INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

Feature Subset Collection in High Dimensional Data: Fast Technique

V.M.Suresh[1], P.Vennila[2]
[1]Assistant Professor, [2] PG Scholars
[1]Department Of Information Technology, [2]Department Of PG-Computer science and Engineering
[1, 2] E.G.S.Pillay Engineering College
[1, 2] Nagapattinam, Tamilnadu, India
[1]vmsureshme@gmail.com, [2] vennilapremi@gmail.com

Abstract—*Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness. A fast clustering-based feature selection algorithm, FAST works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. To ensure the efficiency of FAST, Adopt the efficient minimum-spanning tree clustering method feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes,the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER.real-world high dimensional image, microarray, and text data, demonstrate that FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.*

Index Terms—*Feature subset selection, filter method, feature clustering, graph-based clustering*

1. INTRODUCTION

Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. Embedded method incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. *Wrapper method:* This methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high.the generality of the selected features is limited and the computational complexity is large *filter method.* This methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. filter method to reduce searching space. The filter methods are usually a good choice when then number of features is very large. Thus, we will focus on the filter method in this paper, *hybrid method* This methods area combination of filter and wrapper methods..In cluster analysis, graph-theoretic methods have been well studied and used in many applications.

The general graph-theoretic clustering is simple. we apply graph theoretic clustering methods to features. we adopt the minimum spanning tree (MST) based clustering algorithm propose a Fast clustering-bAsed feature Selection algorithm (FAST).The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features.

The proposed feature subset selection algorithm FAST was tested image, microarray, and text data sets. the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers

OBJECTIVE OF THE PAPER

i) Among high dimensional data how are selecting relevant features, removal redundant features, irrelevant data.
ii) Identify the best subsets by using feature selection.
iii) Ensure the efficiency of FAST, adopt the efficient minimum spanning tree method.
iv) Dimensionality reduction techniques can be categorized mainly into feature extraction and feature selection.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

2. RELATED WORK

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible

(i) Irrelevant features do not contribute to the predictive accuracy

(ii) Redundant features do not redound to getting a better predictor for that they provide many feature subset selection algorithms

Many algorithms effectively eliminate irrelevant features but fail to handle redundant feature. A well known example is Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted Relief-F extends Relief.

RELIEF ALGORITHM

ReliefF[13]searches for nearest neighbors of instances of different classes and weights features according to how well they differentiate instances of different classes. The other three feature selection algorithms are based on subset evaluation.This method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features also affect the speed and accuracy of learning algorithms.

CORRELATION BASED FEATURE SELECTION

CFS[5] exploits best-first search based on the evaluation of a subset that contains features highly correlated with the tar- get concept, yet uncorrelated with each other. is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other.

FAST CORRELATION BASED FILTER SOLUTION

FCBF[15],[17] is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. , we intro- duce a novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality.

THE CONSIST METHOD

The Consist method[3] searches for the minimal subset that separates classes as consistently as the full set can under best-first search strategy**.**

FOCUS-SF

FOCUS-SF[1] is a variation of FOCUS. FOCUS has the same evaluation strategy as Consist, but it examines all subsets of features. Considering the time efficiency, FOUCS-SF replaces exhaustive search in FOCUS with sequential forward selection.

HIERARCHICAL CLUSTERING

Hierarchical clustering has been adopted in word selection in the context of text classification [2],[4].Hierarchical clustering also has been used to select features on spectral data. Van Dijk and Van Hullefor[14] proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. [11] presented methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information.

3. FEATURE SUBSET SELECTION ALGORITHM

A) FRAMEWORK AND DEFINITIONS

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines [7], [9].“Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.”[6]

