A Novel Web Page Classification Model using an Improved k Nearest Neighbor Algorithm

J. Alamelu Mangai, Satej Milind Wagle, and V. Santhosh Kumar

Abstract—With the World Wide Web (WWW) being used to its full potential, automatic classification of web pages into web directories has become more significant. These web directories help the search engines to provide users with relevant and quick retrieval results. In this paper a novel approach to web page classification is implemented by combining the k nearest neighbor classifier (kNN) and association rule mining algorithm. The web pages are preprocessed and discretized before inducing the classifier. The proposed method for web page classification uses a) a feature weighting scheme based on association rules and b) a distance weighted voting scheme. This distance weighted voting scheme enables the model to work for any value of k, being odd or even. Experiments done on a benchmarking data set namely, WebKB have shown that the web page classification accuracy by the proposed method is significantly better than many of the existing web page classification methods.

Keywords—web page classification, discretization, kNN classifiers, association rules, WebKB.

I. INTRODUCTION

WEB Page classification (WPC) is the technique of segregating web pages into specific categories so as to group together web pages of similar significance. Through classifying web pages, retrieval of precise and exact information is possible, which would not only enable faster searching of data but also maintain quality of information retrieved. WPC has strong connection with natural language processing, data mining, text mining, machine learning, and information retrieval and knowledge management. Many machine learning algorithms have been tweaked for web page classification. The k nearest neighbor (kNN) classifier and association rule mining are the two popular data mining classification methods. The kNN classification algorithm is simple and easy to implement. However it suffers from many issues like 1) there is no systematic approach for choosing the best value of k, 2) the simple majority voting scheme degrades the classification accuracy, whenever there is an equal class distribution. Association rule mining task generates rules that help to identify associations/correlations between items in a transaction database. These rules are generated using two interestingness measures called min_support and min_confidence as threshold. In this paper, the performance of the kNN classifier is improved using a new feature weighting scheme and a new distance weighted voting scheme. The feature weighting scheme uses the rules generated for the web page data set using min_support and min_confidence as threshold. The rest of the paper is organized as follows: Section 2 highlights the related work, proposed work is described in Section 3, details of the experiments done are summarized in Section 4, and Section 5 highlights the results and findings.

II. RELATED WORK

Many approaches for automatic WPC have been witnessed over years in literature. The structure of the web document and the images present in them are used to classify them in to various categories in [1]. The performance of the web page classifier is improved using feature selection subsets. A minimum number of highly qualitative features are found by integrating cfs subset evaluator with term frequency method [2] and Ward’s minimum variance [3]. The association between the blocks in a web page [4] is used to frame a query with content based classification framework to classify a web page. Visual features of a web page like color and edge histograms, Gabor and texture features [5] summaries generated by human experts are used in [6]. These approaches of web page classification cannot be applied in situations which suffer from hardware and software limitations. Further, they require lot of human expertise and are computationally complex. The various technologies in web information extraction have been explored in [7] and the authors have expressed their concern that many researchers start with the complex approaches directly rather than trying out the simpler ones first. It is proved in [8] that Naïve Bayes, NB and C4.5 decision tree models are fast consistent, easy to maintain and accurate in the training courses domain. NB classifier based on Independent Component Analysis [9], Hidden Naïve Bayes [10] with Symmetrical Uncertainty for word selection perform better. These approaches are summarized in Section 4, and Section 5 highlights the results and findings.
k-means clustering to improve the performance of the kNN classifiers [12]. A survey of improving kNN for classification is presented in [13]. Genetic Algorithms is combined with traditional kNN to identify the k nearest neighbors to the test sample directly, rather than comparing the test data with every training data [14]. The performance of kNN classifiers is improved by a feature weighting scheme using association rules [15]. However the authors have evaluated the algorithm on structured data sets from UCI repository.

Motivated by these facts this paper proposes a method for WPC using the content of the web page. The web page classification model proposed uses an improved version of the traditional kNN. The distance measure is improved by a feature weighting scheme using association rules. This work also uses a distance weighted measure instead of simple majority voting [16], to avoid equal class probability distribution.

III. PROPOSED WORK

In this work web page classification is implemented in two steps namely, preprocessing and classifier induction. The preprocessing steps involve feature extraction, feature selection and discretization. The highlights of all the preprocessing steps are stated below.

A. Preprocessing

1) Feature Extraction

As web pages are high dimensional data, if features are not properly selected, the time taken to model the classifier will increase. This work attempts to minimize the time taken to build the classifier by using a series of preprocessing steps. The preprocessing steps for improving the performance of web page classifiers are illustrated below.

