报告正文

参照以下提纲撰写，要求内容翔实、清晰，层次分明，标题突出。**请勿删除或改动下述提纲标题及括号中的文字。**

（一）立项依据与研究内容（建议 8000 字以内）：

1．项目的立项依据（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）：

1.1 研究意义

多标记分类在文本分析、多媒体内容理解、推荐系统、生物信息学、医疗诊断、机器人导航等实际领域有着广泛的应用。随着信息技术与互联网技术的不断发展，各个领域大规模数据的不断涌现，用于多标记分类的数据特征也在悄然变化[1]。一方面，多标记数据中的标记呈现出超高维（ultrahigh dimensions）的特性（即：标记成千上万，甚至达数百万个）。比如，维基百科已建立了收录有数百万个类别的分类索引，一个网页可能隶属于其中的多个类别[2]；再比如，在中医健康状态辨识中，人体生命某一阶段的状态可由反映其生理病理属性的证型体现，而证型的种类不可胜举。依据“望、闻、问、切”四诊信息进行辨证，人体往往表现出多种证型兼挟[3,4]。另一方面，丰富的语义信息使得表示标注对象的特征信息通常也异常丰富。在解决实际任务时，特征信息常常从真实世界收集的数据中提取，用以获取其结构化表示（如特征向量）。那么，结构化的特征表示空间同样具有很高的维度。据统计，大量公开的用于科学研究的标准数据包含庞大体量的标记，同时特征维度有几万，几十万甚至上百万之多。这类标记与特征超高维的数据信息为人们感知和认识世界提供了窗口。然而，目前的技术手段如：传统的多标记学习方法[5-9]甚至已有的大规模多标记学习方法[10-18]难以有效对这类数据进行挖掘并做出预测。因此，适应超高维标记与特征数据环境构建性能与效率兼具的多标记分类学习方法具有重要的理论研究意义，且能为实际应用领域数据处理提供强有力的技术支撑。

1.2 超高维多标记数据分类学习面临的挑战

多标记数据呈现出的标记与特征超高维特性使得传统的数据分析理论、方法及技术面临着可计算性、准确性等严峻挑战。具体为：

（1）多标记数据中标记的超高维性。从标记层面来看，学习模型需要从超高维标记数据中为目标对象搜寻其相关标记。多标记学习技术能自然地表达和探索这些丰富的标记信息，因此受到广泛关注。该项技术常用的学习策略大致可分为三类：一阶策略，二阶策略和高阶策略[5]。然而，这些策略衍生出的学习方法难以应对标记空间过大所造成的学习困境。比如，一阶方法考虑为每个标记分别构建一个二类分类器进行学习。当涉及的标记（和特征）数量过大，这类方法将不再可行。以含有 325000 个标记的 WikiLSHTC-325K 数据集为例，一阶方法在此数据集上的训练时间可能需要数月[18]，且该类方法产生的模型规模（model size）巨大，比如：用线性 SVM 训练时模型规模达到 870 GB[18]），致使预测阶段也非常耗时。另外，二阶策略和高阶策略考虑利用两两标记甚至二个以上标记共有的关系进行分类建模。当标记空间包含数量众多的标记时，标记关系的学习速度会非常缓慢。因此，**如何面向超高维标记数据构建可快速计算的分类模型是一项具有重要意义的研究问题。**除了解决可计算问题，提高超高维标记的分类性能也至关重要，而利用标记信息是构建具有泛化能力的分类模型的重要手段。根据超高维标记空间的特点，标记信息的表现方式可归纳如下：首先，标记之间具有相关性[5,19,20]。矛盾的是，探索标记相关性会制约学习模型的可计算性，但是以合适的方式利用标记组成的子集之中的共现性或者样本相似标记的局部性质，有利于引导标记在可接受的时间代价下完成自身的学习任务。其次，标记存在类不平衡性[21,22]，即是指在标记集合中样本隶属于标记的数量远低于所有标记的数量。最后，超高维的标记信息呈现出长尾分布(long-tail distribution)的规律[11,23-25]，只有极少部分标记具有较多的正样本可供训练，大部分标记可利用的正样本数有限甚至极少。依据这些特性，**如何在多标记学习建模的过程中对标记信息进行探索以提高分类的准确性是另一项重要研究问题。**

（2）多标记数据中特征的超高维性。通常情况下，多标记数据具有很高的特征维度，并广泛存在于文本分析、多媒体内容理解、医疗诊断等实际应用领域中[26]。而标记空间维度的不断扩增更是加剧了学习对象表征的复杂性。考虑到多标记数据中充斥着无关、冗余和噪声特征，降维技术有助于解决这一“维数灾难（curse of dimensionality）”问题，并于近些年在多标记学习领域得到了长足的发展[27-33]。其基本思想是从原始特征空间中学得一个低维特征表示，该低维表示具有远远小于原始特征空间的维度，同时又能保留原始特征空间中的有用信息。事实证明，降维技术的引入不仅能显著节省计算成本和存储资源，而且还有

助于适应算法模型的学习机制，提高其泛化性能。这些性质对于面向大规模数据的多标记分类学习任务是非常重要的。不过，超高维标记引发的计算难题在有监督的降维过程中依然存在。再者，现有的无监督降维方法可以达到特征提取的目的，却在学习过程中忽略了标记信息。因此，**如何快速生成对标记有辨识力的低维特征表示以用于超高维标记的分类值得深入研究。**

综合上述分析可知，标记和特征的超高维度给多标记数据处理和分析在可计算性、准确性等方面带来了挑战。故而，迫切需要对多标记分类学习方法进行更为精细的设计，使之具有对超高维标记与特征数据的处理能力。这有利于推动多标记学习的理论研究，也有利于立足实际需求实现其应用价值。

1.3 国内外研究现状

1.3.1 超高维标记数据分类建模研究

多标记学习技术面向多义性对象建模而生，目前已有一些文献[5,34,35]对相关研究工作进行了综述。然而，现实生活中不断产生和累积的数据对多标记学习模型的发现及其高效计算方法的设计提出了更为苛刻的要求。基于此，一些新兴的多标记学习研究主题开始出现，面向超高维标记数据的多标记分类学习便是其中之一。为方便起见，该研究主题接下来简称为超高维标记分类学习。

