Since 1986 he has been serving as a consultant to both government and industry in GaSa systems, communication systems, electro-optic systems, radar, geo-location systems, and detection and recognition ATR algorithm development. He recently pioneered the development of the Soldier 911 geo-location reporting system.

Feature-based target recognition with a Bayesian network

Jun Liu
Kuo-Chu Chang
George Mason University
School of Information Technology and Engineering
Center of Excellence in Command, Control, Communications, and Intelligence (C^3I)
Fairfax, Virginia 22030
E-mail: jliu@c3i.gmu.edu

Abstract. The problem of target classification with high-resolution, fully polarimetric, synthetic aperture radar (SAR) imagery is considered. We propose a framework of using a Bayesian network for feature fusion to deal with the difficult problem of SAR target classification. One difficult problem in SAR feature identification and fusion for target classification is that the features identified may not be independent and that it is not easy to find the "right" fusion rule to combine them. The Bayesian network model when constructed properly can explicitly represent the conditional independence and dependence between various features and therefore provide a sound and natural framework for feature fusion. This paper summarizes our recent work in SAR target recognition using a feature-based Bayesian inference approach. The approach works on the selected features which are chosen so that the separability of the original data are well maintained for later classification. Once the original data are mapped into feature space, the probabilistic model between features and the target is estimated and represented by a Bayesian network, which is then used to calculate the probabilities that a target belongs to one of the given classes based on the observed features. A comparison between the above technique and the traditional statistical approaches such as nearest mean and Fisher pairwise is illustrated based upon performance on a fully polarimetric ISAR (inverse SAR) image data set. Note that although the feature set used in the paper is obtained from the same sensor, the concepts of feature selection and Bayesian network formulation discussed in the paper are not restricted to this case only. They can be applied for multisensor feature-level fusion as well. © *1996 Society of Photo-Optical Instrumentation Engineers.*

Subject terms: sensor fusion; feature level fusion; Bayesian network; synthetic aperture radar (SAR) imagery; target recognition.

Paper SF-002 received July 3, 1995; revised manuscript received Sep. 6, 1995; accepted for publication Oct. 3, 1995.

1 Introduction

The objective of an automatic target recognition (ATR) system is to detect and recognize targets from sensor data. One of the important components of an ATR system is its classifier. The function of the classifier is to categorize input measurements that represent detected targets according to target type. The classifier output corresponding to each input is an estimate of correct category label, based on the observable characteristics of the input.

In general, a feature-based classifier consists of two major parts, a feature selection and a classification mechanism. For the purpose of target classification, the features selected do not necessarily have physical meaning. The only goal in designing features is to preserve class discriminant information of the data while ignoring information that is irrelevant to the discrimination task. Once a feature is identified, it will define a transformation to map input measurements into feature space, which usually has a much lower dimension than that of the input space. This will greatly simplify the classification problem.

A number of attributes that are present in ISAR images can be exploited to discriminate between targets and clutter false alarms. They are size, shape, signal strength, polarimetric properties, spatial distribution of reflected signal, and so on. However, only a few of them can be used to discriminate among classes of targets. It is very difficult to develop discrimination features for exploiting these attributes in any optimal fashion. Furthermore, the features identified may not be independent and it is not easy to find the "right" fusion rule. In fact, past experimental results[1] showed that adding features does not necessarily improve performance if they are not handled correctly. In this study, in the first stage, we examined up to twelve features against the data; some were studied before,[1] and others are new. Out of the twelve features, we then selected the best discrimination features for classification. The main idea in this stage is to select the most useful information or processed results, while ignoring the irrelevant or bad ones. Although the feature set used in the paper is computed from the data of the same sensor, the concepts of feature selection and decision making discussed here are not restricted to this case only.

Research was supported by ARPA under Contract No. N66001-93-C-6010.

In the second stage, a Bayesian network is used for classification. A Bayesian network is a directed, acyclic graph in which the nodes represent random variables, and the arcs between the nodes represent probabilistic dependence between the variables. The Bayesian network model when constructed properly can explicitly represent the conditional independence and dependence between various features and therefore provide a sound and natural framework for feature fusion. Much attention has been drawn to this technology in the past few years and it has been successfully applied both to tasks of assessment under uncertainty and tasks of decision-making under uncertainty.[2–4] Recently it has also been applied to multisource intelligence fusion.[5] In this paper, for feature-level fusion, we first identify the network topology for various target parameters such as class and orientation and sensor data features. Instead of working directly on the input measurements, the transformed data on the chosen feature spaces are used as input. With the selected topology, we then learn (estimate) the probabilistic relationship between various variables in the Bayesian network. The conditional probability that a target belongs to a class given observed features is then computed based on a probabilistic inference algorithm using the network. Finally, the target is assigned to the class with the highest probability. Note that the idea proposed here is general and can be applied to multisensor domain directly. In fact, the idea of applying a Bayesian network to multisensor fusion has become more and more popular in the fusion community.[6]

This paper is organized as follows: Section 2 describes target sets and radar image data, Sec. 3 introduces the network algorithm and the classification process. Sec. 4 presents the performance results, and finally, Sec. 5 contains some concluding remarks.

2 Data Description

The database used for this study was obtained from MIT's Lincoln Laboratory; it consists of four representative targets: a Dodge van, a Chevrolet Camaro, a Dodge pickup truck, and an International Harvester bulldozer. Each of the four targets was put on a platform, and the millimeter-wave radar image data collected using a 35-GHz normal frequency at a fixed 5.5-deg depression angle while the platform turned over a complete 360-deg azimuth. These inverse synthetic aperture radar (ISAR) images are 1-ft. range-processed with full polarizations, namely, horizontal transmit, horizontal receive (HH); horizontal transmit, vertical receive (HV); and vertical transmit, vertical receive (VV). The images of these vehicles are available at 0.04-deg azimuth intervals and each is associated with a viewing angle. The original image of the vehicles has the size of 32×20 pixels. Figure 1 shows an example of four target images using single-channel HH at a 0.4-deg azimuth.

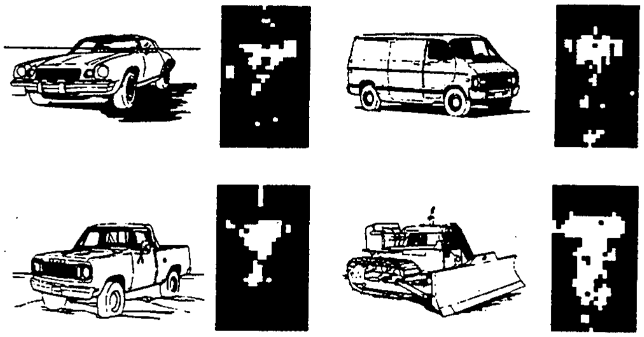

Fig. 1 ISAR images of an HH channel for four targets at 0.4 azimuth with a 1-ft.×1-ft. resolution.

The fully polarimetric ISAR data were first filtered by a polarimetric whitening filter (PWF),[7] and then normalized and compressed by a window slicing technique to a 15×9 dimension.[8] In this study, 5280 processed images were picked up from the database for each target. It should be mentioned that the targets look confusing from different angles. In particular, an image for a target at one angle may look like the image of another target at the same or a different angle, while some images from adjacent angles for the same target may look much different. This made the classification task very difficult.

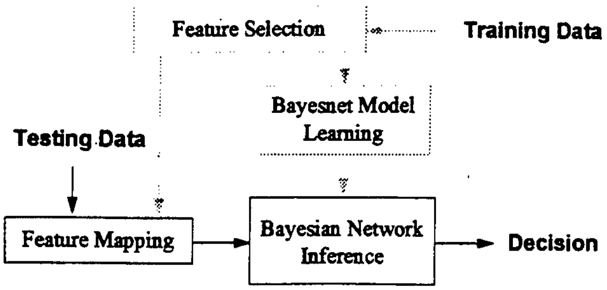

Fig. 2 Feature-based classifier with a Bayesian network.

3 Feature-Based Classification Procedure

Our feature-based classifier is composed of the following stages. As shown in Fig. 2, the observed measurement is first transformed into feature space based on preselected features. Then the transformed data are input into the Bayesian network for probabilistic inference. The result is a set of estimated conditional probabilities that the observed target is from one of the classes given the observed features. Finally, the decision-making procedure simply compares these estimated conditional probabilities, and the observed target is assigned to the class with the highest conditional probability. Note that the feature selection and Bayesian network model learning modules are based on the training data and are done a priori off-line.

The most difficult part to build into this classifier is feature selection. Once features are chosen, the second step is to learn the probabilistic models between features and the targets and represent the model by a Bayesian network. For simplicity, we have assumed a simple two-level network topology where the observed features are assumed to be conditionally independent given the target class and image azimuth angle* (see Fig. 3). Given the network topology, the first task is to estimate the conditional probabilities of the observed features given the target parameters. These

*Note that the observed features are not conditionally independent given only the target class.

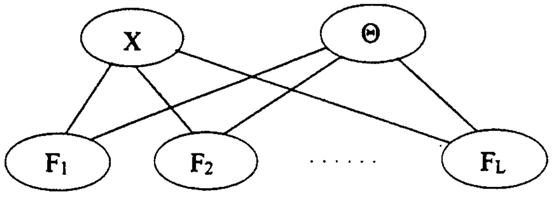

Fig. 3 A probabilistic model between features and target states.

estimated conditional probability distributions are used to describe the variable relationships in the Bayesian network. In the rest of this section, we will describe feature selection, Bayesian network modeling and decision making in detail.

3.1 *Feature Selection*

As mentioned before, the useful features here are those that preserve class discriminant information on the data while ignoring information that is irrelevant to the discrimination task. Since, at present, no method exists for developing discrimination features in any optimal fashion, one way is to test all the proposed features on the data set to see how well they can separate the targets of different classes. The good features, which have a better ability to separate different target classes, are maintained for the further use of classification, and the poor ones are discarded.

In this study, a total of twelve features were examined, of which three (standard deviation, fractal dimension, and weighted-rank fill ratio) were developed by MIT's Lincoln Laboratory to discriminate targets from nature-clutter false alarms in their ATR system,[9] and the rest are new features. Some features are contrast-based, and the other spatial-distribution related. They are

(i) The standard deviation (SD) feature is a measure of the fluctuation in intensity in an image. It is computed from the typical estimator for the standard deviation by using the power (expressed in decibels) of all the pixels in an image. If the radar image in power is denoted by $P(r,a)$, then the log standard deviation σ can be estimated as:

$$\hat{\sigma}=\left(\frac{S_2-S_1^2/N}{N-1}\right)^{1/2}, \tag{1}$$

where

$$S_1=\sum_{r,a} 10\log_{10} P(r,a), \tag{2}$$

and

$$S_2=\sum_{r,a} [10\log_{10} P(r,a)]^2. \tag{3}$$

N is the total number of pixels in the image, which is $9\times 15=135$.

(ii) The *fractal dimension* (FD) feature provides a measure of the spatial distribution of the brightest scatterers in the images. To calculate the FD feature of an image, the first step is to convert the image to a binary one. To do so, the K brightest pixels in the image are selected and their values are converted to 1, while the rest of the pixel values are converted to 0. Then the fractal dimension can be estimated as:

$$\dim=-\frac{\log M_1-\log M_2}{\log 1-\log 2}=\frac{\log M_1-\log M_2}{\log 2}, \tag{4}$$

where M_1 is the number of 1-pixel-by-1-pixel boxes needed to cover the image, and M_2 is the number of 2-pixel-by-2-pixel boxes needed to cover the image. Obviously, $M_1=K$.

