JMIM: A Feature Selection Technique using Joint Mutual Information Maximization Approach

Saner Rajlakshmi Sanjay
Research Student
Department of Computer Engineering
K. K. Wagh Institute of Engineering Education & Research, Nashik, Maharashtra, India.

S. S. Sane, PhD
Vice Pricipal& Head of Department
Department of Computer Engineering,
K. K. Wagh Institute of Engineering Education & Research, Nashik, Maharashtra, India.

ABSTRACT
The process of feature selection is generally used to minimize the size of dataset, to overcome the problem of over fitting and to increase the classifier efficiency. We proposed the JMIM i.e. Joint Mutual Information Maximization algorithm to extract feature and for creation of feature subset efficiently. These algorithms are based on joint mutual information. It follows maximum of minimum strategy. In this paper our aim is to work on utilization of JMIM algorithm, then we compare upcoming outcome with the previously highlighted problems in existed feature selection system. In utilization of JMIM algorithm, we are expecting that our simultaneous processing of feature set selection process will reduces time required for overall execution. As a part of our contribution the process distributed over different clouds that helps in execution and triggers the process.

Keywords
mutual Information, feature selection, classification, joint mutual information, parallel computing.

1. INTRODUCTION
Feature selection strategy is nothing but a pre-processing step or it can be used in conjunction with machine learning algorithms used for classification and regression purpose. Feature selection is mainly classified in to 3 categories: Wrappers [4] are Wrappers uses induction learning algorithm and follows the black box for feature selection mechanism. Since the process of estimation of the 2N subsets proves beneficial for solving the problems like NP Hard, the establishment of the sub-optimal subset is carried out. This is done by retaining the search algorithms discovering the subset. This mechanism achieves better generalization but computation cost is very high if dataset is large. Embedded [5] method uses machine learning techniques such as bound on leave-one-out error of Support Vector Machine-SVM [6]. It is used to spreadinformationof specific formation of the class. This method is much slower than filter and selected features are dependent on learning machine. Filter [7] method use uses variable ranking techniques. Ranking approaches are used due to their easiness and moral achievement is described for applied solicitations. This technique is completely independent of learning machine and data. This method is robust against over fitting but may get fail to extract appropriate feature subset for classification and regression problems. We have investigated various techniques of identification of feature subset creation. We also identified the problem area where there is need a solution over identification of redundant and irrelevant features. In this paper, for feature selection we proposed JMIM algorithm and to analyze its performance in parallel computing environmentto overcome the problem of overestimation that occurs while employing the methods such as cumulative summation, while identifying whole correlation with one or many previously selected features, forward search to approximate the solution. Our system provides a mechanism for choosing the most relevant features in classification tasks. Our system works efficiently to improve efficiency in parallel computing that is required for analyzing high dimensional dataset and parallel identification of significance of candidate features. JMIM with parallel computing would perform better than JMIM without parallel computing. In implementation stage of our system, whole system is segmented in multiple models such as data discretization, feature subset selection, data classification and accuracy calculation. The huge data need to processed for feature set selection and hence workload distribution i.e. divide and conquer strategy is used. The technique named JMIM is used for identification of feature set as well as to evaluate the accuracy of feature set classifier is applied such as Naive Bayes classifier and 3-nearest neighbor. We use parallel computing to improve efficiency of the system. In this strategy data is distributed on 2 nodes and simultaneous processing is done on 2 different processors. The generated result from each node is then merged and output is generated.

2. RELATED WORK
In 1994, Roberto Battiti suggested [2], in neural networks, the concept of information theory is used to build learning algorithms along with analyzing the functionality of classifier. But, the presented method varies from the pruning techniques as observed in the learning stage. This is mainly because of the dimensionality function reduction which is executed before learning processes are executed.In machine learning, entropy is used in such a way that the mutual information introduces the compatible features for boolean formulas represented in a tree structure where the selection process is carried out on the features using the greedy techniques. Hence, Battiti suggests a practical approach to analyze the applicability of the MI in a practical way from supervised training of neural networks. Here mutual information measures the dependency between any random variables. Entropy, which is derived from machine learning, is an additional method to estimate during learning phase by using back propagationalgorithm. So, the benefit of using this algorithm is it eliminatesirrelevant features in the entire process. This consequently increases the performance and generates relevance feedback. Also, thedependencies of various features which are available for the neural classification are mentioned in the final results.

