# Significance Measures and Data Dependency in Classification Methods

A.F. Nejad and T.D. Gedeon
Department of Artificial Intelligence
School of Computer Science and Engineering
The University of New South Wales
Sydney 2052 AUSTRALIA
Email: {akbar,tom}@cse.unsw.edu.au

## *ABSTRACT*

*This paper is a comparative study to reveal the factors that have caused the contradictions among the previous comparative studies. The emphasis has been placed on the significance analysis of input variables, causality analysis and comparative biases. The role of attributes in the classification task with respect to their information bearing nature, and how this is approached by the classification algorithms are studied. We show that the ranked order of significance variables generally differs among the classification methods. We argue that data dependency is a key factor in the analysis of the contradictions among the previous comparative studies. The important role of class prototypes and significance measures, particularly the notion of positional causality introduced in this paper, facilitates future investigation on data transparency, particularly in neural networks. Positional causality means the ability to explain the significance of a variable in each partition of the problem space. Unlike the positional indicators, we argue that global indicators are not reliable. We show that global indicators do not satisfy the reasoning expectations of a human expert, but may cause confusion. We conclude that neural networks are reliable tools with which to analyse the positional causality of problem features.*

## 1. Introduction

Classification is a human need because of the limited capacity of the live brain. The common strategy of a classification method is to minimise misclassification costs of new cases to the characteristics of samples whose category is already known. There are many theories in cognitive psychology concerning the way the human learning system categorises concepts. This has led to many computer implementations of different learning and categorisation theories. There is still a need for a unified, general and complete theory to explain human categorisation.

Classification is a multi-disciplinary science. So there is a need for an overall strategy to choose the best of the existing methods, or at least finding some methods which act to complement each other in the aim of finding the ideal categorisation method.

Since the popular reappearance of neural networks in the 1980's there have been many comparative studies comparing classifiers based on neural networks with statistical techniques and tree-based symbolic methods. Among the variety of existing neural network models, the error back-propagation algorithm has been most frequently used in practical applications.

The previous comparative studies include theoretical and experimental comparisons. The goals have been to evaluate the performance, accuracy, transparency, speed, structure, biological and cognitive plausibility, limitations and abilities of the learning systems. Most of the studies have been done by statisticians and few have considered more than three methods [4].

In general, the results of these studies have not been consistent. This has led to sharp judgments like those of Minsky and Kosko (see [3]). In machine learning the works of Quinlan [15, 16] and Weiss & Kulikowski [20] are relevant. In statistics the works of White [21], Levine et al [7], Geman et al [5], Ripley [17], Sarle [18] and Flexer [4] have pointed out interesting findings. The most interesting and useful statistical work may be found in Cheng et al [2] and its subsequent commentary by top researchers. The valuable work of Michie et al [10] is the result of applying some 20 classification methods on about 20 data sets. These experiments have been performed under the *StatLog* European project.

All the researchers have tried to avoid possible biases in their work, for example by applying the methods on the same training as well as test data sets. The results are not consistent resulting in contradictory statements. According to Ripley, Flexer, Michie et al and many other statisticians, the statistical methods are best due to their rich underlying theories. They believe that *"[t]he modelling-based approach traditional in statistics and pattern recognition can be*

0-7803-2768-3/95/$4.00 © 1995 IEEE

*at least as effective, and often more so*"[17]. Michie et al [10] indicate that a statistical method always ranked first for their data sets.

Quinlan [15] summarises the conclusions of such comparison studies as: "*tree-based symbolic methods and neural networks tend to be more robust across tasks than most other techniques*" and "*tree-based and network classifiers usually have similar accuracy (but with networks slightly ahead). Networks however, require orders of magnitude more computation to develop.*"

The above mentioned inconsistencies have made it difficult to extract a general conclusion from such studies. The difficulty arises from multi-factorial biases governing the studies. The aspects of data transparency and opacity are commonly mentioned as a disadvantage of neural networks in many studies.

The purpose of this paper is to answer the questions:

- Why have the comparison studies not been consistent? What are the hidden biases?
- What aspects of comparison have been misunderstood or are at least ambiguous?
- What are the significance measures and central tendencies in classifiers?
- Has a typical variable the same significance for different classification methods?

