# GRAPH-BASED MULTI-SENSOR FUSION FOR ACOUSTIC SIGNAL CLASSIFICATION

*Umamahesh Srinivas*[†] *Nasser M. Nasrabadi*[⋆] *Vishal Monga*[†]

[†]Department of Electrical Engineering, Pennsylvania State University, USA
[⋆]U.S. Army Research Laboratory, Adelphi, MD, USA

## ABSTRACT

Advances in acoustic sensing have enabled the simultaneous acquisition of multiple measurements of the same physical event via co-located acoustic sensors. We exploit the inherent correlation among such multiple measurements for acoustic signal classification, to identify the launch/impact of munition (i.e. rockets, mortars). Specifically, we propose a probabilistic graphical model framework that can explicitly learn the *class conditional correlations* between the cepstral features extracted from these different measurements. Additionally, we employ symbolic dynamic filtering-based features, which offer improvements over the traditional cepstral features. Experiments on real acoustic data sets show that our proposed algorithm outperforms conventional classifiers as well as recently proposed joint sparsity models for multi-sensor acoustic signal classification.

***Index Terms***— Acoustic signal classification, discriminative graphs, multiple measurements, symbolic features.

## 1. INTRODUCTION

The automatic classification of transient acoustic signals is of relevance both in military [1] and civilian [2–5] applications. Acoustic signals are highly non-stationary in nature, and joint processing of time- and frequency-domain information is necessary to effectively model the inherent signal structure. Coupled with the presence of ubiquitous real-world distortions during acquisition, this makes acoustic classification a challenging task. Many types of features have been proposed to capture discriminative signal information which is useful for classification [1, 2, 6–9]. Powerful classifiers widely used in machine learning, such as neural networks [10] and support vector machines (SVMs) [11, 12], have been employed to classify these features. All these classification schemes have been designed to utilize information from a single sensor recording measurements of the physical events.

Exploiting information from multiple sensing sources for the purpose of robust signal classification is an area of active research interest. Advances in acoustic sensing have facilitated the simultaneous acquisition of multiple measurements of the same physical event via co-located acoustic sensors. Zhang *et al.* [13] have recently demonstrated that multichannel acoustic signal classification offers better robustness than its traditional single-channel counterparts. Their work has also introduced sparsity-based classification, a seminal algorithmic advance first proposed for face recognition [14], to the acoustic signal processing community. Central to sparse representation-based classification (SRC) is the assumption that every signal representation can be approximately represented as a linear combination of similar training samples from the same class. The coefficient vector obtained as the solution of the resulting sparse recovery problem, when the dictionary contains training from *all* classes, naturally encodes the class identity. The joint sparse representation-based classification (J-SRC) scheme in [13] extends this by solving a matrix row-sparsity problem which enforces the common class association of the multiple measurements.

In this paper, we propose a general framework to fuse information from multiple sensors for acoustic classification using probabilistic graphical models. Features corresponding to measurements from multiple co-located acoustic sensors provide complementary yet correlated information useful for classification. We learn the *class conditional correlations* among these multiple feature sets in a two-stage framework. First, we learn a pair of *discriminative tree graphs* (one for each class) for each distinct set of features [15]. The sets of (as yet) unconnected tree graphs per class are representative of naive Bayes classification schemes, where different feature sets are considered to be independent. Then, we add new edges to these disjoint graphs iteratively via boosting [16], thereby learning correlations across different features. Finally, we learn a discriminative classifier that captures inter-feature correlations most crucial for discrimination.

Additionally in this paper, we utilize symbolic dynamic filtering-based features (SDF), inspired by work in [17]. We consider acoustic data collected from the multichannel tetrahedral acoustic sensor array developed by the U.S. Army Research Laboratory. Four co-located microphone sensors simultaneously measure acoustic activity for the launch and impact of rockets and mortars. Experimental comparisons reveal the merits of exploiting multi-sensor information as well as improvements over the joint sparsity scheme [13].

