A Fuzzy Classifier Based on Modified Particle Swarm Optimization for Diabetes Disease Diagnosis

Hamid Reza Sahebi[1], **Sara Ebrahimi**[2]

[1] **Department of Mathematics, Ashtian Branch, Islamic Azad University**
Ashtian, Iran
sahebi@mail.aiau.ac.ir

[2] **Department of Mathematics, Ashtian Branch, Islamic Azad University**
Ashtian, Iran
ebrahimi@mail.aiau.ac.ir

Abstract

Classification systems have been widely utilized in medical domain to explore patient's data and extract a predictive model. This model helps physicians to improve their prognosis, diagnosis or treatment planning procedures. Diabetes disease diagnosis via proper interpretation of the diabetes data is an important classification problem. Most methods of classification either ignore feature analysis or do it in a separate phase, offline prior to the main classification task. In this paper a novel fuzzy classifier for diagnosis of diabetes disease along with feature selection is proposed. The aim of this paper is to use a modified particle swarm optimization algorithm to extract a set of fuzzy rules for diagnosis of diabetes disease. The performances of the proposed method are evaluated through classification rate, sensitivity and specificity values using 10-fold cross-validation method. The obtained classification accuracy is 85.19% which reveals that proposed method, outperforms several famous and recent methods in classification accuracy for diabetes disease diagnosis.

Keywords: *Diabetes disease diagnosis; Particle swarm optimization; Fuzzy classifier.*

1. Introduction

Diabetes is a major health problem in both industrial and developing countries, and its incidence is rising. It is a metabolic diseases characterized by high blood glucose levels, which result from body does not produce enough insulin or the body is resistant to the effects of insulin, named silent killer [1]. The body needs insulin to use sugar, fat and protein from the diet for energy. Diabetes increases the risks of developing kidney disease, blindness, nerve damage, blood vessel damage and it contributes to heart disease [2]. Early detection of diabetes is important to increase the chance of successful treatment. Such detection is often formulated as a binary classification problem[3]. Classification is a method of supervised learning producing a mapping from a feature space onto classes encountered in the classification problem. Classification problems are encountered in various domains including medicine [4], economics [5], and fault detection[6], etc. In order to improve classification performance, a large number of methods have been developed. In the classification task, the aim is assigning the patterns (case, record, or instance) to related classes, out of a set of predefined classes, based on the values of some attributes (called predictor attributes) for the patterns. There are various important categories of classification techniques including statistical techniques, neural networks, and rule based classification techniques. In recent times, many machine learning techniques have been considered to design automatic diagnosis system for diabetes. This paper specifically focuses on the use of fuzzy modeling method to detect diabetes disease which relies on discovering human comprehensible knowledge. Fuzzy approaches have become one of the well-known solutions for the classification problems. Fuzzy logic [7] improves classification and decision support systems by their allowing the use of overlapping class definitions and their powerful capabilities to handle uncertainty and vagueness. Fuzzy systems present two main advantages. First, these systems allow researchers to work with imprecise data and provide a comfortable approach to represent missing values. Second, these systems possess an interpretable rule-based structure. The performance of fuzzy classifier system depends on the "if-then" rules and their numbers that are generated from numerical data or human experiences. In the literature, several methods have been proposed to for building optimal fuzzy classifiers. In recent years Evolutionary Algorithms (EAs) have been widely used to optimize fuzzy classifiers. In the literature several EAs like Genetic Algorithm (GA), particle swarm optimization (PSO), ant colony optimization (ACO) and Artificial Bee Colony (ABC) have been proposed to produce fuzzy classification system. Rani and Deepa [8] proposed a particle swarm optimization approach for the optimal design of the fuzzy system for the classification task. In the proposed approach, both rule base and the membership functions are evolved simultaneously with the objective of maximizing the correctly classified class and minimizing the number of rules. In this paper, some modifications of a standard particle swarm optimization are introduced. This modified version of PSO is proposed as a new tool for building a compact fuzzy rule based classifier without any a priori knowledge. The proposed method has been evaluated using the public Pima Indian Diabetes (PID) dataset which is available at the University of California, Irvine web site [9]. The results indicate that this method can detect the diabetes disease signature with competitive or even better accuracy than the results achieved by earlier works. Also this method has considerable comprehensibility, since it produces a few number of short length fuzzy if-then rules. Classification systems have been used for Pima Indian diabetes disease diagnosis problem as for other clinical diagnosis problems. There have been a lot of studies reported in the literature in which the researchers have used the Pima Indian diabetes data set to evaluate their works. These studies applied different methods to the given problem and achieved high classification accuracies. Peng et al [10] applied data gravitation based classification algorithm and obtained 76.56 classification accuracy. In [11] a generalized discriminates analysis and least square support vector machine was used for diagnosing Pima Indian diabetes. In this work the authors have reported 79.16% classification accuracy. In Temurtas et al. [12] a multi-layer neural network and a Probabilistic Neural Network (PNN) were used for diagnosing Pima Indian diabetes. They have reported respectively 79.62% and 78.05% in terms of correct classification rate. Kayaer and Yıldırım [13] achieved 80.21% classification accuracy using general regression neural network (GRNN) for diagnosing Pima Indian diabetes. They have also reported 77.08% classification accuracy using multilayer neural network with LM algorithm. They have used conventional (one training and one test) validation method. Leon and IV [14] used a Fuzzy Neural Network (FNN) and they got 81.8% accuracy. Beloufa and Chikh [15] proposed an efficient and reliable modified ABC algorithm for diabetes disease. The proposed modified ABC has been used as an evolutionary algorithm to create an optimal fuzzy classifier. This method obtained 84.21 % classification accuracy. Kahramanli and Allahverdi [16] have used a hybrid of artificial neural network (ANN) and fuzzy neural network (FNN) and they have obtained 84.2 classification accuracy. Ganji and Abadeh [17] proposed a classification algorithm called FCSANTMINER which was obtained by combination of Ant Colony Optimization and Fuzzy Logic for diabetes disease diagnosis. The obtained classification accuracy is 84.24%.

2. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart in 1995 [18, 19]. PSO is motivated by social behaviors such as bird flocking and fish schooling. In PSO, a population, also called a swarm, of candidate solutions are encoded as particles in the search space. PSO starts with the random initialization of a population of particles. Particles move in the search space to search for the optimal solution by updating the position of each particle based on the experience of its own and its neighboring particles.

PSO is based on the principle that each solution can be represented as a particle in the swarm. Each particle has a position in the search space which is represented by a vector $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$, where D is the dimensionality of the search space. Particles move in the search space to search for the optimal solutions. Therefore, each particle has a velocity, which is represented as $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$, during the movement, each particle updates its position and velocity according to its own experience and that of its neighbors. The best previous position of the particle is recorded as the personal best *pbest*, and the best position obtained by the population thus far is called *gbest*. Based on *pbest* and *gbest*, PSO searches for the optimal solutions by updating the velocity and the position of each particle according to the following equations:

$$x = x + v \tag{1}$$

$$v = v\omega + c_1 r_1 (p_{best} - x) + c_2 r_2 (g_{best} - x) \tag{2}$$

Where ω is inertia weight, which is to control the impact of the previous velocities on the current velocity. c_1 and c_2 are acceleration constants. r_1 and r_2 are random values uniformly distributed in [0,1]. The velocity is limited by a predefined maximum velocity, $v_{\max}$ and $v \in [-v_{\max}, v_{\max]}$. The algorithm stops when a predefined criterion is met, which could be a good fitness value or a predefined maximum number of iterations.

The BPSO initialized the positions and velocities of the particle swarm randomly by

$$x_{ij} = \begin{cases} 1 & r > 0.5 \\ 0 & otherwise \end{cases} \tag{3}$$

$$v_{ij} = -v_{\max} + 2rv_{\max} \tag{4}$$

where r is random value uniformly distributed in [0,1]. Also, the positions x_{ij} for each variable was calculated by

$$x_{ij} = \begin{cases} 1 & S(v_{ij}) > r \\ 0 & otherwise \end{cases} \qquad (5)$$

here $S(.)$ represented the logistic function, and it served as the probability distribution for the position x_{ij}.

