**学号：2015200795**

**北京化工大学**

**硕士研究生学位论文开题报告**

**论文题目：基于集成分类器的油气层识别方法研究 及应用**

**学 院 名 称：信息科学与技术学院**

**专 业：计算机科学与技术**

**研究生姓名： 赵姗姗**

**导师姓名：耿志强**

**开题日期：2016年9月26日**

|  |  |  |
| --- | --- | --- |
| **考核**  **成绩** |  | |
| **审核**  **小组**  **成员**  **以及**  **职称** | **姓 名** | **职 称** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

目录

[一．课题来源及项目名称 1](#_Toc6804)

[二．文献综述 1](#_Toc3030)

[2.1 研究背景 1](#_Toc28435)

[2.2 自组织映射算法 2](#_Toc29060)

[2.3 分类方法 2](#_Toc12838)

[2.3.1 支持向量机 3](#_Toc29453)

[2.3.2 Fisher判别式分析 4](#_Toc22502)

[2.3.3 K最近邻分类算法 5](#_Toc25312)

[2.4 参考文献 5](#_Toc17619)

[三．研究计划 7](#_Toc18441)

[3.1 选题的目的及意义 7](#_Toc28204)

[3.2 主要研究内容 7](#_Toc11168)

[3.3 研究方案 8](#_Toc26138)

[3.3.1 技术方案 8](#_Toc5848)

[3.3.2 实施方案所需要的条件 8](#_Toc31240)

[3.3.3 拟解决的关键问题 8](#_Toc28266)

[3.4 课题难点分析 9](#_Toc31965)

[3.5 预期研究成果及创新点 9](#_Toc32364)

[3.6 工作计划进度 9](#_Toc25592)

# 一．课题来源及项目名称

课题来源：自选题目

项目名称：基于集成分类器的油气层识别方法研究及应用

# 二．文献综述

## 2.1 研究背景

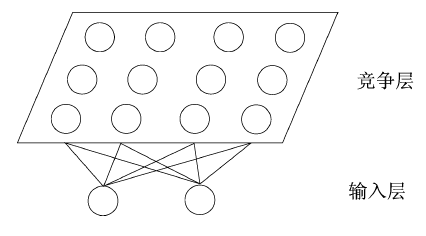
录井技术[1]是油气勘探[2]开发活动中最基本的技术，是发现、评估油气藏最及时、最直接的手段，综合解释评价技术以地化录井、轻烃录井、气测录井、核磁录井等为主要技术手段，通过对各项录井资料的综合分析，实现对油气水层的准确评价。利用录井资料正确识别油气层在油气藏的勘探与开发中具有重要的现实意义, 它将直接影响油气勘探的效率与成功率。

地球化学录井技术（简称地化录井）[3]是油藏有机地球化学的应用技术，通过热解分析技术，即对样品利用程序升温热蒸馏的方法，随钻测出烃源岩和储集岩中烃信息，对生油岩和储集岩进行评价。注水开采的过程实际上就是注入水逐渐置换和驱替孔隙中的原油的过程，孔隙中含水饱和度逐渐上升、含油饱和度逐渐降低是这一过程的必然结果，我国油田大多数处于开发中后期, 油井产液中的水含量越来越高。在注入开采的过程中含油量下降，油组分相应发生变化，地化录井中岩石热解参数也会变化。随含水上升，气相色谱峰值、峰型以及主要特征参数，以及热解参数，都呈现了有规律的变化，因此地化色谱是可以反映油层水淹变化特征的。地化录井对储层评价方法可分为人工解释和定量分析，储层评价的最终依据是试油结论。由人工解释油气层主要是依赖录井解释工程师丰富的实际经验, 存在着很大的偶然性和误差。国内外的工程师对油气层的识别也提出了多种解决方法,其中包括原始数据法、含油饱和度比值法、图版解释法、贝叶斯判别分析法、灰色关联度分析法等综合评价方法, 但是，从识别所花费的时间和效果方面来说都不是很理想[4]。