(iii) The *weighted-rank fill ratio* (WRFR) feature measures the percentage of the total energy contained in the brightest scatterers of an image. Using the notation of Eq. (2), this feature is defined as follows:

$$\eta=\frac{\Sigma_{\text{k brightest pixels}}P(r,a)}{\Sigma_{\text{all pixels}}P(r,a)}. \tag{5}$$

(iv) The *counting* (CNT) feature is obtained by counting the number of pixels in the image that exceed a specific threshold, then dividing by the total number of pixels of the image.

(v)–(xii) The following eight features are designed to measure the spatial distribution of the brightest pixels in the images. First we convert the image to binary by using *amplitude thresholding,* in which all pixel values exceeding a specified threshold are converted to 1, and the remaining pixel values are converted to 0. Assuming the converted binary image is denoted as $B(i,j)$, then these features are defined as:

$$M_X=\frac{1}{N}\sum_{i=1}^{9}\sum_{j=1}^{15} i\times B(i,j), \tag{6}$$

$$M_Y=\frac{1}{N}\sum_{i=1}^{9}\sum_{j=1}^{15} j\times B(i,j), \tag{7}$$

$$S_{XX}^2=\frac{1}{N-1}\sum_{i=1}^{9}\sum_{j=1}^{15} (i-M_X)^2\times B(i,j), \tag{8}$$

$$S_{YY}^2=\frac{1}{N-1}\sum_{i=1}^{9}\sum_{j=1}^{15} (j-M_Y)^2\times B(i,j), \tag{9}$$

$$S_{XY}^2=\frac{1}{N-1}\sum_{i=1}^{9}\sum_{j=1}^{15} (i-M_X)(j-M_Y)\times B(i,j), \tag{10}$$

$$W_X=\max\{i:B(i,j)=1\}-\min\{i:B(i,j)=1\}, \tag{11}$$

$$W_Y=\max\{j:B(i,j)=1\}-\min\{j:B(i,j)=1\}, \tag{12}$$

$$W_{XY}=W_X\times W_Y, \tag{13}$$

where N is the total number of pixels in the image.

Features need to be evaluated since using similar features does not guarantee a better discrimination performance, and sometimes adding features can even degrade performance. One way to evaluate whether a feature is good for discriminating targets of different classes is to test it on the training data set in a heuristic manner. The other way is to use the optimal feature set selection approach,[10] which uses a Bhattacharrya upper bound[11] on the probability of classification error to determine which subset of any l features of the available L features has the lowest upper bound. It can be seen in Sec. 4 that this latter approach is an effective way to select features and its performance is very close to what we obtained using an "optimal manually chosen" feature set.

3.2 *Probabilistic Modeling of a Bayesian Network*

Our objective is to recognize targets from the observed data. Bayesian networks show great promise for performing this function since they can be used to represent complicated probabilistic relationships among variables of interest. Furthermore, many efficient algorithms have been developed for drawing inferences from the evidence.[12–15] For the current Bayesian network model, let the class status of an observed target be a random variable X, and assume the target belongs to one and only one of K classes; then, obviously, X is a discrete random variable, and without loss of generality, we can assume it takes on a value $1,2,...,K$ (in our case $K=4$). In a radar image, an important factor that greatly affects the appearance of the target is the target orientation. Let Θ denote the azimuth angle when a target is imaged, which can have values from 0 to 360 deg. If we discretize the 360-deg space into M small sectors and each has an equal interval, Θ can be treated as a discrete random variable. Finally, based on a set of selected features, denoted as random variables $F_1,F_2,...,F_L$, discrete or continuous, a simple two-level probabilistic model between features and target states (class and azimuth) can be obtained as in Fig. 3.

As shown in Figure 3, each node represents a random variable, and each line indicates the conditional probabilistic relationship between the connected nodes. Note here that the network topology implicitly assumes that the observed features are conditionally independent given the target class and orientation. However, in general, this is not the case, and a more complicated network topology is needed to model the problem. In a Bayesian network, the conditional probability distribution of a child given all of its parents is assumed to be given before any probabilistic reasoning can be drawn. In our case, this is to say that given target type X and radar azimuth angle Θ, the feature F_l is distributed with the known distribution $P(F_l|X,\Theta)$, $l=1,2,...,L$. In reality, the conditional distributions $P(F_l|X,\Theta)$ need to be elicited by expert knowledge or physical models or estimated with the training data. It should be mentioned that for some continuously distributed features, since their distributions are hardly close to any of the well-known parameterized probability distributions, they must be estimated with nonparametric methods. We will discuss this in more detail in the next section.

With the Bayesian network, the class probabilities of observed targets can be computed with any probabilistic inference algorithm.[12–15] Basically, the question becomes one of how to calculate the conditional probability that an observed target is from a class at an azimuth angle given the observed features, e.g., $P(X,\Theta|F_1,F_2,...,F_L)$. For the current simplified model, since the features $F_1,F_2,...,F_L$ are conditionally independent given X and Θ, it can be shown that the required conditional probability can be obtained as:[†]

$$P(X,\Theta|F_1,F_2,...,F_L)=\frac{1}{C}\prod_{l=1}^{L} P(F_l|X,\Theta), \tag{14}$$

where C is a normalizing constant. If there are K classes of targets and M sectors of angles, based on Eq. (14), we can obtain a total of $K\times M$ probability estimations. These are then used to make a decision.

3.3 *Decision Making*

In this step, the observed target is assigned to an appropriate class. What we obtained from Eq. (14) is a set of $K\times M$ probabilities, e.g., how likely an observed target is from class k and at azimuth sector m, $k=1,2,...,K$, and $m=1,2,...,M$. To make a decision, there are two basic decision rules.

Decision rule 1. Among $K\times M$ estimations, find the one with the highest probability value, then assign the target to the class associated with this estimate. It can be seen that, at the time the target class is determined, so can the target azimuth angle. However, this may not be necessary, and hence we have the second decision rule.

Decision rule 2. Instead of estimating $P(X,\Theta|F_1,F_2,...,F_L)$ using Eq. (14), we estimate $P(X|F_1,F_2,...,F_L)$. This can be obtained by the following equation:

$$\begin{aligned} P(X|F_1,F_2,...,F_L)&=\sum_{\Theta} P(X,\Theta|F_1,F_2,...F_L) \\ &=\frac{1}{C}\sum_{\Theta}\prod_{l=1}^{L} P(F_l|X,\Theta). \end{aligned} \tag{15}$$

From Eq. (15); we can obtain K probability estimates. Again, we choose the one with the highest value, then assign the target to the class associated with this estimate.

When making a probabilistic inference, an interesting consideration is to treat the radar azimuth angle as known. This may happen if the radar azimuth angle at which the target is imaged can be determined by other sources of information. If this is the case, namely the radar azimuth angle is known to be in the m'th sector, the conditional probability that the target is from a specific class given the observed features and $\Theta=m$ can be obtained using the following equation:

$$P(X|F_1,F_2,...,F_L,\Theta=m)=\frac{1}{C}\prod_{l=1}^{L} P(F_l|X,\Theta=m), \tag{16}$$

where C is a normalization constant. Again, the decision making is based on K calculated probability estimates.

[†] Again, in general, the network is more complicated and an efficient algorithm such as SPI [16] can be used to compute the conditional probability distributions.

Table 1 Averaged correct classification rates (ACCR) in % using OLPARS.

Optimal feature set	Nearest mean classifier			Fisher pairwise classifier	
	weight	training	testing	training	testing
all twelve features	E	39.7	25.0	86.2	68.3
	var.	66.8	66.2	rej. 20	rej. 830
	cov.	80.2	64.2		
1,3,6,7,12	E	38.3	25.0	83.8	72.4
	var.	66.0	67.9	rej. 20	rej. 44
	cov.	81.5	76.0		
1,3,6,7	E	71.7	71.0	81.9	82.0
	var.	66.9	67.1	rej. 17	rej. 49
	cov.	80.4	80.2		
1,3,6	E	69.1	67.8	78.4	78.4
	var.	69.0	68.5	rej. 26	rej. 26
	cov.	76.7	76.9		
1,3	E	63.1	62.9	72.8	72.2
	var.	65.0	65.0	rej. 59	rej. 141
	cov.	72.4	72.1		

4 Performance Evaluation

The data set used for this study is first randomly and uniformly split into two data sets, the training data set (1728×4) and the testing data set (3552×4). Using the training data set, the conditional probability distributions $P(F_l|X,\Theta)$ are first estimated. They are then used to define the Bayesian network and later to calculate conditional probabilities $P(X,\Theta|F_1,F_2,\ldots,F_L)$ for classification. The testing data set is input to the system to examine the classification performance, and the results are reported in terms of the averaged correct classification rate (ACCR), which is defined as the ratio of the number of correctly classified observations to the total number of observations in the testing data set.

4.1 *Estimation of Conditional Probability Distributions*

In this approach, the conditional probability distributions are estimated by a smoothing kernel approach, which is the most thoroughly developed approach in literature. Assuming we have K classes of targets ($K=4$ in our problem), L features, and the angle space is decomposed into M sectors, there will be a total of $K\times L\times M$ distribution functions to be estimated. For the l'th feature F_l, k'th class, and m'th sector of angles, $P(F_l|X=k,\Theta=m)$ is estimated by using only that image data from the k'th class of target and m'th sector of angles. The data are first transformed based on the feature F_l. If we consider $F_l^1,F_l^2,\ldots,F_l^n$ as n-transformed observations based on the feature F_l, and they are from the k'th class and m'th sector of angles, the kernel estimate has the form

$$\hat{P}(F_l|X,\Theta)=\frac{1}{n}\sum_{i=1}^{n} k(F_l-F_l^i), \qquad (17)$$

where $k(.)$ is a probability density function symmetric about the origin. Here we use a Gaussian density with variance σ^2, which is the only parameter. The parameter should be chosen so that the ACCR is maximized.

4.2 *Classification Results*

In this section, for the purpose of comparison, we first present some test results by using traditional classifiers—nearest mean and Fisher pairwise. We then show some test results to illustrate the key issues discussed in previous sections. The first part of the results (Table 1) is obtained by means of a software package called OLPARS (On-Line Pattern Analysis and Recognition System).[10] The system provides users with a convenient tool to realize a variety of traditional pattern recognition methods, especially optimal feature set selection, as mentioned in the previous section. The traditional classifiers such as nearest mean, Fisher pairwise and so on can also be designed and evaluated easily.

In Table 1, the left-most column refers to the optimal feature set selected in terms of the Bhattacharrya upper bound. The number of features in the feature set should be given in order to do the feature selection. For example, if we decide to use only two features, OLPARS reports that the best feature set is {1,3}, i.e., the first feature SD and the third feature WRFR as defined in Sec. 3.1. Corresponding to each of the optimal feature sets, Table 1 presents the classification results using the nearest mean and Fisher pairwise classifiers for both the training and testing data sets (note: the classifier parameters are estimated by the training data set only). In the table, "E," "var," and "cov." represent the different distance measures, Euclidean, weighted by a diagonal variance matrix, and weighted by a covariance matrix, respectively, and "rej." refers to rejection. As can be seen from the table, the feature set {1,3,6,7} has the overall best performance.