In 1995, James Dougherty et al. [3] suggested various supervised algorithms for machine learning. These algorithms demand a discrete feature space for generation of the results. The discrete feature space defines the techniques also

conducting the empirical evaluation of these techniques. Each feature has a discretized independent of other features and of algorithm's performance. The objective was to pursue wrapper methods that search through the space indicating the number of intervals per attribute. These methods are then compared by discarding the discretion technique with the entropy based techniques, i.e. the unsupervised methods over the supervised methods. It is observed that the performance improved significantly when the naive bayes algorithm along with the discretized features was used. In other words, the increase in induction algorithm performance is noticed when the features are discretized. If the performance never significantly degraded, it is capable of locally discretizing features.

In 2000 Anil K. Jain et al. [4] projected the following: Since a pattern can be in any form like a finger print, or any hand written word or it can even be a speech signal etc, it can be stated that pattern can be any entity which could be defined by a name. It is thus classified in two different tasks: 1] supervised and 2] unsupervised. These classes are defined be the system designer on the basis of which the patterns are classified. During this classification, various issues are to be taken care of. For example, the definition of classes that consist of patterns, the environment, representation of the pattern, extraction & selection of the feature, analysis of the cluster, the designing and learning of the classifier, etc. Pattern recognition is one of the fastest evolving fields in the area of technology. Hence, the task of maintaining the balance between this extraction and classification becomes a complicated task.

In 2003, Isabelle Guyon and et al. [5] introduced the variables and feature selection. This research work represents the gravity of the multiple issues that reveal the proportion of the relevance in variable and feature selection. Supervised learning focuses on the generation of results by treating the features more extensively, as compared to the unsupervised learning. This indirectly proves to be a drawback for supervised learning methodologies. Recently, the techniques which use the pragmatic point are seen improving the performance for variable and feature selection. Methodologies like (a) Sophisticated wrappers and (b) Embedded wrappers are preferred. These techniques improve the prediction results when differentiated with the basic ranking and correlation techniques. These methods are specially used to improve for testing on wide variety of data sets. Variable and feature selection are used to focus the particular areas where the variable datasets are accessible. These areas are namely the field in which the text documents are processed, the analysis of the gene expression is done, or the field where the combinatorial chemistry is involved. Hence, the presented mechanism defines three main modes for variable selection which are: (a) to improve the predictor performance (b) Quicker and proficient predictors and (c) A enhanced understanding towards the process along with the obtained data.

Again in the same year, Chris Ding et al. [6] suggested the following: a wide use of discriminant analysis technique has been observed in the field of bioinformatics. It is the technique of selecting one feature instead of using all available variables in the data. The advantages of such feature selection technique are (a) reduction in dimension for decrease in the computational cost (b) noise reduction improving the accuracy of classification. These are then characterized in gene expression microarray classes. Here, small genes are selected from the large sets of gene expression data sets for the classification accuracy of phenotypes. Out of the two basic methods that are used for the feature selection, i.e. filters and wrappers, filters are used. Since the working mechanism of filters consist of locating intrinsic characters that are maximum relevant. A broader representation of the relevance framework is used to display the relevance phenotype characteristics which are attained by using the usual ranking approaches.

