# A Feature Selection Algorithm based on Mutual Information using Local Non-uniformity Correction Estimator

Ahmed I. Sharaf
Dept. of Computer Science
Computers & Informations Faculty
El-Mansoura University, Egypt

Mohamed Abu El-Soud
Dept. of Computer Science
Computers & Informations Faculty
El-Mansoura University, Egypt

Ibrahim El-Henawy
Dept. of Computer Science
Computers & Informations Faculty
Zagazig University, Egypt

*Abstract*—**Feature subset selection is an effective approach used to select a compact subset of features from the original set. This approach is used to remove irrelevant and redundant features from datasets. In this paper, a novel algorithm is proposed to select the best subset of features based on mutual information and local non-uniformity correction estimator. The proposed algorithm consists of three phases: in the first phase, a ranking function is used to measure the dependency and relevance among features. In the second phase, candidates with higher dependency and minimum redundancy are selected to participate in the optimal subset. In the last phase, the produced subset is refined using forward and backward wrapper filter to ensure its effectiveness. A UCI machine repository datasets are used for validation and testing. The performance of the proposed algorithm has been found very significant in terms of classification accuracy and time complexity.**

***Keywords*—*Feature subset selection; irrelevant features; mutual information; local non-uniformity correction***

## I. Introduction

In many applications of machine learning, the number of samples and dimensions of most datasets have grown rapidly [1]. Since the computational power, processing time and classification accuracy depend on the size of data therefore, reducing the dataset represents a challenge for researches. The primary motivation of reducing dimensions of data and minimizing the set of features is to decrease the training time and to enhance the classification accuracy of the algorithms [2], [3], [4]. Feature subset selection provides an approach for dimensions reduction and data minimization by replacing the original set of features with a compact subset that acts similar to the original one. This approach has been used in several applications in engineering, economy and medical sciences [5], [6], [7], [8].

Feature subset selection is categorized into two main approaches in terms of evaluation strategy [1]: First, the wrapper approach which depends on searching the whole search space to find the optimal subset [9]. This approach finds every combination of subsets to determine the accuracy by the classifier predication function. Thus, the quality of this subset is calculated without any modification of the learning algorithm. Since the produced subset is optimized for a particular classification algorithm therefore, the main advantage of the wrapper approach is the high accuracy. On the other hand, searching every combination consumes the computational power. The wrapper approach may also suffer from over-fitting to the learning algorithm. This drawback may also occur, when any parameter changes in the learning model [10].

Second, the filter approach depends on ranking each feature according to a specific evaluation function using distance, information and statistical measures. Many techniques have been proposed to calculate the feature relevance including: Fishers Discriminate Ratio [11], the Single Variable Classifier [12], Mutual Information [13], the Relief Algorithm [14], Rough Set Theory [15] and Data Envelopment Analysis [16]. The main advantage of the filter approaches is the computational efficiency and scalability in terms of the data dimensionality. Even though the filter approach is faster than the wrapper it suffers from lack of information between the features and the classifier. This approach may also select irrelevant or redundant features because of the limitation of the evaluation function [17].

Information theory [18] has been applied in many filter approaches to determine the relevance and redundancy of features. In feature subset selection process, mutual information is used to measure relevance and redundancy of features effectively. It has been applied by many researchers to characterize the information content of features [13], [19], [20]. The primary contribution of this research is to generate a compact feature subset with high accuracy and to keep the time complexity as minimum as possible. This paper is organized as: Section II introduces the related work and the limitations of the previous work. In Section III, the preliminaries and essential knowledge of information system, mutual information, conditional entropy and feature significance are discussed. In Section IV, the proposed algorithm is illustrated in detail. In Section V, the experiment and final results are presented. Finally, the paper is concluded in Section VI.

