

# Matrix Factorization for Multi Label Classification



**Researcher:**

Waqar Hassan

Roll No. 351-MSEE-FET-S14

**Supervisor**

Dr. Syed Zubair

**Department of Electronic Engineering,  
Faculty of Engineering and Technology  
INTERNATIONAL ISLAMIC UNIVERSITY  
ISLAMABAD**

**2016**



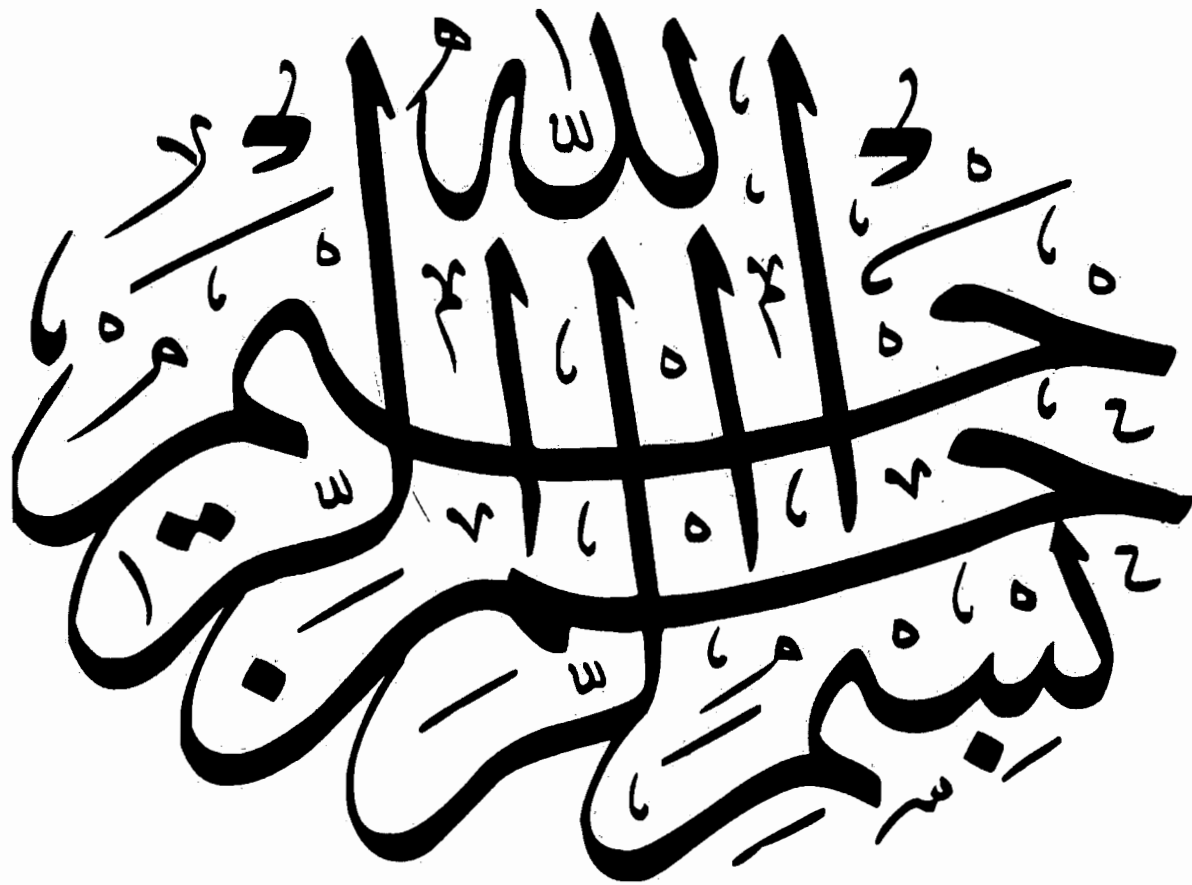
Accession No TH-16923

kygma

MS  
624.171  
WAM



K-Singular  
Multi nstance multi label.  
Segmentation.



**In the Name of ALLAH**

**Who is the Most Gracious and Merciful**

## CERTIFICATE OF APPROVAL

**Title of Thesis:** Matrix Factorization for Multi Label Classification

**Name of Student:** Waqar Hassan

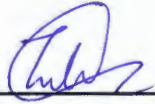
**Registration No:** 351-FET/MSEE/S14

Accepted by the Faculty of Engineering and Technology,  
INTERNATIONAL ISLAMIC UNIVERSITY, ISLAMABAD, in partial fulfillment of the  
requirement for the Master of Philosophy Degree in Electronic Engineering with specialization in  
communication and signal processing.

**Viva voce committee:**

**Dr. Syed Zubair**

Supervisor



---

**Dr. Ihsan ul Haq**

Internal Examiner



---

**Dr. Muhammad Usman**

External Examiner



---

**Prof. Dr. Muhammad Amir**

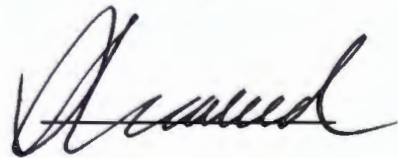
Chairman DEE



---

**Prof. Dr. Naveed Aqduus Malik**

Dean FET



---

## Declaration



Certified that the work contained in this thesis entitled

### **Matrix Factorization for Multi Label Classification**

is totally my own work and no portion of the work referred in this thesis has been submitted in support of an application for another degree or qualification of this or any other institute of learning.

**Waqar Hassan**

351-FET/MSEE/S14

## Acknowledgements

First of all, I am heartily thankful to ALLAH for leading to me this path. Special gratitude to my supervisor Dr. Syed Zubair for the continuous support of my research work, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. My next gratitude to Mr. Rehan Ahmad and Mr. Sharjeel Abid Butt, who always supported me during the studies and for his encouragement, insightful comments, and hard questions. Thanks to all of my friends who encourage me for this struggle.

My humble gratitude to the Department of Electronics Engineering , Faculty of Engineering and Technology , International Islamic University for providing all the available resources for the research and simulation environment.

Lastly, I dedicate this thesis to my family, especially to my mother, father, brothers, sister and all the little angles in the home. All of them are always a big source of praying for me.

## **Dedication**

This Thesis is dedicated to

Prophet Muhammad (P.B.U.H), the most reformer of the world

To my

Beloved Parents, brothers and sisters for their lot of prayers for me

To all the teachers

From the School to University, A beacon for guidance

To all the friends

Great support, and for giving the joyful moment of the life.

# Table of Contents

Table of Figures.....	ix
List of Tables .....	x
List of Abbreviations .....	xi
Abstract.....	xii
Chapter 1 - Introduction.....	1
1.1 Introduction.....	1
1.2 Preliminaries .....	5
1.2.1 Machine learning.....	5
1.2.2 Supervised Learning.....	5
1.2.3 Unsupervised Learning.....	6
1.2.4 Single Instance Single Label.....	7
1.2.5 Single Instance Multi Label .....	7
1.2.6 Multi Instance Single Label .....	9
1.2.7 Multi Instance Multi Label .....	11
1.3 The Learning Framework .....	12
1.3.1 Multi Instance Multi Label .....	12
1.3.2 Traditional Supervised Learning.....	12
1.3.3 Multi Instance Learning.....	13
1.3.4 Multi Label Learning.....	13
1.4 Sound origination process of birds .....	14
1.4.1 Syrinx .....	14
1.4.2 Trachea .....	16
1.4.3 Larynx, mouth and beak.....	16
2 Chapter 2 – Literature Review .....	18
2.1 Literature Review .....	18
3 Chapter 3 – Proposed Work .....	26
3.1 Problem statement .....	26



3.2	Proposed work.....	26
3.3	Expected goals and objectives.....	26
3.4	Research Methodology .....	27
3.5	Block diagram .....	28
3.5.1	Source Separation .....	29
3.5.2	Preprocessing and Noise Reduction .....	30
3.5.3	Segmentation .....	32
3.5.4	Feature Extraction .....	35
3.6	Classification .....	41
3.6.1	Fisher discrimination dictionary learning .....	41
3.6.2	Model of FDDL.....	43
3.6.3	Optimization of FDDL.....	44
3.6.4	Classification scheme .....	45
3.6.5	K-Singular value Decomposition (K-SVD).....	45
4	Chapter 4 - Experiment.....	47
4.1	Data Set.....	47
4.2	Experimental setup .....	51
5	Chapter 5 - Results.....	53
6	Chapter 6 - Conclusion .....	56
7	References.....	58

## Table of Figures

Figure 1.1: Single instance single label .....	7
Figure 1.2: Single instance multi label .....	7
Figure 1.3: Multi instance single label .....	10
Figure 1.4: Multi instance multi label .....	11
Figure 1.5: Sound originating process of bird .....	14
Figure 1.6: Schematic view of songbird syrinx.....	16
Figure 2.1: Hierarchical levels of song .....	19
Figure 3.1: Block diagram of proposed work.....	28
Figure 3.2: Before noise removal.....	32
Figure 3.3: After noise removal .....	32
Figure 3.4: Segmented spectrogram.....	34
Figure 3.5: Illustration of discriminative fidelity term .....	43
Figure 4.1: Song meter data collection locations .....	48
Figure 4.2: Recordings in the dataset, counted in different ways.....	49
Figure 5.1: Recognition rates of FDDL.....	55

**List of Tables**

Table 4.1: The 10 bird species in the data set ..... 50  
Table 5.1: Recognition rates for FDDL ..... 53  
Table 5.2: Classification rates of K-SVD..... 55

## List of Abbreviations

1. MIML: Multi instance multi label
2. SIML: Single instance multi label
3. MISL: Multi instance single label
4. SISL: Single instance single label
5. FDDL: Fisher discrimination dictionary learning
6. SVM: Support vector machine
7. IPM: Iterative projection method
8. DL: Dictionary learning
9. SVD: Singular value decomposition
10. MIL: Multi instance learning
11. MLL: Multi label learning
12. FFT: Fast Fourier Transform
13. MFCC: Mel-Frequency cepstral coefficients
14. DCT: Discrete cosine transform
15. ICA: Independent Component analysis

## Abstract

The audio recordings collected from field contain sounds of different species of birds. Each recording has multiple birds' sounds as they are vocalizing simultaneously. This type of problem belongs to a multi instance multi label framework. The classification of set of species which are present in an audio recording in this framework has received little study so far. In the proposed work of bird's species classification we use dictionary learning techniques for classification. We use two techniques namely Fisher discrimination dictionary learning (FDDL) and K-Singular value decomposition (K-SVD). To separate the multiple bird's calls from a recording we use source separation toolkit. In the preprocessing step, noise removed spectrogram of each audio recording is formed. We also used 2-D time frequency segmentation which separate overlapping calls of bird in time. After segmentation, we find 38 dimensional feature vector of each segment of recording. Then these feature vectors are used by dictionary learning techniques to learn class specific dictionaries. In Fisher discrimination dictionary learning (FDDL), the learned dictionary and sparse coefficients both are discriminative and reconstruction error of each class is used for the classification of bird species. While in case of K-Singular value decomposition (K-SVD), the dictionary is learned using SVD decomposition and the classification is performed on the basis of sparse representation of test signals. The discriminative dictionary used in FDDL play a vital role in the classification of birds species and give better results than K-SVD. The audio data set contains 10 species of birds collected from H.J. Andrews Experimental Forest using omnidirectional microphones. The implemented work has many application e.g. in conservation planning, modelling of species distribution and also in long term monitoring of remote sites.

