D. J. Cornforth, H. F. Jelinek, M. C. Teich, and S. B. Lowen, "Wrapper Subset Evaluation Facilitates the Automated Detection of Diabetes from Heart Rate Variability Measures," in Proc. International Conf. on Computational Intelligence for Modelling Control and Automation (CIMCA'04), edited by M. Mohammadian (University of Canberra, Australia, July 2004), pp. 446-455.

Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures

D. J. Cornforth[1], H. F. Jelinek[1], M. C. Teich[2] and S. B. Lowen[3]
[1]Charles Sturt University, Albury, Australia
E-mail: {dcornforth, hjelinek}@csu.edu.au
[2]Boston University
Boston, Massachusetts 02215, USA
E-mail: teich@bu.edu
[3]Brain Imaging Center, McLean Hospital,
Belmont, Massachusetts 02478, USA
E-mail: lowen@ mclean.org

Abstract

Diabetes affects almost one million Australians, and is associated with many other conditions such as vision loss, heart failure and stroke. Any improvement in early diagnosis would therefore represent a significant gain with respect to reducing the morbidity and mortality of the Australian population. In this study we apply signal processing and automated machine learning to analyse heart rate variability measures. These data are well suited to the diagnosis of cardiac dysfunction, but here we use the same measures to detect diabetes. By applying appropriate methods we were able to select the most relevant features to use as input to a variety of classifier algorithms. We compare sensitivity and specificity results obtained from these classifier algorithms. Results suggest that the detection of diabetes is feasible from heart rate variability measures.

Wrapper subset evaluation facilitates the automatic detection of diabetes from heart rate variability measures

Abstract

Diabetes affects almost one million Australians, and is associated with many other conditions such as vision loss, heart failure and stroke. Any improvement in early diagnosis would therefore represent a significant gain with respect to reducing the morbidity and mortality of the Australian population. In this study we apply signal processing and automated machine learning to analyse heart rate variability measures. These data are well suited to the diagnosis of cardiac dysfunction, but here we use the same measures to detect diabetes. By applying appropriate methods we were able to select the most relevant features to use as input to a variety of classifier algorithms. We compare sensitivity and specificity results obtained from these classifier algorithms. Results suggest that the detection of diabetes is feasible from heart rate variability measures.

1. Introduction

Both the NSW Department of Health and Commonwealth Government have identified diabetes to be a significant and growing global public health problem with the expected incidence in Australia to increase from 4% to 10% by 2010 [1]. In Australia approximately 1 million people are affected by diabetes and health care costs associated with treatment of complications amounts to approximately $7 billion dollars [2]. Vision loss, heart failure and stroke contribute significantly to the morbidity and mortality of the Australian population. Diseases of the circulatory system such as coronary heart disease and stroke were listed as the underlying cause of death in 55.7% of deaths in 2000 where diabetes was an associated cause [3]. In addition some form of nervous system damage such as cardiac autonomic neuropathy (CAN) that affects the function of the heart and blood vessels, occurs in up to 60-70% of people with diabetes [4].

Diabetes may lead to subtle changes in heart rate variability that become increasingly more compromising, manifesting in arrhythmia and cardiac failure in the majority of individuals with diabetes. Early detection of this pathology would allow timely intervention and thus lessen morbidity and mortality. This would improve the quality of life for people with diabetes and at risk for heart failure and stroke, as well as contributing to a substantial saving to the health care system [5].

To assess CAN the variation in the interval length of the ECG between successive beats can be analyzed in terms of heart rate variability (HRV). From this sequence of intervals, many secondary measures may be derived, making use of fractal and wavelet analysis techniques. It is well known as a tool for detecting cardiac dysfunction [6, 7].

Discrimination of various categories or classes (such as “cardiac dysfunction” or “normal”) is a well-studied class of machine learning problems. Here the key is to determine some relationship between a set of input vectors that represent stimuli, and a corresponding set of values on a nominal scale that represent category or class. The relationship is obtained by applying an algorithm to training samples that are 2-tuples $<\boldsymbol{u}, z>$, consisting of an input vector $\boldsymbol{u}$ and a class label z. The learned relationship can then be applied to instances of $\boldsymbol{u}$ not included in the training set, in order to discover the corresponding class label z [8]. A number of machine

learning techniques including genetic algorithms [9], and neural networks [10] have been shown to be very effective for solving such problems.