轻烃录井[5]是应用样品预处理技术与气相色谱分离分析技术直接以岩屑、井壁取心、岩心、钻井液为研究对象，经检测器检测可得到轻烃组分色谱图，色谱图反映出了储层样品中轻烃（C1～C9）的含量、组成及分布特点，应用地球化学的理论，进行解释、评价储层含油性的录井方法。轻烃的组成包含有正构烷烃、异构烷烃、环烷烃和芳香烃等103种化合物，这些是石油和天然气的重要组成部分。随储层含水程度的升高，轻烃不同组分将会发生下列变化：油中的总烃含量降低，油中的芳烃含量大大降低，环烷烃化学性质稳定变化不明显，带季碳原子(DM)的异构烷烃与环烷烃的比值会显著降低。现有的轻烃录井对储层评价方法（如：图谱直观法、图版识别法和参数法等）主要是依赖工程师选取标定原油或油层样品各轻烃参数，建立标准化模型，这一过程费时、费力，且带有人的经验和主观性，因此不是十分理想。

## 2.2 自组织映射算法

自组织映射算法[6,7]（Self -organizing Feature Map，简称 SOM），它通过学习能够提取一组数据中的某种内在规律或重要特征，按照离散时间的方式进行分类。该网络是一个由全连接的神经元阵列组成的无导师、自组织、自学习网络，是一种可以构成对输入数据有选择地给予反应的网络，它不但能够识别输入区域临近的区域，还能够研究输入样本的分布特征和拓扑结构。当外界输入不同的样本时，网络中哪个位置的神经元兴奋起初是随机的，但自组织训练后会在竞争层形成神经元的有序排列，功能相近的神经元分部较近，功能不同的神经元离得较远，即网络可以把任意高维的输入样本映射到低维空间， 并且使得输入数据内部的某些相似性质表现为几何上邻近的特征映射。这样，就在输出层映射成一维或二维的离散图形，并保持其拓扑结构不变。这种分类反映了样本集的本质区别，而且这个过程是自动完成的，大大减弱了一致性准则中的人为因素。



**图1 SOM神经网络结构**

SOM网的基本原理[8,9]是：竞争学习（向量归一化、寻找获胜神经元）；网络输出与权值调整。也就是说，自组织特征映射的形成主要有以下三个过程：

（1）竞争。对于每个输入向量，网络中的各个神经元开始计算各自判别函数的值，其实这个判别函数是神经元之间竞争的基础。使判别函数取得最大值的特定神经元就是胜利者，即获胜神经元。

（2）合作。获胜神经元决定了兴奋神经元拓扑邻域的空间位置，从而为相邻神经元提供了合作基础。

（3）权值调整。兴奋神经元通过适当调整它们的权值来增加其对于该输入向量的判别函数值。所作的调整增强了获胜神经元对以后相似的输入向量的响应。

## 2.3 分类方法

数据分类是一个两阶段过程，包括学习阶段（构建分类模型）和分类阶段（使用模型预测给定数据的类标号）。

### 2.3.1 支持向量机

支持向量机SVM[10-12]的核心思想是寻找一个满足分类要求的最优分类超平面, 该超平面在保证分类精度的同时, 应该使分类间隔最大化。目标检测问题可以简化为一个二分类问题, 这里主要讨论支持向量机如何解决二分类问题. 给定训练样本集，，，超平面为. 为使分类面对所有样本正确分类并具备分类间隔, 要求其满足.由此可得分类间隔为, 于是构造最优超平面的问题便转化为如下带约束的最小值问题:



引入Lagrange 函数：

其中为Lagrange 系数. 约束最优化问题由Lagrange 函数的鞍点决定, 并且最优化问题的解在鞍点处满足对ω和b的偏导为0. 将该QP问题转化为如下相应的对偶问题:

;



经计算, 最优权值向量和最优偏置分别为





其中. 因此得到最优分类面, 而最优分类函数为

，其中。

支持向量机的核心部分便是核函数[13,14]。维数低的空间内,向量集一般难以进行划分,只有将其映射在维数高的空间内才可以解决。可是这样就会导致计算方面更加的复杂,此时核函数的出现恰好能够处理此问题。只有选取到恰当的核函数,才能够获取在维数高的空间内的分类函数。于支持向量机理论内,选取相宜的核函数就会使算法获得更好的性能。如何选择合适的核函数是支持向量机中未有理论依据的一个难点问题，目前的核函数都是在实验中去尝试,常用的核函数有：