### 1.1. Relation to Prior Work

Our work is motivated by very recent work [13] that exploited, for the first time, multi-sensor information for acous-

tic signal classification. In [13], a joint sparsity approach is introduced to capture feature correlations. The inherent discriminative ability of sparse feature representations leads to a simple *reconstruction* residual-based class assignment scheme. Our work is the first to propose a *discriminative graphical classifier* that explicitly learns inter-feature dependencies for the purpose of multi-sensor acoustic signal classification. Additional novelty in our approach is claimed through the choice of symbolic dynamics-based features [17], which have been recently shown to offer better robustness under noise and other real-world distortions, compared to traditional cepstral features.

## 2. ACOUSTIC FEATURE REPRESENTATIONS

### 2.1. Cepstral Features

The power cepstrum [6] of a signal $x[n]$ is given by:

$$\left|\mathcal{F}\left(\log\left(|\mathcal{F}(x[n])|^2\right)\right)\right|^2, \tag{1}$$

where $\mathcal{F}(\cdot)$ represents the Fourier transform. Intuitively, the power cepstrum captures the rate of change of information content across different frequency bands. It has been widely used in speech and other acoustic signal processing tasks.

### 2.2. Symbolic Dynamic Filtering-based Features (SDF)

Symbolic time series analysis [17] has been proposed as an effective means of encapsulating time-series information. The central idea is similar in spirit to cepstral features, in the sense of capturing change across frequencies. The robustness exhibited by symbolic features to real-world noise in applications such as anomaly detection [17] has motivated us to apply them for acoustic signal classification. The schematic of symbolic dynamic filtering is shown in Fig. 1.

First, the coefficients from a wavelet decomposition of a given time signal are subjected to amplitude quantization. The amplitude range is divided into cells and each cell is represented by a unique symbol from an alphabet $\mathcal{A}$ of predetermined cardinality $|\mathcal{A}|$. A symbol sequence is obtained from the signal by replacing each wavelet coefficient with its corresponding symbol. The actual partitioning may be realized from a training set of signals either using simple uniform partitioning or maximum entropy partitioning (MEP). In this paper, we use the MEP method. Next, a state transition probability matrix (of dimension $|\mathcal{A}| \times |\mathcal{A}|$) is generated by computing the (frequentist) probabilities of transition between all pairs of symbols. Finally, the eigenvector corresponding to the unique unity eigenvalue of this matrix is chosen as the feature vector corresponding to that time-series signal. Our implementation differs slightly from the original technique [17] in that we directly quantize the amplitude range of the time signals, bypassing the wavelet decomposition.

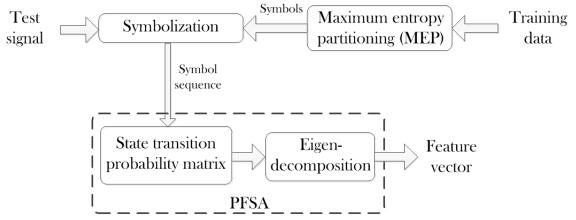



**Fig. 1**. Symbolic dynamic filtering-based features [17].

## 3. GRAPH-BASED MULTI-SENSOR FUSION

### 3.1. Discriminative Graphical Models

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a collection of nodes $\mathcal{V} = \{v_1, \ldots, v_r\}$ and a set of (undirected) edges $\mathcal{E} \subset \binom{\mathcal{V}}{2}$, i.e., the set of unordered pairs of nodes. A probabilistic graphical model is obtained by defining a random vector on $\mathcal{G}$ such that each node represents one (or more) random variables and the presence of edges indicates conditional dependencies. The graph structure thus approximates the joint probability distribution function by a product of terms representing pairwise and marginal statistics. Graphical models offer an alternate visualization of the correlations between the individual random variables in a multivariate probability distribution. They also enable us to draw upon the rich resource of efficient graph-theoretic algorithms to learn complex models and perform inference. Their use in applications has been motivated by practical concerns like insufficient training to learn models for high-dimensional data and the need for reduced computational complexity in realtime tasks [18, 19].

