Information and Rough Set Theory Based Feature Selection Techniques

Liam Cervante and Xiaoying Gao

School of Engineering and Computer Science
Victoria University of Wellington, P.O. Box 600, Wellington, New Zealand
{liam.cervante,xgao}@ecs.vuw.ac.nz

Abstract. Feature selection is a well known and studied technique that aims to solve "the curse of dimensionality" and improve performance by removing irrelevant and redundant features. This paper highlights some well known approaches to filter feature selection, information theory and rough set theory, and compares a recent fitness function with some traditional methods. The contributions of this paper are two-fold. First, new results confirm previous research and show that the recent fitness function can also perform favorably when compared to rough set theory. Secondly, the measure of redundancy that is used in traditional information theory is shown to damage the performance when a similar approach is applied to the recent fitness function.

1 Introduction

In many situations a large number of features are introduced in order to describe the target objects in the universe. This large number of features allows for different concepts and patterns to be identified and can help with numerous problems, such as classification. Often, however, too many features can contribute to "the curse of dimensionality", a major obstacle in classification. In addition to this, the presence of noisy or highly correlated features can decrease performance. Feature selection is an important and well known technique for solving the above problems [1]. Feature selection can be described as follows: given a set of n features, G, find a set of m features, F, such that $m < n$ and $F \subset G$. F should be representative of G, and should have eliminated any irrelevant or redundant features hence increasing efficiency and enhancing classification accuracy.

Any feature selection algorithm has two key aspects: the search strategy and the evaluation criterion (fitness function). The evaluation criterion measures how good the selected features are, this information can then be used for a number of things. For example, it can be used to guide the search, by choosing the next feature or highlighting a good path to explore, or to decide when to stop. Evaluation criterion can be categorized into wrapper approaches and filter approaches. Wrapper approaches embed learning algorithms, such as a Naive Bayes classifier, into the evaluation criterion, while filter approaches use mathematical models as estimates of goodness. Wrapper approaches usually achieve better results than

T. Yoshida et al. (Eds.): AMT 2013, LNCS 8210, pp. 166–176, 2013.

filter approaches, but the cost of the learning algorithm makes them computationally expensive and can lead to a loss of generality as the algorithm will pick features that perform well for that classifier and that particular training set [2].

Filter approaches rely on the mathematical model used to estimate the goodness of the selected features rather than the actual measure used in wrapper approaches. The performance of the algorithm is then dependent on how good an estimate the mathematical model provides. Many different models have been proven effective, including information measures [3] and rough set theory [4]. A filter approach that achieves a good estimate has the ability to perform favorably when compared to wrapper approaches. If it is a good estimate it will likely select good features, and possibly match the wrapper approach, but do so using fewer resources. This means that a filter approach could achieve a more complete search than a wrapper approach in the same amount of time, the potential for finding better subsets is then increased.

The second aspect of any feature selection algorithm is the search strategy. To perform a complete search of every possible feature subset would be unfeasible. A dataset with instances described by only 30 features has 2^{30} possible feature subsets, each of which would need to be evaluated. To overcome this issue more complex search strategies have been devised. Greedy algorithms exist but these have the problem of getting stuck in local minima and maxima [1]. Evolutionary techniques can also be used to perform the search. Increasing the complexity of the search strategy aims to be able to perform a more complete search.

Previous work presented a new information theory function that performed well compared to traditional information theory when using particle swarm optimization as the search technique [5]. This paper focuses on comparing and evaluating that new filter based evaluation criterion on a larger scale. Comparison with traditional information theory [3,6] is performed again and rough set based [7,8] techniques are introduced and compared with both the information theory techniques. Comparison between information theory and rough set theory lacks landmark research, the recent technique has also only been compared with information theory and not rough set theory.

The contributions of this paper are two fold. Firstly, comparison between the information theory and rough set theory approaches show that the recent fitness function produces favorable results when compared to both alternatives. Secondly, the results show that previous results could have been improved further by not considering the measure of redundancy seen in both the information theory approaches.

