分类号：_____ 密级：_____

UDC：_____ 编号：_____

工学硕士学位论文

基于 PPG 的无创连续血压预测模型研究

硕士研究生：苗　宇

指导教师：张志强　教　授

学科、专业：计算机科学与技术

论文主审人：谢晓芹　副教授

哈尔滨工程大学

2018 年 1 月

哈尔滨工程大学
学位论文原创性声明

本人郑重声明：本论文的所有工作，是在导师的指导下，由作者本人独立完成的。有关观点、方法、数据和文献的引用已在文中指出，并与参考文献相对应。除文中已注明引用的内容外，本论文不包含任何其他个人或集体已经公开发表的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者（签字）：苗宇

日期：2018 年 3 月 13 日

哈尔滨工程大学
学位论文授权使用声明

本人完全了解学校保护知识产权的有关规定，即研究生在校攻读学位期间论文工作的知识产权属于哈尔滨工程大学。哈尔滨工程大学有权保留并向国家有关部门或机构送交论文的复印件。本人允许哈尔滨工程大学将论文的部分或全部内容编入有关数据库进行检索，可采用影印、缩印或扫描等复制手段保存和汇编本学位论文，可以公布论文的全部内容。同时本人保证毕业后结合学位论文研究课题再撰写的论文一律注明作者第一署名单位为哈尔滨工程大学。涉密学位论文待解密后适用本声明。

本论文（□在授予学位后即可　□在授予学位 12 个月后　□解密后）由哈尔滨工程大学送交有关部门进行保存、汇编等。

作者（签字）：苗宇　　导师（签字）：张志强

日期：2018 年 3 月 13 日　　2018 年 3 月 14 日

分类号：_____ 密级：_____

UDC：_____ 编号：_____

工学硕士学位论文

基于 PPG 的无创连续血压预测模型研究

硕士研究生：苗　宇

指导教师：张志强　教　授

学位级别：工学硕士

学科、专业：计算机科学与技术

所在单位：计算机科学与技术学院

论文提交日期：2018 年 1 月

论文答辩日期：2018 年 3 月

学位授予单位：哈尔滨工程大学

Classified Index:

U.D.C:

A Dissertation for the Degree of M. Eng

Study on noninvasive continuous blood pressure prediction model based on PPG

Candidate: Miao Yu

Supervisor: Prof. Zhang Zhiqiang

Academic Degree Applied for: Master of Engineering

Specialty: Computer Science and Technology

Date of Submission: January, 2018

Date of Oral Examination: March, 2018

University: Harbin Engineering University

摘　要

众所周知，心脏是人体血液循环的中心，承担着使人体各器官正常运转的重任。血压是由心脏产生的人体非常重要的物理信号之一，正常人的血压在多种因素的调节下保持稳定；而对于老人或者病人来说，血压的波动往往会超出正常范围，而这种血压的非正常波动，对于判断一个人的生理状态十分有帮助。因此，如何有效地测量血压在医疗和日常生活中具有重要意义。

在日常生活中，目前最常用的血压测量设备是电子血压计，其原理是基于柯氏音和示波法，在使用电子血压计的过程中，需要对被测者施加压力，这种方法操作繁琐、不能连续监测，而且容易对被测者造成不适，所以有必要研究出更好的方法来进行无创连续的血压监测。得益于传感器技术的发展，人们可以很容易的获取人体脉搏的Photoplethysmogram(PPG)信号，很多研究也都基于 PPG 信号进行了血压的预测。因为 PPG 信号是一组连续波动的波形数据，通过从 PPG 信号的波形中提取出与血压显著相关的特征，然后对这些提取到的特征建立回归模型，从而预测出血压。

目前大部分研究者都是通过线性回归的方式建立 PPG 特征与血压的关系模型，而通过实验可以看出线性方式获得的模型在预测准确性上存在不足，所以本文将通过多种回归方法，对数据进行建模更好的预测血压。在回归模型的基础上，本文通过线性回归，神经网络等机器学习的方法，对 PPG 信号与血压进行线性和非线性回归建模，然后比较了各个方法的预测效果。针对血压的特点，本文提出了两种优化手段来提高血压预测的准确性。首先针对不同人群间存在的差异，通过聚类的方法将原始的数据分类，再对每一个类别建立回归模型，这样可以一定程度减小模型的误差。另外，通过分析数据，可以发现高压与低压间存在一定的相关性:其间的差值一般稳定在一定的范围,所以本文利用这一特点，提出了改进的梯度提升算法，对基本的回归模型进行优化。最终的实验结果表明,本文的方法能够快速有效地改善血压预测的效果,预测结果具有较低的误差。

关键词： 血压监测；Photoplethysmogram；机器学习；聚类；梯度提升

Abstract

As we all know, the heart is the center of the blood circulation of the human body, and it bears the important task of making the organs of the human body run normally. Blood pressure is one of the body's most important physical signals produced by the heart. The blood pressure in normal person remains stable under a variety of factors; for the elderly or patients, the fluctuations of blood pressure tend to exceed normal range, while the abnormal fluctuations in blood pressure are helpful in determining a person's physical state. Therefore, How to measure blood pressure effectively is of great significance in medical treatment and daily life.

In daily life, the most commonly used blood pressure measurement equipment is electronic sphygmomanometer, which the theoretical basis is korotkoff sounds Oscillographic method. In the use of an electronic sphygmomanometer, the instrument needs to exert pressure on the subject. This method is cumbersome and can not be continuously monitored, and it is easy to cause discomfort to those who have been tested, so it is necessary to seek a better method for continuous noninvasive blood pressure monitoring. Thanks to the development of sensor technology, people can easily obtain Photoplethysmogram (PPG) signals of human pulse, and many studies have also made prediction of blood pressure based on PPG signals. since the PPG signal is a set of continuously fluctuating waveform data, A regression model was developed by extracting features that were significantly related to blood pressure from the waveform of the PPG signal, then the blood pressure can be predicted.

At present, most researchers establish the relationship between PPG features and blood pressure by linear regression. Through the experiment, we can see that the model obtained in the linear way has some shortcomings in the prediction accuracy, so this paper will model the data by multiple regression methods. On the basis of the regression model, in this paper, linear and nonlinear regression models of PPG signals and blood pressure are modeled by means of linear regression, neural networks and other machine learning methods, and then the prediction results of each method are compared. In view of the characteristics of blood pressure, this paper presents two optimization methods to improve the accuracy of blood

pressure prediction. First of all, according to the differences among different groups of people, the original data are classified by clustering method, and then the regression model is established for each category, which can reduce the error of the model to a certain extent. In addition, through the analysis of data we can find that there is a certain correlation between systolic pressure and diastolic pressure: the difference between them is generally stable in a certain range, Therefore, this paper uses this feature to propose an improved gradient boosting algorithm to optimize the basic regression model. The final experimental results show that the proposed method can quickly and effectively improve the effect of blood pressure prediction, and the prediction results have lower errors.

Keywords: blood pressure monitoring; Photoplethysmogram; machine learning; clustering; gradient boosting

目　录

第 1 章 绪 论 1
1.1 研究背景与意义 1
1.2 国内外研究现状 2
1.3 论文的研究工作 4
1.4 论文的组织及内容安排 5
第 2 章 相关技术及研究 7
2.1 PPG 信号相关知识 7
2.2 回归模型 9
2.2.1 线性回归模型 9
2.2.2 非线性回归模型 11
2.4 模型优化 15
2.5 现有方法的不足 16
2.6 本章小结 17
第 3 章 模型实现与优化算法 19
3.1 问题描述与模型实现流程 19
3.2 基于聚类的优化算法 22
3.2.1 聚类优化算法思想 22
3.3.2 算法实现 25
3.3 基于梯度提升的优化算法 27
3.3.1 梯度提升优化算法思想 27
3.3.2 限定阈值的梯度提升方法 29
3.3.3 函数映射的梯度提升方法 31
3.4 本章小结 32
第 4 章 实验与结果分析 33
4.1 实验数据 33
4.2 实验环境及评价指标 33

4.3 不同模型预测效果的比较 34
4.4 优化算法的效果 38
4.4.1 基于聚类的优化效果 38
4.4.2 基于梯度提升的优化效果 44
4.5 本章小结 50
结　论 51
参考文献 53
攻读硕士学位期间发表的论文和取得的科研成果 58
致　谢 60

第 1 章 绪 论

1.1 研究背景与意义

在如今 21 世纪，人民的生活水平日渐提高，人们开始更多地注重自身的健康问题，从身高、体重、体脂到血压、血氧、心率，人们希望能够更全面地了解自身健康状况。在几十年前，要想获得这些指标，人们必须要到医院里进行体检，并且需要有专业的设备和专业的医护人员共同协作完成。而归功于科学技术的发展，越来越多轻便的设备已经省略了繁杂的专业操作，变得更加便捷，所以人们可以在家里自己动手完成各种指标的检测。而在众多的健康指标中，血压是人们最关心的指标之一。

心脏是人体血液循环的中心[1]，人们意识到心脏是血液系统的中心要归功于 William Harvey 的论文《心脏的概念》。这一发现打破了之前认为的肝脏是血液系统中心的论调，使人们真正开始了解血液循环。之后的研究者提出了血压的概念，并且逐步确认了血压对于人体健康的重要性，心脏通过有规律的搏动产生血压，进而向全身供血，完成人体的新陈代谢，所以，血压是人体非常重要的生理信号之一。正常范围内的血压才能保证血液正常循环流动，许多因素共同作用下才能使血压保持正常，从而人体的各个器官与组织能获得足够的血量，进而保持人体正常运转。人体血压含有两个重要的数值：收缩压（SP）和舒张压（DP），医学上通过这两个量来判断人体血压的正常与否。收缩压和舒张压与心血管疾病存在强烈的相关性[2]，而且血压过低过高都会造成严重后果，一般情况下的正常血压范围：90mmHg<收缩压<140mmHg，60mmHg<舒张压<90mmHg。有家族遗传高血压病史的人群、肥胖人群、盐分摄入过量的人群和过度酗酒的人群比较容易患有高血压；而长期卧病在床的人群、年轻女性、体质较弱的人群、更年期妇女和老年人比较容易患有低血压。大多数由血压引起的疾病都是可防可治的慢性病，如果能够早期发现、及时治疗、合理用药，则可以有效地减缓器官损伤，提高生存质量。那么基于此目的，我们就需要能够实时地获取自身的血压，及早地发现血压异常，并及时地采取有效的措施以防病情加重，所以如何快速有效地测量人体血压具有重要意义。

在日常生活的场景下，人们测量血压不仅关心测量的准确性，而且也十分在意测量操作过程的复杂程度和设备的舒适度，尤其是当人们期望连续监测的情况下，这两点就显得更为重要。连续的血压测量，能够实时地监测血压的动态变化，而且可以帮助评估

降压药对病人的效果。大多数传统的血压测量方式，往往操作十分繁琐，而且很多都会对被测者身体造成损伤，或引起被测者不适，因此也就不能多次地使用以达到连续监测的目的。归功于硬件技术和软件技术的快速发展，各种传感器正在变得越来越小巧，软件操作也变得越来越简单，所以现在很多的智能可穿戴设备已经逐渐走向大众化。因此如何有效地将这些智能设备运用到血压测量中，是目前很多研究人员十分关心的课题，所以本文将围绕如何准确有效地监测血压展开研究。

1.2 国内外研究现状

血压测量方法，可以分为直接测量和间接测量[3]。从另一个角度，现代血压监测根据是否对被测者造成损伤，又可以分为有创血压监测和无创血压监测，其中无创血压监测又可以进一步分为连续监测和间歇监测。有创血压监测可以迅速且有效地刻画出动脉血压的瞬时变化，对于患有体外循环，血管痉挛和休克等症状的人群，具有更加准确的测量结果，因此能够更有效地反映患者的血压情况。但另一方面，其有创的特性，会对人体造成损害，是一种侵入式监测手段，操作不当容易引起出血，感染等并发症。而对于日常家庭中的使用来说，显然更倾向于无创血压监测。

直接测量法通过在心脏附近或者大动脉上插入接有压力传感器的导管的方式来监测血压产生的信号，此方法在临床上可以进行连续测量。这种方法可以直接测得血压，数据准确，但其技术要求较高，且有一定创伤性，所以不适用于日常生活中的应用场景。间接测量法不通过直接利用仪器测量动脉血压的方式进行，而是通过测量与血压相关的其他人体物理信号，如血管壁的搏动或血管的容积变化情况等，从而间接得到血压值。因为此方法的操作起来比较简便，所以在医疗上具有比较多的应用。多数的间接测量法都是间歇式的测量，这其中以柯氏音法和示波法[4,5]为代表。柯式音法也是目前人们生活中最常用的血压检测方法之一，其基本原理是将袖带固定在被测者的手臂，袖带另一端连接水银柱，测量开始时，关闭出气阀门，然后手动向袖带内打气，直到血管内的血流被阻断，然后再适当松开阀门进行放气。在放气过程中，将听诊器的听筒放在手臂与袖带之间的动脉上方，仔细听脉搏搏动的音。开始时由于袖带产生的压力大于血压，脉搏将被阻断，几乎听不到声音或声音很小；随着袖带内压力的下降，袖带内压力和血压间的差距逐渐变小，在一个时间点上会听到脉搏搏动的声音，之后声音持续增大，直到听到连续且平稳的声音时，之后声音再逐渐减小，最后声音变调、直至消失。当脉搏音出

现时，此时刻对应的水银汞柱的刻度即为收缩压，当脉搏音趋于平稳时，对应的水银汞柱的刻度即为舒张压。虽然柯氏音法是一种无创间接的血压测量方法，但是在袖带加压的时候，会对被测者的血管产生压迫，因此只能间歇式的测量，不适合连续监测，而且这种方法是以测量过程中每搏血压相同为前提，测量出来的血压值只能表示一个瞬时的血压值。

1963 年，Pressman 提出了动脉张力测定法[6,7]。其理论基础是当具有内在压力的血管被施加在其上的外部压力压扁时，血管壁四周的应力发生改变，当所施加的外力达到某一个临界值时，血管内的压力与施加的外力相同，则此时通过其他手段测量施加的外力就可以获得动脉血压。通过施加一个范围适中并且可以调节的外力，以及一个足够小巧并且能够精确定位的压力传感器，外加一些辅助系统，就可以完成对血压的连续测量。这种方式测量得到的血压值具有较高的准确性，而且能满足长时间测量血压的要求。但是这种方式仍然存在缺点，所用的传感器需要对位移具有较高的灵敏度，测量过程中长时间保证传感器与人体的相对位置固定也是比较困难的。而且在连续测量的过程中，被测试者需要长时间受到加压设备的影响，不利于被测者的正常工作。因此，动脉张力测定法不适合长期监测以及运动中监测，同时在操作难度上还需要进一步的降低。

1973 年 Penaz 提出了容积补偿法[8,9]。原理是对动脉施加外力，使动脉血管处于去负荷状态，这时，外部施加的压力等于人体内动脉的压力，同时血管一直保持固定的容积，而不会随着血压的变化而变化。因此，通过额外的调节系统，逐步改变外部施加的压力，过程中一直保持动脉血管容积不变，那么，此时只需要测量出外部施加压力的大小，就能得到动脉的血压值。通过此方法，能够实现每搏血压值的连续监测，而且能有效地描绘血压变化的图像。但是同样因为外加压力的存在，对被测者的连续测量会导致其静脉充血，从而降低测量方法的准确性，而且也会对被测者的正常工作造成影响。

前面的几种方法在测量血压时，都需要多种操作繁琐的设备协作，才能完成测量，而对于连续监测来说，这些方法还是显得不够方便，因此后来的研究开始通过测量脉搏，然后间接地推测出血压。

脉搏是心脏射血时血液对动脉血管产生的压力变化造成的，因此脉搏与血压具有一定的关联。基于此原因以及传感器技术的发展，相关工作者提出了通过脉搏波信号来间接测量血压的方法。目前，多数的脉搏信号采取了利用光电传感器得到的 Photoplethysmogram(PPG)[10]。传感器通过测量血液含氧浓度引起的光强变化，收集相应

的波动信号。利用 PPG 实现血压测量基本可以分为两类。第一类是脉搏波速测定法[11,12]，采取两路 PPG 或 PPG 结合心电仪，测得脉搏传导时间，计算出脉搏波速，然后利用生理上的相关关系推导出血压；另外一类方法则尝试通过提取 PPG 信号中所包含的特征点[13,14,15,16]，通过建立回归模型来预测血压。

第一种方法是通过测得脉搏波速，然后进一步通过脉搏波速来推导血压，这种方法一般需要对被测者事先进行校准才可以得到较好的结果。文献[17]对脉搏波速方法的可行性进行了验证。通过计算心电图的峰值和 PPG 信号上对应点之间或两个安置在不同部位的 PPG 信号之间的时间延迟(脉搏传导时间)[18,19]，来计算出脉搏波速，从而可以推测出血压。文献[20]还研究了 PPG 信号每个周期内的归一化谐波面积和波形下降时间，并与脉搏传导时间一起计算与血压的相关性，结果表明归一化谐波面积同样与血压有较高的相关性。文献[21]在脉搏波速的测定中引入了生物阻抗信号，使测量准确性进行了提高。文献[22]利用机器学习方法对脉搏传导时间，心率以及 PPG 信号中包含的特征量进行了学习，免去了传统方法需要校准的缺陷。

文献[23]分析了脉搏波含有的特征量及其与其他心血管信号的相关性，并提出了一些特征提取方法，可以为后面的工作提供相关的指导。文献[14]计算了脉搏波每个周期的波形上升时间和波形下降时间，并计算了其分别与相应周期的收缩压和舒张压的相关系数，得出了上升时间和收缩压相关性较高，下降时间与舒张压相关性较高的结论。文献[24,25]分别从 PPG 信号中提取出不同的特征，并运用相关的回归分析方法，对血压进行了预测。随着机器学习的发展，对于 PPG 信号特征与血压间的回归分析，机器学习方法通常能取得更好的效果[26]。

心电结合 PPG 的测量技术，需要利用心电仪，影响了连续血压监测的便捷性；而两路 PPG 的测量技术，会由于人体的运动造成测量的不准确；PPG 特征参数的方法虽然可以在对人体影响最小的情况下方便地预测血压，但预测的准确性还有待提高。本文将在 PPG 特征参数血压预测方法的基础上结合聚类及梯度提升算法对预测模型的效果进行提升。

1.3 论文的研究工作

本文的主要研究内容是将机器学习方法以及回归分析手段应用到血压预测当中。

首先，因为 PPG 光电传感器轻便，无创的特性，能够实现连续无创以及操作便捷

的血压监测，本文基于 PPG 脉搏信号展开研究。可以从开源的网络数据库获取原始数据，同时要保证数据中的 PPG 信号与血压值保持一致的对应关系。

PPG 信号是一组连续周期性变化的数字，通过波形分析手段以及相应处理工具，从原始的 PPG 信号中，提取出与血压显著相关的特征点，然后将每组特征点与其对应的血压值看作是机器学习的一个样本。在提取特征点的时候也要注意处理原始信号中的异常值，通过相应的手段对异常值进行处理。

然后，将所有得到的样本组合起来，作为机器学习训练数据或测试数据。利用训练数据训练得到模型，然后在测试数据上对模型的准确性进行评估。在此过程中，可以选取多种机器学习方法共同测试，之后通过实验比较每种方法的优劣性。

其次，本文提出了两种改进算法，对现有的机器学习方法进行调整，使其更适合本文需要处理的脉搏和血压的情境，从而提高预测结果的准确性。比如可以通过聚类将样本中近似同类的数据收集起来，然后分别对每个类别进行训练，这样可以提高算法的预测准确性。

最后的实验部分，为了对比本文提出的优化方法与原始算法的效果，实验中分别使用原始的算法和改进的算法，从多个方面对模型预测准确率进行了检验。实验结果表明，本文提出的改进算法可以有效地提高血压预测的准确性。

1.4 论文的组织及内容安排

本文的主要研究内容是通过机器学习方法建立脉搏与血压间的关系，从而间接推断出血压，主要提出了基于聚类的回归模型和基于梯度提升的回归模型。本篇文章的主要内容如下：

第 1 章 绪论，介绍了血压预测的研究背景和重要实际意义，然后介绍了目前普遍应用的血压预测方法及其存在的不足，并简略的介绍了本文的研究工作。

第 2 章 相关技术及研究，主要介绍了 PPG 信号的特点、适合解决回归问题的机器学习方法，并分析了其各自的优缺点，以及一些模型优化手段。

第 3 章 模型实现与优化算法，首先详细说明了本文所需要解决的问题以及对血压预测模型的建立过程。然后提出了基于聚类的优化算法，由于不同人群之间存在差异，在将他们统一建立回归模型的过程中会产生较大的误差，而通过聚类手段能够消除这种差异导致的模型准确性降低的情况。再之后，提出了基于梯度提升的优化算法，由于血

压通常都是收缩压与舒张压成对出现，而且他们之间往往存在者紧密的相关性，通常一个人的收缩压和舒张压之间的差异都会稳定在一个适当的范围，因此利用这一特点，再结合梯度提升这种集成学习方法，对模型进行优化。

第 4 章 实验与结果分析，首先说明了本文实验所处的硬件环境，然后介绍了实验数据集的来源以及评价模型预测准确性的指标。之后对实验采用的方法进行介绍，对常用的几种算法进行了比较，并且对本文提出的优化方法进行评估，结果证明本文提出的优化方法能够获得更好地预测准确性。

