Pre-publication draft of a paper which appeared in the Proceedings of the Tenth Biennial Canadian Artificial Intelligence Conference (AI-94), pages 63-70. San Francisco: Morgan Kaufmann.

Incorporating canonical discriminant attributes in classification learning

Simon P. Yip
Dept. of Computer Science
Swinburne University
Hawthorn 3122, Australia

Geoffrey I. Webb
School of Computing and Maths
Deakin University
Geelong 3217, Australia

Abstract

This paper describes a method for incorporating canonical discriminant attributes in classification machine learning. Though decision trees and rules have semantic appeal when building expert systems, the merits of discriminant analysis are well documented. For data sets on which discriminant analysis obtains significantly better predictive accuracy than symbolic machine learning, the incorporation of canonical discriminant attributes can benefit machine learning. The process starts by applying canonical discriminant analysis to the training set. The canonical discriminant attributes are included as additional attributes. The expanded data set is then subjected to machine learning. This enables linear combinations of numeric attributes to be incorporated in the classifiers that are learnt. Evaluation on the data sets on which discriminant analysis performs better than most machine learning systems, such as the Iris flowers and Waveform data sets, shows that incorporating the power of discriminant analysis in machine classification learning can significantly improve the predictive accuracy and reduce the complexity of classifiers induced by machine learning systems.

1. Introduction

Attribute-based or selective inductive classification learning algorithms aim to develop procedures capable of correctly classifying instances of disjoint classes. The condition parts of the classifiers are based on the values of attributes provided in the examples. These algorithms have not in general supported the derivation of conditions based on relationships between attributes. "It is obvious that if the class description is outside the description space that is defined in terms of available attributes or features, then it can only be learnt by extending that space. Indeed, it is possible that the relevant attributes or best features that could be used in the class description may not be explicit or included in the examples" [Elio and Watanabe, 1991]. The issue of constructing new attributes or features is closely related to constructive induction [e.g. Rendell and Seshu, 1990; Michalski, 1983a, Bloedorn and Michalski, 1991]. This paper describes methods of constructing new attributes by incorporating discriminant analysis.

Discriminant analysis is another popular classification method [e.g. Klecka, 1980]. There are two major types of discriminant analysis. Parametric methods assume normal distribution of the attributes while the nonparametric methods have no such assumption. Though discriminant analysis is a powerful classification method, unlike symbolic machine learning, the classifiers it develops do not have the semantic appeal of decision trees and rules. The latter offers modularized clearly explained formats for describing a decision procedure and are compatible with a human's reasoning procedures and expert system knowledge bases. Unlike parametric discriminant analysis, machine learning systems do not depend on the assumption that the attributes are normally distributed and uncorrelated. Previous research has shown that both symbolic machine learning and statistical techniques produce superior classifiers to those produced by the other on differing data sets [Weiss and Kapouleas, 1989; Holte, 1993; Breiman et al, 1984].

This paper describes techniques for incorporating parametric discriminant analysis in symbolic machine learning. Previous machine learning systems which attempt to incorporate parametric discriminant analysis include CART [Breiman et al., 1984] and LMDT (Linear machine decision tree) [Utgoff and Brodley, 1991]. In these systems, linear combinations of attributes are searched and evaluated before each node of a decision tree is created. In CART, which uses piecewise linear discriminants, the computation cost increases tremendously as the number of attributes and nodes increases. In LMDT, a complicated encoding and weight training system is implemented at each node. Another approach is used by SWAP1 [Weiss and Indurkhya, 1991] where discriminant functions are transformed to binary attributes. One binary attribute per class is added to the attribute space. Each binary attribute represents the classification result of a discriminant function. The system reports rules such as:

If (LD1 & (x > 109)) then class=1;

where *x* is a continuous attribute and *LD1* represents the condition that the instance is classified by a set of linear

discriminant functions as class *1*. The use of such attributes greatly reduces the ease with which the rule is comprehended. In the above approaches, the discriminant functions are based on the equation:

$f_i(\boldsymbol{x}) + ln(P(C_i)) > f_j(\boldsymbol{x}) + ln(P(C_j))$ " $i \neq j$

For each class C_i, $f_i(\boldsymbol{x})$, a linear function of the set of attributes, $\boldsymbol{x}$, is derived. An unknown case is classified by applying the functions and choosing the class whose linear score is the largest. Another discriminant analysis technique is canonical discriminant analysis which is based on a different type of function. This paper reports methods of deriving and incorporating canonical discriminant attributes in classification learning.

