ADVANCED TOPICS IN SCIENCE AND TECHNOLOGY IN CHINA

Kaizhu Huang
Haiqin Yang
Irwin King
Michael Lyu

Machine Learning

Modeling Data Locally and Globally

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

Springer

Kaizhu Huang
Haiqin Yang
Irwin King
Michael Lyu

Machine Learning

Modeling Data Locally and Globally

With 53 figures

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

Springer

图书在版编目 (CIP) 数据

机器学习:局部和整体的学习= Machine Learning: Modeling Data Locally and Globally: 英文 / 黄开竹，杨海钦, 金国庆, 吕荣聪著. —杭州：浙江大学出版社，2008.4

ISBN 978-7-308-05831-5

Ⅰ.机… Ⅱ.①黄…②杨…③金…④吕… Ⅲ.机器学习-英文 Ⅳ.TP181

中国版本图书馆 CIP 数据核字 (2008) 第 048968 号

机器学习:局部和整体的学习

黄开竹　杨海钦　金国庆　吕荣聪　著

责任编辑	张　琛　伍秀芳
出版发行	浙江大学出版社 网址：http://www.zjupress.com Springer-Verlag GmbH 网址：http://www.springer.com
排　　版	浙江大学出版社电脑排版中心
印　　刷	杭州富春印务有限公司
开　　本	787mm×960mm　1/16
印　　张	11.25
字　　数	404千
版 印 次	2008年4月第1版　2008年4月第1次印刷
书　　号	ISBN 978-7-308-05831-5 (浙江大学出版社) ISBN 978-3-540-79451-6 (Springer-Verlag GmbH)
定　　价	70.00元

浙江大学出版社发行部邮购电话 (0571)88072522

Preface

Machine Learning: Modeling Data Locally and Globally delivers the main contemporary themes and tools in machine learning including probabilistic generative models and Support Vector Machines. These themes are discussed or reformulated from either a local view or a global view. Different from previous books that only investigate machine learning algorithms locally or globally, this book presents a unified and new picture for machine learning both locally and globally. Within the new picture, various seemly different machine learning models and theories are bridged in an elegant and systematic manner. For precise and thorough understanding, this book also presents applications of the new hybrid theory.

This book not only provides researchers with the latest research results lively and timely, but also presents an excellent overview on machine learning. Importantly, the new line of learning both locally and globally goes through the whole book and makes various learning models understandable to a large proportion of audience including researchers in machine learning, practitioners in pattern recognition, and graduate students.

The Chinese Univ. of Hong Kong,
Jan. 2008

Kaizhu Huang
Haiqin Yang
Irwin King
Michael R. Lyu

Contents

1 Introduction

The objective of this book is to establish a framework which combines two different paradigms in machine learning: global learning and local learning. The combined model demonstrates that a hybrid learning of these two different schools of approaches can outperform each isolated approach both theoretically and empirically. Global learning focuses on describing a phenomenon or modeling data in a global way. For example, a distribution over the variables is usually estimated for summarizing the data. Its output can usually reconstruct the data. This school of approaches, including Bayesian Networks [8, 13, 30], Gaussian Mixture Models [3, 21], and Hidden Markov Models [2, 25], has a long and distinguished history, which has been extensively applied in artificial intelligence [26], pattern recognition [9], and computer vision [7]. On the other hand, local learning does not intend to summarize a phenomenon, but builds learning systems by concentrating on some local parts of data. It lacks the flexibility yet surprisingly demonstrates superior performance to global learning according to recent researches [4, 16, 15]. In this book, a bridge has been established between these two different paradigms. Moreover, the resulting principled framework subsumes several important models, which respectively locate themselves into the global learning paradigm and the local learning paradigm.

In this chapter, we address the motivations of the two different learning frameworks. As a summary, we present the objectives of this book and outline the main models or the contributions. Finally, we provide an overview of the rest of this book.

