The Influences of Number and Nature of Classifiers on Consensus Feature Selection

Kyriacos Chrysostomou, Sherry Y. Chen, and Xiaohui Liu
School of Information Systems, Computing and Mathematics,
Brunel University, Uxbridge, Middlesex, UB8 3PH, UK

Abstract - Wrapper feature selection approaches are widely used to select a small subset of relevant features from a dataset. However, Wrappers suffer from the fact that they only use a single classifier when selecting the features. The downside to this approach is that each classifier will have its own biases and will therefore select very different features. In order to overcome the biases of individual classifiers, we propose Consensus Feature Selection (CFS), which combines different classifiers for feature selection. In this way, selecting classifiers for use in the combinations is very important. Therefore, we investigate how the number and nature of classifiers influence the number of features selected and the classification accuracies that these features generate. In terms of number of classifiers, results showed that few selected more relevant features whereas many selected few features. In addition, 3-classifier combinations selected features that led to highest accuracies. In terms of nature of classifiers, decision trees identified most number of features whereas Bayesian classifiers identified least number of features. However, features selected by Bayesian classifiers led to accuracies higher than the other classifiers.

Keywords: Feature selection, wrappers, consensus, decision tree, Bayesian network.

1 Introduction

The performance of most data mining algorithms can be deteriorated by features that do not add any value to learning tasks. Feature selection can be used to limit the effects of such features by seeking only the relevant subset from the original features [1]. This subset of the relevant features is discovered by removing those that are considered as irrelevant or redundant. By reducing the number of features in this way, the time taken to perform classification is significantly reduced; the reduced dataset is easier to handle as fewer training instances are needed (since fewer features are present) [2], subsequently resulting in simpler classifiers which are often more accurate.

Because of such benefits, a lot of researchers have devoted their time to developing techniques for performing feature selection. Basically, these techniques can be broadly divided into Filters [3] and Wrappers [4]. In general, previous research indicated that Wrappers tended to perform better than Filters. For example, [5] compared the performance of Wrapper and Filter methods for categorical clustering over several datasets. Their results showed that Wrapper methods were superior to Filters. Other research such as [6] also found similar results. Wrapper methods are useful, but they only use one classifier to select the most relevant set of features. The problem of using a single classifier is that each classifier is of a different nature and will have its own biases. Due to these differences, each classifier will select a different feature subset which may contain different number of features and may also lead to different levels of accuracy. In other words, using a single classifier for feature selection may affect the number of features selected and the accuracy levels of the features. Therefore, there is a need to consider combining several different classifiers. By doing so, the biases of each individual classifier can be overcome by deriving a consensus from the features selected by these different classifiers.

In this vein, we propose a new data mining method called Consensus Feature Selection (CFS). CFS uses multiple classifiers to conduct feature selection. In other words, CFS combines the selected features from different classifiers in order to obtain a mutually agreed set of relevant features. As such, the selection of classifiers is very important in CFS. More specifically, the number and the nature of classifiers may influence relevant features selected. To this end, this study investigates two issues: 1) how the number of classifiers influences the number and accuracy of selected features, and 2) how the nature of a classifier influences the number of and accuracy of selected features. These two issues are examined using a dataset consisting of users’ preferences for the design of search engines. Specifically, the dataset is fed as input to the proposed CFS method, which is detailed in Section 2, in order to select relevant sets of features. The results from these sets of relevant features are presented in Section 3 with particular emphasis on how the number and nature of classifiers influence number of features present in each set and the classification accuracies generated using the selected features. Finally, conclusions based on these results are outlined in Section 4.
2 Consensus Feature Selection (CFS)

As discussed in Section 1, CFS uses multiple classifiers. In this study, four classifiers are applied, namely Bayesian Networks (BN) [7], Decision Trees (DT) [8], k-Nearest Neighbor (KNN) [9] and Support Vector Machines (SVM) [10]. These four classifiers were chosen because of their different biases, for example, BN focuses on features that maximize/minimize some scoring metric when building a network structure, whereas KNN focuses on features and instances that are deemed the ‘closest’ by some imposed distance metric.

