LEARNING AND OPTIMIZATION OF THE KERNEL FUNCTIONS FROM INSUFFICIENTLY LABELED DATA

M. EHSAN ABBASNEJAD

UNIVERSITI SAINS MALAYSIA
2010

LEARNING AND OPTIMIZATION OF THE KERNEL FUNCTIONS FROM INSUFFICIENTLY LABELED DATA

by

M. EHSAN ABBASNEJAD

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

May 2010

ACKNOWLEDGEMENTS

In the name of God, the Most Compassionate, the Most Merciful

All praise belongs to God for the countless blessings He has bestowed upon me.

Gratitude beyond words to my parents whose love, support and assistance has been a great help in every single step in my life. I will be forever grateful for their faith in me.

Then, I would also like to express my deepest gratitude towards my thesis advisor, Dr. Dhanesh Ramachandram, for his continuous wisdom, encouragement, friendship, guidance and support over this course of my study. I am forever grateful for the chance to work in his research group.

I would also like to thank my co-supervisor Associate Professor Dr. Mandava Rajeswari for all the useful ideas she has given me from the discussions that we had. Her assistance and invaluable advice throughout my research are highly appreciated.

Not forgetting my friends and lab-mates at the Computer Vision Research Group at Universiti Sains Malaysia. They have been a tremendous source of support and a lot of fun to work with. I will forever remember our moments together. Thank you very-very much Mozaher, Alfian, Osama, Ong, Idayu, Adel, and Anusha.

Special thanks to School of Computer Sciences USM, for providing me with an amazing environment during the course of my research. I would like to thank the USM members for giving me the chance to gain my degree, and for providing me and other students excellent facilities for doing our researches.

Thank you

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbb{R}$	Set of real numbers
A	Matrix
$A^\top$	Transpose of a matrix
I	Identity matrix
$\mathbf{1}$	All 1s vector
n	number of instances, real value
x	Input, a vector
$\mathcal{X}$	Input space, a set of points
y	output/label value
$\mathcal{Y}$	Label vector of a dataset
$\mathcal{H}$	Hilbert (feature) space
$\alpha, \beta, \gamma, \ldots$	Real-valued numbers (vectors)
$\langle x, x' \rangle$	Inner product of x and x'
ϕ	Mapping to feature space
$k(x, x')$	Kernel funtion of x and x'
K	$n \times n$ kernel matrix
$\kappa, \mathcal{K}$	Optimal kernel function/matrix
d	Dimension of input space
$\text{sgn}(x)$	Equals 1, if $x \geq 0$ else -1
$f(x)$	Real-valued funciton
$\|.\|_p$	p-norm, default is 2
$\|A\|$	2-norm of a matrix
$\text{tr}(A)$	Matrix trace
$\exp$	Exponentional function
$\log$	Logarithm function
$\det(A)$	Determinant of a matrix A
$A \succeq 0$	Positive semidefinite matrix A

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ERM	Empirical Risk Minimization
FDA	Fisher Discriminant Analysis
KNN	K-Nearest Neighbour
LP	Linear Programming
MMD	Maximum Mean Discrepancy
PCA	Principle Component Analysis
QCQP	Quadratically Constraint Quadratic Programing
QP	Quadratic Programing
RKHS	Reproducing Kernel Hilbert Space
SDP	Semidefinite Programing
SVM	Support Vector Machine
TLK	Transferred Learning of the Kernel

LIST OF PUBLICATIONS

Abbasnejad M. E. and Ramachandram D. and Mandava R. (2010). An unsupervised approach to learn the kernel functions: From global influence to local similarity, *Journal of Neural Computing and Applications*

Abbasnejad M. E. and Ramachandram D. and Mandava R. (2009). Optimizing kernel functions using transfer learning from unlabeled data, *International Conference on Machine Vision*

Abbasnejad M. E. and Ramachandram D. and Mandava R. (Submitted). A Survey of the State of the Art in Learning the Kernels, *Machine Learning*

PEMBELAJARAN DAN PENGOPTIMUMAN FUNGSI KERNEL DARIPADA DATA YANG TIDAK CUKUP BERLABEL

ABSTRAK

Dalam semua teknik pembelajaran, kaedah inti menjadi semakin popular kerana kecekapan, kejituan serta kebolehannya untuk mengendalikan data berdimensi tinggi. Masalah asas yang berkaitan dengan teknik pembelajaran ini adalah pilihan fungsi inti. Oleh itu, inti sebagai suatu prosedur yang fungsi inti dipilih bagi set data tertentu adalah penting. Dalam tesis ini, dua pendekatan dicadangkan untuk mempelajari fungsi inti: pembelajaran pindah dan pendekatan tanpa-selia. Pendekatan pertama menggunakan pengetahuan yang dipindahkan daripada data tidak berlabel. Data tidak berlabel digunakan seiring dengan data berlabel bagi mengoptimumkan inti dengan menggunakan analisis beza-layan Fisher dan selisih min maksimum. Kejituan pengelasan menunjukkan bahawa bilangan contoh ujian yang diramal dengan betul daripada inti asas. Di samping itu, inti yang dioptimum dibandingkan dalam dua set data yang melibatkan imej satelit dan data sintetik, yang menunjukkan bahawa pendekatan yang dicadangkan mampu memberikan keputusan yang lebih baik. Pendekatan kedua merupakan suatu kaedah tanpa-selia untuk mempelajari gabungan linear daripada fungsi inti. Di sini, struktur intrinsik data tidak berlabel ditaabir melalui suatu langkah yang dikenali sebagai pengaruh, yang dikira dengan membina suatu graf berpemberat. Langkah pengaruh dalam ruang sifat secara kebarangkaliannya berkaitan dengan ruang input yang membolehkan suatu masalah pengoptimuman dapat diselesaikan. Masalah

