Feature Selection and Discretization based on Mutual Information

Sadia Sharmin[1], Amin Ahsan Ali[2], Muhammad Asif Hossain Khan[2], Mohammad Shoyaib[1]
[1]Institute of Information Technology, [2]Computer Science and Engineering
University of Dhaka, Dhaka, Bangladesh
bit0426[at]iit.du.ac.bd, {aminali, asif, shoyaib}[at]du.ac.bd

***Abstract*—Feature selection and discretization have been considered to be an important research topic in the field of pattern recognition and data mining. However, addressing both these issues at a time is rarely discussed in the existing research. In this paper, these issues have been addressed by developing a heuristic namely discretization and selection of features based on mutual information (DSM). Experimental results on 15 datasets show that the proposed DSM outperforms a number of state-of-the-art feature selection or discretization algorithms. On average, its accuracy surpasses that of the best performing state-of-the-art algorithms by 5% using Support Vector Machine. Moreover, for datasets with a large number of features, it shows promising accuracies even with far less number of features than the other competing algorithms.**

Keywords-Feature selection; discretization; relevance; redundancy; complementarity

I. INTRODUCTION

Feature selection methods have been explored considerably in both machine learning and data mining. It facilitates data visualization, data understanding and most importantly improves classification accuracy [1]. In most of the cases, discretization is necessary prior to the feature selection process. Furthermore, it is necessary to add discretization (even if the data is already discretized) to simplify the data, to accelerate the learning process as well as to reduce the noise in data [2].

In feature selection, the goal is to find a subset S of a given feature set F such that S contains most relevant features while discarding the irrelevant and redundant ones. It is also necessary to include a feature f_i to an already selected feature set S, if it contains complementary information though it may be redundant. Therefore, to select a feature f_i the following three conditions should be considered –maximize the relevance with class variable C, minimize the redundancy and maximize the complementary information among the selected features S [3], [4]. From now this property will be referred to as RrC.

On the other hand, in feature discretization, we have to discretize a given feature f_i into n discrete intervals $\{[d_0, d_1], [d_1, d_2], \dots [d_{n-1}, d_n]\}$ where d_0 and d_n are the minimal and the maximal values of feature f_i. To get higher classification accuracy, such a discretization process should come up with a minimum number of intervals for which the feature f_i shows the best relevance with C and thus, represents the given data with minimum information loss. Furthermore, all these intervals must be adjusted for all features such that the RrC property is attained.

If the problem of feature selection and discretization is to be solved simultaneously, the RrC property should be achieved for both cases. However, it is well-known that feature selection is an NP-hard problem [5] and finding optimal discretization is NP-complete [6]. In this paper, a heuristic algorithm has been proposed that attempts to solve these two problems simultaneously while maintaining a low computational complexity with better performance in comparison to other state-of-the-art algorithms.

Different criteria have been proposed in the literature to assess the quality of a feature selection and discretization algorithms. Such as, distance [7], dependency [8], consistency [9], mutual information (MI) [10], [11], [12] etc. are used for feature selection. Among them, mutual information is widely used as it can measure the dependencies between random variables as well as assess the "information content" of features in complex classification tasks [10]. A more detailed discussion on this topic can be found in [5]. For discretization, different heuristics such as entropy [13], likelihood [14], [15], rough sets [16], [17], MI [36] and statistical χ^2 test [18], [19], [20] are commonly used. Beside these MI is also used for wide variety of applications such as [43], [44].

Feature selection methods can be categorized into two types based on their evaluation policy - classifier dependent (Wrapper and Embedded) and classifier independent (Filter) methods [21]. The examples of wrapper based methods are [40], [41], [42] etc. As wrapper methods utilize a classifier, they may become biased to the selected classifiers. Also, these methods are prone to over-fit the data [3] especially when the amount of data is insufficient. Conversely, filter methods are better as they are independent of classification algorithms. Among the filter methods, MI based methods have gained popularity [21].

