Received: July 25, 2017

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

JFIM: A Novel Filter Feature Selection approach Using Joint Feature Interaction Maximization

Sinciya Ponnupilla Omana[1*] **Jeyaa Celin Jesu[2]**

[1]*Noorul Islam University, Kumara coil, India*
[2]*Hindustan College of arts and Science, Chennai, India*
* Corresponding author's Email: sinciyapo@gmail.com

Abstract: As internet plays an important role in day to day life, data increases rapidly. Various applications involves in using this data, which is very complex and high dimensional. Applications requires accurate classification of these data. Data mining is the promising technology which provides various classification algorithm. The existing classifications techniques provides less accuracy in classifying high dimensional data. It also has the overhead of increased execution time. To meet these problems feature selection plays a significant role in classification. Classifiers efficiency is improved with effective feature selection and the selected features has high discriminative power. This paper proposes a novel mutual interaction based feature selection technique called Joint Feature Interaction Maximization (JFIM). The basic technique in this approach is forward greedy methodology which identifies the best feature subset. The resulting feature subset has less correlation between features and high correlation between multiple feature and classes. The resulting selected feature subset has less redundancy. The technique is implemented and tested with four well known bench mark dataset of varied dimensionality from UCI repository. A comparative study is made with two existing feature selection methods viz. Interaction Gain Feature Selection (IGFS) and Mutual Information Feature Selection (MIFS). The classification efficiency of our approach shows improved accuracy and reduced execution time.

Keywords: High dimensional data, Feature interaction maximization, Joint entropy, UCI repository, Information theory.

1. Introduction

Data mining technology is used in varied applications that involves high dimensional dataset. In all applications, features plays an important role in improving classification task. However, not all features are essential, as many of them are redundant or even irrelevant, resulting in reducing the performance of an algorithm. Thus classifiers can be improved by better feature selection methods. Feature selection aims to solve this problem by selecting only a small subset of relevant features from the original large set of features. By removing irrelevant and redundant features, feature selection can reduce the dimensionality of the data, speed up the learning process, simplify the learnt model, and/or increase the performance [1]. All feature selection methods find the optimal feature subset by considering an evaluation function and a search procedure. This evaluation function analyses how the selected subset can be useful in categorizing between classes. It can be classified into two main groups: filters and wrappers. Filter methods are independent of any classifier, which measure the relevancy and redundancy of feature subsets without any classifier support; whereas wrappers depends on the classification algorithm which uses the classifier's performance as the estimation measure.

Feature selection is considered as difficult not only because of the large search space, but also feature interaction problems. This work mainly highlights the information measure which is centered on the notion of mutual interaction. The existing feature selection methods are based on information theory such as information gain, Mutual

information feature selection and conditional mutual information feature selection which do not consider mutual interaction between features. The interactions can be two way, three way or multi way. Certainly, an important feature may be neglected considering the correlation between feature and target class, but the same feature could significantly improve the classification accuracy, when it is combined with other features. Unwanted removal of such features may miss an optimal feature subset and leads to poor classification performance. Thus in the proposed work, a novel joint feature interaction based evaluation function named JFIM (Joint Feature Interaction Maximization), that overwhelms the problems of the existing algorithm is suggested. In which the features which have high interaction information with previously selected features and high relevancy among the features is considered. The method calculates the interaction between the candidate feature and each feature in the subset which is already selected; the minimum interaction is added to the mutual information between the candidate feature and the class label; and the feature with maximum summation is selected and added to the subset S. It has the strongest relevance to the class label and the highest minimum interaction with the selected features. The advantage of this criterion is its ability to select the features that have the highest discrimination power.

The content of the paper is organized as follows. Section 2 describes Information theory concepts. Section 3 reviews literatures in filter feature selection methods. Section 4 gives the idea of feature interaction. Section 5 presents the proposed Joint Feature Interaction Maximization (JMIM) algorithm. Section 6 outlines the conducted experiments. Section 7 gives experimental results and discussion. Finally section 8 gives the conclusion.