pengoptmuman dirumus dalam dua konveks yang berbeza, iaitu: pemprograman linear dan pemprograman separa tentu, yang bergantung pada jenis gabungan inti yang dipertimbangkan. Di sini, dua sumbangan dapat dirumuskan. Yang pertama, pendekatan tanpa-selia baru untuk mempelajari fungsi inti . Yang kedua, suatu kaedah untuk mentaabir kesamaan yang diwakili oleh fungsi inti melalui pengiraan pengaruh global di setiap titik pada struktur set data. Pendekatan yang dicadangkan ini menekankan pilihan inti yang bebas atau tidak bergantung pada algoritma pembelajaran berasaskan inti. Penilaian empirik daripada pendekatan yang dicadangkan pada pengelasan imej dan teks menunjukkan keberkesanan algoritma dalam mendapatkan lebih banyak hasil yang jitu.

LEARNING AND OPTIMIZATION OF THE KERNEL FUNCTIONS FROM INSUFFICIENTLY LABELED DATA

ABSTRACT

Amongst all the machine learning techniques, kernel methods are increasingly becoming popular due to their efficiency, accuracy and ability to handle high-dimensional data. The fundamental problem related to these learning techniques is the selection of the kernel function. Therefore, learning the kernel as a procedure in which the kernel function is selected for a particular dataset is highly important. In this thesis, two approaches to learn the kernel function are proposed: transferred learning of the kernel and an unsupervised approach to learn the kernel. The first approach uses transferred knowledge from unlabeled data to cope with situations where training examples are scarce. Unlabeled data is used in conjunction with labeled data to construct an optimized kernel using Fisher discriminant analysis and maximum mean discrepancy. The accuracy of classification which indicates the number of correctly predicted test examples from the base kernels and the optimized kernel are compared in two datasets involving satellite images and synthetic data where proposed approach produces better results. The second approach is an unsupervised method to learn a linear combination of kernel functions. Here, the global intrinsic structure of the unlabeled data is inferred through a measure called influence, which is computed by constructing a weighted graph and performing a random walk upon it. The measure of influence in the feature space is probabilistically related to the input space that yields an optimization

problem to be solved. The optimization problem is formulated in two different convex settings, namely linear and semidefinite programming, depending on the type of kernel combination considered. Here, the contributions are twofold: first, a novel unsupervised approach to learn the kernel function, and second, a method to infer the local similarity represented by the kernel function by measuring the global influence of each point towards the structure of the dataset. The proposed approach focuses on the kernel selection which is independent of the kernel-based learning algorithm. The empirical evaluation of the proposed approach on image and text classification shows the effectiveness of the algorithm in obtaining more accurate results.

CHAPTER 1

INTRODUCTION

Machine learning is a multidisciplinary field that concerns with the design and development of algorithms that allows computers to learn, predict and generalize the previous experiences. Computers are able to observe the events and we are hoping to use the observations to predict future events with the help of machine learning. The observations are used to formulate a hypothesis about a specific incident. Using machine learning techniques, computers have so far depicted a significant level of learning ability. Currently, the most important applications of machine learning are in computer vision (for example recognizing objects in a picture), information retrieval (for example identifying the titles related to one specific news article), natural language processing (for example recognizing part of speech in a document) and many newly emerged applications like finding genes in DNA sequences, mining the social networks and others.

Machine learning algorithms Bishop (2006) (or learning algorithms in short) use observations, or *training examples (samples)* as in the machine learning parlance, as previous experience to learn to improve their behavior in future. Training examples are usually considered in a batch known as a dataset. The objective of a learning algorithm is to build a *hypothesis* or *model* that forms an appropriate insight into the nature of the problem. Any model obtained from data, maps the training examples (*input space*) to

their prediction value (*output or target space*). The process of using the instances of a dataset to learn the prediction model for that dataset is called the *training* phase. In the training phase, learning algorithms seek to find the nature of the underlying problem by investigating the instances of data. In general, having more training examples will increase the chance of correct prediction as the learning algorithm is more familiar with various aspects of the observations. However in many cases, the dataset contains misleading instances that hinder the learning process due to the imperfection in measurement devices. Consequently in any learning algorithm, the possible existence of noise should be considered.

