International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

Comparative Analysis of Weighted Emphirical Optimization Algorithm and Lazy Classification Algorithms

P. Suganya[1], C. P. Sumathi[2]
[1] Ph.D Scholar, Bharthiar University, Coimbatore, Tamilnadu , India.
[2] Associate Prof. & Head, Dept. of Computer Sci, S.D.N.B Vaishnav College for Women, Chromepet, Chennai.
Email: suganyadgvc@gmail.com

Abstract- Health care has millions of centric data to discover the essential data is more important. In data mining the discovery of hidden information can be more innovative and useful for much necessity constraint in the field of forecasting, patient's behavior, executive information system, e-governance the data mining tools and technique play a vital role. In Parkinson health care domain the hidden concept predicts the possibility of likelihood of the disease and also ensures the important feature attribute. The explicit patterns are converted to implicit by applying various algorithms i.e., association, clustering, classification to arrive at the full potential of the medical data. In this research work Parkinson dataset have been used with different classifiers to estimate the accuracy, sensitivity, specificity, kappa and roc characteristics. The proposed weighted empirical optimization algorithm is compared with other classifiers to be efficient in terms of accuracy and other related measures. The proposed model exhibited utmost accuracy of 87.17% with a robust kappa statistics measurement and roc degree indicated the strong stability of the model when compared to other classifiers. The total penalty cost generated by the proposed model is less when compared with the penalty cost of other classifiers in addition to accuracy and other performance measures.

Keywords: Optimization, Parkinson, Classification, Discretization, accuracy, Kappa, Lazy Classifiers.

I. INTRODUCTION

.

Parkinsons Disease Dataset : Parkinson's disease is the mutual form of Parkinsonism meaning parkinsonism with no external recognizable cause. Generally classified as a nervous disorder, PD also gives upswing to several non-motor types of indications such as sensory scarcities, cognitive difficulties, and sleep problems. The following Table 1 is taken from UCI repository for related work and the attributes are described the Table.

Number of Instances	:	195
Attribute Characteristics	:	Real/Integer
Number of Attributes	:	23
Missing Values	:	Nil
Class1 (Parkinson Disease)	:	147
Class2 (without Parkinsons)	:	48

Table 1: UCI Attributes

Patient Details	Field Name	Description
Name	Name	ASCII subject name and recording number
Vocal Frequency	MDVP: Fo(Hz)	Average vocal fundamental frequency
	MDVP: Fhi(Hz)	Maximum vocal fundamental frequency
	MDVP: Flo(Hz)	Minimum vocal fundamental frequency
Fundamental Frequency	MDVP: Jitter(%),	Several measures of variation
	MDVP: Jiter(Abs)	
	MDVP: RAP	
	MDVP: PPQ	
	Jitter: DDP	
Amplitude	MDVP: Shimmer	Several measures of variation
	MDVP: Shimmer(dB)	
	Shimmer: APQ3	
	Shimmer: APQ5	
	MDVP: APQ	
	Shimmer: DDA	
Ratio of noise	NHR	Two measures of tonal components in the voice
	HNR	
Class	Status	Health status of the subject (Healthy-C0 / Parkinson's - C1)
Complexity	RPDE	Two nonlinear dynamical measures
	D2	
Signal	DFA	Signal fractal scaling exponent
Nonlinear Measures	spread1	Three fundamental frequency variation
	spread2	
	PPE	

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

II. PROPOSED WORK

The proposed research work focuses on the dataset to analyse the performance measure of the Weighted Emphirical Optimization classifier with various classifiers such as Bayesnet, NaiveBayes, zeroR, oneR . Parkinson dataset is normalized, discretized for improvement handling of the data to the various classifiers. The data flow diagram Fig 1 indicates the flow of the research work.The performance measure such as accuracy, sensitivity, specificity, precision, kappa and roc are calculated using confusion matrix. The data flow diagram Fig 1 indicates the flow of the research work.