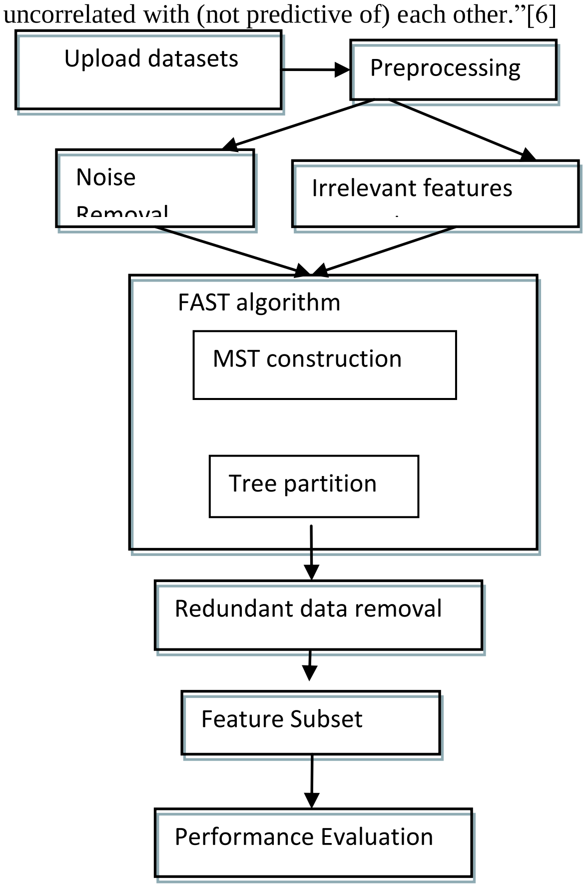

Fig. No.1: Feature subset selection algorithm

John et al. [8] presented a definition of relevant features. Suppose F to be the full set of features, Fi # F be a feature, $Si = F - \{Fi\}$ and $S'i \subseteq Si$.

Relevant feature:

Fi is relevant to the target concept C if and only if there exists some $s'\ i$, fi and c, such that, for probability $p(S'\ i = s'\ i, F\ i = fi) > 0$, $p(C = c \mid S'\ i = s'\ i, F\ i = fi) \neq p(C = c \mid S'\ i = s'\ i)$.

Markov blanket:

Given a feature $Fi \in F$, let $Mi \subset F\ (Fi \notin Mi)$, Mi is said to be a Markov blanket for Fi if and only if $p(F-Mi-\{Fi\},C \mid Fi,M\ i)=p(F-Mi-\{Fi\},C \mid Mi)$.

Redundant feature:

Let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket. The symmetric uncertainty (SU) [12] is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes.

The symmetric uncertainty is defined as follows

$SU(X,Y)=2\times Gain(X|Y)\ H(X)+H(Y)$

H(X)$=-\sum x\in X\ p(x)\log 2\ p(x)$.

$Gain(X|Y)=H(X)-H(X|Y)$

$=H(Y)-H(Y|X)$.

$H(X|Y)=-\sum y\in Y\ p(y) \sum x\in X\ p(x|y)\log 2\ p(x|y)$.

T-Relevance:

The relevance between the feature $Fi \in F$ and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by $SU(Fi,C)$.

F-Correlation:

The correlation between any pair of features Fi and Fj ($Fi,F\ j \in F \wedge i \neq j$) is called the F-Correlation of Fi and Fj, and denoted by $SU(Fi,F\ j)$.

F-Redundancy:

Let $S = \{F1,F2,...,Fi,...,F\ k<|F|\}$ be a cluster of features. if $\exists Fj \in S, SU\ (Fj,C) \geq SU(Fi,C) \wedge SU(Fi,F\ j) > SU\ (Fi,C)$ is always corrected for each $Fi \in S$.

R-Feature:

A feature $Fi \in S = \{F1,F2,...,Fk\}$ ($k<|F|$) is a representative feature of the cluster S (i.e. Fi is a R-Feature) if and only if, Fi = argmax$Fj\in S\ SU(Fj,C)$.

B) ALGORITHM AND ANALYSIS

The proposed FAST algorithm logically consists of tree steps:

(i) removing irrelevant features,
(ii) constructing a MST from relative ones, and
(iii) Partitioning the MST and selecting representative features.

4. PROPOSED SYSTEM

A novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features. FAST algorithm, it involves (i) the construction of the minimum spanning tree (MST) from a weighted complete graph; (ii) the partitioning of the MST into a forest with each tree representing a cluster and (iii) the selection of representative features from the clusters.

i) The best proportion of selected features,
ii) The best runtime,
iii) The best classification accuracy

Fast Algorithm:

inputs: D($F1$, $F2$, ..., Fm, C) - the given data set θ - the T-Relevance threshold.

output**:** S - selected feature subset .