2 Background

2.1 Information Theory

Information theory as developed by Shannon [6] presents a way to quantify the level of uncertainty in random variables. From this we can derive the amount of information gained and shared between random variables.

Entropy, in information theory, can be described as the level of uncertainty in a random variable. Let X be a random variable with discrete values, the entropy of X, $H(X)$, is:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

where $p(x) = P(X = x)$, the probability density function of X.

The joint entropy of two random variables, X and Y, can be described as:

$$H(X, Y) = H(Y, X) = -\sum_{x \in X,\ y \in Y} p(x, y) \log_2 p(x, y) \tag{2}$$

The joint entropy quantifies the degree of uncertainty in two random variables.

Gaining knowledge of a certain variable can often reveal information about others, this is measured by the conditional entropy. Assume that the variable Y is known then the conditional entropy of X given Y, $H(X|Y)$, is:

$$H(X|Y) = -\sum_{x \in X,\ y \in Y} p(x, y) \log_2 \frac{p(x)}{p(x, y)} \tag{3}$$

The conditional entropy can also be calculated using the entropy and joint entropy:

$$H(X|Y) = H(X, Y) - H(Y)$$

Finally, the information shared between two variables is defined as mutual information. The amount of information is shared between variables X and Y, the mutual information, $I(X; Y)$, is defined as:

$$I(X; Y) = -\sum_{x \in X,\ y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{4}$$

As with the conditional entropy, shown above, mutual information can be defined using the other measures of entropy:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

Information theory can be used to build a filter based evaluation criterion that can be used to calculate the information shared between features, for redundancy, and between the class value and the features, for relevance.

2.2 Rough Set Theory

Rough set theory, as developed by Pawlak [7], provides a formal approximation of a conventional set. The rough set is described by the lower and upper approximations of the conventional set. Let **U** be the universe, the set of instances, and

let $\mathbf{A}$ be the set of attributes that describe the instances. Also, let $a(x)$ specify the value of attribute $a \in \mathbf{A}$ in instance $x \in \mathbf{U}$.

For any $P \subseteq \mathbf{A}$ we can define the indiscernible equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbf{U}^2 \mid \forall a \in P.\ a(x) = a(y)\} \tag{5}$$

If $(x, y) \in IND(P)$ we say that x and y are indiscernible according to P. We can use the above relation to define equivalence classes, these are denoted $[x]_P$. This means that $y \in [x]_P \Leftrightarrow (x, y) \in IND(P)$.

Let $X \subseteq \mathbf{U}$ be the set we want to represent with P. We can define the upper and lower bounds of X according to P:

$$\underline{P}X = \{x \mid [x]_P \subseteq X\} \tag{6}$$

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\} \tag{7}$$

The rough set is then the tuple: $\langle \underline{P}X, \overline{P}X \rangle$. We can quantify the *accuracy* of the rough set using:

$$\alpha_P(X) = \frac{|\underline{P}X|}{|\overline{P}X|}$$

Which effectively measures how well the attributes in P separate the target set, X, from the rest of $\mathbf{U}$. If P is poorly chosen then few instances will be in the lower bound while many will be in the upper bound.

As with information theory, rough set theory can be used as filter approach to evaluation. Well selected features will separate the classes that instances can be assigned to.

3 Fitness Functions

This section presents evaluation criterion based on information theory and rough set theory and explains how they can be applied to the data. Fitness functions can be used to compare potential features, all three of the following functions work by adding the potential feature to the test and evaluating the new set together. As with the rough set theory the feature selection framework uses a set of instances that make up the universe $\mathbf{U}$, a set of features, F, that describe the instances, and each instance has a class value, c.

3.1 Paired Mutual Information

Peng et al. present an filter based fitness function based on information theory [3]. Peng et al. use mutual information to estimate the relevance and redundancy of the features as they are selected, the fitness function attempts to maximize the relevance and minimize the redundancy. Each feature, and the class label, needs to be treated as a random variable and the probability density functions can be calculated using a training set. For example, consider a particular feature,

$f \in F$, that can take three values: $\{0, 1, 2\}$. The number of times f takes the value 0 can be used to calculate $P(X = 0)$ and hence calculate the entropy and mutual information.