## II. Related Work

Authors proposed many approaches for the enhancement of feature subset selection using several methods. Mutual information was first proposed by Battiti, et al. [13] to improve the selection process by providing a novel algorithm called Mutual Information Feature Selection (MIFS). The MIFS used mutual information among features and between each feature and the decision class to determine the best $k$ features from the original set. The MIFS used the traditional greedy algorithm to select the optimal candidate set. The MIFS introduced the concept of relevance and redundancy using mutual information. Battiti proved that mutual information could be very useful for feature selection problems, and illustrated that his proposed MIFS is suitable for any classification issues. However, this method is not suitable for non-linear ones. Kwak and Choi [21] analyzed the work presented by Battiti and proposed an enhancement of the MIFS method. Kwak and Choi introduced MIFS-U that enhanced the estimation of information between input features and decision classes obtained from the MIFS. However, they neglected the behavior of the selected features together and focused on individual features.

Peng and Long [19] proposed a different method for solving feature selection problem based on min-redundancy and max-relevance mRMR. This method consists of two steps: in the first step, the best candidate elements are selected using the mRMR first order incremental criteria. In the second one, the wrapper filter is applied to search the obtained candidate set using backward and forward selections algorithms. However, this method searches the complete search space to find the compact subset of features which is a high computational cost. Therefore, it is necessary to reduce the search space with any reduction method or refine the candidate feature set.

The presented methods are all incremental methods that search for one feature at a time according to specific criteria. This strategy neglects the relationship among feature groups and could select one element to represent the group if it is better than the other candidates.

## III. Preliminaries

In this section, a brief introduction to information theory, basic principles and concepts are presented. An Information System $IS$ is defined as quadruple such that $IS = (U, A, V, f)$ where $U$ denotes a non-empty set contains the whole set of objects, $A$ denotes the finite non-empty set of features, $V$ represents the combination of all feature domains $V = \bigcup_{a\in A} V_a$ and $V_a$ is the domain of a specific feature $a \in A$, and $f$ represents the mapping function $f : U \times A \to V$ that produces a unique values of each feature with each object belongs to the universe. Let there is subset called $P$ such that $P \subseteq A$, then for each $P$ there is an associated indiscernible relation defined as $IND(P) = \{(u, v) \in U \times U \mid \forall a \in P, f(u, a) = f(v, a)\}$, it is clear that $IND(P)$ is an equivalence relation on the universe $U$ for $P \subseteq A$. The universe is divided into a various number of classes (granules) by this relations such that $U/IND(P) = \{[u]_p \mid u \in U\}$ where $[u]_p$ is the equivalence class calculated by $u$ with respect to subset $P$. For any given $P \subseteq A$, there is a binary relation called $SIM(P)$ that defined as follows $SIM(P) = \{(u, v) \in U \times U \mid \forall a \in P, f(u, a) = f(a, v)\}$. Let $S_p(u)$ is the maximal set of instances that possibly indistinguishable the universe $U$ by the set $P$ such that $S_p(u) = \{v \in U \mid (u, v) \in SIM(P)\}$. A member $S_p(u)$ from $U/SIM(P)$ is called an information granule. [22].

The information entropy among random variables is defined as the required amount of information to describe this variable $x$ [18], [23]. The entropy of a discrete random variable $X = (x_1, x_2, , x_n)$ is denoted as $H(X)$ and defined as follows:

$$H(X) = -\sum_{i=1}^{n} P(x_i)lg(P(x_i)) \tag{1}$$

Where $x_i$ represents the possible values of $x$ and $P(x_i)$ states for the probability of $x_i$. In the case of discrete random variable then:

$$P(x_i) = \frac{Number\_of\_instances(x_i)}{total\_number\_of\_instances} \tag{2}$$

The base of the used logarithm is two because the unit of measuring entropy are bits. For any two discrete random variables called $X$ and $Y$ with corresponding probability distribution $P(x; y)$. The conditional entropy is defined as:

$$H(X|Y) = -\sum_{x_i\in X}^{X_m}\sum_{y_i\in Y}^{Y_n} P(x, y)lg(P(x, y)) \tag{3}$$