## 2. Classification methods review

We have used three of the most widely used classification methods to meet the goals of this study, from the areas of machine learning, connectionism and statistics.

### 2.1. C4.5

Among the best known supervised machine learning algorithms are CART [1], FOIL, ID3, C4.5, and M5 [14]. C4.5 learns by induction from experiences and exemplars [14]. It is an optimised version of induction trees (ID3) capable of applying an appropriate level of pruning and tree selection strategies to allow for better generalisation.

### 2.2. BP

Multilayer perceptrons (MLPs) trained by the error back-propagation algorithm (BP) [9] have been studied extensively, and are one of the most popular and successful neural network models. We have used MLPs trained with one layer of five hidden units with the sigmoid activation functions. The normalised input and output values have been used to train and test the network.

### 2.3. LDA

Statistical methods are supported strongly by mathematical theories. They are well defined and well formulated classification approaches applied by researchers in many scientific domains for many years. Linear discriminant analysis (LDA) is the most commonly used statistical technique, and was introduced by Sir Ronald Fisher.

It is reported that the results of LDA and a few of other statistical methods are more or less the same [10]. LDA uses the posterior probability to make decision in assigning a new sample to a category based on its discriminant score.

## 3. Significance Measures in Classification Methods

An investigation has been required to find an appropriate and reliable method for measuring the significance of variables. In the following subsections we describe how we have measured them for each classification method.

### 3.1. Significance Measures in C4.5

C4.5 uses information theory to assign significance to attributes. According to information theory [19], the information conveyed by a randomly selected pattern $P_j$ from a sample of n patterns is equal to:

$Info\left(P_j\right) = -\log_2\left(Pr\left(C_i\right)\right) = -\log_2\left(\frac{Size\left(C_i\right)}{n}\right)$,

where $Size\left(C_i\right)$ is the number of patterns with class $C_i$, which is the class assigned to $P_j$ and $Pr\left(C_i\right)$is the probability of a random pattern being in class $C_i$. Thus the conveyed information (entropy) of a sample $T$ will be

$Info\left(T\right) = -\sum_{i=1}^{m} Pr\left(C_i\right) * \log_2\left(Pr\left(C_i\right)\right)$ where m is the number of classes.

Quinlan [15] defines a significance measure which he names the *'gain criterion'*. It is defined as

$gain\left(x\right) = Info\left(T\right) - \sum_{k=1}^{k_x}\left(T_k\right)$, where $k_x$ is the number of sub-samples related to test $x$ and $Info\left(T_k\right)$ is the information conveyed by the sub-sample $T_k$. Each test $x$ is a test of partitioning the problem space in accordance with a special variable.

It has been suggested to report the most important attributes according to the level of the tree in which the variable appears [8]. We argue that this would not be an appropriate method due to two reasons:

- I. The output trees of C4.5 are often unbalanced. For example, if there is only one node in the left branch and all other variables are in the right branch of the tree.
- II. The number of cases classified by each node and its children are not equal in different branches. For example a node may classify 60 cases while a node in the same or lower level and its children may only classify 12 cases.

Therefore, we suggest the following method:

- 1. The variable at the top of the tree or every sub-tree is the most important one and conveys the highest entropy among the attributes.
- 2. For each leaf node, we compute the tree benefit of retaining the node, according to $\alpha=\beta-\gamma$, where $\beta$ is the number of patterns which would be misclassified only if the node were removed, and $\gamma$ is the number of patterns which would be classified correctly only if the node were retained.
- 3. The tree benefit of retaining each parent node is computed by adding the $\alpha$'s of its children.
- 4. If the variable is repeated several times in different levels of a sub-tree, only the $\alpha$'s of the node in highest level is assigned to the variable.
- 5. If a variable is repeated in two sub-trees in which neither is a descendent of the other, their $\alpha$'s are added together.
- 6. The steps 3 to 5 are repeated for all of the nodes in higher levels of the tree.
- 7. The variables appearing in the tree are sorted according to their $\alpha$.

Note that we have chosen the best tree selected by the C4.5 algorithm and have reported the result of its classification. If the result of some of the other trees have been of some interest we have emphasised the trial number of the algorithm.

### 3.2. Significance Measures in BP

In the case of multilayer perceptrons, a few ways have been suggested to measure the most significant variables. These methods are sensitivity analysis, causal index, and hidden index.