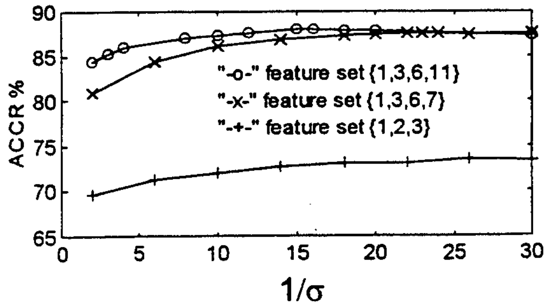

Fig. 4 ACCR vs. $1/\sigma$ for {1,3,6,11} {1,3,6,7}, and {1,2,3}, M=48, and using decision rule 2.

The second part of the results is obtained using feature-based classification with a Bayesian network. Figure 4 shows a comparison of classification results among the optimal feature sets, i.e., {1,3,6,7} selected by OLPARS, the manually selected set {1,3,6,11} and the previously proposed feature set {1,2,3}. The best ACCRs are 87.55% for {1,3,6,7} and 88.00% for {1,3,6,11} respectively, and the latter feature set performs slightly better than the former one. The two feature sets selected by the two different approaches have only one differing feature. However, the optimal feature selection approach is more computationally efficient. In Fig. 4, it also can be seen that the overall performance of feature sets {1,3,6,7} and {1,3,6,11} is much better than that of the previously proposed set {1,2,3}.

Figure 5 gives a comparison between decision rules 1 and 2. We find the second rule is better than the first. However, using the first rule, the target azimuth angles can be predicted simultaneously.

Figure 6 displays the impact of the number M when the azimuth angle space is partitioned. It can be seen that when M is increased, performance improves as well.

The last result, shown in Fig. 7, is obtained using Eq. 16, assuming the radar azimuth angle can be determined from other information sources. In this model, performance is about 5% more accurate than that of decision rules 1 and 2. The best ACCR rate is 93.3%. This result is not surprising because more information is assumed to be available in this model.

By comparing the two experimental results, it can be seen that the feature-based Bayesian net classifier performs noticeably better than the traditional classifiers—nearest mean and Fisher pairwise.

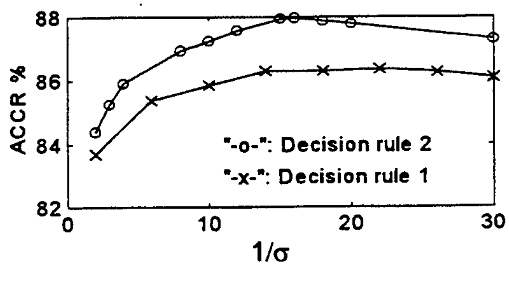

Fig. 5 ACCR vs. $1/\sigma$ for decision rule 1 and 2, M=48, and using {1,3,6,11}

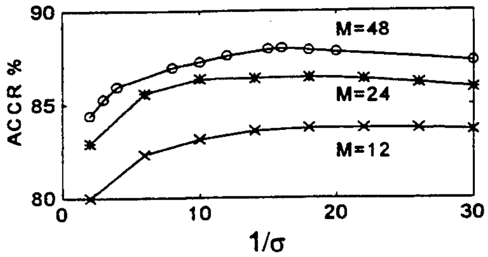

Fig. 6 The impact of M, using decision rule 2 and {1,3,6,11}.

5 Conclusions

In this paper, we applied the Bayesian network technology for feature-level fusion. Bayesian networks show great promise for multiattribute fusion since they can be used to represent complicated probabilistic relationships among variables of interest. We first identify the network topology for various target parameters such as class and orientation and sensor data features. The network topology explicitly represents the conditionally independent or dependent relationships among various features. Instead of working directly on the input measurements, the transformed data on the chosen feature spaces are used as input. With the selected topology, we then learn (estimate) the probabilistic relationship among various variables in the Bayesian network. The conditional probability that a target belongs to a class given observed features is then computed based on a probabilistic inference algorithm using the network. The network model used here is relatively simple. In a separate but related research,[16] we also studied the problem of Bayesian network construction using neural learning techniques where the network is more general and complicated.

In the current approach, performance depends on a number of factors, including selection of a set of workable features, choosing an appropriate probabilistic model to describe the observations, handling approximation of the required conditional probability distributions, and so on. When classifying the SAR image, not only does the Bayesian network model lead to a better performance than certain types of traditional classifiers, for example, nearest mean and Fisher pairwise, it also possesses a certain degree of flexibility to handle other target parameters such as orientation. The orientation can be predicted at the time the target is classified. In the case where the orientation is

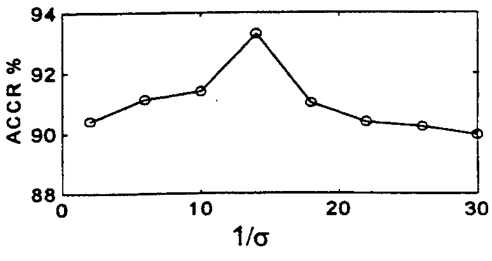

Fig. 7 ACCR vs. $1/\sigma$, assuming Θ is known, using {1,3,6,11}, and M=48.

known, the classification accuracy is improved significantly. The decomposition of azimuth angle space and the parameter of the conditional probability distribution estimates are two main factors that could be adjusted to improve performance.

Although the feature set used in the example is obtained from the same sensor, the concepts of feature selection and decision making discussed in the paper are not restricted to this case. The idea of applying a Bayesian network to feature-level fusion can also be applied to a multisensor domain directly. In fact, the idea of applying a Bayesian network to multisensor fusion has become more and more popular in the fusion community. The evaluation of our results demonstrates the usefulness of the proposed approach.

Acknowledgment

This research was supported by ARPA under Contract No. N66001-93-C-6010.

References

1. L. M. Novak, G. J. Owirka, and C. M. Netishen, "Performance of a high-resolution polarimetric SAR automatic target recognition system," *Lincoln Laboratory J.* **6**(1), 11–24 (1993).
2. R. A. Howard and J. E. Matheson, "Influence diagrams," in *The Principles and Applications of Decision Analysis,* Vol. II, R. A. Howard and J. E. Matheson, Eds., Menlo Park: Strategic Decisions Group, Menlo Park, CA (1981).
3. S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application in expert systems," *J. Roy. Stat. Soc. B* **50** (1988).
4. R. D. Shachter, "Intelligent probabilistic inference," in *Uncertainty in Artificial Intelligence,* L. N. Kanal and J. F. Lemmer, Eds., North-Holland, Amsterdam (1986).
5. K. C. Chang, "Multiple intelligence correlation and fusion with Bayesian networks," *The Seventh Joint Service Data Fusion Symposium* (1994).
6. K. C. Chang and Clayton Stewart, "Application of Bayes nets in sensor fusion," SPIE Euro OPTO, Oct. (1993).
7. L. Novak and C. Netishen, "Polarimetric synthetic aperture radar imaging," *Int. J. Imaging System Tech.* **4**, 306–318 (1992).
8. K. C. Chang and Y. C. Lu, "High resolution polarimetric SAR target classification with vector quantization," to appear in *Proc. FUZZ-IEEE/IFCS '95*, Japan.
9. M. C. Burl, G. J. Owirka, and L. M. Novak, "Texture discrimination in synthetic aperture radar imagery," *Proc. IEEE 23rd Asilomar Conf. on Circuits, Systems, and Computers*, p. 399 (1989).
10. *OLPARS User's Manual, Version 7.3,* PAR Government System, Inc., 1010 Prospect St., Suite 200, La Jolla, CA (1995).
11. P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* Wiley, New York (1973).
12. R. Shachter, A. Del Favero, and B. D'Ambrosio, "Symbolic probabilistic inference: A probabilistic perspective," *Proc. AAAI* (1990).
13. M. Henrion, "Propagating uncertainty by logic sampling in Bayes' networks," Technical Report, Department of Engineering and Public Policy, Carnegie-Mellon University, Pittsburgh, PA (1986).
14. J. H. Kim and J. Pearl, "A computational model for combined causal and diagnostic reasoning in inference systems," *Proc. 8th Internat. Joint Conf. Artificial Intelligence* (1985).
15. K. C. Chang and Robert Fung, "Symbolic probabilistic inference with both discrete and continuous variables," *IEEE Trans. SMC* (1995).
16. K. C. Chang and Sulin Yang, "Bayesian network construction with neural learning," submitted for publication.
17. D. E. Kreithen, S. D. Halversen, and G. J. Owirka, "Discriminating targets from clutter," *Lincoln Laboratory J.* **6**(1), 25–51 (1993).

Jun Liu received his BS in applied mathematics from the Beijing Polytechnic University in 1981, and an MS in remote sensing and cartography from the Graduate College of Chinese Science and Technology University, Beijing, in 1985. From 1985 to 1992 he was an assistant researcher at the Institute of Remote Sensing Application, Chinese Academy of Sciences, working in the areas of information system and field spectral analysis for remote sensing, image processing, and application and evaluation of inertial navigation systems in aerial photography. Since 1993 he has been a graduate research assistant with the Center of Command, Control, Communications and Intelligence at George Mason University, Fairfax, VA where he is a PhD candidate. His current research interests include image processing, image feature detection, and target recognition and classification as well as Bayesian networks.

Kuo-Chu Chang received a BS in communication engineering from the National Chiao-Tung University, Taiwan, in 1979 and MS and PhD degrees in electrical engineering from the University of Connecticut in 1983 and 1986, respectively. From 1983 to 1992, he was a senior research scientist in the Advanced Decision Systems (ADS) division, Booz-Allen & Hamilton, Mountain View, California. He joined the Systems Engineering Department of George Mason University in 1992 as an associate professor. His research interests include estimation theory, optimization, signal processing, and data fusion. He is particularly interested in applying unconventional techniques to conventional decision and control systems. He has published more than 50 papers in the areas of multitarget tracking, distributed sensor fusion, and Bayesian probabilistic inference. He is currently an editor on navigation/tracking systems for *IEEE Transactions on Aerospace and Electronic Systems.* Dr. Chang is a member of Etta Kappa Nu and Tau Beta Pi.

A synthetic aperture radar
hybrid ATR system*

Andrew Hauter and Kuo Chu Chang

George Mason University, Center of Excellence in Command, Control,
Communications, and Intelligence (C3I), Fairfax, VA 22030

ABSTRACT

A hybrid automatic target recognition system is presented that exploits advances in two new fields in detection theory and signal analysis. The first is in the area of Universal Classification that offers asymptotic optimal solutions to non-Gaussian properties of signals and the second is in the field of multi-resolution analysis (MRA) that uses the automatic *feature* isolating properties of the wavelet transform. The Universal Classifier is used as the first stage of a hybrid ATR system that efficiently shifts through large quantities of imagery locating regions of interest that contain "target-like" features. The target chips of interest are then passed through the MRA to be classified at the final stage. Wavelets are adequate to the study of unpredictable signals with both low frequency components and sharp transitions. As a result, there has been recent interest in applying this new signal processing field to the target recognition problem. But few have combined the natural feature extraction capability of time-frequency methods in the classification stage. In this approach, we utilize the sub-space “crystals” from a specific decomposition and operate a classification strategy against each crystal of the transform. The complete ATR system is presented as well as performance examples using both real synthetic aperture radar (SAR) data and data generated using the Xpatch signature prediction code.