# Chapter 1 - Introduction

## 1.1 Introduction

Classification is defined as a method to categorize every new sample into different classes on the basis of similarities of their characteristics with some predefined criteria. This technique has a key role in the machine learning application where the classes of samples have to be predicted. There are different types of classification, one of them is single instance-single label in which one sample may belong to one class only. This classification is good for the cases where one sample belong to only one class at a time but it is not the case for all the time as there are some samples which may belong to many classes in different time period [1]. Other category of classification is single instance multi-label classification, in which one sample may belong to many classes at same time. If we consider the example of bird classification many birds may articulate at the same time, so in this scenario single label technique will not give good results. In that case multi class labels will give desired results for the classification of the birds species [2].

If we are designing any type of machine learning system including audio classification system first step to follow is the division of the dataset. Data is divided into two categories i.e. training dataset and testing dataset. This division of the data is required because every machine learning task requires two steps one is training of the system and other is the testing of the system. Training of the system is done by the training data samples with their corresponding labels so that the system learns the relation between the features of the samples to the corresponding labels and set its characteristic accordingly. In second stage, the trained system is assessed using test data for classification. In most of machine learning tasks facts is not usually present in some standard format on which we can do learning and evaluation of the system, in that case data is converted to

some standard format by doing some preprocessing. If we consider the case of the birds sound classification system, voices of the birds are recorded so it may contain some noise in it like the voice of wind, water etc., so before using such type of samples we will have to do some preprocessing in order to remove unwanted background voices. In this way at the end all the samples only contain the voice of the birds.

To extract features, feature extraction technique can be used that contains information on the basis of which diverse samples of the bird's species can be differentiated. While selecting feature extraction technique one need to be careful to adopt that technique which extract all useful and relevant information of the sample that train the classifier successfully. As we initially divided the data into training and testing dataset so the training of the system is done through training data set and testing of the system is done through testing samples [3].

Using the voice recordings of different species of birds for their automatic classification is the vital problem. Habitat destruction, falling biodiversity, and change of climate impose the production of precise and proficient tools that monitor or observe the birds [2]. Birds could be very important in the life cycle, they provide very use full information about ecosystem as their presence gives the information about the presence of plants, insects etc., in a habitat. They are very sensitive to ecological changes and respond rapidly to that. Therefore, birds are said to be the indicator of the ecological changes, but it is very tedious and difficult to observe the population and activities of the birds.

In our culture sounds of the birds have a vital impact on different aspects of human life. For many people the sound of the bird may be the sign for starting of the new day and for others these are the sign of the starting new season. People find it as a source of pleasure to listen the sounds of the birds and watching them all around world. Many experts of the bird's sound can easily classify hundreds of the sound of different species of birds very easily. Similarly, the main motive behind our work is to project a programmed bird classification and identification system which can classify different species of birds and also can identify the species of birds only by the sound track of the birds.

Syrinx, a special distinctive type of organ which is responsible for the sound production in the birds, have complex function and nature. Classification of birds from their voices is the challenging task because of a very large spectrum produced by their voices. If we look at the categories of the bird songs, there are two type of birds' sounds songs and calls. Elements and syllables and different levels of phrases which are the future classification of these two categories. One or more element combines to make a syllable. If we are considering the continuous time signal, then element can provide use full information than syllables and phrases and it can be used as the testing and training data. Hence, in this way the efficiency of the classification system can be greatly improved. Syllable provides less information about mortal and regional variation in comparison of phrases.

Bird's specie recognition is usually a pattern recognition problem. Bird's songs are termed as patterns which are represented completely by the feature vector. First of all, features are extracted by the songs produced by different species of birds. When new species comes for the recognition, features are extracted in the similar way and compared through the system which contains features of all species [3].

Bird recognition is also done traditionally by human experts. But this traditional method is tedious and time consuming and also have some limitations of temporal and spatial resolution and collecting the data for the recognition. A survey on the traditional method show that most of the time, voice of the birds are recorded using microphone and then these recordings are evaluated by human experts or some machine learning algorithm. Whereas in audile population analyses, microphones can be used to record the sounds of bird, then machine learning techniques can be used to observed these recordings for the evaluation of bird species characteristics e.g. existence of species in a recording, species profusion, age and masculinity etc. Thus human spectator analyses can deliver fewer determination than audile analyses, and in remote sites human survey is not a good practice. For example, if human want to take and record observation on a cliff sides continuously for two weeks it is not feasible or practicable for him, but by using microphone observation can easily be collected on remote sites [4].

Single instance single label (SISL) is the common type of machine learning algorithm in which one sound track is presented by one feature vector and for one feature vector one label is assigned



to it. Such type of single instance single label are classified by well-known SISL algorithms, support vector machine and logistic Regression. Minimizing the complexity of such type of SISL classification problems of image and audio we usually do some transformation and discards all the irrelevant information from the data. For a complicated learning data set many machine learning problems in the image and audio processing divides it into many parts by applying some preprocessing techniques. If the data set contains many sources in it then it may require many feature vectors to completely describe it. Consider the image which is usually characterized as collection of region in the same way audio data can be divided into different parts. This type of complex problems where multi instances are required to represent the data set give inspiration for the Multi instances learning [6]. As the data in this case is represented by two or more feature vectors therefore it is different from basic supervised learning algorithms. On the other hand image can also be represented by multiple labels, for example to label all objects existing in an image. That type of problems are solved in the multi-label learning, where every object characterized through a feature vector and assign a tag to each feature vector. Consider the example of the bird's sound recording, recording can also have the different type of bird's sounds where each bird sound can be exemplified through feature vector and a single label is assign to each feature vector, this type of idea inspires the multi instance multi label problem.

Here in this proposed work we followed some steps in order to solve bird classification problem. First of all we apply a source separation technique on the sound recordings of the birds in order to separate different type of sources present in the audio recording of the bird sounds. Spectrogram is found from the resulted data of source separation and then Fast Fourier Transform is applied on each frame. To modify the sound of interest and remove extra sounds which is not of our interest filtering is applied on it. Spectrogram has diverse segments that express different utterance syllable of bird's sounds. These syllables are completely characterize by the set of feature vector which is then used in the classification problem [7]. We also equipped with Fisher Discriminant Dictionary Learning (FDDL) and dictionary learning approach for classification in our proposed work.

## **1.2 Preliminaries**

In this section we will give complete overview of the previous work in the field of machine learning and specifically about supervised machine learning.

### **1.2.1 Machine learning**

Every time when we incorporate the world we learn a new reality about it. We cooperate with the environment close to us, and get emotion or reactions through our perceived capabilities. Humans have ability to pinpoint the trends and conditions and also have ability to identify our activities rendering to our situations that suit well to a particular condition. That process could be called as learning.

Learning is inclined to intelligent machines and humans. Main problem arises that how to design the algorithm for machines which can copy the human learning nature for new data comes on it. People in the past few decade tried a lot to equip such algorithms for machines which can give them the ability to copy human intelligent. On the other hand, designing such an algorithm which can give complete ability to machine to copy human intelligence is a difficult task. Although scientist and engineers all over the world started to develop such kind of techniques in which machine can copy human intelligence and they are somehow succeeded for some situation where machine can learn the trends from the data and can accurately predict for new data.

Some of these algorithms are commercialized and are effectively working in the corresponding fields. Business can be more productive, cheap and useful and many fields from medicine to marketing can be improve and enhance by equipping machines with such type of intelligence algorithms because machines are more productive then humans as they do tasks more faster than human in short interval of time [8].

### **1.2.2 Supervised Learning**

Supervised learning is the sub category of machine learning. In this type of learning, machine can learn concepts from labeled samples. Basic learning procedure of this type of machine learning algorithm is given by set of data which is termed as training dataset. In this dataset each sample

possess a unique characteristic which is represented by the set of feature vectors in such a way that can make it differentiable from others. To each of these feature vector a corresponding targeted output is assigned to them which is called its label. By using these training data set, model learn the trends and relations between data and built a module that will accurately classify the new coming data. So, if any new data comes it can accurately predict the trends and can classify them into the corresponding class.

Let us consider an example which will show all these concepts very clearly. Suppose we are given a set of different iris followers and we are asked to differentiate between different classes of iris follower so what we will do? First step for us will be to look at the different samples of follower and try learn the basic difference between the classes. Once we learn the difference between the different classes of flowers by observing different attributes like color, Patel length, Patel width etc., then we will be able to differentiate any new samples into the corresponding class.

Here, we can appropriately characterize some fundamental circumstances of supervised machine learning. An example comprises of  $N$  feature vector and every feature vector is assigned an appropriate label to it  $\{j | l(j)\}$ , where  $j = \{j_1, j_2, \dots, j_N\}$  represents the feature vectors and  $l(j)$  represent the associated label assign to each feature vector. Nature of features depend upon the sample which are characterizes by them, it can be some numeric data or character data.

The main step is to find  $l(j)$  by interpreting the given set of feature vectors that process is known as classification [8].

### **1.2.3 Unsupervised Learning**

Unlike supervised learning task, unsupervised learning assume the function which can best estimate the hidden concepts in the data. There is no check on the performance no labels are assign to corresponding feature extraction, main reason of the differentiation between supervised and unsupervised learning algorithms. Many data mining based algorithms are developed by unsupervised learning algorithms.

