Feature Weighting Method Based On Instance Correlation Using Discretization

E. Chandra Blessie  
Department of Computer Applications,  
Nehru College of Management,  
Coimbatore, Tamil Nadu, India

E. Karthikeyan  
Department of Computer Science,  
Government Arts College,  
Udumalpet, Tamil Nadu, India

Abstract: In Machine Learning Process, several issues arise in identifying a suitable and quality set of features from which a classification model for a particular domain to be constructed. This paper addresses the problem of feature selection for machine learning through discretization approach. RELIEF is considered to be one of the most successful algorithms for assessing the quality of features. RELIEF algorithm selects the near instance and far away instance and assigns weight to the selected instance by sampling method. Sampling method will deviate in selecting the relevant features. So, a new feature weighting method is proposed which gives high correlation between the instances and the main objective is to select features that have highly correlated instances with the class. Experimental analysis shows better performance of the new algorithm in comparison with the existing RELIEF algorithm. The data set is taken from UCI ML repository for experiment. Results show that the new method can be successfully used with classifiers.

Keywords: Machine Learning, Discretization, Classification, RELIEF algorithm, Correlation.

I. INTRODUCTION

Many real world data contains historic and irrelevant information. The performance of the classifier will degrade with the irrelevant information. Feature selection technique solves this problem. In machine learning, one of the preprocessing steps used is feature selection. It is a procedure of selecting a subset of features so that the feature set is optimally reduced.

Feature weighting can be viewed as a generalization of feature selection. In feature selection, feature weights are restricted to 0 or 1 (a feature which is used or not). Feature weighting allows finer differentiation between features by assigning each a continuous valued weight. Relief is an algorithm that uses an instance based approach to assign weights to features.

Kira and Rendell [1] proposed RELIEF algorithm which is an efficient and capable algorithm in assessing the quality of significant and relevant features. RELIEF deals with both discrete and continuous features, but it cannot deal with incomplete data and is limited to two-class problems only. RELIEFF, the extension of RELIEF algorithm deals with multi-class problem, and also noisy and incomplete data.

In this paper, we suggest a new Discretization based feature weighting method in order to avoid the use of sampling technique and to avoid static input of sample size parameter. For continuous features, instead of taking samples of size $m$ for estimating the relevance of features, Discretization can be applied to partition the feature into $n$ intervals and an instance can be taken from each of the $n$ intervals [6]. The main objective is to select the highly correlated instances with the class for assigning weight.

The paper is organized as follows. Literature review is given in section II. In section III, the proposed method is discussed and the pseudo code of the new method is given. Performance evaluation by experiment is given in section IV. Conclusion is given in section V.

II. LITERATURE REVIEW

Feature selection or attribute selection is a technique used to reduce the number of features in a high dimensional data set. By reducing the number of features in the dataset, the data mining algorithm’s accuracy, efficiency, and scalability can be improved [2]. The two main approaches to feature selection are the filtering and wrapper methods. In the filtering method, the features are selected independently to the data mining
algorithms used. Features considered irrelevantly will be filtered out [3]. The wrapper method selects features by using the data mining algorithm selected as a function in the evaluation process [3].

RELIEF and the extended versions first select samples of size m from the training dataset. Then, an instance was chosen randomly from the sample of data to find the nearest hit and nearest miss instance. RELIEF [1], RELIEFF [3] and many algorithms use random sampling technique. In [5], selective sampling is used with kd-tree. Moreover, the frequency in sampling selected is also with uncertainty [5]. The author [4] proposed mean-variance model to select the instances. Mean and variance will deviate the weightage values.

RELIEF ALGORITHM

RELIEF algorithm is the predominant algorithm in filter model. It is efficient in estimating the quality of features. The basic idea of the original RELIEF algorithm is to estimate the quality of features according to how well their values distinguish between instances that are near to each other.

RELIEF searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from the different class, called nearest miss M. The next step is to calculate the differences of features between instances and updates the weights. The features having weights greater than the threshold are selected as relevance features. The weights are calculated by using the formula,

\[ W(F) = W(F) - \frac{\text{diff}(F, X, H)}{m} + \frac{\text{diff}(F, X, M)}{m} \]  

III. PROPOSED CONCEPT

Discretization based feature weighting method is introduced which will select the instances that are highly correlated with the class. Discretization [7] is a technique to group continuous features into a finite set of adjacent intervals in order to generate features with a small number of distinct values. A single value is assigned for each interval. Discretization of features can reduce the learning complexity and help to understand the dependencies between the features and the class feature [7].