The fitness of a set of features, $G \subseteq F$, is calculated as:

$$Fitness(G) = D(G) - R(G) \tag{8}$$

$$D(G) = \frac{1}{|G|} \sum_{f \in G} I(f; c) \tag{9}$$

$$R(G) = \frac{1}{|G|^2} \sum_{f \in G, g \in G} I(f; g) \tag{10}$$

Equation (9) quantifies the average mutual information shared by each selected feature and the class label. Equation (10) quantifies the average mutual information shared by each of the selected features with every other feature, hence providing a measure of redundancy. This paper provides a comparison between the fitness function with (MI-R-P) and without (MI-NR-P) the measure of redundancy and shows that it is important to improving the accuracy.

3.2 Group Mutual Information

Where the function presented above considers pairs of features, our previous research presented a second information theory criterion that attempts to evaluate the features as a group [5]. The fitness function requires the joint entropy to be calculated over larger sets of random variables:

$$H(X_1, ..., X_n) = - \sum_{x_1 \in X_1} ... \sum_{x_n \in X_n} p(x_1, ..., x_n) \log_2 p(x_1, ..., x_n) \tag{11}$$

As above the fitness function attempts to maximize the relevance and minimize the redundancy.

$$Fitness(G) = D(G) - R(G) \tag{12}$$

$$D(G) = IG(c \mid G) \tag{13}$$

$$R(G) = \frac{1}{|G|} \sum_{f \in G} IG(f \mid \{G/f\}) \tag{14}$$

The criterion here attempts to quantify the amount of information gained about one random variable, or feature, given knowledge of a set of others. It can be calculated by:

$$\begin{aligned} IG(c \mid G) &= H(c) - H(c|X) \\ &= H(c) - (H(c \cup X) - H(X)) \\ &= H(c) + H(X) - H(c \cup X) \end{aligned}$$

Cervante et al. use this function as a way to evaluate sets of features generated by a search using particle swarm optimization. As with the paired fitness function,

the group evaluation is done using both with the measure of redundancy (MI-R-G) and without (MI-NR-G), the experimental procedure also involved a simpler process than particle swarm optimization to focus comparison on the fitness functions. As the group fitness function considers the selected features as a group rather than in a paired average it should provide a better estimate.

3.3 Probabilistic Rough Set Approximations

Rough set theory provides a natural way to evaluate feature sets, given the definitions. We can partition the universe using the class labels, each partition becomes a target set. The lower bound of each target set then measures the number of instances that have been completely separated from instances of other classes. Assume the universe, $\mathbf{U}$, has been partitioned into target sets: $\{U_1, ..., U_n\}$. An evaluation criterion for a subset of features, $G \in F$, is then:

$$Fitness(G, \mathbf{U}) = \frac{\sum_{U_i \in \mathbf{U}} |\underline{G}U_i|}{|\mathbf{U}|} \tag{15}$$

The evaluation criterion measures the number of instances that have been separated from instances of other classes by the features, a score of 1.0 means that G completely divides the classes.

It is possible that a minority of instances could share identical attributes but have different class labels. In practice this could happen due to minorities that are exceptions to common patterns or even human error in inputting the data. This being the case, having even one instance labeled differently means that it becomes impossible to find a satisfying assignment because of one mistake. To overcome this problem we can introduce probabilistic rough set approximations [8]. Rather than having a strict lower bound, we can relax it with varying degree using a value α. For a given target set X and a set of features G, we define the function μ to be:

$$\mu_G(x) = \frac{|[x]_G \cap X|}{|[x]_G|} \tag{16}$$

So, μ quantifies the proportion of $[x]_G$ that is also in the target set. Using this we can define the lower approximation of X according to G:

$$\underline{apr}_G X = \{x \mid \mu_G(x) \geq \alpha\} \tag{17}$$

Note that when α is set to 1.0 this calculation becomes the same as the lower bound calculation above, since every instance in $[x]_G$ must also be in X making it only true when it is a subset.