Different methodologies can be used in unsupervised learning algorithms like clustering that comprises k-means models and ranked clustering etc.

### 1.2.4 Single Instance Single Label

We have talked about the traditional machine learning algorithms and single instance single label (SISL) is one of them. In this type of learning realistic world situation can be characterize by a feature vector and one label is assigned to each feature vector, this example is illustrated in figure 1.1. We assume that  $W$  represents all the characterizing feature vector of the situation and  $R$  represent the corresponding labels for each situation. Major step is to equip a function  $f : W \rightarrow R$  for the provided database  $\{(w_1, r_1), (w_2, r_2), \dots, (w_m, r_m)\}$ , where  $w_i \in W$  are the characterizing vectors and  $r_i \in R$  are the assign labels for each corresponding feature vector [9].



Figure 1.1: Single instance single label

### 1.2.5 Single Instance Multi Label

Figure 1.2 shows the concept of single instance multi label in which any realistic situation is completely characterize by a feature vector and each feature vector is associated with the many labels.

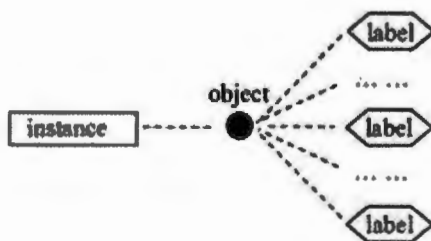


Figure 1.2: Single instance multi label

As discussed earlier all the traditional supervised learning frame work are very effective and efficient and give higher accuracy in almost every field. But these algorithms have the limitation i.e. these are efficient in the case where the instances are complex and it may belong to different situation at the same time. If consider some examples let say we train a model for text characterization but testing sample may contain more topics like sports, London Olympics marketing so all these topics belongs to different situations and it may require different labels and possibly different feature vectors to completely characterize them.

To solve the problem described above where the same instance may belong to different situations and concepts at the same time one direct solution is to extract a feature vector describing each situation and assign labels to it. In relating to supervised learning task one feature vector may extract from the situation which can completely characterize the situation and then assign multiple labels to that feature vector. The main aim is to equip a function which easily predict the labels of the unseen samples.

Multi label learning and supervised learning are similar if multi label learning problem is characterize by only feature vector and only one label assign to it. But if we change the multi label learning task like this to assign a single label to it this will make learning task much more difficult. In reality, problem with learning the model with the multi label learning will occupy much larger size of the output space and the output space increases exponentially with the size of the label. Let say if we have 20 label than the output space will be in exponential power of twenty.

If in a typical multi-label supervised learning algorithm all samples are restricted to have only one single label at a time it will ultimately depreciate this supervised learning algorithm for multi-labeling. The attempt to simplify multi-label learning to have single label at a time will complicate the task and make it even more difficult. The main challenge to deal with in multi-label learning is its irresistible output space size, i.e. the corresponding output label set increases exponentially with increase in its respective class label. For example, for a 20 class label set its output label set will possibly be increases by one million.

In order to deal with this challenge of exponential increase in output label set it is required to link the label between them during learning procedure. For example if an image has the label of Brazil it can't be assigned the label of rain forest or if a document is labeled with politics can't get the label of entertainment at the same time and so on. In other words, to attain multi-label learning techniques the exploitation of correlation information label is considered to be decisive [10].

### **1.2.6 Multi Instance Single Label**

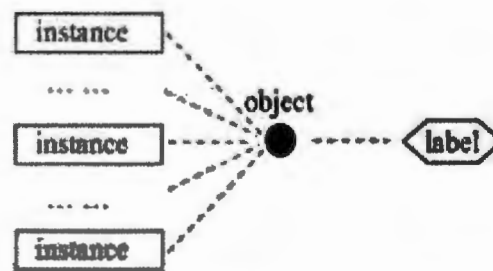
MISL (Multi instance single label) has wide application in real life problems such as image classification or drug activity classification hence, considered one of the most effective method for theoretical problems.

While learning an algorithm for MISL for a real world problem each object classify on the basis of multiple feature vectors or bags of features. Here, each feature vector is assigned a single class label but the instances of the bag of features are not labeled separately. The problem to design a model for real world problem is to train a model that learn relations of given sample data and predict accurately the label of unknown data in the future.

A type of learning or more specifically the supervised learning framework type is Multi instance learning, the multi instance building makes its model from the examples which are given labeled examples. The given examples which are labeled examples used along with using some algorithm, for solving the problems of classification and the regression. Most of the time the classification can be binary as it may be either positive label or either it can be negative label. The main aim of the model designing is to make a model such as per the training dataset given so that the classification of the model can be in the class from which it belongs. Assigning of label in training phase to every class is known as supervised learning algorithms.

The training sample's nature variation explains the difference between the multi learning and the tradition learning. The characterization in multi instance learning is done by feature vector which is multi set, In short we can say that description of object is done by features of multi vectors. The multiple set of feature extraction is shown in Figure 1.3. The corresponding label for sample is assigned and the target is to predict the unknown data which is named as testing data.

Now coming toward the a real life problem, the problem targeted here is of a jailer, the jailer wants to open the locked door and number of key chains are Z in number with keys of large number in each. The key chains are bags in this example and the features in it are keys. The importance will be given to that keychain which contains the key of the locked door which is to be opened, remaining chains of keys are unimportant. In this problem the aim of model designing is to be the model such that it can classify as useful chain or un-useful chain.



*Figure 1.3: Multi instance single label*

The application of the Multi-instance learning techniques is found in multiple fields. Some of them from image processing are image retrieval, detection and recognition of face, classification of text etc. An image is characterized by feature vectors as multiples patches can be obtained from the image and each patch has some information which is used as feature vector. Likewise, extraction of sections from the document file and later the extracted information can be used as feature vector, though it depends on which classification is to be done. Web pages can also be characterized as links of web with some labels so they can be collected and classified on the basis of labeling.

Researchers have done so many researches on multi instance single label algorithm designing, it is subjective to real time objects which can be represented by number of samples and related to single label for each class. Let us take an example, for the network training the bags contain features which are large in number. Labeling of a bag is done as in the existence of a positive sample in a bag, a bag is said to be positive otherwise it can be negative. In this example the goal of designing a model is to classify the samples into positive and negative as discussed above. Dietterich et al. [6] gave an idea regarding the system designing for accurate prediction of drug activity[1].

### 1.2.7 Multi Instance Multi Label

In real world problem, practically existing object is presented in the system of MIML (multi instance multi label) learning by the multiple feature vectors. The Labels associated with the feature vectors are the multi class labels, it can be seen in Figure 1.4 given below:

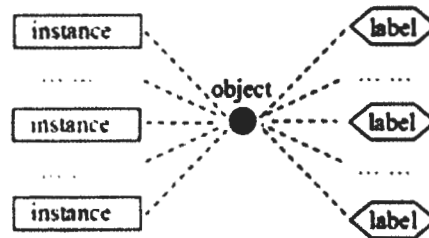


Figure 1.4: Multi instance multi label

The real time objects are problematical as they contain the situation which are not constant instead they vary at the specific time. Let an example image is taken which have multiple classes as like of the Lions grass etc. If we look forward to the text documents it can be novel or book regarding some tale of traveling. The classification of web pages can be as the sports page, discovery pages, news page etc. Now the example of the image retrieval is taken, if we want to retrieve the image of only lion from the whole image we will be interested in the features regarding the lion, but an issue arises when object comprises of the multiple concepts. While considering the solution for the given problem we have to keep in mind the whole possible concept, the subsets of all the possible solution will be collected and from the solution subsets the solution is selected so that's why in this way while learning the network difficulty faced is less.

MIML is very useful in the learning problem of single label object. Sometime a strange case arises as an object belongs to situation which are multiple with large set of information so it will be a difficult task for using the whole information with a single label learning. To resolve this we look at either we can transform to sublevel as each sublevel is dealing with the situation as per the information. In this way we do the network learning first, learning is done for the sublevels and then these sublevels are recombined to do leaning for the complex one [1].

Now moving further towards the traditional supervised form of leaning, this form is the worst form of multi label learning and multi instance learning. Thus, by correspondence of multi instance



multi label problem finding in the SISL framework can be solved by the use of bridge that can be the multi-label learning or multi-instance [9].

### 1.3 The Learning Framework

Let we go through the learning framework now; following consideration are to be made:

Feature vector is represented by  $W$ .

Class labels are represented by  $R$ .

#### 1.3.1 Multi Instance Multi Label

In MIML (Multi instance multi label) learning a function is to be learnt which is  $Y : 2^W \rightarrow 2^R$ .

The learning of the function is from the dataset given below:

$$\{(W_1, R_1), (W_2, R_2), \dots, (W_m, R_m)\},$$

Where

$R_i \subseteq R$  Represents the set of labels  $\{r_{i1}, r_{i2}, \dots, r_{i,t_i}\}$   $r_{ik} \in R$  ( $k = 1, 2, \dots, t_i$ ).

$W_i \subseteq W$  Represents feature vector set  $\{w_{i1}, w_{i2}, \dots, w_{i,p_i}\}$ ,  $w_{ij} \in W$  ( $j = 1, 2, \dots, p_i$ ).

$p_i$  Represents the number of features presented in  $W_i$  which is feature vector.

Number of Labels  $R_i$  is represented by  $t_i$

Now the comparison of MIML and the supervised learning is done. For comparison multi-instance based learning and multi-label based learning are selected in traditional supervised learning.

#### 1.3.2 Traditional Supervised Learning

Function for learning of Single instance single label is  $Y : W \rightarrow R$ . The learning is from the dataset given which is  $\{(w_1, r_1), (w_2, r_2), \dots, (w_m, r_m)\}$

Where  $w_i$  associated label is  $r_i \in R$  and  $w_i \in W$  is feature vector.

### 1.3.3 Multi Instance Learning

Function for learning of MISL learning is  $Y : 2^W \rightarrow R$ . The learning is from the dataset given which is  $\{(W_1, r_1), (W_2, r_2), \dots, (W_m, r_m)\}$

Feature vector is  $\{w_{i1}, w_{i2}, \dots, w_{i, n_i}\}$ , and set is  $W_i \subseteq R, w_{ij} \in W (j = 1, 2, \dots, n_i)$ ,

Label of  $W_i$  is represented  $r_i \in R$  and  $n_i$  represents the total number of feature vector in  $W_i$ .

