

Performance Comparison of Attributes Selection for Machine Learning Task

Thu Zar Phyu¹, and Nyein Nyein Oo²

Department of Information Technology Engineering, Yangon Technological University, Yangon, Myanmar

¹thuzarphyu85@gmail.com, ²nno2005@gmail.com

Abstract: Feature or attribute selection is a topic that concerns selecting a subset of features, among the full features, that shows the best performance in classification accuracy. It performs as a preprocessing step to improve the classification task. The main objective of feature selection is to find useful features that represent the data and remove those features that are either irrelevant or redundant. Reducing the number of features in a dataset can lead to faster software quality model training and improved classifier performance. This paper presents a new method for dealing with feature subset selection based on conditional mutual information. The proposed method can select feature subset with minimum number of features, which are relevant to get higher average classification accuracy for datasets. The experimental results with UC Irvine datasets and Naïve Bayes classifier showed that the proposed algorithm is effective and efficient in selecting subset with minimum number of features getting higher classification accuracy than the existing feature selection methods.

Keywords: feature selection, classification, irrelevant, redundant, datasets.

1. Introduction

Machine learning algorithms automatically extract knowledge from machine readable information. Unfortunately, their success is usually dependent on the quality of the data that they operate on. If the data is inadequate, or contains extraneous and irrelevant information, machine learning algorithms may produce less accurate and less understandable results, or may fail to discover anything of use at all. Therefore, feature selection performs as a preprocessing step to improve the machine learning task. Feature selection is a pre-processing technique that finds a minimum subset of features that captures the relevant properties of a dataset to enable adequate classification. Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. Feature subset selection can result in enhanced performance, a reduced hypothesis search space, and, in some cases, reduced storage requirement. [1] Generally, features are characterized as: (i) Relevant: features which have an influence on the output and their role cannot be assumed by the rest, (ii) Irrelevant: features not having any influence on the output, (iii) Redundant: a feature can take the role of another [1]. Feature selection aims at selecting the most relevant feature subset for more efficient training and improved accuracy. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. In recent years, data has become increasingly larger in both number of instances and number of features in many applications such as genome project, text, image retrieval, and customer relationship management [2]. Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. High dimensional data can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. Therefore, feature selection becomes very necessary for machine learning tasks when facing high dimensional data nowadays.

The paper is organized as follows. In the next section, related works are described. Section III gives brief overview of the system design. Section VI contains attributes selection methods used in the experiment. Section

V gives a brief description of datasets and learning algorithm used in experiment. Section VI describes experimental evaluation and discussion of the obtained results. Conclusion is shown in section VII.

2. Related Work

Tang and Mao [3] presented feature selection algorithm for mixed data with both nominal and continuous feature. In this paper, a criterion avoids feature-type transformation through carefully decomposing the feature space along values of nominal features in the mixed feature subset and measuring class separability based on continuous features in each subspace generated and then combining these measures to produce an overall evaluation. The search algorithm, named mixed forward selection (MFS), is different from traditional search algorithms because it considers the feature-type in both subsets generation and comparison. Liu and Zheng [4] present another feature selection method named filtered and supported sequential forward search (FS_SFS) in the context of support vector machines (SVM). In comparison with conventional wrapper methods that employ the SFS strategy, FS_SFS has two important properties to reduce the time of computation. First, it dynamically maintains a subset of samples for the training of SVM. Because not all the available samples participate in the training process, the computational cost to obtain a single SVM classifier is decreased. Secondly, a new criterion, which takes into consideration both the discriminant ability of individual features and the correlation between them, is proposed to effectively filter out nonessential features. An improved forward floating selection (IFFS) algorithm for selecting a subset of features is presented by Nakariyakul and Casasent [5]. The proposed algorithm improves the state-of-the-art sequential forward floating selection algorithm. The improvement is to add an additional search step called “replacing the weak feature” to check whether removing any feature in the currently selected feature subset and adding a new one at each sequential step can improve the current feature subset. Zahra Karimi and Mohammad Mansour [6] presented a hybrid feature selection methods by combining symmetric uncertainty measure and gain measure. Both SU and gain measures for each feature-class correlation were calculated first and then rank feature according to average score value. High ranked feature greater than a threshold values was selected. They evaluated their system using KDD dataset and Naïve Bayes algorithm. The average detection rate of their method is 98.28, which is higher than other comparable methods. In this paper, we proposed a new method for dealing with feature subset selection based on conditional mutual information. The experimental results with UC Irvine datasets and Naïve Bayes classifier is presented in section IV.