Keywords: Synthetic Aperture Radar, Automatic Target Recognition, Universal Classification, Wavelet Multiresolution Analysis

1. INTRODUCTION

The problem concerned with recognition of targets in SAR imagery is an ongoing challenging research topic that is important for military applications. Some of the best performing algorithms reported in government and industry are related to simple template matching algorithms[1]. This is surprising, since theoretically we know that a matched filter is optimal only under Gaussian class descriptions. One reason for such success is the efficient use of spatial information that is ignored by many applications. But one is led to believe that improvements in performance can be made.

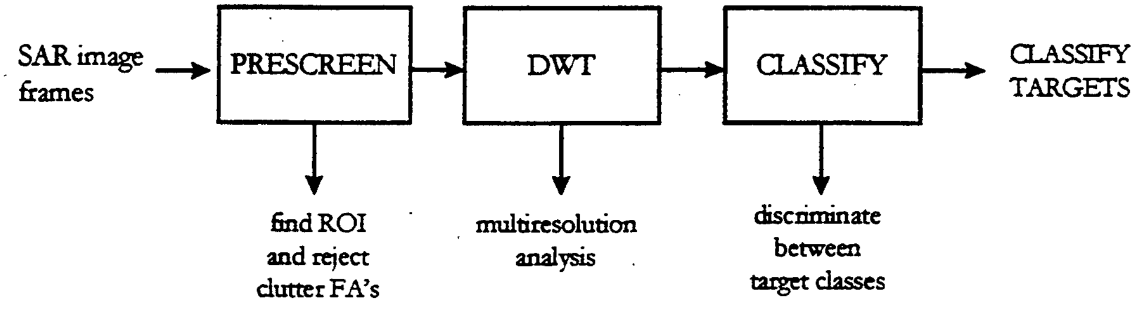

Figure 1. Hybrid SAR ATR System

* Research supported in part by AFOSR under contract #F49620-93-1-0492

In this paper we introduce a unique hybrid ATR system (Figure 1). The first stage exploits recent developments in the theory of Universal Classification[2,3,4]. Prescreening SAR data requires the ability to examine large amounts of data without imposing numerous assumptions on the environment. Universal classification has been shown to be asymptotically optimum (in the amount of data) for classifying extremely general forms of data. The second stage transforms the regions of interest (ROI) from the pre-screener using a multi-resolution analysis (MRA) for the third and final classification stage. The classifier is a Gramm-Schmidt image classifier[5] which discriminates correlation properties of the transformed signal. The organization of the paper is as follows. In Section 2, we introduce the pre-screener and demonstrate on strip map SAR data. In Section 3, we discuss the motivation and approach behind the wavelet decomposition. In Section 4 we present the classifier and discuss some preliminary results using an Xpatch synthetic target data set and finally we summarize in Section 5.

2. UNIVERSAL CLASSIFIER PRE-SCREENER

The general theory underpinning this new branch of classification has been used in universal data compression to develop the Lemple-Ziv (LZ) compression algorithm (compress on UNIX machines). This highly robust algorithm has had an overwhelming impact on the field of data compression and storage. It is believed that the classification extensions of these approaches will have a similar practical impact on classification, and in particular on SAR classification.

The basis for this class of algorithms is the utilization of statistical distance measures between the observed data and training data. Of course, this is similar to correlation based methods in Gaussian environments, but the universal methods incorporates the optimum non-linearities when the environment deviates from this simplistic assumption.

Defining the classification problem as follows:

$$\begin{aligned} H_1&: X \sim S_1(X) \\ H_2&: X \sim S_2(X) \\ &\;\vdots \\ H_N&: X \sim S_N(X) \end{aligned}$$

where we assume that the class densities are unknown. In the absence of a precise statistical model for the classes, we assume the existence of a sequence of training data from the source. We proceed by quantizing and forming the types (empirical density estimates) $P_{S_1}, P_{S_2}, \ldots, P_{S_N}$.

Mathematically, this problem can be modeled as an M-ary hypothesis testing problem. Our Universal Classifier implementation is a generalization from Cover[6], $h(x) = \min_i \left[d_{KL}(P_X, P_{S_i}) \right], i = 1,2,\ldots N$. Where $d_{KL}(P_X, P_Y)$ is the Kullback-Leibler distance between the types P_X and P_Y.

To demonstrate, we apply the classifier to the discrimination between two clutter types (grass and trees) of Lincoln Laboratories ADTS 1-foot resolution SAR strip-map data. Because the classifier is asymptotically optimal the more information available either through higher resolution imagery or as in our example of exploiting the fully polarimetric images the error rate goes to zero (Figure 3). The first row indicates the output decision statistics from testing on several hundred square kilometers of clutter data for varying ROI operators 3x3, 5x5, 20x20. The second row is the respective receiver operating characterics. The last column applies a Bayesian fusion rule across four[1] polarizations that outperforms even the larger windowed 20x20 discriminator. The straight line ROC indicates no errors given the sample size.

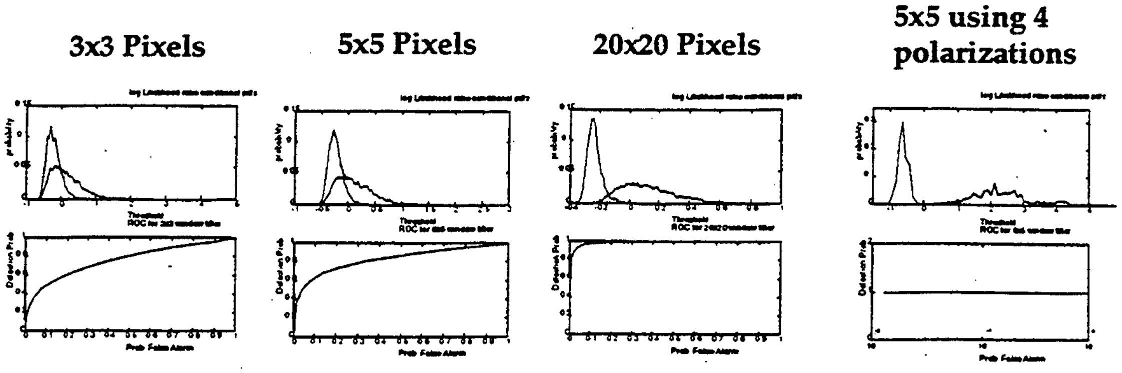

Figure 2. Universal Classification Performance using increasing spatial and polarization information

A typical Constant False Alarm Rate (CFAR) detector, which is the typical pre-screening method, uses a single pixel test that operates no better than the first column case. Even though our experiment was performed on two classes of clutter, target classes have even stronger energy returns and would converge (to zero error) at a faster rate. If a target, however, is imbedded in clutter, for example trees, there exist no algorithm that can extract a target where no signature information exists. But this method, nevertheless, would do very well against partially obstructed or occluded targets. The performance would degrade in proportion to the loss of signature. This algorithm, in its present form is not designed to discriminate between targets, but extensions exist. One such method has been implemented by Warke and Orsak[7] that has achieved very high classification rates to the face recognition problem.

3. WAVELET DECOMPOSITION

In this section we investigate the second stage to our ATR system. What can wavelets contribute? By definition, wavelets should offer some potential because of their properties of providing good localization in both the spatial and frequency domains. There has been a number of recent attempts to exploit such benefits in the pre-screener or detection stages of the ATR problem[(8, 9)] and a few attempts to tackle the challenging later stages of an ATR process[(10)]. This paper is strictly an attempt to exploit the attractive properties of wavelets, the authors are neither wavelet experts nor crusaders. Wavelets have generated a tremendous interest in both theoretical and applied areas. This is due to the interesting properties of a wavelet representation in terms of scale and location. In mathematics and engineering it has been known for some time that techniques based on Fourier series and Fourier transforms are not quite adequate for many problems. One of the assumptions is that the original time-domain function is periodic. As a result, the DFT has difficulty with functions that have transient components, i.e., components which are localized in time. This is particularly true with images that have frequent texture transitions. The other problem with

[1] Analysis indicated that for this data, the cross polarization pairs were not redundant.

trigonometric methods is that the transform does not convey any information pertaining to translation of the signal in time other than the phase of each Fourier coefficient. The "compact support" of the wavelet transform allows a representation that is not only localized in frequency but in time as well. There are a large body of useful overview papers concerning wavelets[11,12,12]. MRA[13] provides a means of constructing orthonormal bases of wavelets spanning the space of finite energy signals. These wavelets can be grouped by their scaling constant into disjoint subsets spanning orthogonal subspaces. These subspaces corresponding to different scales can be thought to represent different resolution levels.

The discrete wavelet transform of a function *f* with respect to ψ is

$$W_{m,n} = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(x)\psi^{*}\left(a_0^{-m} - nb_0\right)dx \qquad (1)$$

where $f, \psi \in L^2(\Re)$, and $\Re$ is the set of real numbers. In general for the DWT $a = a_0^m$, where m is an integer and $a_0 \neq 1$. The most common choice is to set $a_0 = 2$. For the translation parameter, $b = nb_0 a_0^m$ where $b_0 > 0$. ψ is called the mother wavelet and is assumed to be admissible for all functions, $f \in L^2(\Re)$ if there exists real numbers, $A > 0$ and $B < \infty$, such that

$$A\|f\|^2 \leq \sum_{m,n} \left|\langle f, \psi_{m,n} \rangle\right|^2 \leq B\|f\|^2$$

With the surprising and fortunate recent discovery of many such ψ functions such that *A=B=1* by Mallat[9], Daubechies[13], and others; the discrete wavelet transform offers an orthonormal bases in $L^2(\Re)$. In practice we use a discrete time version of (1) and implement it using a 2-dimensional version of the "pyramid scheme" described by Mallat[9]. This is a fast *O(N)* algorithm for an *N*-pixel image. For choices of ψ, we investigated the *haar, daublet, symmlet,* and *coiflet*. Each mother wavelet is depicted in Figure 3. We found that a *coiflet* with six taps using a three scale decomposition (decomposition grid in Figure 4) provided us with good discrimination with the Xpatch data[2]. The properties offered by this choice (in decreasing importance) include orthogonality, compact support, nearly symmetrical, and vanishing moments. Even though we haven't rigorously compared the performance of the wavelet filter options yet, we suspect that any choice would have been satisfactory (with the exception of the *haar*). This is a reasonable assertion based on their similar properties and the natural *automatic* feature isolating properties of the transform. We also incorporate a reflection boundary procedure that reflects the original signal at the boundary and then periodically extends it using the algorithm given by Brislawn[14]. This is not necessary for the orthogonal wavelets, but helped prevent the non-symmetrical ones from drifting in phase.

[2] No formal quantitative analysis has been examined as of yet.