在本文的最后，对我们所做的工作进行概要地总结，说明本文的提出的血压预测模型和优化方法的价值和意义，并提出了现阶段存在的不足以及以后可以改进的重点内容。

第 2 章 相关技术及研究

2.1 PPG 信号相关知识

光电血管容积图（photoplethysmogram,PPG），是通过光电容积描记法（Photoplethysmography）获得的一种波形信号。光电容积描记是一个简单的、低成本的光学技术，可用于检测组织微血管的血流量的变化，而且它是一种作用在皮肤表面的非入侵式测量方法[27]。利用氧合血红蛋白和血红蛋白对光线具有不同的的吸收率的特性，光线在穿过血液时，会产生不同程度的衰减，这种技术可以测量出光线在进出血液前后强度的变化，进而可以换算出血红蛋白在血液中的含量[29]。此方法能够对人体进行连续无创的测量，并且使用方便，反应快速。光电体积描记器是一种具有一个光源（发光二极管）和一个接收器（光电二极管）的十分简单的设备[28]。按照光源和接收器的相对位置，可以将 PPG 传感器实现方式分为穿透方式和反射方式，如图 2.1 所示。穿透方式通常测量时光源和接收器可以放在被检测组织的一侧（比如分别放置在手指的两侧），通过接收透射的光来工作，因为这种方式可以捕捉相对比较高质量的信号，因此得到了广泛的应用。但是，这种方式对被测者而言，如果佩戴时间过长，会造成不适而且影响正常工作。另一种反射方式，将接收器放在被测组织的另一侧，

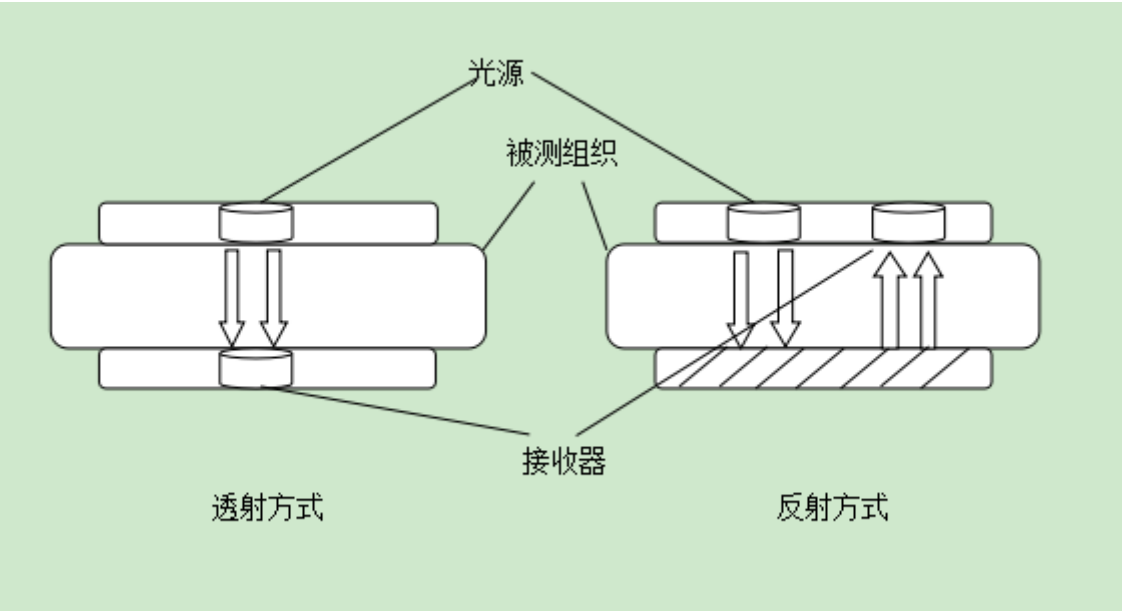


图 2.1 传感器实现方式

通过接受反射光来进行工作。这种方式可以有效地解决以上问题，而且传感器的佩戴位

置可以更加灵活。但这种方式也存在着不足，由于人体组织下方的散射较高，造成了反射信号质量没有透射信号高。所以选择哪种实现方式要根据具体需要来确定。

近年来，由于低成本、简单和便携技术的需求，低成本和小半导体元件的广泛应用，以及基于计算机的脉搏波分析技术的进步，PPG 技术得到了广泛的关注。目前这种技术，已经广泛应用在各种医疗设备中，可以用来测量血压，血氧以及心输出量等，而且还可以评价患者的自主神经功能和检查某些心血管方面的疾病。而且因为机器智能化的普及，PPG 也已经广泛应用于多种智能可穿戴设备。

PPG 信号通常包含直流分量和交流分量，交流分量通常是周期变化的波形，频率一般与心率相等，大约在 1Hz 左右，交流成分被叠加在一个与组织类型和平均血容量有关的较大的直流分量上。将 PPG 传感器至于手腕即可获得 PPG 形式的脉搏波信号，脉搏波可以反映出被测者心血管功能方面的许多信息，而且理论上脉搏的形成与血压是密切相关的，所以可以通过脉搏来预测血压。最常见的脉搏波曲线一般分为 2 种，如图 2.2 所示。通常的一个周期内的脉搏波曲线都由前半部分的上升波和后半部分的下降波组成，前半部分的上升波形反映了心室在射血的时候动脉的扩张过程，后半部分反映射血后期的血管的回缩。随着心室的舒张，心室内的压力低于主动脉的血压，于是血液产生倒流，造成主动脉瓣的关闭，然后形成下降的波形。然后，因为主动脉瓣关闭，使得倒流的血液又能够继续向前流去，在下降过程中又会形成一个较小的上升，可以称为重搏

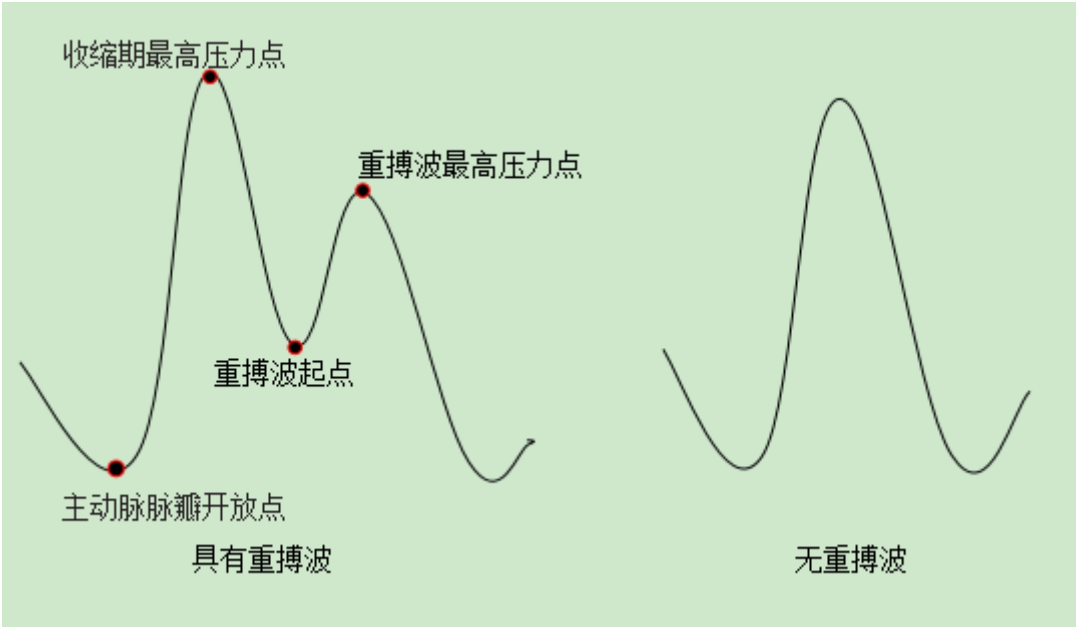


图 2.2 PPG 脉搏波两种形式

波或者降中波。造成重搏波消失的原因，一般是被测者患有心血管方面的疾病或者年龄较大导致血管老化。图中标记的点，通常认为是脉搏中具有代表性的点，波形开始上升的起始点为主动脉脉瓣开放点，上升的最高点为收缩期最高压力点，大多数研究都认为这两个点与血压的收缩压和舒张压具有较高的相关性，从而重点关注这两个点，本文后续的工作也是基于这两个特征点来进行。

2.2 回归模型

前面介绍了PPG脉搏信号的获取方式和信号的特点，后面的工作就是要利用PPG信号建立回归模型，进行回归预测。回归预测就是指，给定已知的自变量$\boldsymbol{X}$，利用$\boldsymbol{X}$来预测位置的因变量Y，而用来预测的模型可以是线性模型也可以是非线性模型，一般都需要根据问题的性质来选择。回归预测通常要遵循特定的流程，最重要的就是要确定问题的已知信息和求解目标是什么，然后才能定义回归中的自变量和因变量。如预测的具体目标是血压，那么血压就是因变量Y。根据实际情况，寻找与预测目标相关性较高并且容易获取的影响因素，令其作为自变量，必要时还需要进行一定的筛选。然后就可以通过具体手段建立回归预测模型，基于统计学或者基于机器学习来建立回归模型。然后的工作就是检测模型的准确率，也可以通过均方误差或均方根误差评价模型的效果。此时的回归模型还需不需要继续优化，要取决于回归模型得到的误差是否满足行业标准。先前得到的预测模型必须在大多数指标的检测下能够具有良好的效果，才可以将最后有效的模型用于实际计算。下面给出本文问题的基本描述。

假设从PPG信号中提取到的特征记为$\boldsymbol{X}=[x_1,x_2,\cdots\cdots,x_k]^{\mathrm{T}}$，将对应的标准血压记为$Y$（可以表示舒张压或者收缩压），目标就是找到一个合适的模型$\boldsymbol{\Phi}$，使得模型产生的误差$\varepsilon=|Y-\boldsymbol{\Phi}(\boldsymbol{X})|$最小。

2.2.1 线性回归模型

线性回归，顾名思义是通过建立线性方程表示自变量$\boldsymbol{X}$和因变量Y的关系。方程的形式通常为

$$f(\boldsymbol{X}_i)=\beta_0+\boldsymbol{\beta}_1\boldsymbol{X}_i \tag{2-1}$$

其与因变量的关系为

$$Y_i=\beta_0+\boldsymbol{\beta}\boldsymbol{X}_i+\varepsilon \tag{2-2}$$

其中 Y_i 表示因变量或者待预测标准值，X_i 表示自变量或者特征向量，β_0 和 $\boldsymbol{\beta}$ 是模型参数，$\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots\cdots,\beta_k]$，分别表示斜率和截距，$\varepsilon$ 表示模型的预测误差[30]。在解决实际问题时，通常我们会先得到用于训练模型的训练集，可以用 D 表示为 $\{[\boldsymbol{X_1},Y_1],[\boldsymbol{X_2},Y_2],\cdots\cdots,[\boldsymbol{X_n},Y_n]\}$，为了方便可以将 $f(\boldsymbol{X_i})$表示为 $\boldsymbol{X_i\beta}$，$\boldsymbol{\beta}=[\beta_0,\beta_1,\cdots\cdots,\beta_k]$，$\boldsymbol{X}=[1,x_1,x_2,\cdots\cdots,x_k]^{\mathrm{T}}$，则多个样本间的关系可以用矩阵的形式表示：

$$F(X)=\begin{bmatrix} 1, & x_1^1 & \cdots\cdots & x_k^1 \\ 1 & x_1^2 & \cdots\cdots & x_k^2 \\ \cdots\cdots & & & \\ 1 & x_1^n & \cdots\cdots & x_k^n \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots\cdots \\ \beta_k \end{bmatrix}=X\beta \tag{2-3}$$

最常用的求解线性回归方程的手段为最小二乘法，即通过是误差平方最小来求解未知参数。则误差可以表示为：

$$J(\beta)=\frac{1}{2N}\sum_{i=0}^{N}\left(F_\beta(X_i)-Y_i\right)^2=\frac{1}{2N}(X\beta-Y)^T(X\beta-Y) \tag{2-4}$$

其中 N 表示样本总数。则求解目标就是通过训练数据找到使误差最小的 $\boldsymbol{\beta}$，如果矩阵满秩则可以直接通过求导的方式计算出最优的 $\boldsymbol{\beta}$，否则需要通过梯度下降法（或牛顿法）来求解最优的 $\boldsymbol{\beta}$。梯度下降的一般过程为，首先初始化一个 $\boldsymbol{\beta}$，然后计算误差方程的对 $\boldsymbol{\beta}$ 的每个分量的偏导数，然后设定一个较小的学习率，每次令 $\boldsymbol{\beta}$ 向导数的负方向按照设定的学习率进行移动，直到 $\boldsymbol{\beta}$ 趋于稳定。初始值和学习率的设定对于梯度下降来说比较重要，如果选取的不好，则可能使结果陷入局部最优点，而不能到达全局最优点。

最基本的采取最小二乘法的线性回归，经常会出现的一个问题是过拟合，而过拟合会导致模型的泛化性变差。简单来说过拟合指的是学习的到的模型只在训练集上表现良好，而在测试集上表现较差。过拟合可能是由于数据噪声太多引起，也可能是因为模型参数选取过多，造成模型过于复杂。所以为了解决由于参数过多引起的过拟合问题，一种可采取的方案是在误差方程里加入正则项。

机器学习有一个原则就是，要能够一边维持较小的误差，同时避免模型的参数过大。使误差尽量小就是为了让模型能够准确地拟合学习的数据，避免模型参数过大即添加正则项，是为了防止模型过拟合，因为我们训练模型的本意不是想让训练误差最小，而是希望它能够在测试数据上表现得更好，也就是能够正确处理新来的数据。那么正则项就是用来限值模型参数的复杂程度的，这也正符合了“奥卡姆剃刀”的原则，“尽量用简单的模型来解释已知的数据”。正则项一般包含了 L0 范数，L1 范数和 L2 范数。其中

L1 范数指向量中非 0 元素的个数，如果用 L0 范数来约束模型参数的话，那得到的结果就是使模型参数向量中部分分量为 0，这就使得模型参数变为稀疏的，也就限制了模型的复杂程度。但是 L0 范数存在一个问题，就是很难求解，所以 L1 范数作为 L0 范数的最优近似，就代替了 L0 范数。L1 范数指向量中所有元素的绝对值之和，同样也可以实现模型参数的稀疏化。模型稀疏的好处就是可以实现自动的特征选择并且增强了模型的可解释性，参数中减小的分量就代表相应的特征所占的权重变小，如果分量为 0，则代表去除了相应的特征，从而实现了特征的选择。最后一种 L2 范数指的是向量中所有元素的平方和然后开方，相比于 L1 范数，L2 范数对于模型参数的限值通常不会使某个分量变为 0，而是会限制在一个比较小的范围内，所以也就不具有特征选择的作用，但同样可以用于解决过拟合的问题。

将 L1 范数和 L2 范数分别应用于最小二乘线性回归，可以得到两种优化的算法，分别称为 Lasso 回归和 Ridge 回归，Lasso 回归的误差方程就变为了：

$$J(\beta)=\frac{1}{2N}\sum_{i=0}^{N}\left(F_{\beta}(X_i)-Y_i\right)^2+\lambda\sum_{j=0}^{k}\left|\beta_j\right| \tag{2-5}$$

其中公式的后一项表示 L1 范数惩罚项，而 Ridge 回归的误差方程变为：

$$J(\beta)=\frac{1}{2N}\sum_{i=0}^{N}\left(F_{\beta}(X_i)-Y_i\right)^2+\lambda\sum_{j=0}^{k}\beta_j^2 \tag{2-6}$$

其中公式的后一项表示 L2 范数惩罚项，超参数 λ 作为正则项的惩罚系数，可以用来调节正则项在误差方程里的权重，相当于调整限制模型参数复杂度的程度。

虽然还有其它一些方法可以用于线性回归预测，但是如果问题的本质决定了它很难通过一个线性方程来表示的话，那无论选择什么形式的线性回归方法，都很难建立起正确的预测模型。所以，在这种情况下，就需要通过建立非线性模型来求解。

2.2.2 非线性回归模型

在实际问题中，很多情况下变量之间不是线性关系，而是更复杂的非线性关系，比如一些常见的关系可以表示为对数函数，指数函数，多项式函数等。在求解非线性回归问题时，通常需要通过观察数据分布情况，然后事先假定一个非线性函数来表示变量之间的关系。对于如何求解非线性函数的参数，一个简单的方法就是将复杂的非线性关系通过变量转换，将其变为线性关系[31]。比如，$Y=aX^b$，可以通过变换$Y'=\ln Y$和$X'=\ln X$，将

关系变为$Y'=a'+bX'$。但是这种方法，需要的前提条件是变量之间的关系能够事先通过观察得出大概性质，而如果变量太多或者变量间的关系太复杂，那么就无法应用这种方法。

对于比较复杂的非线性关系，可以通过一些功能更强大的机器学习算法来求解。在机器学习中的一些解决分类问题的算法大部分也可以用来求解回归问题[32]，如K近邻算法，决策树，人工神经网络，支持向量机等。下面选取几种算法简单说明其实现原理。

K近邻算法，在很多情况下是一个简单且有效的非参数分类算法。它的基本思想是将每个样本点看作一个样本空间中的点，样本的维度就是点的维度，对于训练集来说，每个样本点已经具有了自己的分类标签。然后对于测试集中的一个无标签的样本点，根据一个特定的距离函数，计算其与训练集中的每个点之间的距离，然后选取其中与测试样本最近的k个点，将其中所占比例最大的分类标签作为该测试样本的标签。对于回归问题，则可以选择k个最近样本点的均值（或者带权均值）作为该测试样本的预测输出。K近邻算法的主要缺点是它的效率比较低，而且结果的好坏依赖于k的选取[33]，所以在一些对效率要求比较高的应用中，K近邻不是很适用。

支持向量机是一种有监督的学习方法，可以同来解决分类和回归问题。当我们要解决线性回归问题的时候，通常有很多有效地方法可以选择，而且线性回归算法要比非线性回归算法简便得多，所以支持向量机的核心是找到一个非线性函数，将原来不能处理的样本数据转换到可处理的高维空间，从而将问题转化为线性问题[34]。如图2.3所示，原来在样本的低维特征空间不能进行处理的样本，通过Φ函数被变换到更高维特征空间之后，变为了可以进行线性分类的样本。然后通过使分类间隔达到最大来选取分类的超平面，即可得到最优的分类器。在这里同样遇到的问题是如何得到非线性映射函数Φ，在原始空间中一般很难直接观察出分布的特点，从而Φ的选取也十分困难，支持向量机的解决方法是通过选取核函数代替显式的Φ。核函数是支持向量机的核心思想，核函数可以把高维空间中两个点的距离计算转换到在原始样本空间中计算，因此不必知道Φ的具体形式即可直接求解分类平面。核函数的一般形式如下：

$$K(x,y)=<\phi(x),\phi(y)> \tag{2-7}$$

选取不同的核函数，分类的效果往往也不相同，所以这也是支持向量机的一个研究重点。基本的常见核函数一般包括了 S 形的 Sigmoid 核函数、线性变换的线性核函数以及呈圆形对称的高斯核函数等。对于分类，通过符号函数可以将输出限定在 1 或-1，对于回归，直接将样本值带入求得的方程即可得到预测输出。支持向量机的一个不足之处

在于对大规模训练样本的处理上效率较低。

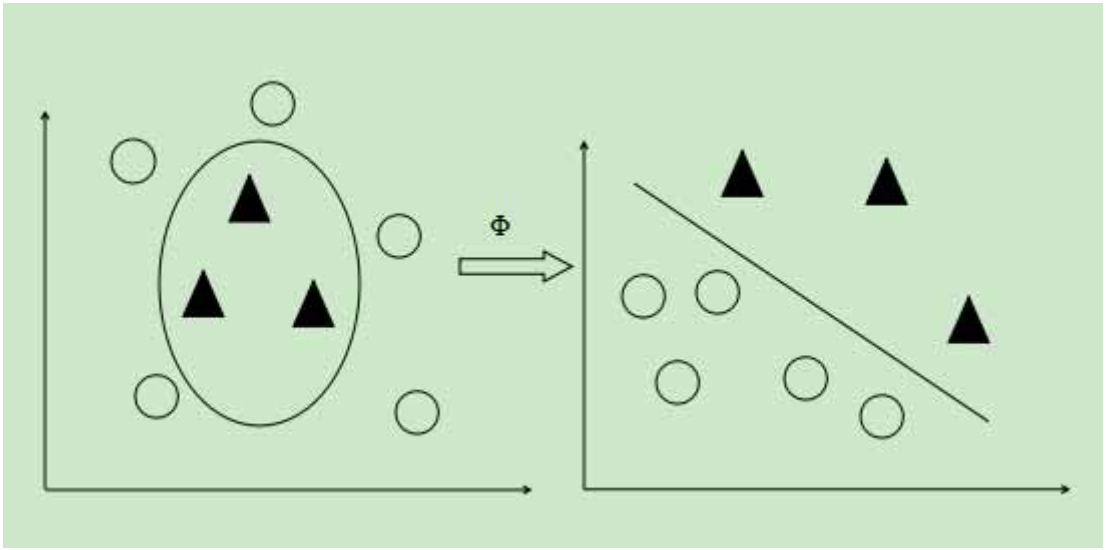


图 2.3 SVM 映射

人工神经网络技术近年来快速发展，已经广泛用于人工智能，模式识别，回归预测等领域。神经网络模拟人脑神经元的工作方式，虽然起源于生物神经学，但现在已经在统计领域占据了重要地位[35]。人工神经网络发展到现在，已经在多个领域展现出了强大的能力，而且根据网络结构、学习算法或神经元数量等标准，人工神经网络又可以分为许多类别。比较常用且原理较为简单的神经网络模型是应用反向传播的前向神经网络，也叫作 BP 神经网络。