One of the major criteria of MI-based feature selection methods is to maximize the max-dependence, i.e., maximize the multidimensional (joint) MI between the features and the class. This is a hard problem to solve [11]. Several approximations are found in the literature that approximate the max-dependence with only relevance [22] or with relevance and redundancy of the features [10], [11]. In spite of having redundancy, some features provide complementary information about the target class. The importance of complementary information has been discussed in [4], [23], [12], [24].

978-1-5090-6004-7/17/$31.00 ©2017 IEEE

Therefore, current state-of-the-art algorithms suggest the use of RrC property for feature selection. However, the value (score) calculated using RrC property usually increases with the addition of a new feature which results in a large number of features being selected. This issue has not been adequately addressed in the previous work.

The major goals of feature discretization methods (e.g., ChiMerge [18], class-attribute interdependence maximization (CAIM) [25] and Modified Chi2 [20]) are to keep the number of intervals as few as possible to lower the inconsistency rate and to increase the classification accuracy while minimizing the computation time to find the best discretization intervals [2]. A satisfactory trade-off between the number of intervals and the achieved accuracy have been demonstrated by the methods such as Chi2 [19], Minimum Description Length Principle (MDLP) [13] [2] etc. Among all the discretizers, CAIM generates the smallest number of intervals for a given continuous feature [26].

All the aforementioned methods are static that do not take into consideration the interdependence among the features while discretizing them. The most popular dynamic discretizers are ID3 [27], ITFP [28], A* [29]. However, they are usually dependent on the learning algorithms. That is why most popular discretization methods are static in nature [2].

To address these issues an algorithm namely Discretization and Feature Selection based on Mutual Information (DSM) has been proposed in this paper that satisfies the RrC criteria where

- The discretization method is global, supervised, dynamic, and top-down in nature and
- The feature selection is filter based that selects small number of necessary features with higher accuracies.

The overall contribution of the paper is thus as follows:

- First, discretization and feature selection is addressed jointly.
- Second, an incremental gain strategy is introduced to select a small number of features that are highly informative.
- Third, a dynamic discretization method is proposed that is independent of the classification algorithm.

II. Literature Review

There has been a significant amount of work on feature selection and discretization, but these works mainly address these two problems separately. Therefore, in this section these two problems are discussed individually.

A. Feature Selection

For selecting a subset of features S from a given set of feature $F = \{f_1, f_2, \dots, f_n\}$, Battiti [10] proposed to maximize the joint mutual information $I(S; C)$, which is defined in (1).

$$I(S;C) = I\{f_1, f_2, \dots, f_k ; C\}$$
$$= \sum_{f_1,\dots,f_k} \sum_C P(f_1, f_2, \dots, f_k ; C) \log \frac{P(f_1, f_2, \dots, f_k ; C)}{P(f_1, f_2, \dots, f_k) P(C)} \quad (1)$$

Here, $I(S; C)$ represents the largest dependency with the target class C. However, it is not feasible to compute $I(S; C)$ directly and can be considered to be an NP-hard problem [5]. To address this problem, Battiti introduces an approximation of $I(S; C)$ in mutual information based feature selection (MIFS) [10] under the assumption that the features are pairwise class-conditionally independent. It is a greedy selection algorithm which adds one feature to S at a time. The first feature is added to S by considering the highest relevance with C and the next feature $f_i \in F$ is chosen when it maximizes (2).

$$I(f_i; C) - \beta \sum_{f_s \in S} I(f_i; f_s) \quad (2)$$

Here, $I(f_i; C)$ represents the relevance between f_i and C while $I(f_i; f_s)$ denotes the redundancy between two individual features f_i and f_s. In (2), β is a user–defined parameter. The redundancy in (2) grows in magnitude with respect to the relevancy with the increase of number of selected features. Thus, the relevance becomes insignificant compared to the redundancy and may select irrelevant features [30]. Peng *et al.* [11] proposed a feature selection framework called minimal-redundancy-maximal-relevance (mRMR) which overcomes this limitation of MIFS partially. They set the average number of the selected feature as β assuming that the features are pairwise independent which is shown in (3).

$$I(f_i; C) - \frac{1}{|S|} \sum_{f_s \in S} I(f_i; f_s) \quad (3)$$

This sort of approximation is also validated in [37] under the assumptions of lower-order dependency between the features.

