MATEC Web of Conferences 42, 06002 (2016)
DOI: 10.1051/matecconf/20164206002

Performance Comparison of Feature Selection Methods

Thu Zar Phyu[1] , Nyein Nyein Oo[2]

[1,2] *Department of Information Technology, Yangon Technological University, Yangon, Myanmar*

Abstract. Feature Subset Selection is an essential pre-processing task in Data Mining. Feature selection process refers to choosing subset of attributes from the set of original attributes. This technique attempts to identify and remove as much irrelevant and redundant information as possible. In this paper, a new feature subset selection algorithm based on conditional mutual information approach is proposed to select the effective feature subset. The effectiveness of the proposed algorithm is evaluated by comparing with the other well-known existing feature selection algorithms using standard datasets from UC Iravine and WEKA (Waikato Environment for Knowledge Analysis). The performance of the proposed algorithm is evaluated by multi-criteria that take into account not only the classification accuracy but also number of selected features.

1 Introduction

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. This reduces the dimensionality of the data and allows learning algorithms to operate faster and more effectively. Feature selection aims at finding a feature subset that can describe the data for a learning task as good as or better than the original dataset. This paper presents a new approach to feature selection based on Mutual Information that uses a correlation based heuristic to evaluate the worth of features. A “feature” or “attribute” or “variable” refers to an aspect of the data. Usually before collecting data, features are specified or chosen. Features can be discrete, continuous, or nominal. Generally, features are characterized as relevant, irrelevant and redundant. Relevant features which have an influence on the output and their role cannot be assumed by the rest. Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example. Redundancy exists whenever a feature can take the role of another [1]. The main objective of feature selection algorithm is to select the subset of features that are independent each other and sufficiently relevant for learning process.

The paper is organized as follows. In the next section, related works are described. Section III contains feature selection methods used in the experiment. Section IV gives brief overview of the system design. Section V gives a brief description of datasets and learning algorithm used in experiment and also describes experimental evaluation. Final section contains discussion of the obtained results.

2 Related works

H. Liu et al. proposed a consistency based feature selection mechanism. It evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes. Consistency of any subset can never be lower than that of the full set of attributes; hence the usual practice is to use this subset evaluator in conjunction with a Random or Exhaustive search which looks for the smallest subset with consistency equal to that of the full set of attributes [1]. M. Hall presented a new correlation based approach to feature selection (CFS) in different datasets and demonstrated how it can be applied to both classification and regression problems for machine learning [2]. Anirut Suebsing, Nualsawat Hiransakolwong proposed Euclidean distance measure to use as score in feature selection for KDD dataset. High score features that is greater than defined threshold value were selected as best feature subsets [3]. Zahra Karimi, Mohammas Mansour, 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 [4]. A. Chaudhary, et. al. presented the performance evaluation of three feature selection methods with optimized Naïve Bayes is performed on mobile device. Correlation based method, Gain Ratio method and Information Gain method methods were used in this work [5].

3 Feature selection methods in experiment

Feature selection can reduce both the data and the computational complexity. It can also get more efficient and find out the useful feature subsets. 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.

3.1 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

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

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 [4].

3.2 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.

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

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 [4].

3.3 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(different value of f/different class) - P (different value of f/same class) (3)

This approach has shown good performance in various domains [6].

3.4 Proposed feature selection method

First calculate $I(C;X_i)$ for all $X_i \in X$. Then, the ith 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. 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 $Xi \in X$. Then, the ith 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.

Among non-linear correlation measures, many measures

Require: $D(F;C) = D(F_1; F2; : : : ; F_N;C)$

Step 1. Initialization

Set S= “empty set”, set X= “initial set of all F features”

Step 2. For i = 1 … 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 $X = X\backslash \{X_i\}$, $S = \{X_i\}$. Find feature $X_i \in X$ that $I(C, X_i) \cong 0$. Set $X = X \backslash \{X_i\}$

Step 4. Repeat

(a) Computation of the Conditional MI

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

(b) Selection of the next feature:

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

Set X = X \ {X+}, S = S ∪ {X+}.

(c) Removal of feature

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

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

Until (X is [ ]).

4 Overall system design

After proposed algorithm is implemented, the evaluation is started. We aims to evaluate the performance of the proposed algorithm using different benchmark datasets from WEKA (Waikato Environment for Knowledge Analysis) and UC Iravine repository and the proposed algorithm is compared with the well-known feature selection methods; Info-Gain, Relief-F and Symmetrical Uncertainty in terms of number of features reduced and learning accuracies. In WEKA and UCI, there are many standard benchmark datasets provided for data mining algorithms evaluation. In order to evaluate how good the selected features are, we apply Naive Bayes and j48 classifiers to evaluate the classification quality of the features selected by each of the four algorithms. Fig. 1 shows the overall system design of the system.

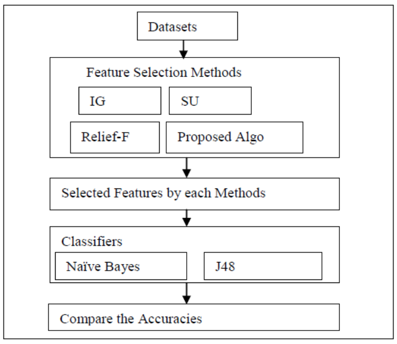

Figure 1. Overall System Design

5 Learning algorithm, datasets and experiment

5.1 Naïve bayes classifier

The naive Bayes algorithm employs a simplified version of Bayes formula to decide which class a novel instance belongs to. The posterior probability of each class is calculated, given the feature values present in the instance; the instance is assigned the class with the highest probability. 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:

$$f_i(X) = \prod_{j=1}^{N} P(x_j \mid c_i) P(c_i) \tag{4}$$

where $X = (x_1, x_2, ..., x_N)$ denotes a feature vector and $j = 1, 2, ..., N$, denote possible class labels. The training phase for learning a classifier consists of estimating conditional probabilities $P\ (x_j \backslash\ c_i)$ 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.