2. Information theory

For a discrete random variable $X = (x_1, x_2 \ldots x_N)$, its entropy is denoted as $H(X)$, where x_i refers to the different values that X can take [2]. $H(X)$ is given as:

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

Where:
$p(x_i)$ is the probability mass function.

When the variable X takes discrete values, $p(x_i)$ is defined as

$$p(x_i) = {}^{X}/_{N} \quad (2)$$

X is the number of instances with value x_i
N is the total number of instances
$H(X)$ takes the value between 0 and 1.

The joint entropy for a two discrete random variable X and T is defined as:

$$H(X,Y) = \sum_{j=1}^{M}\sum_{i=1}^{N}(P(x_i, y_i)\log(p(x_i, y_i) \quad (3)$$

$P(x_i,y_i)$ is the joint probability mass function of the variables X and Y.

The conditional entropy of the variable X given Y *is* defined as:

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

The conditional entropy is the entropy left in C when variable X is considered; it gives either less or equal value to the entropy. Both conditional and joint entropy is bounded and it is given by:

$$H(X,Y) = H(X) + H({}^{Y}/_{X}) \quad (5)$$

3. Literature survey

A number of filter based feature selection methods are suggested in literature. All these methods aims to bring the best selective features. One of approach called Information gain (IG), which is considered as a simple method and it finds the dependence between feature and class. It follows a univariate approach. This method finds the best subset by evaluating mutual information with the class. Lots of application domains are attracted by this approach because of its simplicity and less computational cost. The major disadvantage faced by this method is independency among the features, which does not give better solution always. Another problem is certain features which are chosen may be highly correlated and carries repeated values with the class label. Another method called Joint Mutual Information (JMI), where A subset S is said to be the best feature subset, if there is a joint mutual information with subset of features in S and the class label, where S = {f1,f2,…..,f_k}. In addition to computational cost, JMI is impossible to calculate because of the size restriction on the number of instances necessary to compute the high dimensional probability function. A number of mutual information based on greedy algorithms has been suggested to resolve this problem. Instead of considering feature pairs, MIFS selects the features and added to the subset one by one [3]. The selected

feature is the one that satisfies the following goal function:

$$MIFS = \arg \max_{F \in F-S} (I(F_i ; C) - \beta \sum_{f_s \epsilon S} I(F_i ; F_S)) \quad (6)$$

Where F_i the candidate attribute and F_s is an attribute which is already selected and it belongs to the subset S. The mutual information between F_i and C denotes feature relevancy, and the mutual information between F_i and F_s represents redundancy; β is a user-defined variable for redundancy. MIFS selects the best subset of features when $0.5 \leq \beta \leq 1.0$.

Joint mutual information (JMI) is a variation of MIFS to maximize the cumulative summation of JMI with selected feature subset [4]. This algorithm tries to reduce feature redundancy for a great extent. The feature selection criterion as follows:

$$JMI = \sum_{f_i} (f_i, f_j, C) \quad (7)$$

MIFS-U [5] is an updated version of MIFS by altering the feature-feature interaction term.

$$MIFS - U = arg \max_{F \in F-S} (I(C; F_i) - \beta \sum_{f_s \in S} \frac{I(C, f_s)}{H(f_s)} I(F_i; F_s) \quad (8)$$

MRMR, [6] is also another modification of MIFS, which selects the features which satisfy minimum redundancy and maximum relevance.

$$mRMR = arg \max_{F \in F-S} -\frac{1}{s} \sum_{f_s \in S} I(F_i; F_S) \quad (9)$$

4. Feature interaction

Feature Subset selection is an efficient method to remove unwanted data in high dimensional data classification. For selecting the best subset of feature set, it needs a large search space (O (2 * K)), where K is the number of features) [7]. Researchers have given different estimations to find features which are relevant (e.g., feature relevance is evaluated by the complete correlation between distinct features and the class [8]. There are lot of confusions arises when considering correlation between feature and class. When a single feature is considered, the feature class correlation may be irrelevant but when the features are combined, it becomes very relevant.