The mutual information is defined as the amount of information that variable $X$ contains about variable $Y$ and is represented as follows:

$$I(X; Y) = \sum_{X_i\in X}^{X_m}\sum_{y_j\in Y}^{Y_n} P(x, y)lg\frac{P(x, y)}{P(x).P(y)} \tag{4}$$

The mutual information indicates the level of shared information between two random variables. Mutual information could be used to decrease computation by representing a relation between the entropy and the conditional entropy as follows:$I(X; Y) = H(X) - H(XY) = H(Y) - H(YX) = H(X)+H(Y)-H(X, Y)$. The high value of mutual information means that the two random variables are closely related to each other. Otherwise, if the mutual information value equals to zero, then the two variables are very independent of each other. Replacing $Y$ with $F_n$ and $D$ defines both the feature to class and the feature to features terms respectively. Although the mutual information is a stable measure of obtaining the uncertainty, it is not a monotonic function. Therefore, Dai, et. al [24] presented a monotonic mutual information measure for incomplete decision tables as follows:

$$H(D|B) = -\sum_{i=1}^{|U|}\sum_{j=1}^{m} \frac{|T_B(u_i)| \cap |Y_j|}{|U|} log\frac{|T_B(u_i) \cap |Y_j|}{|T_B \cap (u_i)|} \tag{5}$$

Dai, et. al proved that this new formulation is a monotonic function that could be used for measuring uncertainty effectively. Mutual information is also used to determine the significance of a specific feature $b_i \in B$ such that $B \subseteq C$ with respect to $D$ as follows:

$$sig(b_i, B, D) = I(D; B) - I(D; B - \{b_i\}) \tag{6}$$

The value of $sig(b_i, B, D)$ represents the change of mutual information if the feature $b_i$ is removed from the subset $B$. The higher value of the mutual information is, the more significant the feature is. If $sig(b_i, B, D) = 0$ then the feature $b_i$ is dispensable.

## IV. A Mutual Information based Uncertainty Measure

In this section, the ranking function is introduced to obtain the uncertainty of knowledge. The main properties and features are presented to illustrate the validity of the ranking function. In order to obtain the uncertainty for a target decision, measuring the feature dependency and the redundancy among features. Let $IS = (U, C \cup D)$ is a given information system, the uncertainty of knowledge is formulated as follows:

$$h(c) = \frac{\sum_{i=0}^{m} I(C, D_i)}{\sum_{i=0}^{m} I(C, D_i) + \sum_{j=0}^{n} I(C, C_j)} \tag{7}$$

Where, $I(C, D_i)$ and $I(C, C_j)$ represents the mutual information between the decision and a specific feature and the mutual information between a certain feature and the other features respectively.The proposed function $h(C)$ represents a relation between feature redundancy and decision dependency

**Property 1.** *Let $IS = (U, F \cup D)$ is an information system such that $U$ represents the all space of objects, $F$ is condition classes (features) set and $D$ is the decision set. For $\forall A, B \subseteq C$, if $A \subseteq B$ then $h(A) \leq (B)$.*

