IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Performance Analysis of Classification Algorithms on Diabetes Dataset

K. Saravanapriya[1*], J. Bagyamani[2]

[1*] Dept. of MCA, Sacred Heart College (Autonomous), Tirupattur, India
[2] Dept. of Computer Applications, Chikkanna Government Arts College, Tiruppur, India

*Corresponding Author: rajpriya2109@gmail.com, Tel.: +91-99658-14162

Available online at: www.ijcseonline.org

Received: 22/Aug/2017, Revised: 04/Sep/2017, Accepted: 19/Sep/2017, Published: 30/Sep/2017

***Abstract*—** Healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor' [1]. Today in this hectic lifestyle, one of the major threats to human health is Diabetes Mellitus. Valuable knowledge can be discovered from application of data mining techniques in the Health care System particularly in Diabetes Database. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. This paper aims to analyze the performance of the classification techniques in diabetes data set.

***Keywords*—**Diabetes Mellitus, Data Mining, Classification, Naïve Bayes, Random Forest, J48, JRIP, Multilayer Perceptron, KNN, Support Vector Machine, RBF Network, Weka.

I. INTRODUCTION

Diabetes is the seventh leading cause of death according to U.S. death certificates and also, it is the major cause for heart stroke, kidney failure, non-traumatic lower-limb amputations and blindness [2]. According to the Lancet study 2016 [3], China, India and USA are among the top three countries with a high number of diabetic population. In India, the disease has increased drastically from 11.9 million in 1980 to 64.5 million in 2016. Prevalence of diabetes has more than doubled for men from 3.7% to 9.1%. It also has increased by 80% among women from 4.6% to 8.3%. According to an ongoing Indian Council of Medical Research – IDIAB study among 14 states in the country, Tamil Nadu tops the list in the prevalence of diabetes. While this is 13.8% in the urban areas, rural Tamil Nadu has almost 9%. The study says that there is no much difference in the prevalence rates between urban and rural areas in Tamil Nadu. The average is almost 10% for the whole state. This means that of the estimated 35 million adults in the state, 3.5 million are diabetic. In a highly urbanized area like Chennai, it is 24.5% among those aged above 20 years, 40% of the age group are 50 and above, and 35% is pre-diabetic.

There are three major types of diabetes such as Type 1, Type 2 and Gestational diabetes. Type1 diabetes is also called as insulin-dependent diabetes. It used to be called as juvenile-onset diabetes, because it often begins in childhood. The most common form of diabetes is type 2 diabetes, accounting for 95% of diabetes cases in adults. Type 2 diabetes also called adult-onset diabetes, but with the common symptoms of obese and overweight. Type 2 diabetes is also called non-insulin-dependent diabetes. Diabetes that is triggered by pregnancy is called gestational diabetes. Data mining techniques identify valid patterns and relationships that provide useful information. The objective is to help the physicians to ease and improve their work in an efficient way to frequently assess the patient's details and predict the treatment in advance, so that the mortality rate due to the cause of this disease can be reduced. Though knowledge discovery in databases has conveyed many implications in domains such as fraud detection, targeted marketing etc., it is more essential to apply of data mining techniques towards the health sector. This paper focuses on the performance analysis of classification techniques in PIMA Indian diabetes data from the UCI machine learning repository set that refers to females of ages from 21 to 81. The data set has 768 instances with 9 input attributes including 2 class attributes *such as number of times pregnant, plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure (mm/Hg), triceps skin fold thickness (mm), 2-hour serum insulin (μU/ml), body mass index (kg/m), Diabetes Pedigree Function, Age (year), Status (0-Healthy, 1-Diabetes*). Out of the nine condition attributes, six attributes describe the result of physical examination, rest of the attributes are of chemical examinations. The tool used is Weka from Waikato. The tool we have used here is the Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University Of Waikato, in New Zealand. It is free software licensed under the GNU General Public License [4]. Initially the data is pre-processed to eliminate the missing values. The class label has been divided into two numerical variable with a 1 means tested_negative and 2 stands for tested_positive.

This paper has been divided into four section. Section I gives an overview about the classification techniques that has been

adopted here. In this paper, we have chosen some 8 prominent classification techniques for the comparison. Section II tells about the measures used in Weka tool. Section III deals with Results and Findings that had been obtained. Section IV compares the different classification techniques and their time taken to build the model and the instances that had been correctly classifies using the tool the fifth section concludes the paper.