However when the features are combined, the mutual interaction tells that the irrelevant combination of variables gives better performance in the classification task. The interaction between features is given as a heuristic test called interaction gain [9]. We can say that the information gain or mutual information between the variables is a two way interaction whereas mutual interaction is a three way interaction. It is given by:

$$I(F1; F2; C) = I(F1, F2; C) - I(F1; C) - I(F2; C) \quad (10)$$

So the interaction gain can be defined as the difference between actual decreases in entropy obtained by mutual information of combined features F1F2 and the expected decreases in entropy by considering individual features F1 and F2. The high value of interaction gain indicates that more information was achieved by combining features than the information gained by individual features. If there is a high amount of dependence among the features, the interaction gain is positive and if there is low amount of dependence among the features, interaction gain is negative. If the framework does not affect the dependence among the features then interaction is zero. Interaction gain is similar to mutual information between three random variables [10]. Most of the existing feature selection methods based on mutual information such as MIFS, MIFS-U, and DISR select the features based on relevancy and redundancy. These methods aim to maximize the relevancy and to minimize the redundancy. But the problem is redundancy term goes high when the features are increases. This is because it selects the features only by considering dependency between features and class. To avoid these problems, Joint mutual information, selects the features by considering the relevance of feature and class when the subset of features was selected. However, this method also over weights the importance of some features.

So a novel feature selection method called Joint feature interaction Maximization (JFIM) is suggested in our work. JFIM selects the best set of features based on the following new criterion:

$$JFIM = arg \max_{F \in F-S} (\min \sum_{f_s \in S} I(f_i; f_s; C)) \quad (11)$$

Where:

$$I(f_i; f_s; C) = I(f_i; f_s; C) - I(f_i; C) - I(f_s; C)$$

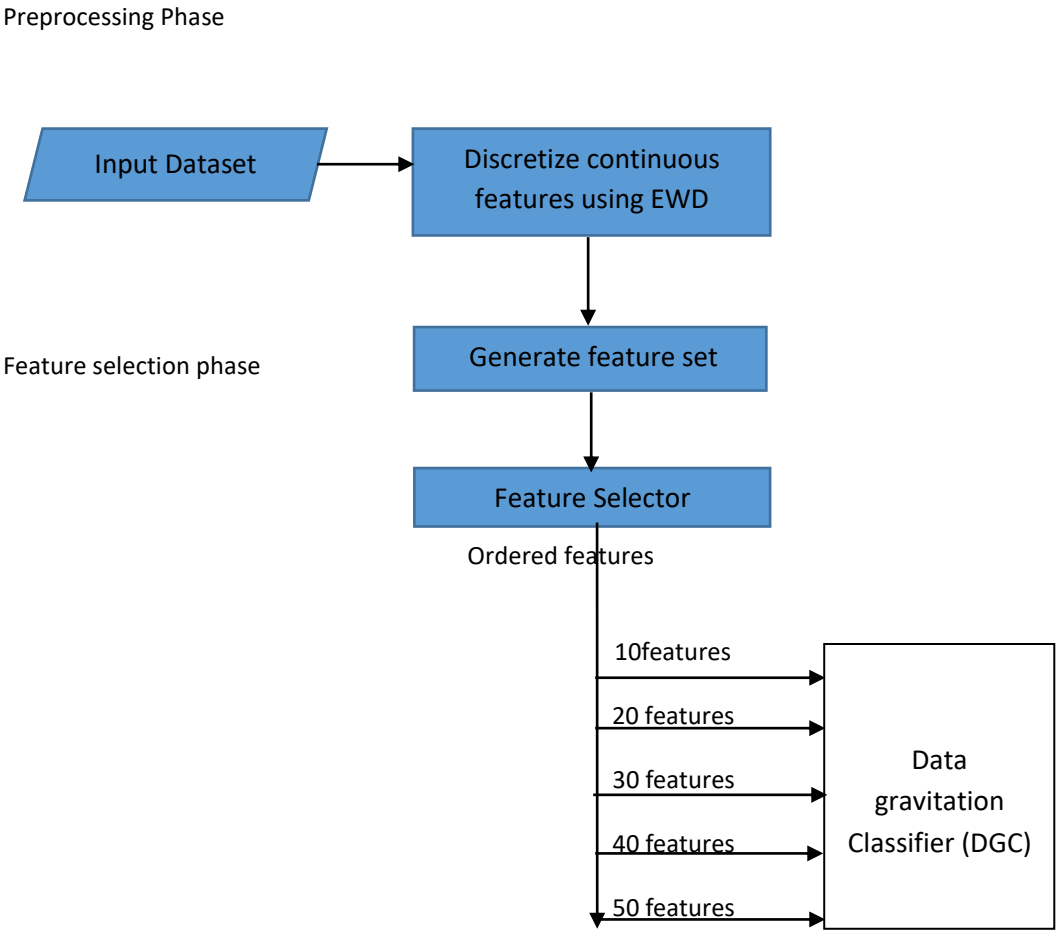

Figure.1 Framework for proposed feature selection algorithm

5. Proposed feature selection algorithm

The architecture of proposed feature selection framework is described in the Fig. 1. It includes 2 phases.

- Pre-processing phase
- Feature Selection phase

The features selected for classification may be binary, continuous or categorical. In pre-processing step, discretize the features if it is continuous or categorical using equal width discretization [11].

Algorithm: A novel MI based filter feature selection technique for classifying High Dimensional data

Step 1: Initialize three sets F, X and C.

Set $F \leftarrow$ Initial set of n features

Set $X \leftarrow$ Empty set for output

Set $C \leftarrow$ Set of class labels

Step 2: Compute the mutual information between features and the class labels for all features. For $\forall fi \in F$, Compute $MI(f_i; C)$

Step 3: Select the first feature f_i that maximizes $MI(f_i; C)$. Add that feature to the output set X.

Set $F \leftarrow F \setminus \{f_i\}$;

Set $X \leftarrow X \cup \{f_i\}$;

Step 4: Forward greedy selection: Repeat this step until the required number of features is selected.

i. Compute the mutual information between feature and the class labels for all feature pairs (f_i, fs). I.e. For $\forall fi \in F$ *and* $fs \in X$, Compute $IGFS(f_i; f_s)$

ii. (Selection of the next feature). Select the next feature that satisfies JSRM, and it is given by

$$JFIM = \arg\max_{F \epsilon F - S} (min \sum_{f_s \in S} I(f_i ; f_s; C))$$

Set $F \leftarrow F \setminus \{f_i\}$;

Set $X \leftarrow X \cup \{f_i\}$;

Step 5: Selected subset of the features is the output set X.

6. Experimental analysis

The performance of the proposed work is evaluated with two other feature selection algorithms such as IGFS and MIFS. These

algorithms are preferred for comparison because of the following three reasons:

a) Existing works stated that these algorithms give better accuracy as compare to other methods. [12, 13].
b) The proposed method takes the advantage of interaction information used by IGFS and the maximum and minimum approach.
c) This method also analyses interaction among the features instead of MI.

These methods are applicable to different application domains including transactional analysis, text data analysis, gene microarray and medical datasets with data of different dimensions including both high dimensions and low dimensions.

Data gravitation based classifier is used to estimate the quality of selected feature subsets. This classifier has been implemented in JCLEC software [14]. The accuracy results are used to assess the quality of selected subset of features. Tenfold cross validation is applied for both feature selection and feature validation. I.e. 90% of data is used for training while 10% is used for validation. This process is repeated 10 times [15].

6.1 Dataset description

Four datasets are collected from the UCI Origin [16] are used for experimentation are shown in (Table 1). Similar data sets are used in different literatures [17]. All the selected datasets are labelled in terms of number of features, instances and classes.