In 2004, Francois Fleuret [7] came up with his another algorithm. The classification of feature selection methodologies is widely done into two main categories, i.e. filters and wrappers. The main objective behind this research work was to design a proficient filter for statistical, as well as computational tasks. Here, the main aim was to select some binary features from the vast features for the process of classification. Using conditional mutual information (CMI) algorithm, this method beats the other existing standard classification algorithms. Since Naive Bayesian classifier is used, the obtained error rate is similar to the boosting technique or the SVM technique. Hence, faster results are generated. This technique can further be developed for tuning the learnt structures for adapting themselves according the explicit complications.

Later in 2008, Ali El Akadi et al. [8] presented: The process of selecting features involves generating the results where the irrelevant and redundant features are also involved. These irrelevant and redundant features reduce the enactment of the classifier in both the terms i.e. prediction precision and rapidity. Here, pattern recognition consists of examining all the possible subsets and choosing a particular subset that satisfies the classification criterion. In this research work, an efficient feature selection technique is suggested which not only takes the mutual information into consideration but also the interaction between them. Here, SVM is used to differentiate and evaluate the performance of the various standard classifier algorithms. Thus, the main benefit of this algorithm is it uses the interaction between the features without affecting the computational complexity.

In 2008, Patrick Emmanuel Meyer et al. [9] also suggested their techniques: In the analysis of micro array data sets, huge features are to be characterized. This gradually increases the noise between the features, linear and non-linear dependencies etc., whereas the selection is made only from a smaller set of samples. This research work generates results by using the basic filtering method in order to attain higher efficiency during the selection of micro array data. This research uses the Double Input Symmetric Relevance (DISR) which considers the variable complementarity. But, in this process the Dispersion Sum Problem (DSP) is observed. In order to find a solution for this problem, sequential replacement and backward elimination process is used. Since, the computational cost of this method is typically high; it is made affordable by computing and storing the DISR matrix.

Ahead in 2009, Imran S. Bajwa1 et al. [10] proposed their algorithm: A feature selection technique like Principal Component Analysis (PCA) uses the features of image for image recognition. These features represent the image discretely. This methodology involves the extraction of the principal image features from an integrated class. Kernel Principal Component Analysis (KPCA) provides higher efficiency and accuracy with faster result generation capability. The efficiency of KPCA does not vary over the huge and complex data sets. There are different approaches to analyse image processing. It is used to classify the single

layered and multi-layered clouds. In weather forecasting applications, such type of technologies are used.

In 2010, Asha GowdaKaregowda et al. [11] presented their research: The main objective for selection of features using the unsupervised technique is to generate the smallest feature subset. This subset has to be capable of discovering the various clusters from the data that satisfy the selection criteria. On the other hand the selection process using the supervised technique focuses on maximizing the precision of the classification results. Selecting features for classification or for the supervised learning is way simpler than that for the clustering process. Class label information is used for the classification, whereas the the genetic algorithms are used by the wrapper methodology. It proves that there is no such standard wrapper which can be used for multiple data sets.

In 2010 again, Harold W. Kuhn [12] introduced a method which combines the linear duality elements and the graph theory tools. It very effectively describes how a matrix which consists of 0's and 1's can be used to solve the assignment problem. Though the Hungarian method is insignificant method for reducing the standard assignment issues, over mere the combination of a 0-1 matrix; this was the only known solution for the assignment problem till then.

Further in 2011, Hongrong Cheng et al. [13] suggested a new algorithm for selecting features. This algorithm, based on the CMI, uses the greedy approach for features selection called Conditional Mutual Feature Selector (CMFS). MIFS is an algorithm which is used for much classification application. They ignore feature synergy, MIFS. There are two main factors on which this suggested technique is focused, i.e. the information interaction and the CMI. On the basis of link between these two factors, the reaction synergy of features and redundancy are calculated. Thus the discriminative features are identified. It decreases the probability of mistaking important feature in searching process. Redundancy interaction of features and redundancy, both can be detected using the CMIFS algorithm. This is mainly because this algorithm avoids the evaluation of the FR information and prefers the FCR information calculation. Unnecessary features are removed and the main objective is extracting informative features in searching process. This proves to be its main advantage.