#### 3.2.1. Sensitivity Analysis

The method of sensitivity analysis measures the change in an output unit due to a small change in an input variable. It is a non mathematical technique to measure significance in neural networks. For each input variable $x_i$, the impact of a small change in its value on an output value $y_j$ is determined by the magnitude of the change in the output value. The bigger the change in output value, the more significance is assigned to the input unit. The main difficulty with this method is initialising the other variables of the input vector in an appropriate way.

These methods have been analysed and applied to a few data sets to investigate the regularities of the internal representations in neural networks [13].

One way of computing Sensitivity Analysis (SA) is to perturb one input slightly with other inputs held constant at 0, 0.5, or 1. This approach is inefficient because SA usually leads to completely different results in separate regions.

Another approach is to compute SA for one input over the range of values for one or a few other inputs. This method is also inefficient because having a separate perturbation for each input over its range is time consuming and makes the analysis of the result difficult.

As an appropriate solution to this problem we have used the canonical of each separate cluster, which we obtained by training a Bidirectional Neural Network [11]. We also averaged the effects of two different positive and negative dithers.

#### 3.2.2. Causal Index

The causal index [6, 22] is a mathematical method. The significance of an input variable $i$ with respect to an output variable $j$ is calculated according to the following formula:

$$c_{ij}=\frac{dx_j}{dx_i}=f'(U_j).\,f'(U_h).\sum_{h=1}^{H} w_{jh}\, w_{hi} \qquad (1)$$

where $c_{ij}$ is the causal index measure of the significance of input $i$, $U_x$ is sum of the inputs to unit x, and $H$ is the number of hidden nodes. The proposers assume the product of first two terms $f'(U_j).\,f'(U_h)$ is constant, allowing (1) to be simplified to $c_{ij}=\sum_{h=1}^{H} w_{jh}\, w_{hi}$ (2)

#### 3.2.3. Hidden Index

Nejad and Gedeon [13] demonstrated the dependency of the simplified equation of causal index to the network structure and learning parameters. Thus, the assumption that product of first two terms is constant, is unfounded. We then introduced our method called the hidden index. Hidden indices are the columns of the matrix $C$ which is calculated according to: $C=M*M'$, where $M$ is the weight matrix of the trained network and $M'$ is its transpose.

Suppose $HI_{avg}$ is the average of the columns related to input units and $HI_\theta$ is the column in $C$ for biases. We showed that $HI = HI_{avg} - HI_{\theta}$ is less dependent on the structure of the network and shows more similar significances to sensitivity analysis values than causal index measure (at least for our data sets).

In this experiment we have used sensitivity analysis because used appropriately, this will provide more efficient answers less sensitive to the network structure than the causal index or hidden index. We set the dither to 0.05 and measured the sensitivity of inputs with respect to each output neuron around its class prototype. The class prototypes have been found by BDNN method [16]. The most important variables reported are those to which the network is most sensitive.

For each data set, five neural networks have been trained with different parameters. The best neural network has been selected according to the results on the test data. We have applied the sensitivity analysis for each class in our data sets. This has been repeated for the single continuous output encoding all output classes. For example, for the SFM data set a list of seven variables will be reported: one for each class and one for a single continuous output node.

### 3.3. Significance Measures in LDA

We had a similar problem in selecting a method to measure significance in the case of statistical approach. The correlation coefficients could not be used because of the possibility of existence of nonlinear relationships among the variables. For example, two variables may be highly related together but not linearly, so their correlation coefficient may be very small. In multivariate regression the coefficients of the regression equation can not be reliable due to the use of different measure units for each variable.

Standardised coefficients, Beta coefficients, and T test of significance provide better results in analysing the most significant variables in the multivariate regression method. Even these criteria may not be a good measure of significance. This happens due to some problems such as multicollinearity. The problem occurs when there are some significant correlations among the independent variables. Thus, the T test of significance may not assign a high value of significance to a really significant variable, if in stepwise entering of variables, one of the highly correlated variables was already entered into the equation. Moreover, this could not be used in the case of classification by the method of linear discriminant analysis due to the classes being discrete.