2. Incorporating canonical discriminant analysis

Canonical discriminant analysis is a dimension-reduction technique related to principle component analysis and canonical correlation. [e.g. Klecka, 1980]. It derives combinations of attributes to maximise the difference of the centroid of different classes. This research investigates incorporating canonical discriminant analysis in inductive classification learning. A canonical discriminant function is a linear combination of the discriminating attributes. It has the following mathematical form:

$$f_{km} = u_o + u_1X_{1km} + u_2X_{2km} + ... + u_pX_{pkm}$$

where f_{km}=the value (score) on the canonical discriminant function for case m in the class k; X_{ikm}= the value on discriminant attribute X_i for case m in class k; u_i=coefficients which produce the desired characteristics in the function.

The maximum number of unique functions that can be derived is equal to the number of classes minus one or the number of attributes, whichever is fewer. The coefficients (the u's) for the each function are derived so as to maximise the distance between the class centroids. A class centroid is a imaginary point which has coordinates that are the class's mean on each of the attributes. In discriminant analysis, classification is a separate activity. The canonical discriminant functions can be used to predict the class to which an unseen case most likely belongs. Several classification procedures exist, but they all use the notion of comparing the case's position to each of the class centroids in order to locate the closest centroid. Since canonical functions aim to maximise the distance between class centroids, they can be utilised to transform the instance space (space containing training instances for learning) so as to maximise the linear separability of cases. As symbolic machine learning systems seek to develop linear partitions of the instance space, in this research, we incorporate the canonical function(s) as additional attribute(s) in the attribute space before submitting the expanded data set to inductive classification learning. Two classification learning systems are employed, C4.5 [Quinlan, 1993] and Einstein [Webb, 1992a]. C4.5, is decision tree based while Einstein, based on the algorithm DLG [Webb, 1992b], a variant of Aq [Michalski, 1983a], induces decision rules. To illustrate, suppose we have the following data:

X	Y	Z	Class
2	9	11	P
4	8	12	P
7	3	15	P
5	12	20	N
15	7	12	N
11	9	10	N
2	8	17	Q
3	10	15	Q
7	6	20	Q

With C4.5 [Quinlan, 1993], the decision tree induced is:

```
X > 7 : N (2.0)
X <= 7 :
|  Z <= 15 : P (4.0/1.0)
|  Z > 15 : Q (3.0/1.0)
```

With Einstein [Webb, 1992a], the rules induced are:

If (X <= 7 & Y <= 9 & Z <= 15) then class=P [3]
If (X >= 5.00 & Y >= 7.00) then class=N [3]
If (6<=Y<=10 & Z >= 15) then class=Q [3]

To incorporate canonical discriminant analysis, we can perform the following. By applying canonical discriminant analysis (available from most statistical packages such as SAS®, [1990]), canonical functions are derived. With three classes, two canonical discriminant functions are derived. The raw coefficients of the first canonical function for attributes *X*, *Y* and *Z* are 0.711, 0.903 and 0.226, respectively. Since the relative values of canonical attributes are the focus of classification, we can ignore the constant term in the canonical function. Thus, the value of the first canonical attribute (CAN1) for the first case is thus equal to: 2*0.711 + 9*0.903 + 11*0.226 = 12.04. Similarly, we can derive other canonical attribute values. The expanded data set is as follows:

X	Y	Z	CAN1	CAN2	class
2	9	11	12.04	3.28	P
4	8	12	12.78	3.45	P
7	3	15	11.08	3.95	P
5	12	20	18.91	5.74	N
15	7	12	19.70	3.13	N
11	9	10	18.21	2.79	N
2	8	17	12.49	4.83	Q
3	10	15	14.56	4.37	Q
7	6	20	14.92	5.42	Q

Submitting the expanded data set to a classification learning algorithm, the following concise trees or rules are derived:

With C4.5 [Quinlan, 1993], the decision tree induced is:

```
CAN1 > 14.92 : N (3.0)
CAN1 <= 14.92 :
|  CAN2 <= 3.95 : P (3.0)
|  CAN2 > 3.95 : Q (3.0)
```

With Einstein [Webb, 1992a], the rules are:
If (CAN1 <= 12.78 & CAN2 <= 3.95) then class=P [3]
If (CAN1 >= 18.21) then class=N [3]
If (CAN1 <= 14.92 & CAN2 >= 4.37) then class=Q [3]

From the theoretical perspective, incorporating canonical discriminant attributes is a form of empirical constructive induction. According to the framework for constructive induction developed by Rendell & Seshu, [1990], creating new attributes from existing attributes is termed feature construction. Feature construction can supplement the deficiency of selective induction in learning *hard* concepts. A concept is hard if its attributes have accurate class membership information but the concept cannot be learned by selective inductive methods. Hard concepts are characterised by dispersed and oddly shaped *peaks* in the instance space. Feature construction is the process of bringing together uniform regions that are dispersed in the instance space. Examination of the scatter plots on the above example supports the theoretical perspective.