In evaluating the performance of any classifier, the accuracy is the most common measure used. In order to avoid bias in reporting this figure, it is necessary to report the accuracy on data not seen by the classifier during training. This requires splitting the available data into two sets, the training set and the holdout set. The classifier is trained on the first dataset then evaluated by its performance on the holdout set. Obviously the holdout set must be chosen carefully to prevent a source of bias here. The most popular way of testing involves the cross validation method [11] where the dataset is divided into a number of subsets. A number of tests are performed where each subset in turn becomes the holdout set. In this way the classifier is tested against all available data.

In order to apply these techniques to real problems, it is usual to obtain a number of measures, or features, which can form the input vector $\boldsymbol{u}$, and to obtain the corresponding class label z. In this case of medical diagnosis the class label is usually supplied after clinical evaluation by a specialist. It is to be expected that of the many measures available, some are better than other at discriminating between the classes. Methods for choosing the best feature set for detecting cardiac dysfunction have been demonstrated by Teich and co-workers [7]. However, in this study we are concerned with detecting diabetes with respect to the occurrence of CAN, which can precede the identification of hyperglycaemia by many years.

It is well known that using too many features can actually degrade accuracy of the prediction, so optimising the accuracy of such methods involves a choice not only of classifier algorithm, but also of the appropriate features. Kohavi [12] has studied the automatic selection of features and concluded:

- The optimum feature subset will depend on the classifier model chosen
- Therefore the subset may be considered a parameter of the model
- The evaluation of feature subsets will be biased in a favourable direction unless it uses independent data.

In addition, Kohavi suggests a wrapper approach where the actual classifier algorithm is used to evaluate the features selected, and perform a search for the set of features that maximises classifier accuracy.

2. Machine Learning Algorithms

A number of automated classifier algorithms are available using the excellent Weka toolbox [13]. These are briefly discussed below and include the Decision Table, Nearest Neighbours, Decision Tree Induction, Kernel Density and Naïve Bayes algorithms. We also used an implementation of the CMAC neural network, which is not included in the Weka toolbox.

The *Decision Table* algorithm divides the dataset into cells, where each cell contains identical records. A record with unknown class is assigned the majority, or most frequent, class represented in the cell. The goal of training is to find a minimum set of features that are optimal in predicting the class [14].

The *Nearest Neighbours* algorithm [15] simply stores samples. When an example is presented to the classifier, it looks for the nearest match from the examples in the training set, and labels the unknown example with the same class. In practice the algorithm looks at the nearest *k* neighbours, where *k* is a parameter set by the user. We have used this algorithm with *k=1* and *k=3*.

The *Decision Tree Induction* algorithm [16] uses the C4.5 algorithm to form a tree by splitting each variable and measuring the information gain provided. The split with the most information gain is chosen, and then the process is repeated until the information gain provided is below a threshold.

The *Kernel Density* algorithm [17] estimates a probability distribution of the data separately for each class. Each point is represented by a kernel function, and the kernels for all points in the class are summed to provide a composite function. An unknown point is evaluated by each composite function separately, and the class function corresponding to the highest probability is chosen.

The *Naïve Bayes* algorithm [18] assumes that features are independent. From the correlation analysis above, we know this is untrue, but the algorithm performs surprisingly well. It estimates prior probabilities by calculating simple frequencies of the occurrence of each feature value given each class, then returns a probability of each class, given an unclassified set of features.

The *Cerebellar Model Articulation Controller* (CMAC) [19] with the *Kernel Addition Training Algorithm* [20] divides the input space using multiple overlapping grids and builds a probability density function for each class. An unknown input then can be associated with an appropriate probability value for each class. The input is assigned to the class having the largest probability.

Included in the Weka toolbox is a Wrapper Subset evaluator. This takes as a parameter the name of the class being used for the discriminant function. The wrapper does a search in the list of features for the set that gives the lowest error on the given classifier.

3. Methods

The aim of these experiments is to determine whether it is possible to make any predictive model from these data. If it is possible, this is evidence that the disease is related to changes in the heart rate parameters measured.

We used heart rate variability data from a joint Israeli-Danish study [21], which consists of 46 adult people, 22 with known diabetes and 24 controls, which had not been diagnosed with diabetes. The dataset provides the measurements shown in table 1.

Table 1. Features used from the heart rate variability dataset.

