Using the Symmetrical Tau (τ) criterion for feature selection in decision tree and neural network learning

Fedja Hadzic and Tharam S. Dillon

Faculty of Information Technology, University of Technology Sydney, Australia

email: (fhadzic, tharam)@it.uts.edu.au

Abstract - *The data collected for various domain purposes usually contains some features irrelevant to the concept being learned. The presence of these features interferes with the learning mechanism and as a result the predicted models tend to be more complex and less accurate. It is important to employ an effective feature selection strategy so that only the necessary and significant features will be used to learn the concept at hand. The Symmetrical Tau (τ) [13] is a statistical-heuristic measure for the capability of an attribute in predicting the class of another attribute, and it has successfully been used as a feature selection criterion during decision tree construction. In this paper we aim to demonstrate some other ways of effectively using the τ criterion to filter out the irrelevant features prior to learning (pre-pruning) and after the learning process (post-pruning). For the pre-pruning approach we perform two experiments, one where the irrelevant features are filtered out according to their τ value, and one where we calculate the τ criterion for Boolean combinations of features and use the highest τ-valued combination. In the post-pruning approach we use the τ criterion to prune a trained neural network and thereby obtain a more accurate and simple rule set. The experiments are performed on data characterized by continuous and categorical attributes and the effectiveness of the proposed techniques is demonstrated by comparing the derived knowledge models in terms of complexity and accuracy.*

Keywords: feature selection, rule simplification, network pruning

1. Introduction

The data collected for various industrial, commercial or scientific purposes usually contains some features irrelevant to the concept of interest. When an induction algorithm is used to obtain a knowledge model about the concept the presence of these features interferes with the learning mechanism because of the noise introduced, and as a result the learned models tend to be more complex and less accurate. It is important to employ an effective feature selection strategy so that only the necessary and significant features will be used to learn the concept at hand. By concentrating on only the important aspects of the domain the derived knowledge models will be improved in terms of accuracy and comprehensibility.

Feature selection strategies can be roughly categorized into filter and wrapper based approaches. Filter approach is done independently of the learning algorithm and the irrelevant features are filtered out prior to learning. Common technique is to evaluate the features based upon their capability of predicting the target attribute and then to choose a subset of features with sufficiently high values. One such approach is the 'Relief' algorithm [6] that assumes two-class classification problems, and is inspired by instance-based learning. Relief detects those features statistically relevant to the target concept by assigning a relevance weight to each feature. It conducts in random sampling of the instances from the training set during which the relevance values are updated. The updating of relevance values is based on the difference between the selected instance and the two nearest instances of the same and opposite class. Another filter approach is "FOCUS" [1] which exhaustively examines all subsets of features and selects the minimal subset that is sufficient to determine the target concept for all instances in the learning set.

In a wrapper based approach [7] the feature selection algorithm exists as a "wrapper" around the induction algorithm. The algorithm conducts a search for a good subset of attributes using the induction algorithm itself as part of the evaluation function. The benefits of using the induction algorithm itself for evaluating feature subsets is that there will be no inductive bias introduced by a separate measure. On the other hand the major disadvantage is the computational cost associated with each call to the induction algorithm for evaluating the feature set [3]. More recently a hybrid algorithm named FortalFS [12] has been proposed, which uses results of another feature selection approach as the starting point in the search through feature subsets that are evaluated by the induction algorithm. In [4] a genetic algorithm SET-Gen was described for solving the problem of feature subset selection. A population of best feature subsets is kept and genetic operators are applied in order to create new feature subsets, which are evaluated according to the predefined fitness function. The fitness function favors those subsets that produce smaller decision trees, use less input features and retain predictive accuracy.