*Proof:* *Assume the universe of objects $U = \{x_1, x_2, ..., x_n\}$, the classification of $U$ induced by subset $A$ is $U/T_A(X) = \{T_A(X_1), T_A(X_2), ..., T_A(X_n)\}$, the classification induced by the set $B$ is $U/T_B(X) = \{T_B(X_1), T_B(X_2), ..., T_B(X_n)\}$ and the classification produced by the decision $D$ is $U/IND(D) = \{D_1, D_2, ...., D_m\}$. Since $A \subseteq B$ then the classification produced by the subset $B$ is better than the classification produced by the subset $A$ or simply $T_B(X) \subseteq T_A(X) i.e : |T_B(X)| \leq |T_A(X)|$. This equation is reformulated as follows: $T_A(X) = (T_A(X) \cap D_j) \cup (T_A(X) \cap (U - D_j))$ and similarly with subset $T_B(X) = (T_B(X) \cap D_j) \cup (T_B(X) \cap (U - D_j))$. From the inequality, we have $(T_B(X) \cap D_j) \leq (T_A(X) \cap D_j)$ and $(T_B(X) \cap (U - D_j)) \leq (T_A(X) \cap (U - D_j))$ respectively. Consequently from the monotonicity of $f(x, y) = -x log \frac{x}{x+y}$ [24], we obtain that $-\frac{|T_B(X) \cap D_i|}{|U|} log \frac{|T_B(X) \cap D_i|}{|T_B(X)|} \leq -\frac{|T_A(X) \cap D_i|}{|U|} log \frac{|T_A(X) \cap D_i|}{|T_A(X)|}$. Hence, $H(D|B) \leq H(D|A)$ according to definition (x). Substituting with equation(x) then $h(A) \leq h(B)$.* ■

**Property 2** (Maximum Value). *Let $IS = (U, C \cup D)$ is a given information system, the maximum value of $h(f)$ is one and occurs when $P(f, f_i) = P(f)P(f_i), \forall 0 < i < n - 1$ where $n$ is the number of features.*

**Property 3** (Minimum Value). *Let $IS = (U, C \cup D)$ is a given information system, the minimum value of $h(f)$ is zero and occurs when $P(f, D_i) = P(f)P(D_i), \forall 0 < i < m - 1$ where $m$ is the number of decision classes.*

### A. Feature selection algorithm

In feature selection process based on mutual information, the tolerance classes must be computed for the complete decision system. This process is an exponentially time consuming that affects the total time performance. In order to design an effective feature selection algorithm based on mutual information for a decision system, a fast algorithm for assembling granules from a given decision system is introduced initially. This algorithm is mainly based on decomposition and mutual information estimation. Computing the mutual information is a very complex task especially for large dimensions or samples. Determining mutual information using the traditional method is a time-consuming task with an exponentially time complexity $O(n^2)$. Therefore, a non-parametric mutual information estimator based on Local Non-uniformity Correction (LNC) is used [25]. The main idea of LNC based algorithm is to calculate an average correction term $\overline{LNC}$ for all each point $x_i \in X$. The correction term $\overline{LNC}$ is used to adapt the value of Kraskov estimated mutual information [26]. The correction term is computed based the volume of max-norm rectangle $V(i)$ produced by the PCA analysis of the $k_{th}$ nearest neighbors of each point $x_i$.

**Algorithm 1** Mutual Information Estimation using LNC

```
 1: Input: $X = \{x^1, x^2, ...x^m\}$ where $X$ is the sample of point, $d$ is the dimension size, $k$ is the k-nearest neighbor and $\alpha$ is a threshold.
 2: Output: $\hat{I}_{LNC}(X)$
 3: Calculate $\hat{I}_{KSG}(X)$ using KSG estimator with $k$ nearest neighbors.
 4: for all $x_i \in X$ do
 5:     Find the $k_{th}$ nearest neighbors of $x^i$ as $\{{knn_1}^i, {knn_2}^i, ..., {knn_k}^i\}$
 6:     Apply PCA on the $k_{th}$ nearest neighbors
 7:     Calculate the volume corrected rectangle $V(i)$, and volume of max-norm rectangle $\overline{V(i)}$
 8:     if $\overline{V(i)}/V(i) \leq \alpha$ then
 9:         $LNC_i = log(\overline{V(i)}/V(i))$
10:     else
11:         $LNC_i = 0$
12:     end if
13: end for
14: Calculate $\overline{LNC} = \sum_{i=1}^{m} LNC_i/m$
15: $\hat{I}_{LNC}(X) = \hat{I}_{KSG}(X) - \overline{LNC}$
16: return $\hat{I}_{LNC}(X)$
```