A training example is commonly represented in the form of a list of all the values extracted from the training object called *feature vector*. There are other representations like graphs or strings, but in this research the emphasis is on the commonly known vector representation. This representation of instances paves the way to treat them as points or vectors and perform well-established mathematical and physical analysis.

Machine learning algorithms are usually organized into several categories based on the type of training examples. Apart from the presented categories, other types of learning algorithms like reinforcement learning are also considered which are beyond the scope of this thesis.

- **Supervised learning** is the category of learning algorithms that deals with the *labeled data*. By labeled data, we mean that each entry in a dataset is assigned with an appropriate value indicating its status. Supervised learning is usually formulated as minimization of the errors based on the model's prediction and

the correct label of a given example.

- **Unsupervised learning** tackles the problem of learning in the cases where no labeled data is available. Two common classes of unsupervised algorithms are dimensionality reduction and clustering. In dimensionality reduction, one intends to find a space in which data is represented with a lower dimensionality than its original space. The low-dimensional representation of data preserves some aspects of the higher dimensions, for example, the distances between pairs of points. In clustering, the goal is to put the instances with similar attributes in one category. The most challenging problem faced in clustering is the determination of suitable attributes from the underlying data and how to assess similarity between them. As an example, consider an algorithm that needs to classifying three birds *ostrich*, *crow* and *swan* in two categories; should ostrich and crow be assigned to the same class based on their similarity on color or contrast, or should crow and swan be assigned the same class based on the fact that both birds can fly?

- **Semi-supervised learning** is a new class of learning algorithms that attempts to use the advantages of supervised and unsupervised methods. Since labeled examples may be difficult or expensive to obtain in some cases, the goal of semi-supervised learning is to assist the training phase to formulate stronger hypothesis. In this case, few labeled and a considerable number of unlabeled examples are jointly used to train the learning algorithm. In such algorithms, it is assumed that the unlabeled data naturally surrounds the labeled ones, thus, they strengthen the hypothesis about the label value of a particular example. This property of the dataset is usually called the *cluster assumption*.

1.1 Kernel methods

In recent years, *kernel(-based) methods* or *kernel machines* have been actively studied by the machine learning community Scholkopf and Smola (2001); Shawe-Taylor and Cristianini (2004); Cristianini and Shawe-Taylor (2000); Klaus-Robert Muller and Scholkopf (2001); Smola et al. (2007). There are supervised, unsupervised and semi-supervised variations of kernel-based algorithms. These algorithms are the potential solutions to many problems because of their lower error rate compared to other learning methods, relatively fast training time and elegant compatibility with high dimensional data. The kernel machines work by *nonlinear mapping* of data points (vectors) to a higher, or possibly infinite, dimensional space such that building the hypothesis model for the problem is easier. In order to illustrate how mapping to a higher dimensions influence decision function, consider an example taken from Scholkopf and Smola (2001) in which a 2-dimensional point in the input space is mapped to a 3-dimensional space as shown in Figure 1.1:

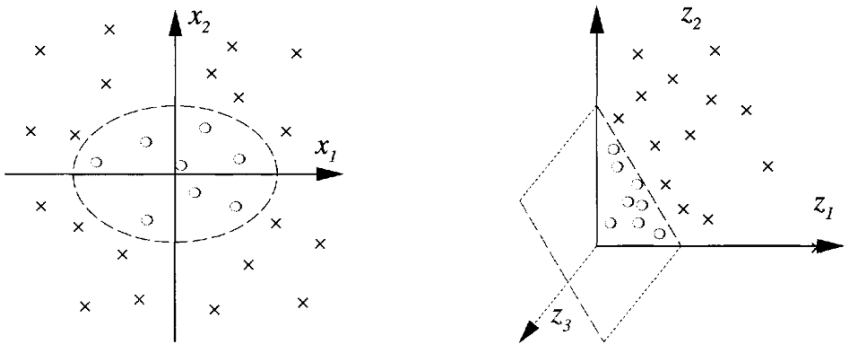

Figure 1.1: Circles from crosses are separated with a hyperplane in a 3-dimensional mapped space (Source: Scholkopf and Smola (2001))

$$([x]_1, [x]_2) \mapsto ([x]_1^2, [x]_2^2, \sqrt{2}[x]_1[x]_2) \tag{1.1}$$

where $[x]_i$ denotes the ith element in the vector. As it is shown in Figure 1.1, while the dataset in the input space is only separable with relatively complex decision function (an ellipsoid like function), a hyperplane which is geometrically simpler in the mapped space perfectly discriminates circles from crosses.

The kernel function represents the inner product of the points in this high dimensional space. Thus, the kernel function amounts to the angle between the vectors of training examples which can be interpreted as the *similarity* between them. Hence, any algorithm that needs a measure of similarity can use a kernel function for that purpose. (See Chapter 2 for a comprehensive introduction to the kernel functions)

The hypothesis model in kernel machines is defined based on a pairwise relation between data points in a higher dimensional space. As it will be shown in Chapter 2, the hypothesis model is solely dependent on a weighted sum of a kernel function in a dataset which makes its selection crucial. Furthermore, it is particularly interesting to know that in the ideal case mapped data in a higher dimensional space is linearly separable. Hence, the well-established linear analysis can be performed in a higher dimensional space which consequently amounts to a nonlinear prediction of the target values (like the example in Figure 1.1). The nonlinear analysis in kernel methods is of a significant value because many of the real-world problems are not linear in nature.