超高维标记分类学习研究的是从一个庞大的候选标记集合中为目标对象找到一个相关的标记子集，常用的学习方法主要有基于嵌入（embedding-based）的方法，基于树（tree-based）的方法，以及一阶多标记学习方法，现分别介绍如下：基于嵌入的方法首先利用超高维的标记空间得到一个低维嵌入，然后构建特征空间到该低维嵌入的线性映射，最后生成所有标记的输出结果。这类方法的一个关键性假设是标记空间低秩（low rank）[36]。而这一假设却与超高维标记数据中的长尾分布规律相违背，导致学得的低维嵌入难以近似（或恢复）原有的标记信息。针对这一问题，Bhatia 等[11]提出了一种局部嵌入方法。该方法首先将训练样本划分为若干子集，然后在每个子集上学习相应标记空间的嵌入式低维表示，从而降低分类学习过程中标记长尾分布引发的性能退化问题。Xu 等[12]依据标记出现的频率将标记划分到不同子集来达到目的。对于出现频率高的头部标记（head labels）组成的子集，作者采用基于嵌入的方法构建多标记分类学习机制，而为出现频率低的尾部标记（tail labels）另设线性分类器进行学习。后来许多方法都沿用了以上二种思路进行学习建模[13,14]。然而，这类方法往往忽略了标记空间与特征空间之间的潜在关联，生成的低维嵌入无法保证超高维标记与特

征空间数据分布的一致性。基于树的方法则是从训练数据中学习归纳出一个树型层次结构，其基本思想是首先对根节点进行初始化，然后根据特定的节点划分方案（比如：优化 F 值[15]和 nDCG 排序损失[16]）逐一划分每个非叶子节点的样本空间。最后，为每个叶子节点分别构建一个基分类器，每个基分类器只负责对应的叶子节点所包含的少量标记。基于树的方法通常具有很高的计算效率，训练时间可以达到数据规模的次线性甚至更低的复杂度。然而，构建一个最优树是极其困难的，往往需要集成大量的树才能生成泛化性能较强的学习模型。最后，一阶多标记学习方法简单直接的学习策略，使其在超高维标记分类学习中受到重视。如第 1.2 节所述，该类方法在标记信息利用和计算效率等方面存在不足。为此，一些研究人员尝试提出改进方法，使之能高效处理超高维标记数据。举例来说，Jain 等[17]首先利用负样本抽样技术重组每个标记的训练数据来构建其相应的基分类器，然后对训练的多个基分类器进行评价以从部分候选标记中搜寻测试样本的相关子集。Yen 等[18]提出一种弹性网络约束的稀疏线性模型，在求解过程中采用原始对偶（primal-dual）的思想提高稀疏学习的优化效率。上述提及的方法在面向超高维标记数据的多标记分类学习中均取得了不错的效果，但并未涉及对超高维特征引发的不确定性进行建模。同时，以上方法存在的不足之处，使得有必要进一步完善和发展超高维标记分类方法。

也有研究人员采用标记选择的方法来减少超高维标记空间维度，用以构建高效的分类机制。例如，Wei 等[37]探索标记（特别是尾部标记）对常见超高维标记分类评价指标（如：PSP@k 和 PSnDCG@k）的影响来缩小用于建模的标记规模。Bi 和 Kwok[38]采用随机采样策略搜寻一个重要标记子集来近似表示原始标记信息。Balasubramanian 和 Lebanon[39]构建了一个组稀疏约束的优化框架学习标记的置信度矩阵，以此进行标记选择来分步设计特征空间到超高维标记空间的映射。以上提及的方法同样未研究特征空间的高维性及其与超高维标记空间之间潜在关联的影响。

1.3.2 联合超高维标记分类与特征降维的研究

针对特征维度过高这一问题，降维技术能够从原始特征空间中搜寻一个有效的低维特征表示，用以改善超高维标记分类的计算效率甚至性能表现。特征数据的降维方式主要有特征选择和特征提取。近年来，机器学习、数据挖掘与模式识别领域出现了大量的降维方法用以处理多标记数据[26,40]，比如：基于稀疏学习的方法利用稀疏正则项约束学习目标来实现降维[27,28]；基于信息理论的方法借

助信息理论刻画数据不确定性，从而构建特征评价函数来达到目的[30,31]。Zhang 和 Zhou[32]利用线性和非线性二种映射方式来生成低维特征表示，使之与标记空间的相关性最大。不过现有方法大多聚焦于标记规模较小的数据预处理，而无法在超高维标记数据中做到有效提取。

联合超高维标记分类与特征降维展开研究，研究人员也做了一些探索性的工作。这方面的工作还比较少，主要研究成果有：Jalan 和 Kar[41]提出了一种特征聚合（feature agglomeration）方法来加速超高维标记分类。对于每个特征，该方法首先聚合标记信息（或者直接利用训练样本在该特征下的原始信息）来构建其表示向量。在此基础上，该方法对所有特征进行层次聚类（每个聚类具有同等规模），然后从每个聚类中聚合一个“超级”特征（“super”-feature）用于后续分类学习。Wei 和 Li[42]提出了一种模型压缩方法来处理超高维标记与特征数据。通过设计L_0范数约束的学习目标，该方法预设需要参与训练的特征和标记个数，使得基于压缩数据生成的学习模型与原始数据训练的模型有相当的性能表现。Liu 和 Tsang[43]提出了一种预算感知（budget-aware）的方法进行特征选择，并同时优化训练误差损失和间隔，从而构建具有非线性预测能力的决策树模型。鉴于标记长尾分布的特性，该方法进一步利用香农-范诺稀疏编码来减少标注。同样采用决策树进行分类建模，Si 等[44]将其嵌套在 Gradient Boost 这一框架下来处理超高维标记数据。另外，考虑到超高维稀疏特征会引发决策树构建的复杂性，作者尝试用低秩分解、主成分分析、随机映射三种方式对特征数据进行降维。上述方法探索了特征聚合、特征选择、特征提取等多种途径来解决超高维标记分类模型构建中的特征空间“维数灾难”问题，这对于本项目研究工作的开展具有重要的启发作用。然而，这些方法局限在用已有的传统多标记分类方法或超高维标记分类方法产生预测结果。同时，这些方法仅考虑简化模型（特别是利用特征数据降维）来提高计算效率，在学习过程中忽略了充分利用标记信息，使得相关研究在准确性和可计算性等方面还有很大的提升空间。