Using this approximation we update the fitness function to be:

$$Fitness(G, \mathbf{U}) = \frac{\sum_{U_i \in \mathbf{U}} |\underline{apr}_G U_i|}{|\mathbf{U}|} \tag{18}$$

We can set α to be 1.0 to mimic the strict lower bound definition and use a variety of values for α to hopefully improve performance. In experimental conditions, three conditions for α were considered, $\{1.0, 0.75, 0.5\}$, with $\alpha = 1.0$ as the baseline.

Table 1. Summary of results at peak accuracy

Criterion	Number of Features	Training Accuracy	Testing Accuracy
MI-NR-P	7	0.897	0.919
MI-R-P	4	0.939	0.945
MI-NR-G	4	0.939	0.945
MI-R-G	8	0.920	0.925
RS-1.0	12	0.932	0.922
RS-0.75	4	0.939	0.945
RS-0.5	4	0.939	0.945

4 Experimental Results

In order to evaluate the different criteria we tested them using a dataset from the UCI repository: Chess (King-Rook vs. King-Pawn). The dataset is split 52:48 into two classes. There are a total of 3196 instances and 36 attributes. We ran a simple forward selection algorithm and tested the accuracy on the training and testing set, derived from the dataset, as each feature was added to show how well each criterion estimated the goodness of potential features. A naive bayes classifier was used and the accuracy on the training set when using all features was 0.869 and on the testing set was 0.870.

8 different criteria were tested, the crisp rough set with $\alpha = 1.0$ (RS-1.0) and the two mutual information functions with no measure of redundancy (MI-NR-P and MI-NR-G for the pair and group criterion respectively) were tested, these are considered the baseline. Next we demonstrated the effect of adding the measure of redundancy in the mutual information criterion (MI-R-P and MI-R-G) and the effect of relaxing the value for alpha (RS-0.75 and RS-0.5). Each criterion peaked at a considerably lower number of features before the performance began degrading as more features were added. The results for the best performing features and how many were included can be seen in Table 1.

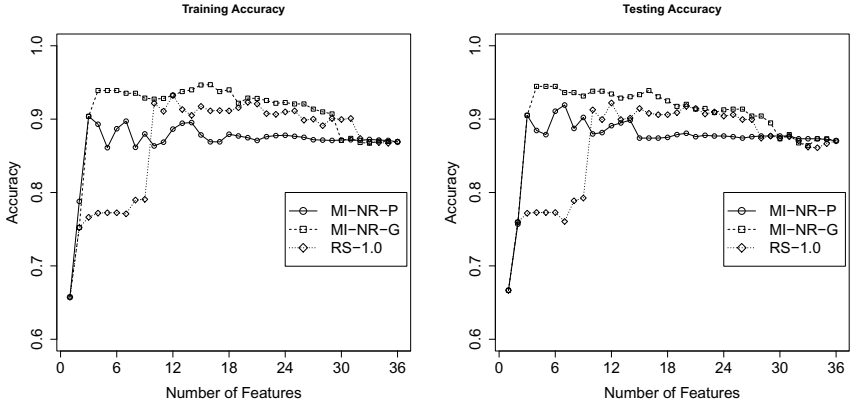

Fig. 1. Performance of the three baseline classifiers: MI-NR-P, MI-NR-G, RS-1.0

The following figures show the change in performance as more features are added. Figure 1. shows a comparison of the three baseline functions. Both the information theory criterion gain a large increase in accuracy very quickly, which the rough set function does not. However, after 12 features are selected the rough set function overtakes and gets better accuracy than the paired mutual information one achieves. Our group fitness function stays ahead of both.