It should be noted that normalization of MI can reduce its bias towards the multi valued features. One such normalization is found in normalized mutual information feature selection (NIMS) [30] where the average of normalized mutual information is calculated for measuring the redundancy. The drawback of mRMR and NIMS is that during the computation of redundancy they do not consider the class variable and thus, may lose the necessary information for classification [21].

To solve this issue Brown *et al.* shows how to accommodate the complementary information with a common parameterized criteria [3] which was proposed in joint mutual information (JMI) [12]. According to his formulation the criteria is presented in (4)

$$I(f_i; C) - \frac{1}{|S|} \sum_{f_s \in S} \big(I(f_i; f_s) + I(f_i; f_s | C) \big) \quad (4)$$

Here, $I(f_i; f_s | C)$ helps to incorporate the complementary information. A unifying theoretical framework has also been established in [4] and concluded that inclusion of RrC property shows more satisfying results. Furthermore, Brown *et al.* [3] performs a rigorous experiment which evaluates 17 filter criteria proposed in different researches and concludes that JMI is the best considering accuracy, stability, and flexibility.

After the selection of features, filter based methods adopt different machine learning algorithms for classification. In the existing researches, the popular algorithms are C4.5 decision tree (DT) [31], [29], Support Vector Machine (SVM) [23], [11], Naïve Bayes [11], Random Forest [5] and Linear Discriminant Analysis [5].

B. Discretization

Discretization methods can be divided into two categories: unsupervised and supervised. The well-known unsupervised methods are Equal width and Equal frequency discretizers. These methods do not perform well as they lose a large amount of information when the data are not evenly distributed [26]. Addressing this issue, supervised methods are introduced where class information is used to discretize the data.

Supervised methods can be grouped into Static and Dynamic. Static methods can further be separated into top-down and bottom-up. In the bottom-up category, the popular discretization methods are ChiMerge [18], Chi2 [19] and Modified Chi2 [20]. However, all of these bottom-up methods have computational complexities that are usually worse than that of the top-down methods [32]. Among the discretization methods that follow the top-down approaches, CAIM is one of the best performing methods [2]. It is a supervised discretization algorithm which maximizes the class-attribute interdependence and is defined by the criteria presented in (5).

$$CAIM(C, D|F) = \left(\sum_{r=1}^{n} \frac{max_r{}^2}{M_{+r}}\right)/n \quad (5)$$

where, n is the number of intervals, q_{ir} is the total number of values belong to the i^{th} class that are within the interval $(d_{r-1}, d_r]$, max_r is the maximum value among all q_{ir} values at the r^{th} column of the quantization matrix, M_{+r} is the total number of values of feature f_i that are within the interval $(d_{r-1}, d_r]$. The advantage of CAIM is that it generates minimal number of discrete intervals and achieves reasonable accuracies for different datasets [26]. However, by definition it has a tendency to give priority to the class containing the highest number of instances and may fail for the unbalanced datasets. Besides, the number of intervals produced by CAIM is close to the number of class which introduces biasness towards the outcome of discretization [33]. To address these issues, a new version called ur-CAIM [33] is proposed which outperforms the original CAIM.

Apart from CAIM there are few other popular top down approaches. Fayyad and Irani [13] introduced a supervised discretization procedure that evaluates every interval and selects the one for which the class information entropy value is minimal. This process continues recursively for multiple intervals till the termination criteria based on MDLP [34] is achieved. The advantage of this method is that it has small number of intervals and low inconsistency rate [35]. Mutual Information Discretization [36] is another supervised top down method that selects the interval for which MI between the class and the feature is maximized.

All of these aforementioned methods are static that discretize each feature independently without considering its dependency with other features. Gama *et al.* [29] proposed a dynamic discretization method considering the dependencies among the features. They formulate their method representing all possible discretization of features as a hierarchy and perform A* search to find the one which has the lowest error rate as well as the smallest number of intervals. Beside this, there are other popular dynamic algorithms such as ID3 [27], ITFP [28]. However, the major limitation of dynamic method is that they are usually classifier dependent and thus have higher computational complexity.