综上所述，本项目拟开展基于超高维标记与特征数据的多标记分类建模关键技术研究。（1）借鉴已有的大规模机器学习模型简化的理论和实践研究，探索超高维标记与特征数据的预处理机制。（2）以经预处理的超高维多标记数据为支撑，重点研究依据标记相关性及其长尾分布等标记数据特性进行建模带来的可计算问题，以望设计出性能与效率兼具的算法模型。（3）进一步地，探索建模过程中超高维标记数据诱导的子模型之间的独立性，设计分布式学习框架

以实现并行计算。最终形成较为完善的超高维标记与特征数据环境下的多标记分类学习方法。（4）最后，以实际领域的超高维多标记数据（如中医健康状态辨识相关的四诊参数和状态信息，开源文本数据）为应用背景，一方面用于验证本项目的研究成果，另一方面考虑将研究成果转化为可落地的应用系统（如中医健康状态辨识系统），从而推动本项目的学术价值与应用价值。

参考文献：

[1] M. Wang, W.J. Fu, X.N. He, S.J. Hao, X.D. Wu. A survey on large-scale machine learning[J]. IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2020.3015777.

[2] R. Agrawal, A. Gupta, Y. Prabhu, M. Varma. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages[C]. In Proceedings of the 22nd International World Wide Web Conference, Rio de Janeiro, Brazil, 2013, pp. 13-24.

[3] 徐玮斐, 刘国萍, 王忆勤, 燕海霞, 郭睿. 近 5 年中医证候诊断客观化研究述评[J]. 中医杂志, 2016, 57(5): 442-445.

[4] 李灿东，杨雪梅，甘慧娟，赖新梅，周常恩，陈梅妹. 健康状态辨识模型算法的探讨[J]. 中华中医药杂志, 2011, 26(6): 1351-1355.

[5] M.L. Zhang, Z.H. Zhou. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (8): 1819–1837.

[6] Y. Zhu, J.T. Kwok, Z.H. Zhou. Multi-label learning with global and local label correlation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1081-1094.

[7] J. Zhang, C.D. Li, D.L. Cao, Y.J. Lin, S.Z. Su, L. Dai, S.Z. Li. Multi-label learning with label-specific features by resolving label correlations[J]. Knowledge-Based Systems, 2018, 159: 148-157.

[8] 刘慧婷, 冷新杨, 王利利, 赵鹏. 联合嵌入式多标签分类算法[J]. 自动化学报, 2019, 45(10): 1969-1982.

[9] 刘海洋, 王志海, 张志东. 基于 RelieF 剪枝的多标记分类算法[J]. 计算机学报, 2019, 42(3): 483-496.

[10] Y.Y. Shen, H.F. Yu, S. Sanghavi, I.S. Dhillon. Extreme multi-label classification from aggregated labels[C]. In Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 8752-8762.

[11] K. Bhatia, H. Jain, P. Kar, M. Varma, P. Jain. Sparse local embeddings for extreme multi-label classification[C]. In Advances in Neural Information Processing Systems 28, Montreal, Canada, 2015, pp. 730-738.

[12] C. Xu, D.C. Tao, C. Xu. Robust extreme multi-label learning[C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 1275-1284.

[13] Y. Tagami. AnnexML: Approximate nearest neighbor search for extreme multi-label classification[C]. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, 2017, pp. 455-464.

[14] W.W. Liu, X.B. Shen. Sparse extreme multi-label learning with oracle property[C]. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, 2019, pp. 4032-4041.
[15] K. Jasinska, K. Dembczynski, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, E. Hüllermeier. Extreme F-measure maximization using sparse probability estimates[C]. In Proceedings of the 33nd International Conference on Machine Learning, New York, NY, 2016, pp. 1435-1444.
[16] Y. Prabhu, M. Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning[C]. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, 2014, pp. 263-272.
[17] H. Jain, V. Balasubramanian, B. Chunduri, M. Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches[C]. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 2019, pp. 528-536.
[18] I.E.H. Yen, X.R. Huang, P. Ravikumar, K. Zhong, I.S. Dhillon. PD-Sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification[C]. In Proceedings of the 33nd International Conference on Machine Learning, New York, NY, 2016, pp. 3069-3077.
[19] S.J. Huang, Z.H. Zhou. Multi-label learning by exploiting label correlations locally[C]. In Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012, pp. 945–955.
[20] 何志芬, 杨明, 刘会东. 多标记分类和标记相关性的联合学习[J]. 软件学报, 2014, 25(9): 1967−1981.
[21] M.L. Zhang, Y.K. Li, X.Y. Liu. Towards class-imbalance aware multi-label learning[C]. In Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015, pp. 4041-4047.
[22] 万建武, 杨明. 代价敏感学习方法综述[J]. 软件学报, 2020, 31(1): 113-136.
[23] T. Wei, W.W. Tu, Y.F. Li. Learning for tail label data: A label-specific feature approach[C]. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019, pp. 3842-3848.
[24] W.W. Liu, I.W. Tsang, K.R. Müller. An easy-to-hard learning paradigm for multiple classes and multiple Labels[J]. Journal of Machine Learning Research, 2017, 18: 94:1-94:38.
[25] R. Babbar, B. Schölkopf. Data scarcity, robustness and extreme multi-label classification[J]. Machine Learning, 2019, 108(8-9): 1329-1351.
[26] Y. Li, T. Li, H. Liu. Recent advances in feature selection and its applications. Knowledge and Information Systems[J], 2017, 53(3): 551-577.
[27] L. Jian, J.D. Li, K. Shu, H. Liu. Multi-label informed feature selection[C]. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, NY, 2016, pp. 1627-1633.
[28] J. Zhang, Z.M. Luo, C.D. Li, C.E. Zhou, S.Z. Li. Manifold regularized discriminative feature selection for multi-label learning[J]. Pattern Recognition, 2019, 95: 136-150.

