A Comparative study on various Classification algorithms for flying insect classification based on their spatio-temporal features

S.Arif Abdul Rahuman[1]**, Dr. J. Veerappan**[2]

[1]*Prof/CSE,Universal college of Engineering & Technology, Tamilnadu, India.* [1]jnellai@gmail.com

[2]*Prof/ECE Vel Tech High Tech* Dr. Rangarajan Dr.Sakunthala *Engineering College, Tamilnadu, India*

Abstract - Mining insects' sound is an open research issue. Nowadays some insects are harmful to mankind and some are beneficiary. The main objective of such mining is to determine and quantify the insects' density available in sensible regions of a city. Recognition of insects based on their sound containing spatio-temporal data is also a hard problem. Various classification methods on such time-series data are available. Identifying the best classification algorithms among all existing methods is a challenging task. This paper presents a performance based comparative study of the most widely used classification algorithms. Moreover, the performance of these algorithms have been analyzed by using same data sets. The data relevant to the flying insects often changes over time, and classification of such data is a challenging task. The proposed framework of recognizing insects uses Support Vector Machine (SVM) classifier and its performance is compared against statistical classifiers like Bayesian classifier, k- Nearest Neighbor classifier, and Fisher Linear Discriminant Classifier and Soft computing classifiers such as Fuzzy classifier and neural network classifiers. These algorithms are compared for their performance with the same dataset. And the paper concludes that SVM classifier outperforms other classifiers in recognizing the insects based on their sound.

Keywords - Support Vector Machine, Bayesian Classification; k-Nearest Neighbor, Fisher Linear Discriminant Classifier, Fuzzy classifier, Neural network classifier.

I. INTRODUCTION

Data mining is the technology that uses evaluation techniques such as statistics, machine learning and pattern recognition in order to perform analysis on huge volume of data or databases. Nowadays there is enormous volume of data available everywhere across the globe. The survey shows that volume of data keeps on increasing year after year. And databases with Terabytes of data in enterprises and research facilities are available throughout the world. The database contains invaluable information and hidden knowledge and there is no automatic method for extracting relevant information; that is, it is practically impossible to mine them.

Automated tools are needed nowadays for the knowledge discovery in databases to control the flood of data that depends on ever-growing databases in each and every field. This paper focus only on the database containing sound of insects, and such data are time variant. Knowledge discovery data (KDD) has the preprocessing, data mining and post processing phases. KDD is the iterative or cyclic process that involves sequence of steps of processes and data mining is the core component of the KDD process. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. Such discovered patterns using data mining techniques are valuable for the process of enterprise's decision making.

The literature survey shows that several algorithms exist to discover knowledge from large datasets. And several methodologies also exist in such algorithms and one among them is the classification. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Popular classification techniques include Bayes classification, k-NN classifier, Fisher Linear Discriminant classifier, Fuzzy classifier, Neural network classifier etc. It uses a training set of data that contains observations to identify which categories each observation should be placed in. And then the testing phase that classifies the data. There are a lot of open issues related to machine learning and statistics-based techniques related to classification.

DOI : 10.18231/2454-9150.2018.1224

In this paper, we focus on the statistical classification methods and soft computing methods of classification, and their performances are evaluated by implementing those algorithms with same dataset. Performance evaluation of classification model is important for understanding the quality of the model, to refine the model, and for choosing the appropriate model with respect to the dataset. The performance evaluation criteria used in classification models include the classification accuracy based on true positive, true negative, false positive, and false negative classification.

The paper is organized in such a way that section 2 deals with the survey on several existing papers about classification, section 3 contains classification algorithms, section 4 contains the performance evaluation and section 5 concludes the performance of the best suited classification algorithm with respect to the benchmark and synthetic dataset and deals with the future scope of study.

II. Related Work

Skyler Seto, Wenyu Zhang, Yichen Zhou on “Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition.”[1] used a classifier based on DTW to classify human activity. It used a template selection approach instead of feature extraction. And estimated time series similarity measure. The implementation of this method shows an increased computational cost.

Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, Eamonn Keogh, “Flying Insect Classification with Inexpensive Sensors.”[2] used a classifier that classifies flying insects based on Circadian Rhythm and Insect Geographic Distribution using Bayesian Classifier. And this paper used Euclidean distance to measure the nearest neighbor. And with the increase in number of species there is an increasing difficulty in classification.