Discretization usually is used as a pre-processing step of feature selection and is mostly found independently i.e., either discretization or feature selection is proposed in the existing literature. Thus it is necessary to develop a method that dynamically discretizes the features and selects features that achieves high accuracy.

III. Proposed Methododlogy

For a given a set of features $F = \{f_1, f_2, \dots, f_n\}$, the objective of this proposal is to select a subset $S = \{s_1, s_2 \dots, s_k\}$, with appropriate discretization levels $D = \{d_1, d_2, \dots, d_k\}$ so that a small set of features with a minimum number of intervals are found and achieve the highest classification accuracy with minimum computational time. To solve this problem, a heuristic namely discretization and selection of features based on mutual information (DSM) is proposed here. The overall process is shown in the DSM algorithm:

Algorithm : DSM

Input: $F = \{f_1, f_2, \dots, f_n\}$, C

Output: S with appropriate discretization $D = \{d_1, d_2, \dots, d_k\}$

1. **Begin**
2. **for each** $f_i \in F$
3. Calculate MI for f_i to find d_i using $Procedure$ 1
4. **End for**
5. Sort F in decreasing order based on their MI values
6. Select f_1 and the corresponding d_1 from F with max MI
7. $S \leftarrow S \cup f_1$; $D[1] \leftarrow d_1$; $F \leftarrow F \setminus S$
8. Initialize $T \leftarrow 0$
9. **for each** *i = 2 to n*
10. Calculate $Gain$ for f_i with d_i using $Procedure$ 2
11. **If** $(Gain > T)$ **then**
12. $T \leftarrow T + \epsilon$; $S \leftarrow S \cup f_i$; $D[i] \leftarrow d$
13. **End If**
14. $F \leftarrow F \setminus f_i$
15. **End for**
16. **Return** S and their respective D
17. **End**

DSM first finds the discretization level of each feature individually based on its relevancy ($Line\ 3\ of\ Algorithm$) by calculating its mutual information with C using (6).

$$MI = I\left(f_i^{d_i}; C\right) \quad (6)$$

The maximum number of intervals (max_d) is determined by the total number of continuous values of feature F. For a single feature, the number of intervals that produces the highest MI needs to be selected. However, the analysis on a large number of datasets reveals that the value of MI becomes almost steady after a certain number of intervals as shown in Fig 1.

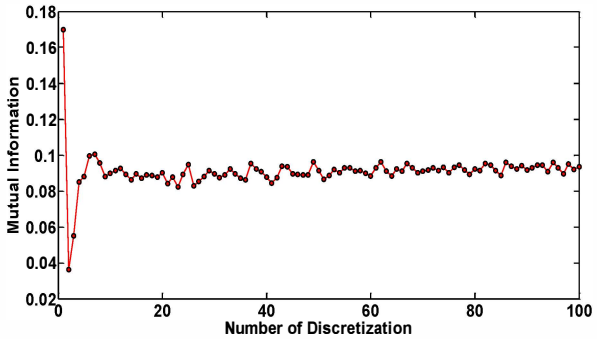

Figure 1. Mutual information at different discritization levels

As a result, the interval point that causes to start stabilizing the MI value is selected so that the interval size remains smaller. Finally, the features set F is achieved with their individual discretization level and the respective relevancy information for which the minimum number of intervals are found. This process is described in $Procedure\ 1$

Procedure 1

Begin
Initialize $MI \leftarrow -inf$
  for all $d_j = 2$ to max_d
    Discretize f_i with d_j
    Calculate MI_{temp} for feature f_i
    If $(MI_{temp} > MI)$ **then**
      $d \leftarrow d_j$
      If $(MI_{temp} - MI) \sim 0$ **then**
        $MI \leftarrow MI_{temp}$
        Break
      End If
      $MI \leftarrow MI_{temp}$
    End If
  End for
Return MI with its discretization d
End