[29] B.L. Guo, C.P. Hou, F.P. Nie, D.Y. Yi. Semi-supervised multi-label dimensionality reduction[C]. In Proceedings of the IEEE 16th International Conference on Data Mining, Barcelona, Spain, 2016, pp. 919-924.
[30] J.S. Lee, D.W. Kim. SCLS: Multi-label feature selection based on scalable criterion for large label set[J]. Pattern Recognition, 2017, 66: 342-352.
[31] J. Zhang, Y.D. Lin, M. Jiang, S.Z. Li, Y. Tang, K.C. Tan. Multi-label feature selection via global relevance and redundancy optimization[C]. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020, pp. 2512-2518.
[32] Y. Zhang, Z.H. Zhou. Multilabel dimensionality reduction via dependence maximization[J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(3): 14:1-14:21.
[33] 白盛兴, 林耀进, 王晨曦, 陈晟煜. 基于邻域粗糙集的大规模层次分类在线流特征选择[J]. 模式识别与人工智能, 2019, 32(9): 811-820.
[34] E. Gibaja, S. Ventura. A tutorial on multilabel learning[J]. ACM Computing Surveys, 2015, 47(3): 52:1-52:38.
[35] G. Tsoumakas, I. Katakis, I.P. Vlahavas. Mining multi-label data[M]. In Data Mining and Knowledge Discovery Handbook, 2nd ed., 2010, pp. 667–685.
[36] H.F. Yu, P. Jain, P. Kar, I.S. Dhillon. Large-scale multi-label learning with missing labels[C]. In Proceedings of the 31th International Conference on Machine Learning, Beijing, China, 2014, pp. 593-601.
[37] T. Wei, Y.F. Li. Does tail label help for large-scale multi-label learning[J]? IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2315-2324.
[38] W. Bi, J.T.Y. Kwok. Efficient multi-label classification with many labels[C]. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013, pp. 405-413.
[39] K. Balasubramanian, G. Lebanon. The landmark selection method for multiple output prediction[C]. In Proceedings of the 29th International Conference on Machine Learning, Edinburgh, United Kingdom, 2012, pp. 983-990.
[40] R.B. Pereira, A. Plastino, B. Zadrozny, L.H.C. Merschmann. Categorizing feature selection methods for multi-label classification[J]. Artificial Intelligence Review, 2018, 49(1): 57-78.
[41] A. Jalan, P. Kar. Accelerating extreme classification via adaptive feature agglomeration[C]. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019, pp. 2600-2606.
[42] T. Wei, Y.F. Li. Learning compact model for large-scale multi-label data[C]. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, 2019, pp. 5385-5392.
[43] W.W. Liu, I.W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions[J]. Journal of Machine Learning Research, 2017, 18: 81:1-81:36.
[44] S. Si, H. Zhang, S.S. Keerthi, D. Mahajan, I.S. Dhillon, C.J. Hsieh. Gradient boosted decision trees for high dimensional sparse output[C]. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017, pp. 3182-3190.

2. 项目的研究内容、研究目标，以及拟解决的关键科学问题（此部分为重点阐述内容）；

2.1 研究目标

本项目的研究目标可概括为二个方面：其一，针对已有的多标记分类学习方法及其面向大规模数据的扩展研究难以满足超高维标记与特征数据处理的时间与精度要求，研究该类数据的分类建模关键技术，以解决标记与特征维度高，标记信息利用不充分等难题，为超高维多标记数据分类问题提供新的研究途径；其二，面向中医健康管理应用领域，以真实中医数据为载体，探索将本项目的研究成果用于构建中医健康状态辨识系统，创造实际应用价值。

2.2 研究内容

对于模型构建，首先研究超高维标记与特征数据的预处理方法，实现降低数据规模来达到模型简化的目的，此为本项目研究的基础；然后，依据超高维标记数据的特性，重点探索用于多标记学习的知识发现与分类建模方法。最后，研究Apache Spark 环境下所构建模型的并行计算框架设计，为进一步提高其可计算性提供技术保证。获得的成果为实际领域的超高维多标记数据分类建模提供有效手段。本项目以中医健康状态辨识为例，展开相关应用研究。具体研究内容如下：

2.2.1 模型简化：超高维多标记数据的预处理

围绕标记和特征数据超高维的特性，设计相应的数据预处理方法来提高数据分析的可计算性，为后续构建高效且具有泛化能力的分类模型创造有利条件。

（1）研究符合超高维标记数据特性的标记选择方法：标记空间的超高维度为分类建模带来了挑战。根据超高维标记数据的特性，搜寻一个重要标记子集来表示整个超高维标记空间是可行的应对方案。一方面，在多标记学习中，标记过多会引起区分一个标记的训练样本增加。在标注的训练数据有限的情况下（考虑到标注成本），大部分标记的正类样本数并不充足，这不利于标记的预测。标记关系的利用对难以区分的标记数据特别有效。换言之，一个标记与其他标记的相关度越高，这个标记越有可能获得正确的预测，反之亦然。因此，标记相关性是标记选择的一个重要标准。另一方面，标记数据呈现出长尾分布的规律，大部分标记可利用的正样本数远远小于训练样本总数。尽管大量研究表明：利用尾部标记进行学习能够提升超高维标记的分类性能[23-25]。然而，当标记出现频率极低时，这类尾部标记对分类性能的影响有限甚至毫无影响。因此，标记出现频率是标记选择的另一个重要标准。综上所述，需研究高效的标记相关性度量方法来衡