- Each web page is converted to a text file.
- The best features from each web page are extracted and web page – feature matrix is constructed. Each feature is weighted using its term frequency-inverse document frequency. These features will be the best representative features of a web page category. Each web page is represented as a vector with the weight of a feature in a web page calculated as

\[
  w_{ij} = \frac{tf_{ij} \cdot idf_i}{max\{f_j\}} = \frac{tf_{ij} \cdot \log_2 (N/df_i)}{
\]

where \( N \) is the number of web pages in the collection. \( tf_{ij} \) is the frequency of term \( i \) in web page \( j \), \( df_i \) is the web page collection frequency of a feature. \( max\{f_j\} \) is the frequency of the most common term in the web page.
- The web pages which have all feature weights as zero are removed.

- Duplicate and conflicting web pages are also removed.

2) Feature Selection

The number of features is further reduced using a novel technique using Ward’s minimum variance measure as stated in our earlier work [3]. The Ward’s measure is used to identify clusters of redundant features. The best representative feature in each such cluster is retained and the others are removed.

3) Discretization

After feature selection, the web pages are transformed into a n-dimensional vector, where n is the number of features. The weights of each feature are continuous values. These are transformed into discrete values using a discretization algorithm as in [17].

B. Proposed Classification Framework

The traditional kNN classifier is improved for web page classification using a feature weighting scheme using the interestingness measures \( min_{sup} \) and \( min_{conf} \) of the association rules. For a rule of the form, ithaca=16 →class=student, support of the rule gives the fraction of transactions having both the rule’s antecedent and consequent. Confidence of the rule is the ratio of number of transactions having both antecedent and consequent to the number of transactions having the rule’s antecedent. Classifying a new web page involves two steps namely 1) calculate the feature weights using association rules and 2) predict the category of the test web page using the modified k nearest neighbor algorithm.

1) Calculate the feature weights

Input: Minimum support \( min_{sup} \), minimum confidence \( min_{conf} \), the web page feature vectors with discrete features and \( n \) the number of features.

Output: The weight of each feature in a vector, \( weight \).

Method

1. Generate association rules with every feature value and web page category combination.
2. Calculate the support and confidence of each rule.
3. Find the maximum support \( max_{sup} \) and maximum confidence \( max_{conf} \) of each attribute.
4. for \( i = 1 \) to \( n \) number of features repeat
   4.1 if \( max_{sup} < min_{sup} \), or \( max_{conf} < min_{conf} \), then \( weight_i = 0 \)
   else \( weight_i = 1/(1 - max_{sup} \)

2) Improved kNN for classifying a new web page

Input: The feature weight vector weight, the web page feature vector with numeric features, value of \( k \) and the test web page.

Output: The predicted category of the test web page.

Method:

1. Find the \( k \) nearest neighbors to the test web page using the feature weighted distance formula as below

\[
distance(X,Y) = \sqrt{\sum_{i=1}^{n} weighted\_distance((x_i - y_i)^2)}
\]

2. A distance weighted voting on the class of the \( k \) nearest neighbors is used to predict the category of the test web page.

a. Assign a weight to each of the \( k \) nearest neighbor as below

\[
w_i' = \begin{cases} 
  \frac{d(x', x_i^{NW}) - d(x', x_k^{NW})}{d(x', x_k^{NW}) - d(x', x_1^{NW})} & \text{if } d(x', x_k^{NW}) \neq d(x', x_i^{NW}) \\
  1 & \text{if } d(x', x_k^{NW}) = d(x', x_i^{NW})
\end{cases}
\]
where \( x' \) is the test web page, \( x_k^{NN} \) is the \( k^{th} \) nearest neighbor to the test web page, \( x_1^{NN} \) is the first nearest neighbor to the test web page and \( x_i^{NN} \) is the \( i^{th} \) nearest neighbor to the test web page.

b. Predict the class \( y' \) of the test web page as

\[
y' = \arg \max_y \sum \{w_i' \cdot \delta(y = y_i^{NN}) \}
\]

The function \( \delta(y = y_i^{NN}) \) is defined as:

\[
\delta(y = y_i^{NN}) = \begin{cases} 
1, & y = y_i^{NN} \\
0, & y \neq y_i^{NN}
\end{cases}
\]

where \( y_i^{NN} \) is the class label of the \( i^{th} \) nearest neighbor, \( y \) is the set of all class labels.

