International Journal of Bio-Science and Bio-Technology
Vol.7, No.1 (2015), pp.39-50
http://dx.doi.org/10.14257/ijbsbt.2015.7.1.05

Analytics for the Quality of Fertility Data using Particle Swarm Optimization

Puneet Singh Duggal [1] , Sanchita Paul [1] and Priyanka Tiwari [2]

Department of Computer Science, BIT Mesra [1] , Index Medical College & Research Center, Indore [2]
duggal@gmail.com, sanchita.07@gmail.com, himmut@gmail.com

Abstract

In today's faced paced life, diseases and medical problems which were only confined to the elderly are slowly becoming common among the younger lot. These medical problems which are induced due to factors such as behavioural habits, eating habits, environmental factors, allergies and the lifestyle of individuals are termed as lifestyle diseases. Male fertility has slowly been degrading due to this. It has been a surge of cases of fertility and mortality degradation of semen which are correlated to the unhealthy and undisciplined lifestyles of the individuals. Studies have been conducted in the past to analyze the scenario through medical and clinical tests.

Non-medical behavioural and lifestyle aspects were studied and analyses were based on it. It was found out that non medical aspects also effect male fertility and there is an correlation between the two. In this paper the seminal quality is determine with the help of clustering techniques and validate using different classification techniques.

Keywords: *Medical Data Analytics, Particle Swarm Optimization, Clustering, Classification*

1. Introduction

The purpose of this paper is paper is to find the performance of different classification methods with Particle Swarm Optimization (PSO) to the male fertility dataset. This approach detects the problems faced by different Artificial Intelligence (AI) methods like Decision Tree, Support Vector Machine and Multi-Layer Perceptron from the ANN Domain [1-3]. This paper does a qualitative and quantitative comparative study among the mentioned methods to conclude into merits of using PSO for analytics of Fertility data.

This is achieved through classification via clustering. In the first step clustering techniques are used over data to produce different clusters. Clustering techniques uses distance measures to divide the data into different sets [4]. In this paper we are using PSO technique for clustering the data set [5].

The paper defines the description of the dataset followed by short description about the different AI methods used for classification. It then gives an insight about the PSO algorithm and how it is used for data clustering. It is proceeded by application of the mentioned and proposed methods for classification of the fertility dataset. The experimental results of the study are then compared on different mathematical performance parameters. The results which are compared gives us an understanding of the performance of the applied methods. The quality aspects are compared with respect to the quantitative values we get from the AI methods and the applied PSO method. Continuing, the testing carried out to analyze our result which leads to the final conclusion is also being described.

ISSN: 2233-7849 IJBSBT

Male Fertility dataset is used for the study, as medical data analytics is an upcoming domain and different methods and models are being developed and required for getting insights about the problems faced in the healthcare sector [6-7]. The qualitative aspect of the fertility dataset is classified into healthy and non-healthy factors with respect to the lifestyle changes and its effect on the quality of semen.

Data analytics techniques are used for the entire mentioned process which is the process of unearthing and displaying of meaningful patterns in data. It is relevant in the domains with bunch of collected information, Analytics is the methodology of analysis, it is not concerned with the individual analysis

or analysis steps. It used descriptive and predictive models based upon applications of statistics, Machine learning, Programming and OR techniques to find quantitative measurements about the performance of the data. Analytics uses this insight to guide decision making. Analytics also favours data visualization to communicate insight [2]

Data analytics using PSO is used in the study which has gained importance as nature inspired, swarm based techniques are giving close and accurate results [8]. PSO is a well documented and referred method applied in various domains [10]. It is being used as an optimization technique, clustering technique and segmentation technique [9, 11]. In this paper PSO is being used as a clustering technique [9] and compared to some well established AI techniques.

PSO method gives us an advantage over other methods as firstly, it is a very well researched method used in versatile domains. Secondly is has been proved to be faster than other methods and lastly, it tends to give a more accurate results.