(l)线性核:

(2)多项式核:

(3)高斯核:

(4)S型核：

支持向量机具有较好的分类性能, 但它只能对两类样本进行分类, 而实际应用中往往需对多个类别分类。因而需要将SVM推广到多类分类的问题中去, 这些支持向量机多类分类算法基于两种方式: 一种是将所有类别的样本组合在一起进行分类, 称为一次性求解法; 另一种是分解重构法, 即将多类问题转化为多个两类问题, 其中心思想是每次将工作集划分不同的两个类别, 在此基础上构造出支持向量机两类分类器, 最后根据某一组合策略将分类器进行组合实现多类分类。一次性求解方法是一种通过一个优化问题解决多类分类问题的方法,实际是对标准SVM中二次优化问题的一种推广,一次性求解方法选择的目标函数十分复杂, 计算复杂度高, 不适于实际问题的应用。因而, 在解决多类问题时一般不使用该方法。

一对多法（one-versus-rest）解决k类问题需要训练k个支持向量机两类分类器,即n个子分类器，“一对多”方法在测试阶段只使用k 个分类器, 其分类速度相对较快, 但由于每次构造分类器都要将整个工作集作为训练样本, 当工作集过大时,会影响训练速度。“一对多”方法所采用的最大优选策略会产生多个最大值情况, 这使得分类器会判断测试样本点同时属于多个类别或不属于任何一个类别, 从而影响分类精度。

一对一法（one-versus-one）解决k类问题时,需要训练k( k- 1) /2 个分类器，得到 k( k- 1) /2 个决策函数，“一对一”方法在训练过程中虽然构造k( k- 1) /2 个分类器, 但每次训练样本只取工作集中的两类, 训练速度比“一对多”方法快。但是, 如果单个两类分类器不规范化, 会导致整个k 类分类器趋于过学习。由于对未知样本作k( k- 1) /2 次决策, 当类别数k 增加时就会影响分类速度。决策阶段采用投票法, 可能存在多个类的票数相同的情况, 从而使未知样本同时属于多个类别, 影响分类精度。此外,“一对一”方法还有泛化误差无界的缺点。

### 2.3.2 Fisher判别式分析

线性判别式分析(Linear Discriminant Analysis, LDA)[15-17]是模式识别的经典算法，它是在 1996 年由 Belhumeur 引入模式识别和人工智能领域的。线性判别分析的基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果。它能够保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离，即模式在该空间中有最佳的可分离性。经典的线性判别分析中使用的是 Fisher 准则函数[18]，所以线性判别分析又被称为 Fisher 线性判别分析(Fisher LDA / FLDA) .

由于FDA是一种线性方法,因此,并不-定能够适用于非线性的数据处理。将核方法与Fisher判别相结合是一种有效的解决方案,也就是基于核的Fisher判别分析(Kernel Fisher Discriminant Analysis, KFDA)[19,20]。KFDA 方法的中心思想是将FDA方法与核方法相结合,以获取非线性处理能力，首先通过一个非线性映射, 将输入数据从原始空间映射到一个高维的线性可分的特征空间中, 然后在这个特征空间中进行线性Fisher判别分析, 从而实现相对于输入空间的非线性判别分析。

通过一个非线性映射Φ，将输入数据映射到一个高维的特征空间F 中，即Φ：Ｒｄ→Ｆ；ｘ→Φ（ｘ），然后在这个特征空间中进行FDA分析，从而实现相对输入空间的非线性判别分析。此时，问题转变为在特征空间F中求解以下问题



式中和；和是Ｆ中相应的类间散度矩阵和类内散度矩阵。

显然特征空间Ｆ 的维数很高，甚至是无穷维，直接求解变得不可能。因此KFDA 采用核函数K(xi，xj）＝＜Φ（xi），Φ（xj）＞对原始问题进行求解，则对任意测试样本ｘ在特征空间Ｆ 的Φ（x）到方向的投影为