1.1 Learning and Global Modeling

When studying real world phenomena, scientists are always wondering whether some underlying laws or nice mathematical formulae exist for governing these complex phenomena. Moreover, in practice, due to incomplete information, the phenomena are usually nondeterministic. This motivates to base probabilistic or statistical models to perform a global investigation on sampled data from the phenomena. A common way for achieving this goal is to fit a density on the observations of data. With the learned density, people can then incorporate prior knowledge, conduct predictions, and perform inferences and marginalizations. One main category in the framework of global learning is the so-called generative learning. By assuming a specific mathematical model on the observations of data, e.g. a Gaussian distribution, the phenomena can therefore be described or re-generated. Fig. 1.1 illustrates such an example. In this figure, two classes of data are plotted as $*$'s for the first class and $\circ$'s for the other class. The data can thus be modeled as two different mixtures of Gaussian distributions as illustrated in Fig. 1.2. By knowing only the parameters of these distributions, one can then summarize the phenomena. Furthermore, one can clearly employ this information to distinguish one class of data from the other class or simply know how to separate two classes. This is also well-known as Bayes optimal decision problems [12, 6].

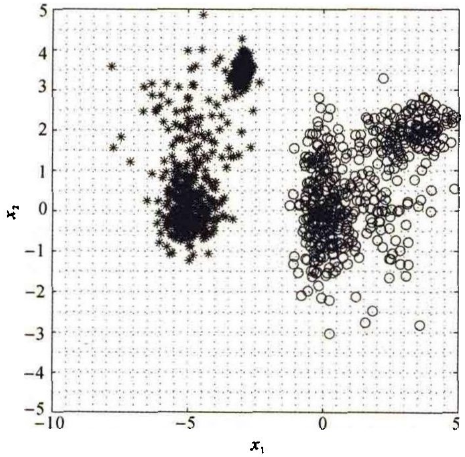

Fig. 1.1. Two classes of two-dimensional data

In the development of learning approaches within the community of machine learning, there has been a migration from the early rule-based methods [11, 32] wanting more involvement of domain experts, to widely-used probabilistic global models mainly driven by data itself [5, 9, 14, 17, 22, 33]. However, one question for most probabilistic global models is what kind of global models, or more specifically, which type of densities should be specified beforehand for summarizing the phenomena. For some tasks, this can be prescribed by a slight introduction of domain knowledge from experts. Unfortunately, due to both the increasing sophistication of the real world learning tasks and active interactions among different subjects of research, it is more

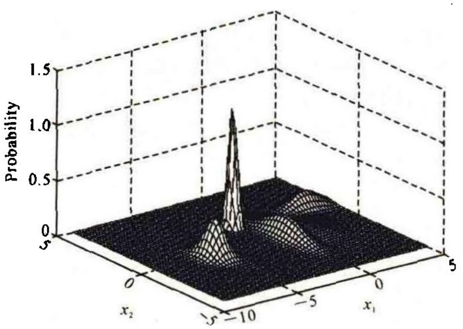

Fig. 1.2. An illustration of distribution-based classifications (also known as the Bayes optimal decision theory). Two Gaussian mixtures are engaged to model the distribution of two classes of data respectively. The distribution can then be used to construct the decision plane

and more difficult to obtain fast and valuable suggestions from experts. A further question is thus proposed, i.e. what is the next step in the community of machine learning, after experiencing a migration from rule-based models to probabilistic global models? Recent progress in machine learning seems to imply local learning as a solution.

1.2 Learning and Local Modeling

Global modeling addresses describing phenomena, no matter whether the summarized information from the observations is applicable to specific tasks or not. Moreover, the hidden principle under global learning is that information can be accurately extracted from data. On the other hand, local learning [10, 27, 28] which recently attracts active attention in the machine learning community, usually regards that a general and accurate global learning is an impossible mission. Therefore, local learning focuses on capturing only local yet useful information from data. Furthermore, recent research progress and empirical study demonstrate that this much different learning paradigm is superior to global learning in many facets.