In our study it is found that PSO when compared to other methods was giving more classification accuracy. Also, the rate of diagnostic power was better as compared to other methods. The specificity was better and the difference between the specificity and sensitivity was less for PSO. The false positive rates and the true negative rates were under the compromised values. It is important to note that in medical data analytics, the method used should give a higher level of accuracy as the results are used for the prediction of diseases. Therefore it becomes even more important to choose the method which will give a higher accuracy. Form data analytics point of view, the methods used should take less time to compute the results and when compared to different methods should prove to be accurate, fast and most importantly satisfy the quantitative quality parameters. PSO becomes an ideal choice because of its adhering to the listed tasks.

In future the authors want to study the medical domain through data analytics by applying Big Data analytics techniques [13]. PSO and other nature inspired swarm based techniques are very well suited for the distributed architecture and handling of high volume unstructured data in Big Data analytics [12, 14].

2. Methodology

The method proposed uses cluster for classification by the use of a meta-classifier. This technique follows the hypothesis that each cluster corresponds to a class. In the paper, fertility data is taken from 100 healthy individuals aged between 18-35 years. Different lifestyle parameters like, smoking, drinking and exercise are taken into consideration other then the medical ones like mortality and number of active sperms in the sample.

The cumulative data is pre-processed using an optional attribute selection process. This intermediate step is used to filter only the relevant and important attributes from the bulk data. Attribute selection can affect the overall results derived from the data.

In the next step, PSO algorithm is applied as a clustering algorithm using the training data. The derived or the predicted clusters are compared to the expected class set. It is used to predict the class labels from the unseen data instances. The class attribute is used to evaluate the obtained cluster as classifiers. The number of cluster should be same as the number of expected classes. Based upon the one to one mapping of the expected and the desired classes, a confuse matrix is derived. The confuse matrix gives the disparities if any, between the desired and the expected classification of data.

For cluster to class progression, PSO algorithm is used which is compared to the proven meta classifiers here viz DT [25, 26], SVM [22, 23] and MLP [27-29]. Figure 1 shows the methodology of the process used.

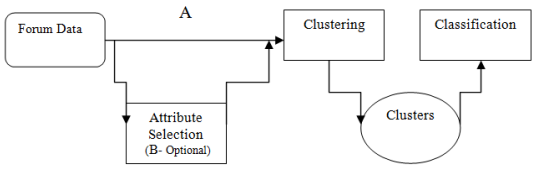

Figure 1. Methodology Adapted

3. Clustering Technique

Clustering is a technique [4] where the datasets is divided into different groups based upon the distance measure from the cluster centre. All the clustering algorithms are based upon the following process.

Step1: Define the number of cluster centers.

Step2: Initialize the cluster centers with a random value.

Step3: Calculate the distance of each instance in the dataset from the cluster centers defined.

Step4: Based on the shortest distance from the defined cluster centers allocate the instance to the cluster which has the shortest distance from the cluster centre.

Step5: Update the value of cluster centers based upon mathematical and logical functions.

Step6: Repeat steps 3-5, until the values of new cluster centers are optimized

4. Particle Swarm Optimization (PSO) based Clustering

PSO is one of the most widely used technique used in data analytics. Particle swarm optimization (PSO) is a population-based stochastic search process, it is emulated on the behavioral analysis of the birds in a group [10]. This algorithm is based upon the notion that each bird in the group can lead to a possible solution. PSO based techniques is a mix of Exploration and Exploitation.

In Exploration stage, each member of the population referred as particle searches for the possible solution. The expected outcome of the algorithm is to identify a position by the particle in the search space which results in the optimized evaluation of the objective function. The possible solution is compared with the defined fitness (objective) function and accordingly the position of the particle is adjusted to be comparable with the fitness function [8-10].

Every bird denoted as particle corresponds to a location in N_d dimensional space, which is moved as in exploration phase in the multi-dimensional search space. After every iteration, The initial point adjust its position and crawls towards best position in the neighborhood of that particle and the ***k*** particle's best position found so far.