### 1.3.4 Multi Label Learning

Function for learning of Single instance multi label learning is  $Y : W \rightarrow 2^R$ . The learning is from the dataset given which is  $\{(w_1, R_1), (w_2, R_2), \dots, (w_m, R_m)\}$

Feature vector is  $w_i \in W$  and  $r_i \subseteq R$  is subset of labels as  $\{r_{i1}, r_{i2}, \dots, r_{i, t_i}\}, r_{ik} \in R (k = 1, 2, \dots, t_i)$ .

Label of  $W_i$  is represented  $r_i \in R$  and  $t_i$  represents the labels number in  $R_i$ .

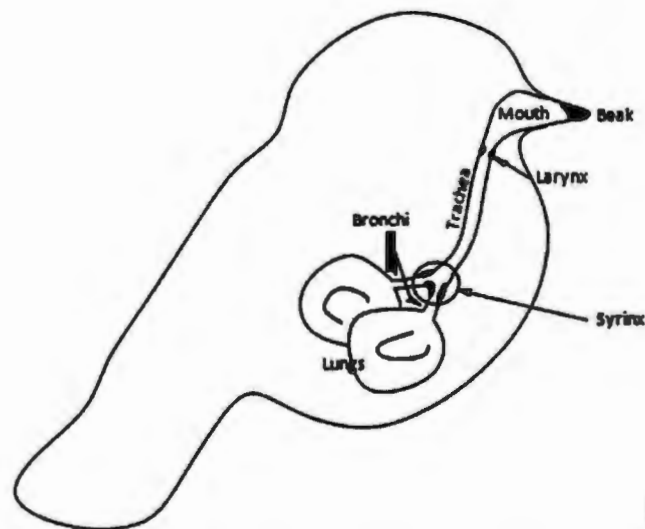
The difference between the learning algorithms can be observed clearly. The multi-learning is loaded because of the ambiguities present in the problems related to the real time. In Multi-instance learning the characterization is because of the multiple feature vector as in space of feature vector ambiguities present. When studying MIML, the ambiguities in both input and output vector spaces are seen. In real time the learning is not as much important as the feature extraction is because if the feature extraction is good, more information will be extracted so it will be easy for the algorithm to learn it and decide on the basis of trained data. In real time the objects are not much easy to understand because of the ambiguity at output and input. For such case of ambiguity, the MIML is very suitable.

If we look forward deeply we came to know that in most of the cases single instance multi label is less practical than multi instance multi label. Let take an example, we have object which is represented by a single feature vector but its linking is multiple labels named as  $l_1, l_2, \dots, l_l$  and  $f_1, f_2, \dots, f_n$  are the naming of the feature vectors. In single features the primary information is easier to work. We can also say that it would be easier for learning in case of MIML. For the relation building between semantic meaning and input pattern, we have to use the set of feature vector for representation of multi label objects. Let consider the MIML illustration example, the

one entity label is  $l_1$  and the feature vectors present in it are  $f_n$ . The label it has is  $l_i$  because of feature  $f_i$  it contains where as both  $f_1$  and  $f_i$  existence generates the  $l_j$  which is label [1].

## 1.4 Sound origination process of birds

To originate the sound in bird main components are as: larynx, beak, bronchi, trachea, lungs, mouth and syrinx. When air is flown from the lungs and transmitted to syrinx over bronchi a sound is produced. Vocal tract modulate the flow of sound from syrinx. Syrinx contains the mouth trachea, beak and larynx. In figure 1.5 the diagrammatically view of process is shown, it varies as per specie in term of dimensions and the size.



*Figure 1.5: Sound originating process of bird*

### 1.4.1 Syrinx

Syrinx the main element in sound initiating method of birds is one amongst the foremost necessary and widely considered organ of birds. In sound originating method though its necessary organ of birds however it additionally offer some necessary information related to the bird's organization as the structure of that organ is completely different in numerous bird species.

They are often categorized by three varieties. These main varieties are unit cartilaginous tube, cartilaginous tube and tracheobronchial. These 3 varieties is as per distinction among cartilaginous tube and cartilaginous tube parts of syrinx. Sound originating system will set at completely different position in cartilaginous tube if is found in cartilaginous tube. Figure 1.6 displays a part of cartilaginous tube area unit animal tissue ring, characteristically comprehensive, with trachea in straight continuance. The cartilaginous tube parts may have paired partial C-shaped animal tissue rings and ends are exposed against one another. The characteristics which are intermediate are same therefor it's terribly troublesome to differentiate of these 3 classes.

In existence the large cluster is of song birds, regarding 4000/9000 from total range of species of birds coated by them. From whole of birds most typically studied are songbirds along with syrinx of birds (song birds). From Figure 1.6 give a view as syrinx of song bird that is varied in structure however relatively constant during this cluster and it's thought-about because the paradigm syrinx. It belongs to tracheobronchial syrinx as a result of it's settled at the junction of two bronchi and of trachea. From lungs the flowing air creates medial typaniform membrane in every bronchial route for trembling via the Bernoulli result in the time when bird sing song. Opposite to cartilage wall the membrane may be vibrating nonlinearly. A symmetrical try of muscles close syrinx controls the voice and emotions of membrane. The controlling of membranes with pressure sort of a reed in wind, however the reed within the wind instruments is closed blown where membranes are the blown open [3].

The endoscopic imaging in latest study shows medial typaniform membrane (MTM) couldn't be the main source for the production of sounds hence it contradicts the MTM theory. According to the Goller for sound production the two soft tissues are responsible. These two tissues are the lateral labia (LL) and medial labia (ML) as that of the human vocal fold. Some experiments are conducted and result deduced is when the air flow through these vibrating tissues a sound is produced. The experiments performed are as the MTM is removed from the birds surgically and birds are let to sing. When the bird produces sound a small change is noted and hence, it is confirmed that bird can live without MTM either the MTM is removed or damaged.

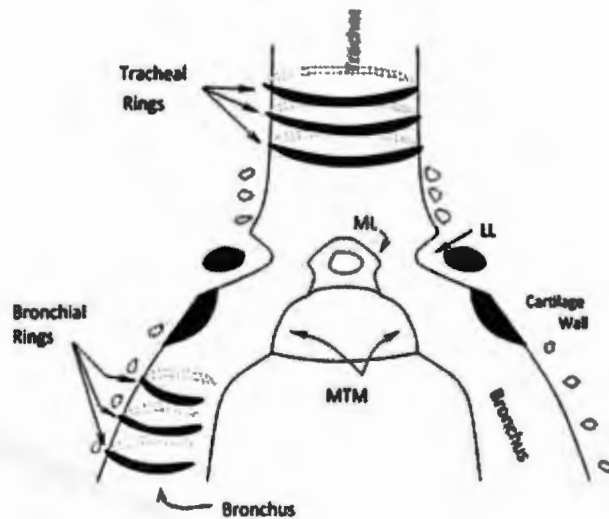


Figure 1.6: Schematic view of songbird syrinx

### 1.4.2 Trachea

Between larynx and syrinx a tube is called Trachea. For sound creator it acts as a resonator. The part of tube are Cartilage rings, usually which are complete. While observing the small passerines to long necked flamingos the cartilage rings of tracheal are ranges from 30 to 350 or less as it depends on the neck size. But it is not specific size as some species exists in which it can be coils or in loop. The difference in the mode of vibration can be enhanced by the loops of trachea. In some species the division of the trachea is in channels which are two in number because of this special feature they have characteristics which are specific.

### 1.4.3 Larynx, mouth and beak

Due to not availability of vocal folds in bird's larynx it is different from as of human. Some studies are available that explain the functionality of the bird larynx but the issue is still disputable that how the sound is originated. It looks as the role of larynx is little while in origination of sound. The bird's mouth is also resonator of cavity as of human but flexibility is less. Some birds tongue bends but most of them have unbending tongue. E.g. parrot produces the sounds as like that of human. Mostly birds handle the area of mouth with tongue.

The analysis of behavior of beak is difficult because of the complexity of the shape of beak and the dynamic nature of it is also a great contributor in not analyzing of dynamic nature. In beak the properties related to acoustics varies as the closing and the opening of beak changes the gap but as per research it is clear that in sound producing the beak has a key part. According to the Hoese et al. by changing either closing or opening the beak the vocal track length changes nonlinearly. Also the vocal track dimensions' changes by the head movement of the birds [3].

## Chapter 2 – Literature Review

### 2.1 Literature Review

Prior work in which machine learning is used for acoustic classification of various species of birds have been briefly explained in this chapter.

For bird species classification and detection there are number of design patterns that are common to different systems. A problem is divided into Segmentation, feature extraction, and classification, the three stages of one of the widely used method. In first step every segment is represented with a feature vector, the segments, syllables or calls are obtained by dividing a spectrogram or audio signal. To forecast the bird species, supervised classification algorithm is applied on feature vectors. Instead of finding the feature vector of individual parts, find the feature vector of the whole recording, and then classify the features. In some cases, prediction is done on the basis of multiple frames or syllables at the same time in other cases, prediction is based upon frames or syllables.

Calls and songs are two types of bird's articulation. Communicating signals which precisely conveying some messages are belonged to Calls, e.g., they can notify about the coming threat of hunter. Those expression of male birds which are linked with defensive singing and mating are difficult decorations of songs. Calls and songs were both considered in their work. Classified bird's articulation consists Phrases, elements and syllables as shown in Figure 2.1. Sequences of syllable having a specific pattern is called a phrase. It is quite possible that in last frame, syllable of phrase might be dissimilar but in most of the cases it is same in the figure. Element is the fundamental part of syllable. In complex cases many elements combine to make a single syllable while in simple

cases it is required for the making of a syllable and there is no difference in both. It is very difficult and tedious to separate element with in syllable [11]. The level of phrase could not be identified in the call sounds because they typically comprise of single syllable. Some species of songs are simple as they do not contain phrase. The smallest unit is Syllable. The sound which is produced by the single blow from lungs is called Syllable. In some cases, this definition is not valid as some birds are capable to produce complex circular breathing while singing. In most of the system which are used for audio classification recording is divided into units referred as syllables.