In further details, instead of globally modeling data, local learning is more task-oriented. It does not aim to estimate a density from data as in global learning, which is usually an intermediate step for many tasks such as pattern recognitions (note that the distribution or density obtained by global learning actually is not directly related to the classification itself); it also does not intend to build an accurate model to fit the observations of data globally. Differently, it only extracts useful information from data and directly optimizes the learning goal. For example, when used in learning classifiers from data, only those observations of data around the separating plane need to be accurate, while inaccurate modeling over other data is certainly acceptable for the classification purpose. Fig. 1.3 illustrates such a problem. In this figure, the decision boundary is constructed only based on those filled points, while other points make no contributions to the classification plane (the decision boundary is given based on the Gabriel Graph method [1, 18, 34]).

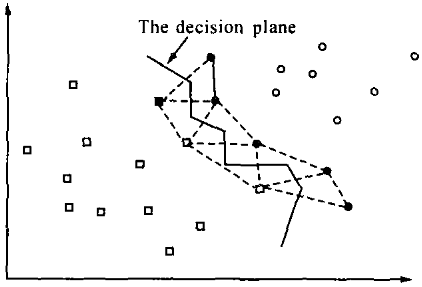

Fig. 1.3. An illustration of local learning (also known as the Gabriel Graph classification). The decision boundary is just determined by some local points indicated as filled points

However, although containing promising performance, local learning appears to locate itself at another extreme end to global learning. Employing only local information may lose the global view of data. Consequently, sometimes, it cannot grasp the data trend, which is critical for guaranteeing better performance for future data. This can be seen in the example as illustrated in Fig. 1.4. In this figure, the decision boundary (also constructed by the Gabriel Graph classification) is still determined by some local points indicated as filled points. Clearly, this boundary does not grasp the data trend.

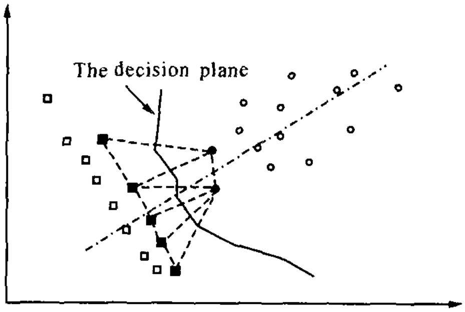

Fig. 1.4. An illustration on that local learning cannot grasp data trend. The decision boundary (constructed by the Gabriel Graph classification) is determined by some local points indicated as filled points. It, however, loses the data trend. The decision plane should be obviously closer to the filled squares rather than locating itself in the middle of filled □'s and o's

More specifically, the class associated with $\circ$'s is obviously more scattered than the class

associated with □'s on the axis indicated as dashed line. Therefore, a more promising decision boundary should lie closer to filled □'s than those filled $\circ$'s instead of lying midway between filled points. A similar example can also be seen in Chapter 2 on a more principled local learning model, i.e. the current state-of-the-art classifier, Support Vector Machines (SVM) [31]. Targeting this problem, we then suggest a hybrid learning in this book.

1.3 Hybrid Learning

There are complementary advantages for both local learning and global learning. Global learning summarizes data and provides practitioners with knowledge on the structure, independence, and trend of data, since with the precise modeling of phenomena, the observations can be accurately regenerated and therefore can be studied or analyzed thoroughly. However, this also presents difficulties in how to choose a valid model to describe all the information (also called the problem of model selection). In comparison, local learning directly employs part of information, critical for the specific oriented tasks, and does not assume models to re-synthesize/restore the whole road-map of data. Although demonstrated to be superior to global learning in many facets of machine learning, it may lose some important global information. The question here is thus, can reliable global information, independent of specific model assumptions, be combined into local learning? This question clearly motivates a hybrid learning of two largely different schools of approaches, which is also the focus of this book.

1.4 Major Contributions

In this book, we aim to describe a hybrid learning scheme to combine two different paradigms, namely global learning and local learning. Within this scheme, we propose a hybrid model, named the Maxi-Min Margin Machine (M^4), demonstrated to contain both the merits of global learning in representing data and the advantages of local learning in handling tasks directly and effectively. Moreover, adopting the viewpoint of local learning, we also introduce a global learning model, called the Minimum Error Minimax Probability Machine (MEMPM), which does not assume specific distributions on data and thus distinguishes itself from traditional global learning approaches. The main models discussed in this book are briefly described as follows.