Particle i as a factor, denotes the position, velocity, best position:
xi : Existing position of the particle
vi : Existing velocity of the particle
yi : Individual finest position of the particle.
Using the mentioned details, a particle's location is attuned according to

$$vi,k(t+1) = wvi,k(t) + c1r1,k(t)(yi,k(t) - xi,k(t)) + c2r2,k(t)(yk(t) - xi,k(t)) \ldots\ldots\ldots\ldots Eq1$$

$$xi(t+1) = xi(t) + vi(t+1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots Eq2$$

where , w = Inertia weight
$c1, c2$= Acceleration constants
r1, j(t), r2, j(t) to U(0,1)
k = 1,...,N_d (The N_d dimensional space)

The velocity of the particle is calculated on the basis of the mentioned based on the assistance

(1) A partial value of the n-1(last position) velocity.
(2) A learning and understanding component which is a function of the distance of the particle from its present best position.
(3) The inter group communication which is a distance measure of the particle when compared and calculated with the personal best particle found.

Individual finest position of the individual in swarm, *i* is computed as

$$yi(t+1) = \begin{cases} yi\,(t) & if\ f\left(xi(t+1)\right) \geq f(yi(t)) \\ xi\,(t+1) & if\ f\left(xi(t+1)\right) < f(yi(t)) \end{cases} \quad \text{--------Eq3}$$

The PSO algorithm is repeatedly applied through equations (1) and (2) until a specific times iterations are executed. Also, the algorithm is ended when the updated velocity after specified number of iterations is close to zero.

4.1. PSO Algorithm for Clustering

```
Using the standard gbest PSO, data vectors can be clustered as follows:

1.Initialize each particle to contain Nc randomly selected cluster
centroids.

2. For t = 1 to tmax do

(a) For each particle i do
(b) For each data vector zp

   i. Calculate the Euclidean distance d(zp,mij) to all cluster centroids Cij
   ii. Assign zp to cluster Cij such that d(zp,mij) = minc=1,..,Nc{d(zp,mic)}
   iii. Calculate the fitness using equation (3)

(c) Update the global best and local best positions

(d) Update the cluster centroids using equations (1) and (2).

where tmax is the maximum number of iterations
```


5. Clustering Technique as a Classifier

The PSO clustering method used for predicting the accuracy of the fertility dataset, which is classified according to the different methods by researchers. Different methods such as Multilayer perception (MLP), Support Vector Machine (SVM) and Decision Tree (DT) are used in the past for finding the accuracy of the fertility dataset [1-3].

6. Performance Parameters

Different performance parameters such as Diagnostic Power, Classification Rate, Sensitivity, Specificity, False Positive Rate, False Negative Rate, Positive Predictive Power, Negative Predictive Power and Kappa values are used for validating our methodology with the existing test results available in the public domain. Parameters which are used for the validation of the data are mentioned in the table 1. The parameters are used for validating the outcomes of the technique used and its comparison with the other techniques used before.

'a',' b', 'c', 'd' values used for the calculation of the parameters are confuse matrix components, viz True Positive (a), False Positive (b), False Negative (c) and True Negative (d). True Positive and True Negative denotes the classified values which are correctly identified by the classifier, False Negative and False Positive denotes the values which were not classified correctly by the classifier. All these values are the resultant from the actual and the predicted calculations of the projected confuse matrix [15]

Table 1. Performance Parameters Used

Parameters	Formula
Prevalence	(a + c)/N
Overall Diagnostic Power	(b + d)/N
Correct Classification Rate	(a + d)/N
Sensitivity	a/(a + c)
Specificity	d/(b + d)
False Positive Rate	b/(b + d)
False Negative Rate	c/(a + c)
Positive Predictive Value (PPV)	a/(a + b)
Negative Predictive Value (NPV)	d/(c + d)
Misclassification Rate	(b + c)/N
Odds-ratio	(a d)/(c b)
Kappa	(a + d) - (((a + c)(a + b) + (b + d)(c + d))/N) / N - (((a + c)(a + b) + (b + d)(c + d))/N)

.

Kappa Values

It can be calculation/measure of the similarity between the understanding of two outcomes of different sets. Kappa is the probabilistic measure of the amount of agreement of the adjusted values with respect to the data values in the main diagonal of the table. Classification methods classify the objects into categories, Table 2 shows the cell probabilities of a 2X2 classification table.

