Isotonic Separation with an Instance Selection Algorithm Using Softset: Theory and Experiments

B. MALAR
Department of Applied Mathematics and Computational Sciences
PSG College of Technology
Coimbatore
Tamilnadu
INDIA
rsakthimalar@gmail.com

R. NADARAJAN
Department of Applied Mathematics and Computational Sciences
PSG College of Technology
Coimbatore
Tamilnadu
INDIA
nadarajan_psg@yahoo.co.in

G. SAISUNDARAKRISHNAN
Department of Applied Mathematics and Computational Sciences
PSG College of Technology
Coimbatore
Tamilnadu
INDIA

g_ssk@yahoo.com

Abstract: In supervised machine learning, a training set containing labeled instances is taken by a learning algorithm to construct a model that is subsequently used for determining the class label of new instances. Isotonic separation is a supervised machine learning technique in which classification is represented as a Linear Programming Problem (LPP) with an objective of minimizing the number of misclassifications. It is computationally expensive to solve the LPP using traditional methods for a large dataset. Characteristics of the training set such as size, presence of noisy data, influence the learning algorithm and classification performance. To resolve this issue, this paper introduces a new linearithmic time algorithm called Soft set based instance selection algorithm (SOFIA) which provides a condensed dataset for a learning algorithm. And, a hybrid classification algorithm, SOFIA-IS which utilizes SOFIA for instance selection and isotonic separation (IS) for classification is introduced. Two sets of experimental studies are conducted on Wisconsin Breast Cancer dataset and the results are reported. First, experiments are performed on SOFIA-IS and the results are compared with isotonic separation and its variants. Then experiments are done on state of the art machine learning techniques by including SOFIA for instance selection and the results are compared with same techniques without SOFIA. Experimental and statistical results show that the condensed sets obtained by SOFIA are optimum, and SOFIA-IS and SOFIA based machine learning techniques are better in terms of classification accuracy, time and space complexity.

Key-Words: - *Isotonic Separation, Soft set, Instance Selection, SOFIA, SOFIA-IS*

1 Introduction

Isotonic Separation [3] is a supervised machine learning technique which attempts to estimate a function g which maps instances from an input space $\mathbf{X}$ to an output space $\mathbf{Y}$, given only a finite number of samples $\{(x_i, y_i) \mid x_i \in \Re^d \text{ and } y_i \in \Re\}_{i=1}^{n}$. It classifies data through an isotonic consistency condition, ie., if $\mathbf{i}, \mathbf{j} \in \mathbf{X}$ and $(\mathbf{i}, \mathbf{j}) \in R$ where R is a quasi relation if and only if $x_{ik} \geq x_{jk}$ for $k = 1, 2.....d$. These isotonic consistency conditions are derived from the domain knowledge of the problem. Isotonic Separation formulates the classification problem as a Linear Programming Problem. It aims to minimize the number of misclassifications by assigning penalty for type-1 and type-2 errors. If there are n instances of which n_1 and n_2 instances are misclassified as Y_1 and Y_2 respectively, then an objective function for the LPP is to minimize $\alpha n_1 + \beta n_2$, where α and β are penalties for misclassification. The relation R becomes the constraints of the LPP. At the worst case, a problem with n instances has $\frac{n(n-1)}{2}$ constraints.

Three major factors that influence the performance of the classifier/LPP are the number of instances n, the number of isotonic constraints m and the dimensionality d. The complexity of the LPP is proportional to the size of the training set. The limitation of traditional LP solver [16] is that it increases the number of decision variables due to slack and surplus variables. Most of the optimization techniques start with an initial feasible solution in the search space and move to the optimum solution using some deterministic transition rule. For a large scale LPP, getting an optimum solution to the linear (or non linear) problems is impractical. Interior point method [22] is infeasible for solving the above LPP because of its excessive memory requirements. It is computationally complex due to its worst case time complexity $O(\max(m,n)^3)$ and the constraint $m > n$ in isotonic separation.

The above linear programming problem can be seen as a maximum flow network model with $n+2$ nodes and $m+2n$ edges [3]. Even though, efficient solution algorithms exist in the literature for maximum flow problems [1][5], the computational time complexity of all algorithms are based on n and m. In isotonic separation $m \leq \frac{n(n-1)}{2}$, it is computationally expensive to solve the maximum flow problem for large values of n.

For huge real time problems, with thousands or millions of instances, the existing machine learning algorithms including isotonic separation are infeasible. Two approaches have been used to resolve this issue: scaling up machine learning algorithm and scaling down the dataset. The first approach aims to propose the faster algorithms with lower consumption of resources to tackle large datasets. The second approach aims to reduce the number of features or instances when the data set is large. Data reduction techniques use different approaches: dimensionality reduction, discretization and instance selection. Dimensionality reduction deals with the selection of optimal subset of features from the given set of features. Discretization deals with the conversion of continuous attributes into discrete values. Instance selection [17] consists of choosing a subset of the total data to achieve the original purpose of the classification technique as if the whole data were being used.

In this paper a new instance selection algorithm, soft set based instance selection algorithm, shortly called as SOFIA is proposed. In this method, the instances that are relevant and important to the classification are included in the training set. Important instances are selected based on the statistical parameters mean and standard deviation. 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. The proposed instance selection algorithm takes linearithmic time complexity and so it can be applied to any type of machine learning algorithms. Since the size of the reduced sub set or the number of instances in the training set is chosen by the researcher, this method is suitable for all types of problems. The objectives of the proposed framework are threefold:

1. Study about the existing isotonic separation method and identify the potential weakness of this method when the dataset set grows in size. To overcome this drawback, a hybrid isotonic separation is proposed. A new instance selection algorithm called *Soft* set based *I*nstance selection *A*lgorithm (SOFIA) is proposed to reduce the dataset. SOFIA finds the important instances of the dataset and constitutes the training set.

2. Propose a new hybrid algorithm, SOFIA-IS which employs SOFIA for instance selection and isotonic separation for classification. Experiments are conducted and the results are compared with variants of isotonic separation with respect to different parameters.

3. SOFIA is hybridized with state of the art machine learning techniques and evaluated through experiments on Wisconsin breast cancer dataset.

The paper is organized as follows: Section 2 reviews literature survey related to instance selection algorithms, Isotonic Separation and soft sets. Section 3 describes the mathematical model of a new instance selection algorithm SOFIA, and the proposed classifier SOFIA-Isotonic separation (SOFIA-IS). It also analyses the properties of SOFIA-IS theoretically. Section 4 presents the experimental results of SOFIA-IS and variants of isotonic separation. Section 5 states the conclusions of the proposed hybrid isotonic separation.

2 Related Work

Isotonic Separation is a data classification technique which has been applied in many fields such as internet content filtering [13], firm bankruptcy prediction [24], and breast cancer detection [25]. Initially, isotonic separation method has been proposed for two category separation and generalized to multi category separation [3]. In this approach, problem size is reduced by removing reflexive and transitively implied constraints from the quasi relation. A continuous outcome isotonic model has also been proposed. The major issues of isotonic separation are identified as feature selection and problem size reduction.