The LNC estimator works typically for any dimensions $d$, for example to compute the mutual information $I(X;Y)$ using the LNC algorithm. Let the dimension parameter $d = 2$, the input array equals to $X = [[x_1, x_2, \ldots], [y_1, y_2, \ldots]]$, the $K_{th}$ nearest neighbors $k = 3$ and threshold $\alpha = 0.25$. The time complexity of Algorithm 1 is determined as follows: for step(3) it is $O(Nk + d)$ which can be approximated to $O(N)$ since both $k$ and $d$ are relatively less than $N$. For step (5) it is $O(k)$, for step (6) it is $O(k^3)$, and the complexity of steps (7-12) are $O(1)$. Then the overall complexity becomes $O(N).(O(k) + O(k^3) + O(1))$ that yields to $O(Nk) \approx O(N)$.

### B. Proposed Method

In this section, the feature subset selection is proposed in detail.

The proposed algorithm consists of three main blocks. In the first block from step(1) to step(12), the proposed measure

**Algorithm 2** A Hybrid Mutual Information based Feature Selection Algorithm

```
Input: $IS = (U, C \cup D)$.
Output: $Red$ The reduced subset.
 1: Calculate $h(c_i), \forall c_i \in C$ using Algorithm 1.
 2: Sort the features in descending order using radix sorting
    and then denote the result by $S = \{a_1, a_2, ...a_n\}$
 3: Calculate $U/SIM(C)$.
 4: for all $a_i \in S$ do
 5:     Calculate $U/SIM(C - \{a_i\})$.
 6:     Calculate $sig(a_i, C, D)$
 7:     if $(sig(b_i, C, D) == 0)$ then
 8:         $\bar{R} = \bar{R} \cup \{a_i\}$
 9:     else
10:         $R = R \cup \{a_i\}$
11:     end if
12: end for
13: Let $R' = R$
14: Calculate $U/SIM(R')$.
15: Construct an input sequence subset $P = \{a_1, a_2, ...a_k\}$
    such that $k = |R'|$ and $k \leq n$.
16: for all $a_i \in P$ do
17:     Calculate $U/SIM(R \cup \{a_i\})$.
18:     Calculate $sig(a_i, R, D)$
19:     if $(sig(a_i, R, D) \neq 0)$ then
20:         $R' = R' \cup \{a_i\}$
21:     end if
22: end for
23: Let $Red = R'$
24: Construct an input sequence subset $Q = \{a_1, a_2, ...a_l\}$
    such that $l = |R'|$ and $l \leq n$.
25: for all $a_i \in R'$ do
26:     Calculate $U/SIM(R' - \{a_i\})$.
27:     Calculate $sig(a_i, R', D)$
28:     if $(sig(a_i, R', D) \neq 0)$ then
29:         $Red = Red - \{a_i\}$
30:     end if
31: end for
32: return $Red$
```

is computed for each feature according to Eq.(7). Since the computation of mutual information is a time expensive process, an effective estimator is used to calculate this formula based on Local Non-uniformity Correction (LNC) and KSG estimator. Afterwards, radix sort is used to sort the features according to the value of $h(c_i)$ in a descending order. The radix sort is used to minimize the total computational cost as it is a linear time sorting. Then, the granules of information are obtained using Pawlak definitions with respect to the complete set of features [27]. Then, the significant of each feature is computed to construct an initial subset of features. If the significance of a feature equals to zero then its considered to be an irrelevant feature. Otherwise, the relevant feature is added to the optimal subset of features called $R$. In the second block from step(13) to step(22), the obtained subset $R$ is refined using forward wrapper filter. The granule of information is calculated to determine the significance of the non-participated features. Each feature is joined to the generated subset $R$ to study it's significant. Once the feature is considered a significant one with respect to $R$, then it should be merged to $R$. The last block from step(23) to step(32) is a backward wrapper, which makes it similar to the second block but with reverse effect.