Table 2. Kappa Measurement Parameters

Rater 1 / Rater 2	Category 1	Category 2	Total
Category 1	P11	P12	P1
Category 2	P21	P22	P2
Total	P1	P2	1

Kappa values are calculated by subtracting the experimental level of agreements $P0 = P11 + P22$ with the expected values of the classifiers P$e = P21.P12 + P12.P21$ [16-17].

The value of Kappa is defined as

$$k = \frac{P0 - Pe}{1 - Pe}$$

K is the ratio of the difference of probability index of the observed and the expected value with the complement of the expected probability.

Kappa has the maximum value of 1, if the level of agreement or the observed outcome is maximum. This makes the numerator part of the equation as large as the denominator. Kappa values can be negative, which means that there is no possible agreement between the expected and the observed values. The different Kappa values are shown in the Table 3.

Table 3. Interpretations of Kappa Parameter Values

Kappa values	Kappa values interpretation
Negative value	Problem in the application of test
Zero	No agreement between the tests and predicted outcome
0 < Kappa < 0.20	Poor agreement between the tests and predicted outcome
0.20 < Kappa < 0.40	Fair to Moderate agreement between the tests and predicted outcome
0.40 < Kappa < 0.60	Good agreement between the tests and predicted outcome
0.60 < Kappa < 0.80	Very Good agreement between the tests and predicted outcome
0.80 < Kappa < 1	Excellent to perfect agreement between the tests and predicted outcome

Accuracy: Accuracy of a model is defined as the total positive instances of the model are divided by the total number of instances. Accuracy parameter provides the percentage of correctly classified instances. The accuracy of model is defined as

Accuracy = (a + d)/N

Sensitivity: This parameter is used to determine the degree of the attribute to correctly classify the person with diseases and is defined as

Sensitivity= a/(a + c)

Specificity: This parameter is used to determine the degree of the attribute to correctly classify the person without diseases and is defined as

Specificity=d/(b + d)

Confuse Matrix: The confuse matrix has been used to determine the relationship between the actual values and predicted values. Table 4 represents the structure of confuse matrix [15].

Table 4. Confuse Matrix Representation

Actual / Predicted	Actual Positive	Actual Negative
Predicted Positive	True Positive (a)	False Positive (b)
Predicted Negative	False Negative (c)	True Negative (d)

7. Description of the Data Used

Fertility dataset contains nine features which are season, age, accident/trauma, childish disease, high fever, surgical intervention, alcohol consumption, smoking habits and number of hours spent sitting per day. Table 5 shows the statistics of fertility dataset. The original dataset contains 100 instances with nine attributes with two classes. The classes are normal and altered fertility rate. But, to find out more relevant features from fertility dataset, feature selection methods are applied to fertility dataset.

It is concluded that some features have lower impact on the overall quality of the data and sometimes non serious data act as a noise to the data adding to the importance of the features. In the data some features may have less impact to predict the fertility rate both experimentally and medically. Due to the above reason, the final experimental dataset contains seven attributes rather than nine attributes.

Table 5. Characteristics of the Data Used

Name of attribute	Attribute role	Attribute type	Attribute statistics	Attribute range	Missing values
Diagnosis	Prediction	Binominal	mode = N (88), least = A (12)	(88), A (12)	0
Season	Regular	Nominal	avg. = −0.072 ± 0.797	[−1.00 ; 1.00]	0
Age	Regular	Real	avg. = 0.669 ± 0.121	[0.500 ; 1.000]	0
Childish diseases	Regular	Nominal	avg. = 0.870 ± 0.338	[0.000 ; 1.000]	0
Accident or serious trauma	Regular	Nominal	avg. = 0.440 ± 0.499	[0.000 ; 1.000]	0
Surgical intervention	Regular	Nominal	avg. = 0.510 ± 0.502	[0.000 ; 1.000]	0
High fevers	Regular	Nominal	avg. = 0.190 ± 0.581	[−1.00 ; 1.00]	0
Alcohol consumption	Regular	Nominal	avg. = 0.832 ± 0.168	[0.200 ; 1.000]	0
Smoking habit	Regular	Nominal	avg. = −0.350 ± 0.809	[−1.00 ; 1.00]	0
Number of hours spent sitting	Regular	Nominal	avg. = 0.407 ± 0.186	[0.060 ; 1.000]	0