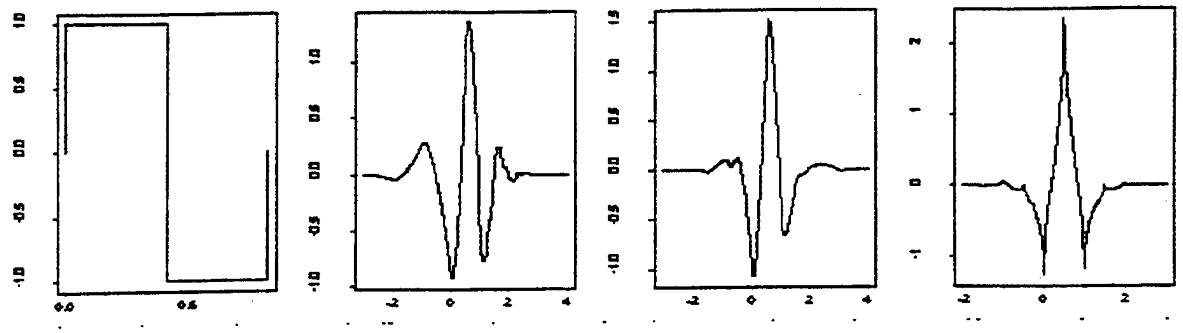

Figure 3. Investigated mother wavelets.

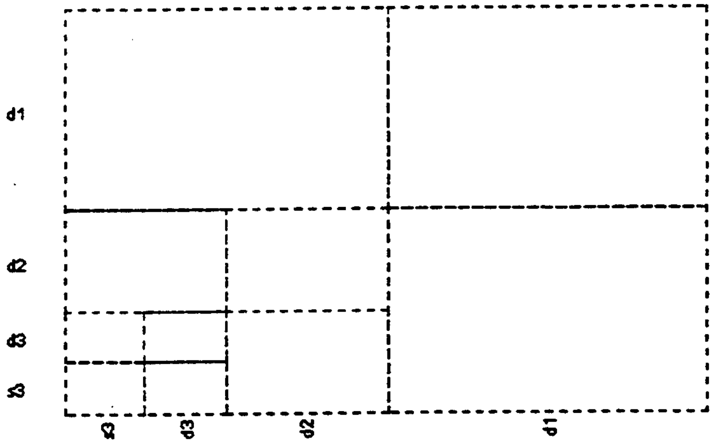

Figure 4. Wavelet scale decomposition

4. GRAM-SCHMIDT IMAGE CLASSIFIER

To exploit this feature information we implement our version of a classical correlation classifier[15] using a signal subspace technique popularized in communications to get target features where each signal is defined by a separate basis function in an orthonormal set. If this set can be described, then the optimum classifier easily follows, hence the minimum probability of error. One method of determining a set of orthonormal basis functions from a signal set is by using the Gram-Schmidt procedure.

Starting with a target set having an angular coverage of *q* degrees. We use a sub-set for designing the classifier. The classifier is designed by processing a target set through a Gram-Schmidt analysis to determine a sub-set that best accounts for the characteristics of all the images within the target space.

Let the design images be denoted by vectors X_1, X_2,..., X_N. The residuals of each image are combined into a matrix formed from the vectors. The decomposition process requires that we choose a subset of images that provides the best representation of the target set. We refer to this subset of images as the Gram-Schmidt (GS) set. The GS algorithm is described in many linear algebra textbooks(16). In our case, an orthonormal set $\boldsymbol{Q}$ is produced by the algorithm from the residuals to make up the components of the newly reconstructed design set Y, such that:

$$Y_i = \sum_{i=1}^{N} Q_i$$

where N is the rank of $\boldsymbol{Q}$.

A subset consisting of N images per target is selected for the classifier design. The classifier operates on the unknown data using the GS-set as follows. A test image X_t is projected onto the N-dimensional space spanned by the GS-set.

The components of the projection are calculated as follows:

$$\Theta_i = (X_t - \overline{X}) \bullet Y_i^T$$

where $\overline{X}_t$ is the residual test image. This test will determine if there is enough energy contained in the test image to fall within one of the target classes. The .scalar components are placed into an N dimensional feature vector:

$$\Theta_t = [\Theta_1, \Theta_2, \ldots, \Theta_N]$$

The max(Θ_i) for $i = 1 \ldots N$ is the distance decision measure corresponding to the best projection onto the GS-set. Hence, if the GS-set completely characterizes the target set then all of the target images should project onto one of the N vectors completely.

4.1 EXPERIMENTAL RESULTS

For our experimental analysis of the discrete wavelet transform Gram-Schmidt classifier (DWTGS) we use a large data base provided by the Model Based Vision Laboratory of Wright Patterson Air Force Base. This data base was produced as an initial test set for the Moving and Stationary Target Recognition (MSTAR) program. Since it contains four target classes we'll refer to it as the MSTAR-4 data. We formed images using 141 phase histories in angle to obtain one foot cross range resolution and 101 frequencies (approximately 500MHz bandwidth) to produce one foot range resolution. No padding or use of powers of two are needed (with 128 frequencies, range resolution would be 0.78 feet, with 50% oversampling, each sample is now 0.66 feet, but resolution is still one foot). We use a 2D Hanning window for 30dB sidelobe suppression. Our initial tests use two of the military objects contained in the MSTAR-4 data. We'll refer to these objects as T1 and M1. We restrict ourselves to 180 degrees in aspect and use a constant elevation angle of 30°. Using a GS set of 10 spanning 180 degrees (basically every 18°) we formed our filter set for each of the two classes. Using independent data we tested this against 180 test images spanning the same aspect coverage. The confusion matrix results are:

	T1	M1
T1	180	0
M1	6	174

The average probability of correct classification is 98.3%. These preliminary results are very encouraging for this data which visually provides very little discrimination potential (Figures 5 and 6). Further analysis is underway, however, to compare the performance with other classification strategies and to assess its relative performance.

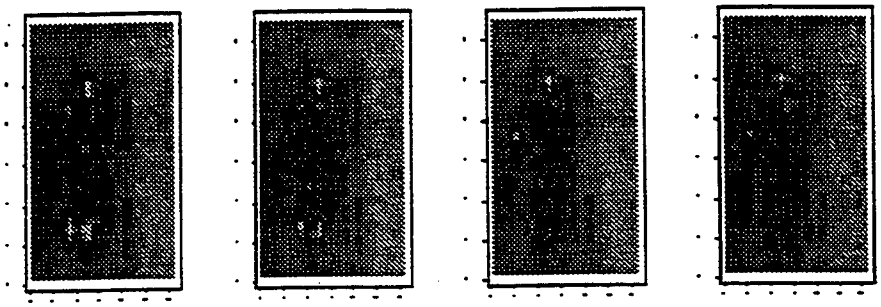

Figure 5. Aspect 0-4 degrees of object T1

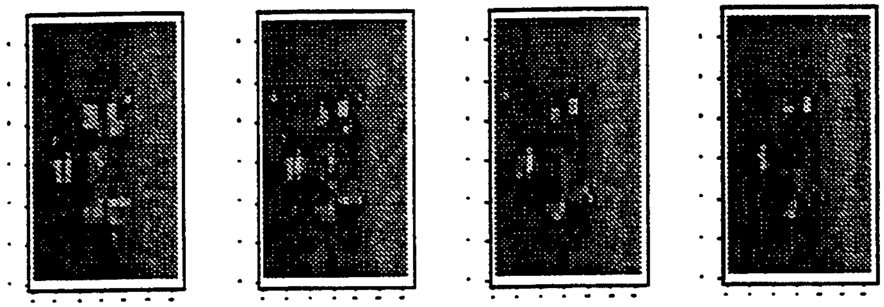

Figure 6. Aspects 0-4 degrees of object M1

5. SUMMARY

We have presented a novel hybrid ATR system using two new techniques in detection theory and signal analysis. The first part uses the so-called Universal Classifier that demonstrates a brand new class of powerful and efficient detectors that are able to significantly outperform traditional methods (in unknown environments). One of the standard approaches in detection analysis dealing with high dimensional problems is to break up a complicated signal into many simple pieces and study each of the pieces separately. This is the philosophy behind our use of wavelet based MRA. The second part of our hybrid system uses wavelet based MRA for demonstrating a new way of extracting feature information for classification. We classifier. Successful experiments implementing these approaches were demonstrated using both real and synthetic SAR data.

6. REFERENCES

[1] L.M. Novak, G.J. Owirka, C.M. Netishen, "Radar target identification using spatial matched filters," Pattern Recognition Journal, Vol. 27, No.4, pp. 607, 1994

[2] N. Warke, G.C. Orsak, "On the theory and application of universal classification to signal detection," Proc. 1994 IEEE Information Theory Workshop on Information Theory and Statistics, Alexandria, VA, October 27-29.

[3] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," IEEE Trans. Inform. Theory, Vol IT-35, pp. 401-408, Mar 1989.

[4] N. Merhav, "Universal classification for hidden markov models," IEEE Trans. Inform. Theory, Vol. IT-37, pp. 1586-1594, Nov 1991.

[5] A. Hauter, K. Chang, S. Karp, "Polarimetric fusion for synthetic aperture radar," Algorithms for Synthetic Aperture Radar Imagery III," SPIE Vol. 2757, Orlando, FL, April 1996

[6] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York,1991

[7] N. Warke, G. Orsak, "An information theoretic methodology for noisy image classification with application to face recognition," 1996 Conference on Information Science and Systems. Princeton, NJ, March 1996

[8] N.S. Subotic, B.J. Thelen, J.D. Gorman, "Multiresolution detection of coherent radar targets," Submitted to IEEE Transactions on Image Processing

[9] D. Wei, H. Guo, J. E. Odegard, M. Lang and C. S. Burrus, "Simultaneous speckle reduction and data compression using best wavelet packet bases with applications to SAR based ATD/R," SPIE Symposium on OE/Aerospace Sensing and Dual Use Photonics, Mathematical Imaging: Wavelet Applications for Dual Use," Orlando, FL, 17-21 April 1995

[10] N. Saito, R.R. Coifman, "Local discriminant bases," Mathematical Imaging: Wavelet Applications in Signal and Image Processing II," SPIE Vol. 2303, Orlando, FL, April 1994.

[11] O. Rioul, M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing, Oct. 1991.

[12] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Inform. Theory, Vol. IT-36, pp. 961-1005, Sep 1990.

[13] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans. On pattern Analysis and Machine Intell, Vol. 11, No. 7, pp.674-693, Jul 1989.

[14] C.M. Brislawn. "Classification of symmetric wavelet transforms," Technical report, Los Alamos National laboratory, Los Alamos, New Mexico, 87545, 1992.

[15] A. Hauter, K. Chang, S. Karp, "Polarimetric fusion for synthetic aperture radar," Algorithms for Synthetic Aperture Radar Imagery III," SPIE Vol. 2757, Orlando, FL, April 1996

[16].J.Wozencraft and I.Jacobs, *Principles of Communication Eng.*, pp. 266-272, Prospects Heights, Ill., Waveland, 1965.

SMC 96'

Efficient Algorithms for Learning Probabilistic Networks

Kuo-Chu Chang and **Jun Liu**

Center of Excellence in Command, Control,
Communications, and Intelligence(C^3I)
School of Information Technology and Engineering
George Mason University
Fairfax, VA 22030, U. S. A.

kchang@gmu.edu, jliu@gmu.edu

ABSTRACT

During past years, several methods have been developed for learning Bayesian networks from a given database. Some of these algorithms, due to their inherent nature, are computational intensive, and others which employing a greedy search heuristic can not guarantee to obtain an I-map (independency map) of the underlying distribution of the data, even if the sample size is sufficiently large. The focus of this paper is on developing efficient methods for learning Bayesian networks. A number of attributes about a Bayesian network and learning metric are identified. Based on these properties, new learning algorithms are developed which can be shown to be computationally efficient and guarantee the resulting network converging to a minimal I-map given a sufficiently large sample size.