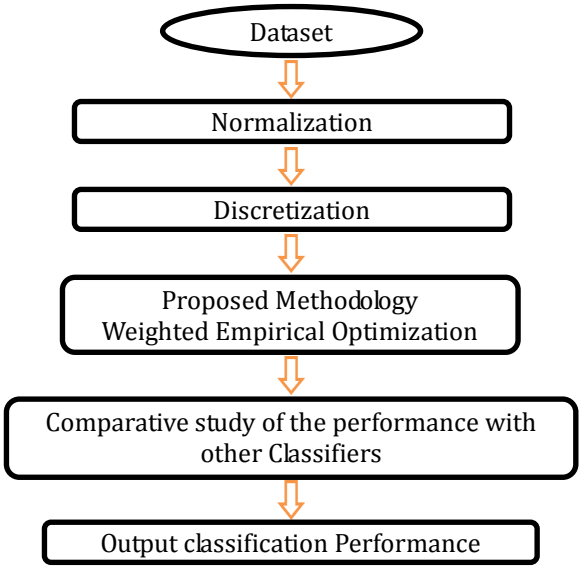


Figure 1: Architecture

A. Preprocessing

Data is prepared for effective processing using preprocessing techniques such as normalization, discretization, outlier removal etc. Removing noise and organizing the data for efficient access to the other context is generally done as the first stage. Preprocessing is required when the dataset consists of meaningless data that is incomplete (missing), noisy (outliers) and variance data. Divya Tomar and Sonali Agarwal [22] considered various methods such as filter, imputation and embedded techniques to switch missing features problem. In Filter technique discard or remove missing features from the dataset while assertion based method and replace the missing features by suitable value. Imbalanced dataset are tackled either by sampling or algorithm alteration method. The preprocessing includes four steps data cleaning, data integration, data transformation, data reduction. In this research the Parkinson dataset is preprocessed with weka using the filter option either supervised unsupervised.

B. Normalization

YogendraKumar Jain and Santhosh Kumar Bhandare23 suggested a matrix of dimension q x p, which is the original dataset. The rows of the matrix denote objects and the column of the matrix denote features. The original data matrix of size q x p, must be first converted by min max normalization technique to be reformed matrix *M* whose size has the same size q x p as the unique data matrix. The min max normalization technique makes the original data matrix M into the detailed range between 0.0 and 1.0. After applying the min max technique on the original data elements has been disconcerted into the small detailed range between 0.0 and 1.0, min max normalization technique is topped on each element of the unique data matrix M into a detailed range such as [0.0,1.0]. The data has to be normalised when seeking for relations.

Min-Max Normalization - This is a simple normalization technique in which we fit the data, in a pre-defined boundary, or to be more specific, a pre-defined interval [C, D]. Formula

$$B = \left(\frac{(A - \text{Mini value of } A)}{\text{Max value of } A - \text{Min value of } A} \right) * (D - C) + C \quad \text{------(1)}$$

C. Discretization

Discretization is the process of conversion of numeric data into nominal data by changing numeric values into distinct sets, in which length is fixed. Jiawei Han, Micheline Kamber, Jian Pei [17] indicate automation generation and encoding techniques in concept hierarchy can be done using discretization the binning is used as a top down technique for data smoothing .The data which are discontinuous is applied with equal frequency binning or equal width and replacing each bin with bin mean or median . The discretization comes under unsupervised technique and it is used for class target attribute in easier way. In weka, the discretization is through equal width binning, where binning divides the scope of possible values into N subscopes (bins) of the same width:

$$\text{Width} = (\max \text{value} - \min \text{value}) / N \quad \text{----- (2)}$$

D. Classification

Classification is a problem where the relations among the attributes or features and the class variable is learnt to test for new data i.e. unlabeled test instance. Suganya[18] evaluated the association with various analyzation with many fields and especially clinical data is most appropriate. With different variation in the features the model is constructed as training phase and when unlabeled test is tried the model with classification predicts accurately. Even in such belongings, a pre-processing phase like nearest neighbor table construction may be accomplished in order to ensure efficacy during the testing stage. The output of a classification algorithm may be presented for a test instance in one of two ways:

Discrete Label: In this case, a label is returned for the test instance.