8. Experimental Setup

The experiments were performed on a Pentium core i3 having virtualization enabled 1.8 GHz CPU with 8 GB RAM based machine. Partial dataset was used for the PSO based clustering for prediction model. The efficiency of the model was tested using an n-fold cross validation method [Figure 2], a 10 fold process was used in the experiments. This was done for the model built during the verification phase while checking the dataset with the mentioned classifiers. This process was executed up to 10 iterations and the each iteration will be consists the different test instance [21].

All the experiments were conducted in Distributed Hadoop Weka, Distributed Weka Base 3.7X [18] and Matlab® 2012a. Weka for Hadoop was used because of its architecture. As in Hadoop, the data is stored distributed as 64-256 MB chunks in the HDFS file system depending on the size of data and the version of Hadoop installed for future analytics in the domain of Big Data Analytics and virtualized distributed systems . The clustering algorithm using PSO was developed using Matlab®.

The approach used has followed the clusters to classes process, where one to one mapping is used in the cluster evaluation to find a minimum-error mapping of clusters to classes. On the basis of which confusion matrix is made, further validation is done using the different performance parameters.

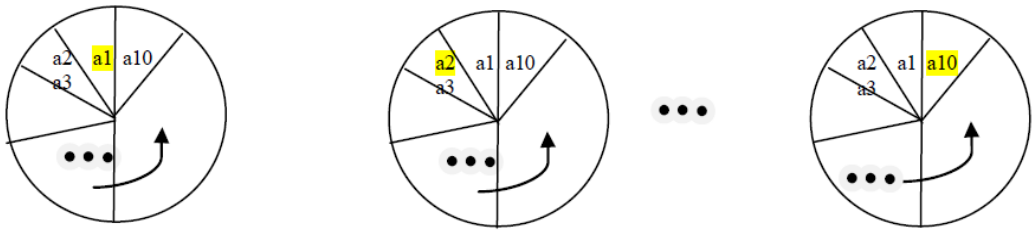

Figure 2. A Systematic Diagram of 10 Fold Cross-Fold Technique

9. Experimental Setup

Table 6. Confuse Matrix for PSO

Confusion Matrix PSO	Actual (+)	Actual (-)
Predicated (+)	81 (a)	5 (b)
Predicated (-)	7 (c)	7 (d)

Table 7. Confuse Matrix for MLP

Confusion Matrix MLP	Actual (+)	Actual (-)
Predicated (+)	80	5
Predicated (-)	9	6

Table 8. Confuse Matrix for SVM

Confusion Matrix SVM	Actual (+)	Actual (-)
Predicated (+)	83	2
Predicated (-)	12	3

Table 9. Confuse Matrix for DT

Confusion Matrix DT	Actual (+)	Actual (-)
Predicated (+)	82	3
Predicated (-)	13	2

In this paper, three classic classification techniques (meta-classifiers) are compared with the Particle Swarm Optimization (PSO) based clustering technique. On the basis of the results generated with the PSO and later through one-on-one mapping with the classified fertility dataset, we come to the conclusion that PSO clustering method can act as a classifier for problems in classification. Three Artificial Inelegance methods SVM, MLP and AI were used for analysis of fertility data.

Dataset was obtained from 100 young male volunteers aged between 18 to 36 years. Through the analysis we can come to the conclusion that lifestyle parameters effect the quality of the semen. The analysis clearly shows a high accuracy rate for the techniques used with respect to the different measurement parameters.