The complexity analysis of the proposed algorithm is determined as follow: the time complexity of step(1) is $O(n|U|)$. For step(2), the radix sort is applied with a complexity of $O(n)$. For step(3), the time complexity is $O(n|U|)$. From step(4) to step(12) the complexity is $O(n) \times (O(|U|) + O(1))$ which is approximated to $O(n|U|)$. From step(13) to step(22), a forward wrapper filter is applied to determine the effect of any irrelevant feature to granules obtained by the classification of subset $R$. The complexity of the forward wrapper is $O(k|U|)$ where $k \leq n$. Then, a backward wrapper is used to refine the generated subset with a complexity of $O(l|U|)$ where $l \leq n$. Hence, to total complexity of proposed algorithm is $O(n|U|) + O(n) + O(n|U|) + O(k|U|) + O(l|U|)$ which approximately equals to $O(n|U|)$.

## V. Experimental Results and Discussion

In this section, datasets description, numeric results and comparative studies are presented.

### A. Dataset Description

Five datasets are used to benchmark and evaluate the proposed approach. These public datasets are used for benchmarking and validation of selection algorithms. A brief discussion of the used datasets is listed in Table I.

TABLE I. Description of Benchmarking Dataset

| Dataset Name | Features Count | Instances Count |
|---|---|---|
| Breast Cancer Wisconsin | 9 | 699 |
| Glass | 9 | 214 |
| Ionosphere | 34 | 351 |
| Iris | 4 | 150 |
| Liver Disorder | 6 | 345 |

The iris dataset is a very popular benchmarking dataset that contains information about iris plant. The Iris dataset contains three classes of 50 instances per class represents an iris category (Setosa, Versicolour, and Virginica). Also, the feature set includes a sepal length, sepal width, petal length, and petal width. The second dataset contains experimental information about breast cancer. Dr. William H. Wolberg collected this dataset from the University of Wisconsin Hospitals, Madison. There are nine features in this dataset out of ten features excluding the sample identifier. Each instance could be classified in the binary decision (benign or malignant). The missing values are replaced with the average value of the corresponding feature in order to prevent exceptions. The Liver Disorder dataset that contains blood measures tests. This dataset contains six features and 345 instances. Each instance could be classified into the binary decision. Finally, the Glass dataset that includes nine features and 214 instances with seven decision classes per instance. The numeric experiment is implemented using Python Scikit-learn [28] package. All comparative studies with the other methodologies are implemented on WEKA [29] software. The experiment is executed on an Intel(R) Core (TM) i7-2400 CPU 3.10GHz platform and MS Windows 10 installed.

### B. Numeric Results And Comparative Studies

In this section, the numeric results of the experiment are presented. The proposed method is implemented on five datasets as described in Table I and achieved a high accuracy over the other methods. The Naive Bayes classifier is used to determine the accuracy of the proposed algorithm. The comparison is based on some standard methods such as Information Gain, Gain Ratio, Chi-Square, Best First Approach and Symmetrical Uncertainty. Also, the mRMR, MIFS, MIFS-ND and MIFS-U are also used for the ionosphere dataset. For the breast cancer dataset, the proposed method achieved the highest accuracy among the other methods as shown in Fig. 1. The minimum accuracy achieved by symmetric uncertainty method, then the best first method. All of the chi-squares, gain ratio and information achieved the same accuracy. The minimum scored accuracy is $91.04\%$ and the maximum accuracy is $92.09\%$.

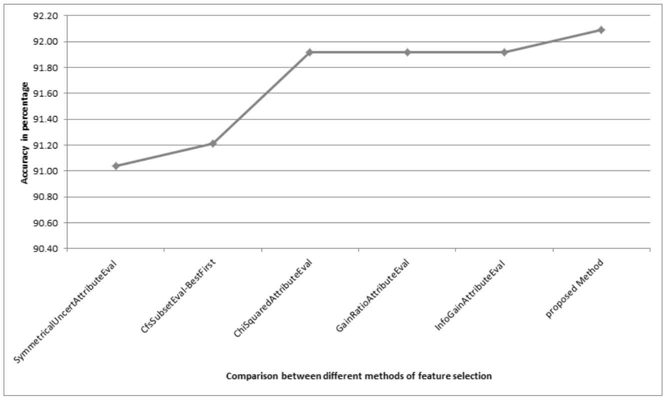

Fig. 1. A comparison between different methods of feature selection for breast cancer dataset.