Traditionally, graphs have been learnt generatively [20], by minimizing the error of approximation to a given distribution. Of more relevance to our problem are advances in discriminative graph learning. We focus on a recent discriminative learning framework [15] wherein a pair of graphs is jointly learnt by minimizing the classification error. Specifically, the tree-approximate $J$-divergence (a symmetric extension of the Kullback-Leibler (KL) divergence) between two distributions $p$ and $q$ is maximized:

$$\hat{J}(\hat{p}, \hat{q}; p, q) = \int (p(x) - q(x)) \log\left[\frac{\hat{p}(x)}{\hat{q}(x)}\right] dx. \tag{2}$$

Based on the observation that maximizing the $J$-divergence minimizes the upper bound on the probability of classification error, the discriminative learning problem then becomes:

$$(\hat{p}, \hat{q}) = \arg \max_{\hat{p}, \hat{q} \text{ are trees}} \hat{J}(\hat{p}, \hat{q}; \widetilde{p}, \widetilde{q}), \tag{3}$$

where $\widetilde{p}$ and $\widetilde{q}$ are the available empirical estimates. It is shown in [15] that this optimization further decouples into two maximum-weight spanning tree (MWST) problems::

$$\begin{aligned} \hat{p} &= \arg \min_{\hat{p} \text{ is a tree}} D(\widetilde{p}||\hat{p}) - D(\widetilde{q}||\hat{p}) \\ \hat{q} &= \arg \min_{\hat{q} \text{ is a tree}} D(\widetilde{q}||\hat{q}) - D(\widetilde{p}||\hat{q}), \end{aligned} \tag{4}$$

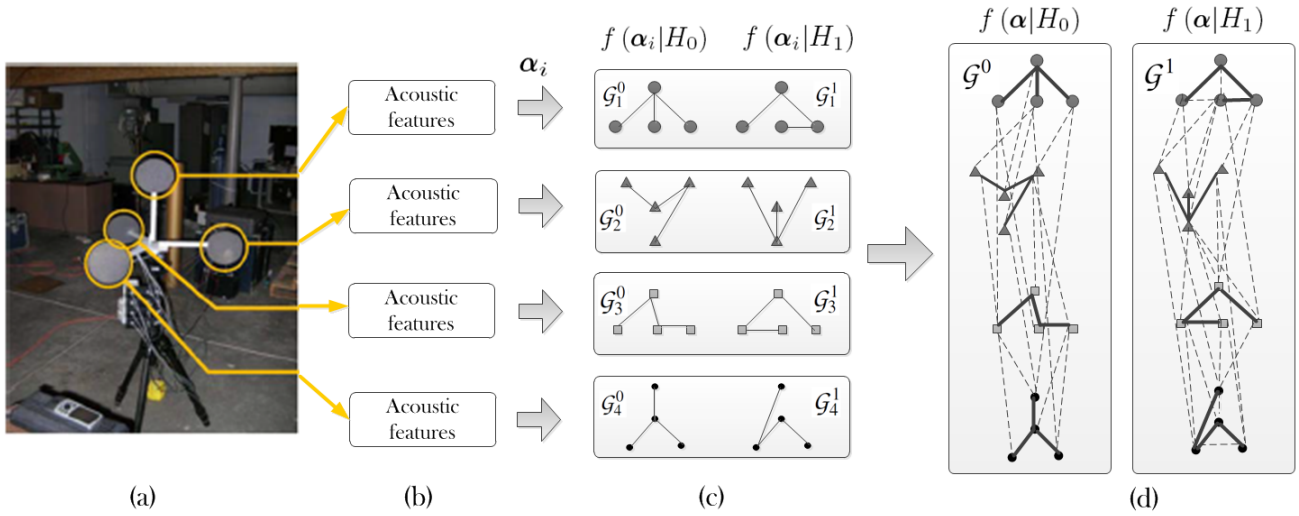



**Fig. 2**. Overall acoustic signal classification framework. (a) Four co-located acoustic sensors. (b) Feature extraction (cepstral or SDF-based features). (c) Individual pairs of trees learnt from each feature set. (d) Thickened graphical models capturing discriminative information via edges (conditional dependencies) across feature sets.