BP 神经网络包含一个输入层与样本的特征向量相连，向量的维度和节点数相同，一个输出层与样本的输出向量，向量的维度和节点数相同，以及多个隐藏层，每层的节点数可以任意改变。每一个节点都和相邻层的所有节点相互连接并且具有特定的权重，下一层的结点接收上一层结点的加权平均作为输入，并且通过一个激活函数限定输入的范围。可以表示为：

$$I = F(\sum_{i=0}^{k} \alpha_i x_i) \tag{2-8}$$

I 表示当前节点接收到的输入，$F(\cdot)$表示激活函数，α 表示权重，x 表示上一层结点的输出。神经网络可用的激活函数有很多种，包括，线性函数、阈值函数、S 型函数、ReLU 函数等。反向传播是一种基于实例来逼近函数的方法[36]，其原理如图 2.4 所示。BP 神经网络的理论基础是梯度下降法，利用梯度下降搜索技术，达到使网络的输出值和期望的

标准输出值的均方误差最小。具体的反向传播算法包括输入信号的前向传播，误差的反向传播两个过程。首先随即初始化权重，输入信号按照从输入层经过隐藏层最后到输出层的顺序传播，通过激活函数引入的非线性，输入信号被映射到其他向量空间，最后得到当前模型的输出，然后计算当前误差，若误差不满足要求，则转入误差的反向传播过程。误差的反向传播是将模型产生的误差通过隐藏层向输入层逐层传播，根据误差调整每层之间的权重，使误差沿梯度方向下降。经过多次地训练，最后确定与最小误差相对应的网络参数(权值和阈值)，则神经网络的训练过程结束。那么现在得到的神经网络就可以对新输入的样本自动进行处理，得到要预测的目标值。BP 神经网络的缺点是训练速度慢，容易陷入局部极小值，而且网络层数、神经元节点个数的选择没有相应的理论指导。

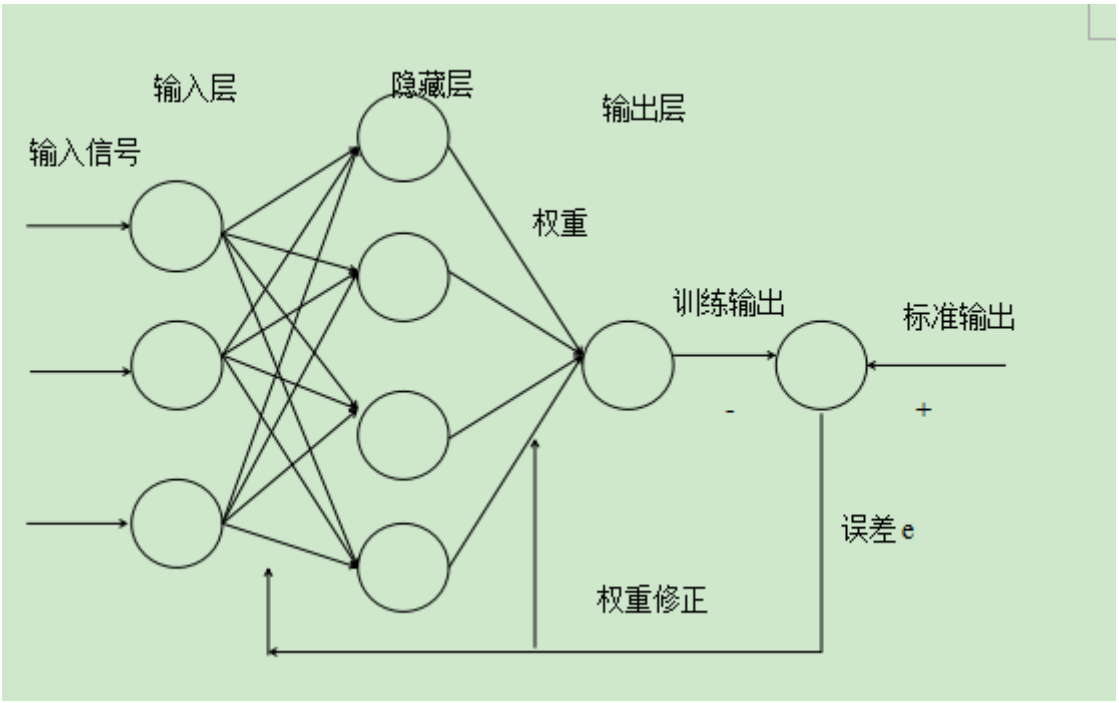


图 2.4 神经网络反向传播

神经网络的一个重要优点是它能自动逼近任何非线性数学函数。当变量之间的关系不知道或者十分复杂时，很难进行统计处理，神经网络在这一方面尤其有用。然而，隐层层数、隐层节点数等与神经网络相关的参数的选择并不直观，寻找最优的神经网络结构是一个非常耗时的过程，而且神经网络的权重选择也是影响预测效果的一个重要因素。虽然在许多问题上神经网络可以表现的很好，但是因为神经网络是一个“黑箱”，它的结果通常难以解释，所以很多统计学专家可能并不会轻易接受神经网络的结果，这

也是对神经网络争论得最多的话题之一。

2.3 模型优化

最基本的模型虽然能解决一些问题，但是随着数据的增多或问题复杂性的提高，基本的模型往往很难直接取得很好的效果，所以需要对模型进行优化，以提升模型的性能。一般提升性能可以从几个角度着手。

从数据的角度，如果数据中的异常点太多或者数据量不足以支撑构建出一个完善的模型，那么就很难获得一个较好的结果。对于所有机器学习算法来说，数据是一切方法能够实现的基础，如果数据能够完美表现出应有的形式，那么大多数模型都可以很好地工作。但事实是，现实中的数据因为多种原因，可能包含各种异常点，或数据不足，或维度过高。对于提高数据质量最直接的办法就是尽可能获取更多的数据，但可能由于客观原因的限制，往往很难获得更多数据，所以就需要根据已有数据，自己对其进行“改造”，平移、截取、旋转等。另一个比较出乎意料的手段是增加噪声，它起到了与正则化方法类似的作用，可以抑制训练数据的过拟合。对于特征维度过高的问题可以运用一些特征选择方法，筛选出相关性高的特征，去掉相关性较低的特征，例如主成分分析法（PCA）或 线性判别分析法（LDA）。归一化是一种简化计算的方式，可以将有量纲的变量，经过某种转换，变成无量纲的变量。从效率上来看，对数据进行归一化处理，可以加快模型的运算速度，而且可以消除各个特征维度量纲不同带来的影响。归一化将数据映射到一个规定的范围内（通常是 0-1 或-1-1），常用的映射转换通常如下：

$$X^*=(X-Xmin)/(Xmax-Xmin) \tag{2-9}$$

除了提到的这些之外，还有很多对数据进行处理的方法，由此可见数据对机器学习来说是至关重要的。

从模型本身的角度，也可以进行调整已获得更好的结果，前文提到的正则化就是对模型的一种有效的改进。大多数的机器学习的方法都会包含几个超参数，这些超参数控制着模型的基本框架，通过调整超参数可以改善模型的效果，常用的选择超参数的手段是通过网格搜索，对范围内的超参数进行枚举，然后利用交叉验证选择出最优的超参数。除此之外，不同的基础模型往往有不同的优化方法，例如，神经网络可以通过调整学习率，Mini-batch，调整网络结构等方式进行优化，而其他模型又有不同的优化方法。

有一类学习方法被称为集成学习（ensemble learning），通过将多个弱分类器通过某种手段进行整合，得到更优的结果。常用的集成学习包括 bagging 和 boosting。Bagging

一般都基于自助采样（Bootstrap sampling），即从原始的包含 n 个样本的训练集中，有放回的随机选择 m 个样本，形成一个新的子集，因为此过程是随机的，所以每次采样得到的集合都与原始训练集有所不同。通过这 k 次不同的采样过程，就得到了 k 个不同的子集，然后再针对这 k 个子集进行训练，得到独立的 k 个弱分类器，最后通过一定的组合方式得到最后的强分类器。随机森林就是基于 bagging 的机器学习方法。Boosting 首先需要对训练集的每个样本赋予相同的权重，然后先训练出一个弱分类器，根据该分类器的结果，更新样本的权重（分类错误的样本更受到重视，应赋予更高的权重），同时为该分类器本身赋予一定权重。然后在调整了权重的样本上，重新进行弱分类器的学习，直到分类器数量达到指定的 k 个，然后将 k 个分类器按权重进行整合得到最终的强分类器。梯度提升决策树就是基于 boosting 的机器学习方法。

除了以上提到了方法，模型优化的手段还有很多，而且要根据具体问题具体判断，只有适合特定问题的模型才是理想的模型。

2.4 现有方法的不足

通过上面的介绍，可以总结出目前的血压预测上存在的几点不足：

1、传统的测量方法或常用的基于柯氏音的的血压测量方法，往往操作繁琐，而且容易造成被测者不适，影响被测者正常工作，难以用来连续监测。

2、现有的利用 PPG 预测血压的方法中，通过 PWV 求血压的方案，需要借助多种设备，测量过程复杂，不适合日常生活中的连续监测。而目前大多数基于脉搏特征与血压建模的方法，都是默认脉搏与血压是线性关系，建立线性方程来求解血压值，但是通过对 PPG 信号样本的观察，其与血压间可能存在更复杂的非线性关系，所以需要更加适合的模型来求解。可以分别应用上文提出的线性回归预测模型和非线性回归预测模型对数据进行建模，并且根据我们的问题本身的特性，对模型进行一定的优化，然后比较它们之间的优劣性。

3、现有的多数利用脉搏预测血压的方法，都直接将不同人群组合在一起作为训练数据，直接进行学习，忽略了不同群体脉搏差异较大的因素，一定程度导致了模型预测准确性低，因为不同群体的脉搏间存在差异，其与血压的关系也可能不同，所以通过一定手段需要解决群体差异的问题。

2.5 本章小结

本章首先介绍了 PPG 信号的获取方式和 PPG 信号的特点，为后续的工作提供依据，然后分别介绍了线性回归预测模型和非线性回归预测模型，并对几种比较有代表性的方法进行了介绍。然后从几个角度，说明了如何对基本的模型进行优化，以提高模型的准确性，这部分可以作为之后建模的理论基础。最后根据目前的研究现状，总结了现有工作的几点不足，对应的将在后面的工作中进行完善。

第 3 章 模型实现与优化算法

本章首先对本文需要解决的问题进行描述，并应用基本的机器学习回归模型给出相应的求解方法，然后通过聚类方法，在前期的训练数据上对模型进行优化。

3.1 问题描述与模型实现流程

在这一节中，对本文的血压预测模型的求解过程进行说明，并且选取了几种不同的模型实现回归预测。

首先本文所使用的数据按照不同的被测者分为多组，每组为 10 分钟内，以 100 Hz 连续采样获得的实数。在第二章的图 2.2 中，已经介绍了 PPG 信号的形式，将每组 10 分钟的数据按照时间关系绘制出来得到的图形与图 2.2 相似。根据之前的相关研究，我们可以从原始的波形中提取出很多特征。观察图形，可以比较直观地看出几个特征，如波形幅值的极值点，极值点之间的时间差和幅值差，波形周期等。结合一些相关文献，还可以提取出很多潜在的特征，如对原始数据求一阶偏差极值点，二阶偏差极值点，波

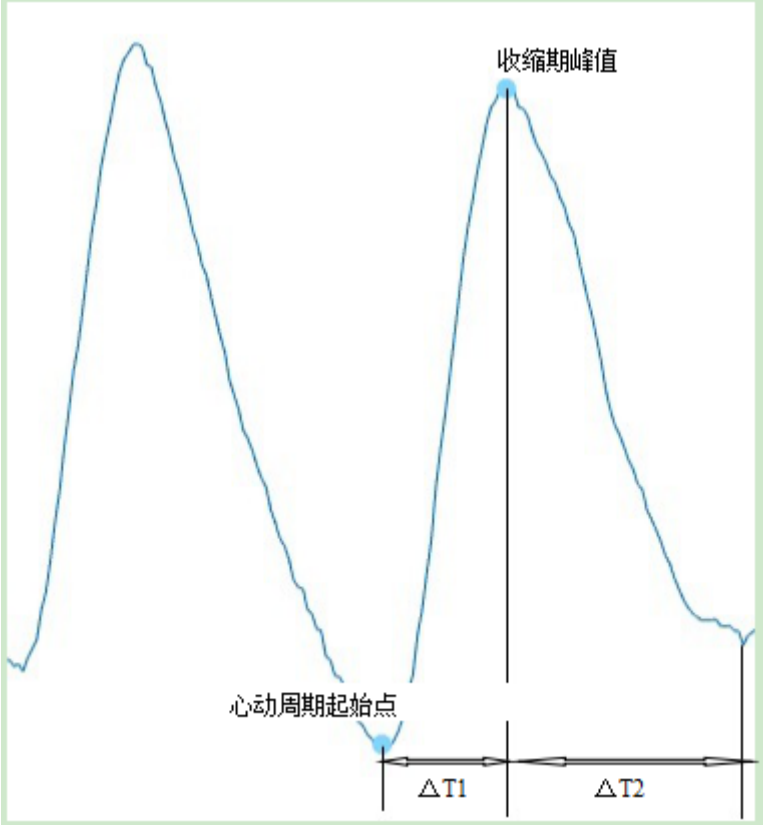


图 3.1 脉搏波主要特征

形围成的面积等。甚至对原始波形进行时域-频域的变换还可以提取更多特征。在所有这些特征值之中，有的特征与血压的相关性比较高，有的特征与血压的相关性比较低。另一方面，考虑到原始波形数据的采样质量以及提取特征的难度和准确性，本文选取了4 个与血压相关性最高同时提取准确性较好的特征作为模型训练的输入向量，这些特征也已经被很多研究证实了其较高的相关性，并且被很多研究应用，具体的特征提取情况可以参考图 3.1。

4 个特征如下：

1) 心动周期起始点振幅：脉搏信号每一个心动周期的起始点，也是每个周期内的最低点，一般情况下与低压的相关性较高。

2) 收缩期峰值点振幅：每个周期内心脏收缩期的最高点，一般情况下与高压的相关性较高。

3) 收缩时间：每个周期内心脏收缩的时间段。

4) 舒张时间：每个周期内心脏舒张的时间段。

特征提取过程如下：

1) 首先，需要计算 PPG 信号的周期，一方面用于求解特征值，另一方面用来在后面剔除异常的特征点。常用的求取周期的方法有快速傅里叶变换和小波变换。傅里叶变换可以将原始的随时间变化的曲线，分解为不同振幅不同频率的正弦波，因为正弦信号的形式比较标准，同时正弦信号具有一些方便统计的特征，所以通过傅里叶变换之后可以更加方便的处理原来的信号，而且将信号从时域转化为频域可以让我们通过另一个角度去分析数据。有些信号在时域上很难看出有用的特征，而在频域上可能就比较容易得到有用的特征。快速傅里叶变换在傅里叶变换的基础上大大简化了计算的复杂度。小波变换可以完成傅里叶变换的功能，同时小波变换也对快速傅里叶变换的一些缺陷有所弥补。对于非平稳信号来说，傅里叶变换不能分辨不同频率部分出现顺序不同的波形，而小波变换可以有效地处理非平稳信号。小波变换将傅里叶变换的基础正弦波更换为有限长的会衰减的小波基，这样就不仅能够有效地获取频率谱还可以定位到时间。对于本文来说，因为只需要周期一个信息，所以本文选择了更便捷的快速傅里叶变换来求周期。通过快速傅里叶变换将时域的 PPG 信号变换为频域的信号，然后频谱中幅值最大的点对应的频率就是整个信号的平均频率，其相应的倒数就是所求周期。

2) 第二步需要求得所有的极大极小值点。对于一般的连续函数来说，求极值可以通

过对原函数求导数并令导数为 0，得到极值点；或者通过梯度下降法或牛顿法求极值。但是对于本文的离散数据来说，我们无法求导，也不能直接建立表示波形与时间关系的算式，所以需要能直接对离散数据求极值点的方法。求极值的方法与求函数极值点的方法相似，样本信号是按采样频率分布的离散点，所以需要进行相应的处理，才能求出极值点。首先，我们需要求出每两个相邻采样点的一阶偏差，然后令所有为正数的一阶偏差为 1，负数的一阶偏差为-1，0 还是 0。然后对修改后的数据再求偏差，即原始数据的二阶偏差，在所得的结果中，如果二阶偏差是-2，则此时间点对应的原始点为极大值点，如果二阶偏差是 2，则此时间点对应的原始点为极小值点。求得极值点后，分别用 4 个数组存储所有极大值，极小值点的幅值和时间坐标，然后通过极小值点的时间坐标来分割一个个周期。结合上一步求出的周期值，每个周期内的前两个极值对应的特征点可以用来求取所需的 4 个特征值，利用周期信息可以直接求出心动周期起始点的振幅，收缩期峰值的振幅，收缩时间和舒张时间。

3) 然后，根据每个周期内极值点的数量和周期可以判断此样本是否包含重搏波。对于非异常点的极值点，如果在一个周期内出现两对极小值或极大值点，则认为波形属于具有重搏波的类型；如果一个周期内只出现一对极小值或极大值点，则认为波形属于无重搏波的类型。根据判别结果可以将波形进行分类，针对不同类型的波形，采取相应的计算方法，得到本文所需要的特征点，并求出相应的特征值，为后续的工作内容提供训练模型所需的样本数据。

4) 最后，在提取特征点的过程中，需要判断每一个得到的特征点是否是正常的采样点。因为一些不当操作或者设备故障会导致异常点的出现，所以需要通过一定的手段剔除异常点。例如，对于幅值异常的点，在对一个用户求特征时，可以记录下所有采样点幅值的数值，然后通过计算得出均值和方差，对每一个求得的极值点，令其与计算得到的均值和方差进行比较，对于相距均值较近的点保留，而相距均值较远的点则认为是异常点，而且可以通过参数控制剔除异常点的严格程度。而对于周期不正常的样本段，可以与上面通过快速傅里叶变换求得的周期值进行比较，保留合理的周期段，剔除相差较大的周期段。除了对脉搏波信号进行去异常点，还要剔除血压值极度异常的点，这一步可以根据医学上的一些统计数据得到一个正常的血压范围，保留范围内的血压值，超出范围的认为是异常点，进行剔除，同时将时间上对应的脉搏特征也去掉。到这一步为止，就完成了特征提取的全部流程。

至此，已经得到了训练模型所需的特征和标准的血压值，下面就可以根据训练数据得到基本的预测模型。

首先对 4 个特征值进行归一化处理，然后令 $\boldsymbol{X}=\{x_1,x_2,x_3,x_4\}$代表最终得到的 4 个特征值，令 $\boldsymbol{Y}=\{y_s,y_d\}$代表标准的高压和低压，预测模型可以表示为：

$$\Phi_\alpha(\boldsymbol{X})=Y' \tag{3-1}$$

Y' 表示模型的预测输出（高压值或者低压值），α 表示模型的超参数。我们的目标是使 Y' 尽可能的接近实际的血压值，为了得到最优的 $\Phi_\alpha(\boldsymbol{X})$，将样本数据分为两部分分别用于训练和预测，并且通过模型的训练找到模型最优的超参数。

在本文中，主要使用 LS（最小二乘线性回归），LASSO（Lasso 回归），SVM（支持向量机）和 ANN（人工神经网络）4 种方法来得到血压预测模型，其中包含了两种线性模型和两种非线性模型，后续将对每种方法得到的模型预测效果进行对比。

3.2 基于聚类的优化算法

3.2.1 聚类优化算法思想

在通过脉搏波形来预测血压时，因为数据采集难度，隐私保护等原因，我们没有办法得到除脉搏波形外的其他因素，例如年龄、性别、生活习惯以及疾病史等因素，而这些因素往往都从某一方面影响血压的高低。通常来说，老年人的血压会高于年轻人，男性的血压高于女性，生活习惯良好的人也会比有不良生活性的人的血压更接近正常范围，这些反映到脉搏上，也会体现出不同的形式，如中医上会提到的脉位，脉数，脉形，脉势等。如果能得到这方面的数据，那么将它们作为训练集的一部分一起训练模型的话，会使得模型考虑到更全面的因素，预测能力更好。但现在无法获得这些信息，所以我们只能在脉搏波数据的基础上对血压进行预测。为了消除这些未知因素的影响，一种方式是将从每一个被测试者身上采集到的数据，单独构造成一组训练数据，然后为每组数据建立一个机器学习血压回归预测模型，这样的话就可以消除不同类型人群间差异带来的影响。但是这样的方式，计算起来比较复杂，而且不具有实际的应用价值。原始的将所有数据提取出特征之后，组合到一起形成一个整体的训练数据集，然后一起进行机器学习血压回归预测模型的训练的方式，没有考虑不同人群间的差异，所以训练得到的结果准确性较差。

对于这个问题，本文采取的方式是：在基本预测模型的基础上，首先对样本的特征

值进行聚类，通过聚类可以将原始数据，按照不同的模式分为几类，也就将那些不同人群间的差异通过反映在脉搏波上的不同模式加以区分，相似的人群被归为了一类，然后对每一种类别单独建立预测模型，这样的模型就适用于这一模式的数据。对于预测也同样根据之前的聚类结果，将每个预测样本分到相应聚类中，通过相应的模型进行计算，得到预测血压值，如图 3.2 所示。