II. CLASSIFICATION IN DATA MINING

Classification is a data mining technique that allocates items in a collection of data to the target categories or classes. It intends to accurately predict the target class for each case in the data. It classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. Here we have considered four main classification algorithms such as Naïve Bayes, Random Forest, Multilayer Perceptron, and J48.

A. Naïve Bayes

The Bayesian Classification is both a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data [5].

B. Random Forest

Random forests or random decision forests is a collective learning method for classification, regression and other tasks, that work by constructing a multitude of decision trees at training time and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set [6].

C. Multilayer Perceptron

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Multilayer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs were very popular machine learning solution in the 1980. It was applied in various fields like Speech recognition, Image recognition, and machine translation software [7].

D. J48

J48 is the decision tree based algorithm and it is the extension of C4.5. With this technique a tree is constructed to model the classification process in decision tree the internal nodes of the tree denotes a test on an attribute, branch represent the outcome of the test, leaf node holds a class label and the topmost node is the root node. Model generated by decision tree helps to predict new instances of data. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple.

E. JRip

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William Cohen (1995) as an optimized version of IREP. It is based on association rules with Reduced Error Pruning (REP), a very common and effective technique found in decision tree algorithms.

F. SVM

SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be.

G. RBF Network

Radial Basis Function Network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control. They were first formulated in the year 1988 by the two researchers Broomhead and Lowe, at the Royal Signals and Radar Establishment [8]. Figure 1 [8] depicts the RBF Network.

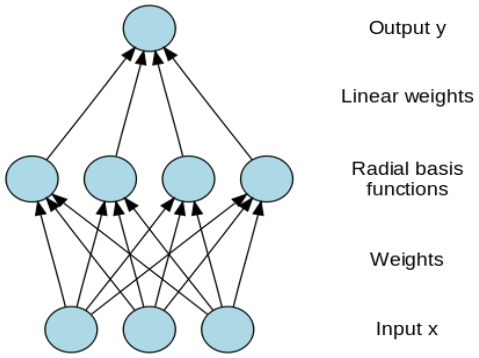

Figure. 1. RBF Network

III. MESURES IN WEKA

A. Kappa Statistics (KS)

The kappa statistic measures the agreement of prediction with the true class. The equation is expressed as in Eq. (1)

$$K = P_0 - Pe \quad (1)$$

Where P_o is the relative observed agreement among raters, and P_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by P_e), $\kappa \leq 0$.

B. Recall (RC)

Recall is the True Positive rate. It is also referred to as sensitivity. It is given by the formula as in Eq. (2)

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

C. Precision (P)

Precision is that what fraction of those predicted positive are actually positive. It is given by the formula as in Eq. (3)

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

D. F-Measure (FM)

F-Measure is a measure that combines precision and recall. It is also known as the measure, because recall and precision are evenly weighted.

$$\text{F-measure} = 2 \text{ X } ((\text{Precision X Recall}) / (\text{Precision} + \text{Recall})) \quad (3)$$

E. Mean Absolute Error(MAE)

The subscript The MAE measures the average magnitude of the errors in a set of calculations, without considering their direction. It measures *accuracy* for continuous variables. It is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

F. Root Mean Squared Error(RMSE)

The RMSE is used to measure the average magnitude of the error. It is the difference between the calculation and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

The MAE and the RMSE are used to identify the variation in the errors in a set of calculations.

The RMSE will always be greater or equal to the MAE. If the difference between them is greater, the *variance* in the individual errors in the sample is greater. If RMSE is equal to MAE, then all the errors are of the same magnitude.

G. Stratified Sampling

Stratified sampling refers to a type of sampling method. With stratified sampling, the researcher divides the population into separate groups, called strata. Then, a probability sample (often a simple random sample) is drawn from each group. Stratified sampling has several advantages over simple random sampling. Here the percentage split is applied by the method of stratified random sampling.

