XVI IMEKO World Congress
Measurement - Supports Science - Improves Technology - Protects Environment ... and Provides Employment - Now and in the Future
Vienna, AUSTRIA, 2000, September 25-28

AUTOMATIC CONSTRUCTION OF A HIERARCHICAL CLASSIFIER

D. Filbert, F. Attia, R. Jahnke
Technische Universität Berlin, R. Bosch GmbH, PNA, Germany

Abstract: The automatic construction of a hierarchical classifier is described. The construction process uses the same classifier to select the features, which is used later for the classification itself. The construction leads to a binary decision tree. Every node is labelled with a feature vector.

The classical statistical approach to feature selection is presented. The add-on algorithm provides feature vectors of minimal length without regarding the classifier. Two applications are described and the results, reached by the classical statistical algorithm and the new hierarchical classifier, are compared.

Keywords: Decision tree, supervised learning, classification, feature evaluation.

1 INTRODUCTION

The classification of products in a production line is an important task of the quality control. The technical diagnosis is more and more used to decide whether the product has no faults or belongs to one of several fault classes.

Acoustic noise or mechanical vibrations are widely measured on the products. The signal processing consists of the measurement by a transducer (i.e. microphone or acceleration sensor), a pre-processing (i.e. filtering and suppression of noise), the feature extraction and the classification. The process of technical diagnosis provides an extremely strong compression of information. The measured signals containing a lot of relevant and irrelevant information are compressed into a few classes, only. The strongest compression takes place during the feature extraction process. Thus, this is the most important task. If relevant information is lost during this process it is impossible to reach good classification results. Therefore, the selection of the most efficient features is very important.

The paper describes the classical feature evaluation and –selection. The new classification is done by a decision tree algorithm called HICLASS (Hierarchical Classifier). Conventional decision trees as well as the tree at hand are presented. The paper finishes with some applications of the new classifier.

Problems of the feature extraction and classification are:

- The classes overlap in the feature space
- The no-fault class consists of a lot of samples
- The fault classes contain only a few samples of the training set
- The training set is selected by experts.
- The experts select features intuitively and in large numbers
- Features are correlated

An automatic construction of a diagnosis system by supervised learning consist of the finding of the structure of the classifier, the selection of the features and the calculation of its parameters (e.g. mean and covariance of a class). In case of supervised learning, described in this paper, the samples are sorted into classes. The samples for the classification are given in form of a training set X of n samples x_j

$$x_j \in X \quad j = 1,...,n.$$

A sample is defined by the feature vector

$$\underline{x}_j = \lfloor x_{j1},....,x_{jr};c_{ji} \rfloor$$

where $x_{j1},...,x_{jr}$ are the values of the features $x_1,...,x_r$ for the object $\underline{x}_j$. The class label c_{ji} is attached to this sample.

$$c_{ji} \in C \qquad i = 1,...,k$$

The training set X can also be interpreted as a set of n points within a multidimensional feature space. The label c_i describes the partitioning of the feature space into the classes.

The aim of the classification is to separate the samples of a test set with a minimum of cost and error rate. This can be achieved by an optimal selection of features and a choice of the most efficient classifier.

1.1 Conventional feature selection.

Feature extraction and classification should be seen as a unity but most design procedures are described separately in literature [1, 2]. The feature extraction is often optimised without regarding the classifier. Optimisation of the feature extraction is done by selecting the relevant features. But, usually features are selected intuitively and in large numbers by experts. A feature has to be judged by how well it separates classes from one another. The aim is to find a subset of features that separates the classes best. There are many measures of separability [3]. Such a measure should be related to the classification error. Most measures are based on Fisher's discriminant. Fisher's discriminant judges a feature depending on the distance of its centroids of the classes and the scatter within the classes. Fischer's discriminant can be used only for one feature and a two-class classification. In practice there will be more than two classes. A possible measure is the F-ratio

$$F = \frac{1/(k-1)\sum_{i=1}^{k}\left(m_i - \overline{m}\right)^2}{1/k(m-1)\sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_{ji} - m_i\right)^2} \qquad (1)$$

with k = number of classes; m = number of measurements for each class; x_{ji} = jth feature measurement for class i; μ_i = mean of all measurements of class i; $\overline{m}$ = mean of all measurements of all classes.