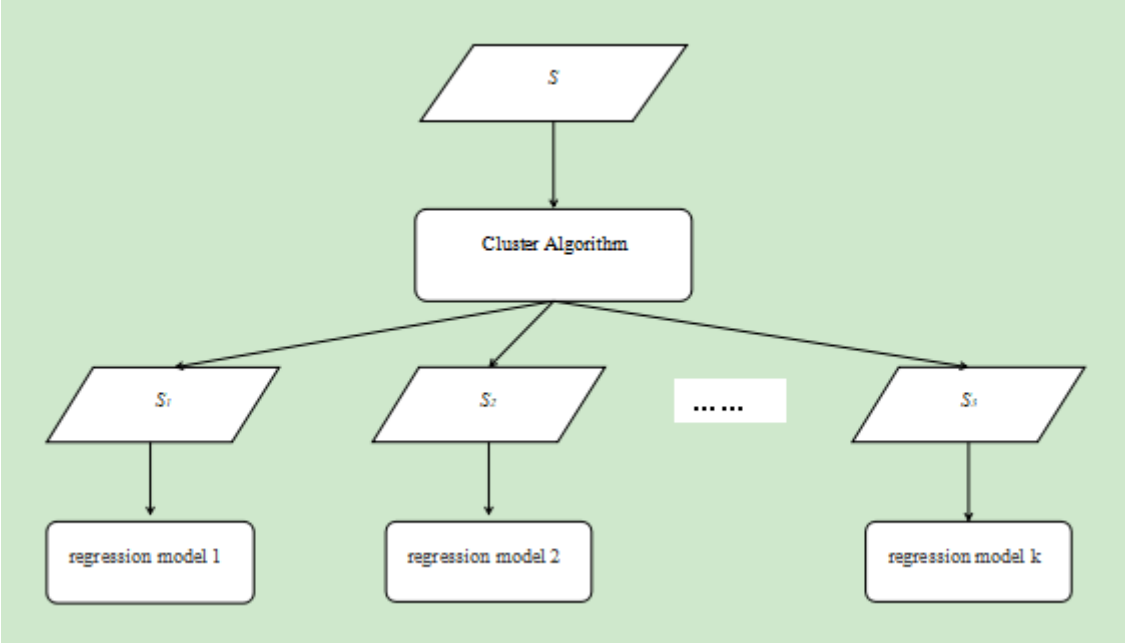


图 3.2 聚类优化流程

聚类算法是研究分类问题的一种重要的数据挖掘算法，聚类与分类的不同点在于聚类所需要划分的类别是未知的，聚类是一组数据潜在模式的一种体现。聚类算法以样本间的相似性为基础，目的是将整个样本集合分成多个聚簇，而每个聚簇内样本间的相似度最大，这样就可以将原始样本集按照一定的规律分成多个独立的类别，实现数据的抽象概括。作为聚类算法中最重要的关键点之一，相似度的衡量有着多种选择。从数据的角度，可以对从原始数据提取特征后构建特征向量集合，两两特征向量直接计算相似度，或者直接在原始的样本数据上求取两个样本的相似度，而在本文中，分别对这两种方式进行了比较。从距离的计算方式来看，常用的距离函数包括：

1) 欧氏距离，欧几里得距离是最常用的距离定义之一，N 维空间内的两个向量之间的欧几里得距离就是这两个点在空间中的实际距离，所以欧氏距离很容易理解，也很直观并且很自然的可以表示距离。

2) 曼哈顿距离，也是一种很常见的距离度量方式，其在几何空间中表示两个点每

一个维度间距离的总和，在度量棋盘中棋子的距离或者度量城市间距离时常用到曼哈顿距离。

3) cos 距离，余弦距离比起其他距离度量方式有一个优势就是，其他距离度量方式，大多都需要事先对数据进行归一化处理，而余弦距离则不需要。余弦距离通常用 1 减去两个向量间夹角余弦值求得，余弦值越趋近于 1，两个向量越相似，余弦值越趋近于-1，两个向量越不相似，即余弦距离越小，两个向量越相似。余弦距离是一种很有效的距离度量方法，经常用在求文本相似度上。

4) correlation 距离，相关性距离常用来表示两个随机变量 X 与 Y 直接的相关程度，用 1 减去相关系数得到的数值就是相关距离。与余弦距离相似，范围也在[1，-1]之间，相关距离越小，两个向量越相似。

通过对这几个距离函数的比较分析并结合之前的相关工作，对本文的问题而言，欧式距离获得的效果比其他几种更好，所以最后本文选取了欧式距离作为计算聚类的距离函数。

以上几种距离度量方式都是基于提取后的特征向量进行距离的计算，而本文的原始样本是一组波形数据，而特征值是在原始波形的基础上提取得到的，所以通过特征值进行聚类的话会丢失一部分原始信息，所以可以考虑直接在原始波形上进行聚类。为了能够对波形数据进行聚类，就要能够比较两个波形序列的相似度，而不同的样本一个周期内的波形序列长度可能不一致，因此传统的欧氏距离不能直接用来求解两个波形序列的相似度，为此引入了动态时间规整方法，此方法能够有效计算不同长度的波形数据的相似度，因此也就能够实现直接对波形数据进行聚类。

动态时间规整[38]，原本是用在求语音序列相似度中，因为不同的人发音习惯不同，语速不同导致了相同音素表现为不同波形序列，为了提高识别率，研究人员提出了动态时间规整算法。该算法基于动态规划的思想，将两个不同长度的序列进行调整，使其长度变得相同并且相似的采样点尽量对其。如图 3.3 所示。在比较两个序列的相似度时，虽然两个序列的长度不相等，但是有时通过观察序列的形状，可以发现其形状是近似的，在这种情况下，时间上对应的采样点，在宏观上可能不是最接近的点，这样直接对时间上对应的点求距离，往往得不到两个序列真正的相似度。因此，需要为每个点找到其真正需要对齐并比较的点，动态时间规整就是用于解决此问题的方法，本文将此方法应用到对脉搏序列进行聚类的步骤中，具体的算法执行步骤，将在下一节详细描述。

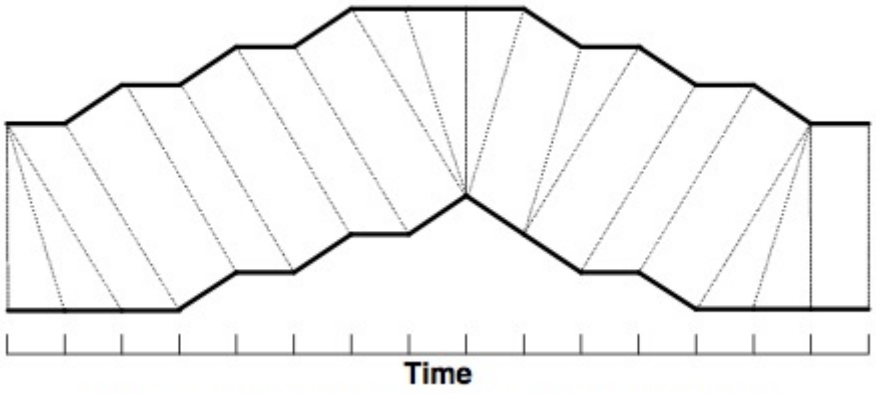


图 3.3 动态时间规整示例

在聚类算法的选择上，本文主要考虑了 k-means 和 k-medoids 两种，根据之前的研究，参考文献[37]指出了 k-medoids 略优于 k-means，而且 k-medoids 对于“噪声”不那么敏感。因此本文之后的实验选取 k-medoids 作为聚类优化方法采用的算法，欧式距离和动态时间规整距离作为聚类优化方法采用的距离函数。

除了聚类采用的算法和聚类距离度量函数的选取之外，对于 k-means 或 k-medoids 来说，k 值的选取也是很重要的一点，需要慎重选择。本文对于 k 的选取，使用了比较直观的方法，初始先选取一个较小的 k 值（如 2），然后执行聚类算法，通过相应的衡量指标，评价当前 k 的取值下的聚类效果，然后逐步增大 k 值，对每一次取值都得到一个聚类效果的评价，直到某个 k 的取值，聚类的效果没有明显的改善，即可停止 k 的增加，选择当前的 k 值作为最终执行优化算法的 k 值。这样的选择过程，既可以保证一定的聚类效果，又可以尽量降低聚类模型的复杂度。

3.2.2 算法实现

聚类优化算法如下：

表 3.1 聚类优化算法

输入：特征样本集合或原始波形样本集合 S，聚类数 k

输出：计算得到的聚类中心 C_k，k 个聚类集

步骤如下：

1. 确定选取的距离函数（欧氏距离或者动态时间规整距离）
2. 选取 k 个聚类中心

3. While(聚类结果未稳定)

4. 对每一个样本，按照当前聚类中心，得到相应类别

5. 重新计算聚类中心

6. 得到 k 个稳定的聚类中心 C_k 和 k 个聚类

算法中，对初始聚类中心的选取，按照 k-means++的初始值选取算法处理。其聚类中心的选取过程为：

1．随机选取第一个聚类中心。

2．计算其余所有样本到最近聚类中心的距离。

3．按照与距离成正比的方式为每个样本分配一个概率，并根据此概率选取下一个聚类中心。

4．重复第 2 步和第 3 步，直到选够 k 个聚类中心。

因为无论是 k-means 还是 k-medoids 聚类方法，都对初始值的选取比较敏感，应用此方法得到的初始聚类中心通常比完全随机选择的聚类中心更合适，在求解时算法收敛更快而且获得的聚类结果更好。

聚类算法停止迭代的条件主要有 3 点：

1．每个样本其所在的聚类标号都不再发生变化（但这种情况要求比较苛刻）。

2．在两次迭代中，所有聚簇的距离总和小于某个阈值。

3．算法的迭代次数已经达到了预先设定的迭代最大值。

得到训练集的 k 个聚类和聚类中心后，分别对每个聚类建立预测模型，这样就得到 k 个独立的预测模型，然后对于测试的样本，先计算出其特征值，然后按照其与聚类中心的距离（如果是采取动态时间规整算法的话，则先计算原始波形数据与聚类中心的距离，再求取特征值），计算当前样本应该属于哪个聚类，并将样本交给对应的预测模型，最后得到其血压预测结果。

动态时间规整的距离求解实现过程如下：

对于给定的两个序列 Q 和 C，假设其长度分别为 n 和 m。$Q=[q_1,q_2,\cdots,q_n]$，$C=[c_1,c_2,\cdots,c_m]$。

1) 如果 $n=m$，则直接计算两个序列的距离；

2) 如果 $n\neq m$，则需要对齐两个序列，可以构造一个 $n*m$ 的矩阵，矩阵的每个位置 (i,j)的元素都表示 q_i 和 c_i 的距离 $d(q_i,c_i)$，一般都是欧氏距离，即 $d(q_i,c_i)=(q_i-c_i)^2$。

所以方法的目标就是找到一条从(1,1)到(n,m)的路径，路径上的点，就是两个序列对齐的点。

可以把选取的这条路径定义为 P，$P=p_1,p_2,\cdots,p_s,\cdots,p_K$，$\max(n,m)\leqq K<m+n+1$，其中每个元素 $p_s=(i,j)_s$，为了找到合适的最优路径，需要满足几个前提条件：

1) 边界条件：即满足 $p_1=(1,1)$，$p_K=(n,m)$。
2) 连续性：路径上任意两个连续的点，后一个点的坐标比前一个点的坐标最多大 1，而不能跨点对齐。
3) 单调性：路径上任意两个连续的点，后一个点的坐标至少比前一个点的坐标大，即随时间向前，而不能向后对齐。

最后可以得到路径选择的策略

$$p(i,j)=\min\begin{cases} p(i-1,j)+d(i,j) \\ p(i-1,j-1)+2d(i,j) \\ p(i,j-1)+d(i,j) \end{cases} \tag{3-2}$$

最后得到 $p(n,m)$的值就是两个序列的相似度。

3.3 基于梯度提升的优化算法

本节主要从集成学习的角度对原始的血压预测模型进行优化，主要选取了梯度提升算法结合本文问题的特点对多个分类器的结果进行集成，旨在获得更优的预测准确性，下面将具体讲解如何应用梯度提升算法以及其实现。

3.3.1 梯度提升优化算法思想

集成方法是通过构造一组分类器，然后对于新的样本点通过多个分类器的预测值加权投票得到结果的学习方法[39,40]。常用的两种集成学习方法包括 bagging 和 boosting，这两种方法从不同侧重点对单一学习器进行了选取和组合，使得集成的效果优于单一学习器。bagging 和 boosting 的区别在于 bagging 每一次的模型都不依赖之前的结果，所以可以并行计算，而 boosting 每一轮的分类器的构建都依赖于上一步的结果，所以计算速度相对较慢。Bagging 方法的代表算法是随机森林，而 boosting 方法的代表算法是 AdaBoost。本文所采取的梯度提升算法（gradient boosting）与 AdaBoost 基本的思想都是“知错就改”，在每一步模型求解时，都将力求补足之前模型的不足，而二者的不同之处在于，AdaBoost 中当前模型的不足由样本权重决定，而梯度提升中当前模型的不足

由残差的梯度决定，从这个方面来看，梯度提升的求解相对更简单一些。

梯度提升是一种可以用来解决回归和分类问题的机器学习技术，它通过集成多个较弱预测模型的形式来得到最终的预测模型。可以将其看作是针对特定损失函数的一种优化算法。梯度提升思想最常用在梯度提升决策树（Gradient Boosting Decision Tree）中，其基本的思想是，先利用原始的训练数据学习，得到初始的一颗决策树，然后在每一颗叶子节点处可以得到当前决策树预测的值，然后利用当前模型预测的残差值去训练下一颗决策树。如果到某一轮学习结束时，残差值仍然很大，就需要继续下一轮的训练；否则，如果预测值基本等于真实值，即可结束训练。最终整个模型的预测输出，就是之前几轮训练过程中每一棵树的判别输出值的叠加。

具体来说，梯度提升通过一种可累加函数的形式表示整体模型：

$$F^i(x)=F^{i-1}(x)+h^i(x) \tag{4-1}$$

以迭代的形式表示模型的求解过程，那么求解的目标就是使 $F^i(x)$最接近真实值，另一个角度看就是使每一步的单一模型的预测值接近真实值与上一步迭代的预测值 $F^{i-1}(x)$的之差（或其他误差函数）。那么整个过程就是，首先利用初始的样本，训练得到一个最初始的模型，然后计算当前得到的模型的误差，然后使下一轮的模型去拟合损失函数关于预测值的反向梯度（对于平方误差来说，梯度值就是真实值与预测值的残差），并按式 4-1 更新模型，直到模型趋于稳定或者达到最大迭代次数，则最终获得的模型就是梯度提升的结果。梯度提升方法的优点就是充分考虑了每一个弱分类器的结果，预测准确性很高，而且每一轮的训练过程都可以选择不同的弱模型，可以充分结合不同模型的优势；其缺点就是每一次的训练过程都要依赖于上一步的结果，所以训练过程不能并行，比较耗时。

除了通过梯度提升方法改进基本模型之外，对于本文来说，我们知道，需要进行预测的量是血压的高压值与低压值，可以容易地发现正常人的血压高压与低压一般都会满足一定的关系，即高低压差值只在一定的范围内波动，即使血压整体有较大变动，其差值也不会变化太大。通常医学上认为，正常人的血压，高压和低压的差值为 20~60mmHg，小于 20 或者大于 60 都可能是因为病理性的原因或者机器采样中产生的误差引起。所以，利用血压间的这一关系，本文可以对梯度提升算法进行一定的修改。

首先，为了表示样本中高压与低压的关系，本文首先通过计算训练样本中高压与低压差值的均值和标准差，这样就得到了当前训练集的高低压差值的大体波动区间，即均

值附近的某个区间，假设为[a,b]。然后按照正常流程进行梯度提升的求解，在每一轮的训练中，考虑一种情况：对于某一个训练样本，假设当前的整体模型得到的高压和低压的预测值为 sp 和 dp，其差值为 d=sp-dp，如果 d 在[a,b]的范围之内，我们有理由认为当前得到的预测值是满足血压变化规律的，对于这个样本模型产生的误差是由于随机误差产生的。所以对于这种样本，在下一轮的训练中，我们可以将其的残差视为 0 或者根据其与均值的远近调整残差值，然后再对那些调整后的残差值进行二次训练，得到下一轮的模型，这样我们就在梯度提升算法的基础上，同时利用了高低血压间的关联，有针对性地对模型进行了改进。

如果随着机器学习模型越来越复杂，虽然在训练集上可能误差会逐渐减小，但是当模型复杂度达到一个临界点时，随着模型复杂度的提高，在测试集上的预测效果可能越来越差。集成学习本身就能一定程度降低过拟合的风险，除此之外，正则化、数据集扩增等手段都可以减少过拟合。在本文中，通过将血压差这一项引入梯度提升算法的训练过程中，我们不仅可以考虑到高压与低压间的相关性，从另一个角度来看，通过修改每一次梯度提升过程产生的残差，相当于根据数据的分布产生了新的数据，可以有效减少模型过拟合的风险。

本文主要采取了两种具体的算法来实现梯度提升优化，将在下一节详细说明。

3.3.2 限定阈值的梯度提升方法

在检测一组数据中的异常点时，有一种方法叫做拉依达准则或 3σ 准则[41,42]。其内容是，如果一组检测的数据只含有随机误差，并且满足或近似满足正态分布的话，那么就可以通过计算得到这组数据的均值和标准差，通过均值和标准差确定一个区间，只要是超出区间内的数值，都认为不属于随机误差，而是应该剔除的粗大误差。在正态分布中，如果假设 σ 代表标准差，μ 代表均值的话，数据分布在$(\mu-\sigma, \mu+\sigma)$中的概率是 0.6826，数据分布在$(\mu-2\sigma, \mu+2\sigma)$中的概率是 0.9544，数据分布在$(\mu-3\sigma,\mu+3\sigma)$中的概率是 0.9974，所以可以看出数据几乎全部集中在$(\mu-3\sigma, \mu+3\sigma)$区间内，数据分布超出这个范围的可能性只有不到 0.3%，因此拉依达准则通常认为$|x-\mu|>3\sigma$ 的样本 x 是应该被剔除的异常，也就是 3σ 准则。对于本文的数据来说，高压值和低压值的分布是满足正态分布的[43]，而高压和低压的差值近似满足正态分布，而且我们的采样数据足够多，可以满足准则的要求，因此可以应用拉依达准则来分析本文中高压和低压间的关系。

在本文中，我们利用训练集求出高压值与低压值之间的差值，将其作为一个随机分布的变量，然后计算出其均值和标准差，然后在这组训练数据上利用机器学习方法进行这一轮的建模，通过模型可以得到一组输出的高压值和低压值，其与训练集一一对应。对于模型输出的预测结果，计算出每一对高压值与低压值之间的差值，通过拉依达准则判断其是否分布在合适的范围。与一般的拉依达准则有所不同的是，在本文中，通过设定一个参数 α，来选择控制范围需要多少倍的标准差，即$|x-\mu|>\alpha*\sigma$ 的结果认为其对应的训练样本是训练效果不佳的点，反之认为此点训练出的血压值已经属于正常的分布区间。所以对于训练结果的血压差值在区间内的样本下一次迭代不再拟合它，可以令其下一轮训练的标准输出为 0；而对于训练结果的血压差值在区间外的样本，下一次迭代时按照正常方法拟合其残差。具体操作过程如下：

表 4.1 限定阈值的梯度提升算法

输入：特征样本集合 $S=\{X1,X2,\cdots\cdots,Xn\}$，标准血压值集合 $P=\{Y1,Y2,\cdots\cdots,Yn\}$，梯度提升迭代次数 k，超参数 α。

输出：得到最终的集成学习模型 F_k

步骤如下：

1. 对于集合 P，计算每对高低压之间的差值，并求解其均值 MD 和标准差 SD
2. 对于训练集，构造初始的预测模型 F_1
3. 令 i 从 2 迭代到 k
4. 求解模型 F_{i-1} 对于每个样本的误差值 E_j
5. 计算模型预测值每对高低压的差值 D_j
6. 如果$|MD-D_j|>\alpha*SD$，令下一轮的 $Y_j=E_j$
7. 否则，$Y_j=0$
8. 然后利用当前的样本数据求解模型 h_i
9. 令 $F_i=F_{i-1}+h_i$
10. 得到结果 F_k

对于每一轮的预测模型，主要是通过判断预测值的高低压差与样本集合的标准高低压差间的绝对差值的范围是否满足一定条件（即上面的 $\alpha*SD$）来决定下一次迭代的标准输出值，同时其基本步骤依旧遵循梯度提升方法的一般流程，所以主要需要控制的参

数是 α 和迭代次数 k，对于 α 和 k 的选取通过交叉验证和网格搜索逐步遍历多个 α 的取值，从中选出最合适的 α 作为最终决定的参数。对于参数 α 和迭代次数 k 的选取将在后面的章节中进行讨论。

3.3.3 函数映射的梯度提升方法

前一种方法，通过判断每一轮训练后的模型得到的预测值的血压差与是否属于特定的区间，来决定下一轮是否对其进行训练，这种方式比较简单直接，但是可能存在因为太过武断造成训练结果不好的问题，所以在此基础上本文又提出了另一种方法。与第一种方法类似，这种方法同样要事先针对原始训练集求出血压差的均值，假设为 μ，然后在梯度提升每一轮的训练过程中，假设对于某个训练样本，根据当前模型预测得到的残差值为 e，高低压的差值为 x，然后根据 x 与 μ 的偏离程度，将下一次构建模型时需要拟合的数据映射到[0,e]的范围内，即 x 与 μ 相差越大，下一次拟合时的标准数据就越接近残差值，反之就越接近 0 值。这种方法就不会简单的根据某一个血压差不满足特定范围直接将残差置 0，而是通过函数映射将残差根据模型得到的血压差进行调整，所以此方法相对更平缓，不管血压差如何分布，每一次迭代后仍然后保留一部分残差信息。其具体的操作过程如下：