3. Overall System Design

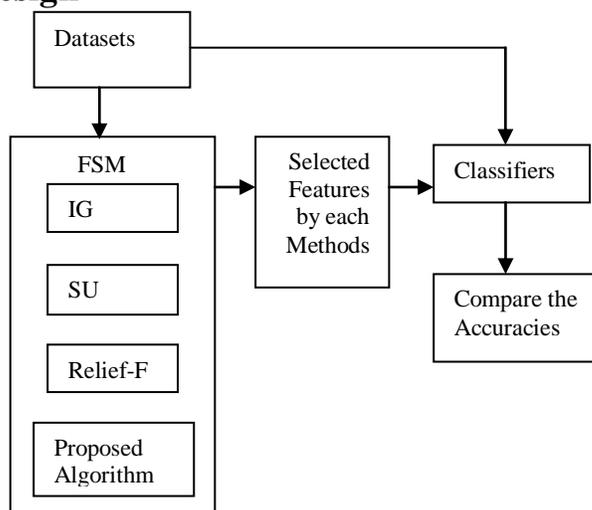


Fig. 1: Overall System Design.

The overall system design of the proposed system is shown in Fig.1. This system is implemented to compare the classification accuracy using full features and selected features produced by each feature selection methods such as Information Gain (IG), Symmetrical Uncertainty (SU), Relief-F and proposed algorithm. In order to evaluate how good the selected features are, Naïve Bayes is applied to each datasets.

4. Attribute Selection Methods in Experiment

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. It is the process of choosing a subset of original features so that the feature space is optimally reduced to evaluation criterion. Feature selection can reduce both the data and the computational complexity. The raw data collected is usually large, so it is desired to select a subset of data by creating feature vectors that Feature subset selection is the process of identifying and removing much of the redundant and irrelevant information possible. This results in the reduction of dimensionality of the data and thereby makes the learning algorithms run in a faster and more efficient manner. Various feature selection methods are available in WEKA (Waikato Environment for Knowledge Analysis) such as Information Gain (IG), Symmetrical Uncertainty (SU) and Relief-F.

4.1. Information Gain (IG)

Information Gain is an important measure used for ranking features. Given the entropy is a criterion of impurity in a training set S , we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y \setminus X) = H(X) - H(X \setminus Y) \quad (1)$$

IG is a symmetrical measure. The information gained about Y after observing X is equal to the information gained about X after observing Y . A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative [6].

4.2. Symmetrical Uncertainty (SU)

Symmetric Uncertainty is one of the best feature selection methods and most feature selection system based on mutual information use this measure. SU is a correlation measure between the features and the class.

$$SU = (H(X) + H(Y) - H(X \setminus Y)) / (H(X) + H(Y)) \quad (2)$$

where $H(X)$ and $H(Y)$ are the entropies based on the probability associated with each feature and class value respectively and $H(X, Y)$, the joint probabilities of all combinations of values of X and Y [6].

4.3. Relief-F

The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature f .

$$w_f = P(\text{different value of } f / \text{different class}) - P(\text{different value of } f / \text{same class}) \quad (3)$$

This approach has shown good performance in various domains [7].

4.4. Proposed Feature Selection Method

Information Theoretical concepts is used to implement the effective feature selection algorithm and then implement the entropy, joint entropy and Mutual Information to produce the first most relevant feature to the class. After that, Conditional Mutual Information is used to reduce redundancy and to produce the other most effective features to the class. First calculate $I(C; X_i)$ for all $X_i \in X$. Then, the i^{th} feature that has maximum $I(C; X_i)$ is chosen as first relevant feature as it provide highest class information among other features. In the next steps, repeat feature selection process until the feature set X becomes empty. Select and remove feature one by one by using the proposed criteria.

$$D(F; C) = D(F_1; F_2; \dots; F_N; C).$$

Step 1. Initialization

Set $S = \text{“empty set”}$, set $X = \text{“initial set of all F features”}$

Step 2. For $i = 1 \dots N$ do

For all features $X_i \in X$ compute $I(C, X_i)$.

Step 3. Selection of the first feature:

Find feature $X_i \in X$ that maximizes $I(C, X_i)$; set $X = X \setminus \{X_i\}$, $S = \{X_i\}$.

Find feature $X_i \in X$ that $I(C, X_i) \cong 0$. Set $X = X \setminus \{X_i\}$

Step 4. Repeat

(a) Computation of the Conditional MI

For all pairs of features (X_i, X_s) with $X_i \in X \setminus S$, $X_s \in S$ computes $I(C, X_i | X_s)$, if it is not yet available.