Numerical Score: In this case, a numerical score is returned for each class label and test in- stance combination. Note that the numerical score can be converted to a discrete label for a test instance, by picking the class with the highest score for that test instance. The advantage of a numerical score is that it now becomes possible to compare the relative propensity of different test instances to belong to a particular class of importance, and rank them if needed. Such methods are used often in rare class detection problems, where the original class distribution is highly imbalanced, and the discovery of some

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

classes is more valuable than others. When there is no prior knowledge about the class target category the problem is clustering and in supervised learning i.e. the classification is clearly known and it result in some area of interest. With classification the splitting up is done on the basis of training data set and it encodes the knowledge it has learnt in groups to predict the class target category.

III. WEIGHTED EMPHIRICAL OPTIMIZATION ALGORITHM

The best optimized data is achieved through various fitness value and the error matrix or confusion matrix is generated .This result is compared with the various performance measures of various classifiers.

Step 1. To accumulate $\sum_{1}^{n} f(x)$ no of various frequency data.

Step 2. Evaluate the data with weightage constrains [Decide on which constraint is needed and evaluate it from the rest of it]

Step 3. To arrive at the optimized data regarding the value of fitness for each weightage constraints. (i.e : $f(x)$)

Step 4. To remember the best optimized data through its fitness value and store it in the given $WEO(x)$.

Step 5. Repeat the Step 3 and 4 again, until the data regarding the $f(n)$ is complete.

Step 6. Exchange the data of weightage in the given $WEO(x)$ to determine the optimal decision making.

A. Lazy Classifier

Lazy classifiers works on classification time which find the training instance closest in Eucidean distance to the given test data and it predicts the outcome of the classifier. If more than one are predicted closely than the first data is taken for more appropriation. The[23] performance of various lazy classifiers with Bayesian classifier has been performed with experimental result analyzing the efficacy of lazy classifiers.

B. IBK Classifier

IBK is a k nearest neighbor classifier where different search algorithms like linear search, KD trees, ball tress, cover trees are used for the finding the nearest neighbors. Euclidean distance is the function used as a parameter function. The other measures are chebyshev, manhattan and minkowski distances. When k=1 the leave one out cross validation for classification is performed.

C. KStar

This method is a generalized distance function based on transformations

1) Transform the instance or attributes through sequence of predefined elemenatary operations
2) Find the probability and randomly choose the attribute
3) The k nearest neighbor rule is applied which decides either in or out of the deciding class.
4) Nominal and numeric values can be transformed in similar manner by defining different transformation sets.

D. LBR Classifier

Lazy Bayesian Rule belonging to Bayesian classifier that process in classification time. The attributes are selected based on independence assumptions. The attributes are discretized before training of the classifier.

IV. EXPERIMENTAL RESULTS

The confusion matrix is tabulated for every classifier with actual and predicted class which is denoted for binary classification. Either supervised or unsupervised the table generalizes the True Positive (TP), False Positive (FP), False Negative(FN) and True Negative(TN) for various calculation of performance for the classifiers.

Table 2: Confusion Matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive(TP)	False Positive(FP)
	No	False Negative(FN)	True Negative(TN)

From Table 2 the various performance measures are listed with formula.

Accuracy : Overall Effectiveness of a classifer

$$TP = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{-------- (3)}$$

Precision : Class agreement of the data labels with the positive labels given by the classifier

$$TP = \frac{TP}{TP + FP} \quad \text{-------- (4)}$$

Recall (Sensitivity): Effectiveness of a classifier to identify positive labels

$$TPR = \frac{TP}{TP + FN} \quad \text{-------- (5)}$$

Specificity : Effectiveness of a classifier to identify negative labels

$$Specificity = \frac{TN}{TN + FP} \quad \text{-------- (6)}$$

FMeasure : Harmonic mean of precision and recall

$$Fmeasure = \frac{2*(precision*recall)}{(precision + recall)} \quad \text{-------- (7)}$$

Kappa :

$$Kappa = \frac{Total\ Accuracy - Random\ Accuracy}{(1 - RandomAccuracy)} \quad \text{-- (8)}$$

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

Type I Error: Inappropriate Elimination of true null hypothesis i.e., False Positive Error is Type I Error. Type I error is finding an effect that is present.