表 4.2 函数映射的梯度提升算法

输入：特征样本集合 S={$X1$,$X2$,……,Xn}，标准血压值集合 P={$Y1$,$Y2$,……,Yn}，梯度提升迭代次数 k，超参数 β。

输出：得到最终的集成学习模型 F_k

步骤如下：

1. 对于集合 P，计算每对高低压之间的差值，并求解其均值 MD
2. 对于训练集，构造初始的预测模型 F_1
3. 令 i 从 2 迭代到 k
4. 求解模型 F_{i-1} 对于每个样本的误差值 E_j
5. 计算模型预测值每对高低压的差值 D_j
6. 令 $bias_j$=|MD-D_j|，令下一轮的 Y_j=map($bias_j$,β,E_j)
7. 然后利用当前的样本数据求解模型 h_i
8. 令 F_i=F_{i-1}+h_i

9. 得到结果 F_k

其中的 map(*bias*,*β*,*error*)有如下定义：

$$\left(1-\frac{1}{e^{\beta\times bias}}\right)\times error \tag{4-2}$$

通过此式，可以将在[0,∞)的模型对样本的预测血压差与样本的标准血压差均值间的绝对差值，映射到[0,error)的范围内（error 为模型对此样本预测得到的残差），β 用来控制映射函数的陡峭程度。此方法同样需要控制两个参数 β 和 k，对于参数 β 和迭代次数 k 的选取也将在后面的章节中进行讨论。

3.4 本章小结

本章首先描述了基本的求解预测模型的流程，选取了几种常用的回归模型；然后提出了基于聚类的优化方法，说明了聚类要解决的问题和使用的聚类方法以及距离函数，并且提出了针对波形数据的动态时间规整聚类方法。之后介绍了集成学习算法的整体思路，比较了 bagging 和 boosting 的异同点，然后重点介绍了本文采用的梯度提升方法。之后，基于梯度提升算法，我们对本文的基础机器学习预测模型进行了改进，同时分析了本文的问题中可能存在的高血压和低血压间的关系，对基本的梯度提升方法进行了调整，使其同时在损失函数和高低压差值上进行优化，更加契合本文的问题。在下一章中，将通过实验将前文提到过的方法一一进行比较，并进行相应的分析，找出最适合的血压预测的方法。

第 4 章 实验与结果分析

本章将通过实验对前文提到过的几种基本方法以及优化算法进行比较验证，从中找出最适合求解本文血压预测模型的方法。

4.1 实验数据

本文选取的实验数据来自于 mimic.physionet.org[44,45]，MIMIC 是由麻省理工学院的实验室为 40000 多名患者的生理健康数据开发的一个公开可用的数据集，其主要包含了人口特征、生命体征、实验室测试、药物以及其他方面的数据。对于获取不到医疗数据的研究人员 MIMIC 数据库是很有价值的。MIMIC 数据库主要包含了两个库：临床数据库和高度解析的波形记录数据库。本文所用的数据包含了从波形数据库得到的 4000 多个文件夹，每一个文件夹代表从一个被测者身上采集到的数据，其中主要包含了 PPG 信号的脉搏波数据和标准血压值数据，PPG 信号的采样频率是 100Hz，并且每段数据持续 600s，标准血压值的采样频率未知，但同样每段持续 600s。为了全面地比较本文提出的方法，我们将所有的数据按照脉搏周期分离，每个周期的数据作为一个样本用于提取特征值，并且找出对应的标准血压值，将特征值与标准血压值一起作为训练模型的样本集。所有的样本按照所含周期数量 10000，20000，…，100000，分为 10 组，然后在不同规模的数据集上运行不同的方法，并比较不同方法的预测准确性。

4.2 实验环境及评价指标

本实验所处环境如下：

1) 64 位 Win10 操作系统
2) 4GB 内存，500G 硬盘
3) 64 位 Matlab2016a 软件环境

对比实验结果所用的指标是均方误差和相关系数。均方误差是一种方便地衡量平均误差的统计量，可以用来评价标准值与预测值之间的变化程度，如式 5-1 所示。

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(observed_i - predicted_i)^2 \tag{5-1}$$

相关系数是显示变量之间线性相关程度的统计量，根据不同的研究对象，相关系数有多种形式，皮尔逊相关系数是最常用的一种相关系数，其公式如下：

$$r(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{5-2}$$

其中的 $Cov(X,Y)$是变量 X 与变量 Y 的协方差，$Var(X)$与 $Var(Y)$是 X 和 Y 的方差。从上式可以看出相关系数的范围是 1 到-1 之间，正数表示两个变量间存在正相关，即一个变量值增大时另一个变量也同时增大，一个变量减小时另一个变量也同时减小；负数表示两个变量间存在负相关；0 表示两个变量不存在线性相关性。当两个变量间的线性关系增强时，相关系数的值就趋近于 1 或-1，反之相关系数趋近于 0。在本文的实验中，分别将标准血压值和模型预测值作为两个变量，带入到相关系数的公式中，求解两个变量的线性相关性，从而可以判断模型预测效果的好坏。

4.3 不同模型预测效果的比较

这一节首先对本文使用的 4 种机器学习回归方法的预测效果进行比较。对 10 组不同规模的样本集合，选择 80%的数据作为训练集，其余的 20%作为测试集，分别用最小二乘线性回归、套索回归、支持向量机和人工神经网络构建血压预测模型。

在进行实验之前，要先通过交叉验证和网格搜索[46,47]的方法找到各个方法所需的最优超参数。通常构建一个机器学习模型时，需要设置一些与模型相关的参数，称为超参数。这种参数与通过训练得到的参数不同，其直接决定了模型的复杂度或学习能力，而且这些参数需要训练之前预先定义，通过选择不同的值，能训练不同的模型，得到的预测效果也不同。常见的几种超参数包括，树的深度或数量、学习率、神经网络层数或节点数、聚类数等等。

为了在参数空间选取出一组最适合目标问题的超参数，通常需要通过一定的模型选择方法来完成最优化问题。选取最优的超参数，不仅能使模型在训练集上得到良好的预测准确率，还能使模型具有一定的泛化能力。常用的求解最优超参数的方法是交叉验证结合网格搜索。

交叉验证是通过分割样本集进行模型训练和验证，以求得可靠稳定模型的一种方法。通常将模型分成三部分，分别是训练集（Training set），验证集（Valid Set）和测试集（Testing Set），在使用一组超参数通过训练集得到预测模型后，先通过验证集对这一

组超参数产生的模型预测准确率进行评定，然后在一定条件下调整超参数，重新训练模型，重复以上过程，直到模型的预测准确率满足一定的限制条件，则认为选取到了较优的模型，然后可以将此模型应用到之后的预测中。交叉验证不仅考虑了训练误差而且也将泛化误差作为选择模型的指标，因此模型的泛化性更好，更不容易产生过拟合现象。交叉验证通常有几种不同的形式，本文中主要通过使用 k 折交叉验证方法选取最优模型。k 折交叉验证，初始时先将训练集分为 k 个子集，将 k-1 个集合的数据用于训练模型，剩余的一个集合作为验证，每个子集只作为验证一次。以上重复执行 k 次，得到 k 次验证的平均结果，即得到这一组超参数对应的模型的预测效果。

有了交叉验证的实现过程，那么就需要对特定范围内的参数取值进行遍历，在所有满足条件的参数组合中，选取交叉验证效果最好的一组，作为最优超参数，即“网格搜索”。网格搜索简单来说就是尝试所有的参数组合，然后进行交叉验证，从而找出最优的一组。网格搜索的方法很朴素直观，但可能效率较低，虽然也有一些较快的方法（比如启发式算法或随机算法），但多数情况下，我们还是倾向于网格搜索。网格搜索处理简单，覆盖范围最全面不会有所遗漏，网格搜索还可以对多组参数并行处理，所以在本文中通过交叉验证与网格搜索进行模型超参数的选择。

本文中主要通过 5 折交叉验证完成超参数的选取，4 种基本模型的超参数选取如表 5.1 所示。

表 5.1 4 种基本模型的超参数

方法	血压类别	超参数	取值
LASSO	*SP*	λ	0.46
	DP	λ	0.21
SVM	*SP*	*C*	64
		γ	100
	DP	*C*	64
		γ	32
ANN	*SP*	*HLs*	[15,15]
	DP	*HLs*	[15,15]

表中的 SP 和 DP 分别表示高压值和低压值，λ 是 LASSO 回归的正则项惩罚因子，本文对 λ 取值为{0.01，0.02，…，1.00}，C 是 SVM 的惩罚因子，γ 是 SVM 核函数系数，本文对 C 的取值为$\{2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$，$\gamma$ 的取值为{20,24,…,100}。HLs 表示神经网络

的隐藏层结构，本文考察了几种不同隐藏层结构的神经网络，针对这些隐藏层进行交叉验证，分析准确率与隐藏层结构之间的关系，最后选取一个较合适的隐藏层结构的神经网络作为最终的预测模型。论文中对 HLs 的取值考察了 16 种组合情况，分别如下：{[5,5], [5,15], [5,25], [15,15], [15,25], [25,25], [5,5,5], [5,5,15], [5,5,25], [5,15,15], [5,15,25], [5,25,25], [15,15,15], [15,15,25], [15,25,25], [25,25,25]}。

通过上面得到的超参数，分别用文中提出的 4 种方法进行模型的构造，四种方法的预测结果如图 5.1 和图 5.2 所示。图中的纵坐标分别表示均方误差（MSE）和相关系数（CF），横坐标表示数据集的规模大小，ANN 表示人工神经网络，SVM 表示支持向量机，LS 表示最小二乘线性回归，LASSO 套索回归。

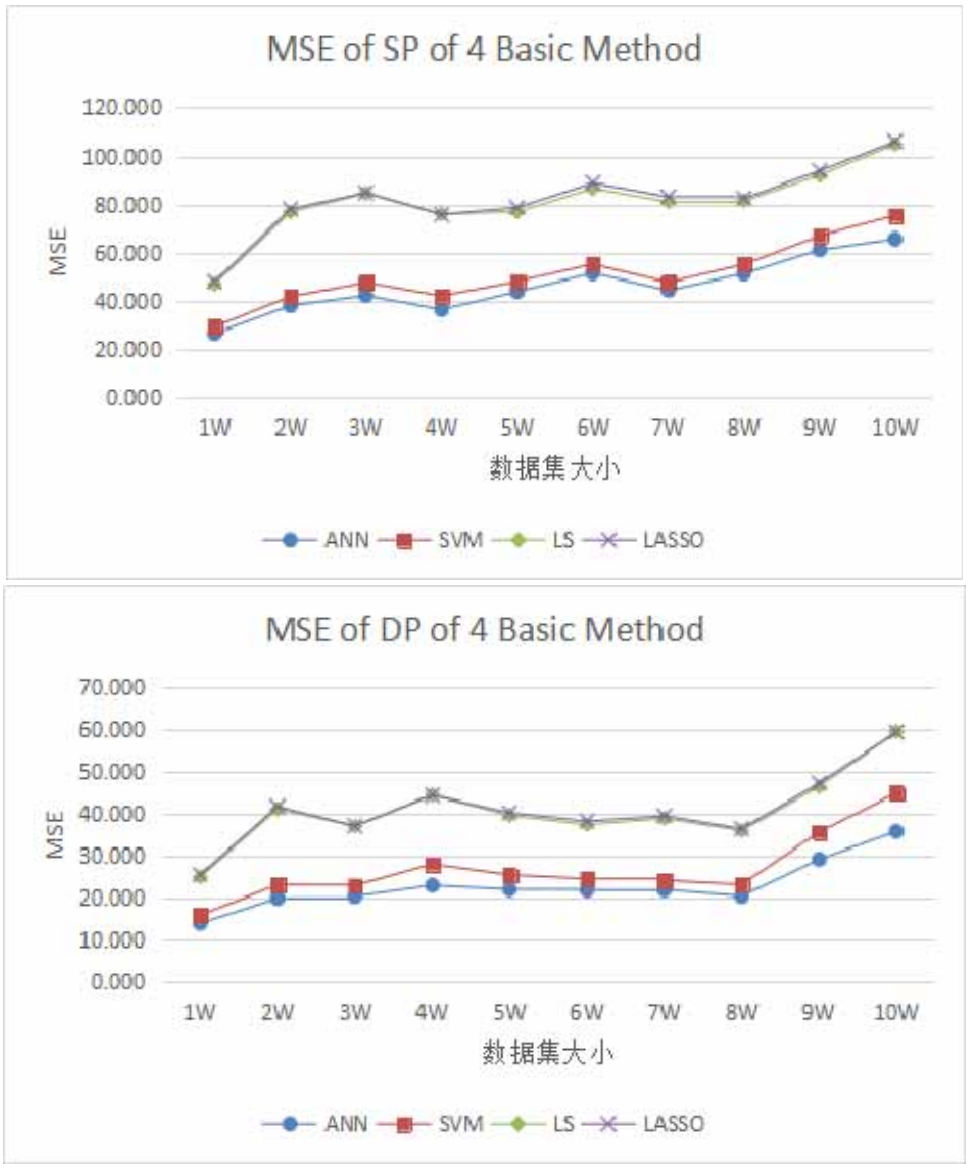


图 5.1 4 种基本模型的均方误差对比

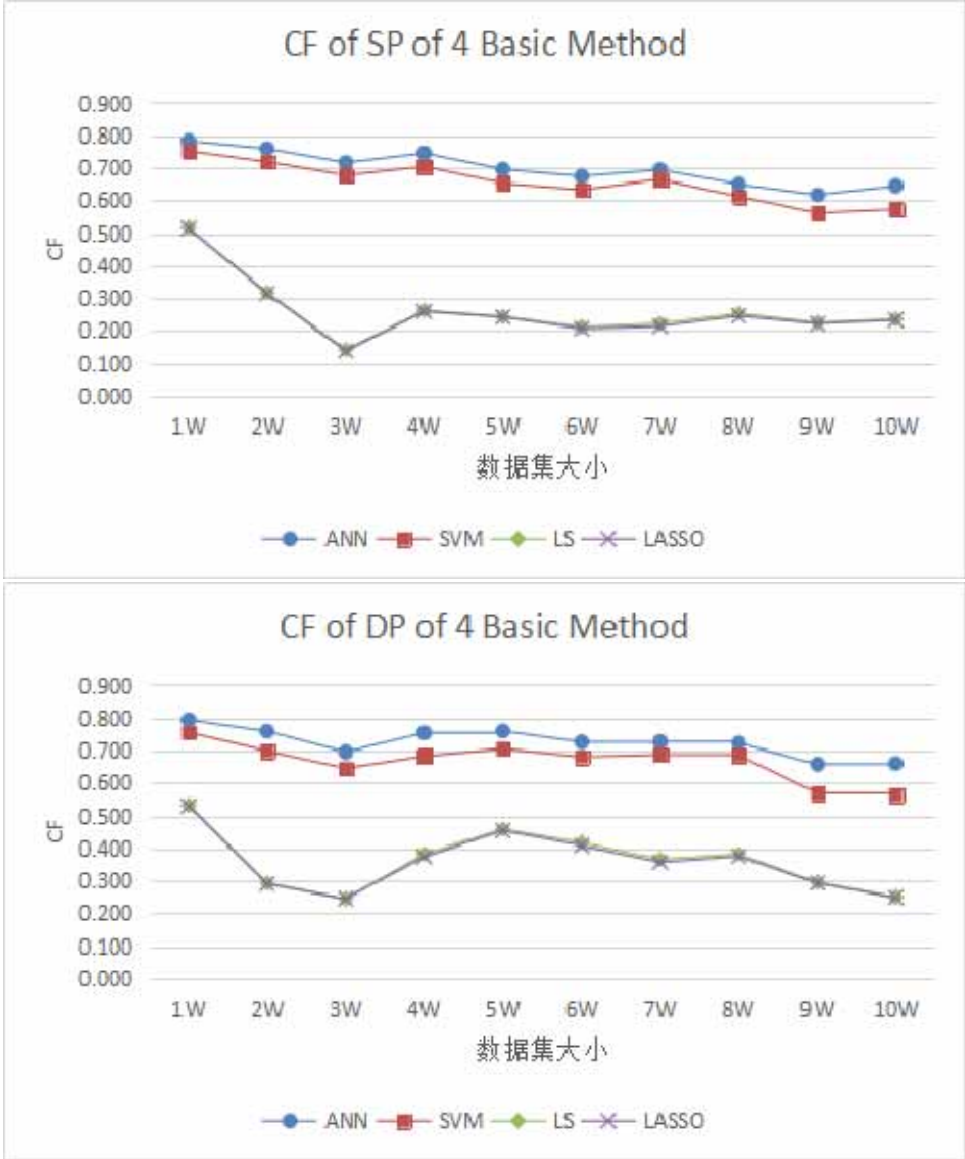


图 5.2 4 种基本模型的相关系数对比

如图所示，分别对高压值（SP）和低压值（DP）绘制了相应的图表，无论是从均方误差还是相关系数来看，支持向量机和神经网络对于血压的预测准确率比最小二乘线性回归和套索回归要好。主要原因是最小二乘线性回归和套索回归只能对线性关系进行建模，模型的适用性和泛化能力较差；而支持向量机和神经网络以提高复杂度为代价能支持更复杂的非线性关系，还可以拟合除线性关系之外的其他复杂关系的各种组合。而对于本文的问题，神经网络的效果还要优于支持向量机一些，一方面可能是因为神经网络比支持向量机更适合于求解回归问题，另一方面也可能是因为神经网络的参数更多，模型更复杂，其结构更适合于本文的血压预测问题。

在算法复杂度方面，最小二乘线性回归和套索回归这两种线性方法的时间复杂度较

低，为 O(I*n*k)，其中 I 表示迭代次数，n 表示样本数，k 表示特征数。相比之下非线性模型的时间复杂度较高，支持向量机的最坏时间复杂度为 O(k*n^2)，其中 k 为特征数，n 为样本数。人工神经网络的时间复杂度为 O(I*n*M)，其中 I 表示迭代次数，n 表示样本数，M 即为神经网络隐藏层的计算复杂度，用 O(M)表示。

4.4 优化算法的效果

这一节主要比较文中提出的优化方法对基本模型的优化效果，并给出相应的分析。

4.4.1 基于聚类的优化效果

在评价聚类优化方法的效果之前，本文先通过实验验证了不同人群之间存在的差异。为了验证此差异，本文采取了两种不同的建模策略：独立模型和统一模型。独立模型表示为每个独立的被测试者的样本集进行建模，然后测试模型效果时也分别应用对应的独立模型；统一模型表示将所有的被测试者的样本统一集中起来建立一个整体的模型。两种策略都选取 80%的数据作为训练集，剩余 20%作为测试集。不同的是独立模型分别将每一个被测试者的样本数据进行分组，统一模型先将所有数据组合成一个大集合，再对大集合分组。实验结果如表 5.2 所示。

表 5.2 统一模型与独立模型对比

	SP RMSE	Ratio	DP RMSE	Ratio
独立模型	e<5	0.709	e<5	0.918
	e<8	0.864	e<8	0.965
	e<10	0.906	e<10	0.977
统一模型	e<5	0.139	e<5	0.290
	e<8	0.263	e<8	0.469
	e<10	0.335	e<10	0.580

表中的 RMSE 表示模型预测值与标准值的均方根误差，Ratio 表示误差在一定范围内的样本所占的比例。从表中可以看出，在不同误差范围内，统一模型的样本占比都要远低于独立模型，也就说明独立模型的预测准确率优于统一模型，此结果也进一步支持

了本文的聚类优化算法。

通过以上方法可以看出不同类型人群之间是存在差异的，所以通过聚类对样本进行分组是有意义的，本文聚类优化方法的超参数选取如表 5.3 所示。

表 5.3 聚类数的选取

方法	血压类别	超参数	取值
CLS	*SP*	*k*	7
	DP	*k*	4
CLASSO	*SP*	*k*	9
	DP	*k*	4
CSVM	*SP*	*k*	2
	DP	*k*	3
CANN	*SP*	*k*	3
	DP	*k*	3

上表里的 CLS，CLASSO，CSVM，CANN 分别是将聚类优化算法应用到 4 种基本方法后的模型，主要对聚类数 *k* 进行选取，得到最优值后，分别将聚类优化方法运用到四种基本方法上，其结果如图 5.3-5.12 所示。

