Risk Minimization Approaches in Signal Processing

by,

Joseph Picone and Jonathan Hamaker
Institute for Signal and Information Processing
Mississippi State University
Mississippi State, Mississippi 39762, USA

Aravind Ganapathiraju
Senior Speech Technologist
Conversay
Redmond, Washington 98052, USA

Abstract

Statistical techniques based on Hidden Markov models (HMMs) with Gaussian emission densities have dominated the signal processing and pattern recognition literature for the past 20 years. However, HMMs suffer from an inability to learn discriminative information and are prone to overfitting and over-parameterization. Recent work in machine learning has focused on models, such as the support vector machine (SVM), that automatically control generalization and parameterization as part of the overall optimization process. SVMs have been shown to provide significant improvements in performance on small pattern recognition tasks compared to a number of conventional approaches. SVMs, however, require ad hoc (and unreliable) methods to couple it to probabilistic learning machines. Probabilistic Bayesian learning machines, such as the relevance vector machine (RVM), are fairly new approaches that attempt to overcome the deficiencies of SVMs by explicitly accounting for sparsity and statistics in their formulation.

In the proposed paper, we will review the past 30 years of research into these new learning machines, and describe how they can be used to solve many traditional signal processing problems. Unifying themes in this work are the concepts of risk minimization and margin maximization, which can be viewed as a generalization of the maximum likelihood principle so fundamental to many signal processing approaches. It is our belief that this information has not been previously explained in a way that makes it accessible to mainstream signal processing researchers, so we believe this paper will have significant tutorial value.

Please direct all correspondence to: Joseph Picone, Institute for Signal and Information Processing, Box 9571, Mississippi State University, Mississippi State, Mississippi 39762, Tel: 662-325-3149, Fax: 662-325-2298, Email: picone@isip.msstate.edu.

A. Overview

Signal processing as a field has evolved significantly in the past 40 years. Simplistic techniques in which a premium was placed on the design of smart feature extraction algorithms (e.g., analog filterbanks) and an understanding of the basic physics of a problem (e.g., physiological models in speech processing) have given way to powerful statistical methods which automatically learn in data-driven modes. In some sense, we can point to a progression of techniques: ad-hoc measurements such as filter bank energies (popular in the early 1960's), Fourier methods (popular in the late 1960's and early 1970's), maximum entropy and least mean square error methods (popular in the late 1970's), temporal modeling techniques such as dynamic time warping (popular in the late 1970's and early 1980's), and Markov models (popular in the mid-1980's and early 1990's). By the mid-1990's, the field of machine learning had grown tremendously, generating many new competing techniques for the generation beyond Markov models. Most of these techniques are rooted in principles of maximum likelihood, and use the Expectation Maximization algorithm to estimate parameters and optimize performance today.

Statistical techniques based on Hidden Markov models (HMMs) with Gaussian emission densities have dominated the signal processing and pattern recognition literature for the past 20 years. However, HMMs suffer from an inability to learn discriminative information and are prone to overfitting and over-parameterization. Artificial neural networks (ANNs) have been proposed as a replacement for the Gaussian emission probabilities under the belief that the ANN models provide better discrimination capabilities. However, the use of ANNs often results in over-parameterized models which are prone to overfitting. Techniques such as cross-validation have been suggested as remedies to the overfitting problem but employing these is wasteful of both resources and computation. Further, cross-validation does not address the issue of model structure and over-parameterization.

Ironically, in the early 1970's, Vapnik introduced a discriminative learning technique based on the principle of risk minimization and maximum margin classification. This learning machine, known as a Support Vector Machine (SVM), has been shown over the years to provide near optimal performance on a wide variety of static classification tasks. Like all good innovations, it has taken a while for researchers to understand the original theory. Today, SVMs are quietly moving into the mainstream of the signal processing literature. However, this approach has serious drawbacks for many problems of great interest to signal processing researchers that involve data that cannot be separated by linear classifiers, or exhibits a large amount of ambiguity. Bayesian methods have previously been popular for dealing with such cases. Naturally, there is an integration of Bayesian methods and margin maximization emerging, and that is an important point of this paper. These approaches can be described from a unified theory based on risk minimization.