The mapping of the input points to a higher dimensional space will form a new geometrical representation of the dataset (for example the mapped dataset in Figure 1.1) which plays an important role in defining the correct model. The selection of the kernel function has a direct relation to this geometrical representation. Therefore, in

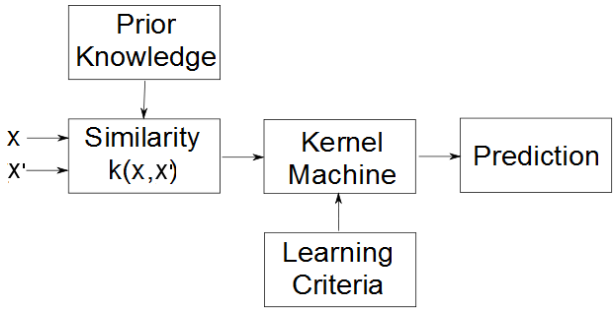

Figure 1.2: The kernel machines perform the prediction in a general framework as illustrated in this figure.

each learning problem a specific mapping function leads to a desirable solution to the learning problem.

In Figure 1.2 the process in which a kernel machine operates is illustrated. The prior knowledge, that is the user's problem-specific knowledge that should be incorporated for informed decision making, when available is reflected in kernel function k while the criteria for learning, that is the principle behind the optimality of an algorithm, is directly influences the kernel machine. It is an important attribute of kernel methods as the prior knowledge of the user can be incorporated independent from the learning algorithm and its optimality conditions. Furthermore, in the kernel machines the decision making is performed solely based on the information obtained from the operation of the kernel function on a pair of examples from the dataset (x and x' in this figure). Examples of kernel-based methods include support vector machine (SVM) (Vapnik, 1999), a well-known supervised algorithm, and kernel principle component analysis (kernel-PCA) (Scholkopf et al., 1998), a popular dimensionality reduction technique. The detailed introduction to these algorithms will be discussed in subsequent chapters.

1.1.1 Model selection

Model selection refers to the stage in which the correct class of hypothesis is determined for any of the machine learning algorithms. Subsequent to the appropriate model selection phase, the learning algorithm finds correct patterns in data and performs an accurate prediction. Model selection is also referred to as one example of a famous philosophical principle known as Occam's razor which states simpler models should be selected over the complex ones and the tradeoff between them should be sought to deter from over-fitting or under-fitting Cherkassky and Mulier (2007). In kernel methods, there are usually two aspects of model selection: firstly, selecting the kernel function and its parameters (typically referred to as *hyper-parameters*) and secondly, the parameters of the learning algorithm itself, if there are any. The former, is specifically very crucial to the performance of the learning algorithm and if not selected appropriately, the tuning in the later part is not effective. In this research, the kernel selection and optimization, as the most important aspect of model selection in kernel machines, is considered.

1.1.2 Optimization

Machine learning techniques normally lead to a maximization or minimization objective. Even the well-known frameworks such as empirical risk minimization (ERM) Vapnik (1999) in statistical learning theory and maximum likelihood (ML) estimation Jain et al. (2000) define the learning criteria as an optimization problem. Hence, the success of machine learning algorithms greatly depends on the performance of the optimization problem. Consequently, a special attention from machine learning commu-

nity has been given to the optimization problems because if a learning algorithm results in an unsolvable optimization problem, the whole learning algorithm fails. Conversely, the triumph of a learning algorithm is certain if its objective leads to an optimization problem with efficient solution. Furthermore, if the optimization does not guarantee the global optimality of the solution, the learning algorithm may not be able to utilize its maximum capability. Various local optimum solutions for a particular learning problem may lead to different models that do not produce stable accuracy for a dataset. Thus, it is an advantage to design a learning algorithm that is capable of producing an optimal solution which is consistent in every run of the problem with same parameters. Additionally, there has been optimization packages recently released that are capable of performing mathematical optimizations with relative efficiency (like `CVX` (Grant and Boyd, 2009)). Specifically in the cases considered in this research, the proposed approaches are designed with the intention of having a unique solution that can be efficiently solved. In the proposed approaches the fulfillment of the constraints are sought which leads to a well-defined optimization. Due to the efficiency in solving the optimization, the proposed approach can be trusted in finding the optimal solution for the problem. In this research, we will discuss several mathematical optimization forms that produce an optimal solution.