For firm bankruptcy prediction, Isotonic Separation method has been applied with a set of 23 financial ratios as features [24]. In this scheme, isotonic separation has been tested and compared with nine other classification techniques. Even though it is a viable technique, this scheme has been trained and tested with small samples.

To filter the objectionable internet content, isotonic separation has been deployed using PICS (Platform for Internet Content Selection) rating scheme [13]. A new approach has been proposed to refine the boundary points during testing. The limitation of this approach is that it is computationally expensive to get the refined boundary points. Also it requires the training set during testing, which consumes more memory, and it works like a lazy learner. This scheme has been trained and tested with a simulated dataset of 300 data points.

An improved version of isotonic separation has been proposed to reduce the number of constraints by finding a maximal subset of data points for each class [25]. Using this maximal subset of data points, the LPP is rebuilt. This method has been trained and tested with 699 points and gave higher accuracy.

Instance selection is performed in two perspectives [8]. In the first approach, it is done for prototype selection to obtain an optimal subset of the training set to increase the accuracy of the 1-NN classifier. In the second approach, it is applied for training set selection.

In this approach, the most suitable instances in the dataset become instances of the training set used by a learning algorithm. The various instance selection algorithms are available in the literature. The main limitation of these algorithms is higher time complexity. Table 1 gives the time complexity of various instance selection algorithms in the literature.

Soft-set theory is a general mathematical tool for dealing with uncertain, fuzzy, not clearly defined objects. In the soft-set theory, the initial description of the object is approximate in nature. Any parameterization methodologies such as real numbers, functions, mappings, words etc., can be used for describing an object. D. Molodotsov presented the notions of soft-set theory, operations and its applications in various fields such as game theory and operations research etc [21]. Soft sets have been used for decision making problem [19]. This has motivated to use the soft-set theory for instance selection.

Table 1. Comparative Study of time complexity analysis of different instance selection algorithms

Algorithm	Time Complexity
ICF [7]	$O(in^2)$
DROP [27][28]	$O(n^3)$
ENN	$O(n^2)$
CNN [8]	$O(n^3)$
RNN [10]	$O(n^3)$
Democratic instance selection [9]	Depending on the instance selection algorithm chosen to embed into the Democratic instance selection.
Divide and Conquer [6]	Depending on the instance selection algorithm chosen to embed into the Divide and Conquer instance selection.
SOFIA	*O(nlogn)*

3 SOFIA - Isotonic Separation

The definitions related to SOFIA and isotonic separation are as follows. The notations and descriptions used in this paper are given in table 2.

Knowledge representation system : It is a pair *Z = (U, A)* where *U* is a nonempty, finite set of instances called the universe and *A* is a nonempty, finite set of primitive attributes.

Soft set : A pair *(F, A)* is called a soft set over *U* where *F* is a mapping of *A* into the set of all subsets of the set $F : A \rightarrow P(U)$ where *P(U)* is the power set of *U*. The soft set $< F, A > f_j(x), 1 \leq j \leq d$ is a parameterized family and gives a collection of approximate description of an instance [21].

Quasi relation : It is a relation which satisfies reflexive and transitive property.

Table 2. Definitions and Notations

Notations	Terms
A	Universe / Dataset
x_i	i^{th} instance
y_i	Class label of x_i
F	A set of Functions
x_{ij}	j^{th} feature of i^{th} instance
C	Number of classes
A_i	Set of instances belonging to class i
d	Number of features
T	Soft Table
S_i	Score of ith instance

3.1 Problem Statement

Given a finite set of objects $\mathbf{A}$, it is partitioned into disjoint classes $\mathbf{A}_0$ and $\mathbf{A}_1$, where $\mathbf{A} = \mathbf{A}_0 \cup \mathbf{A}_1$. Each object $x_i = (x_{i1}, x_{i2}, ..., x_{id})$ has a d-dimensional feature space ie. $x_i \in \mathbb{R}^d$. For each object x_i, define a class label $y_i = g(x_i)$ $where$ $i = 1, 2, .., n$.

$$y_i = \begin{cases} 0 & \text{if } x_i \in \mathbf{A}_0 \\ 1 & \text{if } x_i \in \mathbf{A}_1 \end{cases} \quad (1)$$

The main objective of the SOFIA-IS is to design a two way classifier system that finds the optimal instances from $\mathbf{A}$ and form the condensed training set $\mathbf{A}' = \mathbf{A}'_0 \cup \mathbf{A}'_1$ where $\mathbf{A}'_0$ and $\mathbf{A}'_1$ are the important instances of $\mathbf{A}_0$ and $\mathbf{A}_1$ respectively. It also aims to construct an estimate $\hat{g}$ of g which maps instances from an input space $\mathbf{X}$ to an output space Y. The proposed classifier consists of two phases: instance selection using SOFIA and Classification using isotonic separation.

3.2 Softset based Instance Selection algorithm (SOFIA)

The main aim of any instance selection algorithm is to generate an optimal condensed dataset for classification. SOFIA is a general purpose algorithmic framework that can be applied to any machine learning technique. A model $M = (\mathbf{A}, \Omega)$ of an instance selection problem consists of :

- An input space $\boldsymbol{A}$ defined over a finite set of instances $\mathbf{A} = \{(x_i, y_i) \mid x_i \in \Re^d \ and \ y_i \in \Re\}_{i=1}^n$
- A set Ω of parameters

An optimal dataset is a set of instances that satisfies a condition which includes certain parameters. The first step in the SOFIA is to divide the training set into a number of disjoint class partitions which comprise the whole training set. Each partition contains the instances of same class and the size is not limited. Consider a finite set of instances in the universe **A = (X, Y)**, in which each instance $x_i \in \mathbf{A}$. The set A is partitioned into disjoint sets $\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_{c-1}$ where

$$\mathbf{A}_j = \bigcup_{x_i \varepsilon A} (y_i = j)$$

$$\bigcup_{i=0}^{c-1} \mathbf{A}_i = \mathbf{A}$$

Following the partitioning step, an important phase in SOFIA is determining the optimal set of instances from the given training set. This phase employs a soft set to represent the data in another representation. To create a soft set $< F, A >$ over the class partition A_i, a mapping for each feature $F : \{f_1, f_2, ..., f_d\}$ where

$f_j : A_i \longrightarrow V_i$ where $1 \le j \le d, 1 \le i \le m$ is defined.

$$f_j(x) = \begin{cases} 1 & if\ \mu_j - x_j < x < \mu_j + x_j \\ 0 & otherwise \end{cases} \quad (2)$$

Where μ_j and σ_j are the sample mean and standard deviation for j^{th} feature in the class partition A_i. Soft set is represented in the form of a table named soft table, for the purpose of storing it efficiently. Soft table is a table in which rows are labeled by instances, columns are labeled by features and the entries S_{kj} indicate the membership value of x_k with respect to feature j.

$$S_{kj} = \begin{cases} 1 & if\ \mu_j - \sigma_j < x_k < \mu_j + \sigma_j \\ 0 & otherwise \end{cases} \quad (3)$$

For each instance, compute the score for an instance x_k, denoted by S_k, the number of its parameter values lying within the standard deviation of the corresponding dimension. S_k denotes the total number of parameters that are dominated by an instance k. It is the row sum of an instance x_k, which is calculated by

$$S_k = \sum_{j=1}^{d} S_{kj} \quad (4)$$

The optimal instance is an instance with maximum score and the index of an optimal instance is obtained as follows:

$$l = \arg\max_k S_k \quad (5)$$

Let $\mathbf{A}_i^{'}$ be the reduced optimal subset of the class partition $\mathbf{A}_i$. This set is obtained by collecting all the instances which have high scores. The overall procedure of SOFIA is presented in algorithm 1.