Part 1 : Irrelevant Feature Removal

1) for i = 1 to m do
2) T-Relevance = SU (Fi, C)
3) if T-Relevance >θ then
4) S = S ∪{Fi};

Part 2: Minimum Spanning Tree Construction

5) G = NULL;
6) for each pair of features {$F'\ i$,$F'\ j$}⊂S do
7) F-Correlation = SU ($F'\ i$, $F'\ j$)
8) *Add F′ i and/or F′ j to G with* F- Correlation *as the weight of the corresponding edge*;
9) minSpanTree = Prim (G);

Part 3 : TreePartition and Representative Feature Selection

10) Forest = minSpanTree
11) for each edge Eij ∈ Forestdo
12) if SU($F'\ i$,$F'\ j$) < SU($F'\ i$,C)∧SU($F'\ i$,$F'j$)< SU($F'j$,C) then
13) Forest = Forest – Eij
14) S = ϕ
15) for each tree Ti ∈ Forestdo
16) $Fj\ R$ = argmax$F'\ k\in Ti$ SU($F'\ k$,C)
17) S = S ∪{$Fj\ R$};
18) return S

Time complexity analysis:

The first part of the algorithm has a linear time complexity $O(m)$.The second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is $O(k2)$, and then generates a MST from the graph using Prim algorithm whose time complexity is $O(k2)$. The third part partitions the MST and chooses the representative features with the complexity of $O(k)$.when $k = m$. However, k is heuristically set to be $\lfloor\sqrt{m} * \lg m\rfloor$ in the implementation of FAST. So the complexity is $O(m * \lg 2\ m)$, the time complexity of FAST deviates from $O(m2)$. Thus, on high dimensional data, the time complexity of FAST is far more less than $O(m2)$.The proposed algorithm is compared with five different types of representative feature selection algorithms.

They are (i) FCBF [16], [18], (ii) ReliefF [14], (iii) CFS [5], (iv) Consist [3], and (v) FOCUS- SF [1],set the relevance threshold to be the SU value of the

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

$\lfloor m/\log m \rfloor th$ ranked feature for each data set. For proposed FAST algorithm, we heuristically set θ to be the SU value of the$\lfloor \sqrt{m} * \lg m \rfloor th$ ranked feature for each data set.

5. IMPLEMENTATION

Feature subset selection algorithm,obtain M×N feature subsets Subset and the corresponding runtime Time with each data set. Average |Subset| and Time, we obtain the number of selected features further the proportion of selected features and the corresponding runtime for each feature selection algorithm on each data set. For each classification algorithm, we obtain M×N classification Accuracy for each feature selection algorithm and each data set.

Procedure Experimental Process:

1) M = 5, N = 101
2) DATA = {$D1$,$D2$,...,$D35$}
3) Learners = {NB, C4.5, IB1, RIPPER}
4) FeatureSelectors = {FAST, FCBF, ReliefF, CFS, Consist, FOCUS-SF}
5) for each data ∈ DATA do
6) for each times ∈ [1, M]
7) randomize instance-order for data
8) generate N bins from the randomized data
9) for each fold ∈ [1, N] do
10) TestData = bin[fold]
11) TrainingData = data – TestData
12) for each selector ∈ Feature Selectors do
13) (Subset,Time)=selector(TrainingData)
14) TrainingData′ = se
15) lect Subset from TrainingData
16) TestData′ = select Subset from TestData
17) for each learner ∈ Learners do
18) classifier = learner(TrainingData′)
19) Accuracy = apply classifier to TestData′18

Results and Analysis:

The experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the Win/Draw/Loss record.