Stephan Spiegel, Brijnesh-Johannes Jain, SahinAlbayrak, “Fast Time Series Classification under Lucky Time Warping Distance.”[3] classifies time-series data using DTW. The distance measure used to estimate similarity measure is Lucky Time Warping (LTW). This method works faster when compared to DTW distance measure and the time and Space complexity is linear.

Begum N., B. Hu, T. Rakthanmanon, E. Keogh, “A Minimum Description Length Technique for Semi-Supervised Time Series Classification.”[4] used a Semi-supervised learning (SSL on time-series data. In self-training, a classifier is first trained with a small number of labeled data then classifies the unlabeled data, and adds the most confidently classified object into the labeled set. It provided with limited data for learning and no stopping criteria

AlemGebru, Erich Rohwer, Pieter Neethling, MikkelBrydegaard, “Investigation of atmospheric insect wing-beat frequencies and iridescence features using a multispectral kHz remote detection system.” [5] performed a quantitative analysis on wing-beat frequencies and iridescence features of insects and produces good classification accuracy using Bayesian classifier.

Theodoros Damoulas, Samuel Henry, Andrew Farnsworth, “Bayesian Classification of Flight Calls with a novel Dynamic Time Warping Kernel.”[6] used a Probabilistic Learning is used for classification with temporal features. And in order to improve the classification accuracy spatio-temporal information has to be integrated into the model.

Tuomas Virtanen and Marko Helen, “Probabilistic Model Based Similarity Measures For Audio Query-by-example.”[7]used Probabilistic models to estimate feature distribution using HMM likelihood test with best accuracy.

Frick.T.B, Tallamy.D.W, “Density and diversity of non-target insects killed by suburban electric insect traps” [8] in a survey of insects presented that by the use of electric insect traps can reveal only 31 biting flies, a minute proportion (0.22%) of the 13,789 total insects counted. And more than 104 nontarget insect families were destroyed.

Hao.Y, Campana.B and Keogh.E,, “Monitoring and Mining Animal Sounds in Visual Space” [9] used a novel bioacoustic classification framework to recognize animal sounds in the visual space using the texture of their sonograms with high speed and accuracy.

Yakun Hu, Dapeng Wu, and Antonio Nucci, “Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification” [10] used Fuzzy-clusterin model to identify the speaker among large population using pitch and five vocal source features.

IlyasPotamitis and IraklisRigakis, “Novel Noise-Robust Optoacoustic Sensors to Identify Insects Through Wingbeats” [11] is a hardware-based implementation to identify insects using their wingbeat frequency. Subsequently it can count the number of insects flying through the optoacoustic sensors.

Nguyen.M.N, Li.X.L,, “Ensemble Based Positive Unlabeled Learning for Time Series Classification.” [12] used integrate multiple PU learning classifiers for disease gene predictions with high accuracy and robustness. It integrating multiple biological data sources for training and the outputs of an ensemble of PU learning classifiers for prediction.

Skyler Seto, Wenyu Zhang, Yichen Zhou, “Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition” [13] used a template selection approach based on Dynamic Time Warping that extracts complex features and recognizes

DOI : 10.18231/2454-9150.2018.1224

human activity on real smartphone data with good accuracy.

AlemGebru, Erich Rohwer, Pieter Neethling, MikkelBrydegaard, "Investigation of atmospheric insect wing-beat frequencies and iridescence features using a multispectral kHz remote detection system" [14] has used remote optical classification of insects based on wing-beat frequencies and iridescence features for flight direction of an atmospheric insect.

Begum N., B. Hu, T. Rakthanmanon, E. Keogh, "A Minimum Description Length Technique for Semi-Supervised Time Series Classification" [15] has used small set of human annotated examples for the classification on medical data sources such as electrocardiograms and used a novel parameter-free stopping criterion for semi-supervised learning.

Stephan Spiegel, Brijnesh-Johannes Jain, SahinAlbayrak, "Fast Time Series Classification under Lucky Time Warping Distance." [16] used Lucky Time Warping (LTW) distance for nearest neighbor classification by considering time and space complexity.

Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, Eamonn Keogh, "Flying Insect Classification with Inexpensive Sensors" [17] has used pseudo-acoustic optical sensors to exploit features including intrinsic and extrinsic to the insect's flight behavior, and Bayesian classification approach efficiently classify insect behavior and are robust to overfitting.