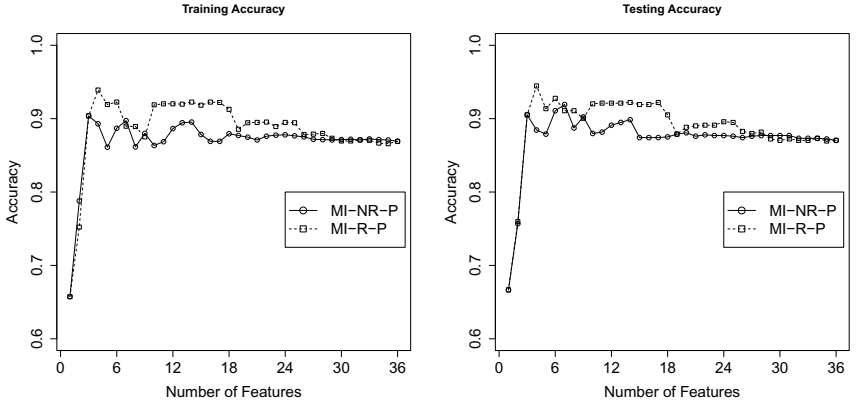

Fig. 2. Performance comparison between using and not using a measure of redundancy in the MI-P criterion

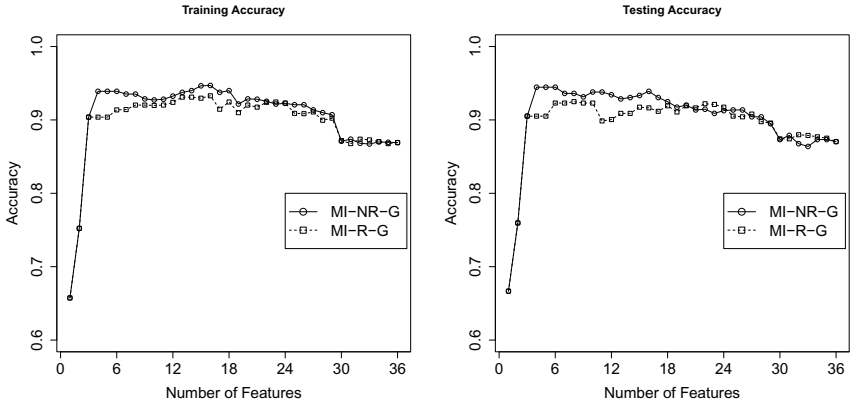

Fig. 3. Performance comparison between using and not using a measure of redundancy in the MI-G criterion

The remaining figures show attempts at further improving the accuracy. Figure 2. shows the introduction of a measure of redundancy to the pair criterion, making it the same as the fitness function as presented by Peng et al. Figure 3. shows the measure of redundancy introduced into our group fitness function. Interestingly, the added redundancy measure decreases the performance of

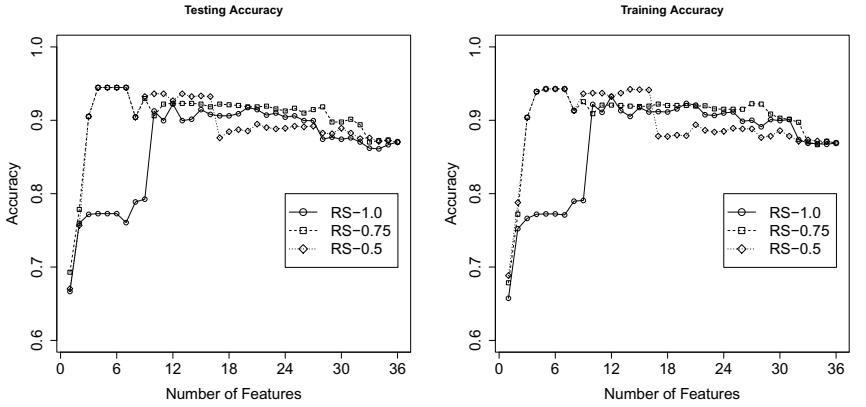

Fig. 4. Performance comparison for varying values of alpha, $\alpha = \{1.0, 0.75, 0.5\}$

the grouped fitness function while it improves the paired fitness. Finally, Figure 4. shows the changes seen with differing values for alpha in the probabilistic rough set approach. Both additional values for alpha show an increase in performance initially but when $\alpha = 0.5$ we see a large drop in performance after 17 features have been selected. With $\alpha = 0.75$ the performance remains strong for much longer.