- *The Maxi-Min Margin Machine model, a hybrid learning framework successfully combining global learning and local learning*

- ◇ *A unified framework of many important models*
  As will be demonstrated, our proposed hybrid model successfully unifies both important models in local learning, e.g. the Support Vector Machines [4], and significant models in global learning, such as the Minimax Probability Machine (MPM) [19] and the Fisher Discriminant Analysis (FDA) [9].
- ◇ *With the generalization Guarantee*
  Various statements from many views such as the sparsity and Marshall and Olkin Theory [20, 23] will be presented for providing the generalization bound for the combined approach.
- ◇ *A sequential Conic Programming solving method*
  Besides the theoretic advantages of the proposed hybrid learning, we also tailor a sequential Conic Programming method [24, 29] to solve the corresponding optimization problem. The computational cost is shown to be polynomial and thus the proposed M^4 model can be solved practically.

- • *The Minimum Error Minimax Probability Machine, a general global learning model*
  - ◇ *A worst-case distribution-free Bayes optimal classifier*
    Different from traditional Bayes optimal classifiers, MEMPM does not assume distributions for the data. Starting with the Marshall and Olkin theory, this model attempts to model data under the minimax schemes. It does not intend to extract exact information but the worst-case information from data and thus presents an important progress in global learning.
  - ◇ *Derive an explicit error bound for future data*
    Inheriting the advantages of global learning, the proposed general global learning method contains an explicit worst-case error bound for future data under a mild condition. Moreover, the experimental results suggest that this bound is reliable and accurate.
  - ◇ *Propose a sequential Fractional Programming optimization*
    We have proposed a Fractional Programming optimization method for the MEMPM model. In each iteration, the optimization is shown to be a pseudo-concave problem, which thus guarantees that each local solution will be the global solution in this step.
- • *The Biased Minimax Probability Machine (BMPM), a global learning method for biased or imbalanced learning*
  - ◇ *Present a rigorous and systematic treatment for biased learning tasks*
    Although being a special case of our proposed general global learning model, MEMPM, this model provides a quantitative and rigorous approach for biased learning tasks, where one class of data is always more important than the other class. Importantly, with explicitly controlling the accuracy of one class, this branch model can precisely impose biases on the important class.

  - ◇ *Containing explicit generalization bounds for both classes of data*
    Inheriting the good feature of the MEMPM model, this model also contains explicit generalization bounds for both classes of data. This therefore guarantees a good prediction accuracy for future data.
- *The Local Support Vector Regression (LSVR), a novel regression model*
  - ◇ *Provide a systematic and automatic treatment in adapting margins*
    Motivated from M^4, LSVR focuses on considering the margin setting locally. When compared to the regression model of SVM, i.e. the Support Vector Regression (SVR), this novel regression model is shown to be more robust with respect to the noise of data in that it contains the volatile margin setting.
  - ◇ *Incorporate special cases very much similar to the standard SVR*
    When considering a consistent trend for all data points, the LSVR can derive special cases very much similar to the standard SVR. We further demonstrate that in a meaningful assumption, the standard SVR is actually the special case of our LSVR model.
- *Support Vector Regression with Local Margin Variations*
  Motivated from the local view of data, another variation of SVR is proposed. It aims to adapt the margin in a more explicit way. This model is similar to LSVR in the sense that they both adapt margin locally.

We describe the relationship among our developed models in Fig. 1.5.

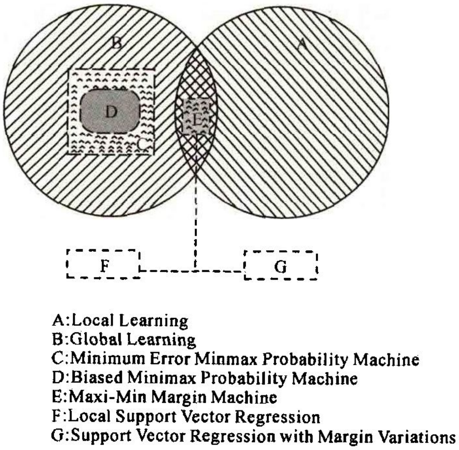
B
A
D
C
E
F
G
A:Local Learning
B:Global Learning
C:Minimum Error Minmax Probability Machine
D:Biased Minimax Probability Machine
E:Maxi-Min Margin Machine
F:Local Support Vector Regression
G:Support Vector Regression with Margin Variations