1. INTRODUCTION

In recent years, research of Bayesian network technology has been mainly focused on two directions, Bayesian network inference algorithm and structure construction and refining. In aspect of structure construction, traditional AI researchers used expert knowledge to construct Bayesian networks. More recently, AI researchers and statisticians have begun to develop new methods for learning these networks. These methods combine prior knowledge with data to produce one or more Bayesian networks. The resulting networks can be used for inference or, in special cases, to infer causal relationships among variables. A Bayesian network structure when associated with a set of conditional probability distributions defines uniquely a joint probability distribution (jpdf) on the n domain variables. However, a given jpdf may have at most as many as $n!$ (minimal) Bayesian network representations. Hence, finding the simplest network representation (a network which has the least number of arcs) among these minimal network representations is considered NP-hard.

When learning a Bayesian network from data, the computational complexity is a major concern. There are a number of learning algorithms relying on a so called conditional independence test. For example, in Srinivas's method [7], the recursive algorithm used for building a sparse Bayesian network of $k+1$ variables based on the existing network K of k variables is to find, for the $k+1$th variable x, a minimal subset Z in K such that x is conditionally independent of K-Z given Z. One way to do this is by generating all possible subsets of K in increasing order of size until Z is found. Such exhaust search requires 2^k independence checks when adding the k+1th node. The total number of independence checks is approximately $O(2^n)$. When the number of domain variables is large the algorithm is computationally intractable. Another learning algorithm, CONSTRUCTOR, developed by M. Fung et al. [4] is also suffering from the similar problem.

The other category of methods to construct Bayesian networks from database is by defining a *search metric* (or *score function*) which is a function of the network structure and the database. Then a search strategy is employed to identify a network structure which has the maximum score among all possible network structures. A typical algorithm of such category is the Bayesian method (K2) developed by Cooper and Herskovits [3]. In the algorithm, they defined a score function, $p(B_S,D)$, which is proportional to the posterior probability $p(B_S|D)$. By giving a predefined node order, the problem of searching a network structure which maximizes the score function becomes to search a parent set for each node that maximizes the local score. Since a greedy method is used when searching the parent set for each node, the algorithm can not guarantee to find a network structure which has maximum score among all structures obeying the given node order. In addition, it can be shown that in general the algorithm can not guarantee to find any I-map[1] [2].

This paper is organized as follows. In Section 2, we introduce notational conventions, definitions, and assumptions used in the remainder of the paper. In Section 3, the new learning algorithms are presented which can be shown to be computationally efficient and guarantee to find a minimal I-map of the underlying distribution of the database when the sample size is sufficiently large. In section 4, we give the conclusion remarks and future work.

[1] A network topology guarantees that nodes found to be separated correspond to independent variables [6].

2. PRELIMINARIES

Throughout the discussion, we consider a domain U of n discrete random variables (r.v.), x_1, ..., x_n. Each variable has a finite number of values. The lower-case letters refer to r.v.'s, and upper-case letters refer to sets of r.v.'s. Let p be a joint probability distribution over U. Let X, Y, and Z be disjoint subsets of U. We use $p(X|Y)$ to represent the conditional probability distribution function (cpdf) for X, given all possible instantiations of Y. We say that X are *conditionally independent* of Y given Z, denoted as $I(X,Z,Y)$, if $p(X|ZY) = p(X|Z)$ (or equivalently $p(X,Y|Z) = p(X|Z)\,p(Y|Z)$) for all possible value assignments of X, Y, and Z.

An observation or a sample over U is a value assignment to all variables in U. A database D of observations over U is a list of observations. In this paper, we assume that the observations in the database are independent of each other. Further more we assume that there are no observations with missing values in the database.

Definition 1 Let B_S be a directed acyclic topological structure for the domain U and B_p is a set of specified conditional pdf $p(x_i\,|\Pi_i)$, $i = 1, ..., n$, where Π_i is the parent set of x_i defined by B_S. Then a *Bayesian network B* is defined to be a pair (B_S, B_p).

A Bayesian network structure B_S specifies for each node x_i a parent set Π_i, i=1, ..., n. Thus, we can write $B_S = \{\Pi_1, ..., \Pi_n\}$. A Bayesian network for a domain U represents a jpdf over U. It can be shown [6] that a given Bayesian network (B_S, B_p) uniquely specifies a jpdf p *of U*, where

$$p(x_1,\cdots,x_n) = \prod_{i=1}^{n} p(x_i|\Pi_i)\,. \qquad (2.1)$$

On the other hand, given a jpdf p on a domain U, we always can find a Bayesian network $B = (B_S, B_p)$ defined on U such that (2.1) holds. Thus, B can be called *a Bayesian network of* the jpdf p. It is well known that a given jpdf may have many Bayesian network representations, that is, they all represent the given jpdf. An important property of a Bayesian network is its node ordering. A node ordering is a priority constraint such that if x_i precedes x_j in the ordering, then we do not allow structures in which there is an arc from x_j to x_i. Now, let $B = \{B_S, B_P\}$ be a Bayesian network of p with an order O: i_1, i_2, ..., i_n. Then it can be shown that

$$p(x_{i_k}|x_{i_1},\cdots x_{i_k-1}) = p(x_{i_k}|\Pi_{i_k})\,, \quad \text{for } k = 1, ..., n. \quad (2.2)$$

It follows that if $B' = \{B_S', B_P'\}$ is another Bayesian network of p with the same order O, then by (2.2)

$$p(x_{i_k}|\Pi'_{i_k}) = p(x_{i_k}|\Pi_{i_k})\,, \quad \text{for } k = 1, ..., n. \quad (2.3)$$

We can always identify a node order for a Bayesian network. Conversely, given a jpdf p and an arbitrary node ordering i_1, i_2, ..., i_n, it can be shown that there must exist a Bayesian network which has the given order and represents p. In fact, based on the chain rule of probability,

$$p(x_1,\cdots,x_n) = \prod_{k=1}^{n} p(x_{i_k}|x_{i_1},\cdots,x_{i_{k-1}})\,. \qquad (2.4)$$

The right hand side of (2.4) defines a Bayesian network of p which is fully connected. Note that a fully connected network carries no information about conditional independence assertions, so it is not very useful. In order to reveal as much conditional independence as possible, it is necessary to construct a Bayesian network which has no redundant arcs under the given node order. We introduce the following definitions:

Definition 2 Let $B = (B_S, B_p)$ is a Bayesian network of p with a given node order. B is called *minimal*, if any arc of B_S is removed, then p can not be represented by B for any B_P. A minimal Bayesian network structure is also called a *minimal I-map*.

Most of Bayesian network learning algorithms are to find a minimal I-map. If a node order is specified, the learning procedure is reduced to find for each node its parent nodes from all candidate nodes which are precedent in order.

3. LEARNING BAYESIAN NETWORKS

Learning a Bayesian network from a database D of observations comprises two tasks: learning the network structure B_S, and after a proper network structure is identified, estimating the set of conditional probabilities B_S. In this paper, the focus is on developing efficient algorithms in learning network structures. Once B_S is obtained, B_P can be estimated from database statistically.

A Conditional Independence Test (CI) Approach

Some of the Bayesian network learning algorithms employ a so called conditional independence test (CI). The typical example include the algorithm developed by Srinivas [7] and the CONSTRCTOR by Fung [4].

Srinivas's method is designed to construct a sparse network. The algorithm begins from an empty network, and, recursively adds one node at each step to the existing network until all n nodes are included in the network. More specifically, let X_k denote the set of k nodes in the existing network at kth step, then for each of n - k remained nodes, namely, $x \in U - X_k$, identify its parent nodes $\Pi_x \subseteq X_k$ such that $p(x|X_k) = p(x|\Pi_x)$. The node which has the least number of parents is chosen to be added to the existing network. At the first step, the algorithm needs to choose a root node based on either an expert prior knowledge or a random selection if no prior knowledge is available. To identify for a node its parent set at kth step requires a total of 2^k CI tests which is the total number of possible subset of X_k. Hence, the total number of CI tests needed for identifying a network is proportional to 2^n.

The algorithm, CONSTRUCTOR, suffers from the same problem. In the algorithm, the idea is to find for each node $x \in U$ a Markov blanket $B_x \subseteq U$. The Markov blanket shields

x from U-$\{x\}$, namely, $p(x \mid B_x) = p(x \mid U)$, and is comprised by x's parents, children, and child's parents. Based on the information provided by the Markov blankets for all nodes, a Bayesian network can then be constructed. Again, it is a nontrivial problem to identify B$_x$'s since it needs 2^{n-1} tests for each node and a total of $n2^{n-1}$ CI tests if an exhaust search is used.

Both algorithms mentioned above are computationally intractable because of their search requirement. In the remaining of this section, we will focus on developing an efficient search method to identify the parent set or the Markov blanket for each node. The problem can be formulated as follows: let $V \subseteq U$ - $\{x\}$, the goal is to find a smallest subset $\Pi_x \subseteq V$ such that

$$p(x|\Pi_x) = p(x|V). \qquad (3.1)$$

Using conditional independence symbol, (3.1) can be denoted as $I(x, \Pi_x, V - \Pi_x)$, namely, x and V-Π_x are conditionally independent given Π_x. Before deriving our efficient algorithm, we introduce some useful properties of the conditional independence which can be found in [6].

Let X, Y, Z, $W \subseteq U$ be mutually disjoint sets of variables. Then we have following properties:

weak union $\quad I(X, Z, WY) \Rightarrow I(X, ZW, Y)$ & $I(X, ZY, W)$,
intersection $\quad I(X, ZW, Y)$ & $I(X, ZY, W) \Rightarrow I(X, Z, WY)$,

where the intersection requires the probability distribution p be absolutely positive. In the follows we assume p is absolutely positive. Based on the properties, we then have the following theorem.

Theorem 1 Let $x \in U$, and $V \subseteq U$ - $\{x\}$ be a set of random variables. Then there exists a subset $\Pi \subseteq V$ such that $I(x, \Pi, V$-$\Pi)$ holds if and only if $I(x, V$-$\{y\}, y)$ holds for $\forall\, y \in V - \Pi$.

Proof: Necessity: for any $y \in V - Z$, let $Y = \{y\}$, $W = V - \Pi - \{y\}$, $Z = \Pi$, and $X = \{x\}$. Then the necessity part follows obviously by the weak union property.