3.3 SOFIA - Isotonic Separation

Given a finite set of objects $\mathbf{A}$, from two disjoint classes $\mathbf{A}_0$ and $\mathbf{A}_1$, SOFIA is applied on $\mathbf{A}$ and the new reduced dataset $\mathbf{A}' = \mathbf{A}'_0 \cup \mathbf{A}'_1$ is obtained where $\mathbf{A}'_0$ and $\mathbf{A}'_1$ are the important instances of $\mathbf{A}_0$ and $\mathbf{A}_1$ respectively. The basic assumption in the isotonic separation[3] is isotonic consistency condition between objects. Based on this, a quasi ordering relation R is constructed.

$$R = \left\{ (i,j) \mid \mathbf{i}, \mathbf{j} \in \mathbf{A}' \text{ and } a_{ik} \ge a_{jk} where\ k = 1,2.....d \right\} \quad (6)$$

For each object i, define a class label $y_i\ where\ i = 1,2,..,n$.

$$y_i = \begin{cases} 1 & \text{if } i \in \mathbf{A}'_1 \\ 0 & \text{if } i \in \mathbf{A}'_0 \end{cases} \quad (7)$$

This classification is considered as an optimization problem and the mathematical model is as follows:

$$\begin{aligned} &\min \\ &\alpha \sum_{i \in A_1} (1 - y_i) + \beta \sum_{i \in A_0} y_i \\ &\text{subject to the constraints} \\ &y_i - y_j \ge 0\ for\ (i,j) \in R \\ &0 \le y_i \le 1\ for\ i \in \mathbf{A}'\ \text{(boundary constraint)} \end{aligned} \quad (8)$$

Algorithm 1. Instance Selection Algorithm

Input : Partitions $\mathbf{A}_j$ where $1 \le j \le |C|$

Output : Optimal training set $\mathbf{A}' = (\mathbf{X}, \mathbf{Y})$

1. For each class partition $\mathbf{A}_j$ *// Constructing Soft table*

a. For each feature a_i

Compute mean μ_i

Compute standard deviation σ_i .

End

b. Compute $\mu_i + \sigma_i$ and $\mu_i - \sigma_i$

c. For each instance (x_i , y_i)

If x_i is in $(\mu_i + \sigma_i, \mu_i - \sigma_i)$

$T_{ij} = 1$

Else

$T_{ij} = 0$

End

d. **For** each instance (row) *i* in the soft set **T**

Find Score S_i

e. Sort the instances in descending order based on S_i

f. Select top *k* rows as important instances of data set.

End

2. $A' = \bigcup_{j=1}^{C} \mathbf{A}_j$

Here y_i is a binary variable and it can be a real variable in (0, 1). In the objective function (8), $\sum_{i\in A_1'} 1-y_i$ and $\sum_{i\in A_0'} y_i$ denote the number of Type-1 and Type-2 errors respectively. Type-1 (Type-2) error occurs when an object actually belongs to $\mathbf{A}_1'(\mathbf{A}_0')$ and system misclassifies as $\mathbf{A}_0'(\mathbf{A}_1')$. α and β are the penalties assigned for type-1 and type-2 errors. The main objective of the proposed classifier is to design a learning algorithm that learns to classify objects with minimal misclassifications. Here $\mathbf{y}^*$ be the optimum solution of LPP (8). Identify the boundary points or undominated points for each class using the below equation.

$$\mathbf{A}_1^* = \left\{ i \mid \mathbf{y}_i^* = 1 \; and \; \text{not}\exists \mathbf{y}_j^* \in \theta^*, i \neq j, \; \mathbf{y}_j^* = 1 \; and \; (i,j) \in R \right\}$$
$$\mathbf{A}_0^* = \left\{ i \mid \mathbf{y}_i^* = 0 \; and \; \text{not}\exists \mathbf{y}_j^* \in \theta^*, i \neq j, \; \mathbf{y}_j^* = 0 \; and \; (j,i) \in R \right\} \quad (9)$$

where $\mathbf{A}_1^* \neq \varnothing \; and \; \mathbf{A}_0^* \neq \varnothing$. The *d*-dimensional space will be separated into three regions based on undominated points:

$$Z_1 = \left\{ i \mid \exists j \in \mathbf{A}_1^* \; such \; that \; (j,i) \in R \text{ where } y_i^* = 0 \right\}$$
$$Z_0 = \left\{ i \mid \exists j \in \mathbf{A}_0^* \; such \; that \; (i,j) \in R \text{ where } y_i^* = 1 \right\} \quad (10)$$

and an unclassified area Z_2.

In the test set, for every object *h*, whose attribute vector is $(a_{h1}, a_{h2}, \ldots, a_{hd})$ classification is done as follows: If *h* lies in the area of Z_1, then *h* belongs to class $\mathbf{A}_1$. If *h* lies in the area of Z_0, then *h* belongs to class $\mathbf{A}_0$.

$$y_h^* = \begin{cases} 1 & \text{if h lies in } Z_1 \\ 0 & \text{if h lies in } Z_0 \end{cases} \quad (11)$$

If *h* lies in the area of Z_2, where none of the objects exist and isotonic consistency condition cannot be able to determine its class label. In that scenario, the distance between the unknown object *h* and undominated points are measured. The object *h* is assigned to the class with minimum distance.

$$D_{h1} = \beta \min \left\{ \sum_{k=1}^{d} \max(a_{hk} - a_{ik}, 0) \; i \in Z_1 \right\}$$
$$D_{h0} = \alpha \min \left\{ \sum_{k=1}^{d} \max(a_{ik} - a_{hk}, 0) \; i \in Z_0 \right\}$$
$$y_h^* = \begin{cases} 1 & \text{if } D_{h1} < D_{h0} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

3.4 Theoretical Analysis

The aim of this proposed work is to obtain an efficient instance selection algorithm that is able to scale up to large and even huge real time problems. So, an analysis of the method is essential.

Lemma 1: Time complexity of SOFIA is $O(n \log n)$.

Proof : For a dataset of *n* instances, *d* features and *k* classes, the first step in the instance selection algorithms is splitting the dataset into *k* partitions where each partition hold the instances belonging to same class. The cost of this partitioning would be *O(n)*. Then to construct a soft set, the calculation of statistical parameters mean and standard deviation need to be calculated for each feature. The time complexity would be *O(mn)*. The construction of soft table takes *O(dn)* operations. To rank the instances in the soft table, the time complexity would be *O(nlogn)*. The total time complexity of the algorithm would be *O(max(nlogn, dn))*. Since the number of features are comparatively less than the number of instances, i.e., *d*<<*n*, the time complexity would be *O(nlogn)*.