A new feature weighting method is proposed. In this, the instance from each interval is selected. The weightage is calculated by finding the difference between the selected instance from each interval with the instances in the most common class and instances with the less common class. This method will increase the correlation between the instances in each class.

In the proposed feature weighting method, the weights are calculated by using the formula,

\[ W(F) = \max\left\{ W_i(\text{diff}(F, X, \sum_{k=1}^{i} (x_k)/m) + \text{diff}(F, X, \sum_{k=1}^{j} (x_k)/m)) \right\} \]  

where \( n \) is the number of instance in the most common class and \( j \) is the number of instances in the less common class, \( X \) is the instance selected from each interval and \( i \) is the number of intervals.

The key ideas of the proposed feature weighting concept are

- The instances of each features is divided into intervals by Discretization technique.
- The Number of instances for calculating the weight is equal to the number of intervals obtained. No need for the expert to specify the sample size parameter.
- In each interval, the difference between the instance values with the most common class and the difference between the instance values with the less common class is taken.
- Weightage is calculated in each interval between the instance selected in the interval and the differences.
- The maximum weight in the intervals is assigned as weight for the feature set.
- The features which has the updated weight greater than the threshold is selected as the relevant feature.

ALGORITHM: PROPOSED ALGORITHM

Let \( D \) be the training dataset with continuous features \( F_i \) S classes.

For every \( F_i \), do:

**STEP 1**

- Find highest \( (d_a) \) and lowest \( (d_b) \) values
- sort all distinct values of \( F_i \) in ascending order
- Find the cut points where the values have different class value and initialize all possible interval boundaries \( B = [d_a, d_b, [d_{a1}, d_{b1}] \ldots [d_{an}, d_{bn}]] \)

**STEP 2**

- For continuous features \( F_i \), assign weight to 0.0 \( W(F_i) = 0.0 \)
- For every \( [d_a, d_{ai}] \) in \( B_i \), where \( i = 1..n \) is the initial boundary points and \( j \) is the number of intervals
- let first instance \( x_i \) be the instance \( i \) in \( [d_a, d_{ai}] \)
- Take the sum of the instances with the most common class \( n \)
  \[ \sum_{k=1}^{n} (x_k) \]
- Take the sum of the instances with the less common class \( j \)
  \[ \sum_{k=1}^{j} (x_k) \]

**STEP 3**

- Find the difference between the instance and the added instance values \( \text{diff}(F, X, \sum_{k=1}^{n} (x_k)/m) \) and \( \text{diff}(F, X, \sum_{k=1}^{j} (x_k)/m) \)
- Update the weight for the features \( W(F_i) = \max\left\{ W_i(\text{diff}(F, X, \sum_{k=1}^{n} (x_k)/m) + \text{diff}(F, X, \sum_{k=1}^{j} (x_k)/m)) \right\} \)
  where \( n \) is the number of instance in the most common class and \( j \) is the number of instances in the less common class, \( X \) is the instance selected from each interval and \( i \) is the number of intervals.

**STEP 4**

- if \( W(F_i) > \tau \) The features are selected as relevant features.
IV. EXPERIMENTAL ANALYSIS

EXPERIMENTAL RESULT

The experiment is implemented using MatLab 7.0. The Breast cancer dataset has 9 features and 100 instances. 25 instances are taken as sample size \( m \) for RELIEF algorithm. In the proposed method, only 15 instances are selected as sample instances. This reduces the computational time as the number of times finding the difference between the selected instance and other instances. 7 features were selected as relevant features while using RELIEF algorithm. When using the proposed method, only 4 features are selected, which will lead to short decision tree. This results in better accuracy by decision tree technique.

For diabetes dataset, 25 sample sizes and 7 features are selected while using RELIEF and the proposed concept selects 9 samples with 4 features. Figure 1 shows the feature quality estimation using four dataset.

![Figure 1: Feature quality estimation](image)

ACCURACY PERFORMANCE

From table 1, the proposed method outperforms the existing method RELIEF using C4.5 classifiers.

<table>
<thead>
<tr>
<th>Dataset/Methods</th>
<th>RELIEF</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>94.42%</td>
<td>94.602%</td>
</tr>
<tr>
<td>Diabetes</td>
<td>74.12%</td>
<td>75.35%</td>
</tr>
</tbody>
</table>

Table 1: Accuracy performance using C4.5 Classifiers

V. CONCLUSION

Sampling method is used to estimate the weight for ranking the features. This new proposed method uses discretization and correlation between instances which will yield better relevant features. The experimental analysis shows that the proposed concept works better than the existing algorithm. This method can also be used for multivalued class features as future extended work.

REFERENCES