Sufficiency: let $V - \Pi = \{y_1, y_2, ..., y_k\}$. Now we have

$$I(x, V\text{-}\{y_i\}, y_i) \quad \text{for } \forall\, i = 1, 2, ..., k. \qquad (3.2)$$

Let $X = \{x\}$, $Y = \{y_1\}$, $W = \{y_2\}$. and $Z = V - \{y_1, y_2\}$, then we have $I(X, ZW, Y)$ and $I(X, ZY, W)$. By the intersection property, we obtain $I(X, Z, WY)$, which is

$$I(x, \mathrm{V}\text{-}\{y_1, y_2\}, \{y_1, y_2\}).$$

Now let $Y = \{y_1, y_2\}$, $W = \{y_3\}$. and $Z = V$-$\{y_1, y_2, y_3\}$. Again using the intersection property, we obtain

$$I(x, V\text{-}\{y_1, y_2, y_3\}, \{y_1, y_2, y_3\}).$$

Repeating this procedure, eventually we obtain

$$I(x, V\text{-}\{y_1, ..., y_k\}, \{y_1, ..., y_k\}),$$

which is $I(x, \Pi, V - \Pi)$. ||

Theorem 2 Let $x \in U$, and $V \subseteq U$ - $\{x\}$ be a set of random variables, $Y = \{y$: $I(x, V$-$\{y\}, y)\}$, and $\Pi_x = V$-Y. Then we have $I(x, \Pi_x, V - \Pi_x)$ holds. Furthermore, Π_x is the smallest conditioning set in the sense that if there is another Π such that $I(x, \Pi, V - \Pi)$ holds, we must have $\Pi_x \subseteq \Pi$.

Proof: Note $Y = V - \Pi_x$, then by the definition of Y and theorem 1, it follows that $I(x, \Pi_x, V - \Pi_x)$ holds. The second part of theorem is proved as following:

$\quad\quad I(x, \Pi_x, V - \Pi_x)$ and $I(x, \Pi, V - \Pi)$
iff $\quad$ (by theorem 1)
$\quad\quad I(x, V$-$\{y\}, y)$ for $\forall\, y \in V - \Pi_x$, and
$\quad\quad I(x, V$-$\{y\}, y)$ for $\forall\, y \in V - \Pi$.
iff $\; I(x, V$-$\{y\}, y)$ for $\forall\, y \in (V - \Pi_x) \cup (V - \Pi) = V - \Pi_x \cap \Pi$.

By the definition of Y, we must have $V - \Pi_x \cap \Pi. \subseteq Y$, which implies $\Pi_x = V$-$Y \subseteq \Pi_x \cap \Pi$. Thus it follows that $\Pi_x \subseteq \Pi$. ||

Theorem 2 provides us an efficient methods to identify the smallest conditioning set for a given variable. To search for a node x its parent set $\Pi_x \subseteq V$, what needs to do is for each y $\in V$ to check if $I(x, V$-$\{y\}, y)$ holds, that is if x and y are conditionally independent given V-$\{y\}$. If the answer is "yes", y must not belong to Π_x, and, if the answer is "no", y must belong to Π_x. Thus the number of CI tests required to identify parent set is equal to the size of V. Back to Srinivas's method, at kth step, we need only $(n - k)k$ CI tests, and totally we only need $\sum_{k=1}^{n-1} k(n-1) \approx n^3/6$ CI tests. In CONSTRUCTOR, we only need $n(n$-$1)$ CI tests.

The other methods to construct Bayesian networks from database is by defining a search metric (or score function) which is a function of network structure B_S and the database D. A network structure B_S which has the maximum score among all possible network structures is identified. In the remaining of this section, we discuss two approaches, and develop an efficient search algorithm.

A Minimum Description Length Approach (MDL)

The MDL method is based on the minimum description length principle [1] which stems from coding theory. The goal of the method is to create a network structure that describes the database as accurately as possible with as few symbols as possible.

Let $U = \{x_1, x_2, ..., x_n\}$, where each x_i can take a value from $\{x_{i1}, x_{i2}, ..., x_{iri}\}$, $r_i \geq 1$, i $=1, ..., n$. Let D_N be a database with N observations over U. Let B_S denote a network structure over U, and for each variable x_i, let Π_i be the set of parents of x_i defined by B_S. Furthermore, for each Π_i, let w_{ij} denote the jth instantiation of Π_i, $j = 1, ..., q_i$, $q_i \geq 0$. Now, let N_{ijk} be the number of observations in D_N in which the variable x_i has the value x_{ik} and Π_i is instantiated as w_{ij}. Finally, let $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Then the description length $L(B_S, D_N)$ of the network structure B_S given the database D_N is defined by

$$L(B_S, D_N) = \log P(B_S) - N \times H(B_S, D_N) - \frac{1}{2} K \log N \qquad (3.3)$$

where $K = \sum_{i=1}^{n} q_i(r_i - 1)$, and

$$H(B_S,D_N) = -\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\frac{N_{ijk}}{N}\log\frac{N_{ijk}}{N_{ij}}.$$

The first term of (3.3) models the prior distribution on network structures. The second term represents the conditional entropy of the network structure B_S, and it is a non-negative quantity. In the third term, the factor K is the number of (independent) probabilities that have to be estimated from the database D for obtaining the probability tables B_P for the network structure B_S. With an increasing number of arcs, a network structure will be able to more accurately describe a jpdf which generates the database, so the entropy term $-N \times H(B_S, D_N)$ increases. However, the cost term 1/2 K log$N$ decreases when more arcs are added. The network structure with the highest quality will balance both these terms, and has the highest score in (3.3). For a problem domain U with n variables, the number of possible network structures is huge, thus an exhaust search is computationally prohibited. An alternative way is to use a greedy search. Unfortunately, the greedy search can not find a global maximum, and for certain type of distributions that generates the database, it can not lead to a I-map no matter how large the database is [2].

To develop an efficient algorithm, we first investigate some properties of the MDL score. Let U be defined as before, and $D_N = \{U_1, U_2, ..., U_N\}$ be N independent samples of U. Let $B = (B_S, B_P)$ be a Bayesian network with a parameter set $B_P = \{\theta_{ijk}\}$ where $\theta_{ijk} = p(x_i = x_{ik} \mid \Pi_i = w_{ij})$, i =1, ..., n, j = 1, ..., q_i, and k = 1, ..., r_i. It can be shown that the posterior log likelihood of D_N given B is:

$$\log p(D_N|B_S,B_P) = \sum_{h=1}^{N}\log p(U_h|B_S,B_P)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\log\theta_{ijk}^{N_{ijk}} = \sum_{i=1}^{n}\sum_{j=1}^{q_i}N_{ij}\sum_{k=1}^{r_i}\frac{N_{ijk}}{N_{ij}}\log(\theta_{ijk}). \quad (3.4)$$

A simple fact associated with the posterior log likelihood is described in the following lemma.

Lemma 1 Let $D_N = \{U_1, U_2, ..., U_N\}$ be N independent samples of U. Let $B = (B_S, B_P)$ and $B' = (B'_S, B'_P)$ be two Bayesian networks over U. If B and B' represent the same joint distribution p over U, then

$$\log p(D_N|B_S,B_P) = \log p(D_N|B'_S,B'_P).$$

Proof: Since B and B' represent the same joint distribution p of U, we have

$$\log p(D_N|B_S,B_P) = \log p(D_N|p) = \log p(D_N|B'_S,B'_P). \quad ||$$

Now let us examine the right hand side of (3.4). The expression in (3.4) can be maximized with respect to θs by using Shannon's inequality, that states $\Sigma_i a_i \log a_i \geq \Sigma_i a_i \log b_i$ for all $a_i, b_i \geq 0$ where $\Sigma_i a_i = \Sigma_i b_i = 1$, with the equal sign holds iff $a_i = b_i$, $\forall\ i$. Hence we have

$$\log p(D_N|B_S,B_P) \leq \sum_{i=1}^{n}\sum_{j=1}^{q_i}N_{ij}\sum_{k=1}^{r_i}\frac{N_{ijk}}{N_{ij}}\log\frac{N_{ijk}}{N_{ij}}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\log\left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}} = \log p(D_N|B_S,\hat{B}_P), \quad (3.5)$$

where $\hat{B}_p = \{\hat{\theta}_{ijk}\}$ with $\hat{\theta}_{ijk} = N_{ijk}/N_{ij}$, is the maximum likelihood estimator (MLE) of B_P with K degree of freedom, where K is defined as in (3.3). Note that the right hand side of inequality (3.5) is $-N\times H(B_S, D_N)$, so the entropy term in MDL score is the posterior log likelihood of D_N given B_S and the conditional probabilities which are estimated by relative frequencies from D_N.

Based on statistical theory of hypothesis testing [8, p.381], if the probability distribution p represented by B is the true distribution that generates the sample D_N, the log likelihood ratio $-2\log p(D_N|B_S,B_P)/p(D_N|B_S,\hat{B}_P)$ will converge, as the sample size $N \to \infty$, to a chi squared with K degree of freedom. Hence we have the following lemma.

Lemma 2 Let $D_N = \{U_1, U_2, ..., U_N\}$ be N independent samples of U with probability distribution p defined over U. Let $B = (B_S, B_P)$ be a Bayesian network defined over U. Then if the joint probability distribution represented by B is identical to p, we have as $N \to \infty$,

$$2(-N\times H(B_S,D_N) - \log p(D_N|B_S,B_P)) \;\to\; \chi_K^2$$

in distribution, where χ_K^2 means a chi square distribution with K degree of freedom. ||

The proof is simple because $-N\times H(B_S,D_N) = \log p(D_N|B_S, \hat{B}_P)$.

When learning Bayesian network, it is desirable to find a minimal I-map of p. To attain this purpose, we need to identify some asymptotic properties associated with the MDL score. Combining Lemma 1 and Lemma 2, we have the following theorem.

Theorem 3 Let $D_N = \{U_1, U_2, ..., U_N\}$ be N independent samples of U with an absolute positive probability distribution p over U. Let $B_S^I = \{\Pi_1^I,\cdots,\Pi_n^I\}$, $B_S^c = \{\Pi_1^c,\cdots,\Pi_n^c\}$ and $B_S = \{\Pi_1,\cdots,\Pi_n\}$ be three network structures obeying the same node order, where B_S^I is a minimal I-map of p, B_S^c is a fully connected network structure, and B_S is obtained by deleting an arbitrary edge of B_S^c. Then for any large positive number $M >> 1$, we have

$$P(L(B_S,D_N) - L(B_S^c,D_N) > M) \to 1 \qquad \text{as } N \to \infty,$$

if the deleted edge does not belong to B_S^I, that is $\Pi_i^I \subseteq \Pi_i$, $i = 1, ..., n$.

In order to prove the theorem, we need the following lemma,

Lemma 3 Let $\{X_N\}$ and $\{Y_N\}$ be random variable sequences which converge in distribution to random variable X and Y, respectively. Let $\{a_N\}$ be sequence that converge to minus infinity, then we have

$$P(X_N - Y_N \le a_N) \to 0 \qquad \text{as } N \to \infty.$$

Proof. For any $M_1 > 0$, and $M_2 > 0$, when N is sufficient large such that $a_N < -M_2$, we have

$$\begin{aligned}
&\{X_N - Y_N \le a_N\} = \{X_N \le a_N + Y_N\}\\
&= (\{X_N \le a_N + Y_N\} \cap \{Y_N \le M_1\}) \cup\\
&\qquad (\{X_N \le a_N + Y_N\} \cap \{Y_N > M_1\})\\
&\subseteq \{X_N \le a_N + M_1\} \cup \{Y_N > M_1\}\\
&\subseteq \{X_N \le M_1 - M_2\} \cup \{Y_N > M_1\}
\end{aligned}$$

which implies

$$\begin{aligned}
&P(X_N - Y_N \le a_N) \le P(X_N \le M_1 - M_2) + P(Y_N > M_1)\\
&= F_{X_N}(M_1 - M_2) + 1 - F_{Y_N}(M_1).
\end{aligned}$$

where F denote the cumulated distribution function. Thus,

$$\lim_{N\to\infty} P(X_N - Y_N \le a_N) \le F_X(M_1 - M_2) + 1 - F_Y(M_1).$$

Letting $M_2 \to \infty$ followed by $M_1 \to \infty$, we obtain

$$\lim_{N\to\infty} P(X_N - Y_N \le a_N) = 0 . \qquad \|.$$

Now we begin to prove Theorem 3.

