Andreas Leonidas Prodromidis
Statement of Research Interests

My research interests are primarily in distributed data mining systems with applications in interdisciplinary domains, such as information retrieval and electronic commerce. My goal is to develop systems and techniques that facilitate the discovery and sharing of useful information that is distributed across different sites in an intranet or internet.

My thesis work investigates data mining techniques that **scale up** to large and possibly physically distributed databases. Specifically, it concentrates on the problem of managing a distributed collection of intelligent learning agents that compute multiple predictive models or classifiers over large and distributed data sources. The classifiers computed are collected and combined by other agents with meta-learning capabilities. Here, meta-learning seeks to compute higher level models, called meta-classifiers, that integrate in some principled fashion the information gleaned by the separately learned classifiers to improve predictive performance. As part of this effort, I developed and evaluated JAM (Java Agents for Meta-learning), an extensible agent-based distributed data mining system that supports the remote dispatch and exchange of agents among participating data sites and employs a set of meta-learning agents that combine the multiple models that are learned.

The design and implementation of useful and practical distributed systems, however, poses an additional level of complexity compared to centralized or single-host systems. The obstacles to overcome in JAM, included dealing with heterogenous platforms, with multiple databases and (possibly) different schemas, with the design and implementation of scalable protocols among the data sites, and the efficient use of the models that are collected from remote data sites. In my dissertation, I addressed these and other aspects as well. As a result, JAM executes on multiple platforms including Solaris, Windows and Linux while special "bridging agents" are used to alleviate schema differences among the databases. The scalability and efficiency of the system is preserved by decomposing and examining the problem at two levels, first by employing distributed and asynchronous protocols at the system architecture level for managing the learning agents across the data sites, and second, by introducing special pruning algorithms at the data site level to evaluate and discard redundant classifiers. Other important problems, intrinsic with data mining systems that are also covered in my thesis include, first, the ability of the system to incorporate and combine newly computed models with existing classifiers (adaptivity to new patterns), and second, the flexibility to snap-in new machine learning methods and data mining technologies (extensibility). An empirical analysis of JAM on the real world data mining task of credit card fraud detection, evaluated the effectiveness of these approaches and demonstrated their potential utility.

JAM release 2.11 is available in the public domain and is being used by many researchers world-wide. The JAM system, however, is far from being complete and many problems remain open for research. One such problem stems out of the fact that meta-classifiers are also classifiers that too can be meta-learned and recursively combined into higher level meta-classifiers structured in multi-level tree hierarchies. While this technique constitutes a generic, scalable and efficient approach to computing broader and more effective models, it further complicates the already hard problem of deciding which classifiers or meta-classifiers to combine. For example, classifiers can be joined in many different ways with other classifiers or meta-classifiers, and meta-classifiers can be combined with other meta-classifiers of similar or different heights. The adverse effect of this approach is large and relatively slow final meta-classifiers. To classify unlabeled instances, predictions need to be generated from the leaves of the multi-level tree (base-classifiers), and propagated through the internal nodes (intermediate meta-classifiers) to the root of the tree (final meta-classifier). The throughput of a meta-classifier is of significant importance in real-time systems, such as e-commerce or intrusion detection systems. In my thesis, I examined pruning techniques for discarding redundant classifiers based on their predictive performance. In the future, I would like to the study the **throughput/predictive performance**

tradeoff, where the speed of the classifiers would also be a factor, to determine the most effective multi-level meta-classifier structure under specific real time constraints.

To further improve prediction throughput, I intend to extend JAM with the capability to generate predictions in a distributed and parallel fashion. By supporting distributed meta-classifiers (i.e. classifier trees with nodes that are not necessarily localized on a single machine), and by augmenting JAM to remotely control the **parallel execution** of classifier agents to obtain their predictions simultaneously, I expect to achieve significant classification speed-up.

In addition to continuing my work on JAM, I am also interested in exploring other research problems. One area of interest that evolved out of my experience in credit card fraud detection, is **cost based learning**. With only a few recent exceptions, accuracy based learning has been the most commonly used technique to compute predictive models. In the credit card fraud domain, however, overall accuracy is inappropriate as the single measure of predictive performance. In cases, where the cost of fraud is not equal to the cost of a false alarm, we should seek for alternative "optimality criteria" that rely on ROC analysis techniques or specific cost models. While I have already examined various cost based methods to compute meta-classifiers that are more sensitive to expensive fraudulent transactions, I am convinced that unless we study this problem at an earlier stage, e.g. when training the "base-classifiers", we will not be able to solve the problem as best as possible. There are two directions to consider here, to develop generic cost-sensitive learning algorithms that can be easily tailored to specific learning tasks, and to apply existing accuracy based learning algorithms but to appropriately transformed data sets and thus compute cost based sensitive predictive models.

One of my long-term research goals is to apply data mining techniques to application domains where they can be very effective but have not been considered yet mainly due to the **data selection problem**. These domains include customer classification and fraud detection in e-commerce systems, seismic waveform analysis and earthquake modeling (gigabytes of collected data are available online waiting to be mined) and e-mail spam detection. The data selection problem refers to the preprocessing, transformation and projection of the available data to expressive and informative features, and is probably one of the hardest, but very important stages in the knowledge discovery process. The process depends on the particular data mining task and requires application domain knowledge. The schema of the credit card databases that I used in my thesis research, for example, was developed by experienced FSTC (Financial Services Technology Consortium) personell to capture important information for fraud detection. The best data mining methods would fail, unless the proper data is collected and carefully translated to appropriate features.

Finally, I plan to pursue research in Information Retrieval (IR), and more specifically, in **Collection Fusion**, the area of collecting and combining the results from multiple IR systems, also known as meta-searching. The objective, here, is to generate a list of more relevant documents to a query than any of the individual IR systems. The approach resembles that of meta-learning. First launch a query to multiple IR systems to search different document databases in parallel and then, integrate in some principled fashion the compiled results in a single list of documents. The area is active and open for research. I believe that by augmenting the existing fusion techniques with other modalities, deduced from additional document lists generated from similar queries, or extracted from the contents of the retrieved documents themselves (e.g. by dispatching specific agents), we can design more elaborate and improved collection fusion methods. I would enjoy delving into this research area to explore numerous possibilities for meta-searching, spanning from simple techniques for combining results from existing IR systems to designing and implementing a complete real-time system for launching agents in parallel to provide additional information for the retrieved documents under user-imposed time constraints.

To summarize, I am interested in collaborating with experts from other disciplines, while continuing to build experimental software systems. My plan is to combine knowledge discovery and data mining techniques with the findings in other fields to provide effective solutions to existing problems.