**Table 1**. Overall classification accuracy for the two-class rocket problem, using cepstral features.

| Method | CRAM04 | CRAM06 | Foreign |
|---|---|---|---|
| SVM | 0.7726 | 0.5845 | 0.8958 |
| CSVM | 0.7354 | 0.6063 | 0.9166 |
| J-SRC | 0.7694 | 0.6510 | 0.9094 |
| MSGM | 0.7812 | 0.6821 | 0.9108 |

**Table 2**. Overall classification accuracy for the two-class rocket problem, using SDF features.

| Method | CRAM04 | CRAM06 | Foreign |
|---|---|---|---|
| SVM | 0.7776 | 0.6079 | 0.8972 |
| CSVM | 0.7514 | 0.6221 | 0.9154 |
| J-SRC | 0.7814 | 0.6883 | 0.9121 |
| MSGM | 0.7966 | 0.6906 | 0.9146 |

**Table 3**. Overall classification accuracy for the two-class mortar problem, using cepstral features.

| Method | CRAM04 | CRAM05 | CRAM06 | Foreign |
|---|---|---|---|---|
| SVM | 0.8480 | 0.8127 | 0.8590 | 0.8364 |
| CSVM | 0.8449 | 0.8280 | 0.7971 | 0.7799 |
| J-SRC | 0.8701 | 0.8626 | 0.8727 | 0.8087 |
| MSGM | 0.8939 | 0.8853 | 0.8879 | 0.8201 |

**Table 4**. Overall classification accuracy for the two-class mortar problem, using SDF features.

| Method | CRAM04 | CRAM05 | CRAM06 | Foreign |
|---|---|---|---|---|
| SVM | 0.8603 | 0.8175 | 0.8623 | 0.8398 |
| CSVM | 0.8498 | 0.8361 | 0.8012 | 0.7846 |
| J-SRC | 0.8837 | 0.8793 | 0.8815 | 0.8161 |
| MSGM | 0.8996 | 0.8907 | 0.8892 | 0.8248 |

where $D(p||\hat{p}) = E_p[\log(p/\hat{p})]$ represents the KL-divergence. From (4), we see that the optimal choice of $\hat{p}$ ($\hat{q}$) minimizes its distance to $\tilde{p}$ ($\tilde{q}$) while simultaneously maximizing its distance from $\tilde{q}$ ($\tilde{p}$). The discussion so far considers tree graphs, which are fully connected acyclic graphical structures. Trees are easy to learn but their sparse edge connectivity limits the model complexity. On the other hand, optimally learning complex graphical models is NP-hard [21]. This inherent trade-off between generalization and performance is resolved in [15] by iteratively thickening the initial graph with more edges via boosting [16] to learn a richer structure.

### 3.2. Feature Fusion via Boosting on Disjoint Graphs

In this section, we introduce our proposed Multi-Sensor-Graphical-Model (MSGM) framework for acoustic signal classification using multiple correlated measurements. An illustration of the framework is shown in Fig. 2. We consider a binary classification problem[1] for ease of exposition. However, the method can be naturally extended to multi-class problems using a one-versus-all strategy. The algorithm consists of offline training followed by an online test stage. The discriminative graphs are learnt in the training stage.

First, features corresponding to training acoustic signals are obtained using techniques discussed in Section 2. For each acoustic signal, four different features $\boldsymbol{\alpha}_i \in \mathbb{R}^N, i = 1,\ldots,4$ are obtained as shown in Fig. 2(b). For each of the four sets of features, a pair of $N$-node discriminative tree graphs $\mathcal{G}_i^0$ and $\mathcal{G}_i^1$, which respectively approximate the class distributions $f(\boldsymbol{\alpha}_i|H_0)$ and $f(\boldsymbol{\alpha}_i|H_1)$, are simultane-

[1] One such problem for acoustic data is rocket launch vs. impact, denoted by hypotheses $H_0$ and $H_1$ respectively.