量标记的重要度，并联合标记出现频率这一因素来设计最终的标记选择方法。

（2）研究标记诱导的快速特征降维方法：降维技术是解决超高维多标记数据中特征空间“维数灾难”问题的主要手段。考虑到该类数据超高维的特性在特征和标记层面都有体现，那么直接采用无监督降维方法（如：主成分分析）能够达到快速获取低维特征表示的目的。然而，这种舍弃标记信息进行学习的方式难以提取出对标记有辨识力的特征。倘若考虑标记信息进行有监督的降维，标记空间的超高维度又会导致计算效率缓慢。因此，研究标记诱导的快速特征降维方法是客观需要，也是超高维多标记数据分类问题研究的基础。目前，大量有监督降维方法已成功应用于多标记数据，并形成了较丰富的理论成果[26,40]，不过因为时空的适应性问题难以适于超高维标记与特征数据环境。因此，借鉴多标记数据降维中较成熟的理论方法，本项目拟研究合理的超高维多标记数据降维方法，以高效地利用标记信息指导特征数据的预处理，从而获取有效的低维特征表示。

2.2.2 探索超高维标记数据特性的多标记分类建模

在重要标记子集研究的基础上进一步探索标记关系来进行多标记分类模型构建，并在建模过程中研究适于标记长尾分布的分类学习机制。

（1）研究结合标记关系的分类模型构建方法：多标记数据中标记之间存在着复杂的语义关系，主要有全局和局部二种关系表现形式。全局标记关系假设标记之间具有某种程度的关联。例如，在图像自动标注中，一幅图像具有类别标记*cloud*，则很可能也具有*sky*这一类别。而局部标记关系关注的是部分样本而非全部样本共有的标记关系。对于标记关系利用，一种方式是利用标记层次结构等外源信息。那么，将标记关系作为先验知识直接告诉算法，就能达到标记关系利用的目的。事实上，这是一种很有效的超高维标记分类学习思路[2]。然而，标记层次结构这样的外源信息在现实生活中较难获得，因而，常见的做法是从数据中统计标记的共现频率或类似指标，并将其作为标记关系纳入学习模型的训练过程。为此，基于已预处理的超高维标记与特征数据，研究并设计结合标记关系的多标记分类模型。一方面，继承标记选择过程中（全局意义上的）标记关系信息的组织方式，探索标记关系诱导的全局分类模式。同时，拟研究高效的局部标记关系度量方法，继而融合局部和全局二种关系信息进行分类模型构建，以取得比单纯全局意义上的标记关系利用更好的学习性能。

（2）研究适于标记长尾分布的分类学习机制：标记长尾分布是超高维标记数据的另一个显著特性，具体表现为：大部分训练样本隶属于少量头部标记，而

只有少量样本隶属于大量尾部标记。这种类不平衡现象会导致训练的分类器倾向于给予尾部标记很低的响应，以减少训练误差。针对这一问题，已有的超高维标记分类方法大多考虑将标记空间划分到多个标记子空间来进行处理[11,12]，从而满足标记空间低秩的要求进行分类建模。事实上，低秩假设在现实应用领域的超高维标记数据中并不成立。因而，推动和发展适于标记长尾分布的分类学习机制这一研究至关重要。本项目拟研究可缓解类不平衡问题干扰的标记（特别是尾部标记）增强方法，从而抑制标记长尾分布对学习性能的消极影响。

2.2.3 Apache Spark 环境下的并行计算框架设计

Apache Spark 在应对超高维标记与特征等大规模数据的计算问题上具有高效和易于扩展的优势。具体说来，通过引入多核处理机制或者集群，Apache Spark 能够轻易实现超高维多标记数据的并行化处理，且占用更少的硬件资源（特别是内存）。而且，在该环境下能够借助一些常用的编程语言（比如：Java 和 Python）进行算法设计，这有利于构建系统化的学习算法库，用以实现超高维标记分类建模。借助 Apache Spark 上述优势，拟研究所构建的算法模型在该环境下实现并行计算的可行性及其更新机制，从而加速超高维标记的分类。

2.2.4 应用：中医健康状态辨识

中医健康状态辨识根据生命过程中某一阶段的表征参数，侧重对反映生理病理属性的证型信息进行辨识。由于生理病理属性如：气血、脏腑、经络、形体官窍等的复杂性，证型种类繁多（目前尚无统一标准）且具有兼挟性。因此，中医健康状态辨识是一个典型的超高维标记分类问题。另外，表征参数信息丰富，主要包含“望、闻、问、切”四诊信息，可能还包含宏观参数（如地点、节气、气候等）与微观理化指标（如血压、体温、血常规、血糖、血脂、尿酸等）。在现实中表征参数往往需要从临床病历中提取。若将提取的信息拼接成统一标准下的向量表达，那么表征参数的特征表示具有稀疏、高维的特点。本项目拟将所设计的算法模型用于中医健康状态辨识研究。

2.3 拟解决的关键科学问题

（1）理论问题：建立超高维标记与特征数据的预处理机制是本项目研究的基础关键点。那么，如何设计标记选择方法以及挖掘标记与特征的隐含关联进行特征数据降维。在降低数据规模的基础上，构建性能与效率兼具的分布式超高维标记分类模型是本项目研究的核心关键点，如何利用标记相关性及其长尾分布等标记数据特性来达到这一目的。

（2）应用问题：探索本项目的理论研究成果应用于中医健康状态辨识的适应性问题，从而促进理论方法的进一步研究，实现精准且高效的辨识系统研发。

3. 拟采取的研究方案及可行性分析（包括研究方法、技术路线、实验手段、关键技术等说明）；

3.1 研究方法与关键技术

经预调研，本项目已有明确的算法设计及其应用方案，具体研究方法与关键技术描述如下：

3.1.1 模型简化：超高维空间数据预处理

（1）研究符合超高维标记数据特性的标记选择方法：超高维多标记数据中大部分标记的正样本数有限甚至极少，导致不同标记的可预测程度往往具有较大差异。基于此，构建符合标记数据特性的标记选择方法。在满足学习性能要求的前提下，搜寻重要（可预测的）标记子集来降低超高维标记数据的规模，从而提高超高维标记数据处理的时空性能。具体地，假设超高维标记数据中包含 q 个标记（记为：$l_1,l_2,...,l_q$），首先利用聚类方法如：k-means，层次聚类，谱聚类等（或标记领域知识）将 q 个标记划分到 k 个组别（记为：$G^{(1)},G^{(2)},...,G^{(k)}$，其相应的聚类中心分别表示为：$c_1,c_2,...,c_k$）。考虑到同一组别中的标记具有相似的特点[19]，那么按照一定比例分别从不同分组中提取出现频率高的标记，就能轻易获取有较多正样本且可通过标记关系预测的标记子集。然而，上述方法利用标记相关性与标记出现频率作为标记选择标准，聚合从不同分组中所选的标记来得到最终的重要标记子集，可能造成可预测的标记被遗漏。针对这一问题，拟另外构建一种全局意义上的标记选择方法。理论上来说，若某一标记与其他标记的相关性越大，且在所有标记中出现频率越高，这个标记越重要。基于此，对于标记空间中的任意标记 l_j（$1\le j\le q$），定义公式（1）计算其重要度。