TheodorosDamoulas, Samuel Henry, Andrew Farnsworth, "Bayesian Classification of Flight Calls with a novel Dynamic Time Warping Kernel" [18] has used Dynamic Time Warping to detect the flight calls of the birds by extracting features like energy value of the call, maximum amplitude, number of peaks, length of the flight call etc.

Tuomas Virtanen and Marko Helen," Probabilistic Model Based Similarity Measures For Audio Query-by-example" [19] specified that the acoustic features are mostly of short-time spectrum. Temporal features are obtained using temporal derivatives using Hidden Markov Model and statistical properties with frame-wise features.

Kaushik H. Raviya et al., [20] presents the comparison on three classification techniques which are K-nearest neighbour, Bayesian network and Decision tree respectively. The objective of this research is to enumerate the best technique from the above three techniques. The evaluation is performed on the three techniques on several datasets and their accuracy and time of execution is considered for the analysis.

III. CLASSIFICATION ALGORITHMS

The overall system architecture of the proposed work is shown in the Figure3.1

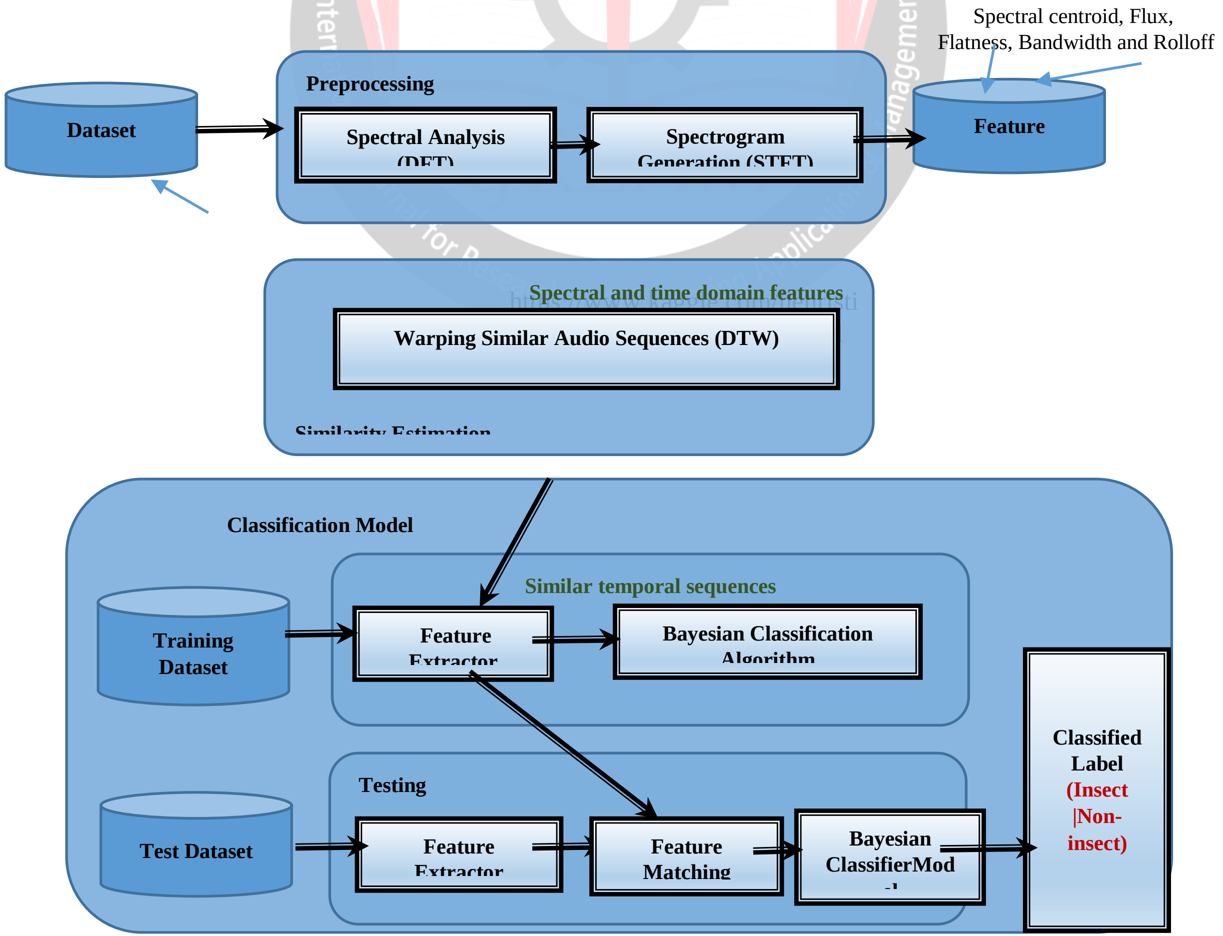

Figure 3.1 Overall system architecture

DOI : 10.18231/2454-9150.2018.1224

Computers still fail badly, compared to humans, while interpreting and recognizing the acoustical sounds. With the selection of correct features and by using appropriate classifier, it is believed that high accuracy of recognition is possible. For the experimental evaluation, one hundred and fifty audio files comprising of insects' sound and another one hundred and fifty audio files with non-insects sound are used as the training data. And another one hundred and fifty files are tested for insect|non-insect classification. All these files are from benchmark dataset (ESC-50, MOSQUITO) and synthetic dataset (Kaggle uploaded).

The feature vectors such as Short Time Energy, Zero Crossing Rate, Spectral Flux, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, and Spectral Flatness are extracted and used from the raw input files (flying insects' sounds) for classification.

A. *Support Vector Machine*

The Support Vector Machine is a binary classifier that finds a maximal margin separating hyper plane between the two classes (insect | non-insect). SVM maximizes the distance to the closest point from each class. SVM is best suited for future data. SVM uses linear kernel function with two input for each training data (data point of feature vector and class label) that provides good separability. SVM classifies the test data faster with good accuracy [2] and the algorithm is stable when compared to other statistical classifiers.