When dealing with the feature selection for neural networks (NN) the problem is commonly referred to as network pruning and it is split into pre-pruning and post-pruning approaches. Pre-pruning is essentially the same as the filter approach and post-pruning approach trains a network to completion and then inspects the links between particular network units in order to determine the relevance between the two [9]. This approach is useful for rule simplification and for removal of attributes whose usefulness has been lost through the learning. Most of the methods for symbolic rule extraction from NN use some kind of pruning technique to increase the performance and produce simpler rules. The contribution of each unit in the network is determined and a unit is removed if the performance of network does not decrease after the removal. This is often referred to as sensitivity analysis in NN and is one of the common techniques for network pruning [9,11].

Symmetrical Tau (τ) [13] is a statistical measure for the capability of attribute in predicting the class of another attribute. Previously it has successfully been used as a feature selection criterion during decision tree construction. The τ criterion was reported to have many promising properties and in this paper we particularly want to demonstrate its capability to handle continuous attributes, Boolean combinations of attributes and the capability of measuring an attribute's sequential variation in predictive capability. We provide an experimental study of some different ways the τ criterion can be used to filter out the irrelevant features prior to learning (pre-pruning) and after the learning process (post-pruning).

The rest of the paper is organized as follows. In section 2 we describe the τ criterion and its promising properties as a feature selection criterion. The three experimental procedures are described in section 3 and experimental results are provided and discussed for each procedure. The paper is concluded in section 4.

2. Symmetrical Tau (τ)

There are many different feature selection heuristics used for various inductive learning methods and some of the common disadvantages are: bias towards multi-valued attributes, errors in the presence of noise, not handling of Boolean combinations and sequential variation in predictive capability [10]**.** Zhou and Dillon [13] have introduced a statistical-heuristic feature selection criterion, Symmetrical Tau (τ), derived from the Goodman's and Kruskal's asymmetrical Tau measure of association for cross-classification tasks in the statistical area. The τ criterion has successfully been used to remove the irrelevant features during decision tree induction and has the following powerful properties:

- Built-in statistical strength to cope with noise;
- Dynamic error estimation conveys potential uncertainties in classification;
- Fair handling of multi-valued attributes;
- Not proportional to the sample size;
- Its proportional-reduction-in-error nature allows for an overall measure of a particular attribute's sequential variation in predictive ability. This determines which attributes have become less useful for prediction and should be deleted (pruned).

- Middle cut tendency separating a node into two balanced subsets;
- Handles Boolean combinations of logical features.

The τ criterion is calculated using a contingency table, which is a table that provides a two-way classification, and may be used if each feature of the sample can be classified according to two criteria. As a result c1*c2 contingency table can be formed, where c1 and c2 are the values of two criteria. If there are I rows and J columns in the table, the probability that an individual belongs to row category i and column category j is represented as P(ij), and P(i+) and P(+j) are the marginal probabilities in row category i and column category j respectively. The Symmetrical Tau measure is defined as [13]:

$$\tau = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\frac{P(ij)^2}{P(+j)} + \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{P(ij)^2}{P(i+)} - \sum_{i=1}^{I} P(i+)^2 - \sum_{j=1}^{J} P(+j)^2}{2 - \sum_{i=1}^{I} P(i+)^2 - \sum_{j=1}^{J} P(+j)^2}$$

For the purpose of feature selection problem one criteria (A) in the contingency table could be viewed as a feature and the other (B) as the target class that needs to be predicted. The τ criterion has the following properties [13]:

- In most cases it is well defined;
- If P(ij) = 1 for some i and j, and all other cells have zero probability then the categories of A and B are known with certainty;
- If τ = 0, then the feature in question has no predictive ability for the category of another feature. For this to occur there must be no P(ij) = 1, and all non-zero probabilities are in a single row or column of the contingency table;
- If τ = 1, then the feature in question has the prefect predictive ability for the category of another feature. For this to occur there cannot be any P(ij)=1 and either: for each j there exists an i such that P(i,j) = P(+j), or for each i there exists a j such that P(ij) = P(i+);
- For all other cases τ falls between 0 and 1;
- τ is invariant under permutations of rows and columns.