To illustrate how the aforementioned classifiers perform feature selection, KNN will be used as an example. In this example, we use the KNN classifier to select the relevant features from preferences of 120 users (instances) collected from a questionnaire that included 90 statements (features), each of which had five possible options ‘very unimportant’, ‘unimportant’, ‘neutral’, ‘important’, and ‘very important’. Relevant features will be selected from these 90 features for a particular target variable which relates to users’ demographic details, i.e., gender. Initially, the KNN uses all user preference data as input. The classifier is then executed to identify the subset of relevant features from user preferences in relation to gender variable (predictor) by using a search strategy, which outlines the way in which feature subsets are searched within the entire set of features. Among the many types of strategies, we chose to use the forward search strategy. The forward search strategy guides the classifier in a forward manner so that few features are initially used and then more are continuously added to form the feature subsets. The benefit of the forward search strategy is that it is effective at selecting feature subsets which are of greater relevance to the target variable than feature subsets selected by other search strategies like backward elimination [11].

Once identified, the feature subsets are evaluated according to an accuracy estimation technique. In this study, $k$-fold cross validation was used as the accuracy estimation technique. Typically, $k$-fold cross validation involves splitting the training data into $k$ approximately equally sized partitions. The chosen classifier is then run $k$ times using $k$-1 partitions as the training set and the remaining partition as the test set. The accuracy results from each of the $k$ runs are then averaged to produce the estimated accuracy. In this study, we choose $k$ to be 10 because this value of $k$ has been widely used when applying cross validation to different data mining tasks [12]. Each of the features will be assigned a value, which indicates the number of times they were chosen by 10-fold cross validation. This value is the output of the process of feature selection, which reveals the level of the relevance of this feature (outcome variable). For example, a feature that was assigned the value 10 by KNN implies that this feature was included in all 10 times/samples of the cross validation procedure. In other words, this feature is highly relevant to the target variable, i.e., gender.

The selected sets of relevant features will then form a matrix where each row will represent a feature contained in the dataset and each column will present the different classifier used. The contents of the matrix included the ranked values of each feature by each classifier, ranging from 0 (indicating that the feature was not selected) to 10 (indicating the feature was always selected by cross validation). In general, the main purpose here is to find agreement between features selected by the classifiers. In this case, the agreement can be calculated using different combination strategies. These can include calculating the maximum value of a feature’s rank, the minimum value, the sum, the mean, and the median. Among these strategies, both the mean and median seemed suitable for our approach because the results of each classifier can be within the range of 0 to 10, which corresponds to the nature of 10-fold cross validation. However, we are dealing with user preference data, which tend to be of a fuzzy nature. Since an outlier can seriously distort the result of the mean, and [13] showed that the median strategy exhibited the best classification results, the median strategy has been adopted in our approach. In this study, we not only take into account features with high relevance, but also include those with low relevance so that an entire set of relevant features can be considered. More specifically, all features that have a median value of 1 or more are chosen as consensus relevant features. These features are then classified using decision tree classifier so as to determine their classification accuracies, i.e., how relevant the features are to the target variable. Decision tree classifier was chosen for this task because it has been showed to provide good classification performance with user preference data [14].