H. Weka results

TP = True Positives: Number of examples predicted positive that are actually positive.
FP = False Positives: Number of examples predicted positive that are actually negative.
TN = True Negatives: Number of examples predicted negative that are actually negative.
FN = False Negatives: Number of examples predicted negative that are actually positive

I. Confusion Matrix

It is a measure that combines precision and recall. It is also known as A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as

another). It is commonly named as contingency table. For a 2x2 confusion matrix, the matrix could be arbitrarily large. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. In Weka confusion matrix, if a, is taken to be positive class (ex. Has Disease).

	a	b	← classified as
Actual a=0	TP	FN	
Actual b=1	FP	TN	

In Weka Confusion Matrix, if a, is taken to be the negative class (ex: no disease):

	*a	b	← classified as
Actual a=0	TP	FN	
Actual b=1	FP	TN	

The **True Positive** (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row.

The **False Positive** (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. In the matrix, this is the column sum of class x minus the diagonal element, divided by the rows sums of all other classes.

Weighted Average (WA) is the average of all measures, each measure is weighted according to the number of instances with that particular class label.

IV. RESULTS AND FINDINGS

The accuracy by the class has been analyzed for each classification technique.

A. Naïve Bayes

From Table 1, Naïve Bayes yields an average of true positive rate with 0.77 and a false positive rate of 0.313. It has a precision values of 0.767, Recall value with 0.77, F- Measure with 0.769 and ROC curve with 0.854. It took 0.09 secs to build the model and it correctly classified 77.02% instances.

Table 1. Naïve Bayes

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.843	0.386	0.824	0.843	0.833	0.854	Tneg
	0.614	0.157	0.646	0.614	0.63	0.854	Tpos
WA	0.77	0.313	0.767	0.77	0.769	0.854	

B. J48

J48 results with an average of true positive rate with 0.762 and a false positive rate of 0.342. It has a precision values of 0.756, Recall value with 0.762, F- Measure with 0.758 and ROC curve with 0.796. It took 0.53 secs to build the model and it classified 76.25% instances correctly. This is shown in Table 2.

Table 2. J48

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.854	0.386	0.843	0.843	0.833	0.854	Tneg
	0.566	0.157	0.646	0.614	0.63	0.854	Tpos
WA	0.762	0.342	0.75	0.762	0.758	0.796	

C. Random Forest

Random Forest results with an average of true positive rate with 0.782 and a false positive rate of 0.342. It has a precision values of 0.756, Recall value with 0.762, F- Measure with 0.758 and ROC curve with 0.796. This is given in Table 3. It requires 1.58 secs to build the model and the correctly classified instances were 78.20%.

Table 3. Random Forest

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.876	0.422	0.824	0.843	0.833	0.854	Tneg
	0.578	0.124	0.646	0.614	0.63	0.854	Tpos
WA	0.782	0.342	0.756	0.762	0.758	0.796	

D. JRip

In Table 4, JRIP yields an average of true positive rate with 0.77 and a false positive rate of 0.287. It has a precision values of 0.773, Recall value with 0.77, F-Measure with 0.772 and ROC curve with 0.741. It took 0.27 secs to build the model and it classified 77.27% instances correctly.

Table 4. JRip

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.82	0.337	0.839	0.82	0.83	0.741	Tneg
	0.663	0.18	0.632	0.663	0.647	0.741	Tpos
WA	0.77	0.287	0.773	0.77	0.772	0.741	

E. Multilayer Perceptron

Multilayer Perceptron gives an average of true positive rate with 0.743 and a false positive rate of 0.293. It has a precision values of 0.756, Recall value with 0.743, F- Measure with

0.748 and ROC curve with 0.772. It took 2.16 secs to build the model and it classified 74.33% instances correctly. This is represented in Table 5.

Table 5. Multilayer Perceptron

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.775	0.325	0.839	0.836	0.805	0.772	Tneg
	0.675	0.225	0.632	0.583	0.626	0.772	Tpos
WA	0.743	0.293	0.756	0.743	0.748	0.772	

F. SVM

SVM in Table 6, results an average of true positive rate with 0.793 and a false positive rate of 0.334. It has a precision values of 0.787, Recall value with 0.793, F- Measure with 0.784 and ROC curve with 0.729. It took 0.41 secs to build the model and it classified 79.31% instances correctly.