In 2012, Gavin Brown et al. [14] presented an efficient technique to crumble the conditional likelihood without disturbing the original interpretation. Using CMI for this purpose helps in implicit statistical assumption of mutual information criteria. Approximations are due to implicit assumptions on a data. Hence using this technique a probabilistic framework is generated which integrates these models. The stability, flexibility and accuracy with small samples of data are best achieved using the JMI. But the information theoretic interpretation for feature selection helps in understanding various nature of the algorithm, like the stability after few changes in the samples of data, the behaviour of the algorithm in limited and extremely small data samples etc. This interpretation is then termed as conditional likelihood optimization.

Later in 2012, VerónicaBolón-Canedo et al. [15] suggested their methods: In this research work, various synthetic samples of data are processed. The main objective behind this is studying the enactment of the selection techniques in the data sets which consists of irrelevant, noisy as well as smaller ratio features. We are very much aware of the three major and basic feature selection and evaluation techniques, namely wrappers, filters and embedded techniques. The most efficient method of these three can be judged on the basis of the accuracy of classification and the success index generated by using these techniques. The efficiency of these methodologies was tested with the four varying classifiers. The effectiveness of the generated results with all the four classifiers easily justifies the uniqueness of the original model. Effective feature selection not only reduces the machine learning complexity, but also increases the accuracy of the predictors. It reflects datasets such as micro-array data and does a commendable challenging task in the area of machine learning where selection of features and variables is very essential.

In 2013, Girish Chandrashekar et al.[16] discussed: The huge variable data sets which are chosen for feature selection result in generation of extremely high dimensional data samples. There are various feature selection techniques which can be used for better understanding of data in area of machine learning, applications for pattern recognition etc. The reduction in computational time, and increase in predictor performance is also observed. The main objective here is elimination of variables. Further this is used for machine learning purpose. Since filter methods are simple and are based on practical approach, they are used for selection of feature process is not only to define the data samples efficiently, but also to increase the prediction accuracy. In this, we also used classifiers i.e. SVM and RBF which is used for feature selection task. Feature selection algorithms can only be done using single dataset. Each underlying algorithm will act differently for different data. It shows that larger the data sample size, higher is the complexity level in machine learning. These algorithms provide various advantages like simplicity, stability, and increased accuracy in classification. In short, it is beneficial to use these methods because they give better understanding of the data samples, improved model for classification etc. It just successfully used for improving predictor performance and for fault prediction analysis of fault model data.

In the latter half of 2013, Cecille Freeman et al. [17] stated how classifier can be used in feature selection: Classification problem does exist in machine learning, but it also improves accuracy. Wrapper is one of the simplest methods of evaluating the feature set. Wrapper works in very simple manner where using the anticipated feature sample the classifier is trained and this classifier accuracy is defined as the suitability of the data sample. In this research work, particular classifiers are used to judge the performance of the feature subset estimation measure using the filters. Hence it is observed that the results vary according to the type of data and the classifier.