The accuracy produced by the SVM using the benchmark dataset[2] is high (77.94%) when compared with other statistical classifiers . Hence, the proposed framework is also implemented with SVM classifier to recognize the insects based on their sound.

The accuracy produced for recognizing insects on the benchmark dataset, ESC-50 is 85.78% and that with kaggle-uploaded dataset is 86.75%. The recognition result shows that SVM outperforms other statistical classifiers.

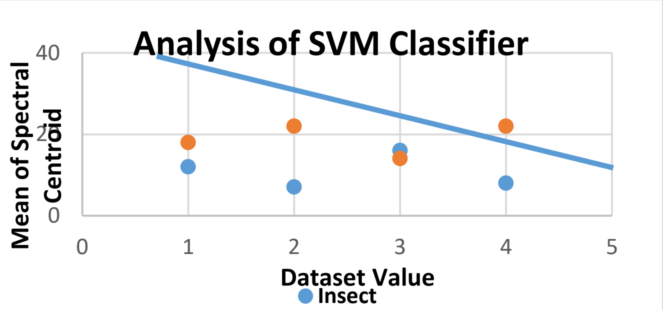

Figure 3.2 Hyperplane generation using SVM for classification

B. *K-Nearest Neighbor Classifier*

The k-NN approach is too simple. This approach directly uses the training data to estimate the probability of observing an unknown data in a class. It first searches the training data and finds the top k nearest neighbors of a class, insect | non-insect. It then computes the probability of observing insect | non-insect class. The training phase stores the feature vectors and class labels of the training dataset. After finding the closest distance pair among the test data and trained data, the class labels are assigned accordingly as insect or non-insect. Classification becomes more difficult as the dimensionality of the feature space increases and hence feature selection is subsequently done.

The classification phase of KNN measures distance between test data and all samples in the training set. Euclidean distance measure is used between the selected features of trained and test dataset. Then it identifies the k nearest neighbors and assigns the class label to the nearest test dataset that matches with the trained dataset.

C. *Fisher Linear Discriminant Classifier*

The Fisher linear discriminant only performs the classification between two classes. The data are projected onto a line, and the classification is performed in this one-dimensional space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class.

The class separability function in a direction, allows for calculating the optimal projection that ensures that the samples belong to each one of the two classes and can be as separated as possible.

D. *Bayesian Classification Algorithm*

Bayesian classifier works on training data set. It can be used to automatically classify all frames in a set of test files as Insect (Class 1), Non-Insect (Class 2). A set of training data i.e. audio files are provided as input and with these data sets probabilities (prior) of all classes with all the conditions were calculated. And the evaluations were stored. As a subsequent process, posterior probability is evaluated with test data.