Table 1. UCI Datasets used in experimentation

Data set	No. of features	No. of instances	No. of classes
Sonar [10]	60	208	2
Musk[10, 12]	166	7074	2
Handwriting [10]	649	2000	10
Libra movement [10, 12]	90	483	15

Table 2. Evaluation results with Sonar dataset

Number of features	JFIM	IGFS	MIFS
10	83.1731	77.2872	72.3944
20	84.1346	78.3	74.7142
30	86.5385	79.0584	79.6503
40	85.5769	81.8327	83
50	86.0577	83.7985	85.64

7. Experimental results

In musk data set, both JFIM and IGFS gives almost same accuracy with 40 features.

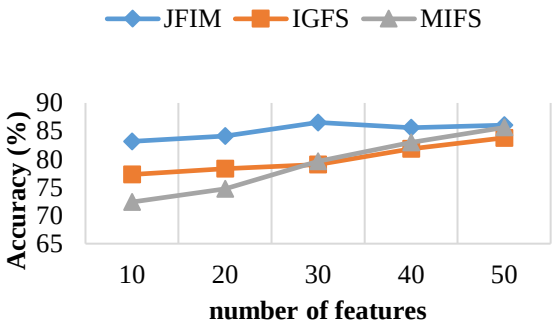

Figure. 2 Average classification accuracy on Sonar dataset

Table 3. Evaluation results with Libra dataset

Number of features	JFIM	IGFS	MIFS
10	60.8333	72.3743	76.5627
20	75	77.5234	78.6729
30	78.6111	77.6514	82.6545
40	84.1667	85.5829	84.8343
50	85.8333	84.3332	85.7694

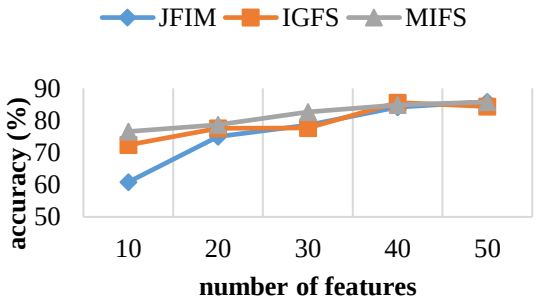

Figure. 3 Average classification accuracy on Libra movement dataset

Table 4. Evaluation results with Musk dataset

Number of features	JFIM	IGFS	MIFS
10	92.6493	87.2452	91.4114
20	93.61	91.3015	91.7922
30	94.7105	94.4561	92.7965
40	95.7566	95.8312	94.0583
50	95.6199	94.7213	95.1534

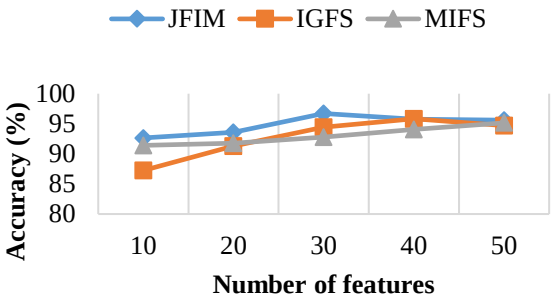

Figure. 4 Average classification accuracy on Musk dataset

Table 5. Evaluation results with Handwritten dataset

Number of features	JFIM	IGFS	MIFS
10	92.4983	93.5724	89.8614
20	93.8059	94.2115	92.7916
30	96.8653	95.5553	94.7445
40	96.6277	96.5912	95.0683
50	96.1459	96.3213	94.1522

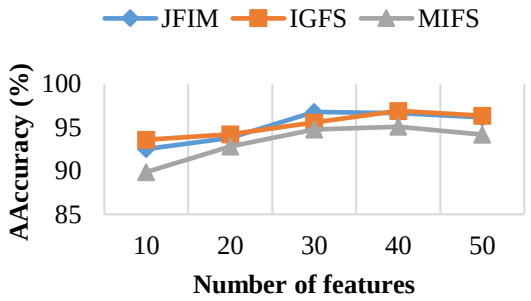

Figure. 5 Average classification accuracy on handwritten dataset

Table 6. Execution time for Sonar, Libra, Musk and Handwritten dataset.