Feature number	Details
1 to 13	Allan Factor using time windows of 1, 1.5, 2.2, 3.2, 4.7, 6.80, 10, 15, 22, 32, 47, 68, and 100 sec [Allan, 1966; Barnes and Allan, 1966].
14 to 21	The Discrete Haar Wavelet Transform using scales of 2, 4, 8, 16, 32, 64, 128, and 256.
22	Wavelet transform power law exponent
23	The Mean of heartbeat counts.
24	The average heart rate (not period)
25 to 29	Other standard measures of distribution used are the Variance, Standard deviation, Coefficient of variation, Skewness, and Kurtosis.
30	SerialCC – Autocorrelation coefficient of sequence of R-R intervals.

These features have been evaluated for their efficacy in discriminating between cardiac dysfunction and normal controls [7], but not as a diagnostic aid for diabetes. The complex interaction between these variables makes it difficult to choose the correct feature subset to use by manual inspection. We therefore conducted a search both for the optimum classifier algorithm and for the optimum feature subset for that model. We used an outside cross validation loop to eliminate the bias. This was implemented as follows:

- Create 46 data sets, each one excluding one data record.
- Perform wrapper subset evaluation on each dataset, for each classifier algorithm using 5-fold cross validation on the remaining 45 records. Repeat 10 times
- For each classifier and exclusion set, choose the most common set of features. Where more than one has equal frequency, pick the fewest number of features. Break any further ties at random.
- Test each classifier by training on 45 records with the chosen feature set, then testing on a single holdout record.
- Evaluate each classifier on the number of records correctly classified out of 46 trials.

We repeated steps 4 and 5 after choosing the most popular feature subset for each classifier, and again after choosing the most popular feature subset from the entire experiment.

4. Results

After the wrapper feature selection, all classifiers except CMAC found the same feature set given the same dataset. There was great variation in the feature subset chosen for different dataset, for all classifiers. The optimum feature subsets found are summarised in table 2, for the first dataset (excluding record #1) and the second dataset (excluding record #2). For the CmacKata algorithm, the feature set was different for each run. For the first dataset the best feature sets indicated were 19, 17, 17, 17, 17, 19, 17, 17, and 17. The most frequent feature set {17} was selected. For the second dataset, the feature set indicated was always {15, 22}.

Table 2. Optimum feature subsets for each classifier (only datasets 1 and 2 shown).

Algorithm	Feature set for dataset 1	Feature set for dataset 2
CmacKata	17	15,22
Decision Table	28	28
IB1	19	17
IBk=3	14,15,16,19,28	14,15,19,28
j48.J48	7,9,28	3,7,9,24,28
Kernel Density	10,12,16,18,20,26,27,28	8,9,21,28
Naïve Bayes	11,16,20,22,26,28	16,20,25

The result of classification using these feature sets is shown in table 3. For most algorithms, the accuracy is very poor considering that a random choice would be expected to assign approximately half of all records (23) to the correct class. The 3-Nearest Neighbours method (IBk=3) is the only classifier that provided a better classification than a random guess. The average number of correctly assigned records for all classifiers is 21, which is not a encouraging result.

Table 3. Results of testing classifiers using feature subsets chosen for each classifier and each dataset using the wrapper method.

Classifier	truePos	falsePos	trueNeg	falseNeg	correct	sensitivity	specificity
CmacKata	5	15	9	17	14	0.227273	0.375
Decision Table	0	1	23	22	23	0	0.958333
IB1	11	16	8	11	19	0.5	0.333333
IBk=3	11	10	14	11	25	0.5	0.583333
j48.J48	7	8	16	15	23	0.318182	0.666667
Kernel Density	8	13	11	14	19	0.363636	0.458333
Naïve Bayes	13	14	10	9	23	0.590909	0.416667

In the second test, we combined feature sets so that one set was used for each classifier. We did this by selecting the most frequently indicated feature set for each classifier, from the results of the wrapper evaluation used in the first test. These sets are shown in table 4, and the results of classification using these sets are given in table 5. There results appear far better, with the majority of the classifiers performing better than a random choice. The average number of correctly assigned records for all classifiers is 30. The 3-Nearest Neighbours classifier (IBk=3) and the Kernel Density classifiers both achieve a relatively high accuracy, correctly assigning 35 out of 46 records to the correct class. Of the other classifiers, most achieve a result better than a random guess. It is interesting to note that the feature sets selected for the two most successful classifiers are very different. This tends to support the findings of Kohavi [12], i.e. that the feature set is best considered a part of the classifier algorithm chosen, and that it is unlikely that an feature set can be chosen that will be optimum for all classifiers.