$$Importance(l_j)=frequency(l_j)\sum_{i=1}^{k}similarity(l_j,c_i), \quad (1)$$

其中，$frequency(l_j)$ 表示隶属于标记 l_j 的正样本数。$similarity(l_j,c_i)$ 表示标记 l_j 与第 i 个聚类中心 c_i 之间的相似性大小，可以看出：一个标记与其他标记的相关性通过累加该标记与每个聚类中心之间的相似度近似得出。那么，利用公式（1）便可得出标记的重要度排序，从而确定一个重要标记子集参与后续学习。

（2）研究标记诱导的快速特征降维方法：在超高维多标记数据中，用于表示学习对象的特征维度通常很高，而且无关、冗余和噪声特征充斥其中。考虑对特征数据进行预处理来降低其规模，具体来说，设计标记诱导的快速特征降维方

法以生成一个有效的低维特征表示，从而解决特征空间的“维数灾难”问题。鉴于基于最大相关性最小冗余性（max-Relevance Min-Redundancy, mRMR）的特征选择方法在处理这一问题上具有坚实的理论基础和优异的性能表现，本项目拟沿用 mRMR 的特征评价方式构建优化学习框架，以高效获取对标记有辨识力的特征子集。需要注意的是，由于待处理的数据特征与标记维度都很高，mRMR 特征评价方式的引入会造成特征-标记相关性、特征-特征冗余性计算上的困难。为此，考虑特征与标记聚类中心之间的相关性来引导标记信息指导的特征选择过程；同时，利用基于锚点的方法学习特征之间的近似相关关系，用于尽可能减少所选特征之间的冗余性。那么，具体的优化学习框架可表示如下：

$$\min_{\mathbf{Z}} \|\mathbf{Z}-\mathbf{C}\|_F^2 + \alpha\,\mathrm{tr}(\mathbf{Z}^T\mathbf{R}\mathbf{Z}), \tag{2}$$

其中，$\mathbf{Z}\in\Re^{d\times k}$（$d$ 为特征个数）表示特征的置信度矩阵。$\mathbf{C}\in\Re^{d\times k}$ 表示特征与标记聚类中心之间的相似度矩阵，$\mathbf{R}\in\Re^{d\times d}$ 表示特征之间的近似相关关系矩阵。通过优化公式（2），生成的 $\mathbf{Z}$ 能够用于确定原始超高维特征空间中的重要特征，据此可以得到其有效的低维特征表示。

采用基于锚点的方法估计 $\mathbf{R}$：首先对特征数据进行聚类生成若干（如：m 个）聚类中心（或随机选择 m 个特征）作为锚点，然后建立锚点与特征之间的相似度矩阵（记为：$\mathbf{S}\in\Re^{d\times m}$）。那么，特征之间的近似相关关系矩阵 $\mathbf{R}=\mathbf{S}\Delta^{-1}\mathbf{S}^T$，其中 $\Delta\in\Re^{m\times m}$ 为对角矩阵，其任意第 j 个元素定义为 $\sum_{i=1}^{d}s_{ij}$。

另外，特征选择能够用于搜寻一个优化的特征子集，也完整保留了特征原有的性质，这使之无法面向超高维稀疏特征数据获得一个稠密的低维特征表示。基于此，借鉴文献[36]中的经验风险最小化学习框架，提供另一种有监督的特征生成方法来实现快速降维，其优化目标定义如下：

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{Y}_{red}-\mathbf{X}\mathbf{W}\mathbf{H}^T\|_F^2 + \alpha(\|\mathbf{W}\|_F^2+\|\mathbf{H}\|_F^2), \tag{3}$$

其中，$\mathbf{X}$ 为特征矩阵，$\mathbf{Y}_{red}$ 为所选重要标记组成的标记矩阵。待求得公式（3）中变量 $\mathbf{W}$ 的最优解，稠密的低维特征表示便可以通过计算 $\mathbf{XW}$ 得到。

3.1.2 探索超高维标记数据特性的多标记分类建模

在超高维空间数据预处理的基础上，考虑多标记分类学习方法的设计。一般来说，探索标记相关性及其长尾分布等标记特性进行建模，有利于降低多标记学习的难度，且分类模型因此通常具有较强的泛化性能。因此，有必要依据这些标记特性建立相应的多标记预测机制。同时，为适应标记空间超高维度的特点，探索这些标记特性进行建模的同时需要顾及算法的计算复杂度。