Number of features	Sonar	Libra	Musk	Hand written
10	0.02	0.01	0.132	0.14
20	0.03	0.02	0.547	0.23
30	0.02	0.02	1	0.12
40	0.03	0.02	1.4	0.16
50	0.05	0.03	3.1	0.28
60	0.08	0.05	5.6	0.84

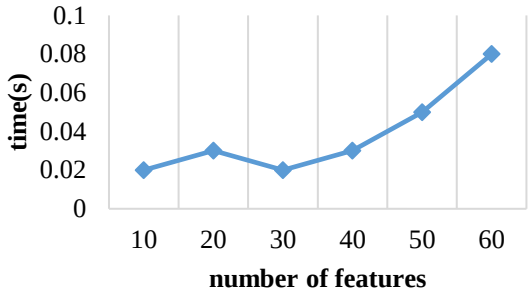

Figure. 6 Execution time on Sonar dataset

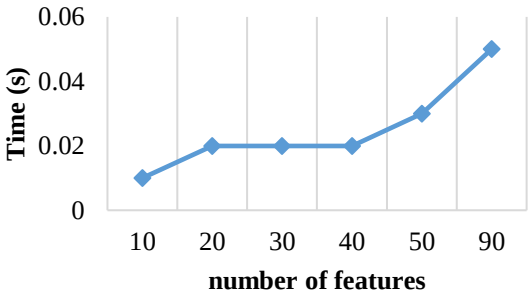

Figure. 7 Execution time on Libra movement dataset

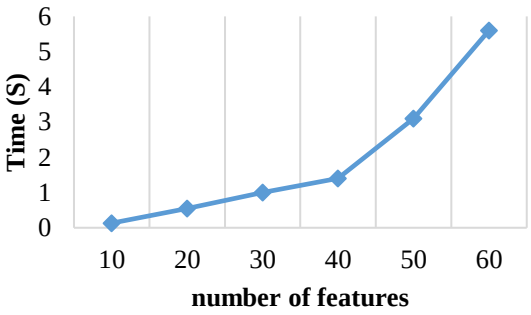

Figure. 8 Execution time on musk dataset

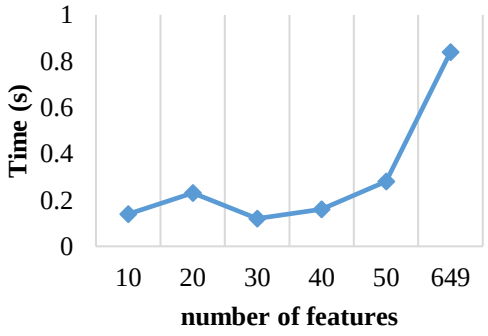

Figure. 9 Execution time on handwritten dataset

8. Discussion

Fig. 2 illustrates the average accuracy results on sonar dataset. It is designed by considering all the subset of selected dataset, ranges from one to 50 features. It is observed that, with 30 features it gives

highest classification accuracy of 86.5385%. The other two methods give the best accuracy with 50 features. When considering all the features of sonar data set, DGC classifier gives the accuracy of 86.5385%. But only 30 features JFIM achieve the same accuracy.

When applied to the Libra dataset (Fig. 3) under the similar backgrounds, The JFIM and IGFS perform better and gives the best accuracy of 85.8333% with 50 features. IGFS achieves this accuracy just with 40 features. With full set of features the classifier gives the accuracy of 80.9444%.

Fig. 4 demonstrates the average accuracy results attained on the Musk dataset. Our JFIM outclasses the others and gives the best classification results. With 40 features our proposed JFIM gives better results. Without feature subset construction, it gives the average accuracy of 95.044%. The other methods achieve these results with more number of features.

Fig 5. Shows best results of handwritten dataset when applying feature selection. With less number of features JFIM gives good results. There is a large variation in the results of with and without feature selection. With only 10 features it gives the accuracy of 92.6493%. When all the features are considered our DGC gives the accuracy results of only 85.203%. Here, JFIM gives the best accuracy with just 30 features. But IGFS gives the same accuracy with 40 features.