TABLE 1: Summary of the 10 benchmark data sets

Data ID	Data name	F	I	T	Domain
1	Chess	37	31	96	Text
2	Mfeat-fourier	77	2000	10	Image face
3	Coil2000	86	98	22	Text
4	Elephant	232	1391	2	
5	Fgs-nowe	320	265	2	Image face
6	Colon	2001	62	2	Microarray bio
7	Arrhythmia	280	452	16	Microarray bio
8	Fbis.wc	2001	2463	17	Text
9	Ar10p	2401	130	10	Image face
10	Pie10p	2421	210	10	Image face

TABLE 2:proportion of selected features of the six feature selection algorithms

Data set	FAST	FCBF	CFS	Relief	Consist	FOCUS-SF
Chess	16.22	21.62	10.81	62.16	81.08	18.92
Mfeat-fourier	19.48	49.35	24.68	98.70	15.58	15.58
Coil2000	3.49	8.14	11.63	50.00	37.21	1.16Text
Elephant	0.86	8.88	5.60	6.03	0.86	0.86
Fgs-nowe	0.31	2.19	5.63	26.66	4.69	4.69
Colon	0.30	0.75	1.35	39.13	0.30	0.30
Arrhythmia	2.50	4.64	9.29	30.00	8.93	8.93
Fbis.wc	0.80	1.45	2.30	0.95	1.75	1.75
Ar10p	0.21	1.04	2.12	62.89	0.29	0.29
Pie10p	1.07	1.98	2.52	91.00	0.25	0.25
Average(image)	3.59	10.04	6.68	79.85	47.56	3.48
Average(microarray)	0.71	2.34	2.50	52.92	0.91	0.91
Average(text)	2.05	3.25	2.64	10.87	11.46	2.53
Average	1.82	4.27	3.42	42.54	5.44	2.06
Win/draw/loss	-	33/0/2	31/0/2	29/1/5	20/2/13	19/2/13

TABLE 3:Runtime of the six feature selection algorithms

Data set	FAST	FCBF	CFS	Relief	Consist	FOCUS-SF
Chess	105	60	352	12660	1999	653
Mfeat-fourier	1472	716	938	13918	3227	660
Coil2000	866	875	1483	304162	53850	1281
Elephant	783	312	905	20991	2439	1098
Fgs-nowe	977	97	736	1072	1360	1032
Colon	166	148	12249	744	1624	960
Arrhythmia	110	115	821	3684	3492	2940
Fbis.wc	14761	16207	66058	79527	579376	479651
Ar10p	706	458	57319	3874	3568	2083
Pie10p	678	1223	77579	7636	4149	2910
Average(image)	1520	4090	905678	3456	2547	2865
Average(microarray)	7543	3567	678945	347256	5907	6834
Average(text)	3792	4617	5902	4167	489032	5734
Average	4532	367832	56743	43167	4321	6743
Win/draw/loss	-	33/0/2	31/0/2	29/1/5	20/2/13	34/0/13

TABLE 4:Accuracy of the navie bayes of the six feature selection algorithms

Data set	FAST	FCBF	CFS	Relief	Consist	FOCUS-SF	Full set
Chess	92.92	92.12	90.13	88.56	89.50	94.34	87.68
Mfeat-fourier	19.48	49.35	24.68	98.70	15.58	15.58	76.07
Coil2000	3.49	8.14	11.63	50.00	37.21	1.16	78.04
Elephant	0.86	8.88	5.60	6.03	0.86	0.86	82.34
Fgs-nowe	0.31	2.19	5.63	26.66	4.69	4.69	63.06
Colon	0.30	0.75	1.35	39.13	0.30	0.30	65.61
Arrhythmia	2.50	4.64	9.29	30.00	8.93	8.93	56.33
Fbis.wc	0.80	1.45	2.30	0.95	1.75	1.75	61.89
Ar10p	0.21	1.04	2.12	62.89	0.29	0.29	72.62
Pie10p	1.07	1.98	2.52	91.00	0.25	0.25	90.67
Average(image)	3.59	10.04	6.68	79.85	47.56	3.48	45.46
Average(microarray)	0.71	2.34	2.50	52.92	0.91	0.91	45.23
Average(text)	2.05	3.25	2.64	10.87	11.46	2.53	67.45
Average	1.82	4.27	3.42	42.54	5.44	2.06	78.44
Win/draw/loss	-	33/0/2	31/0/2	29/1/5	20/2/13	19/2/13	23/2/13

TABLE 5 : Rank of the six feature selection algorithms under different types of data