（1）研究结合标记关系的分类模型构建方法：本项目拟探索全局与局部二种标记关系来进行模型构建。一方面，利用标记选择过程中搜寻到的重要标记及其所属组别（如 k 个组别中标记组成的标记矩阵分别记为 $\mathbf{Y}_{red}^{(1)},\mathbf{Y}_{red}^{(2)},...,\mathbf{Y}_{red}^{(k)}$），拟构建组结构稀疏约束的全局标记关系理解的可解释模型。为此，对于任意标记分组，搜寻对组内相似标记最具区别性的特征子集来生成特征空间到相应标记子空间的映射。那么，联合不同分组标记信息诱导的映射，便可构建一个可感知不同标记之间相关关系的分类器进行学习。从经验上来看，上述方法（从全局上）考虑所有训练样本共有的标记关系进行学习建模，有助于提升多标记学习的性能表现。而且，该方法不需要定量计算两两标记的相关性大小，故而实现的时间代价较低。然而，在实际应用领域中，特别是面向大规模数据计算的应用场景，标记关系往往只适用于局部的小部分样本。基于此，在上述模型构建的基础上进一步探索标记关系的局部性质来实现多标记分类学习。具体而言，挖掘样本之间的近似相关关系来对模型进行修正，使得相似样本具有相似的（标记）预测结果。综上所述，组结构稀疏约束的多标记分类学习框架定义如下：

$$\min_{\mathbf{W}} \sum_{i=1}^{k} LOSS(\mathbf{X}_{red},\mathbf{W}^{(i)},\mathbf{Y}_{red}^{(i)}) + \beta\sum_{j=1}^{g} \| \mathbf{U}^{(j)}\mathbf{X}_{red}^{(j)}\mathbf{W} - \mathbf{X}_{red}^{(j)}\mathbf{W} \|_F^2 + \gamma\sum_{i=1}^{k} \| \mathbf{W}^{(i)} \|_{2,1}, \quad (4)$$

其中，$\mathbf{W}=[\mathbf{W}^{(1)},\mathbf{W}^{(2)},...,\mathbf{W}^{(k)}]$ 为训练的分类器，g 为样本分组个数（对于组别的划分，可通过样本的领域知识或聚类得到），$\mathbf{X}_{red}=[\mathbf{X}_{red}^{(1)};\mathbf{X}_{red}^{(2)};...;\mathbf{X}_{red}^{(g)}]$ 为所选重要特征组成的特征矩阵。$\mathbf{U}^{(j)}$ 表示第 j 个样本分组中两两样本（表征信息或标记信息）的近似相关关系矩阵。考察样本之间的近似相关关系（如 $\mathbf{U}^{(j)}$），可以采用基于锚点的方法进行估计。若所构建的超高维空间数据预处理方法能够将原始特征（或标记）空间约简至较低维度，标记局部关系也可采用 k-d 树快速搜寻每个样本的近邻，从而只计算每个样本与其近邻的相似度来近似表示。

（2）研究适于标记长尾分布的分类学习机制：面向标记长尾分布进行多标记分类学习，训练的分类器往往会对具有大量正样本的头部标记过拟合，从而在预测阶段忽略只有少量正样本的尾部标记的类别信息。而尾部标记在整个超高维标记空间中所占比重极大，若模型能够拟合尾部标记进行学习，这对于分类性能的提升大有裨益。基于此，本项目拟在标记关系诱导的分类模型的基础上，进一步建立适于标记长尾分布的分类学习机制。

考虑到真实世界中，标注工作者通常只标注几个关键点来反映相应标注对象的语义或视觉信息，大量潜在的类别信息（如：模糊的或者稀有的内容）可能已经丢失。那么，挖掘潜在标记信息存在的可能性是解决尾部标记因极端不平衡难

以预测的可行方案。也就是说，面向不均衡数据进行分类学习，已知负样本隶属于某一标记的概率，算法模型（如：*k*NN）更有可能预测出测试样本在该标记上的正确结果。根据这一设想，拟借鉴后验概率（或近似）作为信任度进行标记信息传播的策略，构建可降低类不平衡影响的标记（特别是尾部标记）增强方法来实现多标记分类。为此，在公式（4）的基础上对模型进行改进，使得在标记关系的约束下，可能具有但未知的潜在标记信息能够获取并助力分类器的训练。因此，公式（4）可进一步扩展为：

$$\min_{\mathbf{W},\mathbf{E}} \sum_{i=1}^{k}(LOSS(\mathbf{X}_{red},\mathbf{W}^{(i)},\mathbf{E}^{(i)},\boldsymbol{\delta}^{(i)})+\beta\sum_{j=1}^{g}\|\mathbf{U}\mathbf{E}^{(i)}-\mathbf{E}^{(i)}\|_F^2+\gamma\sum_{i=1}^{k}\|\mathbf{W}^{(i)}\|_{2,1})$$

$$\text{s.t.}\quad e_{pq}=1,\forall(p,q)\in\{(p,q)\,|\,y\_red_{pq}=1,y\_red_{pq}\in\mathbf{Y}_{red}\},\tag{5}$$

其中，$\mathbf{E}=[\mathbf{E}^{(1)},\mathbf{E}^{(2)},...,\mathbf{E}^{(k)}]$为包含潜在标记信息的标记矩阵，其与压缩后的标记矩阵$\mathbf{Y}_{red}$具有同等规模大小。$\boldsymbol{\delta}^{(i)}$为第$i$个标记分组中标记的平衡系数。

3.1.3 Apache Spark 环境下的并行计算框架设计

从算法设计层面上看，所构建的模型，如：公式（4）-（5），将标记数据划分到多个互无交集的标记分组来实现超高维标记分类学习。那么在本项目中，这一分类问题可以分解为多个规模较小、相互独立且与原问题形式相同的子分类问题进行求解。基于此，拟研究 Apache Spark 环境下的并行计算框架设计，从而更新所构建的模型以提高其计算效率。具体实现步骤如下：

分：驱动节点广播降维后的特征数据$\mathbf{X}_{red}$，并将标记分组（$\mathbf{Y}_{red}^{(1)},\mathbf{Y}_{red}^{(2)},...,\mathbf{Y}_{red}^{(k)}$）分别存储到$k$个执行节点。

治：驱动节点为每个执行节点分配并行化计算任务，实现不同标记分组诱导的子模型的参数更新，如：$\{(\mathbf{W}^{(1)},\mathbf{E}^{(1)}),(\mathbf{W}^{(2)},\mathbf{E}^{(2)}),...,(\mathbf{W}^{(k)},\mathbf{E}^{(k)})\}$。

合：待所有参数更新结束，聚合k个执行节点的计算结果。至此，并行计算框架下的分类器（即：$\mathbf{W}=[\mathbf{W}^{(1)},\mathbf{W}^{(2)},...,\mathbf{W}^{(k)}]$）训练完毕。