Table 4. Optimum feature subsets for each classifier, as used for all datasets.

Algorithm	Feature set
CmacKata	15,22
Decision Table	28
IB1	6,10,12,17
IBk=3	13,14,15,16,27,28,30
j48.J48	28
Kernel Density	8,9,21,28
Naive Bayes	16,20,26

Table 5. Results of testing classifiers using feature subsets chosen for each classifier

Classifier	truePos	falsePos	trueNeg	falseNeg	correct	sensitivity	specificity
CmacKata	17	11	13	5	30	0.772727	0.541667
DecisionTable	0	1	23	22	23	0	0.958333
IB1	14	9	15	8	29	0.636364	0.625
IBk=3	16	5	19	6	35	0.727273	0.791667
j48.J48	7	5	19	15	26	0.318182	0.791667
KernelDensity	16	5	19	6	35	0.727273	0.791667
NaiveBayes	16	9	15	6	31	0.727273	0.625

In our third test, we chose the most frequent feature set out of all classifiers and all datasets. This was feature number 28, Skewness. Fig. 1 illustrates the frequency with which each feature was selected across all wrapper evaluation tests. This clearly shows that Skewness was the most frequent overall. However, Fig. 2 illustrates that classes are well interlocked when considering this feature alone, so it is obvious that no single feature would enable accurate classification of the two classes.

The results of the accuracy evaluations are given in table 6. Although most of the classifier algorithms are more successful than a random guess, it is obvious from the result that this has not been as successful as the previous test. The most successful classifier for this test appears to be 3-nearest neighbours.

Table 6. Results of testing classifiers using the same feature subset for every classifier (Skewness).

Classifier	truePos	falsePos	trueNeg	falseNeg	correct	sensitivity	specificity
CmacKata	13	11	13	9	26	0.590909	0.541667
Decision Table	0	1	23	22	23	0	0.958333
IB1	10	13	11	12	21	0.454545	0.458333
IBk=3	13	6	18	9	31	0.590909	0.75
j48.J48	7	5	19	15	26	0.318182	0.791667
Kernel Density	6	2	22	16	28	0.272727	0.916667
Naïve Bayes	8	4	20	14	28	0.363636	0.833333