$$S(x_{ij}) = \frac{1}{1+\exp(-x_{ij})} \qquad (6)$$

The velocities v_{ij} are iteratively updated by the aforementioned formula (2). In order to improve the diversity of BPSO without compromising with the solution quality, in this paper, we introduced the mutation operator, which could further explore untried areas of the search space by

$$x_{ij} = \begin{cases} \sim x_{ij} & r < r_{mut} \\ x_{ij} & otherwise \end{cases} \qquad (7)$$

Where r_{mut} stands for the probability of random mutation. After updating the positions properties in (5), each bits of the solution candidates were mutated with a probability r_{mut}. It was common to set r_{mut}, which indicated that it was expected one bits in each solution candidate would be flipped. This BPSO with mutation was abbreviated as MBPSO.

3. Proposed method

Proposed method operates in two main stages: training stage and testing stage. In training stage, at first, the modified PSO algorithm is applied to generate a set of fuzzy rules via training patterns. A fuzzy classifier consists of linguistic rules which are easy to interpret by the user. The classification problem consists of m training patterns $x_p = (x_{p1}, x_{p2}, ..., x_{pn})$; $p = 1,2,...,m$ from M classes where x_{pi} is the i th attribute value $(i = 1,2,...,n)$ of the p th training pattern. These fuzzy rules are displayed as the following form:

$$R_j = if\ x_{p1} = A_{j1}\ and\ x_{p2} = A_{j2}\ and \dots and\ x_{pn} = A_{jn}$$
$$Then\ x_p\ belong\ to\ class\ C \qquad (8)$$

where R_j is the label of the j th rule, $x = (x_1, x_2, ..., x_n)$ is an n dimensional pattern vector, A_{ij} is an antecedent fuzzy set and $C(c = 1,2,...,M)$ is a class label. We use the fuzzy reasoning method of the winning rule. Each particle of a population in the PSO based method is represented in order to determine a fuzzy classification system. The codification of a fuzzy classification system designed to be evolved by modified PSO algorithm is given in Fig. 1. This solution comprises two types of parameter vectors $S_i = [R_i, f_i]$. One is parameter $f_i = (f_{i1}, ..., f_{ik}, ..., f_{in})$ represented by n binary bits for the selection of features. The other parameter vector $R_i = (R_{i1}, ..., R_{ij}, ..., R_{iMx})$ consists of control vector $mf^i_{jk} \in \{1,2\}$ for specifying the type of membership function (low, high) defined for A_{ij}, and parametric vector. The parametric vector encode four real-valued parameters $p^i_{jk,1}, p^i_{jk,2}, p^i_{jk,3}, p^i_{jk,4}$ that determine the points of a membership functions.

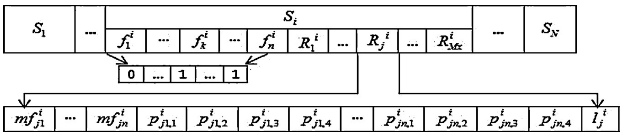

Fig. 1. The coding structure of solution

In the same way as other evolutionary algorithms, we need a function for evaluating the generalized solutions called fitness function. The proposed fitness function takes into account two terms: in the first one, classification rate which measures the classification performance encoded by an individual. The second term takes into account the number of rules so as to favor solutions containing a smaller number of rules. Evaluation function for fuzzy classification system is defined as follows:

$$Fitness(S_i) = c^2(S)(L - Num_R(S_i) + 1) \qquad (9)$$

Where S_i is the ith solution, $c^2(S_i)$ is the classification rate, L is the number of rules generated in the previous stage, and $Num_R(S_i)$ is the number of active rules. The detail of the proposed method is shown in Algorithm 1. Also the flowchart of the proposed method is shown in Figure 2.