1.2 Motivation and problem statement

Many of the algorithms in machine learning are dependent on an appropriate selection of a measure of similarity or dissimilarity (distance). In case of kernel methods, this measure is the kernel function. There have been various kernel (functions) proposed which tend to exhibit diverse characteristics of the mapped space and consequently

explore various aspects of data. Moreover, many of the well-known kernels have parameters to select that must be finely tuned like σ in Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{1.2}$$

In Figure 1.3 the kernel function is kept constant as a Gaussian kernel and the value of parameter σ is changed to show the fact that changes in the parameter value will have a dramatic influence on the decision function. The variance parameter in the Gaussian kernel are set to 0.01, 1, 10, 100 in Figure 1.3b to Figure 1.3f. As it is shown, the variation of a parameter in kernel function changes the decision boundary in SVM dramatically. Even a small variation from 0.01 to 1 has also dramatically change the image of the classification model or decision boundary. The parameter is selected such that each one is ten times more than the former value. Large leaps in parameter value are intentionally selected to show manual selection of an appropriate initial value, update rate and finally the optimal value is extremely complicated. Consequently, it is very challenging to design an algorithm that can automatically determine such value. Choosing an appropriate kernel and possibly its hyper-parameters as the most important aspect of model selection in kernel methods is a challenging task which needs a lot of experience as well as several hours of experiments to overcome.

This issue of selecting the appropriate kernel and its hyper-parameters has given rise to new area of research in model selection known as *learning the kernel*. That is, the kernel is learnt from the dataset at hand as an optimal kernel for the given learning problem. In these methods, an automatic solution to selection of the appropriate kernel is sought.

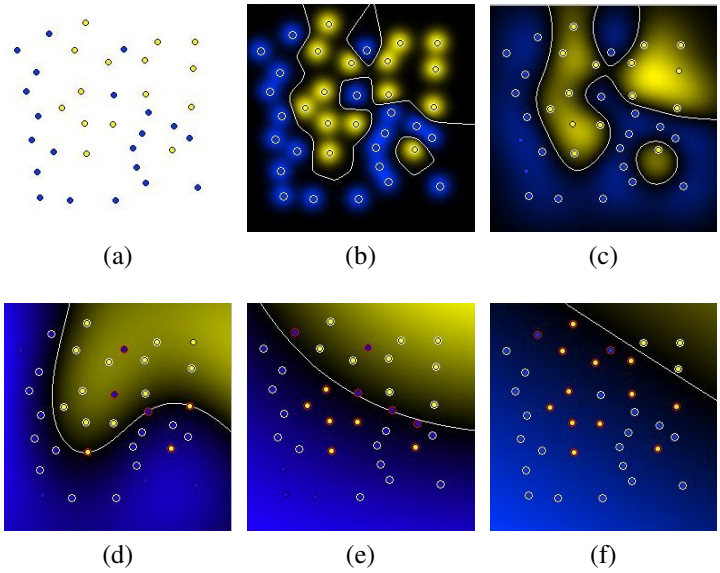

Figure 1.3: The task of classification for a 2-dimensional dataset of blue (dark) and yellow (light) points is shown (Source: Smartlab (2010))

In this research, answers to the following questions are explored:

1. Is it possible to use the data from a similar problem, and learn the kernel from the mixture of this data and a small number of labeled training examples?

2. Is it possible to extend the previous case to completely unsupervised learning of the kernel without using any labeled data?

Supervised methods are widely proposed to overcome the problem of learning the kernel, but they mostly hold the assumption that abundant labeled data are available. In such cases, if the number of training examples is not sufficient, the quality of the optimal kernel obtained is significantly reduced. Needless to say, these supervised methods are not suitable for the cases where no labeled data is available.

Due to difficulty of obtaining labeled training examples, it is useful to have an algorithm that runs with a limited number of training examples. Therefore, it is useful

to have an unsupervised algorithm to learn the kernel functions which can subsequently be used in both supervised and unsupervised learning algorithms.

The answer to the first and second questions surely needs a probe on the nature of the mapping functions which amounts to considering the geometrical representation or structure of the points before and after mapping. Intuitively, it means defining a mapping function for each point should be consistent with the dataset, that is the similarity is defined as a unique measure for a specific set. However, this poses a new question

3. How is it possible to extract the intrinsic structure of a dataset?

Therefore, the answer to the third question will pave the way for better understanding the nature of the underlying problem. The information about the intrinsic structure of the data will help investigating the best optimal for a given dataset. Specifically in the unsupervised case, the structure of the dataset is the only hint that can be used to learn the kernel.

1.3 Objectives

The objectives of this research are:

1. To investigate the algorithms for learning to construct kernels in the case where there are minimal labeled examples.

2. To propose a kernel function constructed from the transferred structure in SVM.

3. To propose an unsupervised approach to learn the kernel function and investigate

its performance in a supervised and unsupervised kernel machines

1.4 Scope

The scope of the presented work is defined as

1. In the proposed algorithms, it is assumed that meaningful values are extracted from corresponding training example and stored in a feature vector. The evaluation of the proposed algorithms is performed on the extracted features.

2. The results of the proposed algorithms are evaluated by the benchmark or synthetic datasets available openly for the researchers. Although all the benchmark datasets are created from real-world problems, it is not intended to look into providing solution for any specific problem. Furthermore, in spite of reporting the general information about the benchmark datasets, the quality in which the dataset has been produced is beyond the scope of this work.