Type II Error: The failure to reject a false null hypothesis False Negative Error. Failing to detect an effect that is present.

Using Weka software the Parkinsons dataset is preprocessed and the following Table 3 is generated by means of various classifiers. The proposed model Weighted Emphirical Optimisation is carried out with Java Software and its confusion matrix is also summarized.

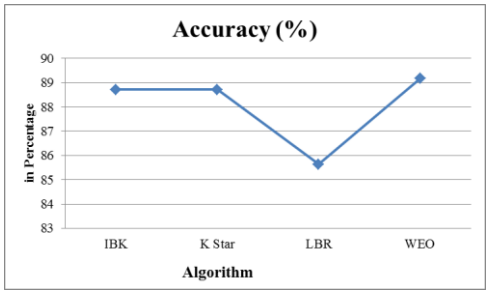
Table 3: Confusion Matrix for various lazy classifiers

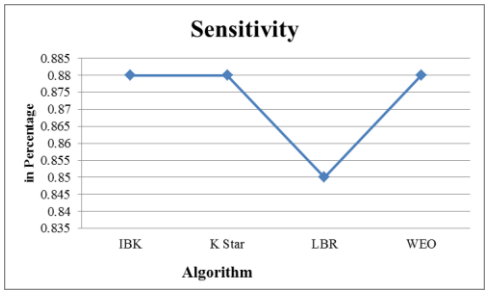
Algorithm	Confusion Matrix			
IBK		**C0**	**C1**	**Sum**
	C0	133	14	147
	C1	8	40	48
KStar		**C0**	**C1**	**Sum**
	C0	133	14	147
	C1	8	40	48
LBR		**C0**	**C1**	**Sum**
	C0	139	8	147
	C1	20	28	48
WEO		**C0**	**C1**	**Sum**
	C0	132	15	147
	C1	8	40	48

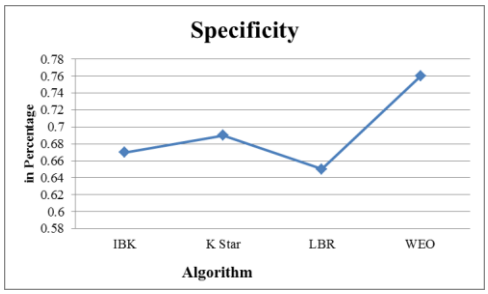
C0 – Indicates Parkinson Disease
C1- Healthy without Parkinson

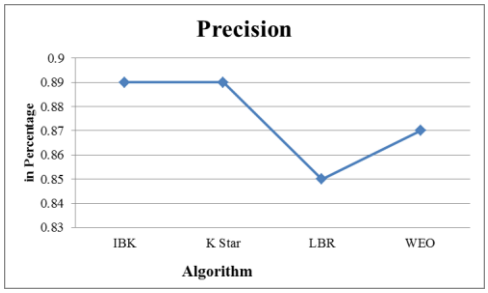
Table 4: Comparative Study of various classifiers with their performance measures

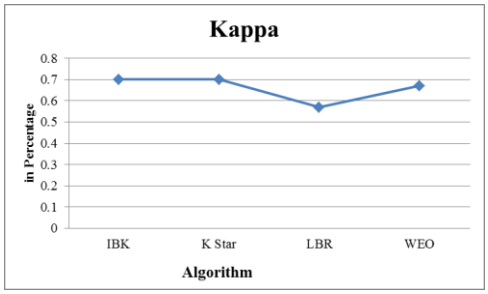
Performance Measure	IBK	Kstar	LBR	**WEO**
Accuracy (%)	88.71	88.71	85.64	**89.17**
Sensitivity	0.88	0.88	0.85	**0.88**
Specificity	0.67	0.69	0.65	**0.76**
Precision	0.89	0.89	0.85	**0.87**
Kappa	0.70	0.70	0.57	**0.67**
ROC	0.90	0.92	0.84	**0.97**
FMeasure	0.88	0.89	0.64	**0.87**