Algorithm. BPSO for diabetes diagnosis

```
Input    $D_{train}$ : Train Dataset
         $D_{test}$  : Test Dataset
```

```
        I: Number of Iteration
        N: Number of population
        $M_x$ :Number of rules
        $r_{mut}$: Mutation rate

Output  Fuzzy Classifier

Begin algorithm
Initialize the population of solutions $S_i$, $i = 1,2,..., N$ randomly.
For each iteration do
    For each particle do
Evaluate the fitness function $Fitness\ (S_i)$ using Eq. (9).
    If the fitness value is better than pbest then
            Set current value as the new pbest
        If pbest is better than gbest then
            Set gbest = pbest
    End for
    For each particle do
        Update particle velocity using Eq. (2)
        Update particle position using Eq. (1)
    End for
    For each particle do
        Mutate particle using Eq. (7)
    End for
End for
Build Fuzzy classifier using solution gbest
End algorithm
```

Algorithm 1: Pseudo code of proposed method

4. Experimental Results

In this section, we first explain the diabetes disease database used in our experiments. Then we present the performance evaluation methods used to evaluate the proposed method. Finally, we give the experimental results and discuss our observations from the obtained results.

4.1 Data source

This paper studies Pima Indian Diabetes dataset (PID), which current data mining approaches have often used for their analyses. This data set was obtained from the UCI Repository of Machine Learning Databases[9]. The reason for using this dataset is that because it is very commonly used among the other classification systems that we have used to compare this study with for Pima Indian diabetes diagnosis problem. The dataset which consists of Pima Indian diabetes disease measurements contains two classes and 768 samples. The class distribution is
Class 1: normal (500).
Class 2: Pima Indian diabetes (268).
All samples have eight features. These features are:
Feature 1: Number of times pregnant.
Feature 2: Plasma glucose concentration a 2 h in an oral glucose tolerance test.
Feature 3: Diastolic blood pressure (mm Hg).
Feature 4: Triceps skin fold thickness (mm).
Feature 5: 2-h serum insulin (lU/ml).
Feature 6: Body mass index (weight in kg/ (height in m) ^2)).
Feature 7: Diabetes pedigree function feature 8: 2-h serum insulin (lU/ml).
Feature 8: Age (years).

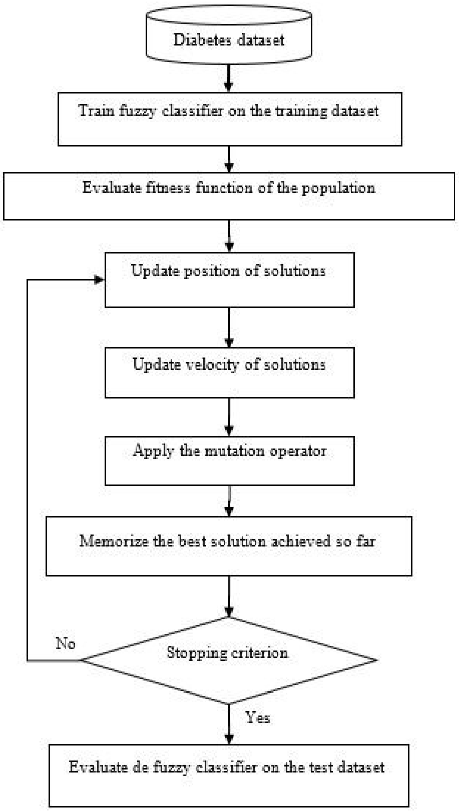
Diabetes dataset
Train fuzzy classifier on the training dataset
Evaluate fitness function of the population
Update position of solutions
Update velocity of solutions
Apply the mutation operator
Memorize the best solution achieved so far
No
Stopping criterion
Yes
Evaluate de fuzzy classifier on the test dataset

Fig 2. The flowchart of proposed method

4.2 Performance evaluation methods

We have used three methods for performance evaluation of diabetes disease diagnosis. These methods are sensitivity (SE), specificity (SP) and classification rate. The respective definitions of these are as follows:

$$classification\ rate = \frac{TP+TN}{TP+TN+FP+FN} \qquad (10)$$

$$sensivity = \frac{TP}{TP+FN} \qquad (11)$$

$$specificity = \frac{TN}{TN+FP} \qquad (12)$$

where TP, TN, FP and FN denote respectively:

- True Positives (TP): Classifies Diabetic as Diabetic.
- True Negatives (TN): Classifies No Diabetic as No Diabetic.
- False Positives (FP): Classifies No Diabetic as Diabetic.
- False Negatives (FN): Classifies Diabetic as No Diabetic

The performance of proposed method is evaluated using 10-fold cross-validation test. In this section, the obtained results are reported. The results with all features and optimal features found through modified PSO search in terms of Classification Rate, number of rules and number of features for all folds is shown in Table 1.