3. Although the proposed approaches are generic in nature, the reported results are only obtained from the evaluation of the respective approach on the noted kernel-based algorithms.

4. The results of running the proposed unsupervised approach is obtained from solving a convex optimization problem. The convexity and some examples of the convex programs will be explained but the detailed description on the methods to solve any specific convex problem will not be discussed.

5. The algorithms proposed in this research emphasize on learning the kernel function as the most important aspect of model selection in kernel methods. Although

other aspects of model selection which usually leads to an algorithm specific tuning are briefly discussed but any comprehensive investigation on those methods are not covered.

1.5 Overview of methodology

The two approaches proposed specifically focus on probing the influence of the structure of the dataset. In cases where sufficient number of labeled examples is not available the structure of the dataset seems to be the only hint on the criteria for measuring the quality of the optimal kernel obtained.

In the first method, limited labeled examples are available which should be utilized. As it is shown in Figure (1.4), the algorithm requires three inputs: $\mathcal{X}$ as the labeled training examples, $\mathcal{Z}$ as the training examples obtained from a similar problem and k as the initial kernel function. A suitable transformation of k is sought that satisfies the objective criteria defined on both $\mathcal{X}$ and $\mathcal{Z}$. There are two sets of criteria considered, firstly on the labeled examples that the similarity and dissimilarity between instances should be preserved (and often magnified). Secondly, the structure of the dataset $\mathcal{X}$ represented by the transformed kernel κ^* should have the highest resemblance with the dataset $\mathcal{Z}$ obtained from a similar problem. This is because the number of labeled training examples is not sufficient to make an informed decision on them consequently an attempt to use the aid of the additional data is made. Ultimately, the algorithm is simplified as an optimization problem that can be easily solved using available optimization techniques.

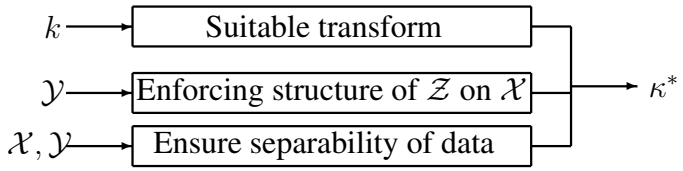

Figure 1.4: The framework in which the optimum kernel κ^* is obtained

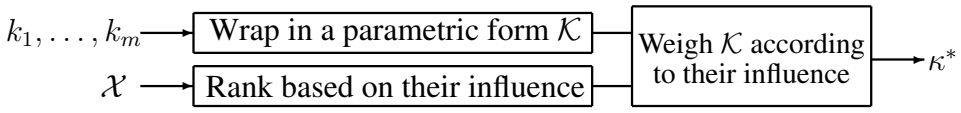

Figure 1.5: Given the dataset $\mathcal{X}$ the optimal kernel function is selected from the parametric representation of the set of kernel functions.

In the second approach as illustrated in Figure (1.5), unlike the first one a parametric kernel based on the set of initial kernel functions is defined. The objective is to find the appropriate weight of the parameters and consequently obtain the optimal kernel such that the similarity of each instance is best measured with respect to the influence values defined on the structure of the dataset. The process used to infer the structural aspects of the dataset is of a great importance which provides the influence of each point in the structure of the dataset. In the second approach the structural aspect is more closely examined as there are no labeled examples available. This approach is proposed in two settings which eventually form two optimization problems.

1.6 Research impacts

The results of this research have a direct impact on the following fields of machine learning:

1. **Transfer learning and multi-task learning** as two very closely related fields of research seek to use the assistance of other data sources to formulate a more

accurate learning algorithm. The optimal kernel learned from a dataset can be transferred to aid other similar learning problems in enhancing the prediction accuracy.

2. **Distance learning** is very closely related to the objective of this thesis. As it will be noted in subsequent chapters, distance metrics can be obtained from a kernel function. Therefore, finding a suitable kernel function provides a distance metric for algorithms that require one.

3. **New family of kernel-based learning algorithms** can be devised if a method to compute the kernel is embedded into the learning problem. In this case, the only thing a user required to do is to select few parameters and the algorithm is able to adapt itself to the data and run with higher accuracy.

1.7 Thesis contributions

This thesis presents two approaches to learn the kernel. Accordingly, two methods to model the structure of the dataset are also considered. Specifically, the contributions of this thesis can be summarized as

1. Two algorithms to learn the kernel functions are proposed in this research. These algorithms learn the kernel functions from the dataset to be subsequently used in the respective learning algorithm.

2. In order to find the optimal kernel, the structure of the dataset has been investigated. In the proposed approaches, two different methods are used to enforce the structural constraints in the algorithm that leads to the optimal kernel.

3. The constraints defined on the learning are modeled in an optimization formulation. In the method considered in Chapter 5, the optimizations with mathematically solid background are considered. In other words, the proposed approach can be considered as an application of the mathematical optimization in machine learning.