Similarly, recent work on machine learning has moved toward automatic methods for controlling generalization and parameterization. SVMs use the principle of structural risk minimization to simultaneously control generalization and performance on the training set. SVMs have some significant shortfalls. First, the SVMs are not probabilistic in nature and, thus, are not able to adequately express the posterior uncertainty in predictions. This is particularly important in signal processing applications in which there is significant overlap in the feature space. SVMs also make unnecessarily liberal use of parameters to define the decision region. Bayesian models which take the same form as an SVM model, but comprehend sparseness in their design using the concept of automatic relevance determination, have been found to provide generalization performance on par with SVMs while typically using nearly an order of magnitude fewer parameters. Therefore, **risk minimization** and **automatic relevance determination** become key concepts in understanding these contemporary approaches. There are few good tutorial papers on these topics that are presented in ways signal processing researchers can easily understand.

B. Outline

An outline of the proposed paper is shown below:

- I. Introduction
  - a. Historical overview of the evolution of signal processing
  - b. Bayesian methods in signal processing
  - c. Discrimination vs. Representation
  - d. Relationship to neural networks
- II. Risk Minimization
  - a. Expected Risk
  - b. Empirical Risk
  - c. VC Dimension
  - d. Principle of structural risk minimization
- III. Support Vector Machines
  - a. Support Vectors and Hyperplanes
  - b. Optimization
  - c. Kernels
  - d. Soft Margin classifiers
  - e. Experimental results
- IV. Sparse Bayesian Methods
  - a. Bayesian Modeling
  - b. The Evidence Framework
  - c. Relevance Vector Machines
  - d. Comparison of RVMs and SVMs
  - e. Experimental Results
- V. Applications in Speech and Signal Processing
  - a. Static pattern classification
  - b. Integration with temporal models
- VI. Summary and Conclusions

In addition to the technical material described above, we will reference readers to publicly available software that implements the ideas discussed above. These resources include data, Java and C++ code that make it easy to replicate the examples cited in the paper.

C. Biography

Dr. Joseph Picone is currently a Professor and Eminent Scholar in the Department of Electrical and Computer Engineering at Mississippi State University. He also serves as the Director of the Institute for Signal and Information Processing (ISIP). He founded ISIP in 1994 with a vision to develop and disseminate public domain speech recognition software. ISIP is now known worldwide as a leading provider of speech recognition software and training. Dr. Picone received his Ph.D. from Illinois Institute of Technology in 1983. He has published over 140 papers on speech processing. He holds 8 patents, is a Senior Member of the IEEE, and is very active in the IEEE and related professional organizations. He has previously worked at Texas Instruments where he was involved in research on speech recognition and compression for military and commercial applications, as well as AT&T Bell Laboratories, where he was involved in similar areas of research.

Dr. Picone has previously published in the *IEEE Proceedings*:

J. Picone, "Signal Modeling Techniques in Speech Recognition," *IEEE Proceedings*, vol. 81, no. 9, pp. 1215-1247, September 1993.

This widely read paper is considered to be one of the definitive references for signal processing in speech recognition. The topic of the current paper proposal is intended to eventually become part of a research monograph on applications of sparse Bayesian methods to speech recognition. Dr. Picone's vitae is available at http://www.isip.msstate.edu/administration/personnel/resumes/picone_joseph.pdf.

Jon Hamaker expects to complete his Ph.D. dissertation on this topic in Spring'03 in the Department of Electrical and Computer Engineering at Mississippi State University under Dr. Picone's supervision. He is responsible for the development of sparse Bayesian methods in speech recognition. His resume can be viewed at http://www.isip.msstate.edu/administration/personnel/resumes/hamaker_jonathan.pdf. His dissertation (which is still under development) is available at http://www.isip.msstate.edu/publications/books/msstate_theses/2003/relevance_vectors/thesis/thesis_v3.pdf. Mr. Hamaker has published 5 journal papers, 22 conference papers, and 3 invited presentations during his graduate studies at MS State. This includes an invited presentation at IBM's Thomas J. Watson Research Center. He also spent two summers as an intern at Microsoft. Mr. Hamaker has been a principal architect and implementer of the ISIP public domain speech recognition tools.

Aravind Ganapathiraju is currently a Senior Speech Technologist at Conversay, where he directs their research and development into speech recognition. He graduated from MS State with a Ph.D. in Computer Engineering in May 2002. His dissertation on the use of Support Vector Machines in speech recognition, which can be viewed at http://www.isip.msstate.edu/publications/books/msstate_theses/2002/support_vectors/, laid the groundwork for much of the material described in this paper. It was his experience in assimilating a vast amount of information from the machine learning community on discriminative training methods that was the inspiration for the proposed paper. It is our feeling that many signal processing engineers will benefit from making this information more accessible to the community. Mr. Ganapathiraju has published four journal papers and 30 conference papers in the area of speech recognition.