Table 6. SVM

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.904	0.446	0.813	0.904	0.856	0.729	Tneg
	0.554	0.096	0.73	0.554	0.63	0.729	Tpos
WA	0.793	0.334	0.787	0.793	0.784	0.729	

G. KNN Classifier

In Table 7, KNN Classifier yields an average of true positive rate with 0.728 and a false positive rate of 0.345. It has a precision values of 0.731, Recall value with 0.728, F-Measure with 0.729 and ROC curve with 0.691. It took 0.01 secs to build the model and it classified 72.79% instances correctly.

Table 7. KNN

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.792	0.41	0.806	0.792	0.799	0.691	Tneg
	0.59	0.208	0.57	0.59	0.58	0.691	Tpos
WA	0.728	0.345	0.731	0.728	0.729	0.691	

H. RBF Network

RBF Network yields an average of true positive rate with 0.801 and a false positive rate of 0.273. It has a precision values of 0.798, Recall value with 0.801, F- Measure with 0.799 and ROC curve with 0.854. It took 1.02 secs to build the model and it classified 80.08% instances correctly. This is illustrated in Table 8.

Table 8. RBF Network

	TP Rate	FP Rate	P	RC	FM	ROC Area	Class
	0.865	0.337	0.846	0.865	0.856	0.854	Tneg
	0.663	0.135	0.696	0.663	0.679	0.854	Tpos
WA	0.801	0.273	0.798	0.801	0.799	0.854	

V. COMPARISON OF CLASSIFIERS

The dataset has been classified with a percentage split of 66 %. If we increase the percentage split we have a raise in the correctly classified instances whereas if we decrease it we have a drastic decrease in the instances. So it is better to stick on with the above said percentage to get the optimal result.

Table 9. Comparison of Classifiers

Algorithm	Time in (secs)	Correct Instances in %	KS	RMSE	RAE in %	RRSE in %
Naïve Bayes	0.09	77.02	0.4631	0.3822	58.97	81.64
J48	0.53	76.25	0.4342	0.4059	69.30	86.72
Random Forest	1.58	78.20	0.4746	0.3092	67.54	83.35
JRIP	0.27	77.01	0.4767	0.4121	79.35	88.04
MLP	2.16	74.33	0.4319	0.4445	70.65	94.95
SVM	0.41	79.31	0.4902	0.4549	45.87	97.17
KNN	0.01	72.79	0.3788	0.5205	60.51	111.20
RBF Network	1.02	80.08	0.5347	0.3785	67.35	80.85

By analyzing all the results from Table 9, according to the time taken to build the model, KNN classifier outperformed in 0.01 seconds and Multilayer Perceptron took about 2.16 seconds to build the same model. Whereas regarding the classification of instances RBF Network outperformed with 80.08%, followed by SVM with 79.31, Random forest with 78.02%. J48 classified the instances with 76.25%. Compared to all these classifiers, KNN classifier performance is quite least. When we observe the measures of the classifiers, if there is a rise in the value of kappa statistic, then there is also a possibility of high rate in the correctly classified instances. This has been visually represented in the below shown charts.

Figure 2 represents the time taken by each techniques to build the model. The chart represents that KNN has taken the least amount of time to build the model.

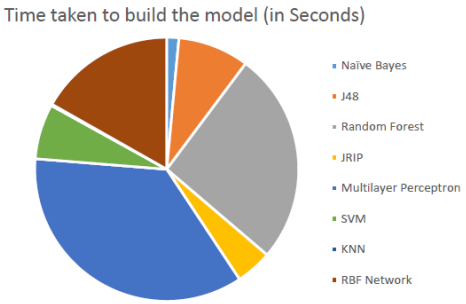

Figure 2. Time taken to build the model (in secs)

Figure 3 signifies the correctly classified instances, of all the techniques being used, RBF Network has the maximum number of correctly classified instances.

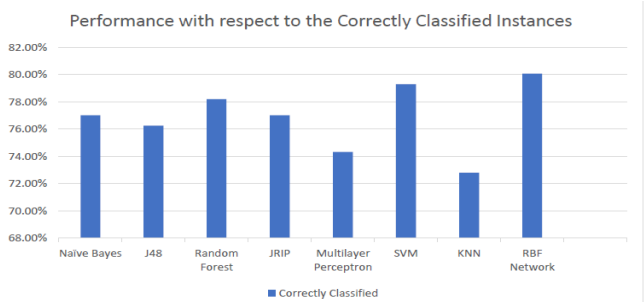

Figure 3. Performances of correctly classified Instances

Figure 4 illustrates the Kappa statistic value of the classification techniques.