At this stage, the features are ranked according to their corresponding MI values and the feature with the highest MI value is considered as the first selected features; i.e. $s_1 = argmax_{f_i} I\left(f_i^{d_i}; C\right)$ $(Line\ 6\ of Algorithm)$. For selecting the subsequent features, DSM tries the other features in the sorted order of MI. When considering feature f_j, its 'gain' is calculated using (7)

$$gain = I\left(f_j^{d_j}; C\right) + \frac{1}{|S|}\sum_{i=1}^{|S|}\left(I\left(s_i; f_j^{d_j}|C\right) - I\left(s_i; f_j^{d_j}\right)\right) \quad (7)$$

If the gain value is higher than the threshold T, it is added in the selected subsets S otherwise discarded ($Line\ 9-14\ of Algorithm$). As the gain is usually increasing while including a new feature, there is a higher chance of including almost all features in S. Thus, the threshold T is adjusted in such a way that it will stop selecting the features when the increment is insignificant which is called incremental gain strategy (IGS). IGS is defined as "Adjust T with a small value ϵ such that every stage of adding a new feature causes the improvement of gain". This process helps to select a small set of features that are highly informative. The overall process of selecting feature subset with their desired discretization is illustrated in $Procedure\ 2$

Procedure 2

Begin
Initialize $Gain \leftarrow -inf$, $p \leftarrow \pm\delta$
  for all $d_j = (d_i + p)$
    Discretize f_i with d_j interval
    Calculate $Gain_{temp}$ for f_i using (7)
    If $(Gain_{temp} > Gain)$ **then**
      $d \leftarrow d_k$; $Gain \leftarrow Gain_{temp}$
    End If
  End for
Return $Gain$ and d
End

For selecting the desired d_j, (7) considers each of the already selected features. The term $I\left(f_j^{d_j}; C\right)$ and $I\left(s_i; f_j^{d_j}\right)$ in (7) measures the relevancy and redundancy of f_j with C and the already selected feature s_i respectively at its discretization level d_j. The term $I\left(s_i; f_j^{d_j}|C\right)$ measures the complementary information of the already selected feature s_i and the current feature f_j at its discretization level d_j. All these terms are normalized by the joint entropy to make them more interpretable and less sensitive to occurrence frequency [38]. Considering the interdependence with the selected features, each feature is evaluated by changing its discretization level to find a better one. To discretize the feature dynamically, its number of intervals is adjusted with $\pm\delta$, assuming the best intervals will appear near its value (d_j) that is obtained from $Procedure\ 1$.

IV. Experimental Result and Discussion

In this section, the description of the dataset is first presented followed by the results and discussions.

A. Implementation Details

15 benchmark datasets from the UCI Machine Learning Repository [39] are used to validate the proposed method. The characteristics of the datasets are presented in the first column of Table I. In this study, DT and SVM (linear kernel) are used to classify the datasets. Furthermore, 5-fold cross-validation is performed to report the average of results in Table I. The same strategy is also followed to generate results of other methods.

TABLE I. CLASSIFICATION ACCURACY FOR DIFFRNENT METHODS. (NO. OF SELECTED FEATURES ARE GIVEN IN PARENTHESIS)