Thus, we extracted the most significant variables from the matrix of pooled within-groups correlations between discriminating variables and canonical discriminant functions for LDA according to the following procedure:

- 1. If the cumulative percent of variance for the first function is large enough (eg, more than 70% for our data), we use the ranked form of variables ordered by the size of correlation within functions. The functions are ordered by their related eigenvalue, so the first one has the highest significance among the functions.
- 2. Otherwise, we use a linear combination of the variables ordered by size of correlation within functions (in our experiments three functions). The percentage variance of the function in the discrimination procedure is used to determine the coefficients of this linear combination.

## 4. Data Sets

The following data sets have been used in our experiments.

### 4.1. Students Final Mark Prediction (SFM)

This data set consists of 153 samples. Each pattern has fourteen input attributes. Four outputs have been used to classify the marks. Each record comprises student information, assessment and subsequent final mark for a sample of students from a first year computer science subject at the University of New South Wales. The major exam component has been omitted, which introduced significant noise.

### 4.2. Geographical Information Systems (GIS)

This data set consists of satellite information from 190 samples from a rectangular grid of 244494 points. Each pattern has 16 input attributes and 5 outputs. 143 records have been used in the training set and 47 records have been used as unseen data. The satellite information has been collected, augmented with ancillary terrain data, and preprocessed in the School of Geography in the University of NSW to classify a large geographical area into some forest supra-type categories (eg. dry sclerophyll and wet sclerophyll).

### 4.3. Gross Domestic Product (GDP)

The original data set consisted of more than 150 records. Each record has fourteen variables which are assigned to some socio-economic statistical information. We removed some patterns due to missing data for more than three attributes. This remaining data set consists of 143 patterns. The data is used to predict the GDP for developing countries.

We have divided the data into a training set consisting of 101 patterns and a test set consisting of 42 patterns.

## 5. Experimental Results

In order to evaluate the generalisation ability, positional causality and data dependency in the three classification methods, we applied them to the above mentioned data sets. Table 1 summarises the results of the experiments. For each data set, we have shown the training error and the test error for each classification method. For each data set, the test data and the training data were the same and no special preprocessing has been applied to the data sets.

Table 1. Classifiers' performance on GDP, SFM and GIS.

| | | LDA | BP | C4.5 |
|---|---|---|---|---|
| **GIS** | Train Er | 26% | 24% | 6% |
| | Test Er | 32% | 30% | 26% |
| **GDP** | Train Er | 34% | 5% | 6% |
| | Test Er | 39% | 38% | 38% |
| **SFM** | Train Er | 14% | 0% | 4% |
| | Test Er | 26% | 28% | 30% |

There are no significant difference between the results of applying these classification methods on the test patterns. This may be because we have avoided any biases or a lack of favoured representation in the data sets for a particular classifier. There are many biases governing comparative methods with respect to the users and the data sets. These biases will be discussed later. However, if we applied appropriate data preparation techniques to the data sets for a particular classification method, we may get much better result for that method. Thus, we might think that the method has been the best or better one (as an example of preparation techniques for neural networks see [12]).

Tables 2 and 3 show the ranked significant attributes in SFM and GIS data sets for each of the classification methods. According to our experiments the order of significant variables may change slightly for each classifier. This may be due to the structure, training parameters, initialisation and data set. This seems natural and we can observe the same phenomenon in people.

Table 2. Ranked significant attributes in SFM.

| LDA | BP | C4.5 |
|---|---|---|
| Mid-Exam | Mid-Exam | Mid-Exam |
| Assign. H2 | Assign. P1 | Assign. H2 |
| Assign. P1 | Assign. H2 | Course |
| Lab-No. 4 | Assign. H1 | Lab-No. 7 |
| Assign. F1 | Lab-No. 10 | Lab-No. 2 |
| Lab-No. 7 | Lab-No. 7 | *** |
| Assign. H1 | Enrol-Status | *** |

Table 3. Ranked significant attributes in GIS.

| LDA | BP | C4.5 |
|---|---|---|
| Topo. Pos. | Topo. Pos. | Topo. Pos. |
| LBT-6 | LBT-5 | LBT-4 |
| Altitude | RA-1 | Altitude |
| GE-1 | TE | LBT-3 |
| LBT-7 | LBT-2 | LBT-6 |
| CA | GEE | LBT-2 |
| TE | LBT-3 | LBT-5 |
| LBT-5 | Altitude | Aspect |

From Tables 2 and 3, we can see that different methods assign different significance to a variable. This is essentially due to the way in which a learning algorithm organises (reorganises) and stores the knowledge.