1.8 Outline of thesis

The organization of the rest of this thesis is as follows:

In Chapter 2, the notion of kernel functions as one of the novelties in machine learning community and various aspects of their definition is discussed. Furthermore, two of the most popular kernel-based algorithms, namely SVM and kernel-PCA, will be discussed. Additionally, the basics of convex optimization as one branch of the mathematical optimization methods will be presented. The concepts explained in this chapter will lay the foundation for further reference in subsequent chapters.

Recently there has been a growing interest in learning the kernels from the training data. In Chapter 3, a review of the state-of-the-art techniques available in the literature will be presented.

In Chapter 4, the case of insufficient labeled data in training SVM and the influence of optimizing kernels in accurate prediction will be investigated.

In Chapter 5, an unsupervised approach to learn the kernel function will be discussed such that the structure of the dataset is scrutinized. The empirical evaluation of

the proposed approach in a supervised (SVM) and an unsupervised case (Kernel-PCA) is presented.

In the final chapter, the thesis is concluded and the possible directions for the future works are discussed.

CHAPTER 2

THEORETICAL BACKGROUND

The introduction of kernel methods to machine learning owes to the recent advances in functional analysis, statistics and optimization. The strong theoretical models of kernel-based algorithms as well as their success in wide range of applications attract a lot of researchers to further develop these methods. The origins of the theoretical foundations of kernel-based learning algorithms can be traced to the work of Aronszajn (1950) where the reproducing kernels were developed based on the foundation laid on the Mercer's theorem Mercer (1909). In subsequent years, Mercer's theorem has proved to be significant in laying the foundation of the kernel-based algorithms. The impact of such theorems had not been fully perceived until later years where the nonlinear methods in machine learning started to emerge. The paradigm shift in machine learning towards nonlinear methods led to the introduction of artificial neural network and decision trees that revolutionized pattern recognition in 1980s. The artificial neural network (ANN) is an example of a non-linear learning algorithm that has found tremendous success in many application domains (Basheer and Hajmeer, 2000). Although artificial neural networks has been successful in many applications, their drawbacks like slow convergence of the optimization and possibility of being trapped in the local optima encouraged researchers to develop new class of nonlinear learning algorithms Vapnik (1999). One of the most important algorithms that proposed in mid-1990s is support vector machine (SVM) that, in contrary to neural networks,

has a strong mathematical foundation and well-formed optimization that can be solved efficiently (Shawe-Taylor and Cristianini, 2004). SVM is an example of kernel-based learning algorithms that put the mathematical foundation laid in previous years into practice and showed their real impact in computer science.

In this research, it is intended to find a kernel function for a specific dataset. This is typically in contrary to the basics of the way kernel methods are designed as the existence of an appropriate kernel function is taken for granted. However, the kernel functions have to satisfy several conditions which represent the principles behind their definition and their conceptual interpretation. Hence, the kernel function constructed from the data is only worthy if it complies with the requirements of the learning algorithms and aligns with the principles of kernel functions. In Section 2.1, the basic characteristics of the kernels, their impact, several of the related theories that justify their effect and immediate conclusions from their principles will be discussed.

Other than the basic concept that kernel functions represent which makes them naturally desirable, their impact in the area of machine learning could not be made possible without the progresses made in optimization. Specifically, optimization problems that have a strong mathematical background are considered as they provide the means for a unique solution to a given problem. Therefore, the convex optimization, as the most important class of mathematical optimization methods, will be discussed in Section 2.2. Consequently, the use of these optimization methods and kernel functions lead to the development of the kernel-based learning algorithms. In Seciton 2.3, two of the most significant of these learning algorithms are discussed, SVM and kernel-PCA, that will be subsequently used in the empirical evolution of the proposed approaches

in Chapter 4 and 5.

2.1 Kernel functions

The increasing complexity of real-world problems has rendered growing demand for non-linear learning solutions. Kernel methods are newly developed alternative of artificial neural networks that project data to a higher dimensional space which exhibits non-linear characteristics for the learning algorithm. In this section, fundamental theories related to the concept of kernels are presented. More comprehensive introduction can be found in Scholkopf and Smola (2001); Cristianini and Shawe-Taylor (2000); Klaus-Robert Muller and Scholkopf (2001); Shawe-Taylor and Cristianini (2004); Smola et al. (2007). Kernels have several important properties that make them favorable in machine learning:

1. Accessibility to high-dimensional feature space at computationally efficient cost, both from time and space perspective.

2. Most of the kernel-based learning algorithms could be presented as convex optimization that does not suffer from local optima and typically solved efficiently.

3. Solid mathematical foundation that justifies the performance of the learning algorithm and enables further developments.

4. Enabling a modular approach to learning. This means, a kernel function is capable of combining with various learning algorithms. Conversely, each learning algorithm is capable of handling any kind of data structure as long as the kernel function is able to resolve its attributes (for example graphs, strings and others).