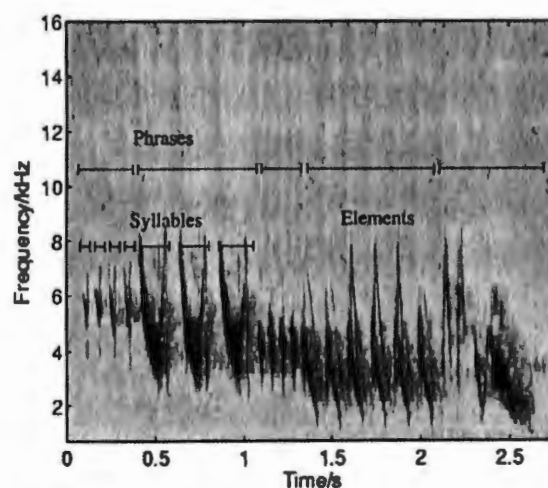


Figure 2.1: Hierarchical representation of the levels of song

Energy based segmentation method is the most widely used technique for segmentation. This technique is linked to loudness. They looked for the high energy spectrum on 2D region. The first step is the calculation of energy (which is in the form of levelled energy packet) as a function of time, to smooth out small deviations it convolved with some kernel. An alternative pattern on the basis of intervals separate the division, approximates the level of energy of noise in the background.

Let us consider an example, smoothed energy envelope, and calculate the global maximum energy  $M_{dB}$  is the first step in Harma and Somervuo [12]. Intervals where the energy is above  $T_{dB} = M_{dB} - 20 \text{ dB}$  threshold are defined as syllables. Form Intervals which are not in a syllable noise energy level  $N_{dB}$  is estimated. Using information based on the rule  $T_{dB} = (M_{dB} - N_{dB})/2$



of the noise level, process iteratively for the refine segmentation, the threshold is updated to a new value.

As it is mention in[3], [11], [13], first of all, the level of noise is estimated in  $N_{dB}$  as the smooth energy packet's global minima. After that  $T_{dB} = N_{dB}/2$  is set to be the threshold (keep in notice that it is divided by 3 because it shows loud sound and also dB is measured in negative). Intervals above a specific threshold, and re-computation from intervals the noise level that are less than the threshold are defined as syllables. By using the identical rule mentioned above, we can calculate the threshold. This procedure is repeated again and again to improve the segmentation.

The values of initial and final time of a syllable attempted to be found by most of the segmentation algorithms. The approach can have considered as a single dimensional approach. Instead, there is always an option to detect a 2D region or to detect bounding box in the spectrogram which match to a syllable. The major benefit of this system is that we can detached the sounds in 2D-segmentation, which overlap in time. Brandes [14] tried to categorize the sounds by species. Species such as frogs, birds and crickets. He kept emphasis on narrow-bandwidth calls and on modulated frequency during his work. On the bases of the average spectrum of the optima, the total range of the frequency was divided into the isolated bands. Then by using an energy threshold division of the 2D per-pixel was performed. The threshold was set by the band of noise-estimation. The major of this method is that, by using this method we get the time overlap segments.

Another researcher named as Forest Briggs [2] also worked on the segment extraction of bird calls or syllables. He took the recording bird calls and then extracted the segments by using a 2D division technique. Random forest classifier was used for the 2D-segmentation. Multi-Instance Multi-Label (MIML) was applied on the soundtrack, to get the species predictions from the track. He also proposed a method known as bag generator. In this technique the bag of vector features was obtained from the audio track. These features were then passed on to the classifier, which has already been discussed before. He also proposed a methodology to convert the audio track into spectrogram. He also applied some filters to get noise free spectrogram. In the next step, 2D-segmentation was applied on these spectrograms, the 38-dimensional feature vector is linked with

each division. After applying the bag generator, he got the classification of species by using the classifier known as MIMLSVM classifier.

Zhou et al directing towards the resemblance of MIML with the old-styled supervised learning technique. He used MIML or SIML technique to resolve MIML difficulties. Two MIML techniques were proposed by him. These techniques are given below:

- MIMLBOOST
- MIMLSVM

### MIMLBOOST

In this method Multi Instance (MI) learning technique is used for linking. First of all in this algorithm the original MIML issue is converted into MIL. This task is performed by converting every MIML model  $(W_i, R_i)$  into  $|R|$  quantity of MI models  $\{([W_i, y], \Phi[W_i, r]) | r \in R\}$ . Wherever  $[W_i, r]$  comprises of  $t_i$  occasions  $\{(w_1^i, r), \dots, (w_{n_i}^i, r)\}$  formed by linking each  $W_i$ 's sample with tag  $r$ , while  $\Phi[W_i, r] = +1$  if  $r \in R_i$  and  $\Phi[W_i, r] = -1$  if  $r \notin R_i$ . After that, MIMLBOOST solves the resultant MIS tag learning duty by recalling the MIBOOSTING technique. This technique converts a MISL learning task into an old-style supervised learning (SISL) problem under the theory that in the bag each feature vector contributes similarly and separately to a bag's label.

### MIMLSVM

In this method Multi Label (ML) learning technique is used as bridge. First of all, this algorithm converts the original multi instance multi learning technique into a SIML problem. This task is performed by converting every MIML model  $(W_i, R_i)$  into ML instance  $(\tau(W_i), R_i)$ . With the help of productive gathering a sack of feature vector is drawn by the function  $\tau(\cdot)$  into a lone sample  $Q_i$ . The clustering in K-medoids could be done on  $\Delta = \{W_1, W_2, \dots, W_N\}$  at the edges of bags and the parts of  $Q_i$  which relate the distance in between medoids and  $W_i$  of the cluster bunch.

Later, the resulting multi label learning task as solved by MIMLSVM technique that converts it into a SISL learning problems by debasing the ML learning task. Both the algorithms which are

mentioned above were applied on database of Images consisting of natural scenes fit in separate classes. These classes consist of sunset, trees, mountains and deserts [9] [15].

Different researcher's worked has proved that the study is usually based on the review of spectrograms of sound. They considered the sound of birds. If we consider a spectrogram of number of bird species, then it would become a very time consuming and difficult task for human to constantly classify the species. So we need to develop a separate system. A system which has the capability to identify the sounds of birds automatically. In order to solve this problem a scholar named as Anderson et al. [16] used the dynamic time wrapping (DTW) techniques to record the sounds of the bird continuously. When we compare the signal directly with the signal spectrograms, Constituent boundaries are identified. In order to extract the feature vector Fast Fourier Transform's (FFT) Log magnitude is used. Testing data is carefully chosen to test the system from truncated cutter and less noisy atmosphere. These data sets were given label by Human experts.

Two scholars, Kogan and Margoliash [5] used two different methods. Those methods are known as :

- DTW method
- Hidden Markov Models (HMM)

They used these methods for the automatic bird song classification from continuous soundtracks. FFT log magnitudes bins range from 0.5-10 KHz. Later they were used in DTW as feature vectors. Six types of feature vector are used to make the performance comparison of HMMs which are listed below:

1. Mel-frequency cepstral coefficients (MFCCs)
2. Linear predictive coding (LPC) coefficients
3. log Mel-filter bank channels
4. LPC cepstral coefficients (LPCCs)
5. linear Mel-filter bank channels
6. LPC reflection coefficients

If we compare the performance of all the above mentioned features, MFCC gave the best results. Because of these good results, this technique was used to collect the feature for the entire experiment. DTW technique gave the best performance towards the end. Though, when we try to extract the feature from short and unclear sound track, it requires more template. This is the only limitation that can be associated with DTW features. Similarly, in some experiments DTW has less accurate results compare to the HMMs. HMMs also have some disadvantages such as it is very likely that it may misclassify the song which has short duration.

In order to identify the bird species automatically Harma proposed a technique. Finding of the sinusoidal designing of syllable was the major functioning belief of this technique [17]. Syllable of each term was approximated by using two terms known as Frequency and Amplitude trajectory. By using this proposed algorithm bird's sound was divided into different segments. By selecting the sum of differences of frequency and the amplitude, we can calculate the distance among two syllables. These type signal models are enough for those soundtracks which comprises of restricted amount of species of birds.

When we try to model the sound of birds we do not get perfect results the major reason behind that is harmonic spectrum structure, both scholars named as Harma and Somervuo proposed a different technique which has the capability to classify the birds syllable into four classes. These classes were made on the bases of harmonic structure [12]. By using this technique, the rate of recognition was improved from 5-20% for most of the bird species. So, after these observations they decided that, in order to get the better recognition rate they will use level of song structure as an alternative of isolated syllables. Hence, they developed one more classification of bird song method. The new technique was built on the syllable pair histograms [18]. We can find the Gaussian syllable prototype automatically in order to stand in for every syllable. These automatic syllable prototypes can be found to represent each syllable by a set of Gaussian. The information which is related to time based structure of each bird song is given to Gaussian syllable. Every prototype which contains two successive syllables is known as syllable pair and it collect the histogram of syllable pair. Every birdsong was modeled with their syllable-pair histogram. A mutual correlation within histogram was calculated to check the similarity in the two histogram.

In order to assess the performance of the system 4 species were selected and 50 birds collectively from all the 4 species were taken, and the number of songs which were recorded are 257.

In order to do the automatic classification of bird's species these are the feature extraction techniques which are being used here (for extraction of feature from soundtrack of birds) descriptive parameters, MFCCs and sinusoidal model[11]. For best result another proposed technique is used in which MFCC feature vector is used with trajectory model for signal-syllable-based recognition. Using birdsong instead of using syllable means they use sequence of successive syllable as an alternative of using single syllable for classification in the prediction which improve the rate of recognition [19].

Frames, syllables, or whole recordings/songs could be targeted from classifiers for bird's sound. For the description of existence of sound for all the levels structure features were used. In some cases, extraction of feature for one level would be done and then accumulated to get of feature for higher level.

For the description of individual frame of a signal frame level feature were used, and frequently use frame's Fourier coefficients. The vector of Fourier coefficient magnitudes is the simplest frame-level feature.

Originally frame-level features were developed for speech analysis is Mel-frequency cepstral coefficients (MFCCs) [20]. Spectrum of a frame is transformed to the Mel-frequency scale in order to calculate MFCCs feature vectors, which is like human perception of pitch. After the use of number of non-uniformly spaced triangle filters, transformation is accomplished. Mel-frequency coefficient (MFC) is obtained from each of the triangle filter response. After the application of discrete cosine transform (DCT) to log of Mel frequency coefficients MFCCs features are calculated. By the use of DCT there is significant reduction in the dimensionality of the Mel-scale spectrum, and also did the removal of correlation with its elements. There are several ways by which we can calculate the MFCCs, one of them is to use number of triangle filters.