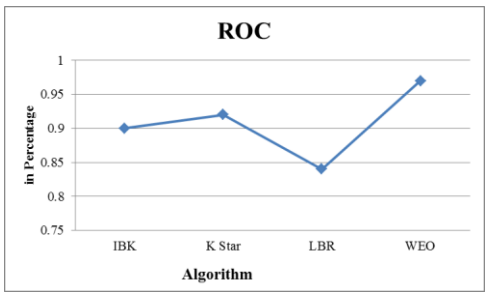












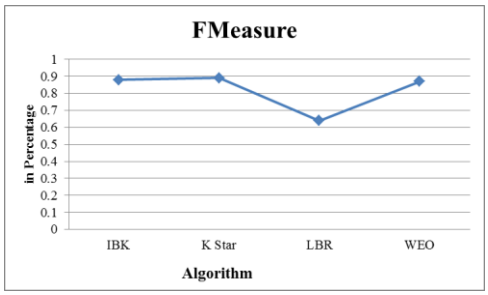


Figure 1: Graph of various classifiers with their performance measures

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

From the above Table 4, the various classifiers are compared using confusion matrix and the accuracy of proposed model showed an accuracy of 89.17% which is relatively higher than the other classifiers. Kappa statistics of the proposed model was much better for imbalanced dataset and the other algorithms. The stability of the model is indicated by the roc curve with 0.97 measurements for the new proposed model shows a good sign with concern to other classifiers. The various measurements are depicted using Figure 1 graph.

V. CONCLUSION

The proposed model exhibited utmost accuracy of 89.17% with a robust kappa statistics measurement and roc degree indicated the strong stability of the model when compared to other classifiers. In medical dataset for classification problem type II error is significant than type I error. Penalty cost for misclassifying an instance with respect to type I error is Rs. X and with respect to type II is Rs.(2*X) . The objective of the proposed work is to minimize False Negative than False Positive. In addition to accuracy the proposed method also considers measures like sensitivity, specificity, precision, kappa, roc and Fmeasure. Since, the dataset is imbalanced where the class C0 represents the person with Parkinson disease and the class C1 represents the person who are healthy as(147,48) the scope of the future work is to remove noise and to handle the imbalanced dataset to use ensemble classifiers to improve the performance measures. Among the algorithms Weighted Emphirical Optimisation algorithm performance is much better than other algorithms on accuracy, sensitivity, specificity, precision, kappa, roc and Fmeasure . On the contrary lazy classifiers produced lesser percentage of misclassified instances than the other algorithms of which False Negative is lesser than False Positive and the model proposed in this work results in a lesser penalty cost than other classifiers which is important for clinical data. In this article the data is imbalanced which has to be balanced to carry out feature selection in future work and also to compare with more classifiers to exhibit the concert of the proposed model.

References

[1] Ahlrichs C, Lawo M. Parkinson's Disease Motor Symptoms in Machine Learning: A Review. 2013. http://arxiv.org/abs/1312.3825.

[2] Genain N, Huberth M, Vidyashankar R. Predicting Parkinson ' S Disease Severity from Patient Voice Features.; 2014. http://roshanvid.com/stuff/parkinsons.pdf.

[3] Mathers C, Fat DM, Boerma JT, Organization WH. The Global Burden of Disease : 2004 Update. Geneva, Switzerland: World Health Organization; 2008.

[4] Chen L. Computer-Aided Detection of Parkinson's Disease Using Transcranial Sonography. In: Dissertation for Fulfillment of Requirements for the Doctoral Degree of the University of Lübeck from the Departments Information Technology. Vol Germany: University of Lübeck; 2013:128. http://www.students.informatik.uni-luebeck.de/zhb/ediss1324.pdf.

[5] Sateesh Babu G, Suresh S. Parkinson's disease prediction using gene expression – A projection based learning meta-cognitive neural classifier approach. Expert Syst Appl. 2013;40(5):1519-1529. doi:10.1016/j.eswa.2012.08.070.