Note that none of the classifiers were supported by expert knowledge in deciding the significant order of variables. This is done by the learning algorithms. This implies that the relative contribution of attributes could differ for each classifier. For example, the student mark on assignment P1 may be very important as represented by LDA and BP, but not by C4.5.

According to our results, the classifiers usually agree on the most significant variables (at least for our data sets). This implies that the main difference should be in measuring the partial correlations after entering the most significant variable. The most important variables have been *topological position*, *Mid-Exam* mark, and *Doctor per 1000 people* in the GIS, SFM, and GDP data sets.

Table 4. Ranked significance attributes in SFM classes.

| DIST | | CRED | | PASS | | FAIL | |
|---|---|---|---|---|---|---|---|
| Mid | 100 | Mid | 100 | Mid | 100 | Mid | 100 |
| P1 | 20 | p1 | 21 | p1 | 23 | p1 | 67 |
| Lab10 | -15 | H2 | 15 | H2 | 23 | Lab10 | 61 |
| H2 | 15 | Lab10 | -14 | H1 | 13 | Lab7 | 59 |
| H1 | 12 | H1 | 12 | Lab10 | -12 | Lab4 | 59 |
| ES | 10 | ES | -10 | Lab7 | 11 | Course | 32 |
| Lab4 | -10 | Lab4 | -9 | ES | -10 | H1 | 27 |

Table 4, shows the positional causalities of a trained neural network for SFM data set. The selected positions have been the cluster centroids obtained by BDNN. The numbers indicate that the contribution significance of the variables in the task of classification varies in each partition of the problem space.

The order of variables are not the same in different trees and different trained networks, at least for our data sets. For example, the most significant variables in one net or tree may be different in subsequent trials with the same or varied parameters.

In the case of C4.5 usually all variables will not appear in tree nodes. Each variable may appear in more than one level for an induction tree. The output trees usually are not balanced. The C4.5 rules are not essentially explanation rules but are discrimination rules.

## 6. Data dependency in classification methods

To train a learning method to efficiently partition a problematic model space, either the sample size should be increased or the model complexity decreased. Increasing the sample size is not always possible, so the model complexity should be decreased. Data preparation methods and model pruning techniques are two ways of reducing the model complexity.

Data representation methods usually either reduce the dimensionality of data or otherwise manipulate data ignoring the dimensionality. Even if the best data preparation methods are used, sample size limitations may hinder the learning model from finding the best fit for a particular problem.

The complexity of classification problems with high dimensionality arises from the fact that high variance in less significant or unimportant variables may change the appearance of similar patterns to completely different patterns and vice versa. Thus, dimensionality reduction methods are very important. Feature extraction, data transformations, principal component analysis and expert knowledge are usually used to decrease the data dimensionality.

LDA and C4.5 do not use all of the independent variables in the model produced. They use a variable only when it significantly increases the classification performance without losing some degree of freedom. In LDA, *F-to-remove* values should be computed to test the removal or contribution allowance of a new variable in the model. Dimensionality reduction is also very important in neural networks.

Referring to the *StatLog* project, Ripely [17] writes *"comparisons with other methods are rare, but when done carefully often show that statistical methods can outperform state of the art in neural networks"*. What Ripley is saying is correct, not only for statistical but also for neural and machine learning methods. To generalise his statement we argue that *when desired data preparation is done carefully for a particular model, the model can often out-perform other models.* We have shown this for the case of neural networks by introducing four preprocessing techniques [12].

## 7. Conclusion

We have revealed the lack of consistency among the previous comparative studies of classification methods. To analyse the contradictions, the biases (both user and data dependent) were considered and use of significance measures in classification methods was investigated.

We showed that global causalities may cause confusion instead of comprehension. Thus, a positional causality is preferred. We investigated the plausibility of determining the positional causality among the classifiers. We concluded that from the classification methods tested, neural networks trained by BP are reliable tools with which to analyse the positional causality of problem features. This shows an advantage of neural networks from the data transparency point of view, for which they are often otherwise criticised.