Dataset (Features/Instances/class #)	WDF		CAIM		JMI		mRMR		CMIM		DSM	
	DT	SVM	DT	SVM	DT	SVM	DT	SVM	DT	SVM	DT	SVM
Dermatology (34/366/6)	0.947	**0.968**	0.850	0.886	0.905	**0.968**	0.905	**0.968**	0.905	**0.968**	**0.950(16)**	0.944(16)
Sonar(60/208/2)	0.748	0.618	0.730	0.795	0.702	0.618	0.702	0.618	0.702	0.618	**0.772(19)**	**0.796(19)**
Glass(9/214/6)	0.675	0.506	0.693	**0.662**	0.684	0.506	0.684	0.506	0.684	0.506	**0.694(8)**	0.600(8)
Wine(13/178/3)	0.881	0.918	0.854	0.967	**0.935**	0.918	**0.935**	0.918	**0.935**	0.918	0.886(10)	**0.968(10)**
Australian (14/690/2)	0.830	0.860	0.782	0.814	0.653	0.860	0.653	0.860	0.653	0.860	**0.854(7)**	**0.874(7)**
BreastTissue(9/106/6)	0.600	**0.725**	0.575	0.575	**0.641**	0.725	**0.641**	0.725	**0.641**	0.725	0.625(6)	0.683(6)
Pima (8/768/2)	0.683	**0.766**	**0.733**	0.725	0.683	**0.766**	0.683	**0.766**	0.683	**0.766**	0.712 (8)	0.746(8)
Yeast (8/148/10)	**0.554**	0.311	0.549	0.577	0.515	0.311	0.515	0.311	0.515	0.311	0.522(6)	**0.578**(6)
Libras(91/360/15)	0.581	0.514	0.568	**0.784**	0.413	0.514	0.413	0.514	0.413	0.514	**0.608(15)**	0.725(15)
Iris(4/150/3)	**0.940**	0.926	0.913	0.880	0.906	0.926	0.906	0.926	0.906	0.926	0.926(3)	**0.953(3)**
Arrhythmia (279/452/16)	**0.738**	0.691	0.724	**0.781**	0.668	0.670	0.668	0.670	0.668	0.670	0.724(14)	0.722(14)
Parkinsons (22/197/2)	0.805	0.845	0.845	0.860	0.840	0.845	0.840	0.845	0.840	0.845	**0.845(14)**	**0.861(14)**
German (20/1000/2)	0.714	**0.762**	0.715	0.707	**0.729**	**0.762**	**0.729**	**0.762**	**0.729**	**0.762**	0.716(8)	0.752(8)
Liver (7/345/2)	0.611	0.710	0.628	0.579	0.611	0.710	0.611	0.710	0.611	0.710	**0.629(6)**	**0.724(6)**
Lung (56/32/3)	0.633	0.533	0.600	0.366	0.633	0.533	0.633	0.533	0.633	0.533	**0.666(12)**	**0.567(12)**
Average	0.711	0.697	0.706	0.713	0.701	0.709	0.701	0.709	0.701	0.709	**0.742**	**0.766**

B. Results and Discussion

The performance of DSM is first compared to the performances without any feature selection and discretization. The column WDF in Table I represents the classification accuracies for the original datasets. It is observed that DSM outperforms in most of the cases even with a small number of selected features. It can also be found that mRMR, CMIM and JMI methods produce same accuracies (using forward selection strategy) with equal number of selected features for these datasets. Here, the number of selected features for each of the feature selection method is kept same as DSM for fair comparison. Under this setting, DSM produces better results in most of the cases. For example, in case of low dimensional features such as *Australian* dataset, DSM select only seven features (using DT) and produces much better results compare to all. Similar observation can also be made for high dimensional feature set such as *Lung* and *Libras* dataset. In a few cases, the discretization method namely CAIM performs better compare to the DSM. However, CAIM uses all the features but DSM uses only a subset of features.

The beauty of the proposed method is that for both small and large feature set, in most of the cases, DSM demonstrates better performances with much lower number of features (e.g., *Sonar*, *Parkinsons* and *Lung* datasets) which validates its effectiveness. The main reason for selecting a lower number of features is due to the incorporation of IGS. IGS also helps to improve the performance because it discards features which fail to increase the value of gain significantly. Moreover, the average accuracies are also much better than the compared methods which indicate the efficiency of DSM.

The graph in Fig 2 demonstrates the performances of JMI (JMI is the best performing one according to [3]) and DSM for different number of features in three datasets namely Dermatology, Lung and Libras. Except for twenty features of Dermatology and ten features of lung, DSM produces better results compared to JMI. Most importantly, DSM shows much stable behavior with the increasing number of features compared to JMI.

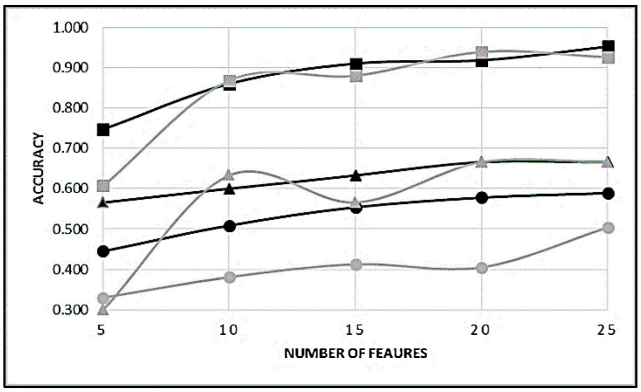

Figure 2. Classification accuracy for different number of features (square, triangle and circle represent *dermatology, lung* and *libras* dataset respectively where black and grey line represent DSM and JMI respectively)