(b) Selection of the next feature:

$$X^+ \leftarrow \arg \max_{X_i \in X} \left[I(C; X_i) - \sum_{X_s \in S} \{I(C; X_i) - I(C; X_i | X_s)\} \right]$$

Set $X = X \setminus \{X^+\}$, $S = S \cup \{X^+\}$.

(c) Removal of feature

$$X^- \leftarrow I(C; X_i | X_s) \cong 0$$

Set $X = X \setminus \{X^-\}$

Until $(X \text{ is } [])$

5. Learning Algorithm and Datasets

5.1. Naïve Bayes Classifier

Naive Bayes classifier greatly simplifies learning by assuming that features are independent given the class variable. More formally, this classifier is defined by discriminate functions:

$$F_i(X) = \prod_{j=1}^N P(x_j/c_i)P(c_i) \quad (4)$$

where $X = (x_1, x_2, \dots, x_N)$ denotes a feature vector and $j = 1, 2, \dots, N$, denote possible class labels. The training phase for learning a classifier consists of estimating conditional probabilities $P(x_j | c_i)$ and prior probabilities $P(c_i)$. Here, $P(c_i)$ are estimated by counting the training examples that fall into class c_i and then dividing the resulting count by the size of the training set. Similarly, conditional probabilities are estimated by simply observing the frequency distribution of feature x_j within the training subset that is labeled as class c_i . To classify a class-unknown test vector, the posterior probability of each class is calculated, given the feature values present in the test vector; and the test vector is assigned to the class that is of the highest probability [6].

5.2. Datasets

Standard seven datasets such as Letter, Glass, Sonar, Arrhythmia, Cylinder-band, Waveform and Vehicle drawn from the UC Irvine were used in the experiments. Letter dataset include 17 features, 20000 instances and 26 classes. Glass dataset includes 10 features, 214 instances and 7 classes. Sonar includes 60 features, 208 instances and 2 classes. Arrhythmia includes 280 features, 452 instances and 16 classes. Cylinder-band includes 40 features, 512 instances and 2 classes. Waveform includes 41 features, 1000 instances and 3 classes. Vehicle includes 18 features, 946 instances and 4 classes. These datasets include discrete and continuous attributes. These seven datasets are available online. Letter, Waveform and Vehicle datasets include more instances than other datasets. A summary of datasets is presented in “Table 1”.

TABLE I: Characteristics of Datasets

No	Dataset	Features	Instances	Classes
1	Letter	17	20000	26
2	Glass	10	214	7
3	Sonar	60	208	2
4	Arrhythmia	280	452	16
5	Cylinder-band	40	512	2
6	Waveform	41	1000	3
7	Vehicle	18	946	4

6. Experimental Result

The objective of this section is to evaluate the algorithms in terms of number of selected features and learning accuracy. Firstly, we consider the number of features reduced by feature selection methods. Reducing the number of features of dataset is important because it can decrease the complexity and reduce the learning time. WEKA (Waikato Environment for Knowledge Analysis) is used to measure the performance of each feature selection algorithm. It is a well-known machine learning tool based on JAVA. Selected feature subsets by each

method are evaluated using Naïve Bayes learning algorithm. The number of features selected by each feature selection methods is presented in Table 2.

TABLE II: Number of features selected by feature selection methods

No	Dataset	IG	SU	Relief-F	Proposed Algo
1	Letter	16	16	16	16
2	Glass	8	9	9	6
3	Sonar	8	22	45	11
4	Arrhythmia	74	112	140	79
5	Cylinder-band	4	21	16	19
6	Waveform	18	20	18	8
7	Vehicle	17	19	19	14

The classification accuracy of selected feature by proposed algorithm is measured using the Naïve Bayes learning algorithm by 10-fold cross validation. Then, features produced by each existing feature selection methods are evaluated by the same classifier. After that the difference of accuracies between the reorganized datasets are calculated and compared. The proposed method was tested for various datasets in UC Irvine repository and compared the performance with other feature selection algorithm. Each values in Table 3 shows average classification accuracies.