3 Results and Discussion

3.1 Number of Classifiers

As described in Section 1, this study investigates how the number of and nature of classifiers influence relevant features selected. This section presents the results related to the effects of the number of classifiers on feature selection. Initially, four different classifiers were adopted to perform the feature selection. These four classifiers were chosen because they represent and belong to the four most popularly used classifier families, i.e., BN, DT, KNN and SVM. In addition, using classifiers from these very different families will help us select different features that can then be combined with CFS to see their influences on feature selection. The four classifiers were combined using an exhaustive approach so that each classifier was used with every other classifier. This will help identify how combinations with different numbers of classifiers affect the features selected. This exhaustive approach led to the construction of 2-classifier combinations (BN+DT, BN+KNN, BN+SVM, DT+KNN, DT+SVM, and KNN+SVM), 3-classifier combinations (BN+DT+KNN, BN+DT+SVM, BN+KNN+SVM, and DT+KNN+SVM), and a 4-classifier combination where all the four classifiers were
combined (BN+DT+KNN+SVM). The number of relevant features chosen by each classifier combination is shown in Table 1. This Table also includes the classification accuracy generated by each combination, which shows how relevant the selected feature subsets are in relation to the target variable. In this case, high classification accuracy implies that the features are very relevant to the target variable whereas low classification accuracy implies that the features are not so relevant. The classification accuracies of the feature subsets selected by CFS shall also be compared to the single classification accuracy generated using the entire set of features present in this study’s dataset, which was found to be 79.17%. In this way, we will be able to determine whether CFS selects feature subsets that are more relevant to the target variable than using all the features in the dataset.

Analyzing Table 1 reveals several interesting issues. These issues are described below.

- **2-classifier combinations select more relevant features**

  On close examination of Table 1, it seems that the 2-classifier combinations identified more relevant features than the 3-classifier and 4-classifier combinations. A possible explanation for this finding may lie within the fact that the median was used to combine the results from different classifiers. The median computes the middle value for a given set of values. In this study, the median was calculated for every feature selected by each combination. To better understand the way in which the median strategy was used, the BN+DT 2-classifier combination will be used as an example for explanation. For a feature, \( f \), BN may show a relevance value of 2, while DT may not find the feature relevant (i.e., 0 is assigned). In this example, the median of \( f \) will be 1. This feature will therefore be classified as a relevant feature because any feature with a median of 1 or more is considered relevant by CFS. This shows that features with low relevance levels have the ability of being classed as relevant, which, in turn, results in more relevant features being identified. On the other hand, using three or four classifiers results in fewer relevant features being identified. This may be due to the fact that each feature will have more relevance values. When using 3-classifier combinations each feature will have three associated relevance values. For a feature to be regarded as relevant, at least two of the three relevance values must be 1 or more so that the resulting median is above 1. This same concept also applies to 4-classifier combinations, but with the exception that three out of the four relevance values for each feature must be 1 or more for a feature to be classed as relevant. Overall, this shows that fewer relevant features can be identified when more than two classifiers are used because more classifiers are required to (mutually) agree on a particular feature.

- **Combinations with BN classifier select fewer relevant features**

  The combinations with the BN classifier select far fewer relevant features. A possible reason for this may lie within restrictions imposed by BN classifiers. BN are parametric classifiers and typically require a priori knowledge of the data for data analyses [15]. This priori knowledge is normally provided by experts. Using the field of Bioinformatics as an example, human biologists are at hand to provide in-depth details about the features and instances presented in genomics data. However, there was no priori knowledge about the data used in this study, except the data itself, so using BN has resulted in fewer relevant features selected.

- **Combinations with DT classifier select more relevant features**

  Combinations with DT classifier selected more relevant features. A deep analysis of the number of features selected by each classifier was conducted, which showed that DT (N=32) selected more relevant features than BN (N=9), KNN (N=22), and SVM (N=21). This may be the reason why combinations that involved the DT classifier were more likely to select a higher number of features. However, the level of tree pruning was found to influence the number of selected features. Two levels of pruning were used, namely relaxed pruning level and strict pruning level. The aforementioned number of features was selected using the DT classifier with the relaxed pruning level. We found the DT classifier with

<table>
<thead>
<tr>
<th>Combinations</th>
<th>BN</th>
<th>DT</th>
<th>KNN</th>
<th>SVM</th>
<th>No. of Relevant Features Selected</th>
<th>Classification Accuracies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-classifier</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>18</td>
<td>79.17</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>80.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>78.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>83.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
<td>82.50</td>
</tr>
<tr>
<td>3-classifier</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>11</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>81.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>83.33</td>
</tr>
<tr>
<td>4-classifier</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>9</td>
<td>82.50</td>
</tr>
</tbody>
</table>

TABLE 1

No. of Features Selected by Different Classifier Combinations and Associated Classification Accuracies
the strict pruning level selected fewer relevant features (N=19). This suggests that the pruning level used by the DT classifier influences the number of relevant features selected.