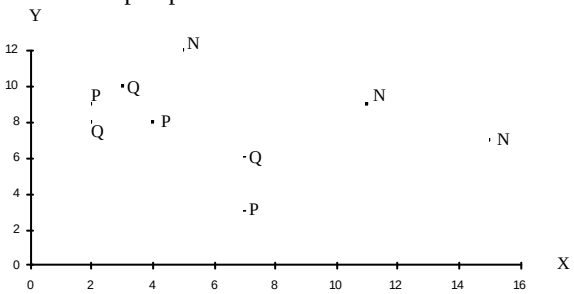


Figure 1: Scatter plot of X-Y

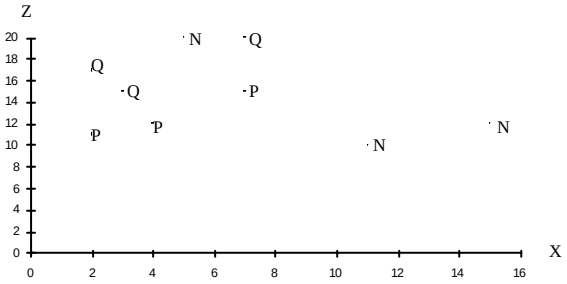


Figure 2: Scatter plot of X-Z

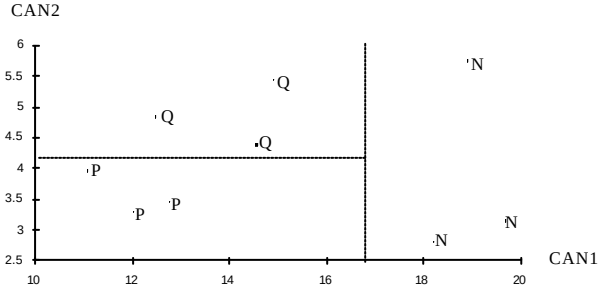


Figure 3: Scatter plot of canonical attributes
(dotted lines showing simple decision surfaces)

In the scatter plots of (X-Y) and (X-Z), i.e. Fig.1 and 2, we can observe that the class members are dispersed. No simple decision surfaces can be found. In the scatter plot of the canonical attributes (Fig.3), the class members are grouped together and simple decision surfaces are easily found.

3. CAF (Canonical attribute finding)

We call the process of deriving and incorporating canonical attributes as CAF. The objective of the procedure is to find combinations of existing attributes that can contribute to the discrimination performance of existing attributes. When they are derived, each of the combinations is transformed into a single attribute and added to the attribute space. The application of CAF is indicated when the predictive accuracy of discriminant analysis for the domain is significantly higher than that obtained by the machine learning system under focus. The algorithm can be expressed as follows:

Algorithm: CAF
Input: a training set of examples
Output: an expanded training set of examples (*ET*)
Begin
    raw canonical coefficients ←canonical discriminant analysis on attributes;
    canonical attribute values ←(attribute values, raw canonical coefficients);
    ET ←Extend the descriptions of examples to include canonical attribute(s) as additional attribute(s);
End.

4. CCAF (Clustering before Canonical attribute finding)

Existing methods [e.g. Breiman et al., 1984; Utgoff & Brodley, 1991] for finding good attribute combinations involve search at each node when constructing the decision tree. Such methods involve high computation costs. If CAF is applied to subsets of data or at each node in building decision trees, the search cost at each node can be significantly reduced, but the computation cost to apply CAF for every node remains. CCAF is a method for tackling part of this problem. This method starts by applying clustering to re-classify the training set before deriving canonical attributes. It is useful when the possible partitions of the data are different from that given in the training examples. Two main categories of clustering methods exist: conceptual clustering [e.g. Michalski, 1983b] and cluster analysis [e.g. Everitt, 1980]. Since the objective is to derive canonical attributes, cluster analysis is used in this research. The common cluster analysis methods are based on agglomerative hierarchical clustering procedures. Each observation begins in a cluster by itself. Two clusters can be merged to form a new cluster that replaces the two old clusters. Various clustering methods differ in how the distance between two clusters is computed. In this study, we use Ward's minimum-variance method.

By re-classifying a training set of examples into clusters before deriving canonical attributes, we can capture the partition information of clusters. CCAF is indicated when the predictive accuracy obtained by discriminant analysis for the domain is significantly higher than that of the machine learning system under focus. In this research, we set the maximum number of clusters to two times the number of different classes. The algorithm can be restated as follows:

Algorithm: CCAF
Input: a training set of examples
Output: an expanded training set of examples (*ET*)
Begin
    clusters ←cluster analysis on attributes;
    raw canonical coefficients ←canonical discriminant analysis on attributes with clusters as classes;
    canonical attribute values ←(attribute values, raw canonical coefficients)
    ET ←Extend the descriptions of examples to include canonical attribute(s) as additional attribute(s);
End.