Theorem 1: Time complexity for SOFIA-IS is $O(p^3)$.

Proof : Given a training set consists of *n* objects in a *d*-dimensional data space, the first phase in the SOFIA-IS is instance selection, to filter the relevant instances in the dataset. Based

on Lemma 1, the time complexity is $O(n \log n)$. Let p be the number of instances selected after SOFIA and $p \le n$. To check the isotonic consistency constraints and construct a relation using (6) is $O(p^2 d)$. If the relation R is represented with a graph data structure, computing time to find out and eliminate transitive pairs is $\theta(p^3)$. So, the time complexity of SOFIA-IS is $O(p^3)$.

4 Experimental Analysis

To evaluate the effectiveness of the proposed SOFIA-IS, experiments are conducted on Wisconsin Breast Cancer Dataset due to the existence of monotonic property. The classification performance of SOFIA-IS is evaluated and compared with variants of isotonic separation. SOFIA is applied on the state of the art machine learning techniques and the results are compared with same without SOFIA. Statistical tests are also done to validate the significance of these algorithms.

4.1 Setup

The Wisconsin Breast Cancer dataset (WBCD) [18] was obtained from the University of Wisconsin Hospitals, containing 699 data points taken from breast cancer patients. Among these, 458 and 241 are diagnosed as benign and malignant respectively. Each data point consists of values from 1-10, stating that a higher value correspond to a more abnormal state of the tumor. The parameters of SOFIA are size of the training set, mean and standard deviation of individual features. From the training set, SOFIA selects y instances in the range of $(\mu_i + T\sigma_i, \mu_i - T\sigma_i)$ for *T = 1, 2, and 3* and *y = n/2, n/3, and n/6* . The description of the different training sets is given in table 3.

Experiments are done using modified ten-fold cross validation. In this scheme, let **A** be the given set of *d*–dimensional data for classification, which is partitioned into $\mathbf{A}_i$ where i = 1, 2, …, 10 and each partition $\mathbf{A}_i$ contains approximately equal number of malignant and benign instances. In each trial i, partition $\mathbf{A}_i$ is used as a test set and $\mathbf{A}$-$\mathbf{A}_i$ is used as a training set. From the training set, an optimal training set is created using SOFIA and a model is constructed using isotonic separation. Subsequently, the model is tested with test set and measures are considered for evaluation. Then mean and standard deviation for the measures (accuracy, precision, recall, F-measure and ROC) of 10-trials are calculated and reported [4].

Table 3. Different models of datasets for experiments

Dataset	*Parameter*	*# instances*
#1	$(\mu+\sigma, \mu-\sigma)$	*n/6* (100)
#2	$(\mu+2\sigma, \mu-2\sigma)$	*(n/6)* (100)
#3	$(\mu+3\sigma, \mu-3\sigma)$	*(n/6)* (100)
#4	$(\mu+\sigma, \mu-\sigma)$	*(n/3)* (200)
#5	$(\mu+2\sigma, \mu-2\sigma)$	*(n/3)* (200)
#6	$(\mu+3\sigma, \mu-3\sigma)$	*(n/3)* (200)
#7	$(\mu+\sigma, \mu-\sigma)$	*(n/2)* (300)
#8	$(\mu+2\sigma, \mu-2\sigma)$	*(n/2)* 300
#9	$(\mu+3\sigma, \mu-3\sigma)$	*(n/2)* 300

To assess the performance of SOFIA on state of the art machine learning techniques such as K-NN [29], Naïve Bayes [29], Support vector machine [3][14], Decision tree [23] and Back propagation network, experiments are performed on Wisconsin breast cancer dataset. For each algorithm, two sets of experiments are done. In the first set of experiments, a machine learning technique is applied without SOFIA using 10 fold cross validation experiments. In the second set of experiments, same machine

learning technique is applied on the condensed dataset obtained by SOFIA as explained in the modified ten-fold cross validation. K-NN experiments are conducted for K = 9. Back propagation is a multi layer feed forward neural network architecture in which nine neurons are used in the input layer and nine neurons are used in the hidden layer. Linear transfer function and sigmoid function is used in input layer and hidden layer respectively. Support Vector Machine experiments are done using SVM Light [15] using penalty parameter as 0.5 and polynomial kernel function. Decision tree experiments are performed using Weka [11] classifier.

One way ANOVA (Analysis of Variance) is done on the cross validation results on each dataset to assess whether the mean test error rates of the different models of SOFIA-IS and isotonic separation variants are different at the confidence level of 0.95. The *p*-value denotes the probability under the null hypothesis that the mean error rates of all the algorithms are same. Smaller *p*-value indicates the rejection of null hypothesis which means that at least one algorithm is different from others. Then, t-tests are conducted to assess the most significant algorithm at the confidence level of 0.95.

4.2 Results and Discussion

SOFIA-IS is tested with different sizes as described in the previous sub section and the results are reported in fig 1. It shows the mean values of accuracy, F-measure and ROC measures on datasets. With a dataset of size 100, Dataset #3 has higher accuracy, f-measure and true positive rate and lower false positive rate than #1 and #2. With a size of 200, dataset #6 has higher accuracy than #4 and #5. Dataset #5 has lower false positive rate. With a size of 300, dataset #9 has higher accuracy and f-measure, dataset #7 has higher true positive rate and dataset #9 has lower false positive rate.

To compare the results of SOFIA-IS with its predecessor isotonic separation, four parameters have been analysed: look ups, structure of the LPP, boundary points, and the classification performance of the classifier. Look ups define the number of times the training instances are accessed to construct a model during training and testing. In isotonic separation, look ups are directly proportionate to the size of the training set. Ryu's model requires some preprocessing of training set to combine similar data points and to find the maximal set of data points before constructing the LPP. It increases the number of lookups in training. In SOFIA-IS, look ups are substantially reduced since the relevant instances are selected by SOFIA and the comparative study of the number of look ups in training and testing is reported in figure 2. During testing, look ups are based on boundary points. In Jacob's method, boundary points are updated during testing. It increases the time complexity of the classifier at run time and an additional overhead at real time. The other methods require constant time complexity to classify the data. The asymptotic behavior of all the above methods during training and testing is shown in figure 3.