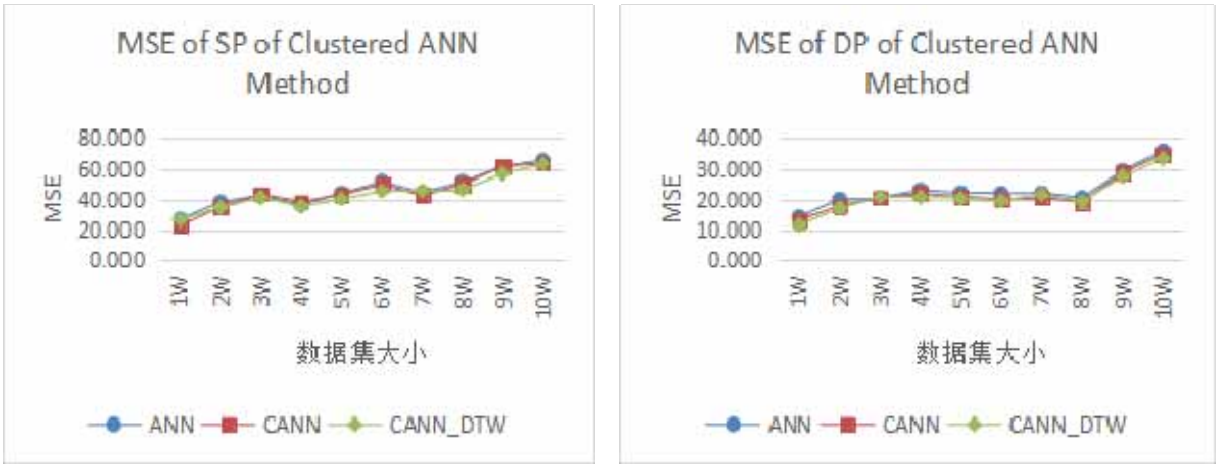
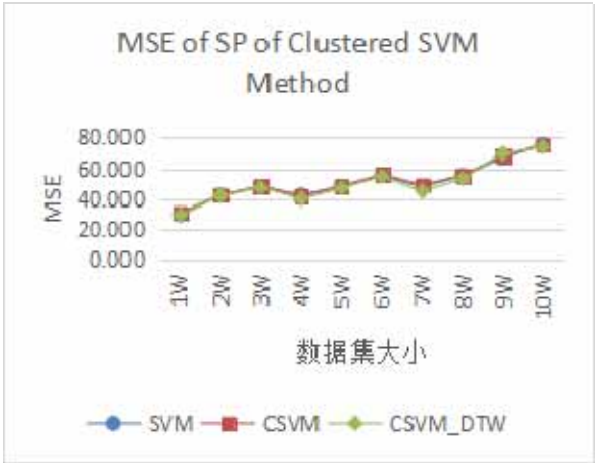


图 5.3 神经网络应用聚类优化后的均方误差对比



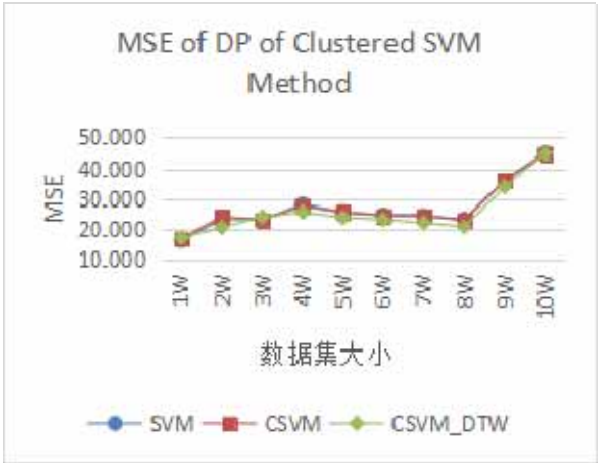
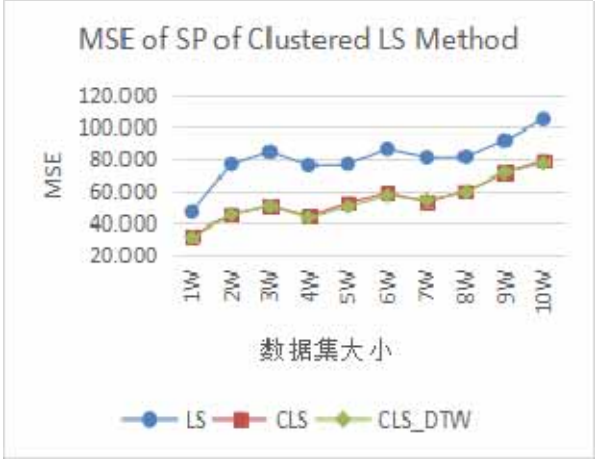


图 5.4 支持向量机应用聚类优化后的均方误差对比



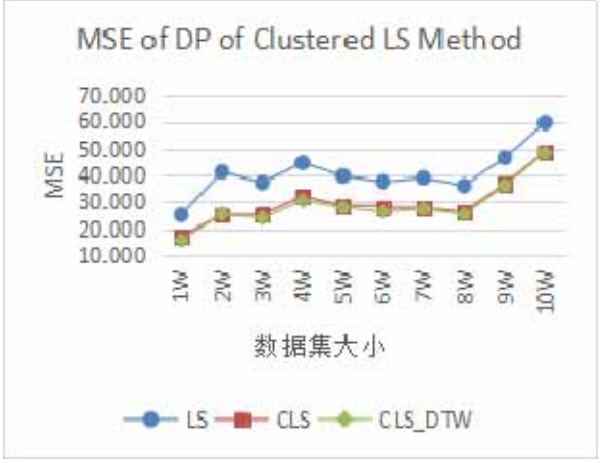
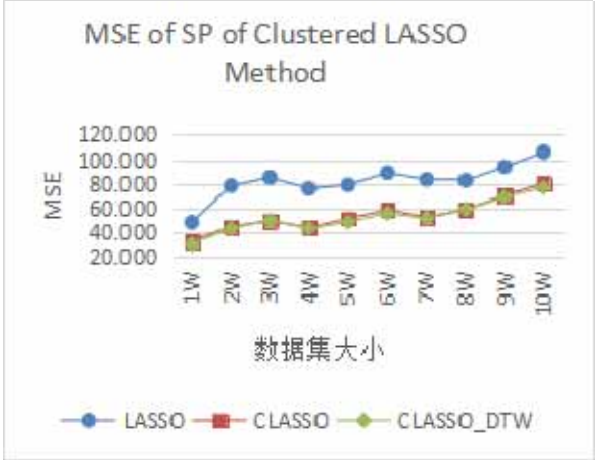


图 5.5 最小二乘线性回归应用聚类优化后的均方误差对比



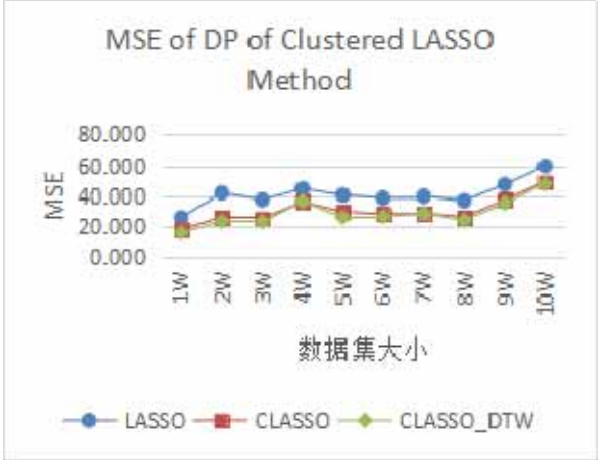


图 5.6 套索回归应用聚类优化后的均方误差对比

图例中各种方法前的“C”字头表示分别应用聚类优化算法（欧氏距离）到 4 种基

础模型上，字尾“DTW”表示采取动态时间规整距离的聚类优化算法。

从均方误差来看，聚类优化算法分别应用到 4 种基础模型时，对于支持向量机和人工神经网络的优化效果不是很明显，通过分析可知，这两种方法由于模型本身足够复杂，相关参数较多，所以从模型内部来看，已经隐性的包含了对数据的分类，所以对聚类优化算法的反馈不明显。而最小二乘线性回归和套索回归两种线性回归模型，在应用聚类优化算法时，获得的算法准确率提升比较明显，且应用动态时间规整时的效果要更优一点。对于两种线性模型来说，其本身的模型复杂度较低，对数据的分类能力相对较差，所以事先对数据进行分类，会很大程度地提升模型的准确性，而应用动态时间规整对原始的波形数据直接求距离会保留更多的原始信息，这样在执行聚类时的可考虑因素更多，所以分类后的类间更无关，类内更相似，再建模也就能获得更好的预测准确率。

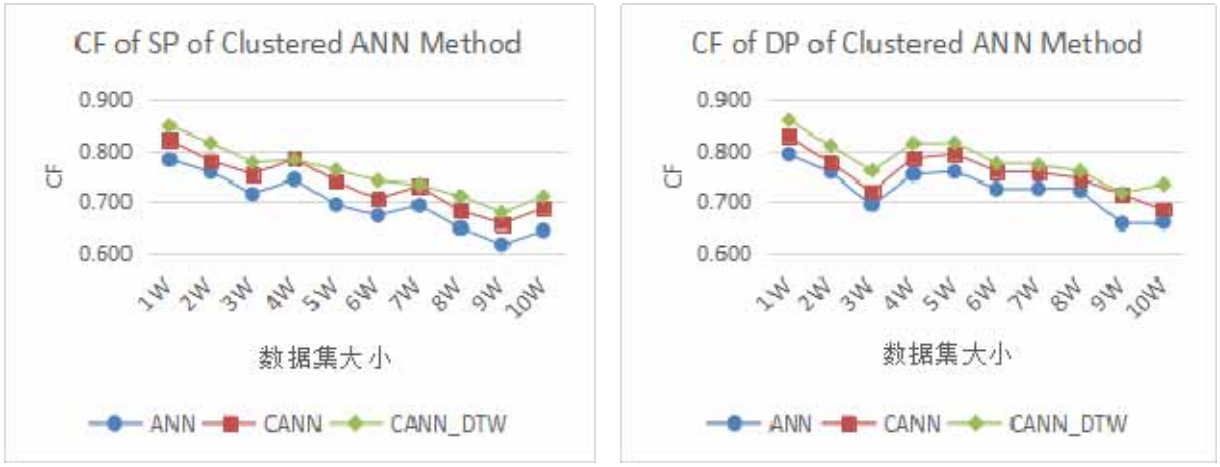


图 5.7 神经网络应用聚类优化后的相关系数对比

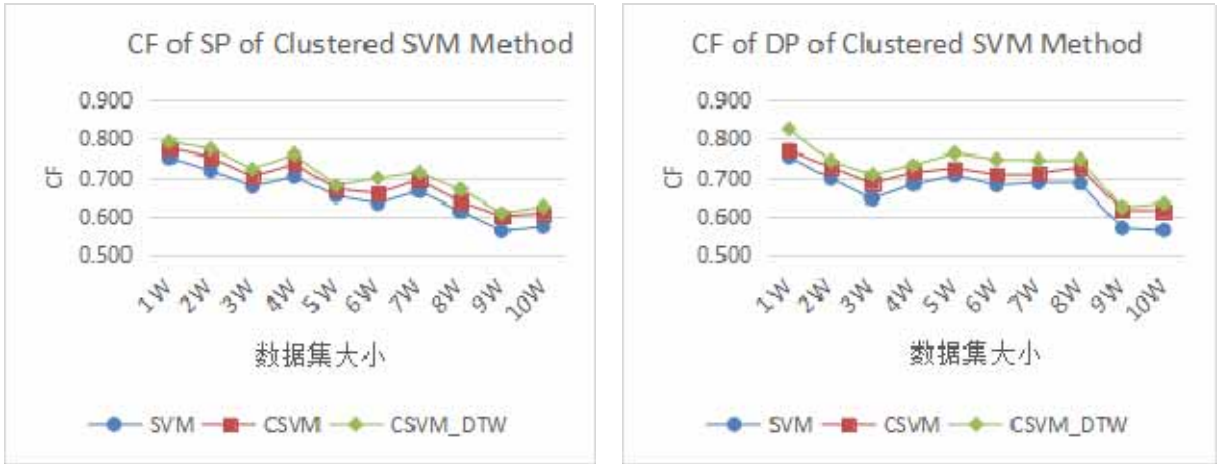
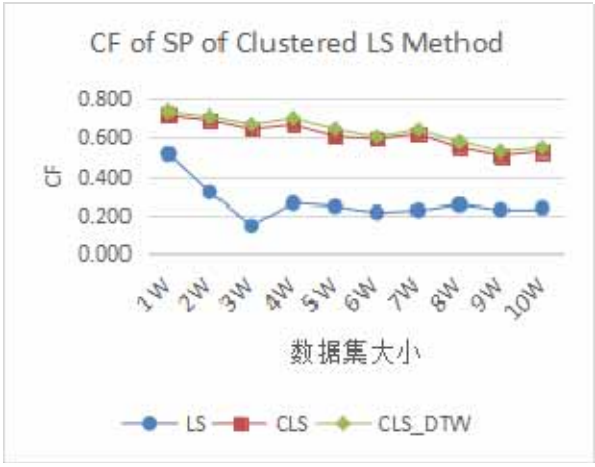


图 5.8 支持向量机应用聚类优化后的相关系数对比



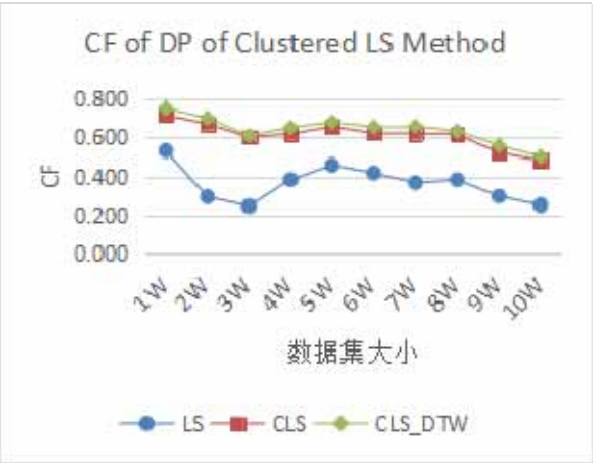
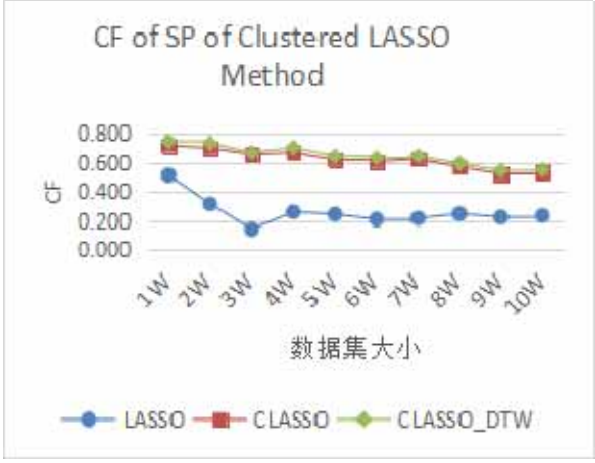


图 5.9 最小二乘线性回归应用聚类优化后的相关系数对比



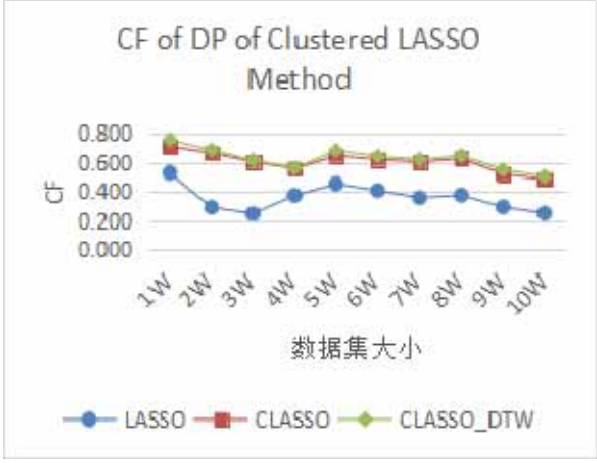


图 5.10 套索回归应用聚类优化后的相关系数对比

在计算相关系数时，对于高压或者低压，在不同大小的数据集上分别选取已知的标准血压值和模型得到的血压预测值的两组向量，令其作为两个随机变量，通过式（5-2）计算得到两者的相关系数。通过观察相关系数，可以看出模型预测值的整体变化情况与标准值是否相似，两者的相关系数越高，说明模型的预测效果越好。从相关系数来看，可以更明显地看出聚类优化算法的效果，应用聚类优化之后，几个模型得到的相关系数会更高，而相比之下也可以看出采取动态时间规整距离之后得到的相关系数更高。下面两图显示了四种方法应用聚类优化后的整体对比情况。

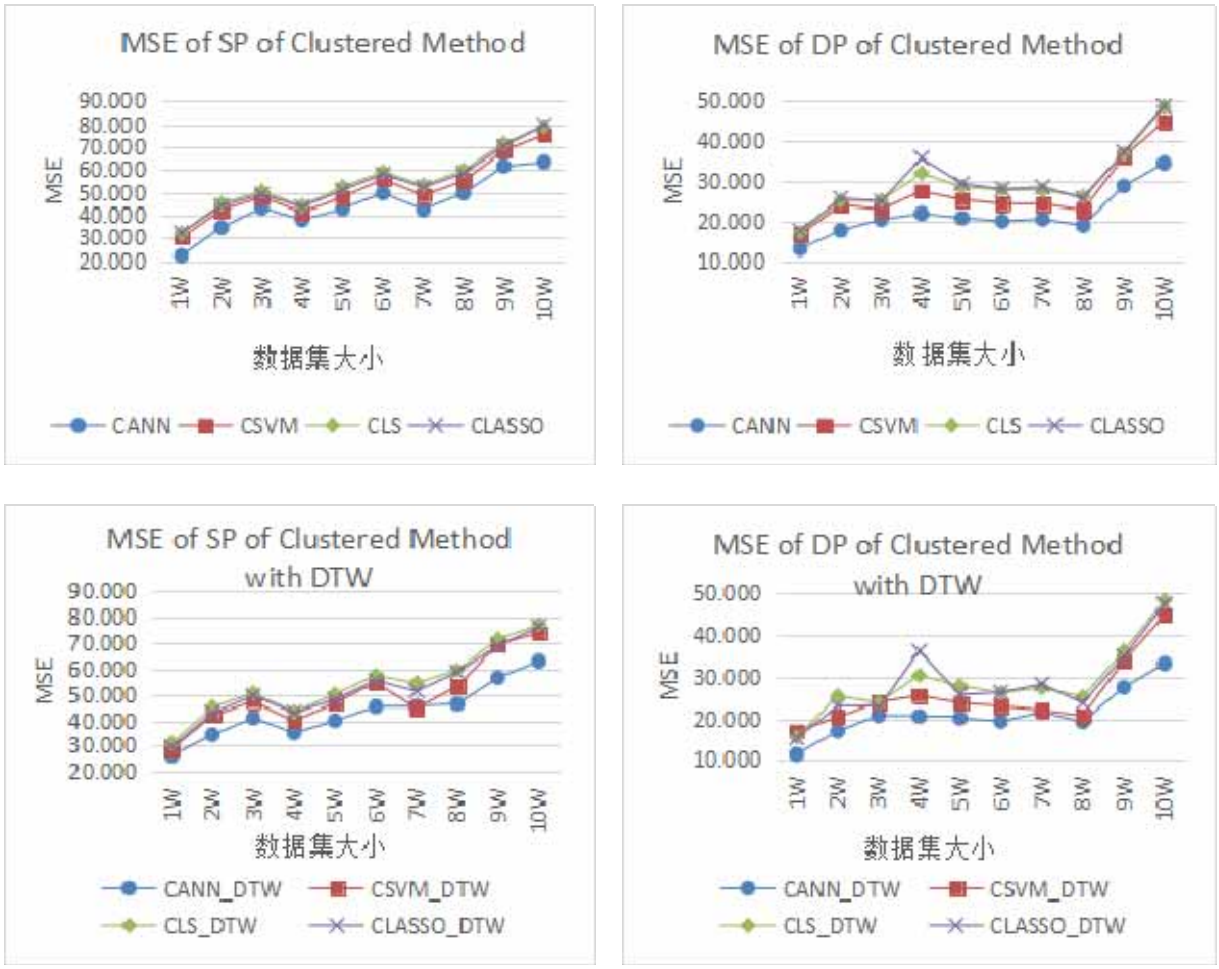


图 5.11 4 种聚类方法（欧氏距离）的均方误差对比

从图 5.11 和图 5.12 中可以看出，应用聚类优化算法后原本 4 种方法之间的差距变小了，而且将距离函数换为动态时间规整后四种模型整体的均方误差在减小，相关系数在增加。从以上的实验结果来看，聚类优化算法，可以有效地提升模型预测准确率。

虽然聚类优化算法取得了一定的效果，但同时也存在一定的不足。聚类优化方法对于本身较为复杂的模型来说优化效果有限，而且基于动态时间规整的算法时间复杂度较高。k-medoids 的时间复杂度是 O ($T*n^2*k*m$)，T 表示迭代次数，m 表示特征数，n 表示样本数，k 表示聚类数。对于动态时间规整来说，复杂度还要更高，欧氏距离的特征数只有 4，而若令 m'表示 PPG 信号每个周期内的样本点数，则动态时间规整求解距离时需要 O(m'^2)的时间复杂度，整个时间复杂度变为 O ($T*n^2*k*m'^2$)。可以看出，总体上聚类优化算法的计算是比较费时的。