The divergence [4] has its roots in the information theory. If the covariance matrices W_i and W_l of two normal distributed classes are equal or can be averaged by W, then the divergence reduces to

$$D_{il} = tr\left[W^{-1}\left(\overline{x}_{ji} - \overline{x}_{jl}\right)\left(\overline{x}_{ji} - \overline{x}_{jl}\right)^T\right] \qquad (2)$$

In the case of more then two classes the divergence is averaged over all pairs of classes

$$D = tr\left[W^{-1}\left\langle\left(\overline{x}_{ji} - \overline{x}_{jl}\right)\left(\overline{x}_{ji} - \overline{x}_{jl}\right)^T\right\rangle\right] \qquad (3)$$

W is called the intraclass covariance matrix. The averaged $\left\langle\left(\overline{x}_{ji} - \overline{x}_{jl}\right)\left(\overline{x}_{ji} - \overline{x}_{jl}\right)^T\right\rangle$ is the interclass covariance matrix **B**. Thus, the separability measure is

$$D = tr\left(W^{-1}B\right) \qquad (4)$$

This measure is applicable for the evaluation of a group of features and several classes. Fukunaga [5] has listed other measures, as well.

1.2 Subset selection

The purpose of feature evaluation is to reduce the number of features in order to gain a set of relevant features containing the information to separate the measured objects into classes with a low classification error. A sure way to select the best subset is to evaluate every possible subset. This means examining a total of $\binom{N}{K}$ subsets, if N is the number of initial features which is to be reduced to a subset of K features. The expense of evaluation is not affordable if N is large and K is intermediate. A better approach is the add-on algorithm. The algorithm starts with the feature holding the highest F-ratio. All N-1 pairs of features comprising the nucleus and one other feature are evaluated using the separability measure D. This process is repeated until the measure does not increase substantially when another feature is added. The evaluation requires only K(2N+1-K)/2 operations. If a subset of K=10 has to be selected out of N=50 initial features, the permutation needs 10^{10} evaluations. The add-on algorithm needs 450 evaluations, only.

2 THE HIERARCHICAL CLASSIFIER

The design of decision trees for classification is a most promising approved approach for real-world applications [6,8]. A tree is defined as a finite set of one or more nodes such that

- there is a unique node designated the root
- the remaining nodes are partitioned into disjoint sets, each of which in turn is a tree called a subtree.

Generally, two types of information in a tree are important, the information about a node stored as a set of parameters of the features for classification (test) and the information relating the node to its neighbours (outcome of the test). The outcome is often a value of the feature. This has the advantage, that only simple limits of the features are necessary but the discriminant function can be lines or hyperplanes parallel to the co-ordinates, only. Decision trees exhibit the structure of the classification problem with high transparency to the user. But, the structure of the tree can not be adjusted by additional training sets. The construction of a decision tree is often described as a task of supervised learning. Classification learning means that a set of observed object is classified a priori. This is the training data set. The goal of the learning is that every object can be classified with a minimum error or costs. The automatic construction of the hierarchical classifier HICLASS unites the design of the structure of the classifier, the selection of the features and the calculation of its parameters, instead of a separate feature selection. The idea is to split the classification procedure into a chain of two-class tests. The outcome of the tests is a distance from the tested object to the class and the rest of the data set. The classifier in Fig. 2 is a binary decision tree. Each object to be classified passes from the root of the tree to a terminal node. The terminal nodes are labelled with the classes. The (k-1) non-terminal nodes are labelled with tests. They partition into a class and the rest of the data set. The information contained is the kind of the classifier (a different classifier can be applied in each non-terminal node), the features used for classification in that node, as well as the parameters (e.g. mean, variance). The hierarchical order of the nodes is very important as a wrong decision in a higher node can not be repaired by a lower one. To decide which features should be used at the root and which at the succeeding nodes several methods are described in literature [6,7]. The entropy measure method is used in the well known ID3 algorithm to evaluate discrete features [7]. For the construction of the decision tree at hand the class that can be separated with minimum error has to be determined. At first, an arbitrary class and an arbitrary feature of the training set are selected. The classification error separating that class from the rest of the training set is calculated. Let $n_{ji}<n_i$ be the number of objects reaching the current node and being correctly labelled with the class label c_i and n_i is the total number of objects of class c_i, then the estimated error rate is

$$e = 1 - \frac{n_{ji}}{n_i} \qquad (5)$$

The classification of the training set is carried out with all possible features in a sequential way. The feature leading to the minimum error is stored. A second arbitrary feature is selected by the add-on algorithm and the same classification is carried out using the sub-group of two features. Again, the group with the minimum error is stored. This procedure is continued until an optimal group out of all features or a pre-set number of features is determined. The data set of the next class is taken to continue the construction of the classification tree. After having tested all classes, the class with the lowest classification error is referred to the root node. The other classes are referred to their nodes depending on the classification error. When this construction has been finished the classification tree consists of nodes described by a number of features and their classification parameters depending on the training set. The number of features may vary from node to node. Certain features will appear at a single node or at several nodes.