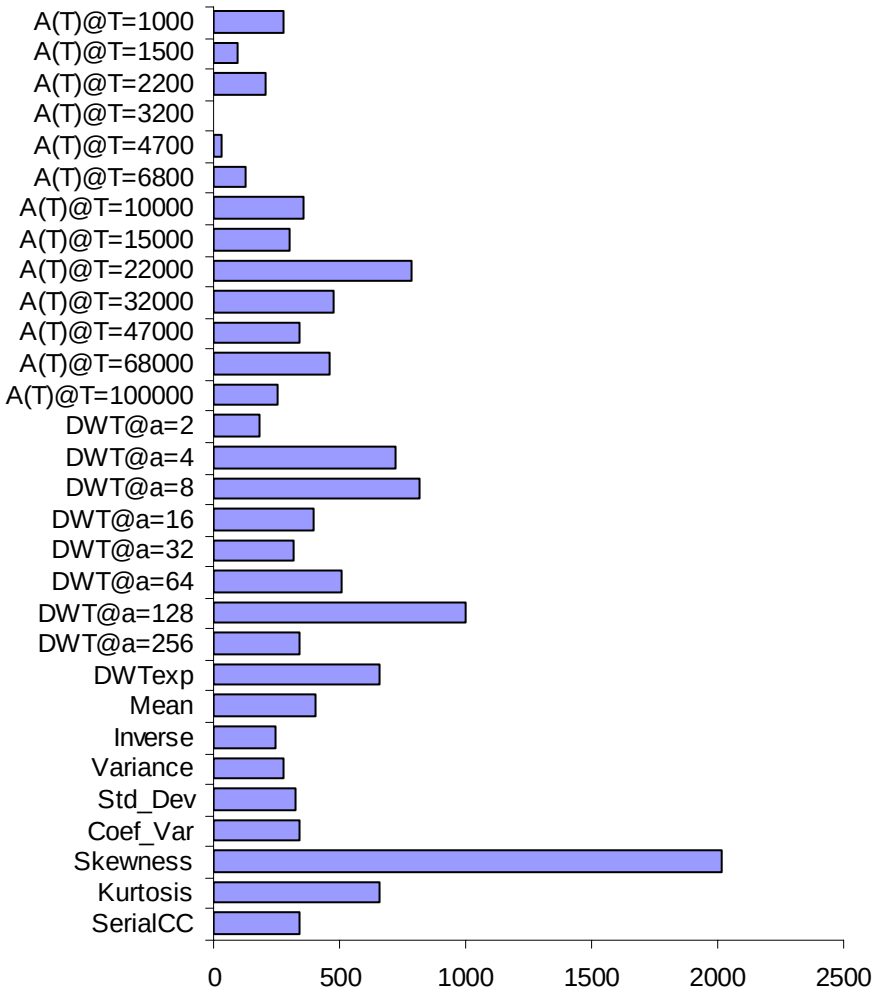

Fig. 1. Examination of the frequency distribution of features chosen from the dataset, for all classifiers, suggest that feature #28, the Skewness of distribution of interval sizes, has some special property in distinguishing the classes.

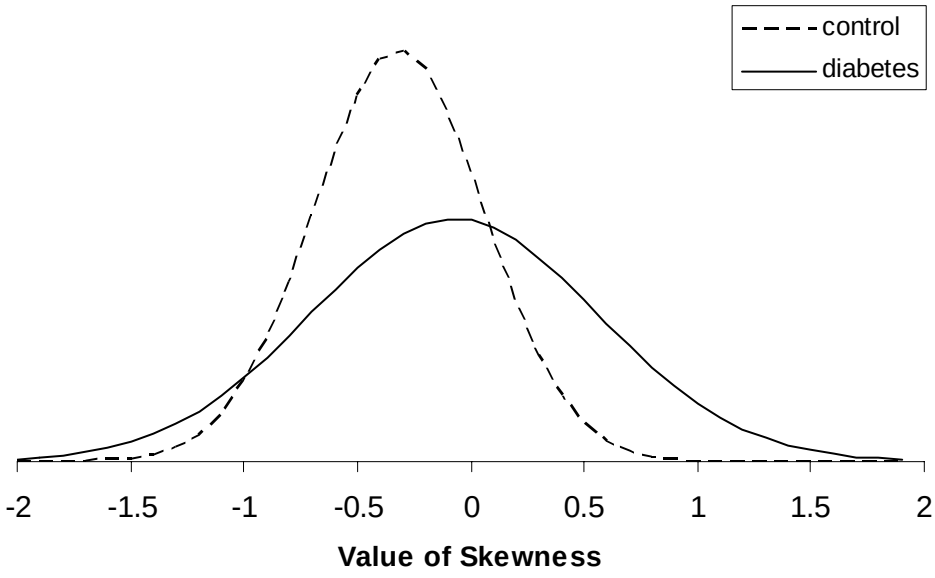

Fig. 2. Normal distributions for the two classes shown using the most popular feature, Skewness.

5. Conclusions

Based on these preliminary results, we suggest the following conclusions:

- It appears that diabetes could be detected from an analysis of heart rate variability, but further study is obviously required.
- Diagnosis of diabetes by this method is not 100% accurate, and depends upon careful selection both of the classifier algorithm used, and the feature subset.
- Selection of a feature set for a particular classifier appears to be more successful than selection of a separate feature set for each classifier and each dataset, or selection of a common feature set for all classifiers.

We have demonstrated the feasibility of detecting diabetes from heart rate recordings with possible Cardiac Autonomic Neuropathy (CAN). The most successful classifier algorithms tested were the 3-Nearest Neighbours and the Kernel Density, achieving a sensitivity of 0.727273 and a specificity of 0.791667. We have been careful to eliminate bias and over fitting in our evaluation of success, to provide some confidence in the results. The results are quite promising despite the small dataset available (46 records). The relative speed with which the analysis presented here can be performed on a modest computer suggests that this technique could be of benefit in screening, especially in rural regions where specialist medical personnel are less accessible than in urban areas.