- **3-classifier combinations generate highest classification accuracies**

According to Table 1, 3-classifier combinations not only generated higher classification accuracies than the accuracy with the original 90 features (79.17%) but also led to accuracies higher than the majority of the other combinations. More specifically, 3-classifier combinations led to the two highest classification accuracies among all combinations. In addition, 3-classifier combinations were found to generate a mean accuracy level (83.75%) higher than the mean accuracy levels of 2-classifier (80.42%) and 4-classifier (82.50%) combinations. Due to such high accuracy levels, features selected by 3-classifier combinations can be said to be more relevant to the target variable. This suggests that using a balance between few (2 classifiers) and many (4 classifiers) classifiers may be more suitable to identify features that are most relevant to the target variable, where the balance is in the form of 3-classifier combinations. Therefore, using three classifiers for CFS may offer a better chance of identifying the most relevant set of features within a dataset.

- **Combinations with DT classifier influence classification accuracies**

In general, we found that combinations with the DT classifier led to higher accuracies than those without the DT classifier. A possible explanation for this may lie within the nature of DT classifiers. DT classifiers are sometimes regarded as another type of feature selection method called ‘embedded methods’ [2]. This means that a feature selection step is ‘embedded’ within the classifier such that the classifier identifies a subset of relevant features from the original set of features and then uses this subset for doing classification. In other words, DT classifiers perform feature selection themselves when classifying data.

The fact that DT classifiers perform feature selection on their own may help identify highly relevant features when used with CFS. This is because features are selected at two different stages. In the first stage, features are selected by the individual DT classifier and in the second stage features are selected by CFS combinations that also include the DT classifier. In this manner, only the features that are selected at both of these stages will form the final feature subset, which is very likely to include features of high relevance. Uncovering highly relevant features with respect to the target variable increases the likelihood of obtaining higher classification accuracies. This may thus explain why combinations with DT generated higher accuracy levels than other combinations.

### 3.2 Nature of Classifiers

This section presents the results related to the effects of the nature of classifiers on feature selection. The nature of three different types of classifier combinations were considered in this section. One is based on the BN family, another on the DT family and the other on the KNN family. For each family, three classifiers of the same nature are used which are showed in Table 2. The reason for choosing and combining these classifiers to do the feature selection is two-fold. First, previous studies have showed that the three classifiers which belong to each family in Table 2 produce very good performance when used to classify user preference data, e.g. [16][17]. Second, combining many classifiers that are of the same nature, like the three combined for each family in this section, will result in very similar features being selected. This is because classifiers of the same nature may have very similar biases which mean that they may behave in similar way, i.e., select similar features. These similar features can thus help reveal the effects of the nature of classifiers on feature selection because the nature of the classifiers used may affect which features are going to be selected. In other words, using many classifiers of the same nature can select features that help provide a better picture of the overall influences of nature of the classifiers on feature selection, compared to using fewer classifiers of the same nature. As such, three classifiers of the same nature were used in each combination.

