Feature Selection for High-Dimensional Remote Sensing Data by Maximum Entropy Principle based Optimization

Shixin Yu   Paul Scheunders
Department of Physics, University of Antwerp
Antwerp, B-2020, Belgium

Abstract

For high-dimensional remote sensing data, the appropriate selection of features has a significant effect on the cost and accuracy of an automated classifier. In this paper, a method for feature selection by Estimation of Maximum Entropy Principle Algorithm, is presented. This method based on the EDA (Estimation of Distribution Algorithm) paradigm, avoids the use of crossover and mutation operators to evolve the populations, in contrast to Genetic Algorithms. It is combined with an approximate application of the Maximum Entropy Principle as the models for representing the probability distribution of a set of candidate solution in the feature selection problem, using the application of automatic learning methods to induce the right distribution model in each generation. Computational comparison is made between EDA in combination with Bayesian networks and EDA in combination with Maximum Entropy Principle. Experiments are performed on AVIRIS dataset.

1 Introduction

Advances in sensor technology for earth observation make it possible to collect large numbers of spectral bands. For example, the NASA/JPL Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) generates image data in more than 220 spectral bands simultaneously. For such high dimensionality, pattern recognition techniques suffer from the well-known curse-of-dimensionality phenomenon. This problem is resulting from the fact that the required number of labeled samples for supervised classification increases dramatically as a function of dimensionality [1].

In this paper, an approximate application of the Maximum Entropy Principle [3] is applied to feature selection problem for high dimentional remote sensing data. This method based on the EDA (Estimation of Distribution Algorithm) [4] paradigm, avoids the use of crossover and mutation operators to evolve the populations, in contrast to Genetic Algorithms. It is combined with an approximate application of the Maximum Entropy Principle as the models for representing the probability distribution of a set of candidate solution in the feature selection problem, using the application of automatic learning methods to induce the right distribution model in each generation.

This paper is organized as follows: Section 2 introduces the EDA paradigm, Bayesian networks and the EBNA search algorithm. Section 3 presents a new algorithm base on an approximate application of the maximum entropy principle. Section 4 summaries the primary results of comparasion.

2 EDA Paradigm and EBNA Algorithm

Genetic Algorithms(GAs) [2] are one of the well known optimization tools, based on population search with two operators: crossover and mutation. In contrast to GAs, the Estimation of Distribution Algorithm (EDA) [4] has neither crossover operator nor mutation operator. The scheme of EDA is as follows:

$D_0 \leftarrow$ Generate N individuals from the initial population randomly. Repeat for l=1,2,..., until a stop criterion is met. $D^s_{l-1} \leftarrow$ Select $S \leq N$ individuals from D_{l-1} according to a method. $p_l(X) = p(X\|D^s_{l-1}) \leftarrow$ Estimate the joint probability distribution of an individual being among the selected individuals. $D_l \leftarrow$ Sample N individuals from $p_l(X)$.

There are many methods proposed in the literature to estimate the distribution probability: Population Based Incremental Learning (PBIL) [5], Compact Genetic Algorithm (cGA) [6], Univariate Marginal Distribution Algorithm (UMDA) [7] and Bit-Based Simulated Crossover (BSC) [8] are four algorithms which assume independence among the features. Population-based MIMIC algorithm using simple chain distribution [9], Dependece trees [10], and Bivariate Marginal Distribution Algorithm (BMDA) [11] are typical algorithms which cover pairwise interactions among the features. Most recently, Inza [12] proposed an algorithm called EBNA, which is based on EDA paradigam, using bayesian networks as the models for representing the probability distribution of a set of candidate solutions. We will compare our proposed algorithm with EDNA in the paper.

3 Maximum Entropy Principle based Optimization

Jaynes [16] pioneered *maximum entropy* (ME) principle to do inference, cheeseman [17] proposed a clever technique to improve the compution efficiency of ME joint probability distribution, following work of Ku and Kullback [18], but no learning method was proposed. Most recently, Yan and Miller [3] proposed an approximate ME method, which encodes arbitrary low-order constraints but while retaining quite tractable learning. we here follow their discussion. For a feature vector $F = (F_1, F_2, \ldots, F_n)$ with $F_i \in \mathcal{A}_i$, the ME joint probability consistent with pairwise probabilities is written as follows:

$$P[f] = \frac{exp(\sum_{m=1}^{N-1} \sum_{n>m}^{N} \gamma(f_m, f_n))}{\sum_{f' \in \mathcal{G}} exp(\sum_{m=1}^{N-1} \sum_{n>m}^{N} \gamma(f'_m, f'_n))}$$

where $\mathcal{G}$ is the full feature space, $\gamma(f_m, f_n)$ is the Lagrange multiplier. If one uses the Lagrangian method to do *direct* optimization, the Lagrangian cost has the form $D - TH$, where H is the joint entropy, D is a cost function encoding the equality constraints on pairwise probabilities, and T is a Lagrange multiplier. Then next minimize $D - TH$ over $\gamma(f_m, f_n)$, but as H and D are *explicitly* expressed as functions of the joint probability with the form of the above equation, so the optimization process is intractable. Yan and Miller [3] proposed to do a restriction on joint probability support in order to achieve tractable learning, i.e. $P[F]$ is restricted to $f \in \mathcal{G}_s \subset \mathcal{G}$. By a Lagangian reformulation, the quantities H and D have the following forms respectively [3]:

	pop. size	learning alg.	error rate	bands
EBNA	1000	IB3	6.79	12
ME	1000	IB3	4.50	15