Proof (Theorem 3). It is equivalent to show

$$P(L(B_S, D_N) - L(B_S^c, D_N) \le M) \to 0 \qquad \text{as } N \to \infty.$$

The difference between two scores is

$$R - N \times H(B_S, D_N) + N \times H(B_S^c, D_N) + \tfrac{1}{2}(K^c - K)\log N\,, \quad (3.6)$$

where $R = \log P(B_S)/P(B_S^c)$ is a constant as $N \to \infty$, $K = \sum_{i=1}^n q_i(r_i - 1)$, and $K^c = \sum_{i=1}^n q_i^c(r_i^c - 1)$. Assume that B_S is obtained by deleting one of the parent nodes of x_i, that is $\Pi_i \subset \Pi_i^c$, and $\Pi_j = \Pi_j^c$, $\forall\, j \ne i$. Since p is absolutely positive, when N is sufficient large, all possible instantiations of Π_j^c must occur, thus $q_i^c > q_i$, and $q_j^c = q_j$, $\forall\, j \ne i$. Note that all $r_j = r_j^c$. We conclude K^c - $K > 0$, and hence, the third term in (3.6) goes to infinity.

On the other hand, since both B_S^c and B_S are super graphs of B_S^I , we can find B_P^c and B_P such that $B^c = (B_s^c, B_P^c)$ and $B = (B_S, B_P)$ both represent p. Then by Lemma 2, as $N \to \infty$,

$$X_N \equiv 2(-N \times H(B_S, D_N) - \log p(D_N | B_S, B_P)) \;\to\; \chi_K^2\,,$$

$$Y_N \equiv 2(-N \times H(B_S^c, D_N) - \log p(D_N | B_S^c, B_P^c)) \;\to\; \chi_{K^c}^2\,.$$

Note that by Lemma 1, $\log(D_N | B_S, B_P) = \log(D_N | B_S^c, B_P^c)$. We conclude

$$\begin{aligned}
&P(L(B_S, D_N) - L(B_S^c, D_N) \le M)\\
&= P(X_N - Y_N \le 2M - 2R - (K^c - K)\log N) \to 0 \text{ as } N \to \infty
\end{aligned}$$

where the limit in the last step follows the Lemma 3. ||

Theorem 3 tells us when search a minimal I-map we can begin from a fully connected network structure, then delete one edge and examine the MDL score. If the deleted edge is not an edge in the minimal I-map, the score will increase with high probability. However a question arises: what would happen if the deleted edge is an edge in the minimal I-map. The following theorem will answer this question.

Theorem 4 Let $D_N = \{U_1, U_2, \ldots, U_N\}$ be N independent samples of U with a absolute positive probability distribution p over U. Let $B_S^I = \{\Pi_1^I, \cdots, \Pi_n^I\}$, $B_S^c = \{\Pi_1^c, \cdots, \Pi_n^c\}$ and $B_S = \{\Pi_1, \cdots, \Pi_n\}$ be three network structures obeying the same node order, where B_S^I is a minimal I-map of p, B_S^c is a fully connected network structure, and B_S is obtained by deleting an arbitrary edge of B_S^c. Then almost surely (a.s.) we have

$$L(B_S, D_N) - L(B_S^c, D_N) \to -\infty\,, \qquad \text{as } N \to \infty,$$

if the deleted edge is an edge in B_S^I, that is, $\exists\, i$ such that $\Pi_i^I \not\subset \Pi_i \subset \Pi_i^c$, and $\Pi_j^I \subseteq \Pi_j = \Pi_j^c$ for $\forall\, j \ne i$.

Proof: As discussed in the proof of Theorem 3, the third term in (3.6), $\frac{1}{2}(K^c - K)\log N \to +\infty$, and the first term R is a constant. Now let us consider the behavior of the entropy term $H(B_S, D_N)$. By the strong law of large numbers, we have that

$$\frac{N_{ijk}}{N_{ij}} \to p(x_i = x_{ik} | \Pi_i = w_{ij}) \qquad \text{as } N \to \infty \text{ a.s.}$$

$$\frac{N_{ijk}}{N} \to p(x_i = x_{ik}, \Pi_i = w_{ij}) \qquad \text{as } N \to \infty. \text{ a.s.}$$

Then we have

$$\begin{aligned}
&-H(B_S, D_N) + H(B_P^c, D_N) \to \text{ a.s.}\\
&\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} p(x_i = x_{ik}, \Pi_i = w_{ij}) \log p(x_i = x_{ik} | \Pi_i = w_{ij}) -\\
&-\sum_{j=1}^{q_i^c}\sum_{k=1}^{r_i^c} p(x_i = x_{ik}, \Pi_i^c = w_{ij}^c) \log p(x_i = x_{ik} | \Pi_i^c = w_{ij}^c)\,.
\end{aligned}$$

By marginalization the above equation can be written as

$$\begin{aligned}
&\sum_{j=1}^{q_i^c}\sum_{k=1}^{r_i^c} p(x_i = x_{ik}, \Pi_j^c = w_{ij}^c) \log p(x_i = x_{ik} | \Pi_j^c = w_{i\sigma(j)})\\
&\qquad -\sum_{j=1}^{q_i^c}\sum_{k=1}^{r_i^c} p(x_i = x_{ik}, \Pi_i^c = w_{ij}^c) \log p(x_i = x_{ik} | \Pi_i^c = w_{ij}^c)\,.
\end{aligned}$$

where $w_{i\sigma(j)}$ conforms to w_{ij}^c. Again by Shannon's inequality, it can be shown the above expression is less than zero because B_S is not a I-map, and there are indexes j and k such that $p(x_i = x_{ik} | \Pi_j^c = w_{i\sigma(j)}^c) \ne p(x_i = x_{ik} | \Pi_j^c = w_{ij}^c)$. Thus we conclude as $N \to \infty$,

$$\begin{aligned}
&-N \times H(B_S, D_N) + N \times H(B_P^c, D_N)\\
&= N(-H(B_S, D_N) + H(B_P^c, D_N)) \to -\infty \qquad a.s.
\end{aligned}$$

Although the third term goes to infinity, because N term dominate $\log N$ term as $N \to \infty$, the difference between two MDL scores will converge to minus infinity almost surely. ||

Based on Theorem 3 and Theorem 4, we have the following algorithm. Given a database with N independent samples distributed with a absolutely positive joint pdf p, to construct a minimal I-map B_S^I of p that obeys a node order, we begin from a fully connected network structure B_S^c that obeys the same node order. Secondly, we examine all edges in B_S^c to determine which edge is in B_S^I and which is not. To do so, we deleting one edge at a time in B_S^c to generate a new network structure B_S. If the score difference between B_S and B_S^c is greater than zero, that is $L(B_S,D_N)-L(B_S^c,D_N)>0$, we believe with certain degree of confidence that the edge just deleted is not in B_S^I due to Theorem 4, otherwise, it is in B_S^I due to Theorem 3. Note that there are only $n(n-1)/2$ B_S 's to be examined, and each one has only one edge difference from B_S^c. Finally, those edges which make the score difference less than zero will be collected and comprise the minimal I-map. This algorithm is computationally efficient and guarantees to converge to the minimal I-map that obeys the given order provided that the sample size is sufficiently large.

A Bayesian Approach (K2)

As mentioned before, The philosophy of Bayesian learning methods, in principle, is based on a so called *score function* which is proportional to the posterior probability $p(B_S|D_N)$ of a network structure B_S given database D_N. A network structure B_S which maximizes the score function is considered to be the most likely structure generating the database. Cooper and Herskovits [3] used $p(B_S, D_N)$ as a score function, and based on a number of assumptions, they proved, for a discrete database D_N, that

$$P(B_S,D_N)=c\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}\,!, \qquad (3.7)$$

where c is the prior probability, $P(B_S)$, for each B_S, and r_i, q_i, N_{ijk}, and N_{ij} are defined as in MDL score. Bouckaert [3] investigated the difference between MDL score and Bayesian score, and concluded that

$$L(B_S,D_N)=\log P(B_S,D_N)+o(1) \qquad (3.8)$$

where $o(1)$ is with respect to N. That is as $N\to\infty$, $o(1)$ tends to a constant. However, the behavior when it tends to the limit depends on both B_S and D_N. From formula (3.8), we can show that the algorithm developed for MDL score can also be used for the Bayesian score. The only difference is that when examining if the deleted edge is in B_S^c, we use the log of Bayesian score. By slightly rewriting Theorem 3 and Theorem 4, it can be shown that the algorithm also converges to the minimal I-map provided the sample size in the database is sufficiently large. In addition, the computational efficiency is the same as in MDL.

4. CONCLUSIONS

In this paper, a number of attributes about a Bayesian network and score functions are identified. A set of theorems is derived based on which we derive two categories of Bayesian network learning algorithms, one for conditional independent test, and the other for MDL score and Bayesian score. Throughout the derivation, we show that the new algorithms are considerably simple and computationally efficient, and guarantee to converge to a minimal I-map of the probability distribution which generates the database, if the sample size is sufficiently large. The algorithms developed in this paper need to be evaluated using a simulation and real data, particularly for a database with a finite sample size. Learning Bayesian network from database is a difficult task because there is a huge number of possible network structures to be considered. Although it is known that given any node order, there must exist a minimal I-map of underlying distribution, an improperly-chosen node order may lead to a network which fails to reveal as much conditional independence as desired. Hence, to develop an algorithm that can, without relying on a prespecified node order, lead to a sparse I-map among all minimal I-map is an important future research direction.

5. REFERENCES

[1] Bouckaert, R. R., Belief network construction using the minimum description length principle. *Pro. European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*,pp.41-48, Springer-Verlag, 1993.

[2] Bouckaert, R. R, Bayesian belief networks: from construction to inference, Ph.D dissertation, University Utrecht, Dutch, 1995.

[3] Cooper, G. and Herskovits, E, A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, pp. 309-374, 1992.

[4] Fung, M., and Crawford, S. L., Constructor: A system for the induction of probabilistic models, *Proceedings of AAAI*, 1990, pp. 762-769, Boston, MA: MIT Press.

[5] Herskovits, E.H., *Computer-based Probabilistic Network Construction*, Doctoral dissertation, Medical Information Sciences, Stanford University, Stanford, CA,1991.

[6] Pearl, J, *Probabilistic reasoning in intelligent systems*. San Mateo, CA, Morgan Kaufmann, 1988.

[7] Srinivas, S., Russell, S., and Agogino, A., Automated construction of sparse Bayesian networks for unstructured probabilistic models and domain information. In M. Henrion, R.D. *et al* (Eds.), *Uncertainty in AI* 5, North Holland, 1990.

[8] George C., and Roger. L. B., *Statistical Inference*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, 1990.