3.1.4 应用：中医健康状态辨识

鉴于用于健康状态辨识的中医数据呈现出的超高维特点，本项目拟将所设计的算法模型用于这一应用研究。首先，以临床病历为本底资料，对导出的数据进行结构化和预处理，以从中获取有效的表征参数和状态信息。在此基础上，借助所设计的算法模型，重点探索人体气血、脏腑、经络、形体官窍等生理病理属性组成的证型之间的兼挟性，继而利用四诊临床症状和体征，以及宏观与微观参数等信息实现辨证，为从中医数据中获取潜在的应用价值和可理解模式提供依据。

3.2 技术路线

围绕本项目的研究目标以及研究内容，根据上述研究方法，实现本项目的总体技术路线如图 1 所示：

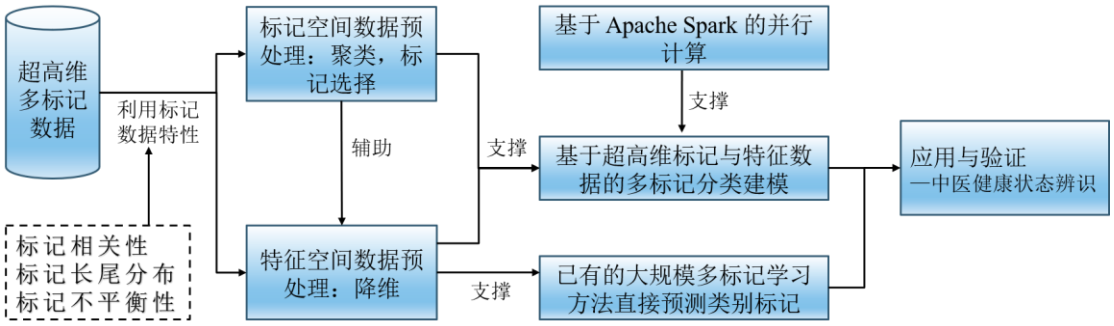

图 1 技术路线图

3.3 可行性分析

3.3.1 研究方案可行性

申请人一直从事机器学习，数据挖掘和模式识别领域相关的研究工作，主要涉及多标记学习，数据降维，以及医疗人工智能三个研究方向，先后参与了多个国家级项目研究，其中：国家重点研发计划子课题，国家自然科学基金促进海峡两岸科技合作联合基金重点项目，国家自然科学基金面上项目各一项，并在国际人工智能联合会议 IJCAI、《IEEE Transactions on Cybernetics》、《IEEE Transactions on Neural Networks and Learning Systems》、《Pattern Recognition》、《Information Sciences》、《Knowledge-Based Systems》、《Applied Soft Computing》、《Expert Systems with Applications》、《Neurocomputing》、《Applied Intelligence》、《模式识别与人工智能》、《中医杂志》、《中华中医药杂志》等国内外重要会议和期刊发表了多篇学术论文，为本项目的开展积累了较多的项目研究经验和技术资源，具体研究方案可行性分析如下：

（1）超高维多标记数据在文本分析，推荐系统，医疗诊断等现实领域广泛存在，近些年得到了学术界和产业界的广泛关注，属于国际热点研究内容。国内外的研究人员针对多标记数据中的超高维标记与特征信息已开展了一些研究工作，可以为本项目的研究提供可借鉴的参考资料。本项目深入分析了超高维标记与特征数据分析的难点问题，并结合当前大规模机器学习最前沿的理论成果，将模型简化、近似计算以及并行计算等技术用于超高维多标记数据理解的分类学习建模，方法上具有一定的可行性。

（2）本项目应用所需的中医数据经由福建中医药大学附属人民医院、附属

第二人民医院和附属厦门市中医院收集并整理，用于实验验证的数据源有所保证。值得一提的是，针对中医健康状态辨识，申请人前期已经做了一些探索性工作，如：常见证型的分类学习（发表在《中华中医药杂志》2019 年第 7 期）、典型疾病如：代谢综合征的微观辨证（获第一届中华医药博士生创新创业大赛“片仔癀”特别奖，2019 年，中国澳门），为基于超高维中医数据作进一步研究打下了良好的基础。除了真实中医数据，大量标准的超高维多标记数据也可免费下载用于实验验证（http://manikvarma.org/downloads/XC/XMLRepository.html），这为本项目的研究提供了充足的测试数据支持。

（3） 申请人在多标记分类学习，多标记数据降维，以及中医健康状态辨识等方面具有比较深厚的工作积累，也发表了一些较好的研究成果，为开展基于超高维标记与特征数据的多标记分类建模关键技术研究积累了一定的经验。详细的前期工作请参考后续的研究基础部分。

3.3.2 研究条件可行性

申请人所在的龙锦益教授团队目前有教授 2 名，副教授 3 名，讲师 3 名。团队成员基本具有海外工作经历，多年来一直从事一线的科研和开发，包含了计算机科学、统计学、医学信息处理等方面的交叉人才，承担了多项国家级项目的基础课题研究工作。该团队拥有“广东省中医药信息化重点实验室”，与本校的中医学院及附属第一医院的中医科有良好的合作。这些为本项目的开展提供坚实的基础与保障，能够保证研究结果的正确、可靠。

申请人对科研事业饱含激情，具有创新精神，敢于迎接挑战。相信：经过三年的努力，有能力解决该项申请的关键科学问题，实现提出的研究目标。

4. 本项目的特色与创新之处；

（1）考虑标记相关性及其出现频率等多个因素，提出了重要标记的度量标准进行可预测的标记选择；探索了 mRMR 特征评价函数的构建和矩阵分解的方法进行快速特征降维，从而获取有效的低维特征表示。

（2）提出了组结构稀疏约束的标记关系理解的多标记分类学习方法，并利用标记信息传播进行标记增强以提升尾部标记的可预测性。在此基础上，研究不同标记分组诱导的并行计算框架加速超高维标记分类。

（3）用于健康状态辨识的中医数据具有标记和特征超高维的特点，符合所设计的模型的应用目标。基于此，构建中医类人认知的健康状态辨识系统，为面向中医健康管理的系统研发提供新的方法学依据。