5 Discussion

We can see that in this case it is possible to find the highest performing subset using all three criterion. MI-R-P, MI-NR-G, and the two probabilistic rough set approaches, RS-0.75 and RS-0.5, reach an accuracy of 0.939 after 4 features have been selected. In addition to this, the group function, MI-NR-G, and the two rough set approaches maintain this accuracy even as more features are being selected and only begin to drop when 7 features or more are selected. The paired function, MI-R-P, drops straight away but then levels, and after 7 features matches the rough set approaches. Our group fitness function, MI-NR-G, maintains a higher accuracy for much longer suggesting that it is finding better features than the alternate approaches, even though the accuracy isn't increased the danger of overfitting is lessened.

The reason of the accuracy drop when adding the redundancy measure to the group criterion could be cause it becomes too concerned with avoiding redundancy, too much weight is given to that consideration. The function already considers the fitness of the group as a whole so adding redundancy could be unnecessary. This is in contrast to the paired function, where the relevance measure does not consider the group and so it is important to add the measure of redundancy. In our earlier paper [5] that first introduced the group evaluation measure only the function that considers redundancy as well as relevance is used, accuracy could have been improved further by not considering the measure of redundancy.

Finally, it is clear that using probabilistic rough sets, RS-0.75 and RS-0.5, instead of the strict lower bound, RS-1.0, can select better features. However, when alpha was set very low, $\alpha = 0.5$, overfitting (searching for too long) becomes a significant risk, after 50% of the features had been selected the performance dropped to less than the performance of the strict definition. Using the strict lower bound likely means that it is difficult for the classifier to generalize while a low value for alpha leads to over generalization. The performance of the rough set theory function is then dependent on the choice of alpha. The optimum value of alpha could differ across multiple situations, the information theory functions do not have this problem as they have no global variables that need to be defined.

6 Conclusion

In conclusion, the results show that our previous group mutual information based approach can achieve a high accuracy and choose features to maintain it for longer. In addition to this, the redundancy measure used in the paired fitness function actually hurt the performance of the group fitness function. Finally, a comparison between these two information theory based approach and that of a rough set theory based approach was undertaken.

This paper presented the mathematical reasoning behind information theory and rough set theory, before showing two traditional evaluation criterion derived from the reasoning. Our group approach was then compared with the traditional approaches, and it is highlighted that previous work could have been improved by not considering the redundancy measure in the recent approach. This is in contrast to the traditional information theory approach in which considerable improvement is gained by including the measure of redundancy.

Results also show that while adding more features decreases the accuracy in the training and testing sets, the new approach can maintain a higher degree of accuracy for longer. This suggests that it is still finding good potential features where the others are not. Previous work compared the group fitness function favorably with the pair fitness function when particle swarm optimization was used as the search technique, future work could consider including a comparison with the rough set theory functions using a more complex search technique.

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
2. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97(1-2), 273–324 (1997)
3. Peng, H.P.H., Long, F.L.F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
4. Zhong, N., Dong, J., Ohsuga, S.: Using rough sets with heuristics for feature selection. J. Intell. Inf. Syst. 16(3), 199–214 (2001)

5. Cervante, L., Bing, X., Zhang, M.: Binary particle swarm optimisation for feature selection: A filter based approach. In: Proceedings of 2012 IEEE Congress on Evolutionary Computation, pp. 881–888. IEEE Press (2012)
6. Shannon, C.E., Weaver, W.: A Mathematical Theory of Communication. University of Illinois Press, Champaign (1963)
7. Pawlak, Z.: Rough sets. International Journal of Parallel Programming 11, 341–356 (1982), 10.1007/BF01001956
8. Yao, Y.: Probabilistic rough set approximations. Int. J. Approx. Reasoning 49(2), 255–271 (2008)