5. Evaluation

Previous studies comparing discriminant analysis with classification machine learning have found that each approach outperforms the other on different sets of data. [Weiss & Kapouleas, 1989; Holte, 1993; Breiman et al, 1984]. Since we are interested in improving the performance of machine leaming by incorporating canonical discriminant analysis, we select the ones on which discriminant analysis performs better. In this research, we use the Iris plants and Waveform data sets [Murphy & Aha, 1994]. The statistical package SAS® [1990] is used for canonical discriminant and cluster analysis. The machine learning systems used are C4.5 [Quinlan, 1993] and Einstein [Webb, 1992a].

5.1 Study 1 (Iris flower data)

In this study, we use the widely examined Iris flower data set with 150 examples. Each example consists of four numeric-valued attributes: sepal length, sepal width, petal length and petal width in centimetres. There are 3 classes of species: Iris setosa, Iris versicolor and Iris virginica. To enable comparison with other learning algorithms, this study uses the "leave-one-out" cross-validation method [e.g. Breiman et al., 1984]. The Chi-square test for correlated samples is used to compare predictive accuracy under different methods and the pair-wise t-test to compare complexity of induced classifiers. In the following tabulation of results, complexity refers to the number of rules or nodes in the classifier; CAF+C4.5, for example, represents the method of treating the data set with CAF before submitting to C4.5:

Method	Accuracy(%)	Complexity
(1) C4.5		
(pruned)	94.67	10.79
(rules)	95.3	4.02
(2) CAF+C4.5		
(pruned)	98	5
compare (1):	(χ^2=5; p≤.05)	(t=81.32; p≤.0005)
(rules)	96.67	3.96
compare (1):	(χ^2=1)	(t=2.77; p≤.025)
(3) CCAF+C4.5		
(pruned)	94	10.8
compare (1):	(χ^2=0.33)	(t=-0.38)
(rules)	94.67	4.05
compare (1):	(χ^2=0.33)	(t=-1.91)
(4) Einstein	96	7.03
(5) CAF+Einstein	96	5.98
compare (4):	(χ^2=0)	(t=36.21; p≤.0005)
(6) CCAF+Einstein	95.3	6.05
compare (4):	(χ^2=0.33)	(t=39.04; p≤.0005)

In the above tabulation, we observe that by deriving and adding canonical attributes in a data set with CAF, the performance of induced decision trees or rules can be significantly improved and the complexity significantly reduced. The results of applying CCAF are insignificant. The best result of incorporating canonical discriminant analysis can be compared to other methods that use leave-one-out evaluation design:

	Accuracy(%)
(5) CAF+C4.5(pruned) [this paper]	98
(6) Linear discriminant [this paper]	98
(7) Quadratic discriminant [this paper]	97.33
(8) Nearest neighbor, k=1 [this paper]	96.67
(9) CART [Weiss & Kapouleas, 1989]	95.3
(10) EACH [Salzberg, 1991]	95.3
(11) Neural net [Weiss & Kapouleas, 1989]	96.7
(12) PVM [Weiss & Kapouleas, 1989]	96.0
(13) SWAP1 [this paper]	97.33
(14) SWAP1+discriminant[1] [this paper]	96.67

In the above comparison to other methods, the predictive accuracy of C4.5 (pruned tree) is improved to equal that of linear discriminant analysis. The effect of CAF can be further examined by plotting the performance vs. training size graph. In this study, 20% of the data set is used as the evaluation set and the training set consists, in turns, of 40%, 60% or 80% of the data set. The performance of the induced trees or rules of each training set is evaluated over 10 runs. The comparative predictive accuracy and complexity is illustrated in the following graphs:

[1] Represents SWAP1 with discriminant analysis option

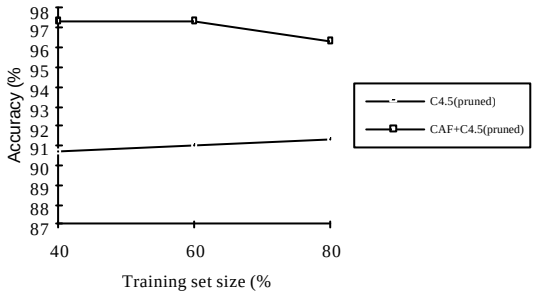


Figure 4: CAF+C4.5(pruned) vs. C4.5(pruned) Accuracy-Training_size-plot on Iris data

In Figure 4, the predictive accuracy of CAF+C4.5(pruned) is significantly better than that of C4.5(pruned) at all three training set sizes ($t_{40\%}$=6.71, p≤.0005; $t_{60\%}$=6.05, p≤.0005; $t_{80\%}$=6.71, p≤.0005). In the above presentation, "$t_{40\%}$", for example, represents the t-value when the training set size equals 40% of the data set.