式中，ｌ为训练样本数；θ 为核函数的参数向量。选取适当的偏置ｂ，可得到特征空间Ｆ 的Fisher判别函数为



### 2.3.3 K最近邻分类算法

K最近邻(KNN)算法[21-23]作为向量空间模型( VSM)[24,25]下最好的分类算法之一，由Cover 和Hart 于1967 年提出，基本思想是: 先计算待分类样本与已知类别的训练样本之间的距离或相似度, 找到距离或相似度与待分类样本数据最近的K 个邻居;再根据这些邻居所属的类别来判断待分类样本数据的类别。如果待分类样本数据的K个邻居都属于一个类别, 那么待分类样本也属于这个类别。否则, 对每一个候选类别进行评分, 按照某种规则来确定待分类样本数据的类别。KNN 算法是目前数据挖掘领域一种比较常见的分类算法, 由于其实现的简单性, 在许多领域有着广泛的应用。但是，作为一种惰性学习方法KNN 算法不需要构建分类模型, 所有的有关分类的计算都是在对新样本数据分类的时候进行的, 因此当样本数据的特征属性的数量较多、样本的容量较大时, 分类的时间代价很大, 分类的效果不是很好, 另外存在易受样本空间密度影响、相似性度量以及类别判断不考虑权重等不足。这会对实际应用产生很大的影响。

## 2.4 参考文献

[1] 方锡贤, 王旭波, 吴振强,等. 不同录井资料在储集层物性和含油气性上的反映[J]. 录井工程, 2015, 26(1):40-45.

[2] Kyi K K, Lynn C S, Haddad S, et al. Integration of Downhole Fluid Analysis and Advanced Mud Gas Logging Reduces Uncertainty in Reservoir Evaluation[C]. Information Processing and Trusted Computing, 2014.

[3] 李玉桓, 刘应忠. 地化录井技术现状与展望[J]. 录井工程, 2011, 22(3):7-10.

[4] Alberty M W, Fink K. The Use of Connection and Total Gases Quantitatively in the Assessment of Shale Pore Pressure[J]. Spe Drilling & Completion, 2014, 29(2): 208-214.

[5] 段毅, 赵阳, 姚泾利,等. 轻烃地球化学研究进展及发展趋势[J]. 天然气地球科学, 2014(12):1875-1887.

[6] Mera K, Ichimura T. User's mentality classification method using self-organising feature map on healthcare intelligent system for diabetic patients[J]. International Journal of Medical Engineering and Informatics, 2009, 2(1): 94-106.

[7] De A, Chakraborty K, Chakrabarti A, et al. Classification of power system voltage stability conditions using Kohonen's self‐organising feature map and learning vector quantisation[J]. European Transactions on Electrical Power, 2012, 22(3): 412-420.

[8] Zhang K, Chai Y, Yang S X, et al. Self-organizing feature map for cluster analysis in multi-disease diagnosis[J]. Expert Systems With Applications, 2010, 37(9): 6359-6367.

[9] Jintun Z. Self-organizing feature map classification and ordination of Larix principisrupprechtii forest in Pangquangou Nature Reserve[J]. Acta Ecologica Sinica, 2011.

[10] Gu B, Sheng V S, Tay K Y, et al. Incremental Support Vector Learning for Ordinal Regression[J]. IEEE Transactions on Neural Networks, 2015, 26(7): 1403-1416.

[11] Chang C, Lin C. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3).

[12] Ghimire D, Lee J. Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines[J]. Sensors, 2016, 13(6).

[13] Lu X, Zou H, Zhou H, et al. Robust Extreme Learning Machine With its Application to Indoor Positioning[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2016, 46(1).

[14] Gu B, Sheng V S, Wang Z, et al. Incremental learning for ν -Support Vector Regression[J]. Neural Networks, 2015.

[15] Jin X, Zhao M, Chow T W, et al. Motor Bearing Fault Diagnosis Using Trace Ratio Linear Discriminant Analysis[J]. IEEE Transactions on Industrial Electronics, 2014, 61(5): 2441-2451.