IV. EXPERIMENTS AND RESULTS

Experiments were done on a benchmark data set called WebKB [18]. This data set contains WWW pages collected from computer science departments of various universities.

TABLE I. THE DATA SET AFTER PREPROCESSING

<table>
<thead>
<tr>
<th>Input Size</th>
<th>No. of Instances</th>
<th>No. of Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>70 - 30</td>
<td>56</td>
<td>5</td>
</tr>
<tr>
<td>100 - 100</td>
<td>92</td>
<td>6</td>
</tr>
<tr>
<td>200 - 200</td>
<td>291</td>
<td>13</td>
</tr>
<tr>
<td>300 - 200</td>
<td>298</td>
<td>9</td>
</tr>
<tr>
<td>300 - 300</td>
<td>414</td>
<td>14</td>
</tr>
<tr>
<td>350 - 150</td>
<td>391</td>
<td>13</td>
</tr>
<tr>
<td>400 - 200</td>
<td>422</td>
<td>15</td>
</tr>
<tr>
<td>400 - 300</td>
<td>557</td>
<td>17</td>
</tr>
<tr>
<td>400 - 400</td>
<td>585</td>
<td>17</td>
</tr>
</tbody>
</table>

For the analysis of the proposed work course web pages are considered as positive examples and student web pages as negative examples. Table I shows the details of the various web page collections after feature extraction and feature selection. The second and third column shows the effect of applying various pre-processing steps on the data sets for dimensionality reduction. The numbers of features are further reduced using Wards minimum variance measure. Table II shows the results of feature selection done on the preprocessed data sets using Wards minimum variance measure.

The discretization algorithm used, transforms the range of values of each numeric feature into a specified number of bins automatically. Table III shows a portion of the 100-100 data set before and after discretization in ARFF (attribute relation file format), supported by the data mining tool, WEKA [19]. The weights of each feature are then calculated using association rules with min_sup and min_conf as thresholds.

The kNN classifier with the modified distance formula is then modeled using 70 – 30% split. The web page classification accuracy of this modified kNN which uses distance weighted voting is compared with the traditional kNN with simple majority voting.

\[ \text{Fig. 1 Comparison of Average Classification Accuracy between kNN and MkNN} \]
TABLE IV. COMPARISON OF CLASSIFICATION ACCURACY BETWEEN kNN AND MkNN

<table>
<thead>
<tr>
<th>Value of k</th>
<th>kNN</th>
<th>MkNN</th>
<th>kNN</th>
<th>MkNN</th>
<th>kNN</th>
<th>MkNN</th>
<th>kNN</th>
<th>MkNN</th>
<th>kNN</th>
<th>MkNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-30</td>
<td>82.35</td>
<td>100</td>
<td>82.35</td>
<td>100.00</td>
<td>82.35</td>
<td>100.00</td>
<td>82.35</td>
<td>100.00</td>
<td>82.35</td>
<td>100.00</td>
</tr>
<tr>
<td>100-100</td>
<td>82.14</td>
<td>96.43</td>
<td>82.14</td>
<td>96.43</td>
<td>82.14</td>
<td>96.43</td>
<td>82.14</td>
<td>96.43</td>
<td>82.14</td>
<td>96.43</td>
</tr>
<tr>
<td>200-200</td>
<td>88.51</td>
<td>92.05</td>
<td>88.51</td>
<td>92.05</td>
<td>89.66</td>
<td>92.05</td>
<td>89.66</td>
<td>92.05</td>
<td>90.80</td>
<td>92.05</td>
</tr>
<tr>
<td>300-200</td>
<td>93.26</td>
<td>93.33</td>
<td>93.26</td>
<td>93.33</td>
<td>92.13</td>
<td>93.33</td>
<td>91.01</td>
<td>93.33</td>
<td>91.01</td>
<td>94.44</td>
</tr>
<tr>
<td>300-300</td>
<td>95.97</td>
<td>92.74</td>
<td>95.97</td>
<td>93.55</td>
<td>95.97</td>
<td>93.55</td>
<td>96.77</td>
<td>93.55</td>
<td>96.77</td>
<td>93.55</td>
</tr>
<tr>
<td>350-150</td>
<td>95.73</td>
<td>100</td>
<td>95.73</td>
<td>100.00</td>
<td>96.58</td>
<td>100.00</td>
<td>96.58</td>
<td>100.00</td>
<td>96.58</td>
<td>100.00</td>
</tr>
<tr>
<td>400-200</td>
<td>96.92</td>
<td>95.38</td>
<td>96.15</td>
<td>95.38</td>
<td>96.15</td>
<td>96.15</td>
<td>96.15</td>
<td>96.15</td>
<td>96.15</td>
<td>96.15</td>
</tr>
<tr>
<td>400-300</td>
<td>94.61</td>
<td>94.41</td>
<td>94.61</td>
<td>94.41</td>
<td>94.61</td>
<td>97.01</td>
<td>94.61</td>
<td>97.01</td>
<td>95.81</td>
<td>97.60</td>
</tr>
<tr>
<td>400-400</td>
<td>92.57</td>
<td>95.43</td>
<td>93.71</td>
<td>96.00</td>
<td>93.71</td>
<td>96.00</td>
<td>94.29</td>
<td>96.00</td>
<td>94.29</td>
<td>96.57</td>
</tr>
<tr>
<td>Average</td>
<td>91.34</td>
<td>95.75</td>
<td>91.38</td>
<td>95.91</td>
<td>91.48</td>
<td>96.06</td>
<td>91.51</td>
<td>96.12</td>
<td>91.85</td>
<td>96.31</td>
</tr>
</tbody>
</table>