Table 1: Primary results and some papameters

$$\begin{aligned} H = \quad & - \sum_{m=1}^{|\mathcal{G}_f|} P[f_m] log P[f_m] - \\ & \frac{1}{N} \sum_{k=1}^{N} \sum_{m=1}^{|\mathcal{G}_f|} P[f_m] \sum_{i=1}^{|\mathcal{A}_k|} P[F_k = i | f_m^{-k}] \\ & \times log P[F_k = i | f_m^{(} - k)] \end{aligned}$$

$$\begin{aligned} D = \quad & \sum_{k=1}^{N} \sum_{l=1, l \neq k}^{N} (\sum_{i=1}^{|\mathcal{A}_k|} \sum_{j=1}^{|\mathcal{A}_l|} P[F_k = i, F_l = j] \\ & \times log \frac{P[F_k = i, F_l = j]}{P_M^{(k)}[F_k = i, F_l = j]}) \end{aligned}$$

Given D and H, the Lagrangian cost $D - TH$ can be formed, and the solution maximizes H under the constraints $D = 0$.

Once the joint probability is obtained, the ME model can be now used to do feature selection in combination with EDA schema.

4 Experiments and Discussion

Experiments were conducted on an AVIRIS dataset, containing 220 bands of 145×145 pixels, that is downloadable from [13], along with a groundtruth image, containing 16 classes. The primary experimental result is shown in Table 1 with some papameters. The *MLC* software [19] was used to execute ID3 algorithm, and for random number generator, the GAlib [20]'s implementation was used.

we only did a primary experiment on 30-band data with 2 classes. When the dimensionality increases, the computation time increases dramatically, so how to solve the computional complexity of ME model remains a topic. We are currently looking for methods, e.g.[21], [22], which might improve the computation efficiency of ME model.

Acknowledgments

The authors wish to thank Dr. I. Inza of the department of computer science and artificial intelligence, University of Basque County, Spain, for the implementation of bayesian networks, and for his helpful suggestions.

References

[1] K. Fukanaga, *Introduction to Statistical Pattern Recognition.* (Academic Press, San Diego, California, 1990).

[2] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[3] L. Yan and David J. Miller,*General statistical inference for discrete and mixed spaces by an approximate application of the maximum entropy principle.* IEEE trans. Neural Networks, vol.11, No.3, May 2000.

[4] H. Muhlenbein, G. Paas, *From combination of genes to the estimation of distributions: Binary parameters* in H.M. Voigt, et al.(Eds.) Lecture Notes in Computer Science 1411: Parallel Problem Solving from Nature - PPSN IV, 1996, pp. 178-187.

[5] S. Baluja, *Population-based increment learning: A method for integrating genetic search based function optimization and competitive learning,* Technical Report CMU-CS-94-163, Carnegie Mellon University, Pittsburgh, PA, 1994.

[6] G.R. Harik, F.G. Lobo, D.E. Goldberg,*The compact genetic algorithm* IlliGAl Report 97006, Urbana: University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, 1997.

[7] H. Muhlenbein, *The equation for response to selection and its use for prediction*, Evolutionary Computation 5 (3) (1997) 303-346.

[8] G. Syswerda, *Uniform crossover in genetic algorithms,* in Proceedings of the international conference on genetic algorithm 3, Arlington, VA, 1989, pp. 2-9.

[9] J.S. De Bonet, C.L. Isbell, P. Viola, *MIMIC: Finding optima by estimating probability densities*, in M. Mozer, M. Jordan, T. Petsche(Eds.) Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, 1997.

[10] S. Baluja, S. Davies, *Using optimal dependence-trees for combinatorial optimization:learning the structure of the search space,* in Proceedings of the fourteenth international conference on machine learning, Nashville, TN, 1997, pp. 30-38.

[11] M. Pelikan, D.E. Goldberg, E. Cantu-Paz, *BOA: the bayesian optimization algorithm*,IlliGAL Report 99003, Urbana: University of Illinois at Urbaba-Champaign, Illinois Genetic Algorithms Laboratory, 1999.

[12] I. Inza, P. Larraaga, R. Etxeberria, B. Sierra, *Feature subset selection by bayesian network based optimizaiton*, Artificial Intelligence, 123(1-2), 157-184, 2000.

[13] *http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html*
19/2, 153, 1997.

[14] D. Landgrebe, Information Extraction Principles and Methods for Multispectral and Hyperspectral data, in *Information processing for Remote Sensing*, ed. C.H. Chen (World Scientific, USA, 2000).

[15] C. Lee and D. Landgrebe, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15/4, 388, 1993.

[16] E.T. Jaynes, Papers on probability, statistics and statistical physics. Dordrecht, The Netherlands, 1982.

[17] P. Cheeseman,*A method of computing generalized Bayesian probability values for expert systems*, in Proc. 8th Int. Joint Conf. AI, vol.1, 1983, pp.198-202.

[18] H.H. Ku and S. Kullback, *Approximating discrete probability distributions* , IEEE trans. Information Theory, vol.IT-15, No.4, pp.444-447, 1969.

[19] R.Kohavi, D. Sommerfield, J. Dougherty, *Data mining using MLC++, a machine learning libray in C++*, Int. Journal of Artifical Intelligence Tools 6(4), 1997, pp.537-566.

[20] *http://lancet.mit.edu/ga/*

[21] S.A. Goldman, *Efficient methods for calculating maximum entropy distributions*, MIT/LCS/TR-391, May, 1987.

[22] A.W. Moore and M.S. Lee. *Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets,* Carnegie Mellon University, Robotics Institute, CMU-Ri-TR-97-27, July, 1997.