	Image Data						Microarray Data					
	FAST	FCBF	CFS	ReliefF	Consist	FOCUS SF	FAST	FCBF	CFS	ReliefF	Consist	FOCUS SF
NB	3	1	2	4	5	5	1	3	2	6	4	4
C4	2	3	1	4	5	5	1	3	2	6	4	4
IB1	4	2	1	3	5	5	1	3	2	4	5	5
RIPPER	1	2	1	6	3	3	1	4	5	6	2	2
Sum	10	8	5	17	18	18	4	13	11	22	15	15
Rank	3	2	1	4	5	5	1	3	2	6	4	4

ISSN 2320-6802
ICIRET-2014

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

	Image Data						Microarray Data					
	FAST	FCBF	CFS	ReliefF	Consist	FOCUS SF	FAST	FCBF	CFS	ReliefF	Consist	FOCUS SF
NB	1	3	2	6	5	4	1	3	2	4	6	5
C4	1	2	6	4	5	1	1	2	4	6	5	3
IB1	3	4	1	6	2	4	2	3	1	6	4	5
RIPPER	5	3	1	6	2	4	1	4	5	6	2	3
Sum	12	12	10	22	14	13	5	12	12	22	17	16
Rank	2	2	1	6	5	4	1	2	2	6	5	4

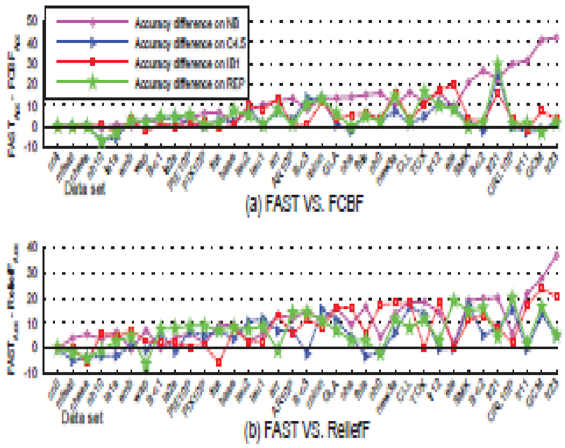

Fig.No.2: Accuracy differences between FAST and the comparing algorithms

6. CONCLUSION AND FUTURE WORK

A novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the image, microarray, and text data.proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions.FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data.

For the future work,plan to explore different types of correlation measures, and study some formal properties of feature space.

REFERENCE

[1] Almuallim H and Dietterich T.G.,Learning Boolean concepts the presence of many Asia conference on knowledge discovery and data -mining pp 109,2000.

[2] Basker L.D.andMcCallum A.K.Distributional clustering of words for text classification.,In Proceedings of the 21st Annual international ACMSIGIR Conference on Research and pacific irrelevant features,Aritificial Intelligence,6991-20,pp 279-305,1999

[3] Dash M,Liu H and Motoda H..Consistency based feature selection,In proceedings of the Fourth Development in information Retrieval,pp 96-103,1998

[4] Dhillon I S.,Mallela S and Kumar R.,A Divisive information theoretic clustering algorithm for text classification,J.Mach.Learn Res.,3,pp 1265-1287,2003

[5] Hall M.A.,Correlation Based Feature Subset Selection for Machine Learning Proceedings of the Fifth IEEE international Confernce on Data Mining pp 581-554

[6] Hall M.A. and Smith L.A., Feature Selection for Machine Learning: Com- paring a Correlation-Based Filter Approach to the Wrapper, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.

[7] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.

[8] John G.H., Kohavi R. and Pfleger K.,Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[9] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif. Intell., 97(1-2), pp 273-324, 1997

[10] Koller D. and Sahami M., Toward optimal feature selection, In Proceedings of International

Conference on MachineLearning, pp 284-292, 1996.

[11] Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162, 2007.

[12] Press W.H., Flannery B.P., Teukolsky S.A. an Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge, 1988.

[13] Robnik-Sikonja M. and KononenkoI.Theoretical and empirical analysis of Relief and ReliefF, Machine Learning,53, pp 23-69, 2003.

[14] Van Dijk G. and Van Hulle M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference on Artificial Neural Networks, 2006

[15] Yu L. and Liu H., Feature selection for High – dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp856-863, 2003.

[16] Yu L. and Liu H., Redundancy based feature selection for micro array data, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004.

[17] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, 10(5), pp 12051224,2004.