The analysis of the ranked list of significant variables showed a main difference among the contribution of independent variables in the task of classification. This was not the case in comparing the generalisation ability of the classifiers. This may imply useful suggestions for designing hybrid systems in the future.

## 8. References

[1] Breiman, L, Friedman, JH, Olshen, RA and Stone, CJ *Classification and Regression Trees,* Belmont, California: Wadsworth, 1984.

[2] Cheng, B, Titterington, DM "Neural Networks: A Review from a Statistical Perspective," *Statistical Science,* Vol. 9, No, 1, 1994.

[3] Firebaugh, MW *Artificial Intelligence a Knowledge-Based Approach,* PWS-KENT Publishing Company, Boston, 1989.

[4] Flexer, A "Connectionists and Statisticians, Friends or Foes," *Proceedings of International Workshop on Artificial Neural Networks,* Malaga, 1995.

[5] Geman, S, Bienenstock, E and Doursat, R "Neural Networks and the Bias/Variance Dilemma," *Neural Computation,* vol. 4, pp. 1-58, 1992.

[6] Hora, N, Enbutsu, I and Baba,K. "Fuzzy rule extraction from a multilayer neural net," *Proc. IEEE,* vol. 2, pp. 461-465, 1991.

[7] Levine, E, Tishby, N and Solla, SA "A Statistical Approach to Learning and Generalisation In Layered Neural Networks," *Proc. IEEE,* vol. 78, no. 10, 1990.

[8] Mansuri, Y, Kim, JG, Compton, P and Sammut, C "A Comparison of a Manual Acquisition Method and an Inductive Learning Method," *The First Australian Workshop on Knowledge Acquisition for Knowledge Based Systems,* pp. 114-132, Pokolbin, 1991.

[9] McClelland, JL and Rumelhart, DE *Explorations in Parallel Distributed Processing,* Cambridge: MIT Press, 1988.

[10] Michie, D, Spiegelhalter, DJ, Taylor, CC (eds.) *Machine Learning, Neural and Statistical Classification,* Ellis Horwood, England, 1994.

[11] Nejad, AF and Gedeon, TD "Bidirectional Neural Networks and Class Prototypes," *Proceedings of IEEE ICNN,* Perth, Australia, 1995.

[12] Nejad, AF and Gedeon, TD "Bidirectional Neural Networks Reduce Generalisation Error," in *From Natural to Artificial Neural Computation*, Mira, J, Sandoval, F, Springer Verlag, pp. 543-550, 1995.

[13] Nejad, AF and Gedeon, TD "Analyser Neural Networks: An Empirical Study in Revealing Regularities of Complex Systems," in *Applic. of Neural Networks to Telecommunication 2*, Springer Verlag, Sweden, pp. 281-289, 1995.

[14] Quinlan, JR "Learning with Continuous Classes," *Proceedings 5th Australian Joint Conference on Artificial Intelligence,* Singapore: World Scientific, 1992.

[15] Quinlan, JR *C4.5: Programs for Machine Learning,* San Mateo, California: Morgan Kaufmann, 1993.

[16] Quinlan, JR "A Case Study in Machine Learning," *Proceedings 16th Australian Computer Science Conference*, Brisbane, 1993.

[17] Ripley, BD "Statistical Aspects of Neural Networks," in Barndorff-Nielsen, OE, Jensen, JL, Kendall, WS (eds.), *Network and Chaos: Statistical and Probabilistic Aspects*, London: Chapman & Hall, 1993.

[18] Sarle, WS "Neural Networks and Statistical Models," *Proc. 19th Annual SAS Users Group Int. Conf.*, Cary, NC: SAS Inst., 1994.

[19] Shannon, CE and Weaver, W *The Mathematical Theory of Communication*,Urbana: University of Illinois press, 1964.

[20] Weiss, S and Kulikowski, C *Computer Systems That Learn,* Morgan Kaufmann, San Mateo, 1991.

[21] White, H, "Learning in Artificial Neural Networks: A Statistical Perspective," *Neural Computation,* vol 1., pp. 425-464, 1989.

[22] Yoda, M, Baba, K and Enbutu, I. "Explicit representation of knowledge acquired from plant historical data using neural networks," *IJCNN,* San Diego, vol. 3, pp. 155-160, 1991.