TABLE III: Classification results on Naïve Bayes Classifier using 10 fold cross validation

No	Dataset	All	IG %	SU %	Relief-F %	Proposed Algo %
1	Letter	64.1	64.1	64.1	64.1	74.4
2	Glass	48.6	48.6	48.1	48.6	73.4
3	Sonar	67.8	69.71	69.7	69.7	86.1
4	Arrhythmia	62.4	69.0	69.46	68.4	77.7
5	Cylinder-band	72.2	65.4	67.8	63.9	80.9
6	Waveform	80.0	80.1	80.1	80.0	81.4
7	Vehicle	44.8	44.8	44.8	44.8	62.6

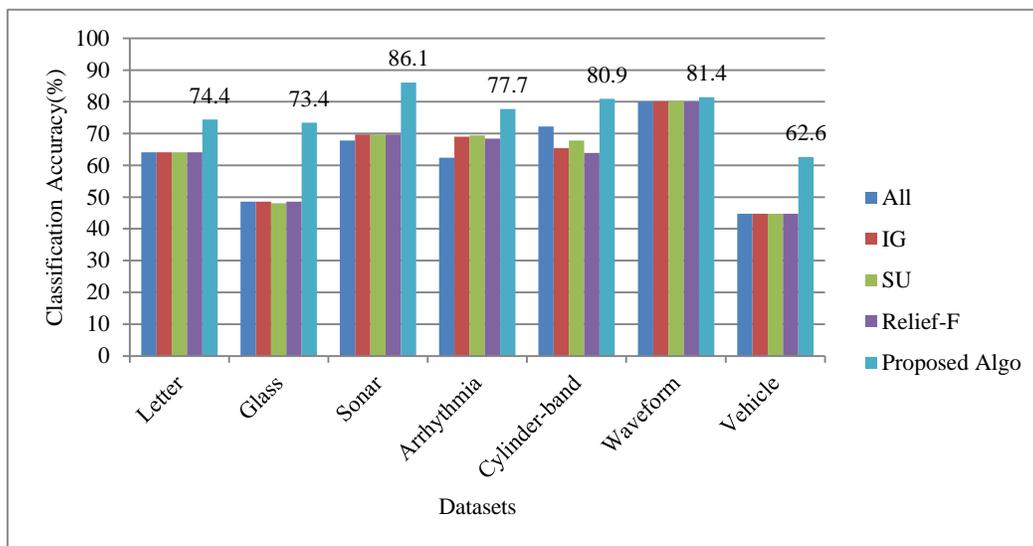


Fig. 2: Average Classification Accuracies on datasets

The chart in Fig 2 shows learning accuracy compared with the proposed algorithm. The vertical lines show the percentage of classification accuracies by each method. The names of datasets are shown at the bottom of the chart. All together 7 datasets were used to test the performances of each feature selection methods. Among all

datasets, waveform consists of 1000 patterns and has 41 features and 3 classes. The proposed algorithm selected only 8 effective features among the 41 features to classification experiments and result analysis. Moreover, the proposed algorithm reduced features from 10 and 18 to 6 and 14 in the Glass and Vehicle datasets. In the remaining datasets, the proposed algorithm reduced feature more than the other algorithms.

7. Conclusion

In this paper, three different feature selection methods and the proposed Conditional Mutual Information based method were compared and tested with public datasets. The proposed method significantly reduced features from 60 and 41 to 12 and 8 and also got higher classification accuracies 86.06 and 81.4 than the other methods in the Sonar and Waveform datasets. Almost all of the methods produced better performance in terms of the feature size and the classification accuracy, the proposed method performed even better than other methods for all datasets with Naïve Bayes learning algorithms. The proposed algorithm not only reduced feature more than most of the other existing algorithms but also produced effective features to get higher classification accuracies.

8. Acknowledgement

The author is deeply grateful to Professor Dr. Myo Min Than, Head of Department of Information Technology Engineering, Yangon Technological University, for extending all the facilities of the department.

The author is extremely grateful to supervisor Associate Professor Dr. Nyein Nyein Oo from Yangon Technological University, for her guidance, advice and encourages for completing this paper.

9. References

- [1] S. B. Kotsiantis, "Feature selection for machine learning classification", Springer Science+Business Media B.V. 2011.
- [2] L. Faivishevsky and J. Goldberger, "Unsupervised Feature Selection based on Non-Parametric Mutual Information", *IEEE in International Workshop on Machine Learning for Signal*, Sept 23–26, 2012.
- [3] T. W. Mao, "Feature selection algorithm for mixed data with both nominal and continuous features", 2007.
- [4] L. Y. Zheng, "FS_SFS: A Novel Feature selection Method for Support Vector Machines and Pattern Recognition" 39:1333–1345, 2006.
- [5] S. Nakariyakul, D.P. Casasent, "An improvement on floating search algorithms for feature subset selection", 2009.
- [6] Z. Karimi and M. Mansour, "Feature Ranking in Intrusion Detection Dataset using combination of filtering", *International Journal of Computer Applications*, Vol78, September 2013.
<http://dx.doi.org/10.5120/13478-1164>
- [7] M. Robnik and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", *Machine Learning Journal*, 2003.
<http://dx.doi.org/10.1023/A:1025667309714>