[16] Fraley C, Raftery A E. Model-Based Clustering, Discriminant Analysis, and Density Estimation[J]. Journal of the American Statistical Association, 2011: 611-631.

[17] Martinez A M, Kak A C. PCA versus LDA[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228-233.

[18] Jin X, Zhao M, Chow T W, et al. Motor Bearing Fault Diagnosis Using Trace Ratio Linear Discriminant Analysis[J]. IEEE Transactions on Industrial Electronics, 2014, 61(5): 2441-2451.

[19] Dong S, Wang Z, Zeng L, et al. Lithology identification using kernel Fisher discriminant analysis with well logs[J]. Journal of Petroleum Science and Engineering, 2016: 95-102.

[20]Duong T. ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R[J]. Journal of Statistical Software, 2007, 21(1): 1-16.

[21] Saez J A, Derrac J, Luengo J, et al. Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers[J]. Pattern Recognition, 2014, 47(12): 3941-3948.

[22] Derrac J, Garcia S, Herrera F, et al. Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects[J]. Information Sciences, 2014: 98-119.

[23] Jiang S, Pang G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems With Applications, 2012, 39(1): 1503-1509.

[24] Turney P D, Pantel P. From frequency to meaning: vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[25] Li H, Ma B, Lee C, et al. A Vector Space Modeling Approach to Spoken Language Identification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(1): 271-284.

# 三．研究计划

## 3.1 选题的目的及意义

油层识别是目前石油勘探与开发的重要环节，录井分析专家和地质学家都十分关心油层的录井解释及其流体性质识别方法的研究。面对油田勘探中愈加复杂的岩性、低渗透率和低电阻率等类型储集层的油层识别评价问题，常规线性、经验性录井解释技术常常表现得无能为力。从本质上来说，基于多参数信息的油气层识别属于高维、非线性的模式识别问题。

本课题将以数据挖掘方法为基础, 确定模型驱动数据挖掘的理论框架；以地化、轻烃色谱图数据和试油结果为源数据，利用自组织映射算法获取样本特征参数，同时结合油品种类的实际特征获得多参数组合；以敏感数据为核心，使用支持向量机、KNN等方法获得多参数组合的预测模型，比较预测模型的精度和效率，对分类效果不好的算法进行改进或者重新设计。由于传统的多分类方法的时间复杂度和空间复杂度较低，因此需要重新设计新的方法，提高分类效率。为避免单一分类算法带来的误差，构建集成分类器。

选取该课题的目的：一是希望改进已有的特征提取方法和多分类算法，为解决工业生产过程中高维、强关联性、非线性及含噪声数据的分类问题提供一种新思路；二是以此为核心算法，设计一个系统，能够对油气层进行较为准确的识别；三是完善系统，以达到实用的目的，解决工业油气勘探中的油气层识别问题。

## 3.2 主要研究内容

关于油气层识别的研究，课题的研究内容主要从以下几点展开。

（1）数据的降噪和特征提取。结合现有的录井背景资料，针对地化、轻烃色谱图数据的高维性、强相关性和含噪声，研究利用自组织映射算法无监督的对训练样本集进行聚类，实现数据去相关性和降维；针对数据的非线性时序关系， 研究数据高斯/非高斯特性表现方式，可采用非线性特征分析方法，提取非线性数据特征、频域序列幅值、均值、方差等统计特征定量，建立定量和定性描述。

（2）研究支持向量机多分类器的结构特征、算法改进、算法结构优化，降低分类过程的时间复杂度和空间复杂度，在保证分类精确度的同时，提升分类效率。

（3）针对数据存在的混叠现象，借助Fisher判别分析思想，将数据经过非线性映射将输入数据映射到一个高维特征空间，结合KNN算法对来分类边界数据，提高预测模型对混叠数据的预测精度。

（4）建立油气层识别系统，系统包含地化数据评价模块和轻烃数据评价等模块，每个模块包括若干种不同的分类方法，利用投票思想建立集成分类器，避免不同分类器间分类结果冲突以及使用单一分类器造成的误差，使用集成分类器可以最大程度的保证分类精确性。