3. Experimental Procedure and Results

In this section we provide our various experimentations done to demonstrate some different ways Symmetrical Tau can be used as a feature selection criterion. In each of the sections the approach taken is described and the experimental results are provided. For experimentation the decision tree algorithm used is C4.5, and for neural network testing we used the standard back-propagation algorithm with 2 hidden layers, learning rate - 0.3, learning momentum - 0.2 and the training time of 500 epochs. The training set was made up of 60% of the available data, and the rest was used as the testing set for the accuracy of the predicted model. Any attributes that serve as a unique identifier of an instance have been removed from the training set. We have used data of varying complexity and attribute characteristics, publicly available from the ‘uci’ machine learning depository [2].

4.1 Filter approach using the τ criterion for feature selection

Here the τ criterion is used to rank the existing attributes according to their capability in predicting the class of the target attribute. Only the attributes with sufficiently high τ values will form a part of the feature subset to be used by the learning algorithm. The relevance cut-off point chosen is where the difference amongst the τ values of the ranked attributes is sufficiently high. The aim is to remove the

irrelevant attributes without decreasing the accuracy of the derived knowledge model. Table 1 summarizes the results obtained when the described method was applied to domains characterized by categorical (top six) and continuous (bottom three) attributes.

	C4.5								Back-propagation NN	
	1 – unpruned		2- pre-pruned		3 -post-pruned		4 - pre- and post-pruned		Full feature set	τ-reduced feature set
Domains	Size	accurracy	Size	accurracy	size	accurracy	size	accurracy	accurracy (%)	Accuracy (%)
Postop	33	47.2	25	75	7	72.2	7	72.2	61.1	63.88
Breast-cancer	45	93.57	45	94.28	31	93.928	23	95	95.7	96.78
Voting	37	95.977	29	97.126	11	97.7	11	97.7	92.5	94.25
Lenses	7	70	5	70	7	70	5	70	70	70
Mushroom	29	100	27	100	29	100	27	100	100	100
Zoo	17	92.68	15	92.68	17	92.68	15	92.68	82.9	87.8
Wine	13	91.6	13	91.6	9	91.6	9	91.6	95.8	98.61
E-coli	51	76.29	51	76.29	43	78.51	43	78.51	71.8	75.5
Glass	51	67.4419	49	69.7674	51	69.7674	41	70.9302	59.3	61.62

Table 1 – Results of applying the τ criterion for filtering out the irrelevant attributes

The comparison of the decision tree results are displayed on the left where unpruned corresponds to the results obtained when the standard C4.5 algorithm is used, pre-pruned when the attribute set has been reduced according to the τ criterion and post-pruned when the post-pruning technique from C4.5 is used. The size and predictive accuracy (%) of the resulting decision trees were compared, and improvements occurred when the attribute set was filtered according to the τ criterion. For the unpruned version comparison (1 versus 2), the resulting decision tree was simpler in all cases except for breast-cancer domain where it remained the same. The decrease of tree complexity was not at the cost of a reduction in accuracy. In fact accuracy was either improved or kept the same. A significant improvement in accuracy was observed in 'post-operative patients' domain, with an increase from 47.2 % to 75 %.

For the pruned version (3 versus 4), the resulting decision tree was simpler in all but three cases where it remained the same. The accuracy increased for breast-cancer and glass domain and remained the same for the rest. When comparing the results obtained either by applying the τ criterion for pre-pruning or the post-pruning approach from C4.5 (2 versus 3) there were four cases in which pre-pruning achieved a simpler tree and other five for post-pruning. Accuracy increased through pre-pruning for two cases and by post-pruning for two (rest is same). Besides this similarity one advantage of pre-pruning is that irrelevant features are detected early in the learning process which avoids poor choices being made for test-nodes in the tree. As both approaches combined achieved the best results in all but one domain (postoperative patients), good practice would be to use a filtering method first followed by a post-pruning method which will detect and delete those attributes that have become useless and possibly interfering for the prediction task.