3. PROPOSED SYSTEM

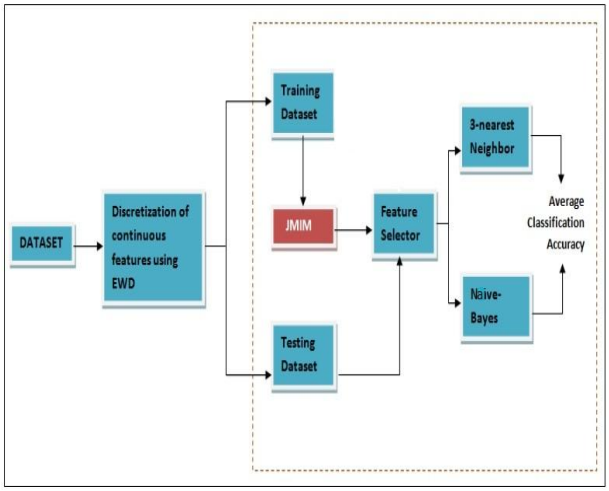

Figure 1: System Architecture

We designed a system for feature selection from dataset having multidimensional attributes and using JMIM feature selection algorithm finalization of candidate feature list is done. On this list classification is applied using Naive Bayes algorithm and 3-nearest neighbor algorithm. JMIM calculates the symmetrical relevance and the most relevant features. As a part of our contribution the process distributed over different clouds that helps in execution and triggers the process. We proposed a new method named JMIM, this method works on feature selection process. On this candidate feature list classifier is applied. Naive-Bayes algorithm and 3-nearest neighbor algorithm is applied for classification and final result is achieved. Dataset selected for this system must be multi-dimensional and variety of attributes that tests the system in all directions. Also over fitting problem is tested.

4. INFORMATION THEORY

In this section, the Information Theory calculates the mutual information and the entropy between the random variables. Entropy of any random variable is nothing but the impurity constraint. It is also defined as the uncertainty value of that variable. It is used to define the average information required for describing that variable. Entropy of any random variable X=(x1,x2,x3,…xN) is denoted by H(X) where xi refers to the possible values that X can take. H(X) is defined as:

$$H(X) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i)) \quad (1)$$

Where, p(xi) = probability mass function.

The value for p(xi) is:

$$p(x_i) = \frac{number\ of\ instants\ with\ value\ x_i}{total\ number\ of\ instants\ (N)} \quad (2)$$

The Joint Entropy of C and X is defined as:

$$H(X,C) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log(p(x_i, c_j)) \quad (3)$$

Where, p(xi, cj) is the joint probability mass function of the variable X and C.

The Conditional entropy of X given C is defined by:

$$H(C|X) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log(p(c_j|x_i)) \quad (4)$$

The relation between Joint Entropy and Conditional Entropy is given as:

$$H(X,C) = H(X) + H(C|X) \quad (5)$$

$$H(X,C) = H(C) + H(X|C) \quad (6)$$

Mutual Information is the amount of information that both variables share, defined as:

$$I(X;C) = H(C) - H(C|X) \quad (7)$$

We use the Mutual Information, since it reduces the amount of uncertainty of variable C. i.e. if the variables are statistically independent; the mutual information is calculated as zero.

Now since Mutual Information is symmentric,

$$I(X;C) = I(C;X) \quad (8)$$

$$I(X;C) = H(X) - H(X|C) \quad (9)$$

$$I(X;C) = H(X) + H(C) - H(X,C) \quad (10)$$

The Joint Mutual Information is defined as:

$$I(X;C|Y) = H(X|C) - H(X|C,Y) \quad (11)$$

$$I(X,Y;C) = I(X;C|Y) + I(Y;C) \quad (12)$$

The interaction between the variables can be termed as the amount of Mutual Information shared by those random variables. Hence, the interaction and the mutual information is defined as:

$$I(X;Y;C) = I(X,Y;C) - I(X;C) - I(Y;C) \quad (13)$$

Using only Joint Mutual Information allows overestimation, hence using this algorithm with the 'maximum of minimum' approach proves beneficial.

5. JOINT MUTUAL INFORMATION MAXIMIZATION

Feature Selection Process:

$$f_{JMIM} = \arg\ max_{f_i \in F-S}(min_{f_s \in S}(I(f_i, f_s; C))) \quad (14)$$

Where,

$$I(f_i, f_s; C) = I(f_s; C) + I(f_i; C/f_s) \quad (15)$$

$$I(f_i, f_s; C) = H(C) - H(C/f_i, f_s) \quad (16)$$

$$I(f_i, f_s; C) = \left[-\sum_{c \in C} p(c) \log (p(c)) \right] - \left[\sum_{c \in C} \sum_{f_i \in F-S} \sum_{f_s \in S} \log \left(\frac{p(f_i f_s, c/f_s)}{p(f_i/f_s) p(c/f_s)} \right) \right] \quad (17)$$

This method uses the following iterative forward greedy search algorithm to find relevant feature subset selection within the feature space.