V. CONCLUSION

In this paper, we have introduced DSM that simultaneously selects and discretizes a given set of features using mutual information. In most of the cases, it is observed that DSM outperforms the state-of-the-art discretization and feature selection methods. It is also a dynamic discretization procedure that selects a small set of features by incorporating an Incremental Gain Strategy (IGS). However, rigorous study is required to find the best possible incremental value for IGS which we plan to address in future.

ACKNOWLEDGMENT

This research is supported by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh. 56.00.0000.028.33.065.16-747, 21-06-2016.

REFERENCES

[1] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

[2] Garcia, Salvador, Julian Luengo et.al., "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning." IEEE Transactions on Knowledge and Data Engineering 25, no. 4 (2013): 734-750.

[3] Brown, Gavin, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection." Journal of Machine Learning Research 13, no. Jan (2012): 27-66..

[4] Vergara, Jorge R., and Pablo A. Estévez. "A review of feature selection methods based on mutual information." Neural Computing and Applications 24.1 (2014): 175-186.

[5] Naghibi, Tofigh, Sarah Hoffmann, and Beat Pfister. "A semidefinite programming based search strategy for feature selection with mutual information measure." IEEE transactions on pattern analysis and machine intelligence 37.8 (2015): 1529-1541.

[6] Chlebus, Bogdan S., and Sinh Hoa Nguyen. "On finding optimal discretizations for two attributes." International Conference on Rough Sets and Current Trends in Computing. Springer, 1998.

[7] Bins, José, and Bruce A. Draper. "Feature selection from huge feature sets." Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. Vol. 2. IEEE, 2001.

[8] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." Journal of machine learning research 5.Oct (2004): 1205-1224.

[9] Dash, Manoranjan, and Huan Liu. "Consistency-based search in feature selection." Artificial intelligence 151.1 (2003): 155-176.

[10] Battiti, Roberto. "Using mutual information for selecting features in supervised neural net learning." IEEE Transactions on neural networks 5.4 (1994): 537-550.

[11] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence 27.8 (2005): 1226-1238.

[12] Yang, H., and John Moody. "Feature selection based on joint mutual information." Proceedings of international ICSC symposium on advances in intelligent data analysis. 1999.

[13] Irani, Keki B. "Multi-interval discretization of continuous-valued attributes for classification learning." (1993).

[14] X. Wu, “A Bayesian Discretizer for Real-Valued Attributes,” The Computer J., vol. 39, pp. 688-691, 1996.

[15] M. Boulle´, “MODL: A Bayes Optimal Discretization Method for Continuous Attributes,” Machine Learning, vol. 65, no. 1, pp. 131-165, 2006.

[16] S.H. Nguyen and A. Skowron, “Quantization of Real Value Attributes - Rough Set and Boolean Reasoning Approach,” Proc. Second Joint Ann. Conf. Information Sciences (JCIS), pp. 34-37, 1995.

[17] G. Zhang, L. Hu, and W. Jin, “Discretization of Continuous Attributes in Rough Set Theory and Its Application,” Proc. IEEE Conf. Cybernetics and Intelligent Systems (CIS), pp. 1020-1026, 2004

[18] Kerber, Randy. "Chimerge: Discretization of numeric attributes." Proceedings of the tenth national conference on Artificial intelligence. Aaai Press, 1992.

[19] Liu, Huan, and Rudy Setiono. "Chi2: Feature selection and discretization of numeric attributes." ICTAI. 1995.

[20] Tay, Francis EH, and Lixiang Shen. "A modified chi2 algorithm for discretization." IEEE Transactions on knowledge and data engineering 14.3 (2002): 666-670.

[21] Bennasar, Mohamed, Yulia Hicks, and Rossitza Setchi. "Feature selection using joint mutual information maximisation." Expert Systems with Applications 42.22 (2015): 8520-8532.