The value of the τ criterion at which the attributes were removed from the training set varied in most of cases, and it was not easy to determine a general cut-off point. The factors that affected the cut-off point for a certain domain appeared to be the attribute-set size and interrelationship between the attributes within the set. For example in the mushroom domain high difference amongst the τ values occurred high in the ranking, and bottom 15 attributes could be removed without affecting the accuracy. On the other hand in the lenses and post-operative patients domain this difference occurred low in the ranking and only the bottom two attributes could be removed. Furthermore some attributes having low τ values proved to be important independently from the attribute-set size due to their interrelationship with other attributes from the training set. In the post-operative patients domain all the attributes had very low τ values, and only when combined they provided high predictive power. All these observations indicate the importance of measuring the predictive capability for Boolean combinations of attributes, which is discussed next.

4.2 Measuring predictive capability for Boolean combinations of features

In order to calculate τ for combinations of features the input data was transformed for each n-combination by combining the attributes and the values that occur in each instance. The τ criterion was then calculated for all n-combinations and the one with the highest τ value was the attribute subset used by the induction algorithm. In some domains (mushroom, zoo, voting) attribute set was too large to calculate all possible combinations, in which case the attributes with low τ values were removed. It should be noted that this could potentially miss the best combination as sometimes an attribute that may have a low τ value could become useful when combined with another attribute. If forming all possible combinations is still infeasible, one could continue to remove combinations at each step by determining a cut-off point for each set of n-combinations formed. Only the promising combinations would be used for forming higher n-combinations, and the combinatory explosion problem could be alleviated to some extent.
The results of the experiment are provided in table 2. Note that the results of applying the C4.5 algorithm with post-pruning to the highest τ-valued combination are excluded from this table as they remain the same to when no post-pruning is done. Besides the domains obtained from the 'uci' depository, we have used a simple noise-free syntactic file for recognizing LED digits in order to check that the τ value will be equal to one for the necessary and sufficient attribute combination. Indeed the combination of five attributes was detected with value of 1 which is the minimal required attribute set to obtain perfect predictive accuracy for this domain. This can be seen from table 2 as the C4.5 algorithm achieves perfect accuracy with the used combination. However, the neural network was incapable of achieving perfect accuracy with this combination. As we are using a type of graph structure to represent the necessary information for τ calculation when a certain attribute combination has a value of 1 the knowledge about the target attribute is contained in the structure itself. Each child node of the attribute combination corresponds to the set of permissible values and the target vector associated with this node shows which class is implied by that particular combination of values. As these rules would involve all attributes from the combination a concept hierarchy formation technique [10] could be applied to obtain a comprehensible conceptual hierarchy for the domain. In this case there would be no need for the use of an inductive learning algorithm to obtain the knowledge model. LED domain is excluded from any further discussion.

	C4.5						Back-propagation NN	
	Unpruned		Post-pruned		Highest τ-combination		Full feature set	Highest τ-set
Domains	Size	Accurracy (%)	Size	Accurracy (%)	Size	Accurracy(%)	Accurracy(%)	Accurracy(%)
Breast-cancer	45	93.57	31	93.928	3	91.4286	95.7	89.64
Voting	37	95.977	11	97.7	3	96.55	92.5	96.55
Lenses	7	70	7	70	5	70	70	70
Mushroom	29	100	29	100	10	98.4923	100	98.49
Zoo	17	92.68	17	92.68	11	98.6829	82.9	82.9
LED	19	100	19	100	19	100	100	89.36

Table 2 – Comparison of results obtained when the highest τ-valued attribute combination is used