This classification is iteratively done by choosing test files with the same statistics as that of the training set, finding the posterior probability on test data based on prior probability of trained data and classifying each frame of each test file as belonging to one of the two classes. This work computes Euclidean distance between the test frame feature vector and the sets of class means and variances to determine the similarity in classification of a class. And then class labels are assigned as Insect /Non-Insect. The sound outside is ignored using wingbeat frequency to improve the classification accuracy. Bayes classifier requires a small amount of training data to estimate the variable values necessary for classification.

E. Fuzzy *Classifier*

Fuzzy classifier transforms quantitative data into fuzzy data through identification of suitable membership function (mean and standard deviation). Thereby it is easy to generate IF-THEN rules to form equivalence classes of interest (insect | non-insect) by using the membership function on the quantitative assessments of seven spatio-temporal features of the audio input (insects sound). The classification is close to the correct solution.

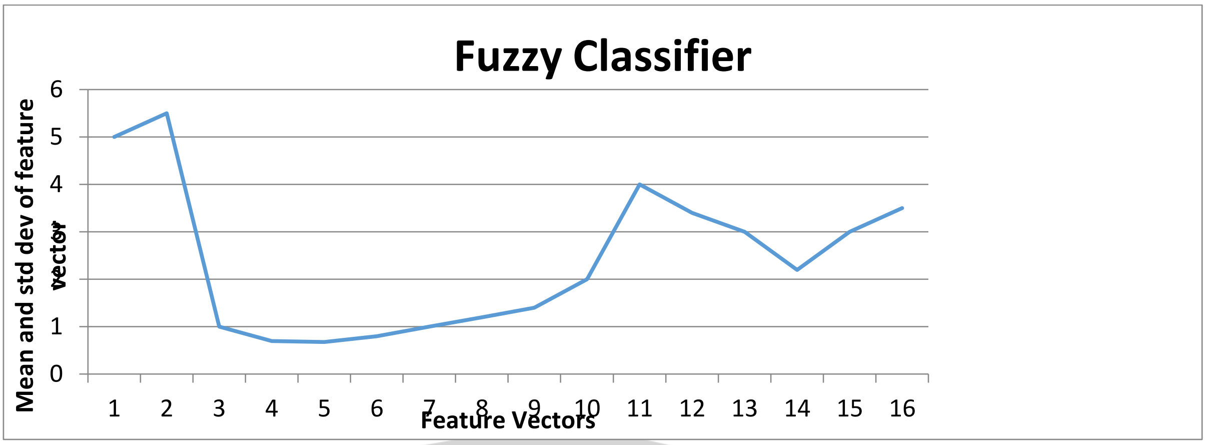

Figure 3.3 Mean and Standard deviation of attribute values using Fuzzy classification

F. Neural Network Classifier

Neural Network classifier uses the Back-propagation algorithm for classification. The algorithm proceeds in five steps. (i) It initializes the network with 7 neurons (feature vectors) in the input layer, 10 neurons in the hidden layer, 2 neurons for the output layer. Each neuron in the input layer is assigned with a set of weights (mean of each feature) for connecting the neuron to the hidden layer neurons. (ii) Neurons are propagated forward by assigning weights (iii) Estimates the error using back propagation (iv) trains the network (v) predicts the output layer.

4. Performance Analysis

The performance analysis of the statistical classifiers and soft-computing techniques applied classifiers on benchmark dataset (ESC-50, MOSQUITO) and uploaded dataset (kaggle) is listed in Table 4.3

Dataset Algorithm	Performance Measures	ESC-50	Kaggle	MOSQUITO
SVM (%)	Accuracy	85.78	86.75	77.94
	Precision	85.72	86.57	77.87
	Recall	85.82	86.71	77.95
k-NN (%)	Accuracy	80.88	83.82	71.56
	Precision	80.73	83.87	71.60
	Recall	81.14	85.71	71.58
Bayesian (%)	Accuracy	85.29	86.09	76.96
	Precision	85.42	86.08	76.89
	Recall	85.56	86.13	76.97
FLDA (%)	Accuracy	77.94	81.50	73.91
	Precision	78.14	83.15	73.17
	Recall	78.50	81.32	74.31
Neural Network (%)	Accuracy	78.431	80.347	73.16
	Precision	77.15	79.871	74.37
	Recall	78.441	81.218	74.36
Fuzzy (%)	Accuracy	74.51	77.457	75.25
	Precision	74.723	77.486	75.56
	Recall	80.63	77.419	74.90