Both false negatives and false positives have merit here, so the results may be even better than indicated in table 5. We suggest that some false positives are actually controls with CAN and some false negatives are diabetics with no CAN. This needs to be clarified with additional tests in future work. With care, the error structure could be shifted towards false positives, which would allow marginal cases to be referred for further investigation.

The success of classifier algorithms using a separate feature set for each classifier supports the findings of Kohavi [12], by showing that the feature set chosen should be regarded as part of the classifier algorithm. From the results presented, the best feature is the Skewness of heart rate intervals, although it is clear that this feature alone cannot provide any useful discrimination of diabetics from controls. It is necessary to use a set of features suited to the chosen classifier.

The success in discriminating diabetes from normal controls in heart rate variability data suggests a methodology that would provide a very simple, cheap and quick test that could be performed in a rural health clinic, and if implemented, would bring great benefits to the rural community in terms of early diagnosis and consequently a reduction in hospitalization and length of stay.

References

[1] Colagiuri, S., Colagiuri, R. and Ward, J., 1998, *National diabetes strategy and implementation* plan, Diabetes Australia, Canberra.

[2] Mathers, C., T. Vos, and C. Stevenson, 1999, *The burden of disease and injury in Australia*, Australian Institute of Health and Welfare: Canberra.

[3] AIHW, 2002, Diabetes: *Australian Facts 2002*, Canberra: Australian Institute of Health and Welfare, ISBN 1-74024-198-3, accessed at http://www.aihw.gov.au/publications/cvd/daf02/.

[4] Rathmann, W., et al., 1993, Mortality in diabetic patients with cardiovascular autonomic neuropathy, *Diabetic Medicine*, 10(9): 820-824.
[5] Pagani, M., 2000, Heart rate variability and autonomic diabetic neuropathy, *Diabetes, Nutrition and Metabolism*, 13(6): 341-346.
[6] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, P. J. Schwartz, and the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology,1996, Heart rate variability - standards of measurement, physiological interpretation, and clinical use, *Euro. Heart J*., 17: 354-381. Also published in *Circulation*, 93: 1043-1065.
[7] M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer, and C. Heneghan, 2001, Heart Rate Variability: Measures and Models, in M. Akay (ed.), *Nonlinear Biomedical Signal Processing, Vol. II, Dynamic Analysis and Modelling*, IEEE Press, New York.
[8] Dietterich, T.G. and Bakiri, G., 1995, Solving Multiclass Learning Problems Via Error-Correcting Output Codes, *Journal of Artificial Intelligence Research*, 2: 263-286.
[9] Holland, J., 1992, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, second edition.
[10] Duda, R.O and Hart, P.E., 1973, *Pattern Classification and Scene Analysis*, New York: John Wiley and sons.
[11] B. Efron, 1983, Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association*, 78 (382): 316-330.
[12] Kohavi, R., and John G., 1996, Wrappers for Feature Subset Selection. In *Artificial Intelligence Journal*, special issue on relevance, 97, (1-2): 273-324.
[13] I.H. Witten and E. Frank, 1999, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.
[14] R. Kohavi, 1995, The Power of Decision Tables, in: *Proceedings of the European Conference on Machine Learning, Lecture Notes in Artificial Intelligence*, 914: 174-189, Springer Verlag.
[15] Fisher, R.A., 1936, The use of multiple measurements in taxonomic problems, *Annual Eugenics* 7 (part II):179-188. Reprinted in Contributions to Mathematical statistics, 1950, Wiley.
[16] J.R. Quinlan, 1986, Induction of Decision Trees, *Machine Learning* 1 (1): 81-106.
[17] B.W. Silverman, 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
[18] T. Bayes, 1763, An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, 53: 370-418.
[19] J.S. Albus, 1975, A New Approach to Manipulator Control: the Cerebellar Model Articulation Controller (CMAC), *Journal of Dynamic Systems, Measurement and Control* 97: 220-233.
[20] D. Cornforth and D. Newth, 2001, The Kernel Addition Training Algorithm: Faster Training for CMAC Based Neural Networks, in: *Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES 2001)*, Otago, New Zealand.
[21] Y. Ashkenazy, M. Lewkowicz, J. Levitan, H. Moelgaard, P. E. Bloch Thomsen, and K. Saermark, 1998, Discrimination of the healthy and sick cardiac autonomic nervous system by a new wavelet analysis of heartbeat intervals, *Fractals*, 6: 197-203.