Algorithm 1. Forward greedy search.

1. (Initialisation) Set $F \leftarrow$ "initial set of n features"; $S \leftarrow$ "empty set."
2. (Computation of the MI with the output class) For $\forall f_i \in F$ compute $I(C; f_i)$.
3. (Choice of the first feature) Find a feature f_i that maximises $I(C; f_i)$; set $F \leftarrow F \backslash \{f_i\}$; set $S \leftarrow \{f_i\}$.
4. (Greedy selection) Repeat until $|S| = k$: (Selection of the next feature) Choose the feature $f_i = \arg max_{f_i \in F-S} (min_{f_s \in S} (I(f_i, f_s; C)))$; set $F \leftarrow F \setminus \{f_i\}$; set $S \leftarrow S \cup \{f_i\}$.
5. (Output) Output the set S with the selected features.

6. MATHEMATICAL MODEL

The system S can be representing in terms of input, output, processing and terminals:

S= {I, O, F, T}

I= Set of Input= {I1, I2}

I1 = Raw input dataset file

I2 = feature count

O = Set of outputs= {O1, O2, O3, O4}

O1 = ARFF File

O2 = feature set

O3 = classification result

O4 = Accuracy evaluation

F = set of functions= {F1, F2, F3, F4, F5, F6, F7, F8, F9, F10}

F1 = Preprocessing of dataset

This preprocessing includes conversion of text _le T to arff file T! A.

Where, A is a matrix containing n attributes and m samples.

F2 = Discretization of n attributes to discrete values.

F3 = Symmetrical relevance Sr Calculation relevance based on Information I of attribute ai in set of attribute A.

F4 = Calculate most Mrrelevance based on Information I of attribute ai in set of attribute A.

F6 = JMIM calculation based on Mr.

F7 = greedy search to select single attribute ai from n attributes.

F8 = merge feature subset.

F9 = Update dataset with selected feature F in feature set.

F10= Naive Bayes classification.

F11 = 3 nearest neighbor classification for updated feature dataset.

F12 = precision P for accuracy evaluation based on classifier C evaluation.

F13=recall R for accuracy evaluation based on classifier C evaluation.

7. EXPERIMENTAL RESULTS

We have developed this system in java using jdk1.7. For distribution environment we have used Remote Method Invocation technique in java. To work with dataset we have used external weka-3.9 source library. The system is build using Net Beans 8.0.1 IDE. The system configuration is core I3 processor with 4 GB RAM.

2 datasets from the UCI Repository (Bache and Lichman, 2013) [18] are used in the experiment. These datasets have been previously used in similar research.

Data Set Description:

Data set	Number Of Features	Number Of instances	Number of classes
Sonar	60	208	2
Libra movement	90	360	15

Data set	Accuracy before Discretization	Accuracy afterDiscretization
Sonar	73.08%	87.02%
Libra movement	73.06%	75.00%

We have discretized our dataset and evaluate the classification accuracy. As we discretize the result accuracy of classification also increases. We have implemented Naive Bayes classifier.

After discretization we have evaluated NJMIM for different sizes of feature set and evaluated the accuracy using Naive Bayes classifier algorithm.

Data set	No. of features	Accuracy
Sonar	77%	79%
Libra movement	76%	78%

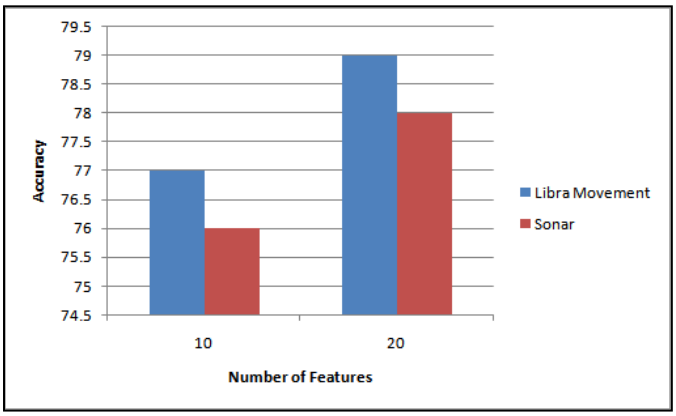

Figure 2:Percentile Classification Accuracy for different datasets with different feature set sizes