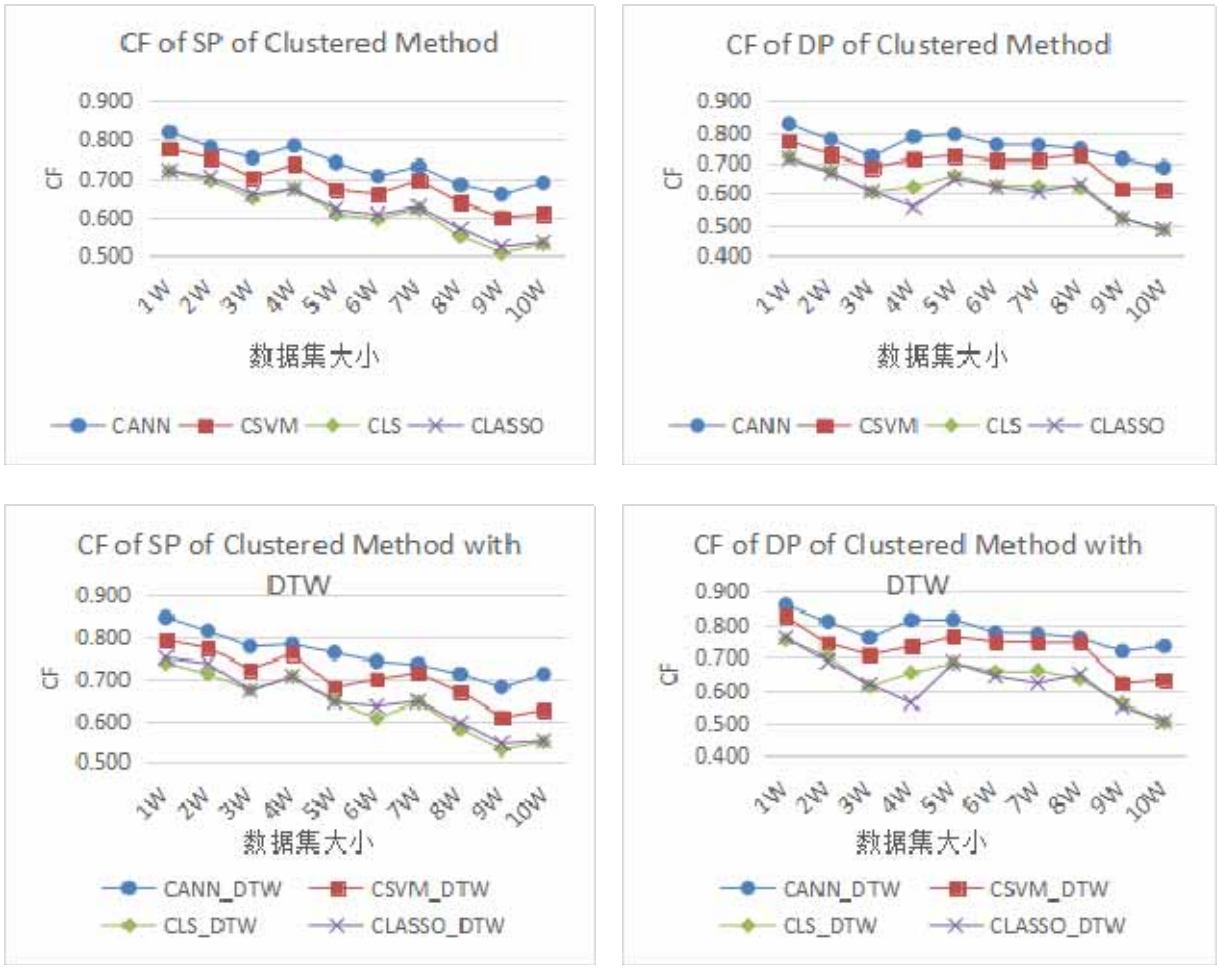


图 5.12 4 种聚类方法（动态时间规整）的相关系数对比

4.4.2 基于梯度提升的优化效果

为了比较梯度提升优化算法的效果，本文首先对梯度提升的几个超参数进行选取，其结果如表 5.4 所示。其中 GBLS1，GBLASSO1，GBSVM1，GBANN1 分别表示将第一种基于阈值的梯度提升算法应用到 4 种基本方法得到的模型，I 表示最大迭代次数，α 表示阈值控制参数，GBLS2，GBLASSO2，GBSVM2，GBANN2 分别表示将第二种基于映射的梯度提升算法应用到 4 种基本方法得到的模型，I 表示最大迭代次数，β 表示映射控制参数。

表 5.4 梯度提升算法的参数选取

方法	血压类别	超参数	取值	方法	血压类别	超参数	取值
GBLS1	*SP*	*I*	2	GBLS2	*SP*	*I*	3
		α	2-6			β	2^{-6}
	DP	*I*	2		*DP*	*I*	3
		α	2^{-5}			β	2^{-5}
GBLASSO1	*SP*	*I*	3	GBLASSO2	*SP*	*I*	3
		α	2^{-4}			β	2^{-6}
	DP	*I*	4		*DP*	*I*	3
		α	2^{-5}			β	2^{-6}
GBSVM1	*SP*	*I*	5	GBSVM2	*SP*	*I*	5
		α	2^{-1}			β	2^{-1}
	DP	*I*	4		*DP*	*I*	3
		α	2^{-1}			β	2^{-1}
GBANN1	*SP*	*I*	3	GBANN2	*SP*	*I*	3
		α	2^{-6}			β	2^{-4}
	DP	*I*	4		*DP*	*I*	3
		α	2^{-6}			β	2^{-5}

根据表中所示的超参数，分别将梯度提升方法运用到四种基本方法上，其结果如图 5.13-5.22 所示。

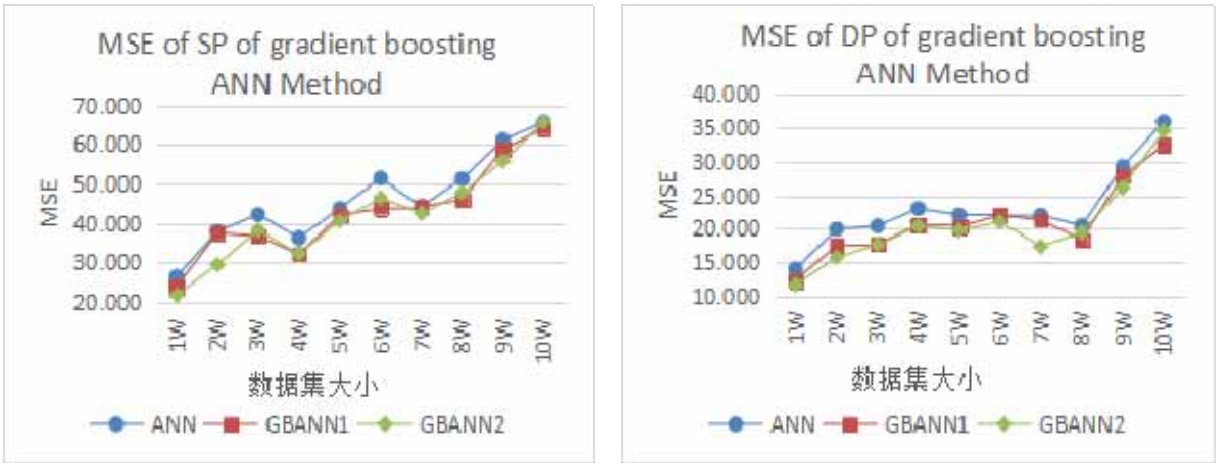
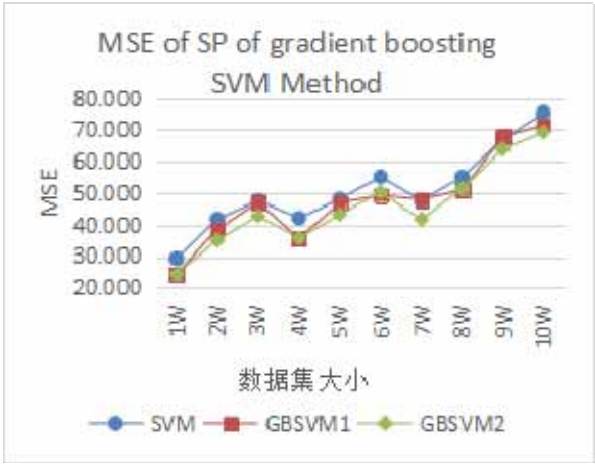


图 5.13 神经网络应用梯度提升优化后的均方误差对比



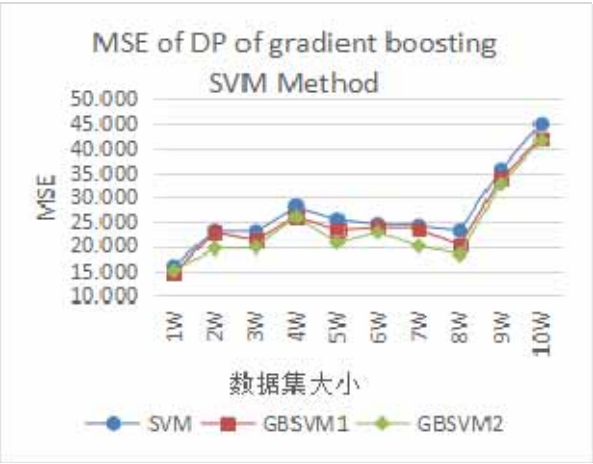
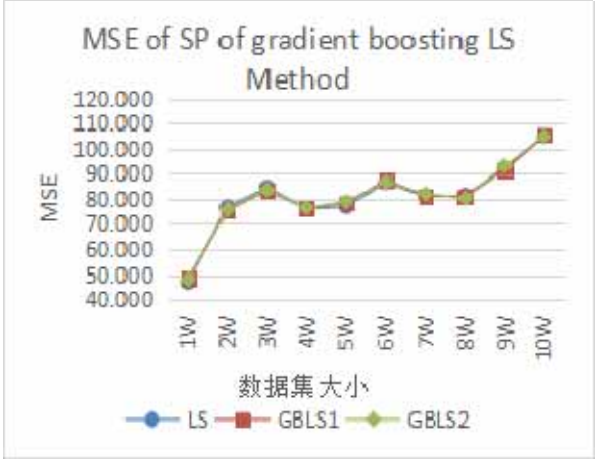


图 5.14 支持向量机应用梯度提升优化后的均方误差对比



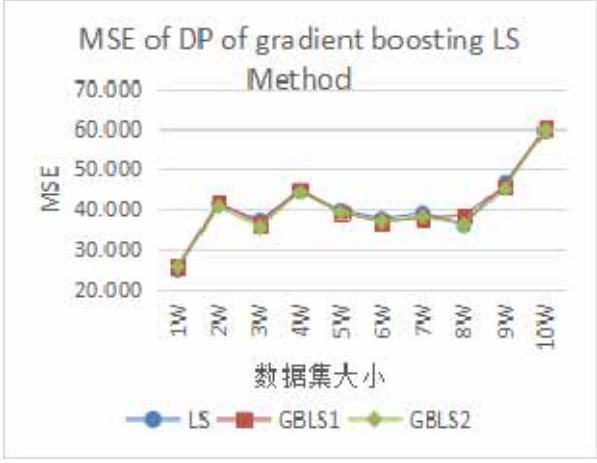
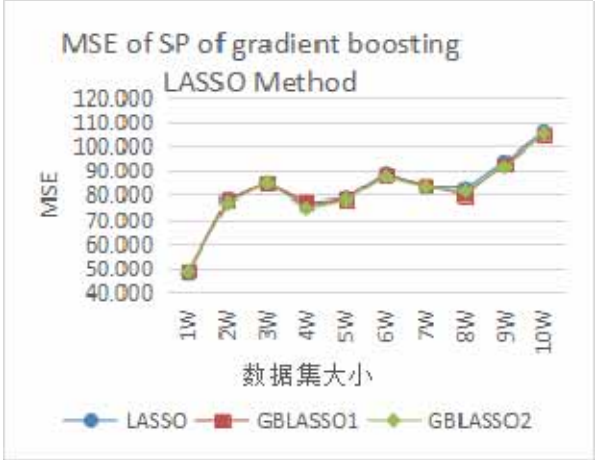


图 5.15 最小二乘线性回归应用梯度提升优化后的均方误差对比



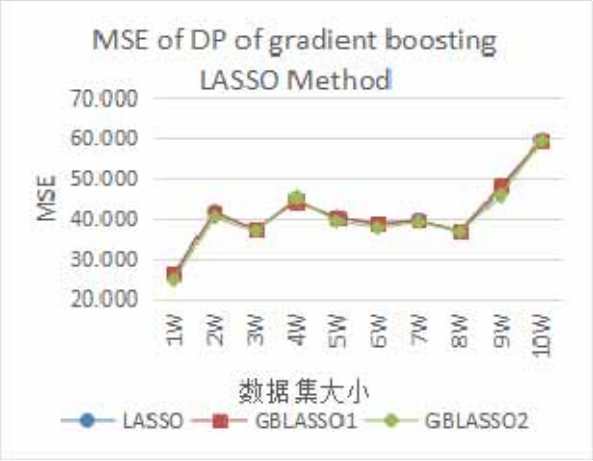


图 5.16 套索回归应用梯度提升优化后的均方误差对比

从均方误差来看，梯度提升算法可以一定程度上提高支持向量机和人工神经网络模

型的预测准确率，两种实现方式中，基于函数映射的方法相对更好。对于最小二乘线性回归和套索回归来说，梯度提升算法几乎没有效果，通过分析这两种线性模型可知，多个线性模型的线性组合仍然是线性模型，所以单纯地通过将梯度提升算法应用到线性模型上时，模型的本质还是线性，所以模型预测效果很难有显著的提高。

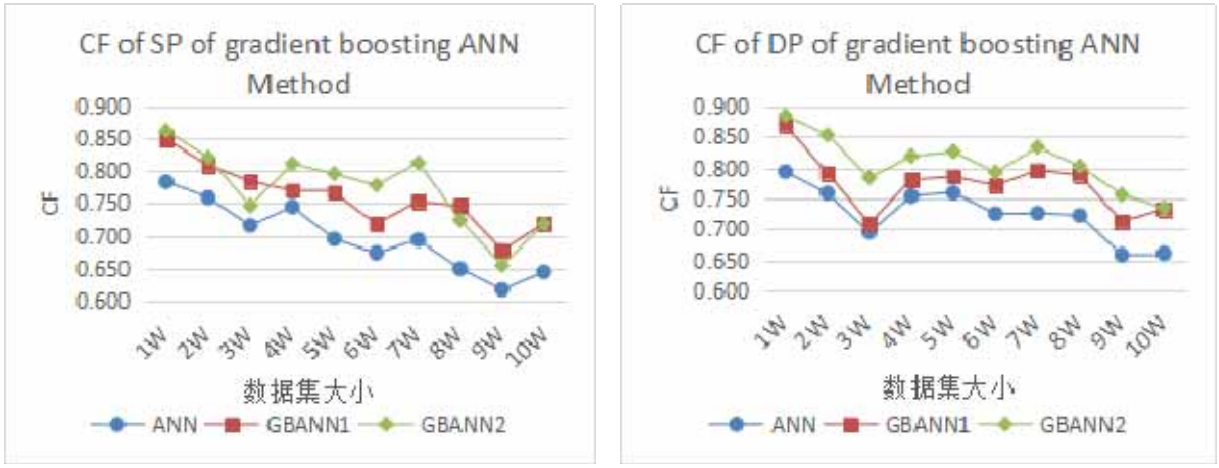


图 5.17 神经网络应用梯度提升优化后的相关系数对比

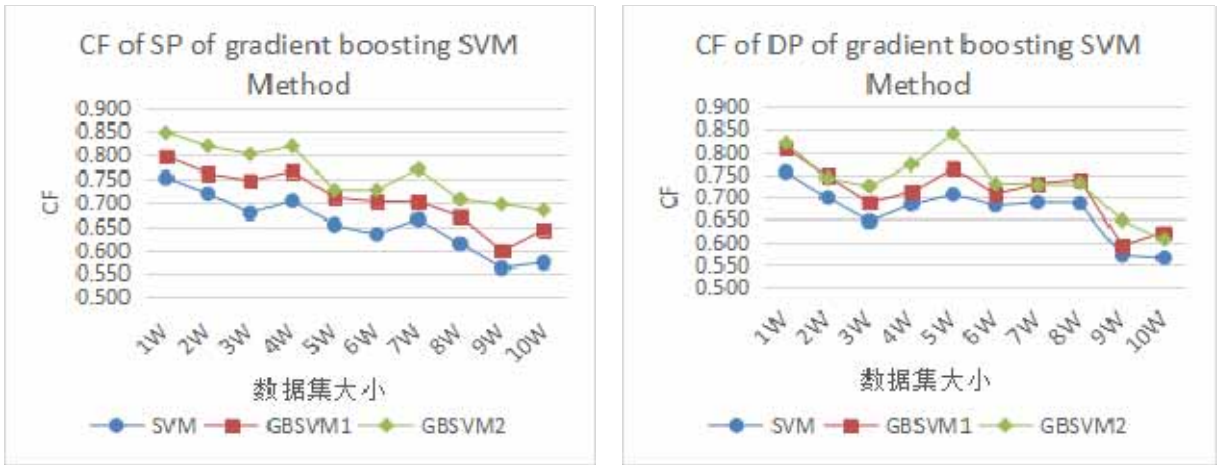
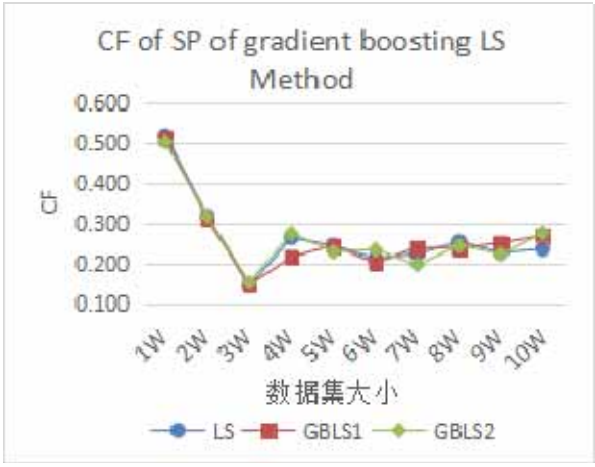


图 5.18 支持向量机应用梯度提升优化后的相关系数对比



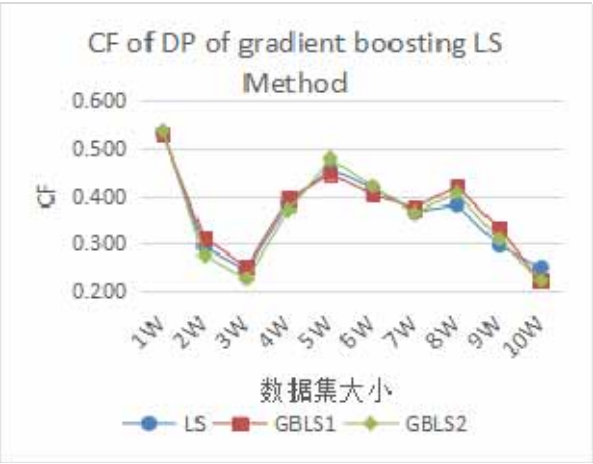
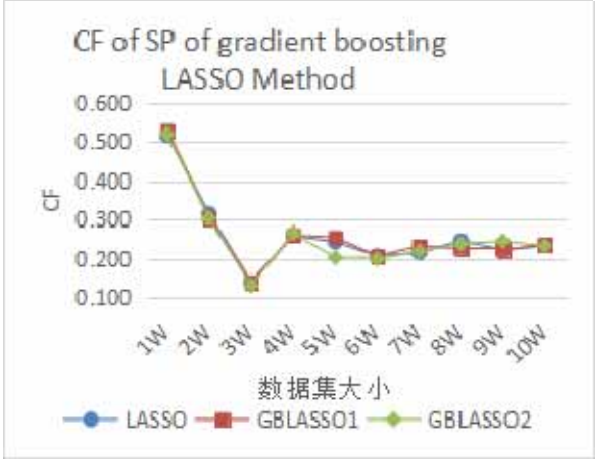


图 5.19 最小二乘线性回归应用梯度提升优化后的相关系数对比



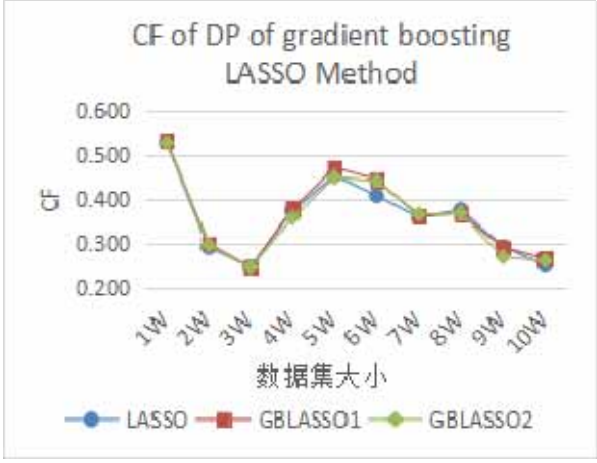


图 5.20 套索回归应用梯度提升优化后的相关系数对比

相关系数显示的结果相比均方误差会更加明显，两种形式的梯度提升优化算法可以对非线性的模型的预测效果有所提高，而且基于函数映射的方式更优。对于线性模型来说，梯度提升算法不会对模型有明显的改善，理由如上文所述。