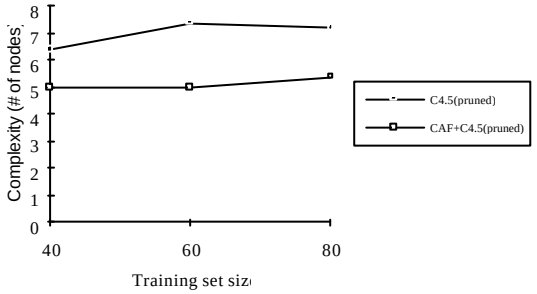


Figure 5: CAF+C4.5(pruned) vs. C4.5(pruned) Complexity-Training_size-plot on Iris data

Figure 5 shows that the complexity of pruned trees induced by CAF+C4.5 is significantly less than that of C4.5 alone, at all three training sizes ($t_{40\%}$=3.25, p≤.005; $t_{60\%}$=4.81, p≤.0005; $t_{80\%}$=4.58, p≤.005). The pattern of the performance vs. training size graphs of C4.5 rules is similar to that of C4.5 pruned trees.

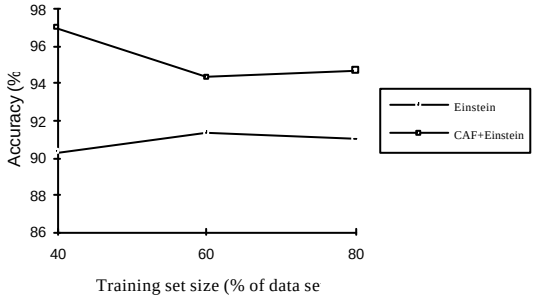


Figure 6: CAF+Einstein vs. Einstein Accuracy-Training_size-plot on Iris data

In Figure 6, we can observe that the predictive accuracy of CAF+Einstein is significantly better than that of Einstein alone at all three training set sizes ($t_{40\%}$=2.33, p≤.025; $t_{60\%}$=2.59, p≤.025; $t_{80\%}$=4.71, p≤.005).

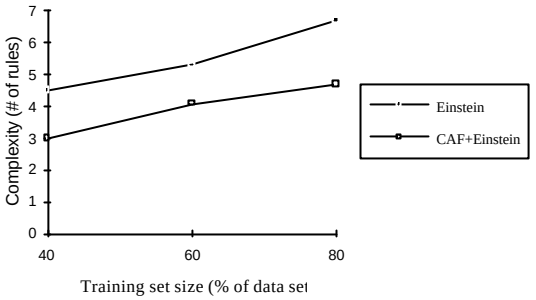


Figure 7: CAF+Einstein vs. Einstein Complexity-Training_size-plot on Iris data

Figure 7 shows that the complexity of rules learned by CAF+Einstein is significantly less that learned by Einstein alone at all three training sizes ($t_{40\%}$=7.75, p≤.0005; $t_{60\%}$=2.45, p≤.025; $t_{80\%}$=7.75, p≤.0005).

In this study, we observe that by deriving and incorporating canonical discriminant attributes in machine learning, we can significantly improve the predictive accuracy and reduce the complexity of classifiers induced from various size of the training data. The predictive accuracy of CAF+C4.5(pruned) can be compared to that of other methods as follows:

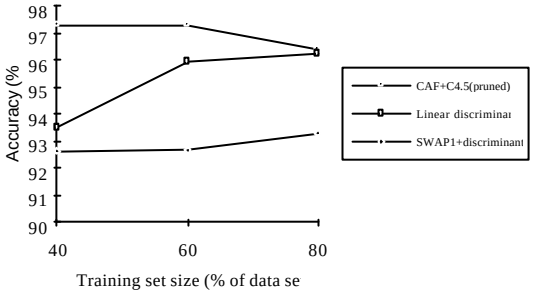


Figure 8: CAF+C4.5(pruned) vs. Linear discriminant vs. SWAP1+discriminant Accuracy-Training_size-plot on Iris data

Figure 8 shows that the predictive accuracy of CAF+C4.5(pruned) is better than that of Linear discriminant analysis ($t_{40\%}$=2.46, p≤.025; $t_{60\%}$=2.08; $t_{80\%}$=0.13) and SWAP1+discriminant ($t_{40\%}$=3.78, p≤.005; $t_{60\%}$=3.5, p≤.005; $t_{80\%}$=5.02, p≤.0005). The performance of CAF+C4.5(pruned) at training set size of 40% is particularly notable.