From the Table 1 we can see that fuzzy rules selected have relatively small number with an average of 12.1 rules (without feature selection). The number of features and rules used by different folds when performing feature selection varies significantly with an average 7.6 rules.

Table 1. Classification rate, number of rules and CPU time of the proposed method.

All features			**With feature selection**		
Classification rate	#Rules	CPU time (s)	Classification rate	#Rules	CPU time (s)
83.11	11	115	83.11	8	152
81.81	14	108	85.71	9	158
85.71	12	104	84.41	7	141
84.41	11	119	87.01	6	139
83.11	13	128	84.41	7	157
79.22	12	115	85.71	8	137
81.81	12	112	87.01	10	134
84.41	11	121	83.11	7	138
83.11	13	109	84.41	6	142
85.71	12	118	87.01	8	138
83.24	12.1	114.9	85.19	7.6	143.6

To compare the performance of the proposed modified PSO, the basic PSO algorithm is developed and the performance comparison is given in Table 1. In our experiments, we used the same population number and the maximum evaluation number for each test problem. As can be seen from the Table 2, the best results were obtained with the modified version of PSO with respect to the mean and maximum classification results. The best result was 86.01 % and the mean result was 83.11 %.

Table 2. The results of comparison

	Basic PSO			Modified PSO		
	Best	Worst	Mean	Best	Worst	Mean
Classification Rate (%)	81.81	76.62	78.54	86.01	83.11	85.19
Sensitivity (%)	79.54	72.87	77.39	85.96	82.84	84.29
Specificity (%)	82.76	77.06	79.95	87.95	84.56	86.39

The convergence performance for the modified ABC [15] and the modified PSO is also studied. This is shown in Fig. 4 using feature selection and all features. From this figure, it is observed that modified ABC algorithm takes more number of generations to converge than modified ABC, the modified ABC is unable to attain convergence before the 140th generations. Even though the proposed modified PSO takes less generations (at 135th generations) the blended mutation operator incorporated into them performs well in tuning the solution and classification rate is very good than the modified ABC.

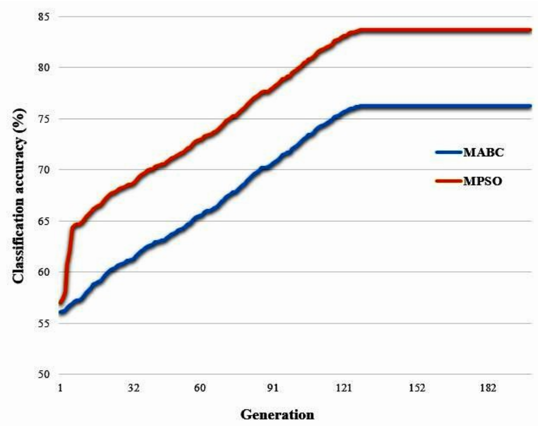

Figure 3. Convergence of modified ABC (MABC) and modified PSO (MPSO)

The classification rate obtained by this and best values of other studies for pima-diabetes disease dataset were presented in Table 3. As we can see from these results, our method using 10-fold cross validation obtains the highest classification accuracy, 85.19 %, reported so far. Therefore, we can draw this conclusion that the combination of particle swarm optimization and Fuzzy Logic would be very effective in detection of diabetes disease.

Table 3. Classification rate obtained with different methods for PID

Method	Classification rate	Study
Data gravitation method	76.56 %	Peng et al. [10]
Generalized discriminant analysis and least square SVM	79.16 %	Polat et al. [11]
Multi-layer neural network	79.62 %	Temurtas et al. [12]
General regression neural networks	80.21 %	Kayaer et al. [13]
Relational fuzzy neural networks	81.80 %	Leon and IV [14]
Modified ABC fuzzy classifier	84.21 %	Beloufa and Chikh [15]
Hybrid of artificial neural network	84.20 %	Kahramanli and Allahverdi [16]
ACO-based fuzzy classification system	84.24 %	Ganji and Abadeh [17]
Modified PSO fuzzy classifier	**85.19 %**	**Our study**