Mostly, the feature of bird sound used in machine learning is MFCCs [5], [11], [13]. It is an argument that Mel-frequency transform is not well motivated for bird sound. To detect human

speech, the Mel-frequency transform is suitable method, as in this high-frequency portion of spectrum is more compressed as compared with the low frequency portion. The most significant information in human speech lies on low frequencies.

Those bird's species which produces the sound at high frequency were degenerated with the help of Mel-frequency transformation. Due to non-robust to noise MFCCs method faced criticism [21], to address this issue many enhanced versions have been proposed.

We can represent a recording in many forms as an assortment of segments and then we can summarize these segments within a feature vector. Those techniques which can be used to extract the feature vector Histogram of frame level is one of them. This is the way how we get the features for a short duration time audio, Briggs [22] gave another probabilistic model, after that Bayes give his risk minimizing classifier, and then he proved that nearest neighbor classifier by using Kullback-Leibler divergence in order to equate histogram of features be thoroughly approximated. Similarly syllable feature's histogram were used in [18], [23], [24].

In the MLSP bird's classification contest of 2013 pattern matching technique has been used to produce feature vectors at the recording level [24]. The set of patterns has been mined automatically based on the training data segmentation into syllables. Every pattern obtained through this method is then compared with the target files. Here, the cross-correlation values will be used as features after normalization and maximization.

## **Chapter 3 – Proposed Work**

### **3.1 Problem statement**

In the context of our multi-instance multi-label classification example such as bird sound classification, multiple birds are vocalizing simultaneously. This requires the extraction of the overlapping birds sounds from their raw sound data. In existing MIML methods, different feature extraction methodologies, e.g. bag of words, are used to convert raw sound data in to meaningful features. However, those feature extraction methodologies lack the ability to separate the overlapping sound segments and thus causing bottleneck in overall system performance improvement. Those features are then fed to any conventional classifier for classification purposes. To address this limited ability of existing features extraction methods for overlapping sound, we will apply matrix factorization methods that have proved to be beneficial in the separation of different sounds. Moreover, matrix factorizations have also successfully been used as classifiers in many classification applications. Hence we will use matrix factorization methods for multi-instance multi-label classification.

### **3.2 Proposed work**

There are different techniques and algorithms used for multi label classification. Our proposed technique for multi label classification is based upon Matrix Factorization methods. In matrix factorization, there are different algorithms such as non-negative matrix factorization (NMF), Independent component analysis (ICA), and dictionary learning (DL) for sparse representation. We will apply dictionary learning techniques for sparse representation for classifiers for the MIML problem.

### **3.3 Expected goals and objectives**

- To asses previous methods for MIML problem.
- To investigate the employability of dictionary learning methods for MIML problem.

- Propose new classifier based on matrix factorization techniques to solve MIML problems.

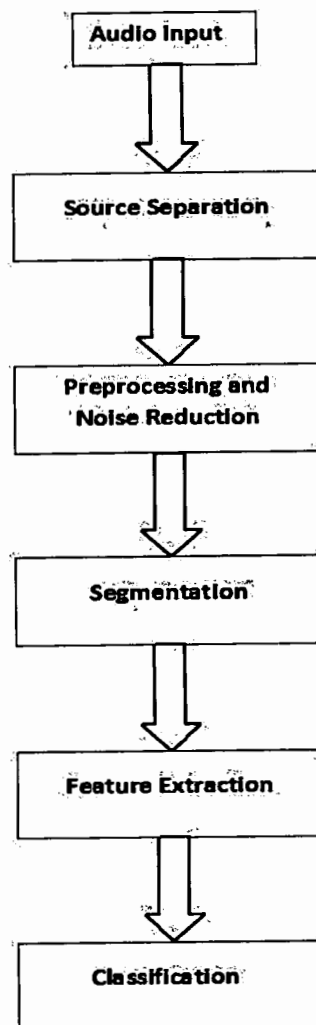
### 3.4 Research Methodology

The audio recording which we use in our work is of length 10 seconds and 10 classes of bird species are present in the recordings, so here the problem with which we are dealing is multi-instance and multi-label. Source separation method is required to find out individual bird call, so this method separates distinct sources that are present in the recordings. Source separation toolkit i.e. FASST toolkit [25] is applied to separate individual birds call. Input of source separation toolkit are these 10 second audio recordings of bird's sounds. In order to segregate birds' individual calls or syllables of birds from a recording of mixture of groups we will use this toolkit. So, first of all by using source separation we do the conversion of a multi label problem into single label problem. Let suppose we have a recording which include three birds, when we apply source separation we will get three different independent recordings which do not depend upon each other as an example of this toolkit. After source separation for each recording we have a single label. For each recording an individual spectrogram is found. Vertical axis comprises frequency information in the spectrogram plot, and brightness representing amplitude, intensity, energy, or loudness and horizontal axis contain the time values. To get spectrogram of audio signal, that signal is divided into overlying frame, each frame consists of a block of successive samples of signal. Then a vector of complex Fourier coefficients is found by applying Fourier transform on each frame. The magnitudes of the coefficients for one frame corresponds to each column of pixels in a spectrogram. In spectrogram instances are shown as segments and species present in recordings as label. Preprocessing on spectrum is applied after finding spectrograms. In step of preprocessing, the portion of spectrogram is cropped in which call or syllables of a bird is present, then noise reduction is applied on these cropped spectrograms. We get filtered spectrogram at the output of preprocessing part. Segmentation is applied on these filtered spectrograms after preprocessing step. The filtered spectrogram is segmented into 2D regions using Random Forest. After segmentation, we found 38 dimensional feature vector for each segment. Mask descriptors, profile statistics, and histogram of gradients (HOG) can be found as a feature vectors, these three types of features describe a piece. As we are using 2D segmentation so we could not find the most commonly used audio feature such as MFCC. These feature characterize type of the segments. For



the classification of the set of species which are present on the recording, a classification algorithm is being used based upon feature vectors. Every step is showed in the block diagram given beneath. Descriptive detail of each step is explained in further topics.

### 3.5 Block diagram



*Figure 3.1: Block diagram of proposed work*

### 3.5.1 Source Separation

Source separation is a problem in signal processing, wherein the amount of signals can be mixed together at a combined signal, and the goal is to restore the original signal component of the combined signal. For example a number of people are talking simultaneously in a room and a hearer is trying to track any one of the negotiations. This can be well handled by human brain to separate or listen the individual voice or talking but in case of digital signal processing or any auditory source separation model it is very difficult. Many audio source separation techniques have been developed. Most of these techniques are established for some specific situation described by their number of sources present and also how many channels involved and may also the sources and mixing process characteristics.

The working of this general flexible source separation framework is presented in “A General Flexible Framework for the Handling of Prior Information in Audio Source Separation”[25].

- First read audio file in Matlab.
- Find the time frequency transform by using “comp\_transf\_Cx”. In this function we find out the content of local sections of a signal i.e. sinusoidal frequency and its phase, as it changes over time so it can be found using short-time Fourier Transform. For computation of short time Fourier Transform, first divide a signal into a shorter time interval i.e. to make segments of signal in to equal length. After finding the frames or shorter segment then apply Fourier transform on each frame or segment.
- After finding time frequency transform of input audio signal we find the mixture structure of the audio signal by using “init\_mix\_struct\_Mult\_NMF\_inst”. The mixture structure could be a Matlab structure that can be used to integrate prior info in to the framework. The structure is a hierarchical organization. Signal representation parameters are called Global parameters. And in the first level of hierarchy these global parameters are defined. Spatial components  $J_{spat}$  and spectral components  $J_{spec}$  are defined on the second level of hierarchy. Each source can be molded by one spectral component, though some sources may also be modeled by numerous spectral components. Moreover, each spectral component must be related with one spatial component, and each spatial component is associated with at least one spectral

component. This implementation could be more general in the sense that the number of spectral components can not necessarily be equal to the spatial components i.e.  $J_{spec} \geq J_{spat}$ . The third level of the hierarchy is factoring each spectral component in one or more factors, which represents, for example excitation and filter structures. In fourth level of hierarchy, each factor can be characterized as the product of three or four matrices. The factor demonstrating excitation structure can either be represented as the product of four matrices i.e.  $W_j^{ex} U_j^{ex} G_j^{ex} H_j^{ex}$  and these matrices represents narrowband spectral patterns, spectral pattern weights, time pattern weights and time-localized patterns or as the product of threes matrices  $W_j^{ex} U_j^{ex} G_j^{ex}$  when  $H_j^{ex}$  is marked by the empty matrix.

- Then we have to estimate the model parameter by using function “estim\_param\_a\_post\_model”. This function takes as input the above calculated mixture structure and finds the parameter which corresponds to that audio source.
- Different audio sources have different spectral and spatial components and different audio sources combined with the different mixing structure. So after estimation of mixing structure, sources are separated by using function “separate\_spat\_comps”. Output of this function is separated audio sources.

### 3.5.2 Preprocessing and Noise Reduction

10-s individual recording were obtained after source separation with the sampling frequency of 16KHz. After recording, the spectrogram of these recordings will be attained by separating the input audio signal to frames and each frame consist of 512 samples and by setting the frame overlay to 50%. Later by the help of Hamming window we apply Fast Fourier transform, to calculate 256 element magnitude spectrum of every frame. Spectrogram’s elements are denoted by  $S(t, f)$ , where  $f$  signify the frequency at that specific time  $t$  and  $t$  represent the index of the frame. In order to get a filter spectrogram, we apply wiener filtering for the noise reduction to the obtain spectrogram.

### 3.5.2.1 Spectrograms

A wave of pressure causes to produce sound. In computer audio data can be stored as quantized signal, in which there is a specific sequence of numbers that actually represents the pressure in term of the function of time. The signal has been sampled at particular frequency. In several techniques the analysis can be done directly on sound as it is comparatively easy to recognize the pattern in spectrogram representation. One of the most fundamental tool for bird sound analysis is spectrogram [4]. In spectrogram time is along horizontal axis, whereas frequency is along vertical axis, and the brightness shows the energy, intensity, loudness or amplitude. More bright points have large values and less bright point has little values. Therefore, it can be concluded that spectrograms are very beneficial technique for imagining the three dimensional data (amplitude, frequency, time).