[22] Lewis, David D. "Feature selection and feature extraction for text categorization." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.

[23] Meyer, Patrick Emmanuel, Colas Schretter, and Gianluca Bontempi. "Information-theoretic feature selection in microarray data using variable complementarity." IEEE Journal of Selected Topics in Signal Processing 2.3 (2008): 261-274.

[24] Cheng, Hongrong, Zhiguang Qin, Weizhong Qian, and Wei Liu. "Conditional mutual information based feature selection." In Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on, pp. 103-107. IEEE, 2008.

[25] Kurgan, Lukasz A., and Krzysztof J. Cios. "CAIM discretization algorithm." IEEE transactions on Knowledge and Data Engineering 16.2 joiny(2004): 145-153.

[26] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Discretization techniques: A recent survey." GESTS International Transactions on Computer Science and Engineering 32.1 (2006): 47-58.

[27] Quinlan, J. Ross. C4.5: programs for machine learning. Elsevier, 2014.

[28] Au, W-H., Keith CC Chan, et. al. "A fuzzy approach to partitioning continuous attributes for classification." IEEE Transactions on Knowledge and Data Engineering 18.5 (2006): 715-719.

[29] Gama, Joao, Luis Torgo, and Carlos Soares. "Dynamic discretization of continuous attributes." Ibero-American Conference on Artificial Intelligence. Springer Berlin Heidelberg, 1998.

[30] Estévez, Pablo A., Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada. "Normalized mutual information feature selection." IEEE Transactions on Neural Networks 20, no. 2 (2009): 189-201.

[31] Yan, Deqin Garcia, Deshan Liu, and Yu Sang. "A new approach for discretizing continuous attributes in learning systems." Neurocomputing 133 (2014): 507-511.

[32] Tsai, Cheng-Jung, Chien-I. Lee, and Wei-Pang Yang. "A discretization algorithm based on class-attribute contingency coefficient." Information Sciences 178.3 (2008): 714-731.

[33] Cano, Alberto, Dat T. Nguyen, Sebastián Ventura, et.al. "ur-CAIM: improved CAIM discretization for unbalanced and balanced data." Soft Computing 20, no. 1 (2016): 173-188.

[34] Rissanen, Jorma. "Modeling by shortest data description." Automatica 14.5 (1978): 465-471.

[35] Liu, Huan, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. "Discretization: An enabling technique." Data mining and knowledge discovery 6, no. 4 (2002): 393-423.

[36] Ferreira, Artur J., and Mário AT Figueiredo. "Feature Discretization with Relevance and Mutual Information Criteria." Pattern Recognition Applications and Methods. Springer International Publishing, 2015. 101-118.

[37] Balagani, Kiran S., and Vir V. Phoha. "On the feature selection criterion based on an approximation of multidimensional mutual information." IEEE Transactions on Pattern Analysis and Machine Intelligence 32.7 (2010): 1342-1343

[38] Bouma, Gerlof. "Normalized (pointwise) mutual information in collocation extraction." Proceedings of GSCL (2009): 31-40.

[39] "UCI machine learning repository,". [Online]. Available: http://archive.ics.uci.edu/ml/. Accessed: Nov. 16, 2016.

[40] Wanderley MF, Gardeux V, Natowicz R, de Pádua Braga A. GA-KDE-Bayes: an evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems. InESANN 2013.

[41] Sharmin S, Arefin MR, Abdullah-Al Wadud M, Nower N, Shoyaib M. SAL: An effective method for software defect prediction. In Computer and Information Technology (ICCIT), 2015 18th International Conference on 2015 Dec 21 (pp. 184-189). IEEE.

[42] Gu S, Cheng R, Jin Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing. 2016:1-2.

[43] Yan HP, Liu WB. The third order correction on Hawking radiation and entropy conservation during black hole evaporation process. Physics Letters B. 2016 Aug 10;759:293-7.

[44] Fattah SA, Lin CC, Kung SY. A mutual information based approach for evaluating the quality of clustering. InAcoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on 2011 May 22 (pp. 601-604). IEEE.