As it can be seen on the left of table 2, the size of the decision tree has been substantially reduced in all domains. However, in most cases this achievement was at cost of a small reduction in accuracy. An interesting observation is that out of all attribute combinations a single attribute had the highest value in breast-cancer (bare nuclei) and voting (physician-fee-freeze) domains. These attributes do indeed contain the most information for distinguishing the classes of the target attribute. The difference in the accuracy by using the full attribute set is only very small in comparison to the large reduction in tree size. In fact for the voting domain better accuracy was achieved by using single attribute rather than the full attribute set if no post-pruning was applied in the C4.5 algorithm, and the NN achieved better accuracy using only one attribute. The question still remains as to why the attribute combination that would increase the accuracy by this small amount did not have higher τ value than the single attribute. This is due to the fact that when using the τ measure for pre-pruning the value measures the 'total' predictive capability of

attributes and not sequential predictive capability, which is essentially what post-pruning is used for. Total predictive capability refers here to the measure calculated over all classes and instances. To measure the sequential variability in predictive capability of attributes the τ criterion would need to be calculated over a subset of classes and hence instances, which is done in the next section. In an attribute combination the extra information that the combined attributes provide may interfere with the main predicting attribute and hence the τ value is small. It would interfere until some class values are distinguished at which stage this interfering attribute may become useful. In other words some attribute with low τ value may have the necessary constraints to distinguish the remaining instances for which the high τ valued attributes did not have sufficient constraints. This claim is supported by the fact that only after post-pruning was applied for the voting domain the accuracy was higher than by using the single attribute. Furthermore, the attribute set in the voting domain had to be reduced for combining which may have missed some potentially useful combinations. Besides the fact that measuring predictive capability for Boolean combinations of features cannot capture the attributes that become useful once many classes have been distinguished, it can still be very useful to detect the most crucial attributes or combinations for a particular domain. Furthermore, in most of cases the difference in accuracy was not sufficiently high to discard the usefulness of the approach.

4.3 Using the τ criterion for rule simplification

The aim of this section is to demonstrate how the τ criterion can be used as a post-pruning approach for neural networks. Due to its capability of measuring attribute's sequential variation in predictive capability it is used to determine the relevance of an attribute to the rule extracted from a NN. In general the method could be applicable to any rule sets where there are clearly defined attributes values that imply a subset of target classes.

Self-Organizing Map (SOM) [8] is an unsupervised neural network that effectively creates spatially organized "internal representations" of the features and abstractions detected in the input space. It is based on the competition among the cells in the map for the best match against a presented input pattern. Existing similarities in the input space are revealed through the ordered or topology preserving mapping of high dimensional input patterns into a lower-dimensional set of output clusters. When used for classification purposes, SOM is commonly integrated with a type of supervised learning in order to assign appropriate class labels to the clusters. After the supervised learning is complete each cluster will have a rule associated with it, which determines which data objects are covered by that cluster.

For this experiment we have used a slight modification of the original SOM algorithm adjusted so that when used in domains characterized by continuous attributes, rules can be extracted directly from the networks links [5]. Once the rules have been assigned to each cluster the supervised learning starts where a cluster with smallest Euclidean distance to the input instance is activated. Each cluster has a target vector associated with it which is updated every time the cluster is activated. During this process the occurring input and target values have been stored for attributes which define the constraints of the activated cluster. The input values that are close to each other are merged together so that the value object represents a range of values instead. The information collected corresponds to the information contained in a contingency table between an input attribute and the target attribute for the instances captured by the cluster.

The τ criterion has been used for the purpose of removing the links emanating from nodes that are irrelevant for a particular cluster. These links correspond to the attributes whose absence has no effect in predicting the output defined by that cluster. The cluster attributes are ranked according to decreasing τ value. The relevance cut-off occurs at the attribute where the τ value is less than half of the previous attribute's τ value. Note that the τ criterion can only be calculated for cluster attributes that contain more than one value and whose cluster was activated for more than one target class. CSOM is then retrained with all the irrelevant links removed and the aim is that the newly formed clusters will be simpler in terms of attribute constraints.