Table 4.3 Performance analysis of various classifier

The performance analysis shows that SVM classifier recognizes the insects with the highest accuracy concerned with other statistical classifiers. SVM is the most appropriate classifier for high dimensional and time series data (acoustical sounds). KNN classifier is simple and not able to compute the distance among the high dimensional feature vectors with good precision. Bayesian is a binary classifier suitable for acoustical sounds, but the noise reduces the classification accuracy. The predicted probability using FLDA leads to misclassification. Soft computing techniques shows good accuracy of recognition of insects and specifically Neural network classifier, a slow classifier shows high accuracy yet Fuzzy classifier becomes cumbersome with high dimensional feature vectors.

IV. Results and Discussions

Data were collected during a period of 15 days from several regions of Tirunelveli district for several species like bees, mosquitoes, beetle, bumble bees, cockroach, etc. The data was collected in a temperature ranged from 70.2◦F up to 75.3◦F, and humidity ranged from 50% to 70%. Most of the data were collected in 12-hour recording sections. We adapted the recording sections to the periods of activity of each insect. Therefore, bee data were collected in periods that included at least a few hours of daylight. For mosquitoes, the recoding sections included dawn and/or dusk, and at least a few night hours. All data were recorded at a sampling rate of 44100Hz, and were later sampled down to 16000Hz, in order to reduce the memory requirements to process and store the data. The sampling rate of 16000Hz is adequate to record insect data, since it can represent frequencies up to 8000Hz, and virtually all insect species have wing-beat rates are lower than 1000Hz. In total, we obtained more than 100 hours of recordings, considering all species. The recordings consist of background noise with occasional "bleeps".

The ESC-50 dataset as shown in Table 5.1 includes 2,000 short clips comprising 50 classes of various common sound events, and an abundant unified compilation of 2,50,000 unlabeled auditory excerpts extracted from recordings available. They are grouped in 5 loosely defined major categories (10 classes per category) such as animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/domestic sounds, and exterior/urban noises.

Table 5.1 Benchmark dataset of insect classification

Name of Dataset	URL	Description	Types of Insects
ESC-50	https://github.com/karoldvl/ESC-50	• 50 classes of various environmental sounds (40 clips per class). • 5-second-long recordings • 44.1 kHz • .ogg files	Animal sounds, water sounds, Human (non speech) sounds
Mosquitoes	[2] Chen, Y. Supporting Materials. https://sites.google.com/site/insectclassification/ (2013).	20,000 Files	Different Type Of mosquitoes

The proposed work is implemented in MATLAB for assigning class labels as insect and non-insect based on their sound. It accepts audio file of the format .au or .wav as input from the benchmark data or uploaded dataset with good performance accuracy of 86.75% and for uploaded dataset the classification accuracy is 85.78%.

V. Conclusion

Classification is the most important data mining technique used for categorizing the data. And it is too difficult to classify the acoustical sounds of flying insects, as they are time-invariant. In this paper, classification models such as SVM, Bayes, kNN and Fisher LDA are analyzed with the same dataset. Classifiers based on soft computing techniques such as Fuzzy classifier and Neural Network classifier are also considered for performance analysis against the same dataset. According to the classification of insect|non-insect the accuracy is computed with respective algorithms implemented. And the computation shows that the Bayesian outperforms other classifiers aforementioned.

In future, in order to improve the accuracy and time of execution combination of classifier models on same dataset can be implemented.

References

1. Banko.M, Brill.E, 2001, "Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing." Proceedings of the first international conference on Human language technology research (pp. 1-5). Association for Computational Linguistics.
2. Batista.G.E, Keogh.E, Mafra-Neto.A, Rowton.E, 2011, "SIGKDD demo: sensors and software to allow computational entomology, an emerging application of data mining." In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 761-764
3. Benedict.M, Robinson.A, 2012 "The first releases of transgenic mosquitoes: an argument for the sterile insect technique." TRENDS in Parasitology, 19(8): 349-355