The kernel function for a given dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ is defined as:

$$k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}, \quad (x, x') \mapsto k(x, x') \;\; \forall x, x' \in \mathcal{X} \tag{2.1}$$

The function k amounts to the inner product of the feature vectors (data points) in a higher dimensional space. The points in the dataset $\mathcal{X}$ known as the *input space* are mapped to a higher dimensional space $\mathcal{H}$ called the *feature space*. The nonlinear function ϕ called *feature map* performs the projection of $\mathcal{X}$ to $\mathcal{H}$, that is

$$\begin{aligned} \phi : \;\; & \mathcal{X} \mapsto \mathcal{H}, \quad \mathbb{R}^d \mapsto \mathbb{R}^D \\ & x = ([x]_1, [x]_2, \ldots, [x]_d) \mapsto \phi(x) = ([\phi(x)]_1, [\phi(x)]_2, \ldots, [\phi(x)]_D) \end{aligned} \tag{2.2}$$

where d and D are the dimension of data while $[x]_i$ and $[\phi(x)]_i$ are the vector entries referring to the *attributes* in the input and feature space respectively. For example in case of Fig. 1.1, the mapping function is defined as

$$\phi : \mathcal{X} = \mathbb{R}^2 \mapsto \mathcal{H} = \mathbb{R}^3$$

Therefore, given a feature map ϕ the inner product of two points $x, x' \in \mathcal{X}$ in the feature space $\mathcal{H}$ is given by kernel function k as

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \tag{2.3}$$

Subsequently, for a given kernel function k and dataset $\mathcal{X}$, the $n \times n$ matrix

$$K = (k(x_i, x_j))_{ij} \tag{2.4}$$

is called the *kernel (Gram) matrix* of k.

Performing the explicit mapping of each point in the feature space is very costly in terms of time and space and therefore almost impossible to implement in practice. The significance of kernel functions is that the inner product is computed *implicitly*, which means for a given kernel function there is a feature space that the inner product in that space is equivalent to the value obtained from the kernel function. Examples of these kernel functions are

1. $k(x, x') = \langle x, x' \rangle$ (Linear kernel)
2. $k(x, x') = (\langle x, x' \rangle + c)^d$ (Polynomial kernel)
3. $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$ (Gaussian kernel)

Each of these kernel functions represents a unique feature space. For example, one can observe that in case of polynomial kernel for an N dimensional input space, the dimensions of the feature space is obtained from Cristianini and Shawe-Taylor (2000)

$$\begin{pmatrix} d + N - 1 \\ d \end{pmatrix} \tag{2.5}$$

Additionally, in case of Gaussian kernel, by applying the Taylor expansion on the exponential function, a polynomial with an infinite degree is obtained:

$$\exp(x) = \sum_{i=0}^{\infty} \frac{1}{i!} x^i \tag{2.6}$$

Therefore, it can be concluded that the Gaussian kernel in fact corresponds to a feature space with infinite dimensions.

The definition of kernel functions as inner product has an interesting property that allows their further use as similarity measures:

Proposition 1 *The inner product amounts to the angle between projected vectors in the feature space, that is*

$$\sphericalangle(\phi(x), \phi(x')) \triangleq \textit{arccos}\left(\frac{\langle \phi(x), \phi(x') \rangle}{\|\phi(x)\| \|\phi(x')\|}\right) \textit{radians} \tag{2.7}$$

Considering normal vectors in the feature space, a larger angle corresponds to smaller similarity value and vice-versa.

Furthermore, Cauchy-Schwarz inequality may be extended to kernels using the definition in Eqn. (2.3):

$$|k(x_1, x_2)|^2 \leq k(x_1, x_1) k(x_2, x_2) \tag{2.8}$$

Although the Cauchy-Schwarz inequality is true for all kernel functions, however, any function that satisfies this inequality does not necessarily introduce a kernel and consequently does not represent an inner product in a higher dimensional space. The necessary conditions for validity of kernel functions will be subsequently discussed.

The necessary conditions for the validity of the kernel functions are the consequence of the mapping of the data points to a feature space with special characteristics for the inner product of the projected points. This space is called the Hilbert space named after the 20th century mathematician David Hilbert.

Definition 1 (Hilbert Space) *A Hilbert space $\mathcal{H}$ is an inner product space with the additional properties that is separable and complete.*

The Hilbert space is simply an abstract vector space that generalizes the notation of Euclidean space. The importance of Hilbert space in the definition of kernel functions is that it ensures the feature space induced from the mapping function is equipped with measurements of length and angle. Additionally, it guarantees the sequence of products to be limited.

The necessary conditions for a function to represent an appropriate feature space is defined in the well-known Mercer's theory (Scholkopf et al., 1999; Scholkopf and Smola, 2001; Williamson et al., 1999) which has a significant influence in establishing the kernel methods in machine learning.

Theorem 1 (Mercer's Theorem) *Suppose k is a bounded continuous symmetric real-valued function such that the integral operator $T_k : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{X})$ is:*

$$(T_k f)(.) = \int_{\mathcal{X}} k(.\,, x) f(x) dx, \tag{2.9}$$