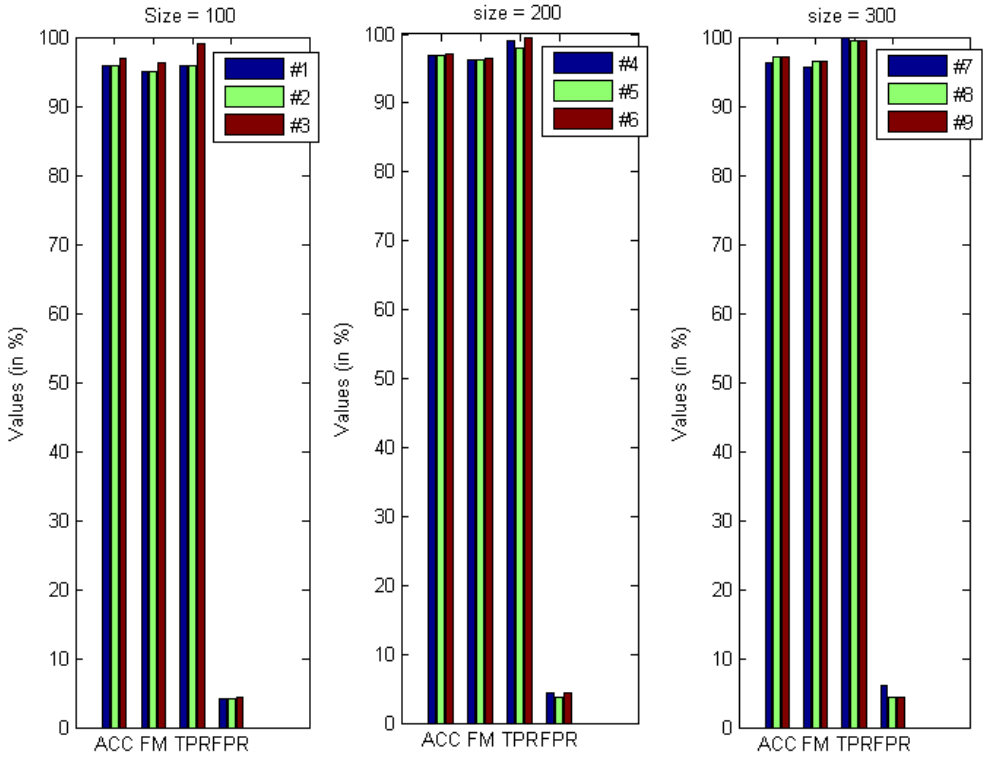

Fig. 1. Results of SOFIA-IS with a) size = *n/6* b) size = *n/3* c) size = *n/2*. X axis denotes the Metrics: Accuracy, F-measure, True Positive Rate and False Positive Rate and Y axis denotes the mean value. Legends denote the dataset ids based on table 2.

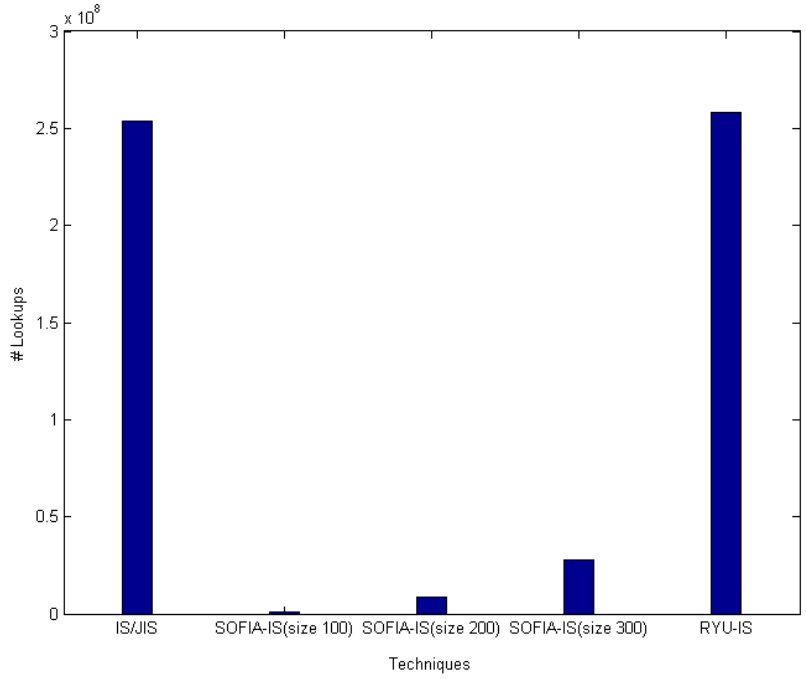

Fig. 2. Comparision of look ups in variants of isotonic separation and SOFIA-IS X axis denotes the dataset and y axis denotes the number of look ups.

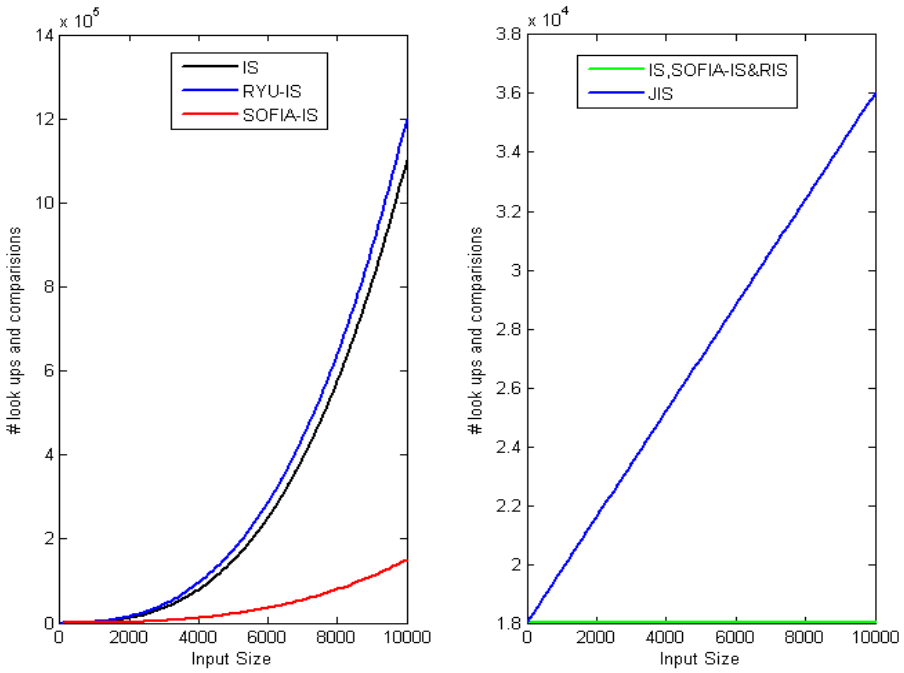

Fig. 3. Comparision of asymptotic behavior of look ups in variants of Isotonic separation and SOFIA-IS X axis denotes the size of the training/test dataset and y axis denotes the number of look ups and comparision

The second parameter, structure of the LPP plays a key role in isotonic separation. The computational complexity is the main issue in solving the LPP when the dataset grows. The parameters that affect the performance of isotonic separation are, number of decision variables (n) and number of constraints in the LPP ($|R|$). Isotonic separation generates an LPP with n decision variables and $O(n^2)$ constraints for a training set with n instances. In SOFIA-IS, the LPP contains p variables and $O(p^2)$ constraints where p denotes the number of instances in the optimal dataset after applying SOFIA and p is always less than n. Figure 5 compares the structure of the LPP in SOFIA-IS, isotonic separation and its variants in terms of variables and constraints. To compare the time complexity of solving the LPP using traditional methods, interior point and Edmond-Karp maximum flow algorithm are considered. The time complexity of interior point method is $O(m^3)$, because in isotonic separation, $n \le m \le \frac{n(n-1)}{2}$. Time complexity of Edmond-Karp algorithm to solve the maximum flow network problem is $O(nm^2)$ which is computationally expensive. The comparative study of these methods for experimental instance is demonstrated in figure 5. From this, it is observed that, due to the reduced constraints and decision variables, SOFIA-IS and Ryu's isotonic separation can take minimum time to solve the LPP.