Initial Clusters			Clusters after pruning and retraining		
C#	Constraints	Target Vector	C#	Constraints	Target Vector
1	0.35 ≤ SL ≤ 0.58 0.24 ≤ SW ≤ 0.58 0.54 ≤ PL ≤ 0.66 0.57 ≤ PW ≤ 0.62	Ivs - 6	1	0.3 ≤ SL ≤ 0.36 0.41 ≤ SW ≤ 0.41 0.58 ≤ PL ≤ 0.59 0.57 ≤ PW ≤ 0.58	Ivs - 2
2	0.63 ≤ SL ≤ 0.64 0.37 ≤ SW ≤ 0.41 0.57 ≤ PL ≤ 0.64 0.5 ≤ PW ≤ 0.54	Ivs- 3	2	0.41 ≤ SL ≤ 0.416 0.29 ≤ SW ≤ 0.29 0.67 ≤ PL ≤ 0.69 0.75 ≤ PW ≤ 0.75	Ivg - 2
3	0.194 ≤ SL ≤ 0.5 0.12 ≤ SW ≤ 0.41 0.33 ≤ PL ≤ 0.627 0.37 ≤ PW ≤ 0.58	Ivs -16	3	0.04 ≤ PW ≤ 0.16	Is - 33
4	0.49 ≤ SL ≤ 0.611 0.2 ≤ SW ≤ 0.41 0.64 ≤ PL ≤ 0.78 0.54 ≤ PW ≤ 0.75	Ivg – 8 Ivs - 1	4	0.25 ≤ SW ≤ 0.75 0.792 ≤ PW ≤ 1	Ivg - 19
5	0.44≤ SL ≤ 0.44 0.41 ≤ SW ≤ 0.5 0.64 ≤ PL ≤ 0.69 0.7 ≤ PW ≤ 0.7	Ivg – 2 Ivs – 1	5	0.36 ≤ SL ≤ 0.694 0.29 ≤ SW ≤ 0.41 0.52 ≤ PL ≤ 0.644 0.49 ≤ PW ≤ 0.58	Ivs - 8
6	0.52 ≤ SL ≤ 0.66 0.33 ≤ SW ≤ 0.58 0.69 ≤ PL ≤ 0.847 0.87 ≤ PW ≤ 1	Ivg -13	6	0.24 ≤ SL ≤ 0.42 0.12 ≤ SW ≤ 0.29 0.42 ≤ PL ≤ 0.576 0.36 ≤ PW ≤ 0.54	Ivs - 9
7	0.778 ≤ SL ≤ 1 0.25 ≤ SW ≤ 0.75 0.831 ≤ PL ≤ 1 0.7 ≤ PW ≤ 0.91	Ivg - 8	7	0.19 ≤ SL ≤ 0.22 0.37 ≤ PW ≤ 0.41	Ivs - 2
8	0.36 ≤ SL ≤ 0.47 0.29 ≤ SW ≤ 0.33 0.66 ≤ PL ≤ 0.69 0.62 ≤ PW ≤ 0.79	Ivg – 3 Ivs - 1	8	0.55 ≤ SL ≤ 0.55 0.2 ≤ SW ≤ 0.29 0.66 ≤ PL ≤ 0.67 0.7 ≤ PW ≤ 0.75	Ivg - 3
9	0.02 ≤ SL ≤ 0.417 0.41 ≤ SW ≤ 0.91 0 ≤ PL ≤ 0.15 0 ≤ PW ≤ 0.16	Is - 32	9	0.417 ≤ SW ≤ 0.5 0.625 ≤ PW ≤ 0.7	Ivg - 7

Table 3: Comparison of initially obtained clusters and clusters after pruning and retraining
Notation: SL – sepal_length, SW – sepal_width, PL – petal _length, PW – petal_width, Ivs – iris-versicolor, ivg – iris-virginica, is – iris-setosa.