As seen in Fig 1, for smaller values of k, the performance of kNN and MkNN is not significantly varying. However, as k increases, due to the combined attribute and distance weighted voting schemes, the performance of MkNN drastically improves over kNN. Table IV shows a comparison of the average classification accuracy of kNN and MkNN. For higher values of k, the proposed method of web page classification using MkNN is better than traditional kNN. The proposed method is also compared with some of the other existing classification methods like decision tree (J48), Naïve Bayes(NB), k-nearest neighbor (kNN with k = 13), Neural network (NN) and support vector machine based classifier(SMO). Table V shows a comparison of classification accuracy of all these methods.

TABLE V. COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER EXISTING CLASSIFIERS

<table>
<thead>
<tr>
<th>Data Set</th>
<th>J48</th>
<th>NB</th>
<th>KNN</th>
<th>NN</th>
<th>SVM</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-30</td>
<td>100.00</td>
<td>100.00</td>
<td>82.35</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100-100</td>
<td>92.86</td>
<td>92.86</td>
<td>89.66</td>
<td>88.51</td>
<td>92.05</td>
<td></td>
</tr>
<tr>
<td>200-200</td>
<td>85.06</td>
<td>89.00</td>
<td>92.13</td>
<td>91.01</td>
<td>94.44</td>
<td></td>
</tr>
<tr>
<td>300-300</td>
<td>91.13</td>
<td>95.97</td>
<td>95.16</td>
<td>96.77</td>
<td>93.55</td>
<td></td>
</tr>
<tr>
<td>350-150</td>
<td>90.50</td>
<td>90.44</td>
<td>94.68</td>
<td>94.61</td>
<td>97.01</td>
<td></td>
</tr>
<tr>
<td>400-200</td>
<td>90.00</td>
<td>92.31</td>
<td>96.00</td>
<td>96.15</td>
<td>96.15</td>
<td></td>
</tr>
<tr>
<td>400-300</td>
<td>87.43</td>
<td>93.41</td>
<td>95.21</td>
<td>95.21</td>
<td>97.60</td>
<td></td>
</tr>
<tr>
<td>400-400</td>
<td>90.86</td>
<td>94.86</td>
<td>93.71</td>
<td>94.29</td>
<td>96.57</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>90.98</td>
<td>94.42</td>
<td>91.42</td>
<td>94.34</td>
<td>96.31</td>
<td></td>
</tr>
</tbody>
</table>

As seen in Table V, the accuracy of the NN model is good for smaller input sets. But, for larger input data sets, it took a longer time to model the classifier. Based on the average classification accuracy, the proposed method performs better than other methods. It is also simple and easy to interpret and implement.

V. CONCLUSIONS

As WWW has become a huge information repository, automatic classification of web pages and maintaining web directories helps the search engines to provide quick and relevant retrieval results for a user query. In this paper, web page classification is implemented by using a modified k-nearest neighbor, MkNN algorithm. This uses an attribute weighting scheme that assigns weights to each attribute using association rules generated from the training web pages. Further, a distance weighted voting scheme that assigns more weight to the most nearest neighbor to the test web page has helped to improve the classification accuracy. The proposed algorithm is tested on web pages from WebKB repository. The experimental results have shown that the proposed method of web page classification has significantly good classification accuracy than many other existing WPC methods.

REFERENCES