**Table 5**. Overall classification accuracy for the four-class problem, using cepstral features.

| Method | CRAM04 | CRAM05 | CRAM06 | Foreign |
|---|---|---|---|---|
| SVM | 0.7669 | 0.7648 | 0.7437 | 0.8036 |
| CSVM | 0.7648 | 0.7441 | 0.6874 | 0.7765 |
| J-SRC | 0.8036 | 0.7847 | 0.7431 | 0.7960 |
| MSGM | 0.8113 | 0.7955 | 0.7522 | 0.8115 |

**Table 6**. Overall classification accuracy for the four-class problem, using SDF features.

| Method | CRAM04 | CRAM05 | CRAM06 | Foreign |
|---|---|---|---|---|
| SVM | 0.7683 | 0.7669 | 0.7481 | 0.8080 |
| CSVM | 0.7682 | 0.7510 | 0.6923 | 0.7818 |
| J-SRC | 0.8092 | 0.7903 | 0.7489 | 0.8016 |
| MSGM | 0.8218 | 0.8067 | 0.7665 | 0.8243 |

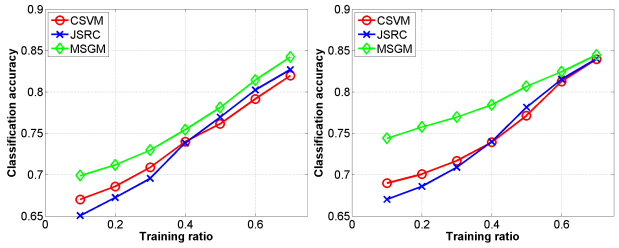
ously learnt (see Fig. 2(c)). The initial disjoint graphs with $4N$ nodes, representing the class distribution corresponding to $H_0$ and $H_1$, are then generated by separately concatenating the nodes of $\mathcal{G}_i^0, i = 1,\ldots,4$ and $\mathcal{G}_i^1, i = 1,\ldots,4$, respectively. These graphs with sparse edge structure are iteratively thickened via boosting [15, 16, 22]. Different pairs of discriminative graphs over the same sets of nodes with different weights are learned in different iterations, and the newly-learnt edges are used to augment the graphs. These new edges uncover correlations among the multiple feature sets that are crucial for discrimination. The final "thickened" graphs $\mathcal{G}^0$ and $\mathcal{G}^1$ are shown in Fig. 2(d).

The classification of a new test sample $\boldsymbol{y}$ is then performed online. Features $\boldsymbol{\alpha}_i, i = 1,\ldots,4$, are extracted from the test signal and concatenated to form $\boldsymbol{\alpha}$. Let $f(\boldsymbol{\alpha}|H_0)$ and $f(\boldsymbol{\alpha}|H_1)$ denote the probability distribution functions for the final graphs $\mathcal{G}^0$ and $\mathcal{G}^1$ learnt for $H_0$ and $H_1$ respectively. The class label of $\boldsymbol{y}$ is determined as follows:

$$\text{Class}(\boldsymbol{y}) = \begin{cases} \text{Launch} & \text{if } \log\left(\frac{f(\boldsymbol{\alpha}|H_0)}{f(\boldsymbol{\alpha}|H_1)}\right) \geq 0 \\ \text{Impact} & \text{if } \log\left(\frac{f(\boldsymbol{\alpha}|H_0)}{f(\boldsymbol{\alpha}|H_1)}\right) < 0. \end{cases} \quad (5)$$

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

Transient acoustic data are collected during the launch and impact of two types of artillery - mortar and rocket. For each event, four measurements are acquired using the tetrahedral acoustic sensor array in Fig. 2 (a), at a sampling rate of 1001.6 Hz. Raw acoustic signals are subjected to pre-processing using maximum spectral detection to localize the segment corresponding to the actual event (typically of duration 1 second). We consider four data sets for our experiments: CRAM04, CRAM05, CRAM06, and Foreign. For each data set, we present results for three different classification scenarios: (i) rocket launch vs. rocket impact, (ii) mortar launch vs. mortar impact, and (iii) the combined four-class problem. We select the first 50 cepstral coefficients as features and use an alphabet of size 30 (leading to a 30-D feature) for SDF.




(a) Cepstral features. (b) SDF features.

**Fig. 3**. Performance as a function of training ratio, for the rocket mortar vs. launch binary classification (CRAM04).

We compare the results of our MSGM approach with three methods: (i) SVM: results reported as the average of the four channels classified independently, (ii) CSVM: SVM on the concatenated vector of features from all four sensors, and (iii) J-SRC [13]. Tables 1-2 present the overall classification accuracy for the two-class rocket problem, while Tables 3-4 present the classification rates for the two-class mortar problem. The classification rates for the overall four-class problem are reported in Tables 5-6. In each experiment, we choose a training ratio of $r = 0.5$ and report average results from five different runs of the experiment. We identify two key trends: (i) our proposed MSGM approach consistently gives better classification performance than competing classifiers, and (ii) for the same choice of classifier, the SDF features lead to better overall classification when compared to cepstral features.