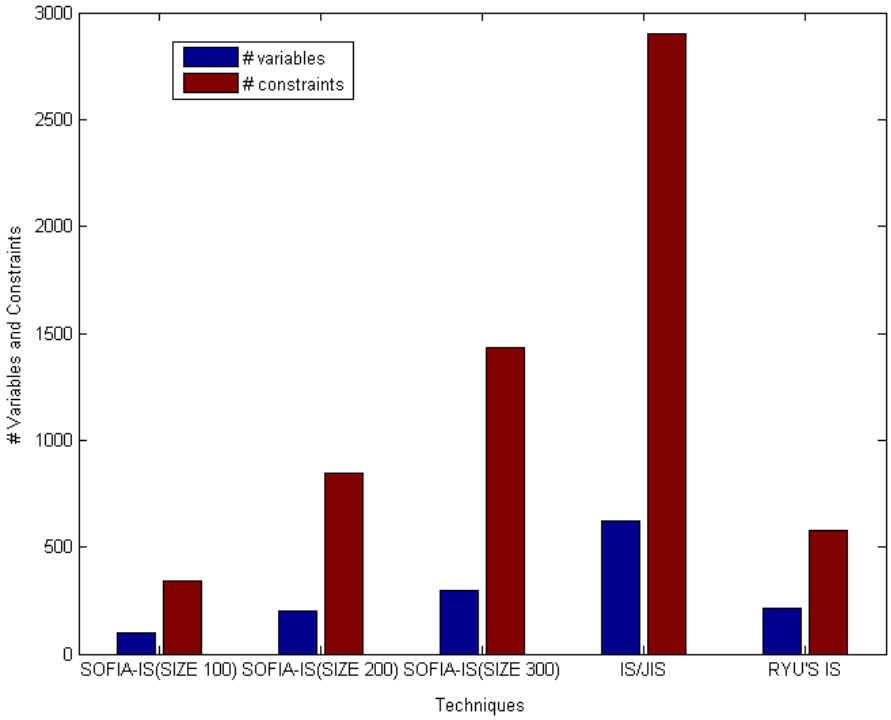

Fig. 4 Comparision of number of constraints in the LPP of variants of Isotonic separation and SOFIA-IS

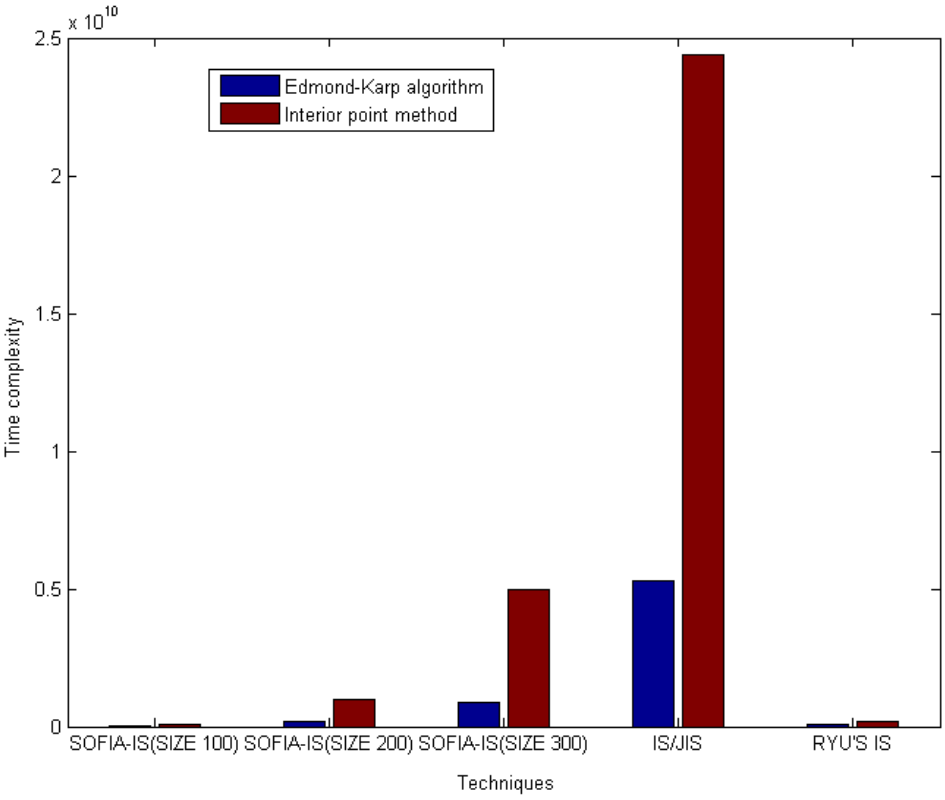

Fig. 5. Comparative time complexity analysis of SOFIA-IS, isotonic separation and its variants

The third parameter, boundary points for each class, is measured from the optimum solution of the LPP. Here, the optimum solution is real and these values are converted into integer values by setting a threshold. Because of this, misclassifications may occur and it affects the generation of boundary points. Figure 6 reports the analysis of misclassifications in the optimum solution of the LPP and the boundary points generated from variants of isotonic separation and SOFIA-IS. Isotonic separation is able to generate more number of boundary points because it considers all the instances in the training set and it has some misclassifications. In Ryu's method, even though there are no misclassifications in the LPP, minimum boundary points are identified. This has happened because of the reduced problem size. Since SOFIA-IS eliminates irrelevant data points from the training set, the least number of boundary points are obtained. However 85% of the boundary points of SOFIA-IS are in isotonic separation. Besides these, new boundary points have been generated by SOFIA-IS.

The fourth parameter, performance of the classifier, plays a key role in determining the efficiency of the classifier. The accuracy of the classifier varies based on the boundary points generated by it. Accuracy, f-measure and ROC measures are calculated and reported in the table 4. It contains mean and standard deviation of all the performance measures. It shows that SOFIA-IS has higher accuracy, f-measure and true positive rate and lower false positive rate than the other methods.