More appropriately, spectrograms is found by dividing the signal in to overlying frames, every frame contains of a block of successive examples of the signal. After that FFT is applied on every frame to generate a vector of complex Fourier coefficients. In a spectrogram respective column of pixels links the magnitudes of coefficients for one frame [2].

### 3.5.2.2 Noise Reduction

For the reduction of noise and escalate the divergence of the sound of bird, we first of all regulate  $S(t, f)$  in range  $[0, 1]$ , and subsequently compute  $S_1(t, f) = \sqrt{S(t, f)}$  to every constituents of the spectrogram. Double iterations of the whitening filter can be applied on that spectrogram. The primary goal is to evaluate frequency profile from low energy frames of noise, and then attenuate every pixel of spectrogram using this frequency profile. The filter is:

- 1) Amount calculated which analogous to every frame's energy at  $t$ , as

$$E(t) = \left(\frac{1}{f_{max}}\right) \sum_{f=1}^{f_{max}} S_1(t, f)^2. \text{ Then we arrange the frames respective to } E.$$

- 2) For all frequencies  $f$ , compute  $P(f) = \sqrt{\mathcal{E} + \sum_{t \in N} S(t, f)^2}$ , where  $\mathcal{E} = 10^{-10}$ . We added  $\mathcal{E}$ , to avoid division by zero.

3) For every value of  $(t, f)$  we practice the noised reduced spectrogram as  $S_2(t, f) = S_1(t, f)/P(f)$ .

Figure 3.3 and Figure 3.2 shows before and after applying whitening filter.



*Figure 3.3: Before noise removal*



*Figure 3.2: After noise removal*

### 3.5.3 Segmentation

Bird song has a configuration which comprises of single articulation syllables, which can be used to make songs. The structure of songs changes among different species of birds, but individual structure of syllables remains constant. Therefore, these level of syllable are the basic core concept of many methods [4].

To segregate each syllable of birds, call segmentation is required in input soundtrack. By energy-based time domain segmentation, low-noise single-bird recordings have successfully been worked out. Though, in our proposed work too much noise is there because the soundtrack which we use

is recorded on the field sound recording system. In our case strong segmentation is required. Therefore, we will use "Random Forest Classifier" for time frequency audio segmentation.

After preprocessing we get filtered spectrogram. After that filtered spectrograms is used for segmentation. Syllables present in the recording of bird call are separated that may overlay in the time. 2-D time frequency segmentation is used for that purpose.

### 3.5.3.1 2D segmentation

In utmost segmentation procedures we estimate the final and initial time of the syllable. However that is observed as 1-Dimensional methodology. Alternatively, we can select the 2-D region or bounding box present in the spectrogram consistent to syllable. The benefit of using two dimensional segmentation method is that, it might help in separating sounds that overlay in time. Then apply a random forest classifier to spot each pixel in the spectrogram as noise or bird sound [26]. In order to do that we link the feature vector to every pixel in the spectrogram which depict the oblong cover adjacent to them. For every value of  $(t, f)$  in the spectrogram, we compute features vector  $x(t, f)$ :

- The spectrum-bin index  $f$ .
- In a oblong surrounding  $S(i, j)$  the spectrogram's components have a value i.e.  
 $i \in [t - t_w, t + t_w], j \in [f - f_w, f + f_w]$ , where  $t_w = 6$  and  $f_w = 12$ .

In order to train the classifier used for segmentation, we manually mark spectrograms as illustration of precise segmentation. The mask  $M(t, f)$  of spectrogram can be well-defined as, white if mask is zero i.e.  $M(t, f) = 0$ , the element of spectrogram in that region is background noise. Black if mask is equal to one i.e.  $M(t, f) = 1$ , the element of spectrogram in that region is syllable or voice of bird. RF classifier is used to get the list of sets  $(x_1, y_1), \dots, (x_n, y_n)$  as the training data. We take these set by randomly select 50,000 points from the spectrograms which is manually interpreted. The points are randomly selected in such a way that there are 10% positive and 90% negative cases. We find feature vector of every selected set as  $x_i = (t_i, f_i)$ . Each example used for training is denoted by  $y_i = M(t_i, f_i)$ . So now this is two class problem with label 0 and 1. Forty trees are used to train Random forest classifier. This method is in fact a collective of decision tree.

Random forest produce the probability  $P(\frac{y}{x})$  for the feature vector associate with each class  $y$ , that is volume of the trees which provides label of  $y$  for particular input  $x$ . The probability of each pixel  $(t, f)$  present in spectrogram is computed by using random forest, i.e.  $P(y = 1/x(t, f))$ . Then we smooth these probabilities by convolving them with Gaussian kernel i.e.  $g(t, f) = P(y = 1/x(t, f)) * P$ . Where Gaussian kernel is denoted by  $P$  with  $\sigma = 3$  for  $17 \times 17$  box. Finally, forecasted segmentation mask  $M(t, f)$  of spectrogram is found by involving the threshold of  $\theta = 0.2$ , as illustrated in Figure 3.4.



Figure 3.4: Segmented spectrogram

### 3.5.3.2 Random Forest

It is the cooperative classifier that involves number of decision trees. For a given set of the training inputs examples  $F$ , each tree  $v_j$  in this classifier regardless of the others is constructed from the bootstrap example that could be selected in place of input examples  $F$ . Trees are constructed by sequentially applying subsequent scheme:

- Select input training examples  $F_i = (s, l)$ , where feature vector is represented by  $s$  and the input is labeled by  $l$ .
- A leaf node is generated with the value  $l$ , if each label  $l$  is analogous.
- Take any subset  $J$  of  $\log_2(m) + 1$  features, where number of features is shown by  $m$ .

- For each feature  $d \in J$ , sort  $F$  on  $d$  and then find threshold value  $\theta_d$  on the base of which we have to be found two sets of  $F$  as follow  $F_{left}$  and  $F_{right}$ , as to maximize the Gini index  $G(F_{left}, F_{right})$ .

For maximum of  $G$  choose feature vector and the threshold value  $(d, \theta_d)$ . The leaf node could be labeled with majority label when all probable value of  $G$  are identical. If this is not a case then use  $F_{left}$  and  $F_{right}$  as an input and again following the similar process.

Each inner node of a Random Forest tree associates to the test of form  $s_d < \theta$ . The leaf node would be formed by navigating the tree for any feature vector  $t$  which wholly comprises the similar class label for each of them. Each decision tree in this classifier gives its impact for classifying each input feature vector  $s$ . The forecasted label of the input feature vector  $s$  is considered to be the fraction of the tree that gives the likelihood for  $l$  [26].

### 3.5.4 Feature Extraction

In pattern recognition or classification, the key principle is that for every input data predict the class based on training data. At this time from soundtrack syllable of bird sound removed and after that calls of birds are used as pattern. After the removal of features from the syllables which were specified, these features were used in classification. Very important information related to syllable is present in features. Features mining can be done in three parts listed below:

1. Extraction of feature from the raw data.
2. Removal of the outline because of which wrong data become clear.
3. Feature normalization is done in which every feature vector is being adjusted to a dynamic range so that it can significantly contribute in classification process.

Non normalized feature can also be used for the training of classifier but that requires more time and data for training.

From the input data numerous feature vectors can be extracted but their condensation is mandatory. The minimization of the number of feature vectors in classification is because of



several reasons. Because of the low number of feature vectors the system which we used will be less complex and also required less computation time to execute. Sensitivity to noise of the classification system decreases at the same time significantly improve the generalization property. Redundant features can may generate noise in it and can also effect the process of classification.

The selection of only relevant feature vector is key point in such type of classification system. Selection of only those feature in which there is very high interclass discrimination power are the main reason but also it has very small intraclass discrimination supremacy. The primary purpose of figuring out these discriminative supremacy of the feature vector could be that they will state us that in what way features differentiate between classes. This power of vector helps us in selecting the feature vector. In some classes some feature vectors have high discriminative power as the number of classes is higher than the collection of feature vectors.

For sound classification or recognition, system required those features in which maximum useful information is present. Human can sense the sound by perceptual features. These features includes brightness, pitch and loudness. There is great deal of association in-between perceptual feature and physical feature. These feature are classified in two classes. Those feature which are in time domain have zero crossing rate and signal energy could directly be calculated from sound wave form. While it cannot be find directly from sound for the spectral feature. For spectral feature first of all audio signal is converted into frequency domain this can be done by applying Fourier transform then from that transformed signal features would be extracted.

The frame based features vectors are very easy to calculate in speech analysis and signal processing because for frames the quantity of data and extent of changes is reduced. First, calculate the overlapped features using syllables then find the features using windowed frames and complete the path of the features of syllable. After that, means and variance have been calculated for these paths and hence actual features have been calculated from the basic features. At the end we have a feature vector comprises of mean and variance of frame based features and intact syllable to calculate the parameters [3].

Before finding the feature vector of a segment we need to crop the spectrogram to that portion where only that segment is present. In this way we have the entire segment cropped in the time but it is not cropped in frequency. In order to obtain unique segments for feature extraction and to avoid overlaps in the segments we made all those cropped spectrogram zero which doesn't lay in the mask. To relate the segment features specific notations are used: for a segment.

- $M_c(t, f)$  is the binary mask
- $S_c(t, f)$  is the cropped noise reduced spectrogram.

Here the value of  $t$  lie between 1 to entire interval of the segment  $T$ .

### 3.5.4.1 Categories of features

Three types of features mentioned below described the feature vector completely:

1. Mask descriptors
2. Profile statistics
3. Histogram of gradients (HOG)

In this case as we are dealing with 2D segmentation hence we do not estimate most common or extensively used features like MFCCs. The information obtained from mask-based features can also be acquired by the shape as these features describe the shape or forms of segment. That statistical profile is same as obtained earlier from bioacoustics for noisy environment based on two dimensional segmentation [27].