The results shows lower values for the specificity and the PPV even when the classification values are good which is also reflected in the confuse matrix. This discrepancy may be due to the dataset skewed towards the population who are healthy. We can see that the actual value of the healthy individuals is 88 as compared to 12 individuals who are unhealthy. This means that if there is imbalance in the distribution of values then, we may get lower Specificity and positive predictive values. These results can be analyzed using table 6-9.

PSO gives the highest accuracy then both MLP and SVM methods, PSO obtains superior Specificity values at 88 % . Therefore, PSO seems to be the preferred method for predicting the quality of fertility data with respect to the environmental factors and lifestyle, this method seems to be useful with the new data also as it can handle generalization potential.

MLP, SVM and further DT gives slightly lower accuracy, DT has an upper hand on others due to its simple visuals and illustration while understanding and interpreting the data. Decision tree models are easy for the non technical researcher, Also, very less data pre-processing is required in it as compared to other techniques which require data normalization.

Using PSO algorithm makes the best use of the above factor as it does not require data preparation and normalization, on top of that, the time taken for calculating through PSO is much less as compared to the other three AI techniques (Refer table 10).

Data pre-processing techniques like feature selection and elimination could be used to further enhance the accuracy of the model.

Soft computing and other computational techniques have been used, in the domain of reproductive data analysis, case here is seminal analysis for forecasting the results of In vitro fertilization / Intracytoplasmic sperm injection, to assess sperm morphology and to predict the presence of healthy sperm in testes of men with non-obstructive azoospermia [1-3].

This paper tries to develop and non-linier methodology for recognition of a logical relationship between the semen quality and life-style / environmental factors. Previous studies have some limitations and lacunas [2, 6-7]. PSO based method allows for an optimal approach towards the complex problem, due to is meta-heuristic approach.

Table 10. Comparison of the Different Techniques Used

Parameters	Statistics/Results			
	PSO	**MLP**	**SVM**	**DT**

Overall Diagnostic Power	0.12	0.11	0.05	0.05
Correct Classification Rate	0.88	0.85	0.85	0.84
Sensitivity	0.9205	0.8989	0.8737	0.8632
Specificity	0.5833	0.5455	0.6	0.4
False Positive Rate	0.4167	0.4545	0.4	0.6
False Negative Rate	0.0795	0.1011	0.1263	0.1368
Positive Predictive Power	0.9419	0.9412	0.9765	0.9647
Negative Predictive Power	0.5	0.4	0.2	0.1333
Kappa	0.47	0.3833	0.2432	0.1351

Using AI methods give high accuracy through different classification techniques, they are not an outright replacement but can surely be an alternative to the already overburdened and expensive laboratory tests. The presented methods and their improvements can at least replace the initial tests for checking the fertility of a population or an individual as well as in selecting the donors for data collection. Stating and testing environmental and lifestyle factors trough the mentioned techniques are able to give an accuracy of up-to 90%, these results supports the argument.

10. Discussion and Future Scope

In future studies, apart from performing the analytics on Distributed platform adhering to Big Data analysis, the effect of imbalanced classes on classification performance can be studied and solved while developing computational methods using Fuzzy Systems, MLP, SVM, DT, PSO and other Swarm based techniques for artificial immune systems, decision support systems, and Quality analysis for medical diagnosis .

Machine learning and data mining and Big Data analytics methods when combined will enhance the correlation between medical data, here seminal data and other non-medical, non-clinical attributes.

More efficient data collection techniques need to be developed and used combined with Data warehousing techniques where large heterogeneous data coming from sensors can be processed. Merging of historical and current data to analyze and predict the medical problems facing mankind needs to be addressed.

Use of Big Data analytics can be of much importance in the field, as Big Data can accommodate very high volume of heterogeneous data and performing analytics on this data may prove to be very useful.

The benefit of using Big Data analytics is that it may handle large and unstructured heterogeneous data with higher efficiency and analytics will be performed with higher accuracy on our systems and also, it will supports performance improvement [13-14].

In conclusion, this is the first time that PSO has been used and compared with MLP, SVM and DT to address the issue of the relationship between life styles of the population and semen quality.

