Problem Statement: How to utilize high-performance computing to enhance knowledge discovery and data mining (KDD) processes in various scientific domains. It requires developing algorithms that can efficiently process massive and complex datasets to extract meaningful insights and patterns.

Specific Aims:

1. Developing scalable algorithms that can handle large datasets while maintaining performance and accuracy.
2. Apply schemes for encoding metadata to strengthen mining algorithms and improve user interaction with KDD systems.
3. Mitigate the challenges related to overfitting and model fitting by chance in data mining systems, particularly in the context of large search spaces.
4. Explore methods for data sampling, reduction, and dimensionality reduction to handle heterogeneous datasets with both categorical and numerical fields.

Timeline:

|  |  |  |  |
| --- | --- | --- | --- |
| Tasks | Month1 | Month2 | Month3 |
| T1: Develop Scalable Algorithms | Research existing methods; Design initial algorithms | Implement Algorithms;  Conduct initial testing | Refine algorithms based on testing results; Optimize for scalability |
| T2: Implements Metadata Encoding | Investigate metadata encoding techniques; Design encoding schemes | Implement encoding schemes; Integrate with mining algorithms | Test metadata encoding functionality; Fine-tune implementation |
| T3: Address Overfitting | Study overfitting challenges; Explore model validation techniques | Develop strategies for mitigating overfitting; Implement validation methods | Test and evaluate overfitting mitigation strategies; Iterate based on results |
| T4: Explore Data Reduction | Research data sampling and dimensionality reduction methods | Implement data reduction techniques; Conduct experiments | Evaluate effectiveness of data reduction methods; Refine approaches for better results |

Appendix:

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. Future generation computer systems, 13(2-3), 99-115.