Apart from the three classifiers of the same nature, each combination also includes the SVM classifier. The SVM classifier was included in all combinations mainly because such classifiers are well known for their highly accurate performance and excellent generalization ability on many different types of datasets [18]. In this way, we cannot only investigate how the nature of different classifiers influences the number and accuracy of selected features, but also we can see if the inclusion and exclusion of the SVM classifier influence the number of and accuracy of selected features by the different combinations.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>No. of Features Selected by Each Individual Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN Family</td>
<td>DT Family</td>
</tr>
<tr>
<td>Bayesian Network (N=9)</td>
<td>C4.5 (N=32)</td>
</tr>
<tr>
<td>Naive Bayes (N=25)</td>
<td>CART (N=27)</td>
</tr>
<tr>
<td>Average One-Dependence Estimators (AODE) (N=17)</td>
<td>CN2 (N=25)</td>
</tr>
</tbody>
</table>

The results obtained using the abovementioned combinations are showed in Table 3. This Table presents the
number of relevant features selected by the combinations with and without the SVM classifier, in addition to the classification accuracy of each combination. On closer examination of this Table, we found several interesting trends. These trends are explained below.

- **More relevant features from the DT family**

As showed in Table 3, more relevant features were obtained from the DT family, regardless of whether the SVM classifier was removed from the combinations. The reason why classifiers belonging to the DT family identified more relevant features than classifiers belonging to the BN and KNN families may lie within the nature of these three types of classifiers. On the one hand, the nature of BN classifiers allows them to identify relationships between features which are typically presented in a graphical network structure. However, as previously uncovered, the number of relationships and thus the number of features identified can decrease when no or little prior knowledge is available about the actual features. On the other hand, the nature of classifiers belonging to the KNN family allows them to determine relevant features by using some distance metric, where the most relevant features are those that are deemed the closest by distance metric. However, the number of most relevant features identified greatly depends on a small number of features, usually referred to as neighbors. This is because the number of neighbors employed represents number of features used to determine the most relevant (i.e., closest) features. For example, using few neighbors may lead to fewer features whereas more neighbors may result in more features. Thus, using an inappropriate number of neighbors may lead to fewer features identified [19].

Conversely, the nature of DT classifiers is very different to the nature of BN and KNN classifiers described above. DT classifiers use a statistically measure to determine the relevance of features, where the most relevant features are those with the highest statistical relevance values [20]. More importantly, the nature of DT classifiers does not require them to have prior knowledge about the features in the data (like BN classifiers) and does not require them to rely on a predetermined number of neighbors for finding the relevant features (like KNN classifiers). The fact that DT classifiers do not need to deal with such issues, which can considerably decrease number of features selected, suggests that DT family combination is more likely to identify a higher number of relevant features than BN and KNN family combinations. This may thus explain why the DT family combination selected more relevant features.

- **More relevant features when removing SVM**

Another interesting trend is that more relevant features were selected when the SVM classifier was removed from the combinations. Such a trend may be explained with regards to the number of classifiers used for feature selection. As previously found, the number of classifiers used for feature selection significantly influenced the number of relevant features identified. Specifically, it was found that fewer classifiers led to more consensus relevant features. Therefore, removing the SVM classifier from each combination resulted in fewer classifiers being used for CFS. Involving fewer classifiers can lead to more consensus features being identified.

- **BN family generates highest classification accuracies**

According to Table 3, the BN family combination generated the highest classification accuracies in comparison to the other classifier families. A possible reason for this may lie within the nature of BN family classifiers. The nature of BN family classifiers typically allows them to capture relationships and dependencies among features in the dataset [21]. However, only the relationships between features with the highest conditional probabilities are considered [22]. In other words, only the features that have high relevance with regards to the target variable are considered. By considering features that are highly relevant to the target variable, BN family classifiers can generate high classification accuracies.

The fact that BN family classifiers generated high accuracies without any prior knowledge about the data used in this study shows that they are not sensitive to this issue. This may be because BN classifiers have a way of dealing with the absence of prior knowledge about features in the data. When no prior knowledge is available, BN family classifiers typically consider each feature to be equally relevant to the target variable [23]. In this way, each feature is treated in the same way which helps avoid any emphasis on certain features in the dataset. This can help limit any

<table>
<thead>
<tr>
<th>Combinations</th>
<th>With SVM classifier</th>
<th>Without SVM classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Relevant</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>Features Selected</td>
<td>Accuracies (%)</td>
</tr>
<tr>
<td>BN Family</td>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>DT Family</td>
<td>13</td>
<td>82.50</td>
</tr>
<tr>
<td>KNN Family</td>
<td>11</td>
<td>84.17</td>
</tr>
</tbody>
</table>
influences features may have on the feature selection process, and in turn the accuracy levels of selected features.