[6] Bind S, Tiwari AK, Sahani AK. A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction. International Journal Computer Science Information Technology. 2015;6(2):1648-1655. http://www.ijcsit.com/docs/Volume 6/vol6issue02/ijcsit20150602163.pdf.

[7] Holkar P, Gatti P, Meher S, Sable P. A Review on Parkinson Disease Classifier Using Patient Voice Features. International Journal Advance Research Electrical Electronics Instrumentation Engineering. 2015;4(5):3827-3830. http://www.ijareeie.com/upload/2015/may/5_A_review.pdf.

[8] Rustempasic I, Can M. Diagnosis of Parkinson ' s Disease using Principal Component Analysis and Boosting Committee Machines. SouthEast European Journal of Soft Computation. 2013;2(1):102-109. http://www.ius.edu.ba/sites/default/files/articles/51-145-1-PB.pdf.

[9] Khemphila A, Boonjing V. Parkinsons Disease Classification using Neural Network and Feature selection. International School of Science Research Innovation. 2012;6(4):15-18. http://waset.org/publications/8538/parkinsons-disease-classification-using-neural-network-and-feature-selection.

[10] Sakar BE, Kursun O. Telemonitoring of changes of unified Parkinson ' s disease rating scale using severity of voice symptoms. In: The 2nd International Conference on E-Health and Telemedicine. Vol Diagonastic Pathology; 2014:114-119.

[11] Armañanzas R, Bielza C, Chaudhuri KR, Martinez-Martin P, Larrañaga P. Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. Artificial Intelligence Medicine. 2013;58(3):195-202. doi:10.1016/j.artmed.2013.04.002.

[12] Sriram TV., Rao MV, Narayana GVS, Kaladhar D, Vital TPR. Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms. In: International Journal of Engineering and Innovative Technology (IJEIT). Vol 3. ; 2013.

[13] Khan SU. Classification of Parkinson's Disease Using Data Mining Techniques. J Park Dis Alzheimer's Dis. 2015;2(1):1-4.

[14] Ramani RG, Sivagami G, Jacob SG. Feature Relevance Analysis and Classification of Parkinson Disease Tele-Monitoring Data Through Data Mining Techniques. International Journal of Advance Research Computation Science Software Engineering 2012;2(3):298-304. http://www.ijarcsse.com/docs/papers/March2012/volume_2_Issue_3/V2I300135.pdf.

[15] Rustempasic I, Can M. Diagnosis of Parkinson's Disease using Fuzzy C - Means Clustering and Pattern Recognition. International University Sarajev , Fac Eng Nat Sci. 2013.

[16] Kaladhar DSVGK, Rao, Nageswara P, Rajana Blvrn. Confusion Matrix Analysis For Evaluation Of Speech On Parkinson Disease Using Weka And Matlab. International Journal of Engineering Science Technology 2010;2(7):2734-2737.

[17] Jiawei Han, Micheline Kamber, Jian Pei , Data Mining Book, Southeast Asia Edition.

[18] P.Suganya and C.P.Sumathi, A Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease, Indian journal of Science & Technology, vol 8, issue 14, July 2015.

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.125-130
ISSN: 2278-2419

[19] Divya Tomar and Sonali Agarwal, International Theory of Database Theory and Automation,Vol.7, No.4 (2014), pp. 99–128, A Survey on Pre-processing and Post-processing Techniques in Data Mining.
[20] YogendraKumar Jain and Santhosh Kumar Bhandare, International Journal of Computer & Communication Technology, Volume 2, Issue VIII, 2011, Min Max Normalization Based Data Perturbation Method for Privacy Protection.
[21] Mike A Nalls and Cory Y MCLean, The Lancet Neurology, Vol 14, Issue 10, October 2015, Pages 1002 -1009, Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population based modelling study.
[22] Roberto Erro and Gabriella Santangelo, Journal of Parkinsonism and Restless legs syndrome,vol 5, Nonmotor symtoms in parkinson's disease: Classification and management.
[23] Vijayarani, S., and M. Muthulakshmi. "Comparative analysis of bayes and lazy classification algorithms." International Journal of Advanced Research in Computer and Communication Engineering 2.8 (2013): 3118-3124.