D. References

Machine Learning:

1. V.N. Vapnik, *Statistical Learning Theory*, John Wiley, New York, NY, USA, 1998.
2. C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 293-297, 1995.
3. C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," http://svm.research.bell-labs.com/SVMdoc.html, AT&T Bell Laboratories, November 1999.
4. E. T. Jaynes, "Bayesian Methods: General Background," *Maximum Entropy and Bayesian Methods in Applied Statistics*, J. H. Justice, ed., pp. 1-25, Cambridge University Press, Cambridge, UK, 1986.
5. D. J. C. MacKay, *Bayesian Methods for Adaptive Models*, Ph. D. thesis, California Institute of Technology, Pasadena, California, USA, 1991.
6. D. J. C. MacKay, "Probable networks and plausible predictions --- a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, 6, pp. 469-505, 1995.
7. T. J. Loredo, "From Laplace to Supernova SN 1987A: Bayesian Inference in Astophysics," *Maximum Entropy and Bayesian Methods*, P. Fougere, ed, Kluwer Publishing, 1989.

8. S. F. Gull, “Bayesian Inductive Inference and Maximum Entropy,” *Maximum Entropy and Bayesian Methods in Science and Engineering*, vol. 1, Foundations, G. J. Erickson and C. R. Smith, eds., Kluwer Publishing, 1988.

9. M. Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning*, vol. 1, pp. 211-244, June 2001.

10. M. Tipping, “The Relevance Vector Machine,” in *Advances in Neural Information Processing Systems* 12, S. Solla, T. Leen and K.-R. Muller, eds., pp. 652-658, MIT Press, 2000.

11. C. Bishop and M. Tipping, “Variational Relevance Vector Machines,” *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pp. 46-53, Morgan Kaufmann, 2000.

12. A. Faul and M. Tipping, “Analysis of Sparse Bayesian Learning,” in T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, pp. 383–389.

Related Publications by the Authors:

13. J. Hamaker and J. Picone, "Advances in Speech Recognition Using Sparse Bayesian Methods," submitted to the *IEEE Transactions on Speech and Audio Processing*, January 2003.

14. A. Ganapathiraju, J. Hamaker and J. Picone, "Continuous Speech Recognition Using Support Vector Machines," submitted to the *IEEE Transactions on Speech and Audio Processing* (in review), June 2003.

15. J. Hamaker, *Sparse Bayesian Methods for Continuous Speech Recognition*, Ph.D. Dissertation Proposal, Department of Electrical and Computer Engineering, Mississippi State University, March 2002.

16. A. Ganapathiraju, *Support Vector Machines for Speech Recognition*, Ph.D. Dissertation, Department of Electrical and Computer Engineering, Mississippi State University, January 2002.

17. A. Ganapathiraju, J. Hamaker, and J. Picone, "Advances in Hybrid SVM/HMM Speech Recognition," presented at the GSPx / International Signal Processing Conference, Dallas, Texas, USA, April 2003.

18. J. Hamaker, J. Picone, and A. Ganapathiraju, "A Sparse Modeling Approach to Speech Recognition Based on Relevance Vector Machines," *Proceedings of the International Conference of Spoken Language Processing*, vol. 2, pp. 1001-1004, Denver, Colorado, USA, September 2002.

19. A. Ganapathiraju, J. Hamaker and J. Picone, "A Hybrid ASR System Using Support Vector Machines," *Proceedings of the International Conference of Spoken Language Processing*, vol. 4, pp. 504-507, Beijing, China, October 2000.

20. A. Ganapathiraju and J. Picone, "Support Vector Machines For Automatic Data Cleanup," *Proceedings of the International Conference of Spoken Language Processing*, vol. 4, pp. 210-213, Beijing, China, October 2000.

21. A. Ganapathiraju, J. Hamaker and J. Picone, "Hybrid HMM/SVM Architectures for Speech Recognition," *Speech Transcription Workshop*, College Park, Maryland, USA, May 2000.

22. A. Ganapathiraju, J. Hamaker and J. Picone, "Support Vector Machines for Speech Recognition," *Proceedings of the International Conference on Spoken Language Processing*, pp. 2923-2926, Sydney, Australia, November 1998.