Fig 6 – 9 shows the execution time for all the four datasets sonar, libra movement, musk and handwritten. From all cases it is shown that with less number of features classification takes less time. When considering all the features, sonar takes execution time of 0.08 seconds, libra takes 0.05 seconds, musk takes 5.6 seconds and handwritten takes 0.84 seconds.

9. Conclusion

Classification of high dimensional data set is used in varied application which involve high retention of voluminous data set. Classifier efficiency is improved with better feature selection methods. This paper proposes a new feature selection method which selects a new set of features by considering the correlation between features and correlation between multiple features and classes. Here the features are ranked based on feature correlation criterion and then top ranked features are selected for evaluation. Four benchmark datasets from UCI are used for experimentation. This method is compared with other two other feature selection methods. For all datasets, JFIM gives better results with less number of features. The results show that the proposed Joint feature interaction maximization gives a better classification performance in terms of both accuracy and execution time. The proposed work has a high impact for classifying high dimensional data sets. Development of a novel classification method with JFIM is under process to apply in variety of application domains.

References

[1] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li., "Conditional mutual information based feature selection analysing for synergy and redundancy", *Electronics and Tele communications Research Institute*, Vol. 33, pp. 210-218, 2011.

[2] R. W. Yeung, "A new outlook on Shannon's information measures", *IEEE transactions on Information theory*, Vol. 37, No.12, pp.466-474, 1991.

[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning", *IEEE Transactions on Neural Networks*, Vol.5, No.8, pp.537–550, 1954.

[4] H. Yang and J. Moody, "Feature selection based on joint mutual information", In: *Proc. of the International ICSC Symposium on Advances in Intelligent Data Analysis*, pp.22–25, 1999.

[5] L. Yu and H. Liu, "Feature selection for high dimensional data: a fast correlation based filter solution", In: *Proc. of the Twentieth International Conference on Machine Learning*, San Francisco, CA, USA, 2003.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy", *IEEE transactions on Pattern analysis Mach. Intelligence*, Vol.27, No.19, pp.1226-1238, 2005.

[7] M. A. Hall, "Correlation based feature selection for discrete and numeric class machine learning", In: *Proc. of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, pp. 359-366, 2000.

[8] C. Freeman, D. Kuli´c, and O. Basir, "An evaluation of classifier specific filter measure performance for feature selection", *Pattern Recognition*, vol.48, No.6, 1812–1826, 2015.

[9] M. Bennasar, R. Setchi, and Y. Hicks, "Feature Interaction Maximization", *Pattern Recognition Letters*, Vol.34, No.11, 1630–1635, 2013.

[10] N. Kwak and C.H. Choi, "Input feature selection for classification problems", *IEEE Transactions on Neural Networks*, vol.13, No.13, pp. 143-159, 2002.
[11] W.J. McGill, "Multivariate information transmission", *Psychometrika*, Vol.19, No.2, pp. 97-116, 1954.
[12] V. Kumar and S. Minz, "Feature Selection: A literature Review", *Smart Computing Review*, Vol. 4, No. 3, pp.211-229, 2014.
[13] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximization", *Expert Systems with Applications*, Vol.42, 8520–8532, 2015.
[14] G. Brown. A. Pocock, M. Zhao, and M. Lujan, "Conditional like likelihood maximization: a unifying framework for information theoretic feature selection", *Journal of Machine Learning Research*, Vol.13, pp. 27–66, 2012.
[15] H. H Yang and J. Moody, "Data visualization and feature selection: new algorithms for non-Gaussian data", *Advances in Neural Information Processing Systems*, Vol.12, pp. 687-693, 2000.
[16] K. Bache and M. Lichman, UCI machine learning repository. Irvine, CA: University of California, *School of Information and Computer Science*. (http:// archive.ics.uci.edu/ml), 2013.
[17] M. F. Zaidi and B. Baharudin, "A Proposed Hybrid Approach for Feature Selection in Text Document Categorization", *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol.4, No.12, pp. 1799-1803, 2010.