The CSOM was trained on the ‘iris’ domain available from the ‘uci’ depository and the comparison of initially obtained clusters and clusters after pruning and retraining is shown in table 3. Please note that the order in which the clusters are displayed in the right column does not reflect the clusters that have been simplified. Due to clarity issues and space limitations the clusters that are only triggered once during supervised learning are excluded from results as they are usually merged into other clusters or deleted due to noise suspicion. As can be seen from table 3 the use of τ criterion for network pruning was successful as the newly obtained clusters (rules) were simplified without increasing the misclassification rate. All the clusters are now implying only one target value and the minimal constraints have been found for certain target classes. Generally speaking a simplified network has better performance and simpler rules are expected to have better generalization power.

5. Conclusion

In this study we have demonstrated some different ways of effectively using the Symmetrical Tau (τ) measure to aid in the feature selection problem. The τ criterion proved to be useful as a filter type approach to feature selection, where in one experiment it was used to filter out single irrelevant attributes, and in other to select the most promising subset of features by determining the predictive capability of feature combinations. The study also gives an example of how the τ criterion can be used for post-pruning in neural networks. The approach simplified the extracted rule set and improved the accuracy by removing the attributes that are irrelevant for a particular output. The experimental results show the effectiveness of the proposed method and indicate its potential as a powerful feature selection criterion in other types of inductive learners.

6. References

[1] Almuallim, H., & Dietterich, T.G. 1991. "Learning with many irrelevant features", *Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 547-552, San Jose, CA: AAAI Press.

[2] Blake, C., Keogh, E. & Merz, C.J., 1998. "UCI Repository of Machine Learning Databases", Irvine, CA: University of California, Department of Information and Computer Science., 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

[3] Blum, A. & Langley, P. 1997. "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol 97, Issue 1-2, pp 245-271.

[4] Cherkauer, K.J. & Shavlik, W.J., 1996. "Growing simpler decision trees to facilitate knowledge discovery", *In Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press.

[5] Hadzic, F. & Dillon, T.S., 2005. "CSOM: Self Organizing Map for Continuous Data", *3rd International IEEE Conference on Industrial Informatics (INDIN'05)*, 10-12 August, Perth.

[6] Kira, K. & Rendell, L.A., 1992. "A practical approach to feature selection", Proceedings of the Ninth International Conference on Machine Learning, pp 249-256.

[7] Kohavi, R. & John, G.H., 1997. "Wrappers for feature selection", *Artificial Intelligence*, Vol. 97, issue 1-2, pp 273-324.

[8] Kohonen, T., 1990. "The Self-Organizing Map", *Proceedings of the IEEE*, vol. 78, no 9, pp. 1464-1480.

[9] LeCun, Y., Denker, J. & Solla, S., 1990. "Optimal brain damage", In Touretzky, D.S., ed.: *Advances in Neural Information Processing Systems*, Vol. 2, pp 598-605, San Mateo, CA, Morgan Kauffman.

[10] Sestito, S. & Dillon, S.T., 1994. *Automated Knowledge Acquisition*, Prentice Hall of Australia Pty Ltd, Sydney.

[11] Setiono, R., Leow W.K. & Zurada, J.M., 2002. "Extraction of Rules From Artificial Neural Networks for Noninear Regression." *IEEE Transactions on Neural Networks*, vol. 13, Issue 3, May pp. 564 – 577.

[12] Souza, J., Japkowicz, N., & Matwin, S., 2005. "Feature Selection wuth a General Hybrid Algorithm", in *Proceedings of the Workshop on Feature selection for Data Mining: Interfacing Machine Learning and Statistics* held in conjunction with the 2005 SIAM International Conference on Data Mining, April 23, Newport Beach, CA.

[13] Zhou, X. & Dillon, T.S., 1991. "A statistical-heuristic feature selection criterion for decision tree induction", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no.8, August, pp 834-841.