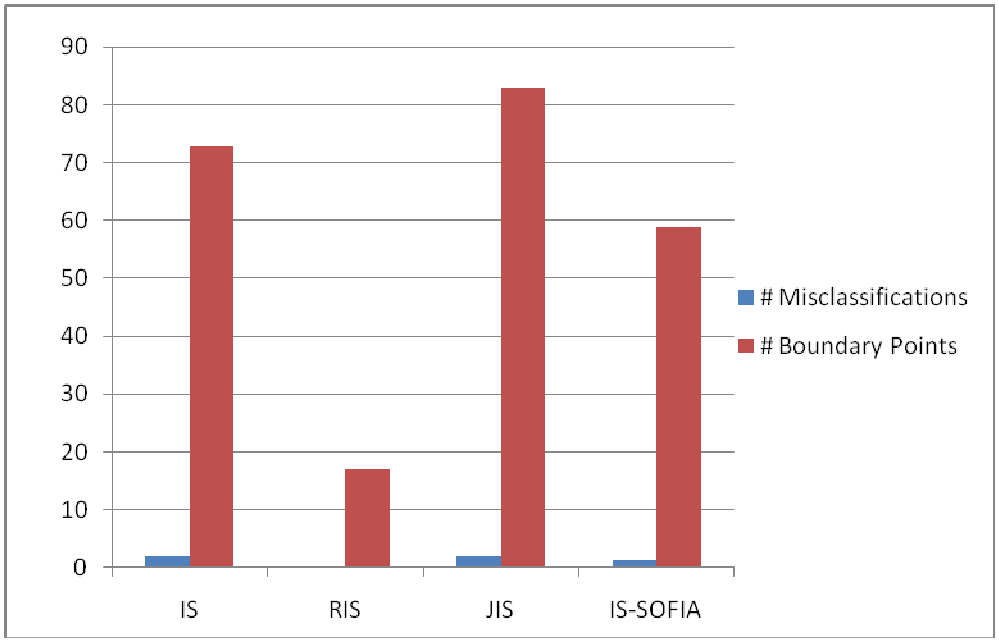

Fig. 6 Comparison of number of boundary points and the number of misclassifications in the training phase of Isotonic separation variants and SOFIA -IS

Table 4. Comparison of F-measure and ROC measures of SOFIA-IS and variants of Isotonic Separation

	Accuracy	**F-measure(%)**	**TP Rate(%)**	**FP Rate(%)**
SOFIA-IS	97.2±2.8	98.6±3.3	99.5±1.6	2.4±2.1
Isotonic Separation	95.8±3.5	94.7±4.4	93.4±4.7	2.7±3.8
Ryu's IS	79.0±1.1	83.1±8.2	99.5±1.5	34.7±18.9
Jacob's IS	96.9±3.7	93.8±1.2	98.2±2.6	4.7±1.2

The distribution of errors during tenfold cross validation is presented as a box plot in figure 7. In Ryu's isotonic separation, error rate is more because of less boundary points. In SOFIA-IS, when T=3, it generates minimum of 0% and a maximum of 4% error rate. This parameter is considered as the best for instance selection, since 75% of the times, this SOFIA-IS generates in the range of 0%-2% error rate. Results of ANOVA are presented in table 5. The probability value shows that there is some significance between the means of the proposed SOFIA-IS and other algorithms. Then post hoc pair-wise t-test is conducted to find the significance between the isotonic separation and SOFIA-IS with different parameters (T=1, 2, and T=3) and the results are shown in table 6. It shows that SOFIA-IS with T=3 is extremely significant than the other models of SOFIA-IS, isotonic separation and its variants at the confidence level of 0.95.

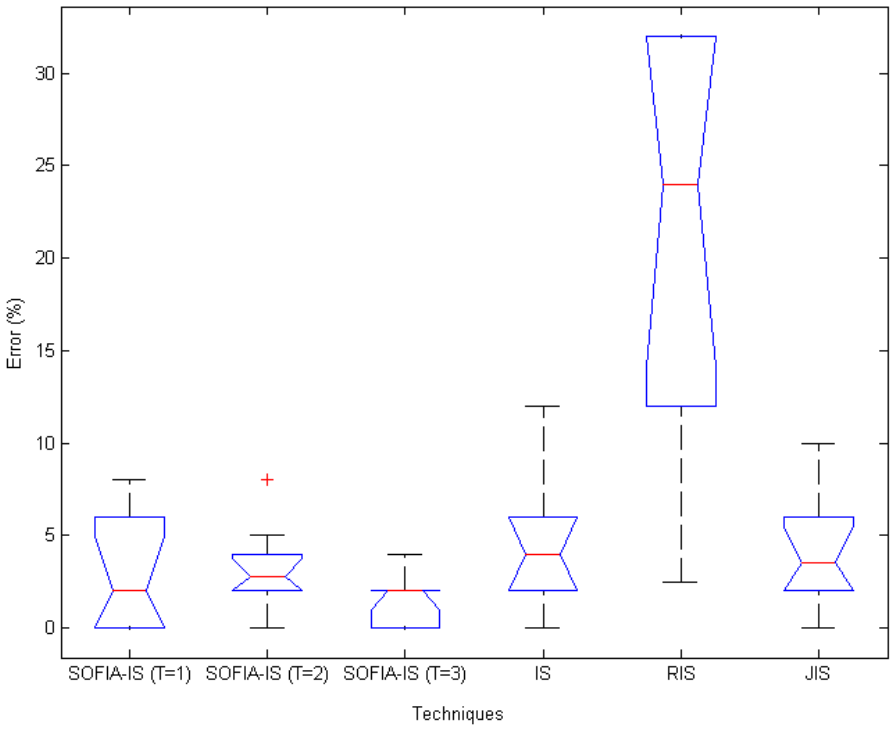

Fig. 7 Results of one way ANOVA between variants of isotonic separation and SOFIA-IS in box plot

Table 5. ANOVA table

Algorithms	**Sum of Squares**	**Degrees of Freedom**	**Mean square error**	**F-value**	**Probability**
Between	2715.13	5	543.03	20.1	2.9e-11
Within	1461.43	54	27.06		
Total	4176.56	59			

Table 6. Post hoc paired t-test against SOFIA-IS with T=3

Methods	*Probability*
SOFIA-IS(T=1)	0.0001
SOFIA-IS(T=2)	0.0001
Isotonic separation	0.0001
Ryu's Isotonic Separation	0.0001
Jacob's Isotonic Separation	0.0001

Table 7 shows the comparative results of the above mentioned machine learning techniques. K-nearest neighbor (K-NN) classifier [12] classifies data based on its k closest neighbors based on Euclidean distance. SOFIA increases the accuracy of the classifier by eliminating the noisy instances in the spherical region of the class. By setting a confidence factor as 0.25, decision tree gives 96.6% accuracy on the test set. As the confidence factor increases, it is observed that the accuracy decreases. SOFIA helps to construct an optimal decision tree. SVM classifies 96.9% of the instances correctly. SOFIA avoids the inclusion of noisy instances become support vectors and this increases the accuracy of the classifier. To validate the algorithms, pairwise t-test is conducted at the confidence level of 0.95 and the results show that SOFIA based machine learning technique is more significant than machine learning technique without SOFIA.

Table 7. Comparative study of different measures of state of the art machine learning techniques with SOFIA and without SOFIA (I denotes the results of classifier without instance selection II denotes the results of classifier with SOFIA. All values are in the form of mean ± standard deviation. * in the t-test column indicates that the corresponding algorithm is more significant than the other at the confidence level of 0.95.)