- Higher classification accuracies when removing SVM

As previously found, removing the SVM classifier from combinations led to more features being selected. Interestingly, the removal of the SVM classifier also led to differences in classification accuracies. More specifically, removing the SVM classifier resulted in higher accuracies compared to when the SVM classifier was included in the combinations. It may be due to the fact that removing the SVM classifier reduces the classifiers used in the combination. In other words, the combination is changed from a 4-classifier combination to a 3-classifier combination. As the findings showed in the previous section, using three classifiers for feature selection resulted in feature subsets that produced higher classification accuracies. This may be the reason why removing the SVM classifier increased the accuracies.

4 Conclusion

This study proposes a novel data mining method called Consensus Feature Selection (CFS) that uses multiple classifier families for feature selection, including Bayesian Networks, Decision Trees, Nearest Neighbor and Support Vector Machines. On the one hand, this approach can overcome the biases of each individual classifier. On the other hand, the selection of classifiers is important when using multiple classifiers. To this end, the goal of this study was to investigate two issues: 1) how the number of classifiers influences the number of and accuracy of selected features, and 2) how the nature of a classifier influences the number of and accuracy of selected features.

The answer to the first issue is that using few classifiers for feature selection led to more relevant features, while many classifiers resulted in fewer relevant features being identified. Although the number of classifiers influenced number of features selected, a balance between few and many classifiers was found to be more suitable for feature selection. The balance was in the form of 3-classifier combinations because these combinations selected features that generated classification accuracies higher than the accuracies of the other combinations. The answer to the second issue is that using decision tree classifiers resulted in a higher number of relevant features being selected while using Bayesian network classifiers led to a lower number of relevant features selected. Although the Bayesian network classifiers selected lowest number of features, these features were shown to be more accurate and thus more relevant to target variable than those of other classifier families. In addition, we found that the nature of the SVM classifier does not have substantial effects on the results of the classifier combinations. In general, these results can be used as recommendations in order to help choose the most suitable classifiers when performing feature selection. By doing so, the classifiers chosen will be better suited to a particular type of feature selection task.

Overall, this study has not only showed the influences of the number of and nature of classifiers on feature selection, in particular CFS, but has also showed the benefits of using the CFS method for selecting highly relevant feature subsets. More specifically, CFS was able to select highly relevant feature subsets from a dataset consisting of users’ preferences of search engines. This was proven in Section 3 (see Table 1 and Table 3) because the accuracy levels of the majority of feature subsets selected by CFS were significantly higher than the single accuracy level generated using the entire set of features in the dataset. These highly accurate feature subsets can prove very useful because they will contain those features that will help differentiate the search engine preferences of different users. Such features can be used to develop search engines that better accommodate the preferences of these users. However, this is only a small-scale study. First, only a small selection of data mining classifiers was combined to carry out the feature selection and only one classifier was used for classification. A greater number of classifiers and classifiers with different nature, such as Neural Networks and Evolutionary Algorithms, need to be considered to identify differences in selected features. In addition, it would be interesting to use classifiers from other families for doing the classification so as to determine similarities or differences in the classification accuracies generated. Second, CFS was used to analyze only one type of dataset, i.e., user preference data. It is necessary to use several other datasets, including artificial datasets, so as to test the true effectiveness of CFS within different data mining tasks. The findings from such studies can be integrated into those of the study presented in this paper so that more comprehensive guidance can be provided for the selection of classifiers in CFS.

5 References


microarray data for cancer classification”. Pattern Recognition, 2006, 39, 12, 2383-2392.