5. Conclusions

Diabetes is a complex and complicated disease characterized by either lack of insulin or a resistance to insulin, a hormone which is crucial for metabolism of blood sugar. This paper presents a fuzzy classifier called MPSOFC which was obtained by combination of particle swarm optimization algorithm and Fuzzy Logic for diabetes disease diagnosis. The novelty of the system lies in in the fact that the network can select good features along with the relevant rules in an integrated manner. Proposed method was also compared with the previous studies. The results indicate that our method achieved the highest classification accuracy among several well-known and recent learning methods. Moreover, results confirmed that the high classification rate of MPSOFC has been obtained by the means of a very simple and interpretable fuzzy rule base. According to obtained results, fuzzy method with modified PSO algorithm could be a powerful tool for diagnosis of diabetes disease.

References

[1]American diabetes association, <http://www.diabetes.org/diabetes-basics> (last accessed November 2014).

[2] E. I. Mohamed, R. Linderm, G. Perriello, N. Di Daniele, S. J. Poppl, and A. De Lorenzo, "Predicting type 2 diabetes using an electronic nose-base artificial neural network analysis", Journal of Diabetes Nutrition & Metabolism, Vol. 15, No. 4, 2002, pp. 215-221.

[3] N. Sean Ghazavi, and T.W. Liao, "Medical data mining by fuzzy modeling with selected features". Artificial Intelligence in Medicine, Vol. 43, 2008, pp. 195-206.

[4] S. Sun, and C. Zhang, "An experimental evaluation of ensemble methods for EEG signal classification". Pattern Recognition Letters, Vol . 28, 2007, pp. 2157–2163.

[5] Y. Zhang, Z. Lu, and J. Li, "Fabric defect classification using radial basis function network". Pattern Recognition Letters, Vol. 31, 2010, pp. 2033-2042.

[6] H. Guo, L.B. Jack, and A.K. Nandi, "Feature generation using genetic programming with application to fault classification". IEEE Transactions on systems, Man, and Cybernetics, Part B, Vol. 25, No. 1, 2005, pp. 89-99.

[7] L.A. Zadeh, "Fuzzy sets". Information and Control , Vol. 8, 1965, pp. 338-353.

[8] C. Rani and S.N. Deepa, "Design of optimal fuzzy classifier system using particle swarm optimization". in: International Conference on Innovative Computing Technologies (ICICT10) (IEEE), 2010: p. pp 1-6.

[9] A. Asuncion, and D. Newman, UCI repository of machine learning datasets. Availablefrom: <http://archive.ics.uci.edu/ml/datasets.html>, 2007.

[10] L. Peng, Y. Chen, and B. Yang, Chen, "A novel classification method based on data gravitation". Neural Networks and Brain, ICNN&B'05., 2005.

[11] P. Kamel, G. Salih, and A. Ahmet,"A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square suport vector machine". Expert Systems with Applications, Vol. 34, No. 1, 2008, pp. 482–487.

[12] H. Temurtas, , N. Yumusak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks". Expert Systems with Applications, Vol. 36, No. 4, 2009, pp. 8610-8615.

[13] K. Kayaer, and T. Yıldırım, "Medical diagnosis on Pima Indian diabetes using general regression neural networks", in the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), 2003, pp. 181-184.

[14] W. D. Leon, "Enhancing pattern classification with relational fuzzy neural networks and square BK-products". Ph.D dissertation in computer science. FL, USA: Springer, 2006, pp 71–74.

[15] B. Fayssal, and Ch. Mae, "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm". Journal of Computer Methods and Programs in Biomedicine, 2013, pp. 92-103.

[16] H. Kahramanli, and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases". Expert Systems with Applications, Vol. 35, 2008, pp 82-89.

[17] G. Mostafa, and A. Mohammad. A, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis". Expert Systems with Applications, Vol. 38, No. 12, 2011, pp. 14650-14659.

[18] J. Kennedy, and R. Eberhart, "Particle swarm optimization", in International Conference on Neural Networks, 27 November-1 December, Perth, 1995, pp. 1942–1948.

[19] J. Kennedy, and R. Eberhart, "A new optimizer using particleswarm theory", in Sixth International Symposium on Micro Machine and Human Science 4–6 October, Nagoya, Japan, 1995, pp. 39-43.