For the glass dataset, the proposed method archived the highest accuracy as shown in Fig. 2. The minimum accuracy achieved by chi-square, gain ratio, information gain and symmetric uncertainty. The best first feature selection scored $49.53\%$ accuracy and the proposed method scored $54.67\%$.

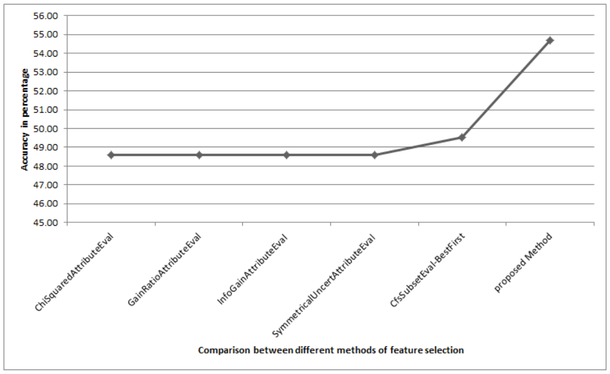

Fig. 2. A comparison between different methods of feature selection for the glass dataset.

For the ionosphere dataset, the proposed method achieved the best accuracy over the other methods as shown in Fig. 3. The MIFS-U achieved the best accuracy for only three features. The MIFS-ND scored the best accuracy over the all other methods. The proposed method scored the best accuracy from the other methods for the both the ten and fifteen features. The minim accuracy achieved for 15 features is $92\%$ by MIFS, and the maximum accuracy is $94.3\%$ by the proposed method.

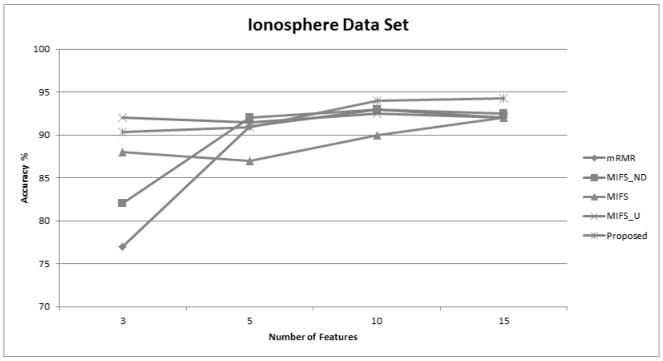

Fig. 3. A comparison between different methods of feature selection for the ionosphere dataset.

For the liver disorder dataset, the proposed method scored the best accuracy as shown in Fig. 4. The minimum accuracy achieved by most of the selections methods (chi-square, gain ratio, info gain and symmetric uncertainty). Then the best first scored accuracy of $58.55\%$. The best scored accuracy $58.84\%$ achieved by the proposed method.

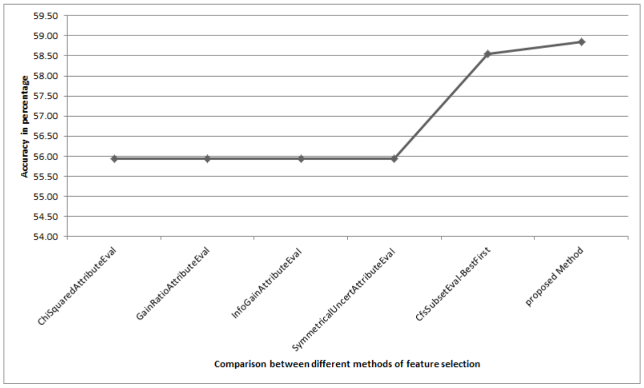

Fig. 4. A comparison between different methods of feature selection for the liver disorder dataset.