5.2 Study 2

5.2.1 Waveform data set

In this study, we use the waveform data set used by the CART system [Breiman et al., 1984]. The data were generated with the program published in the UCI data base [Murphy & Aha, 1994]. It is a three class problem based on the three waveforms $h_1(t)$, $h_2(t)$ and $h_3(t)$ graphed as follows:

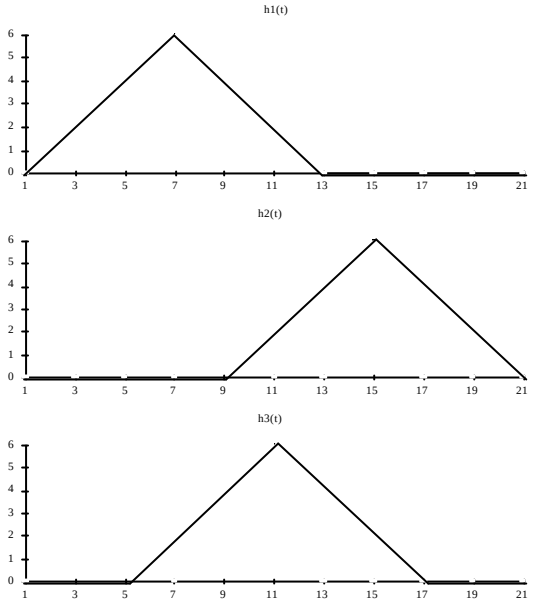


Figure 9: The three underlying waveforms

Each class consists of a random convex combination of two of these waveforms sampled at the integers with noise added. The measurement vectors are of 21 dimensions: $\mathbf{x}=(x_1,..x_{21})$. To generate each vector **x**, a uniform random number u and 21 random numbers $\varepsilon_1,....\varepsilon_{21}$ normally distributed with mean zero and variance 1 are generated:
For class 1 vectors, set: $x_m=uh_1(m) + (1-u)h_2(m) + \varepsilon_m$
For class 2 vectors, set: $x_m=uh_1(m) + (1-u)h_3(m) + \varepsilon_m$
For class 3 vectors, set: $x_m=uh_2(m) + (1-u)h_3(m) + \varepsilon_m$
where m=1,... 21

In order to compare performance with that of other studies, in this research, training sets of 300 examples using prior probabilities of (1/3, 1/3, 1/3) and a test data set of 5000 records are generated. The mean performance over 10 runs are as follows:

Method	Accuracy(%)	Complexity
(1) C4.5		
(pruned)	71.08	57.6
(rules)	70.44	14.9
(2) CAF+C4.5		
(pruned)	76.1	43.8
compare (1):	(t=8.15, p≤.0005)	(t=6.65, p≤.0005)
(rules)	76.58	10.3
compare (1):	(t=8.30; p≤.0005)	(t=2.9, p≤.01)
(3)CCAF+C4.5		
(pruned)	78.47	46.0
compare (1):	(t=11.05; p≤.0005)	(t=5.41, p≤.0005)
(rules)	78.91	11.6
compare (1):	(t=10.42; p≤.0005)	(t=2.41, p≤.025)
(4) Einstein	71.53	11.0
(5) CAF+Einstein	73.83	9.8
compare (4):	(t=6.23, p≤.0005)	(t=4.33, p≤.005)
(6) CCAF+Einstein	73.51	10.1
compare (4):	(t=4.5, p≤.005)	(t=3.0, p≤.01)

In the above tabulation, we observe that by incorporating canonical discriminant attributes in machine learning, the predictive performance can be significantly improved and the complexity of classifiers significantly reduced. The best result of incorporating canonical discriminant attributes in this study can be compared to other learning systems as follows:

Method	Accuracy(%)
(7) Linear discriminant [this paper]	80.72
(8) Quadratic discriminant [this paper]	78.45
(9) CCAF+C4.5 (pruned) [this paper]	78.47
CCAF+C4.5 (rules) [this paper]	78.91
(10) CART [Breiman et al., 1984]	72
(11) Nearest neighbor [Breiman et al., 1984]	78
(12) CART with 55 attributes added [Breiman et al., 1984]	80
(13) CART with linear combination [Breiman et al., 1984]	80
(14) SWAP1[2] [this paper]	73.06
(15) SWAP1+discriminant [this paper]	79.0

In method (12), the 55 new attributes added were based on the averages, $X_{m1,m2}$, over the attributes from m_1 to m_2 for odd values of m_1 & m_2 , where

$$X_{m1,m2} = 1/(m_2-m_1+1) \sum_{m=m_1}^{m_2} X_m,\ m_2 > m_1$$

In method (13), the linear attribute combination algorithm used by CART [Breiman et al, 1984] involves repetitive search for the best combination of attributes at each node to make up the best split when generating the classification tree. The computation cost of this method is high and the attribute evaluation function is system dependent. By using CCAF as an independent pre-machine learning step, the predictive performance of C4.5 rules is increased from 70.44% to 78.91%, and the classifier complexity is reduced from 14.9 rules to 11.6 rules. However, the accuracy performance is still less than that of linear discriminant analysis (t=2.99, p≤.01) and SWAP1+discriminant (t=0.14).

In this evaluation, we showed that by incorporating canonical discriminant analysis as a pre-symbolic classification learning step, the predictive accuracy and complexity of classifiers can be significantly improved when compared to classification learning alone.