METHOD		**Accuracy(%)**	**Precision (%)**	**Recall (%)**	**F-measure(%)**	**TPR (%)**	**FPR (%)**	**t-test**
K-NN	I	95.4±4.4	95.4±4.9	95.0±4.6	94.1±5.6	93.4±6.7	3.3±4.7	
	II	96.7±3.2	96.9±3.1	96.0±3.7	96.7±3.2	97.5±1.5	1.7±3.2	*
Naive Bayes	I	95.4±4.4	95.5±4.6	95.0±4.5	94.1±5.6	93.4±6.7	3.3±4.7	
	II	97.6±2.7	97.0±2.3	97.5±1.8	97.4±2.5	99±1.5	3.8±2.4	*
Decision Tree	I	92.4±5.5	93.4±4.3	93.0±4.6	92.9±4.8	93.0±4.6	8.2±6.4	
	II	94.4±2.5	94.3±3.6	94.0±3.4	94.3±2.9	93.3±3.8	4.7±5.3	*
SVM	I	96.6±2.9	96.8±2.8	96.5±3.2	96.5±3.1	96.5±3.2	3.4±2.9	
	II	98.9±1.2	98.5±1.6	98.9±0.7	98.7±0.9	99.1±0.9	2.1±1.0	*
BPN	I	96.2±3.1	96.4±3.0	96.3±3.1	96.2±3.1	96.2±3.1	3.9±3.3	
	II	97.2±2.9	97.3±2.1	97.4±2.1	96.8±2.9	96.7±4.1	3.8±3.2	*

5 Conclusion

This paper proposes a hybrid isotonic separation where a new instance selection algorithm SOFIA is embedded in isotonic separation technique to select the important instances in the dataset. Its main objective is to provide a solution for the computational difficulty of solving the large scale LPP. SOFIA deploys softset for instance selection and the condensed data set is obtained. Then isotonic separation technique is applied on this training set to construct a model. To prove the efficiency of the proposed SOFIA-IS algorithm, it is tested on Wisconsin Breast cancer dataset using modified ten-fold cross validation and compared with isotonic separation algorithm and its variants. Statistical validation is done using one-way ANOVA followed by post hoc t-test at the confidence level of 0.95. Empirical and statistical results show that SOFIA-IS shows more significance than the variants of isotonic separation. This paper also investigates SOFIA for different machine learning algorithms. It is observed that a classifier with SOFIA classifies the unknown instances more efficiently than the classifier with no instance selection. It also removes the noisy instances from the dataset during the instance selection phase. The parameters that affect the instance selection algorithm are mean and standard deviation of individual features. Due to these parameters, SOFIA can be applied only for trivial classification problems which contain spherical datasets. Topics that remain to be explored in the future include consideration of other parameters such as distance between the instance and the centroid of the class for instance selection for strengthening SOFIA for all types of datasets and problems.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, J. B. Orlin, *Network Flows,* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] R. Chandrasekaran, Y. U. Ryu, V. Jacob, S. Hong, Isotonic separation, *Informs Journal Computing*, 17(4), 2005, 462-474.

[3] J. C. B. Christopher, A Tutorial on Support Vector Machines for Pattern Recognition, *Journal of Data Mining and Knowledge Discovery,* 2, 1998, 121-167.

[4] T. Fawcett, An Introduction to ROC Analysis, *Pattern Recognition Letters,* 27(8), 2006, 861-874.

[5] A. V. Goldberg, Recent developments in maximum flow algorithms, *Proceedings of the1998 Scandinavian Workshop on Algorithm Theory*, Springer-Verlag London, UK.

[6] Aida de Haro-Garcia, Nicolas Garcia-Pedrajas. A divide and conquer recursive approach for scaling up instance selection algorithms, *Datamining Knowledge discovery*, 18, 2009, 392-418.

[7] H. Brighton, C. Mellish. On the consistency of information filters for lazy learning algorithms, *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, 283–288. Springer-Verlag, 1999.

[8] J. R. Cano, F. Herrera, M. Lozano, Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study, *IEEE Transactions on Evolutionary Computation,* 7 (6), 2003, 561–575.

[9] Cesar Garcia Osorio, Aida de Haro-Garcia, Nicolas Garcia-Pedrajas, Democratic Instance Selection : a linear complexity instance selection algorithm based on classifier ensemble concepts, *Artificial Intelligence* 174, 2010, 410-441.

[10] G. W. Gates, The reduced nearest neighbor rule, *IEEE Transactions on Information Theory* 18 (3), 1972, 431–433.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, 11(1), 2009.

[12] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[13] V. Jacob, R. Krishnan, Y. U. Ryu, Internet content filtering using isotonic separation on content category ratings, *ACM Transactions on Internet Technology*, 7(1), 2007, 1-19.

[14] T. Joachims, Text Categorization with Support Vector Machines: Learning with

Many Relevant Features, *Proceedings of 1998 European Conference on Machine Learning (ECML), 1998*.

[15] T. Joachims, SVM Light Support Vector Machine, http://svmlight.joachims.org/. 2002.

[16] Kalyanmoy Deb, *Multiobjective Optimization Using Evolutionary Algorithm,* John Wiley & Sons Inc., NewYork, 2001.

[17] H. Liu, H. Motada, L. Yu, A selective sampling approach to active feature selection, *Artificial Intelligence* 159 (1–2) 2004, 49–74.

[18] C. J. Merz, P. M. Murphy, UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, Irvine, 1998.

[19] P. K. Maji, A. R. Roy, R. Biswas, An application of soft sets in a decision making problem, Computers and Mathematics with Application, 44, 2002.

[20] T. M. Mitchell, *Machine Learning,* McGraw Hill, 1997.

[21] D. Molodtsov, Soft set theory-first results, *Computers and Mathematics with Applications,* 37, 19–31, 1999.

[22] R. Monteiro, I. Adler, Interior Path following primal-dual algorithms part II: Convex quadratic programming. *Mathematical Programming,* 44, 1989, 43–66.

[23] J. R. Quinlan, *C4.5: Programs for Machine Learning,* Morghan Kaufman, San Mateo, CA 1993.

[24] Y. U. Ryu, W. T. Yue, Firm bankruptcy prediction; Experimental comparison of isotonic separation and other classification approaches, *IEEE Transactions on System Man and cybernetics, part A: Systems and Humans* 35(5), 2005, 727-737.

[25] Y. U. Ryu, R. Chandrasekaran, V. S. Jacob, Breast cancer detection using the isotonic separation technique, *European Journal of Operational Research*, 2007. 181, 842-854.

[26] D. L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, *IEEE Transactions on Systems, Man and Cybernetics*, (2), 1972, 408–420.

[27] D. R. Wilson, T. R. Martinez, Instance pruning techniques, *Proceedings of 14th International Conference on Machine Learning*, 403–411, 1997, Morgan Kaufmann.

[28] D. R. Wilson, T. R. Martinez, Reduction techniques for instance-based learning algorithms, *Machine Learning*, 38(3), 2000, 257–286.

[29] X. Wu, Top 10 algorithms in data mining, *Knowledge and Information Systems,* 14 1-37, 2012, Springer.