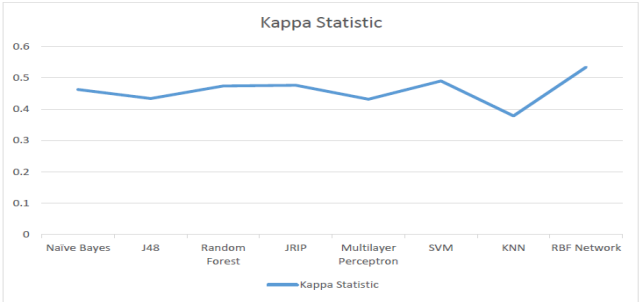

Figure 4. Kappa Statistics

With these results, it clearly states that the performance of an algorithm and the time required to build the model depends on the type of data we feed in as the input and selection of data mining approaches also depends on the nature of the dataset.

VI. Conclusion And Future Scope

The diabetes dataset consists of the labeled features. Various classification techniques have been analyzed and it is inferred that the classification techniques best suits for the prediction of results. Though various techniques have been used, the diagnosis of disease suffers false alarm and detection rate is low. This gives an insight to the researchers to propose a novel approach to reduce the false alarm rate in the situation of incomplete dataset handling.

References

[1] Harleen Kaur and Siri Krishan Wasan, "*Empirical Study on Applications of Data Mining Techniques in Healthcare*", Journal of Computer Science, Volume 2, Issue 2, 2006.

[2] V. Vapnik. *The Nature of Statistical Learning Theory*. NY: Springer-Verlag. 1995.

[3] K. Rajesh, V. Sangeetha, "*Application of Data Mining Methods and Techniques for Diabetes Diagnosis*", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[4] K. Saravanapriya, *A Study on Free Open Source Data mining Tools*", International Journal of Engineering and Computer Science, Volume 3, Issue 12, 2014.

[5] Russell, Stuart, Norvig, Peter, "Artificial Intelligence: A Modern Approach (2nd ed.)". Prentice Hall. 2003 ISBN 978-0137903955

[6] Ho, Tin Kam, "Random Decision Forests", Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282, 1995

[7] Wasserman, P.D.; Schwartz, T, "Neural networks. II. What are they and why is everybody so interested in them now?", pp: 10-15; IEEE Expert, Vol 3, Issue 1, 1988.

[8] Broomhead, D. S.; Lowe, David, "Radial basis functions, multi-variable functional interpolation and adaptive networks (Technical Report). RSRE 4148, 1988.

[9] Tejashri N. Giri, S.R. Todamal, "*Data mining Approach for Diagnosing Type 2 Diabetes*", International Journal of Science, Engineering and Technology", Vol 2(8), 2014.

[10] Dr. M. Renuka Devi, J. Maria Shyla, "*Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus*", International Journal of Applied Engineering Research ISSN 0973-4562 Vol 11, Number 1 (2016), pp 727-730.

[11] Pardha Repalli, "*Prediction on Diabetes Using Data mining Approach*", Oklahoma State University.

[12] R. Sukanya, K. Prabha,*"Comparative Analysis for Prediction of Rainfall using Data Mining Techniques with Artificial Neural Network",* Vol 5, Issue 6, pp 288 - 292, June 2017.

Authors Profile

Ms. K. Saravanapriya is currently pursuing her Ph.D. degree under the guidence of Dr. J. Bagyamani.in the Department of Computer Science, in Periyar University, Salem. Her research interests include Data Mining, Computer Graphics. She has 8 years of teaching experience and 2 years of Research Experience.

Dr. J. Bagyamani is an Associate Professor and Head in the Department of Computer Applications, Chikkanna Government Arts College, Tiruppur, Tamil Nadu, India. She has got 20 years of Teaching Experience and 7 years of Research experience. She has published 15 papers in reputed International Journals, and presented 8 Papers in National and International Conferences and Seminars. She has received Two Best paper awards in International Conferences. She has guided 10 Mphil reesearch scholars. Her areas of research and interests include Data mining, Bi-clustering of Gene Expression Data using Heuristic and Meta-heuristic techniques, Optimization algorithms, Web mining and Image Processing. She has guided 15 M.Phil. Scholars.