5.2.2 Waveform data set with noise

In the evaluation of the CART system, waveform data sets containing the original 21 attributes plus 19 noise attributes were also used. In this study, we use a similar noisy data set generated by the published program of UCI

[2] The data set, which contains noise, is offset by +4, because SWAP1 accepts only positive numbers.

database. Again, a training data set with 300 examples and a testing data set with 5000 cases were generated. Because of system limitations of Einstein and SWAP1, only C4.5 is used in this part. The results, based on 10 runs can be presented as follows:

Method	Accuracy(%)	Complexity
(1) C4.5		
(pruned)	68.63	58.0
(rules)	67.80	12.30
(2) CAF+C4.5		
(pruned)	73.35	38.60
compare (1):	(t=7.72, p≤.0005)	(t=12.93, p≤.0005)
(rules)	72.89	8.8
compare (1):	(t=7.02; p≤.0005)	(t=5.22, p≤.0005)
(3) CCAF+C4.5		
(pruned)	74.45	46.20
compare (1):	(t=7.3; p≤.0005)	(t=5.5, p≤.0005)
(rules) 75.06	12.0	
compare (1):	(t=7.7; p≤.0005)	(t=0.2)

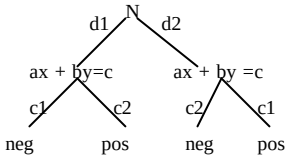
In the above tabulation, we also observe that by incorporating canonical discriminant attributes in machine learning, the predictive accuracy can be significantly improved and the complexity of classifiers significantly reduced. The best result of incorporating canonical discriminant attributes in this study can be compared to other learning systems as follows:

Method	Accuracy(%)
(4) CCAF+C4.5 (rules) [this paper]	75.06
(5) Linear discriminant analysis [this paper]	76.26
(6) Quadratic discriminant [this paper]	70.72
(7) CART [Breiman et al., 1984]	72
(8) Nearest neighbor [Breiman et al., 1984]	38

By using CCAF as an independent pre-machine learning step, the predictive accuracy of C4.5 rules is increased from 67.8% to 75.06% but still slightly less than that of linear discriminant analysis (t=1.65). In this study, very noisy data sets are used. Evaluation shows that by re-classifying the data with cluster analysis and deriving canonical discriminant attributes as additional attributes before submitting to classification learning, the predictive accuracy can be significantly improved and classifier complexity significantly reduced when compared to classification learning alone.

5.3 Study 3

Parametric discriminant analysis is quite robust to violation of the assumption that the attributes are normally distributed. The purpose of this study is to illustrate the incorporation of canonical discriminant analysis on mixed attributes and the merit of CCAF under certain conditions. In this study, we use an artificial data set generated with the following decision tree in mind:



In the above decision tree, N is a discrete attribute; x and y are continuous attributes; $d1$ and $d2$ are different discrete values; a, b, c, $c1$, and $c2$ are continuous values where $a \neq b$; $c1 \neq c2$. To generate an artificial set, we set $d1=2$, $d2=3$, $a=3$, $b=2$, $c1=135$ and $c2=150$. The class values are {pos, neg}. With the above decision tree in mind, we generated 400 cases, with 100 for each of the 4 leaves. For each case, x is assigned a random value between 1 and 40. With 10-fold cross validation, the performance of different methods can be presented as follows:

Method	Accuracy(%)	Complexity
(1) C4.5		
(pruned)	81.0	92.2
(rules) 84.0	37.9	
(2) CAF+C4.5		
(pruned)	74.25	109.2
compare (1):	(t=-3.62, p≤.005)	(t=-2.58, p≤.025)
(rules) 75	40.8	
compare (1):	(t=-4.19, p≤.005)	(t=-1.39)
(3) CCAF+C4.5		
(pruned)	98.75	8.2
compare (1):	(t=8.63, p≤.0005)	(t=19.82, p≤.0005)
(rules) 98.75	4.1	
compare (1):	(t=7.17, p≤.0005)	(t=25.54, p≤.0005)
(4) Einstein	90.5	40.1
(5) CAF+Einstein	85.5	42.3
compare (4):	(t=-1.96)	(t=-1.99)
(6) CCAF+Einstein	100	4.4
compare (4):	(t=6.22, p≤.0005)	(t=52.2 p≤.0005)

	Accuracy(%)
(7) Linear discriminant	43.5
(8) Quadratic discriminant	100
(9) Nearest neighbor (k=1)	100
(10) SWAP1	96.5
(11) SWAP1+discriminant	55.75

As illustrated above, the use of CAF worsens the performance of classifiers obtained by the machine learning systems, but the use of CCAF leads to significant increases in predictive accuracy and decreases in classifier complexity. The improved performance is comparable to that of quadratic discriminant and nearest neighbour methods. The incorporation of linear combinations of mixed attributes may reduce the semantic appeal of classifiers. Alternatively, linear combinations of continuous attributes only can be derived.