A novel significant insight is revealed by Fig. 3, which compares classification performance as a function of training ratio for the rocket two-class problem on the CRAM04 data set. As expected, all methods show a decrease in performance as the number of training samples decreases. However, the proposed MSGM method shows a more graceful degradation, unlike the sparsity-based techniques whose success is dependent on the availability of rich training sets.

## 5. CONCLUSION

We have proposed a graphical model-based feature fusion framework for multi-sensor acoustic signal classification. Experiments reveal its improved classification performance over competing classifiers. In future work, we will: (i) compare performance with state-of-the-art audio event classification techniques, and (ii) explore connections between sparse feature representations and graphical models.

## ACKNOWLEDGMENT

The authors would like to acknowledge the US Army Research Laboratory's Acoustics & EM Sensing Branch for providing the fund and data for the Acoustic transient event classification in support of Deployable Force Protection (DFP), Gunfire Detection Systems (GDS), and Counter Rocket, Artillery, and Mortar (CRAM) programs.

## 6. REFERENCES


[1] M. R. Azimi-Sadjadi, Y. Yang, and S. Srinivasan, "Acoustic classification of battlefield transient events using wavelet subband features," in *Proc. SPIE Defense and Security Symposium*, vol. 6562, 2007.

[2] S. Sampan, *et al.*, "Neural fuzzy techniques in vehicle acoustic signal classification," Virginia Polytechnic Institute and State University, Tech. Rep., 1997.

[3] G. Valenzise, *et al.*, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conf. Adv. Video Signal Surveill.*, 2007, pp. 21–26.

[4] A. Klausner, S. Erb, A. Tengg, and B. Rinner, "DSP based acoustic vehicle classification for multi-sensor real-time traffic surveillance," in *Proc. Europ. Signal Process. Conf.*, 2007, pp. 1916–1920.

[5] S. Z. Li, "Content-based classification and retrieval of audio using the nearest feature line method," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 619–625, 2000.

[6] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[7] R. E. Learned and A. S. Willsky, "A wavelet-packet approach to transient signal classification," *Applied Computational and Harmonic Analysis*, vol. 2, no. 3, pp. 265–278, 1995.

[8] M. Till and S. Rudolph, "Optimized time-frequency distributions for acoustic signal classification," in *Proc. SPIE, Wavelet Applications VIII*, vol. 4391, 2001, pp. 81–91.

[9] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1306–1309.

[10] A. Kundu, G. C. Chen, and C. E. Persons, "Transient sonar signal classification using hidden Markov models and neural nets," *IEEE J. Ocean. Eng.*, vol. 19, no. 1, pp. 87–99, 1994.

[11] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, pp. 209–215, 2003.

[12] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[13] H. Zhang, N. M. Nasrabadi, T. S. Huang, and Y. Zhang, "Transient acoustic signal classification using joint sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2011, pp. 2220–2223.

[14] J. Wright, *et al.*, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[15] V. Y. F. Tan, S. Sanghavi, J. W. Fisher, and A. S. Willsky, "Learning graphical models for hypothesis testing and classification," *IEEE Trans. Signal Processing*, vol. 58, no. 11, pp. 5481–5495, Nov. 2010.

[16] Y. Freund and R. E. Shapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, Sep 1999.

[17] V. Rajagopalan and A. Ray, "Symbolic time series analysis via wavelet-based partitioning," *Signal Processing*, vol. 86, pp. 3309–3320, 2006.

[18] S. L. Lauritzen, *Graphical Models*. Oxford University Press, NY, 1996.

[19] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[20] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 462–467, Mar. 1968.

[21] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, vol. 42, pp. 393–405, Mar. 1990.

[22] T. Downs and A. Tang, "Boosting the tree augmented naive Bayes classifier," in *Intelligent Data Engineering and Automated Learning, LNCS*, vol. 3177, 2004, pp. 708–713.