After construction, the hierarchical classifier can be used for the quality control of products. Experience showed that the number of features used is smaller than found by the algorithm of chapter 2 and the grouping of features is more efficient, because the algorithm starts not by the feature with the highest F-Ratio but with that feature, that is able to partition the data set with a minimum classification error.

3 RESULTS

HICLASS was tested on 2 classification data sets and compared with classification results obtained by classical statistical methods. The first classification task is to construct a classical statistical classifier and HICLASS for the classification of the well known 'IRIS' data set [9]. The data set consists of 150 objects, described by 4 features. The 150 objects are labelled with 3 classes (50; 50; 50). A feature evaluation shows that feature No. 3, has the highest F-ratio. The add-on algorithm provides the grouping of the features of Table I. A simple Euclidean distance classifier was used. The distance classifier classifies the object depending on the distance of the object from the centroids of the classes. The distance from the object, described by its feature vector x_j to the centroid μ_i of class C_i is

$$\underline{d}_{ji} = \left(\underline{x}_j - \underline{m}_i\right)\cdot\left(\underline{x}_j - \underline{m}_i\right)^T \qquad (6)$$

The decision rule is: Decide for the class with minimum distance. The Euclidean distance classifier was also selected for all (k-1) non-terminal nodes of HICLASS. The classification success was calculated for reclassification. The training- and the test set were the same. The results are given in Table I.

Table 1. Classification- success for the ‘IRIS’ data set

STATISTICAL CLASSIFIER				HICLASS			
	Classification-Success %				Classification Success %		
Features	Class No. 1	Class No. 2	Class No. 3	Features	Class No. 1	Class No. 2	Class No. 3
3,1,4,2	**100**	**92**	**86**	4	**100**	**96**	**92**
3,1,4	**100**	**90**	**86**				
3,1	**100**	**86**	**80**				

The comparison of the classifiers shows that HICLASS is superior. Only 1 feature is used in all 3 nodes and the classification success is better. For the statistical classifier the add-on algorithm starts with feature No.3, although feature No.4 is better, referring to HICLASS. The reason is that the algorithm always starts with the feature with the highest F-ratio, which is feature No.3. The three groups of features of the statistical classifier show different results.

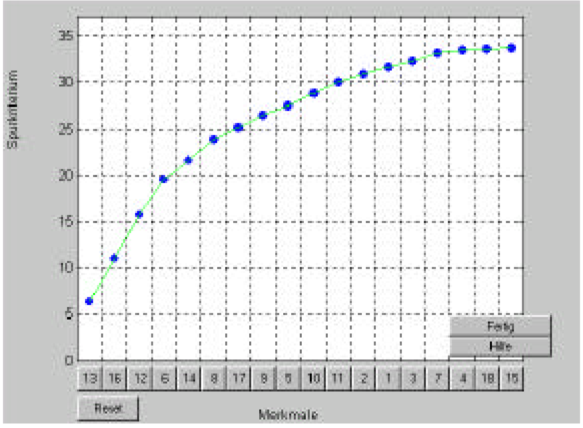

Fig. 1: Separability vs. feature-numbers

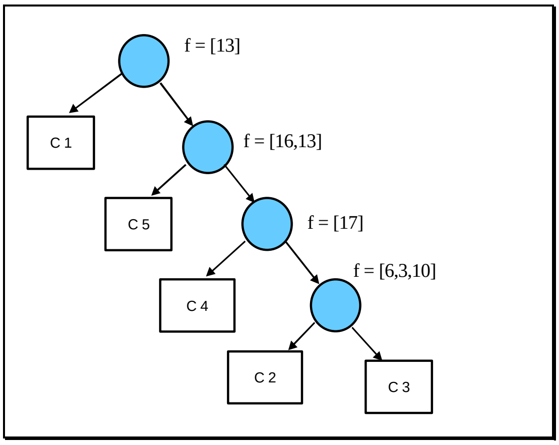

Fig. 2: Decision tree. The nodes are labelled with the feature numbers for classification.

HICLASS was also applied to problems of quality control. The data set measured from a drilling machine consists of 5 classes (number of objects in brackets): [no-fault (40), unusual noise (30), commutator fiering (50), gear fault (49), rotor unbalance (50)]. 18 features (spectral powers, RMS, excess, kurtosis, etc.) were calculated from the measured signals. The add-on algorithm delivers the function of Fig. 1. It is difficult to decide the number of features, as the separability increases with every additional feature. It was decided to use the features until no. 17 (13,16,12,6,14,8,17) for the statistical classifier. HICLASS constructed the decision tree of Fig. 2. Table 2 presents the classification success for the statistical classifier and HICLASS.