8. CONCLUSION

We proposed feature selection algorithms based on mutual information. We used JMIM algorithm for better feature extraction in parallel environment. We use Naive Bayes algorithm and 3-Nearest Neighbor algorithm for classification. We implement JMIM in distributed mechanism. To test the performance of proposed JMIM algorithm we used standard datasets. Hence, our system gradually tends to increase the efficiency and performance in the parallel computing environment.

9. REFERENCES

[1] Mohamed Bennasar, Yulia Hicks and RossitzaSetchi, Feature selection using Joint Mutual Information Maximisation, Expert Systems with Application, Volume 42, Issue 22, 1 December 2015.

[2] Roberto Battiti, Using Mutual Information for Selecting Features in Supervised Neural Net Learning, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 5, NO. 4, JULY 1994.

[3] James Dougherty, Ron Kohavi and MehranSahami, Supervised and unsupervised Discretization of Continuous Features, 1995.

[4] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 22, NO. 1, JANUARY 2000

[5] Isabelle Guyon and Andre Elissee_, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157-1182 2003.

[6] Chris Ding and Hanchuan Peng, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, Proceedings of the Computational Systems Bioinformatics (CSB03) 2003.

[7] Francois Fleuret, Fast Binary Feature Selection with Conditional Mutual Information, Journal of Machine Learning Research 5 (2004)15311555 2004.

[8] Ali El Akadi, Abdeljalil El Ouardighi, and DrissAboutajdine, A Powerful Feature Selection approach based on Mutual Information, IJCSNS International Journal of Computer Science and Network Security,VOL.8 No.4, April 2008.

[9] Patrick Emmanuel Meyer, Colas Schretter, and GianlucaBontempi, Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity, IEEE Journal Of Selected Topics In Signal Processing, Vol. 2, No. 3, June 2008.

[10] Patrick Emmanuel Meyer, Colas Schretter, and GianlucaBontempi, Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity, IEEE Journal Of Selected Topics In Signal Processing, Vol. 2, No. 3, June 2008.

[11] Asha GowdaKaregowda, M.A.Jayaram, and A.S. Manjunath,Feature Subset Selection Problem using Wrapper Approach in Supervised Learning, 2010 International Journal of Computer Applications (09758887).

[12] Harold W. Kuhn, The Hungarian Method for the Assignment Problem, 2010.

[13] Hongrong Cheng, Zhiguang Qin, Chaosheng Feng, Yong Wang, and Fagen Li, Conditional Mutual Information-Based Feature SelectionAnalyzing for Synergy and Redundancy, ETRI Journal, Volume 33,Number 2, April 2011.

[14] Gavin Brown, Adam Pocock, Ming-Jie Zhao and Mikel Lujan, Conditional Likelihood Maximisation: A Unifying Framework for Infor-mation Theoretic Feature Selection, Journal of Machine Learning Research 13 (2012) 27-66 2012.

[15] VernicaBoln-Canedo, Noelia Snchez-Maroo and AmparoAlonsoBetanzos, A review of feature selection methods on synthetic data,DOI 10.1007/s10115-012-0487-8 2013.

[16] Girish Chandrashekar, FeratSahin,A survey on feature selection methods, Computers and Electrical Engineering 40 (2014) 1628 2013.

[17] Cecille Freeman n, Dana Kuli, OtmanBasir, An Evaluation Of Classifier Specific Filter Measure Performance for Feature Selection,Pattern Recognition2014.

[18]https://archive.ics.uci.edu/ml/machine-learning-databases/