ABSTRACT

Isotonic separation is a classification technique that separates *d*-dimensional data in a domain where a meaningful ordering exists not only between class labels, but also on the value set of each feature. Isotonic separation works on both discrete and continuous outcome models.

Given a data set with *n* instances and *d* features, the basic idea behind isotonic separation is to formulate a maximum flow network, or a linear programming problem (LPP). Based on the solution obtained from any one of the above models, the *d*-dimensional input space is partitioned into a number of regions: an isotonic region for each class, and a non isotonic region. The unknown test data will be classified based on the region where it appears in the input space.

The major issues identified in isotonic separation are computational complexity, feature selection, and problem size reduction. Computational complexity measures the time taken by the processor to execute an algorithm. Given a training set with *n* instances in a *d*-dimensional data space, the time complexity of isotonic separation is $O((m+n)m^2)$ where *m* denotes the number of constraints in the LPP and $n \le m \le \frac{n(n-1)}{2}$ which makes it highly complex to use isotonic separation for large datasets. Feature selection aims to select relevant features that contain ordering to construct a model. Problem size reduction focuses on reducing the number of constraints or decision variables in the LPP. A training set with *n* data points and *d* features/dimensions and the satisfaction of isotonic consistency condition in a binary classification problem, isotonic separation develops a maximum flow network model with *n+2* nodes and *m+2n* arcs or a linear programming model with *n* decision variables and *m* constraints. As training set increases, the size of the LPP or

maximum flow problem will increase linearly. It is highly complex to obtain the optimum solution of the above models using traditional methods when the data set grows. Hence, it is essential to develop some heuristic and model based approaches to handle these issues.

This thesis is devoted to binary classification using isotonic separation. The goal of the thesis is three-fold: (1) Provide algorithms to address the issue of increasing time complexity of isotonic separation when the dataset grows. (2) Provide algorithms to address the issue of problem size reduction in the LPP formulated in isotonic separation. (3) Provide a scheme to address the issue of feature selection.

In particular, there are five major works associated with above mentioned goals. They are characterized briefly below:

Evolutionary Isotonic Separation (EIS) is a hybrid heuristic algorithm in which a genetic algorithm (GA) is embedded in the training phase of the isotonic separation. It addresses the issue of increasing time complexity in solving the large scale LPP in isotonic separation. This approach proposes a slack vector to obtain the feasible solution of the LPP formulated in isotonic separation. It provides a comprehensive theory which defines the procedure of obtaining the solution of the LPP and necessary proof. Evolutionary isotonic separation lacks in two factors: it suffers from the setting of optimal convergence criterion and consumption of more time to converge.

Meta-heuristic Isotonic Separation with a new Convergence criterion based Particle Swarm Optimization (MeHeIS-CPSO) is another hybrid heuristic algorithm to overcome the issues encountered in EIS. It is a particle swarm optimization (PSO) based isotonic separation in which an optimum convergence criterion is proposed. Due to this criterion,

convergence time is reduced substantially. The related theory and proof of the proposed work are also discussed.

Graph based Isotonic separation (Graph-IS) is a model based approach in which the given training set is formulated as a graph. A new algorithm called constrained breadth first search (Co-BFS) is proposed to assign labels to vertices. In Co-BFS, a seed node is selected from the graph and the label is assigned. This algorithm iteratively evaluates the adjacent vertices of a node and labels it. These labels form an optimum or near optimum solution to the LPP. After finding the optimum solution, the boundary points are extracted by mining the bipartite graphs from the formulated graph. This work addresses the feature selection problem by using a histogram based approach for selecting the relevant isotonic features.

Cascade Isotonic Separation (Cascade-IS) addresses the issues of computational complexity and problem reduction for large scale problems by proposing cascade-IS, which utilizes cascade architecture in the isotonic separation classifier for constructing a model. The proposed framework splits the dataset into partitions and each partition is assigned to the isotonic separation separately. The output of each model is combined with the next partition in the cascade and is fed to the isotonic separation.

Soft set Based Instance Selection algorithm and Isotonic Separation (SOFIA-IS) is a heuristic approach in which a new generalized instance selection algorithm, soft set based instance selection algorithm, shortly called as SOFIA is embedded in the training part of isotonic separation. In SOFIA, instances that are relevant and important for classification are included in the training set. The relevance and importance of instances are measured using soft set. This reduced dataset is used for constructing a model for isotonic separation. The main advantage of this method is that it is suitable for small, medium, and large datasets.