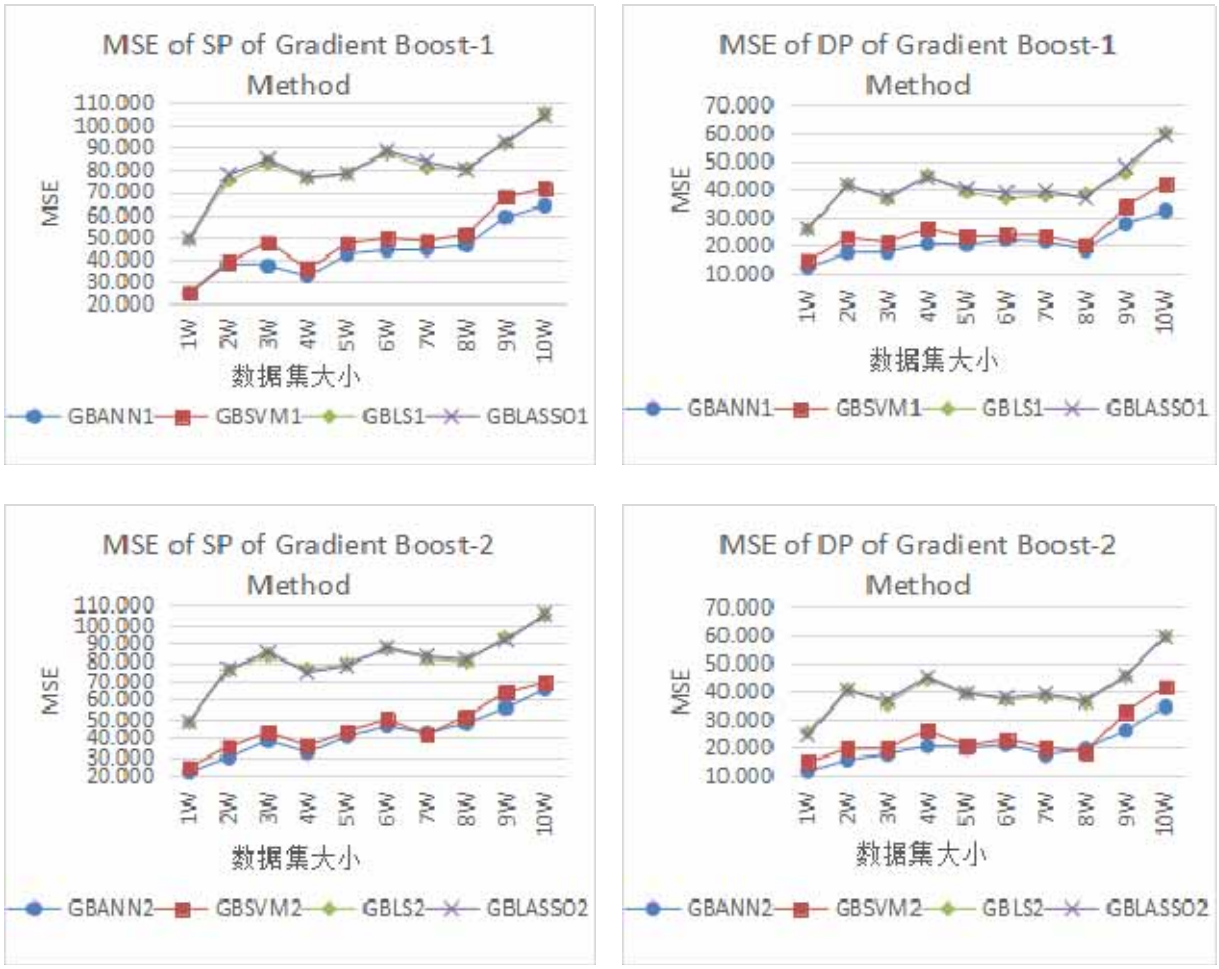


图 5.21 4 种梯度提升方法（基于阈值）的均方误差对比

最后，从两种梯度提升优化算法应用到 4 种基本模型的整个对比结果图 5.21 和图 5.22 中，可以看到两种非线性模型的预测效果仍然明显地优于两种线性模型，两种优化后的非线性模型对比与优化之前，整体来看预测准确率有所提高，这也说明了本文提出的基于梯度提升的优化方法具有一定的优化效果。

在时间复杂度上，基于梯度提升的优化算法要比基于聚类的优化算法快很多，在每次迭代过程中所需的时间复杂度为 O(n)，其中 n 表示样本数量，每次迭代只需要计算算法所需要的残差，预测值的高低压差值等量，所以时间复杂度只与样本数量成正比。梯度提升算法的主要问题是，对于线性模型存在一定的局限性，累加形式的线性模型其得到的结果仍然局限在线性模型，所以模型的本质并没有发生改变，所以预测效果会限制在线性模型带来的预测上限之内，也因此本文的梯度提升算法很难对线性模型的效果有明显的改进。

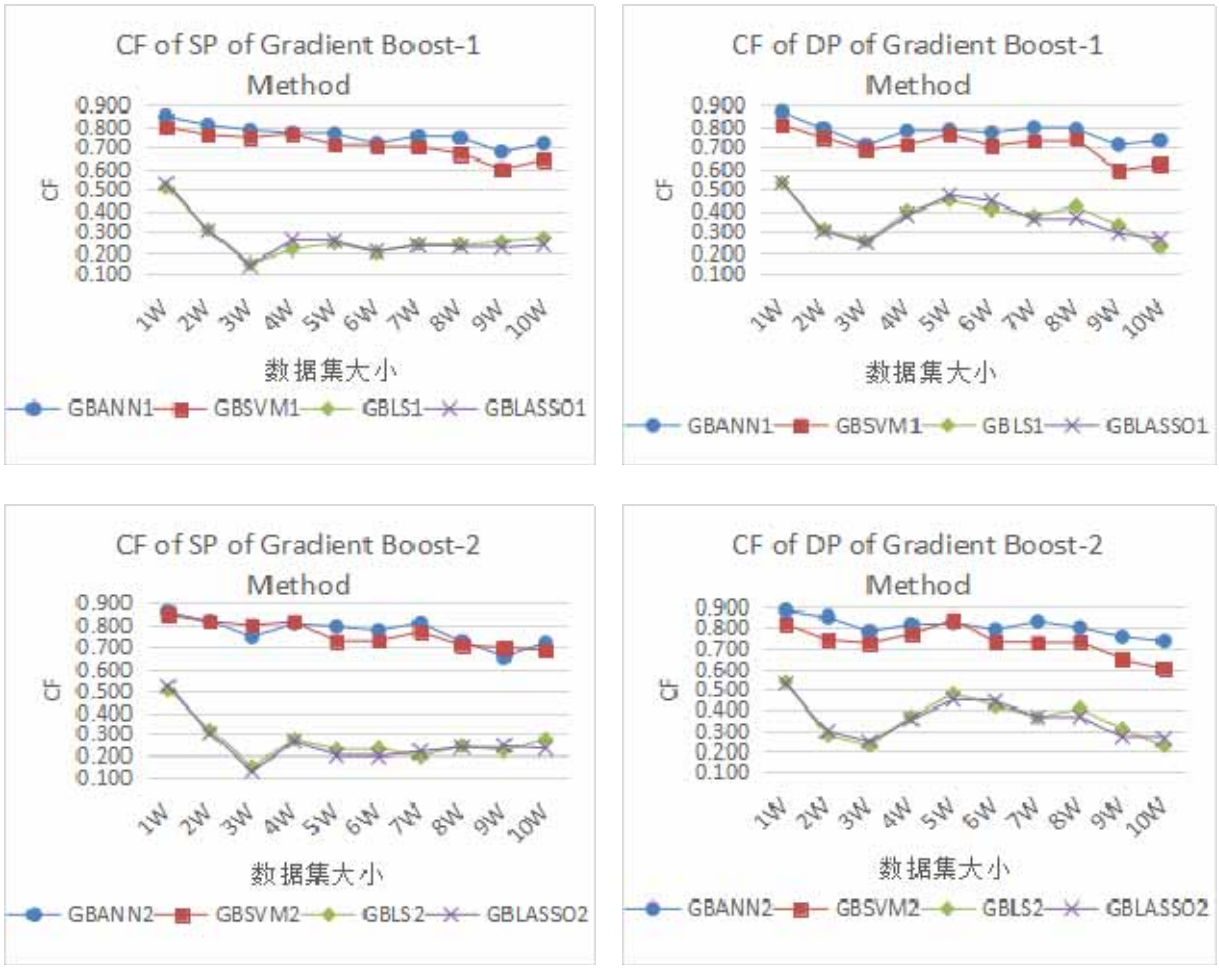


图 5.22 4 种梯度提升方法（基于映射）的相关系数对比

4.5 本章小结

在本章首先通过网格搜索和交叉验证方法对每种基本方法和优化算法的超参数进行了选取，然后通过实验验证了前文提出的几种求解回归的机器学习方法，对比了几种方法的预测效果，之后对本文提出的聚类优化方法和梯度提升方法也进行了验证，两种方法都一定程度上对原有方法的预测准确性进行了提升，比较之下，选择动态时间规整的聚类优化算法和选择基于函数映射的梯度提升算法较优，最后总结并分析了两种方法的优点和不足。

结　论

本论文首先介绍了血压的重要性以及血压测量方面的相关研究，分析了现有的大部分血压测量方法，主要包括了柯氏音法，动脉张力测定法和容积补偿法。这些方法都在一定程度上存在不连续，有创，不方便等问题。为解决现有问题，本文利用从 PPG 传感器采集到的脉搏波信号，通过机器学习的方法建立模型，实现血压的预测。本文的主要实现的工作有：

首先完成了对基本机器学习模型的构建。首先选取了 4 个与血压相关性最高的特征值，通过快速傅里叶变换，取极值点，去异常点，归一化等过程，得到特征值集合和标准血压集合。然后选择了最小二乘线性回归，套索回归，支持向量机和人工神经网络 4 种模型对通过 PPG 脉搏信号预测血压的可行性进行了验证。根据实验可知，人工神经网络效果最优，支持向量机次之，其余两种方法效果稍差。

然后，本文提出了两种不同的优化方法，对基础模型进行改进。第一种方法通过聚类对初始样本进行划分，目的是为了降低不同属性的人群样本数据存在差异带来的预测误差。选择了 k-medoids 作为聚类方法，对距离的选取分别采取了应用于特征值的欧氏距离和应用于原始波形样本的动态时间规整距离。另一种优化方式是梯度提升算法。本文在原始梯度提升算法上，根据高低血压存在关联的特点，进行了改进，分别通过设定阈值和映射函数，决定了梯度提升每一轮残差的选取。

最后通过实验，对本文提出的优化方法效果进行了验证。选取了均方误差和相关系数两个评价指标。实验结果表明，聚类优化方法对于两种线性模型的提升效果比较明显，对另外两种非线性模型的提升效果稍弱。两种距离函数相比，采用动态时间规整的效果略优于欧氏距离。梯度提升优化方法对于每种模型的预测准确率都有一定程度的提高，其中通过函数映射的方式效果更优。

本论文同时也存在一些不足之处。首先论文选取的数据多样性不够丰富，数据大部分来自于医院病房的病人，所以对于健康人员的数据收集不足，存在一定的局限性，另外所有数据都只包含了 PPG 信号和标准血压值，没有关于被测者其他方面的一些个人信息，这也对结果的验证造成了一定影响。另一点不足是，对于动态时间规整距离的求解，其时间复杂度较高，与每个周期内的采样点数的二次幂成正比，因此复杂度与直接通过对特征值进行距离求解的欧氏距离相比会有提高，整体基于动态时间规整的优化算

法实验也更加耗时。

对于本文的方法而言，也有一些可以继续优化和改进的地方。第一，除了本文所选的 4 种方法之外，还有许多机器学习算法可以进行回归问题的求解，比如回归树模型，k 近邻模型等，对多种模型进行建模和实验验证，得到的结果也更加具有适用性。第二，对于特征的选取，可以通过特征工程的一些方法，进行选优，更合适的特征，可能会得到更准确的预测模型。第三，对于本文方法的某些步骤，可以通过采取并行化计算，提高算法执行效率，使算法更加实用，第四，对于 PPG 传感器的硬件方面，也可以进行一定的优化，使得初始得到的数据更加准确。

参考文献

[1] 严爵基. 血压监测的发展历程. 中外医学研究. 2016, 14(19):160-162 页

[2] Vasan R S, Larson M G, Leip E P, et al. Impact of high-normal blood pressure on the risk of cardiovascular disease. New England Journal of Medicine, 2001, 345(18):1291

[3] 齐颁扬. 医学仪器：上册. 第一版. 北京：高等教育出版社, 1990:136-145 页

[4] Drzewiecki G M, Melbin J, Noordergraaf A. The Korotkoff sound. Annals of Biomedical Engineering, 1989, 17(4):325-359

[5] Sapinski A. Standard algorithm for blood pressure measurement by sphygmo-oscillographic method. Medical & Biological Engineering & Computing, 1996, 34(1):82-3

[6] Pressman G L, Newgard P M. A TRANSDUCER FOR THE CONTINUOUS EXTERNAL MEASUREMENT OF ARTERIAL BLOOD PRESSURE. IEEE Transactions on Biomedical Engineering, 1963, 10(2):73-81

[7] Sato T, Nishinaga M, Kawamoto A, et al. Accuracy of a continuous blood pressure monitor based on arterial tonometry. Hypertension, 1993, 21(1):866-74

[8] Penaz J. Photoelectric measurement of blood pressure, volume and flow in the finger. Digest of the 10th international conference on medical and biological engineering, 1973:104

[9] 高树枚, 宋义林, 田中志信, 等. 基于容积补偿法的手腕式血压连续检测系统. 中国医疗器械杂志. 2009, 33(5):323-327 页

[10] Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 2007, 28(3):R1-39

[11] 贺礼荣. 脉搏波速测定方法及临床意义. 内科, 2009, 4(5):766-768 页

[12] Yamashina A, Tomiyama H, Arai T, et al. Brachial-ankle pulse wave velocity as a marker of atherosclerotic vascular damage and cardiovascular risk. Hypertension Research Official Journal of the Japanese Society of Hypertension, 2003, 26(8): 615

[13] Millasseau S C, Ritter J M, Takazawa K, et al. Contour analysis of the photoplethysmographic pulse measured at the finger. Journal of Hypertension, 2006,

24(8): 1449-1456

[14] Yoon Y Z, Yoon G W. Nonconstrained Blood Pressure Measurement by Photoplethysmography. Journal of the Optical Society of Korea, 2006, 10(2): 91-95

[15] Elgendi M. On the analysis of fingertip photoplethysmogram signals. Current Cardiology Reviews, 2012, 8(1): 14-25

[16] 李章俊, 王成, 朱浩等. 基于光电容积脉搏波描记法的无创连续血压测量. 中国生物医学工程学报. 2012, 31(4):607-614 页

[17] Yamashina A, Tomiyama H, Takeda K, et al. Validity, reproducibility, and clinical significance of noninvasive brachial-ankle pulse wave velocity measurement. Hypertension Research Official Journal of the Japanese Society of Hypertension, 2002, 25(3): 359-364

[18] Cattivelli F S, Garudadri H. Noninvasive Cuffless Estimation of Blood Pressure from Pulse Arrival Time and Heart Rate with Adaptive Calibration. International Workshop on Wearable and Implantable Body Sensor Networks, Berkeley, CA, USA, 2009. IEEE, 2009: 114-119

[19] Bhavirisetty R T. Calculation of blood pulse transit time from PPG. Rourkela: National Institute of Technology. 2012

[20] Yan Y S, Zhang Y T. Noninvasive estimation of blood pressure using photoplethysmographic signals in the period domain. International Conference of the IEEE Engineering in Medicine & Biology Society, Shanghai, China, 2005. IEEE, 2005: 3583-3584

[21] Solà J. Continuous non-invasive blood pressure estimation. Universitat Politècnica de Catalunya. BarcelonaTech: Diss.dgenössische Technische Hochschule Eth Zürich Nr, 2011

[22] Kachuee M, Kiani M M, Mohammadzade H, et al. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. IEEE International Symposium on Circuits and Systems, Mountain View, CA, USA, 2015. IEEE, 2015: 1006-1009

[23] 张俊利, 蔺嫦燕, 杨琳等. 脉搏波波形特征信息检测及与部分血流动力学变化相关分

析. 生物医学工程与临床. 2008, 12(2):104-107 页

[24] 吕海姣, 严壮志, 陆维嘉. 一种基于脉搏波的无创连续血压测量方法. 中国医疗器械杂志. 2011, 35(3):169-173 页

[25] Samria R, Jain R, Jha A, et al. Noninvasive cuff'less estimation of blood pressure using Photoplethysmography without electrocardiograph measurement. Region 10 Symposium. Kuala Lumpur, Malaysia, 2014. IEEE, 2014: 254-257

[26] Kurylyak Y, Lamonaca F, Grimaldi D. A Neural Network-based method for continuous blood pressure estimation from a PPG signal. Instrumentation and Measurement Technology Conference. Minneapolis, Minnesota, USA, 2013. IEEE, 2013:280-283

[27] Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 2007, 28(3):R1-39

[28] 周一峰, 刘超英, 黄虎等. 基于光电容积脉搏波描记法的反射型 PPG 信号传感器的设计. 电子世界. 2016(12):161-161 页

[29] Shelley K H, Shelley S. Pulse Oximeter Waveform: Photoelectric Plethysmography. Clinical Monitoring Practical Applications for Anesthesia & Critical, 2001:420-423

[30] Karlsson A. Introduction to Linear Regression Analysis. Technometrics, 2007, 170(3):856-857

[31] Motulsky H J, Ransnas L A. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. Faseb Journal Official Publication of the Federation of American Societies for Experimental Biology, 1987, 1(5):365-74

[32] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 2015, 71:804-818

[33] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification. Lecture Notes in Computer Science, 2003, 2888:986-996

[34] Salcedo‐Sanz S, Rojo‐Álvarez J L, Martínez‐Ramón M, et al. Support vector machines in engineering: an overview. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2014, 4(3):234-267

[35] Paliwal M, Kumar U A. Review: Neural networks and statistical techniques: A review of applications. Expert Systems with Applications, 2009, 36(1):2-17

[36] Hecht-Nielsen R. Theory of the backpropagation neural network. Neural Networks for Perception. 1992, 1(1):593-605

[37] Miao Y, Zhang Z, Meng L, et al. A Cluster Method to Noninvasive Continuous Blood Pressure Measurement Using PPG. Smart Health - International Conference. Haikou, China, 2016. Springer, 2016:109-120

[38] Yi B K, Jagadish H V, Faloutsos C. Efficient retrieval of similar time sequences under time warping. International Conference on Data Engineering, Orlando, Florida, USA, 1998. Proceedings. IEEE, 1998:201-208.

[39] Dietterich T G. Ensemble Methods in Machine Learning. 2000, 1857(1):1-15

[40] Elder J, Elder J. From trees to forests and rule sets: a unified overview of ensemble methods. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 2007. ACM, 2007:6

[41] Petunin Y. Justification of the three-sigma rule for unimodal distributions. Theory of Probability & Mathematical Statistics. 1980

[42] Friedrich Pukelsheim. The Three Sigma Rule. American Statistician. 1994, 48(2):88-91

[43] Huxley, Julian Sorell. Problems of relative growth. Problems of relative growth. Methuen & Co. 1932:893.

[44] Clifford G D, Scott D J, Villarroel M, et al. User guide and documentation for the MIMIC II database. Mimic. 2009

[45] 11.Lee J, Scott D J, Villarroel M, et al. Open-access MIMIC-II database for intensive care research. In International Conference of the IEEE Engineering in Medicine & Biology Society. 2011:8315-8

[46] Huang Q, Mao J, Liu Y. An improved grid search algorithm of SVR parameters optimization. In IEEE International Conference on Communication Technology. 2013:1022-1026

[47] Ji Changming, Zhou Ting, Xiang Tengfei et al. Application of support vector machine based on grid search and cross validation in implicit stochastic dispatch of cascaded

hydropower stations. Electric Power Automation Equipment. 2014, 34(3): 125-131

攻读硕士学位期间发表的论文和取得的科研成果

Miao Y, Zhang Z, Meng L, et al. A Cluster Method to Noninvasive Continuous Blood Pressure Measurement Using PPG. Smart Health - International Conference. Haikou, China, 2016. Springer LNCS, 2016, 10219:109-120

致　　谢

在哈尔滨工程大学度过了两年多的研究生时光，如今已临近毕业，回首这两年的生活，我对专业知识上的认识变得更加深刻，思想变得更加成熟，视野也变得更加开阔，在整个校园生活中得到了很多老师和同学的帮助。在如今论文工作接近尾声之时，我要感谢身边所有鼓励我，帮助我的人。

首先最应该感谢的是我的指导教师，张志强老师。在我的研究生生涯中，张老师一直在督促我们要多学习，多思考，要紧跟最新的研究方向，如此的求学态度也一直在影响着我们每一位学生。每周一次的讨论会，也时刻告诫自己要不断学习，不断钻研，学术不能流于表面，要真正搞懂每一处细节，这样才能真正掌握其本质。除了对学术的追求之外，张老师也不断提醒我们要学会表达，再优秀的学术水平没有较好的的表达能力也很难让别人接受自己的思想，所以表达能力也是至关重要的。再次对于张老师的学术精神和人生态度给我带来的积极影响表示感谢。

对于同在一个实验室的另外三位老师，潘海为老师，谢晓芹老师和韩启龙老师，他们同样是十分优秀的科研工作者，在各自的研究领域也取得了优秀的成绩，对于各位老师在平日在给予我的关照我也十分感谢。实验室的每一位师兄师姐师弟师妹，我们也一起度过了至少一年的时光，在这不短的时间之内，大家不仅在学习上给我提供了很多的帮助，而且在其他方面也一直在照顾和鼓励我。在学习之余，大家一起聚餐，游戏也为我的研究生生活增添了许多的乐趣，在此，对大家表示衷心的感谢。

此外，还要感谢在嘉兴实习过程中每位同事的帮助，虽然你们各位只比我稍长几岁，但是在很多方面都有我需要学习的地方，在工作中与你们的沟通和交流，对我的论文完成提供了很大的帮助。

最后，还要向百忙之中抽空评阅本人论文的所有评审老师表示谢意，希望各位老师能够给出宝贵意见。