DOI : 10.18231/2454-9150.2018.1224

4. Capinera.J.L, 2005, Encyclopedia of entomology. Springer. Epsky ND, Morrill WL, Mankin R, "Traps for capturing insects." In Encyclopedia of Entomology, pp. 2319-2329, Springer Netherlands

5. Chen.Y, Hu.B, Keogh.E, Batista.G, 2013, "DTW-D: time series semi-supervised learning from a single example." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 383-391

6. Cooperband.M.F, Hartness.A, Lelito.A.P, Cosse.A.A, 2013 "Landing surface color preferences of Spathiusagrili (Hymenoptera: Braconidae), a parasitoid of emerald ash borer, Agrilusplanipennis (Coleoptera: Buprestidae)." Journal of Insect Behavior. 26(5):721-729

7. Deakin.M.A, 2010 "Formulae for insect wingbeat frequency." Journal of Insect Science,10(96):

8. Frick.T.B, Tallamy.D.W, 1990, "Density and diversity of non-target insects killed by suburban electric insect traps." Entomological News, 107, 77-82

9. Hao.Y, Campana.B and Keogh.E, 2012, "Monitoring and Mining Animal Sounds in Visual Space." Journal of Insect Behavior: 1-

10. Keogh.E, Pazzani.M, 1999, "Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches." In Proceedings of the seventh international workshop on artificial intelligence and statistics. pp. 225-230

11. Li.Z, Zhou.Z, Shen.Z, Yao.Q, 2009, "Automated identification of mosquito (diptera: Culicidae) wingbeat waveform by artificial neural network." Artificial Intelligence Applications and Innovations, 187/2009: 483–489

12. Nguyen.M.N, Li.X.L, 2012, "Ensemble Based Positive Unlabeled Learning for Time Series Classification." Database Systems for Advanced Applications. Springer Berlin/Heidelberg

13. Skyler Seto, Wenyu Zhang, Yichen Zhou,2015 "Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition." 978-1-4799-7560-0/15/$31 c IEEE

14. AlemGebru, Erich Rohwer, Pieter Neethling, MikkelBrydegaard, 2014 "Investigation of atmospheric insect wing-beat frequencies and iridescence features using a multispectral kHz remote detection system." Journal of Applied Remote Sensing 083503-1 Vol. 8

15. Begum N., B. Hu, T. Rakthanmanon, E. Keogh, 2014, "A Minimum Description Length Technique for Semi-Supervised Time Series Classification." Springer International Publishing: 171-192

16. Stephan Spiegel, Brijnesh-Johannes Jain, SahinAlbayrak,2014 "Fast Time Series Classification under Lucky Time Warping Distance." SAC'14, March 24-28, 2014, Gyeongju, Republic of Korea Copyright ACM 978-1-4503-2469-4/14/03

17. Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, Eamonn Keogh, 2014, "Flying Insect Classification with Inexpensive Sensors." Journal of Insect Behavior September, Volume 27, Issue 5, pp 657–677

18. TheodorosDamoulas, Samuel Henry, Andrew Farnsworth, 2010, "Bayesian Classification of Flight Calls with a novel Dynamic Time Warping Kernel." in The Ninth International Conference on Machine Learning and Applications (ICMLA'10) at Washington

19. Tuomas Virtanen and Marko Helen, 2007,"Probabilistic Model Based Similarity Measures For Audio Query-by-example." in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

20. JonaEnrique Alexandre, Manuel Rosa, Lucas Cuadra, and Roberto Gil-Pita', 2006 "Application of Fisher Linear Discriminant Analysis to Speech/Music Classification." Audio Engineering Society , Paper 6678, Journal of the Audio Engineering Society.

21. Hongtao Zhang, Yuxia Hu, "Extension Theory for Classification of the Stored-grain Insects", 2010 International Conference on Machine Vision and Human-machine Interface

22. Thomas Pellegrini, Jos´ePortˆelo, Isabel Trancoso, Alberto Abad, Miguel Bugalh, "Hierarchical Clustering Experiments for Application to Audio Event Detection", in Proceedings of Interspeech'08, Brisbane, 2008

23. Siti N. A. Hassan, Nadiah S. A. Rahman, ZawZawHtike and Shoon Lei Win,"Advances In Automatic Insect Classification, , Electrical and Electronics Engineering: An International Journal (ELELIJ) Vol 3, No 2, May 2014

DOI : 10.18231/2454-9150.2018.1224