PSO shows the highest prediction accuracy at 88 % whereas MLP and SVM show a much less prediction accuracy of 85%. The clarity, reduction and simplification of the problem while applying DT may also be noted at a much less accuracy of 84%.

References

[1] G. David, J. L. Girela, J. D. Juan, M. Jose Gomez-Torres and M. Johnsson, "Predicting seminal quality with artificial intelligence methods Expert Systems with Applications", (**2012),** pp. 12564-12573.
[2] D. Gil, J. L. Girela, J. De Juan, M. Jose Gomez-Torres, and M. Johnsson, "Predicting seminal quality with artificial intelligence methods", Expert Systems with Applications, (**2012),** pp. 12564–12573.
[3] J. L. Girela, D. Gil, M. Johnsson, M. J. Gomez-Torres and J. De Juan, "Semen Parameters Can Be Predicted from Environmental Factors and Lifestyle Using Artificial Intelligence Methods Biology of Reproduction", (**2013)**, pp. 1–8.
[4] J. Han, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, CA, USA, (**2005**).
[5] I. Mandal, "SVM-PSO based Feature Selection for Improving Medical Diagnosis Reliability using Machine Learning Ensembles", (**2012**), pp. 267-276.
[6] Y. N. Devi and S. Anto, "An Evolutionary-Fuzzy Expert System for the Diagnosis of Coronary Artery Disease", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) vol. 3, no. 4, (**2014**), pp. 1478-1484.
[7] Y. Kumar and G. Sahoo, "Prediction of different types of liver diseases using rule based classification model", Technology and Health Care, (**2013**), pp. 417-432.
[8] F. den Bergh, "An Analysis of Particle Swarm Optimizers", PhD Thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa, (**2002**).
[9] M. Omran, A. Salman and A. P. Engelbrecht, "Image Classification using Particle Swarm Optimization", Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning, Singapore, (**2002**).
[10] Kennedy and RC Eberhart, "Particle Swarm Optimization, Proceedings of the IEEE International Joint Conference on Neural Networks", vol. 4, (**1995**), pp. 1942–1948.
[11] D. W. van der Merwe and A. P. Engelbrecht, "Data Clustering using Particle Swarm Optimization.
[12] R. Steinbrook, "Personally Controlled Online Health Data-The Next Big Thing in Medical Care", *The* New England Journal of Medicine, (**2008**), pp. 1653-1656.
[13] T. B. Murdoch and A. S. Detsky, "The Inevitable Application of Big Data to Health Care", Journal of American Medical Association, no. 13, (**2013**), pp. 1351-1352.
[14] A. O'Driscoll a, J. Daugelaite b and R. D. Sleator, "Methodological Review 'Big data', Hadoop and cloud computing in genomics", Journal of Biomedical Informatics, (**2013**), pp. 1-8.
[15] R. Kohavi and F. Provost, "Glossary of terms", Machine Learning, (**1998**), pp. 271-274.
[16] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The Kappa Statistic", Journal of Farm Med, Columbia University, (**2005**).
[17] J. L. Fleiss, J. Cohan and B. S. Everitt, "Large sample standard errors of Kappa and weighted Kappa", Psychological Bulletin, (**1969**).
[18] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
[19] H. Liu and M. Hiroshi, "Feature selection for knowledge discovery and data mining", Springer, (**1998**).
[20] G. Isabelle and E. André, "An introduction to variable and feature selection", The Journal of Machine Learning Research, (**2003**), pp. 1157-1182.
[21] S. Geisser, "Predictive inference: an introduction", vol. 55, CRC Press, (**1993**).
[22] C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, (**1995**), pp. 273-297.
[23] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST), (**2011**).
[24] C. W. Hsu, C. C. Chang and C. J. Lin, "A practical guide to support vector classification", National Taiwan University, Taiwan (**2003**).
[25] J. R. Quinlan, "Induction of decision trees", Machine Learning, vol. 1, no. 1, (**1986**), pp. 81-106.
[26] K. Polat and S. Gnnes, "A novel hybrid intelligent method based on c4.5 decision tree classifier and one-against-all approach for multi-class classification problems", Expert Systems with Applications, vol. 36, no. 2, Part 1, (**2009**), pp. 1587–1592.
[27] B. D. Ripley, "Pattern recognition and neural networks", Cambridge University Press, (**1996**).
[28] S. Haykin, "Neural networks: A comprehensive foundation", Englewoods Cliffs NJ: Prentice-Hall, (**1998**).
[29] C. M. Bishop, "Neural networks for pattern recognition", Oxford University Press, (**2005**).
[30] M. J. Zinaman, C. C. Brown, S. G. Selevam and E. D. Clegg, "Semen Quality and Human Fertility: A Prospective Study with Healthy Couples", Journal of Andrology, vol. 21, no. 1, (**2000**) January/February, pp. 145-153.
[31] A. J. Gaskins, D. S. Colaci, J. Mendiola, S.H. Swan and J. E. Chavarro, "Dietary patterns and semen quality in young men", Human Reproduction, vol. 27, no. 10, (**2012**), pp. 2899–2907.
[32] E. Carlsen, A. Giwercman, N. Keiding and N. E. Skakkebaek, "Evidence for decreasing quality of semen during past 50 years", BMJ (**1992**), pp. 609-613.