## VI. Conclusion

In this paper, a feature subset selection algorithm is proposed based on mutual information and LNC estimator. Although information theory has used before for solving feature selection. The proposed method ranks the features according to $h(c)$ that represents the relation between features redundancy and decision dependency. A Mutual information estimator based on LNC is used to reduce the overall time complexity. Five dataset from the UCI machine repository are used for testing and validating the proposed algorithm. Naive Bayes classifier is used to compare the obtained feature subsets. Final results are compared to Information Gain, Chi-square, Best first methods. The results of Ionosphere datasets are compared to mRMR, MIFS, MIFS-ND and MIFS-U. The obtained results from the comparative study illustrate the efficiency and accuracy of the proposed method. Reducing total time complexity of mutual information algorithms is also achieved.

## References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16 – 28, 2014.

[2] Y. Zhang, C. Yang, A. Yang, C. Xiong, X. Zhou, and Z. Zhang, "Feature selection for classification with class-separability strategy and data envelopment analysis," *Neurocomputing*, vol. 166, pp. 172–184, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215004609

[3] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, Conference Proceedings, pp. 856–863.

[4] A. Arauzo-Azofra, J. L. Aznarte, and J. M. Bentez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8170–8177, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741001523X

[5] G. M. U. Din and A. K. Marnerides, "Short term power load forecasting using deep neural networks," in *Computing, Networking and Communications (ICNC), 2017 International Conference on*. IEEE, 2017, pp. 594–598.

[6] Y. Zhang, D.-w. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 1, pp. 64–75, 2017.

[7] J. Yu, J. Yu, A. A. Almal, S. M. Dhanasekaran, D. Ghosh, W. P. Worzel, and A. M. Chinnaiyan, "Feature selection and molecular classification of cancer using genetic programming," *Neoplasia*, vol. 9, no. 4, pp. 292–IN3, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1476558607800996

[8] S. Sasikala, S. A. a. Balamurugan, and S. Geetha, "A novel feature selection technique for improved survivability diagnosis of breast cancer," *Procedia Computer Science*, vol. 50, pp. 16–23, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915005062

[9] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[10] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, no. Jan, pp. 27–66, 2012.

[11] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fishers discriminant ratio for text sentiment classification," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[13] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on*, vol. 5, no. 4, pp. 537–550, 1994.

[14] H. Zheng and Y. Zhang, "Feature selection for high-dimensional data in astronomy," *Advances in Space Research*, vol. 41, no. 12, pp. 1960–1964, 2008.

[15] J. Liang, F. Wang, C. Dang, and Y. Qian, "A group incremental approach to feature selection applying rough set technique," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 294–308, 2014.

[16] Y. Zhang, A. Yang, C. Xiong, T. Wang, and Z. Zhang, "Feature selection using data envelopment analysis," *Knowledge-Based Systems*, vol. 64, pp. 70–80, 2014.

[17] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[18] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[19] H. Peng, L. Fulmi, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[20] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[21] N. Kwak and C. Chong-Ho, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[22] Q. Zhang, Q. Xie, and G. Wang, "A survey on rough set theory and its applications," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 4, pp. 323–333, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2468232216300786

[23] R. Togneri and J. Christopher, *Fundamentals of information theory and coding design*. CRC Press, 2003.

[24] J. Dai, W. Wang, Q. Xu, and H. Tian, "Uncertainty measurement for interval-valued decision systems based on extended conditional entropy," *Knowledge-Based Systems*, vol. 27, pp. 443–450, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705111002334

[25] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables." in *AISTATS*, 2015.

[26] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.69.066138

[27] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982. [Online]. Available: http://dx.doi.org/10.1007/BF01001956

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278