6. Discussion and Conclusion

In this paper, we presented two methods: CAF and CCAF, which incorporate the power of discriminant analysis into symbolic machine learning by deriving canonical discriminant attributes and adding them to the original attribute space. The expanded data set is then subjected to classification learning. Evaluation on data sets on which discriminant analysis performs better than most machine learning systems, shows that such techniques can significantly improve the performance of the machine learning systems. Linear combinations of attributes are derived with low search and computation costs. Stepwise discriminant analysis can also be used to reduce the number of terms in the linear combination. Alternatively, terms with coefficients close to zero can be discarded. Experiments on other data sets suggest that the better in accuracy performance of discriminant analysis over selective induction, the more significant is the positive effect on selective induction by incorporating canonical discriminant analysis.

In conclusion, discriminant analysis and symbolic inductive machine learning have been two important techniques in classification learning. Each has its own advantages and limitations. This paper demonstrates methods for combining these techniques. With a pre-machine learning step to derive and incorporate canonical discriminant attributes, we can significantly improve the predictive accuracy and decrease complexity of classifiers obtained by existing symbolic machine learning systems.

Acknowledgments

The authors wish to thank Ross Quinlan for C4.5 and Sholom Weiss for SWAP1.

References

[Bloedorn and Michalski, 1991] Bloedorn, E. and Michalski, R.S. (1991) Data driven constructive induction in AQ17-PRE A method and experiments. In *Proceedings of the third international conference on tools for AI*. San Jose, CA. 30-37.

[Breiman et al., 1984] Breiman, L., Friedman, J.H., Olshen, R.O. & Stone, C.J. (1984) *Classification and regression trees.* Wadsworth International Group, Belmont, California.

[Elio and Watanabe, 1991] Elio, R. & Watanabe, L. (1991) An incremental deductive strategy for controlling constructive induction in learning from examples. *Machine Learning*. 7, 7-44.

[Everitt, 1980] Everitt, B.S. (1980) *Cluster analysis.* 2nd edition, London: Heineman Educational Books.

[Holte, 1993] Holte, R.C. (1993) Very simple classification rules perform well on most commonly used data sets. *Machine learning*, 11, 1: 63-91.

[Klecka, 1980] Klecka, W.R. (1980) *Discriminant analysis.* Sage: Beverley Hills.

[Michalski, 1983a] Michalski, R.S. (1983) A theory and methodology of inductive learning. In Michalski, R.S., Carbonnel, J,G. & Mitchell, T.M. (Ed.) *Machine learning: an artificial intelligence approach.* Springer-Verlag.

[Michalski, 1983b] Michalski, R.S. (1983) Learning from observations: conceptual clustering. In Michalski, R.S., Carbonnel, J,G. & Mitchell, T.M. (Ed.) *Machine learning: an artificial intelligence approach.* Springer-Verlag.

[Murphy and Aha, 1994] Murphy, P.M. & Aha, D. (1994) UCI repository of machine learning databases. Irvine, CA: University of California, Dept. of information and Computer Science.

[Quinlan, 1993] Quinlan, R. (1993) *C4.5 Programs for machine learning.* Morgan Kaufmann.

[Rendell and Seshu, 1990] Rendell, L. & Seshu, R. (1990) Learning hard concepts through constructive induction: framework and rationale. *Computational intelligence*, 6: 247-270

[Salzberg, 1991] Salzberg, S. (1991) A nested hyper-rectangle learning method. *Machine learning*, 6:251-276.

[SAS® ,1990] SAS Institute Inc., Cary, NC: 27513

[Utgoff and Brodley, 1991] Utgoff, P.E. & Brodley, C.E. (1991) Linear machine decision trees. COINS Technical Report 91-10, University of Massachusetts, Amherst, MA.

[Webb, 1992a] Webb, G. (1992) Man-machine collaboration for knowledge acquisition. In *Proceedings of the 5th Australian joint conference on Artificial intelligence.* World Scientific, 329-334.

[Webb, 1992b] Webb, G. (1992) Learning disjunctive characteristic descriptions by least generalisation, *Technical report C92/9, Deakin University*, Geelong 3217

[Weiss and Kapouleas, 1989] Weiss, S.M. and Kapouleas, I. (1989) An empirical comparison of pattern recognition, neural nets and machine learning classification methods. In *Proceedings of the 11th international joint conference on artificial intelligence (IJCAI)* Detroit, MI : Morgan Kaufmann, 781-787.

[Weiss and Indurkhya, 1991] Weiss, S.M. and Indurkhya, N. (1991) Reduced complexity rule induction. In *Proceedings of the 12th international joint conference on artificial intelligence (IJCAI),* Sydney: Morgan Kaufmann, 678-684.