5.2 J48 classifier

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple. While building a tree, J48 ignores the missing values. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them.

Table 1. Characteristics of Datasets

No	Datasets	Features	Instances	Classes
1	Vehicle	18	945	4
2	Page-blocks	11	5473	5
3	Sonar	60	208	2
4	Liver-disorder	7	345	2
5	Cylinder-band	40	512	2
6	Waveform	41	1000	3

5.3 Datasets

Standard datasets drawn from the UC Iravine and WEKA (Waikato Environment for Knowledge Analysis) collection were used in the experiments. These datasets include discrete and continuous attributes. A summary of datasets is presented in “Table 1”. WEKA (Waikato Environment for Knowledge Analysis) contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [7].

5.4 Performance measure for machine learning algorithm

The objective of this section is to evaluate the algorithms in terms of number of selected features and learning accuracy on selected features. In our experiment, we choose three representative feature selection algorithms and proposed algorithm for comparison. The experiment is conducted using WEKA (Waikato Environment for Knowledge Analysis) implementation of all these existing algorithms. All together six datasets are selected from the WEKA (Waikato Environment for Knowledge Analysis) and UC Iravine Machine Learning Repository. The number of features selected by each feature selection methods is presented in “Table 2”.

The performance of any machine learning algorithm is determined by some measures. Any classification is correct if it can be judged by calculating the number of correctly identified class samples (true positives), the number of correctly identified samples that are not members of the class (true negatives) and samples that either were incorrectly allocated to the class (false positives) or that were not identified as class samples (false negatives). All the experiments are performed using ten-fold cross validation method. All the methods produced better performance with reduced feature set than full features, there are no significantly changes in these datasets compare to the existing feature selection algorithm. The classification accuracies using Naïve Bayes significantly increase after applying proposed algorithm with small features in Vehicle, Page-block, Sonar and Waveform datasets. Not only the proposed algorithm reduced feature from 60 to 11 and 41 to 8 in Sonar and Waveform datasets but also the classification accuracies are higher compared to other. In the Cylinder-band dataset, the proposed algorithm reduced feature from 40 to 19; the classification accuracy using Naïve Bayes is 80.9. The accuracy is higher compared to others methods. But the accuracy is the same as other methods using J48 classifier in Cylinder-band dataset. The classification results on each feature selection methods by using Naive Bayes and J48 classifiers are shown in "Table 3" and "Table 4".

Table 2. Number of selected feature by feature selection methods

No	Datasets	IG	SU	Relief-F	Prpposed Algorithm
1	Vehicle	17	19	19	14
2	Page-block	10	11	8	8
3	Sonar	8	22	45	11
4	Liver-disorder	2	2	1	2
5	Cylinder-band	4	21	16	19
6	Waveform	18	20	18	8

Table 3. Classification results on Naïve bayes classifier using 10 fold cross validation

No	Dataset	IG%	SU%	Relief-F%	Proposed Algorithm%
1	Vehicle	61.70	62.65	62.64	64.40
2	Page-blocks	90.8	90.8	91.6	95.5
3	Sonar	78.40	69.7	85.57	87.98
4	Liver-disorder	56.5	56.5	58.0	63.2
5	Cylinder-band	72.2	73.1	71.1	80.9
6	Waveform	80.1	80.1	79.9	81.4

Table 4. Classification results on J48 classifier using 10 fold cross validation

No	Dataset	IG%	SU%	Relief-F%	Proposed Algorithm%
1	Vehicle	72.5	72.2	72.8	73.3
2	Page-blocks	96.9	96.9	96.7	96.9
3	Sonar	77.9	77.9	73.1	86.05
4	Liver-disorder	60.9	60.9	58.0	60.9
5	Cylinder-band	57.8	57.8	57.8	57.8
6	Waveform	76.2	76.2	76.6	76.6

6 Conclusion

We compare different feature selection methods which are newly proposed and tested with public data. Naïve Bayes and J48 classifiers with different feature selection methods are shown in this paper. Existing feature selection algorithms cannot produce effective feature subset for classification in several different areas. Although some algorithm can reduce feature more, their classification accuracy is not good. The proposed conditional mutual information based feature selection algorithm produces the effective and small features with higher classification accuracies in several different datasets. Our proposed algorithm improve the machine learning task by extracting the relevant and effective feature set from original feature set.

References

1. H. Liu and R. Setiono, "A probabilistic approach to feature selection - A filter solution," the *13th International Conference on Machine Learning*, pp. 319-327, 1996.
2. M. Hall, "Feature Selection for Discrete and Numeric Class Machine Learning", Department of Computer Science.
3. A. Suebsing and N. Hiransakolwong, "Euclidean-based Feature Selection for Network Intrusion Detection", International Conference on Machine Learning and Computing IPCST, 2011.
4. Z. Karimi and M. Mansour and A. Harounabadi "Feature Ranking in Intrusion Detection Dataset using combination of filtering", International Journal of Computer Applications,Vol. 78, September 2013.
5. A. Chaudhary, S. Kolhe and Rajkamal, "Performance Evaluation of feature selection methods for Mobile devices", ISSN: 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp. 587-594.
6. M. Robnik and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", Machine Learning Journal, 2003.
7. http://weka.sourceforge.net/doc.dev/weka/attributeSelection