## 3.3 研究方案

### 3.3.1 技术方案

录井数据的油品种类一共分为五大种：稠油、高凝油、轻质油、稀油、重质油。每种油根据水淹程度不同，又可以分为：油层、弱水淹层、中水淹层、强水淹层、差油层等。

1. 首先，用基于SOM的神经网络模型对训练样本集 G 进行聚类；然后，在聚类之后的数据集上计算各个样本与其所属聚类中心的欧氏距离，按一定比例筛选距离聚类中心较近的样本，去除距离聚类中心较远的样本，最终形成相对优良的训练集 G'。
2. 在 G'上用 KSVM 多分类器模型进行训练，最终形成基础的 SOM-KSVM 模型，优化核函数的选择，以达到最好的分类效果。对于数据存在混叠的现象，以及不同类别的边界数据归类模糊问题，采用改进类间距的方法来改变样本投影空间的分布，使得样本具有较好的投影效果，然后通过定义阈值参数来筛选出边界数据，对于边界数据，采用改进的K近邻（KNN）算法来分类，对于非边界数据，采用欧式距离来分类。

（3）建立地化数据评价技术模型和轻体数据评价技术模型。其中，技术模型中包含：数据特征分析法、面积分析法、标志峰分析法，对于不同方法带来的分类结果不同，采用投票法确定最终类别，构建集成分类器。

（4）以地化、轻烃分析评价技术为基础，构成一个综合单项技术的评价系统，为两种评价结果分配不同的权值，共同完成含油性分析。

### 3.3.2 实施方案所需要的条件

（1）所需编程环境：Java和Matlab。

（2）测试数据集：地化、轻烃色谱图数据集。

### 3.3.3 拟解决的关键问题

（1）自组织提取数据特征分量。解决特征提取过程中无法实现根据样本自组织提取，需要监督的问题。

（2）结合KNN算法，改进KSVM和KDA多分类器，构造集合分类器对数据进行准确分类。

（3）构建一个油气层识别系统，用于解决地化、轻烃数据分类的实际问题。

## 3.4 课题难点分析

（1）因为涉及到实际的应用，所以模型中的参数必须是非敏感性，同时整个模型结构在满足功能要求时需设计的尽可简单。

（2）如何为分类算法找到最优的核函数是本课题的难点之一。

（3）因为数据在采集过程中受到仪器、实验条件的干扰是不纯的，所以分类模型算法的设计，必须是具有很强的鲁棒性、抗干扰性。

## 3.5 预期研究成果及创新点

预期研究成果：

（1）提出一种基于SOM-KSVM多分类模型，将该模型应用于油气层识别过程中。

（2）提出一种基于KNN 算法的KDA分类方法，能够对重叠数据和边缘数据进行识别分类。

（3）开发出基于集成分类器的油藏综合解释评价系统，验证所提方法在实际油气识别过程中应用的可行性。

（4）在研究的工作的基础上，撰写硕士论文，发表学术文章1~2篇。

创新点：

（1）采用SOM算法，自组织、自学习的对样本数据进行去噪和特征提取。

（2）对传统的多分类方法进行改进，为SVM和Fisher分类器选取最优核函数；结合KNN算法，改进Fisher判别式分析方法，能够对混叠数据和边缘数据进行识别分类。

（3）针对不同分类方法分类结果的不同，构造集成分类器，采用投票法确定最终投票结果。将以上方法应用到油气层识别过程中，验证提出方法的可行性。

## 3.6 工作计划进度

2016年09月 ~ 2016年10月 整理相关文献，明确课题

2016年11月 ~ 2016年12月 实验方法设计

2016年12月 ~ 2017年02月 实现算法并应用于数据集验证算法的有效性

2017年03月 ~ 2017年05月 对实验结果分析整合

2017年06月 ~ 2017年11月 根据实验结果进一步深入研究

2017年12月 ~ 2018年04月 撰写大论文

|  |
| --- |
| **指导教师意见：**  **指导教师签名：**  **年 月 日** |
| **审核小组意见：**  **审核小组组长签字：**  **年 月 日** |
| **研究生根据审核小组意见对开题报告的改进措施：**  **年 月 日** |
| **备注：** |