Fig. 1.5. The relationship among the developed models in this book

1.5 Scope

This book states and refers to the learning first as statistical learning, which appears to be the current main trend of learning approaches. We then further restrict the learning in the framework of classification, one of the main problems in machine learning. The corresponding discussions on different models including the conducted analysis of the computational and statistical aspects of machine learning are all subject to the classification tasks. Nevertheless, we will also extend the content of this book to regression problems, although it is not the focus of this book.

1.6 Book Organization

The rest of this book is organized as follows:

- Chapter 2
  We will review different learning paradigms in this chapter. We will establish a hierarchy graph attempting to categorize various models in the framework of local learning and global learning. We will then base this graph to describe and discuss these models. Finally, we motivate the Minimum Error Minimax Probability Machine and the Maxi-Min Margin Machine.
- Chapter 3
  We will develop a novel global learning model, called the Mininum Error Minimax Probability Machine. We will demonstrate how this new model represents the worst-case Bayes optimal classifier. We will detail its model definition, provide interpretations, establish a robust version, extend to nonlinear classifications, and present a series of experiments to demonstrate the advantages of this model.
- Chapter 4
  We will present the Maxi-Min Margin Machine, which successfully combines two different but complementary learning paradigms, i.e. local learning and global learning. We will show how this model incorporates the Support Vector Machine, the Minimax Probability Machine, and the Fisher Discriminant Analysis as special cases. We will also demonstrate the advantages of Maxi-Min Margin Machine by providing theoretical, geometrical, and empirical investigations.
- Chapter 5
  An extension of the proposed MEMPM model will be discussed in this chapter. More specifically, the Biased Minimum Minimax Probability Machine will be discussed and applied into the imbalanced learning tasks. We will review different criteria for evaluating imbalanced learning approaches. We will then base these criteria to tailor BMPM into this type of learning. Both illustrations on toy datasets and evaluations on real world imbalanced and medical datasets will be provided in this chapter.

- Chapter 6
  A novel regression model called the Local Support Vector Regression, which can be regarded as an extension from the Maxi-Min Margin Machine, will be introduced in detail in this chapter. We will show that our model can vary the tube (margin) systematically and automatically according to the local data trend. We will show that this novel regression model is more robust with respect to the noise of data. Empirical evaluations on both synthetic data and real financial time series data will be presented to demonstrate the merits of our model with respect to the standard Support Vector Regression.
- Chapter 7
  In this Chapter, we show how to adapt the margin settings locally for the Support Vector Regression differently from the LSVR. We demonstrate how the local view of data can be widely used in various models or even differently applied in the same model. Empirical evaluations are also presented in comparison with other competitive models on financial data.
- Chapter 8
  We will then summarize this book and conduct discussions on future work.

We try to make each of these chapters self-contained. Therefore, in several chapters, some critical contents, e.g. model definitions or illustrative figures, having appeared in previous chapters, may be briefly reiterated.

References

1. Barber CB, Dobkin DP, Huhanpaa H (1996) The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software 22(4):469–483
2. Baum LE, Egon JA (1967) An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull. Amer. Meteorol. Soc. 73:360C-363
3. Bozdogan H (2004) Statistical Data Mining and Knowledge Discovery. Boca Raton, Fla.: Chapman & Hall/CRC
4. Christopher J, Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2):121–167
5. Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. IEEE Trans. on Information Theory 14:462–467
6. Duda R, Hart P(1973) Pattern Classification and Scene Analysis. New York, NY: John Wiley & Sons
7. Forsyth DA, Ponce J (2003) Computer Vision: A Modern Approach. Upper Saddle River, N.J. : Prentice Hall
8. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Machine Learning 29:131–161
9. Fukunaga K (1990) Introduction to Statistical Pattern Recognition. San Diego, Academic Press, 2nd edition