[33] J. Auger, J. M. Kunstmann, F. Czyglik and P. Jouannet, "Decline in semen quality among fertile men in Paris during the past 20 years", N. Engl. J. Med., (**1995**), pp. 281-285.
[34] W. Y. Wong, G. A. Zielhuis, C. M. Thomas, H. M. Merkus and R. P. Steegers-Theunissen, "New evidence of the influence of exogenous and endogenous factors on sperm count in man", Eur. J. Obstet. Gynecol. Reprod. Biol., (**2003**), pp. 49-54.
[35] A. Giwercman and Y. L. Giwercman, "Environmental factors and testicular function", Best Pract. Res Clin Endocrinol Metab, (**2011**), pp. 391-402.
[36] A. C. Martini, R. I. Molina, D. Estofan, D. Senestrari, M. Fiol de Cuneo and R. D. Ruiz, "Effects of alcohol and cigarette consumption on human seminal quality", Fertil Steril (**2004**), pp. 374-377.
[37] A. Agarwal, N. R. Desai, R. Ruffoli and A. Carpi, "Lifestyle and testicular dysfunction: a brief update", Biomed Pharmacother, (**2008**), pp. 550-553.
[38] M. C. Inhorn, "Global infertility and the globalization of new reproductive technologies: Illustrations from Egypt", Soc. Sci. Med., (**2003**), pp. 1837-1851.
[39] W. Lutz, B. C. O'Neill and S. Scherbov, "Demographics", Europe's population 295 at a turning point, Science, (**2003**), pp. 1991-1992.
[40] J. Grant, S. Hoorens, S. Sivadasan, M. V. Loo, J. Davanzo, L. Hale and W. Butz, "Trends in European fertility: should Europe try to increase its fertility rate or just manage the consequences?", Int. J. Androl, vol. 26, (**2006**) , pp. 17-24.
[41] S. O. Skouby, "Contraceptive use and behavior in the 21st century: A comprehensive study across five European countries", Eur. J. Contracept. Reprod. Health Care, (**2004**), pp. 57-68.

Authors

Puneet Singh Duggal, He is associated with in BIT Mesra, Ranchi as a research scholar in the Department of Computer Science & Engineering. His research domains are Data Analytics, Big Data Analytics, Medical Informatics and Cloud Computing.

Sanchita Paul, She is working as an Assistant Professor in the Department of Computer Science & Engineering in BIT Mesra, Ranchi. Her interests are in the areas of Artificial Intelligence, Pattern recognition, Parallel Computing, Automata Theory, Design and analysis of Algorithms. Her Research areas are in the field of Cloud Computing, Natural Language processing, DNA Computing, Machine Learning, Robotics & Nanotechnology.

Priyanka Tiwari, She is a practicing pathologist with expertise in Cancer Diagnostics. She is doing her MD in Clinical Pathology. Her areas of interest are Thyroid Malignancy and Haematology.