Table 2. Classification success for drilling machine data set.

STATISTICAL CLASSIFIER						HICLASS					
	Classification- Success in %						Classification Success in %				
Features	C 1	C 2	C 3	C 4	C 5	Features	C 1	C 2	C 3	C 4	C 5
All 18	**87,5**	**23,3**	**40**	**42,9**	**76**	13,16,17, 6,3,10	**95**	**100**	**92**	**81,6**	**94**
13,16,12, 6,14,8,17	**77,5**	**23,3**	**26**	**55,1**	**86**						

Again HICLASS is superior. As can be seen from Fig.2, the maximal number of features for classification is 6 out of 18 and the classification success is better in all 5 classes. Another important criterion relevant for quality control applications is the computational complexity. The computational complexity determines the speed of the test of the objects. The computational complexity can be expressed by the mean path length

$$L = \frac{1}{n} \sum_{i=1}^{k-1} n_i \cdot t_i \qquad (7)$$

with n = number of samples in the data set; k = number of classes, n_i = number of samples in the data set reaching node i, t_i = number of tests from root to node i. For the drilling machine data set the mean path length is 2.77, which together with the low number of features (i.e. fast tests) gives a good real-time performance. The cpu-time on a 200 MHz Pentium was 5 min. for the learning procedure and the cpu-time for the reclassification of the drilling machine data set was only 0.98 s. The classification performance will be even better as the no-fault class is connected to the highest node. That means, that most objects (>95% for quality control) will pass the first test, only.

4 CONCLUSION

The HICLASS classification tree construction algorithm offers a good and generalised classification with high transparency. As shown, only simple test (i.e. using a simple distance classifier) are necessary. Simple tests by simple classifiers means fast classification, which is very important in real-time classification for quality control in production lines. The automatic construction of the decision tree by HICLASS provides a binary tree with k terminal nodes and (k-1) non-terminal nodes. The add-on algorithm finds the class that can be separated from the rest of the training data set with minimum classification error. It also determines the minimum number of features, necessary for the tests. The classification success is fairly high and usually better than that reached by classical statistical methods. A disadvantage is that the learning procedure is time consuming. But, with modern computers this can be done within minutes and the time consuming learning does not affect the fast classification when HICLASS is in service.

The decision tree shows those classes, that can be separated best, in the higher hierarchy of the tree. In the lower hierarchy are those classes with difficult separation. Therefore it may be an advantage to use simple tests in the higher nodes and more complex tests in the lower nodes to gain better classification errors. The classification error was used as a measure to construct the decision tree by HICLASS. It is also possible to install other measures in the construction algorithm. The cost of a classification of a faulty object into the no-fault class may be rather high for tasks of quality control. It would be of practical interest to use a cost matrix for the selection of classes and features.

REFERENCES

[1] T.Parsons, *Voice and Speech Processing,* Mc Graw-Hill Book Company, New York, 1987.
[2] G. Meyer-Brötz, J. Schürmann, *Methoden der automatischen Zeichenerkennung,* Akademie Verlag, Berlin, 1970.
[3] H. Niemann, *Klassifikation von Mustern,* Springer Verlag, Berlin, 1983.
[4] S. Kullback, *Information Theory and Statistics,* Wiley, New York, 1959.
[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, New York, 1972.
[6] W. Müller, F. Wysotzki, The Decision Tree Algorithm CAL5 based on a Statistical Approach to ist Splitting Algorithm*, Machine Learning and Statistics*, Wiley, 1996
[7] P. Yoh-Han, *Adaptive Pattern Recognition and Neural Networks,* Addison-Wesley Publishing Co. Inc., 1989.
[8] H.I. Frohwein, J.H. Lambert, Y.Y. Haimes, Alternative measures of risk of extreme events in decision trees*, Reliability Engineering and System Safety*, Elsevier, 1999.
[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics, vol.7,*1936.

AUTHORS: Univ. Prof. Dr.-Ing. D. Filbert, Institut für Meß und Automatisierungstechnik, Technische Universität Berlin, Einsteinufer 17, D-10775 Berlin, Germany, Phone +49 30 31422541, e-mail: dieter.filbert@tu-berlin.de
Dr.-Ing. F. Attia, Dept. FV/FLP, Robert Bosch GmbH, PO Box: 10 60 50, D-70059 Stuttgart, Germany, Phone +49 711 0811 7046, e-mail: fawzi.attia@pcm.bosch.de
Dipl.-Ing. Ronny Jahnke, Polytec Noise Analysis GmbH, Am Hardtwald 3, D-76275 Ettlingen, Germany, Phone +49 7243 716152, e-mail: r.jahnke@pna.de