#### 3.5.4.1.1 Mask Descriptors

The feature we are dealing with for segment, based on the mask and provides thorough information about figure of that segment. These features are mentioned below:

1. *minimum-frequency* =  $\min \{f: M_c(t, f) = 1\}$
2. *max-frequency* =  $\max \{f: M_c(t, f) = 1\}$
3. *bandwidth* = *max-frequency* – *minimum-frequency*
4. *segment duration* =  $T$
5. *total area* =  $\sum_{t,f} M_c(t, f)$

6.  $perimeter = \frac{1}{2} (\# \text{ of pixel in } M_c \text{ such that at least one pixel in the surrounding } 3 \times 3 \text{ box is 1 and atleast one pixel is 0 } )$
7.  $non\text{-compactness} = perimeter^2 / total \text{ area.}$
8.  $rectangularity = \frac{area}{(bandwidth \times duration)}$

### 3.5.4.1.2 Profile Statistics

Once we get these features we will estimate another feature vector that contains the whole information about the arithmetic properties of segments for its both temporal and spatial profile. To find the frequency or time profile rows and columns of spectrogram are added together. The time profile is given as:

$$p_t(t) = \sum_f S_c(t, f)$$

And the frequency profile is represent as

$$p_f(f) = \sum_t S_c(t, f)$$

To interpret these frequency and time profile as a function of mass probability we normalized these profiles to 1. Using the Gini index two types of features can be estimated for uniformity of these densities.

$$freq\text{-gini} = 1 - \sum_f P_f(f)^2$$

$$time\text{-gini} = 1 - \sum_t P_t(t)^2$$

Here  $P_t$  and  $P_f$  represent normalized profile mass densities. By computing central moments  $K_{th}$  of both spectral and temporal profile we would estimate further features for each segment. However, based upon the duration of every segment (which vary segment to segment) these features will be calculated in rescaled coordinate system, considering time and frequency change from 0 to 1.

1. *frequency-mean* =  $\mu_f = \sum_{f=1}^{f_{max}} P_f(f)(f/f_{max})$ .
2. *frequency-variance* =  $\sum_{f=1}^{f_{max}} P_f(f)(\mu_f - f/f_{max})^2$ .
3. *frequency-skewness* =  $\sum_{f=1}^{f_{max}} P_f(f)(\mu_f - f/f_{max})^3$ .
4. *frequency-kurtosis* =  $\sum_{f=1}^{f_{max}} P_f(f)(\mu_f - f/f_{max})^4$ .
5. *time-mean* =  $\mu_t = \sum_{t=1}^T P_t(t)(t/T)$ .
6. *time-variance* =  $\sum_{t=1}^T P_t(t)(\mu_t - t/T)^2$ .
7. *time-skewness* =  $\sum_{t=1}^T P_t(t)(\mu_t - t/T)^3$ .
8. *time-kurtosis* =  $\sum_{t=1}^T P_t(t)(\mu_t - t/T)^4$ .

The maxima of time and frequency profile is also estimated.

1. *frequency-max* =  $(\arg \max P_f(f))/f_{max}$ .
2. *time-max* =  $(\arg \max P_t(t))/T$ .

Using this mask value standard deviation and mean of the spectrogram also be calculated by using following formula,

1. *mask-mean* =  $\mu_{tf} = (1/area) \sum_{tf} S_c(t, f)$ .
2. *mask-stddev* =  $\sqrt{(1/area) \sum_{tf} (\mu_{tf} - S_c(t, f))^2}$ .

### 3.5.4.1.3 Histogram of gradients

The histogram gradients feature vector of every segmented part has also been calculated in this study. The cropped spectrogram and its respective mask  $S_c(t, f)$  and  $M_c(t, f)$  will be use as input.

First of all, these segmented spectrograms are convolved with  $7 \times 7$  Gaussian kernel  $K$  where

$$\sigma^2 = 4 \quad \text{To get} \quad S_b(t, f) = S_c(t, f) * K$$

from a blurred spectrogram of segment. The gradients of  $(t, f)$  points are found by convolving Sobel kernel to the  $S_b$  as given below

$$(d/dx)S_b(t, f) = S_b(t, f) * D_x \quad \text{and} \quad (d/dy)S_b(t, f) = S_b(t, f) * D_y$$

where

$$D_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad D_y = D_x^T$$

After calculating this for every pixel of spectrogram of a segment lie in the mask

$$[\text{i.e., } M_c(t, f) = 1]$$

now we move further to calculate

$$\nabla S_b(t, f) = ((d/dx)S_b(t, f), (d/dy)S_b(t, f))$$

Pixels only with  $\|\nabla S_b(t, f)\|^2 \geq 0.01$  endow to the histogram. This histogram consist of 16 equally spaced bins for complete range of the angles  $[0, 2\pi]$ . This feature vector contains normalized count for every bin and its corresponding number of bin.

Therefore, we obtain 16 dimensional HOG (histogram for gradient feature) for every given segment.

### 3.5.4.2 Feature rescaling

All calculated feature vectors for each segment up till now is combined to get single feature vector. These features in vector varies in values and ranges. To overcome this limitation of feature vector we normalize this vector between 0-1 to make it feasible to use for all classifiers.

## 3.6 Classification

Two different learning methods have been used for the classification of different species of bird i.e. K-SVD and Fisher discrimination dictionary learning. To find out discriminative dictionary and sparse coefficients we used feature vectors as training data set. Now using both the coefficient and learned dictionary we calculated recognition rate for our testing dataset (testing features). For FDDL method this classification rate was found out using reconstruction error however, test signal coefficient's sparsity level is being used for K-SVD recognition rate estimation.

### 3.6.1 Fisher discrimination dictionary learning

Here in this work a technique which is known as discriminative DL frame work is used which learn the structured dictionary by fisher discrimination criterion. In this proposed discriminative DL frame work atoms of dictionary is in corresponding to each class labels so for classification the reconstruction error related to each class is used. In order to make work more efficient fisher discrimination criteria is also incorporated with them. In the discriminative DL process, sparse coding coefficients are modified database such that having small value within class scatter and having large value between class scatter. From whole, we make a sub dictionary for specific class which have a better representation of the training samples for corresponding class but should have poor representation for other classes. Reconstruction error and coding coefficient will be found with the proposed fisher discriminative based DL method. This is totally a new classification technique which is develop here and utilized to birds species classification to assess its performance.

We develop a structure dictionary  $D = [D_1, D_2, \dots, D_c]$ , where  $D_i$  represent sub dictionary related to specific class  $i$  and  $c$  denotes the whole set of classes. By using this  $D$ , reconstruction error could be used for classification. Let  $A = [A_1, A_2, \dots, A_c]$  represent the set of training samples, where  $A_i$  represent the sub set of samples (training) that is associated with class  $i$ . The coefficient matrix of  $A$  over  $D$  can be represented by  $X$ , i.e.  $A \approx DX$ . We can write coding coefficient matrix  $X$  as  $X = [X_1, X_2, \dots, X_c]$ , where  $X_i$  represent the sub matrix that contains coding coefficients of  $A_i$  over  $D$ . The FDDL objective function is represented as:

$$J_{(D,X)} = \underset{(D,X)}{\operatorname{argmin}}\{r(A, D, X) + \lambda_1 \|X\|_1 + \lambda_2 f(X)\} \quad (1)$$

$r(A, D, X)$  represents discriminative fidelity term, and sparsity constraint is represented by  $\|X\|_1$ ,  $f(X)$  is discrimination constraint forced on coefficients matrix  $X$  and  $\lambda_1$  and  $\lambda_2$  are scalar parameters.

### 3.6.1.1 Discriminative fidelity term

We can write coding coefficient matrix as  $X_t = [X_t^1, \dots, X_t^j, \dots, X_t^c]$  where  $X_t^j$  represents the coding coefficient of  $A_t$  over the sub dictionary  $D_j$ . The exemplification of  $D_k$  to  $A_t$  is given as  $R_k = D_k X_t^k$ . First, the dictionary  $D$  have ability to well exemplify  $A_t$  that is  $A_t \approx DX_t = D_1 X_t^1 + \dots + D_t X_t^t + \dots + D_c X_t^c = R_1 + \dots + R_t + \dots + R_c$ . Second, as  $D_t$  is associated with the  $i^{th}$  class, it is also understandable that  $A_t$  must entirely characterized by  $D_t$  but not well characterized by  $D_j, j \neq i$ . It is also required that  $X_t^i$  must have some vital coefficients such that  $\|A_t - D_t X_t^i\|_F^2$  is small, while  $X_t^j$  should have nearly zero coefficients such that  $\|D_j X_t^j\|_F^2$  is small. Thus discriminative fidelity term is defined as:

$$r(A_t, D, X_t) = \|A_t - DX_t\|_F^2 + \|A_t - D_t X_t^i\|_F^2 + \sum_{\substack{j=1 \\ j \neq i}}^c \|D_j X_t^j\|_F^2 \quad (2)$$

Figure 3.5 show the explanation of these three terms in  $r(A_t, D, X_t)$ . Left most side shows that even if  $D$  is guaranteed to represent  $A_t$  well,  $R_t$  might diverge much from  $A_t$  so that  $D_t$  could not well represent  $A_t$ . Better discrimination is achieved, by considering another constraint that  $\|A_t - D_t X_t^i\|_F^2$  is small, shown in middle part of Figure 3.5. However,  $A_t$  can also be well characterized by other sub-dictionaries, e.g.  $D_{t-1}$ , due to which discrimination capability of  $D$  is reduced. By using third constraint which is representation of  $D_j, j \neq i$ , to  $A_t$  is small, so this discriminative term can overwhelmed this problem, which is shown in right most side of Figure 3.5.

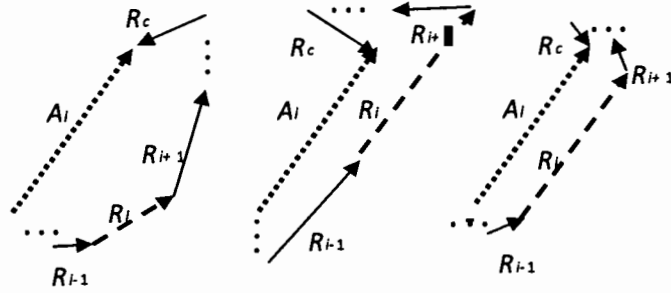


Figure 3.5: Illustration of discriminative fidelity term

### 3.6.1.2 Discriminative coefficient term

To create dictionary  $D$  be discriminative we can also make the coding coefficient  $X$  be discriminative. This is attained by decreasing the within class scatter of  $X$  which is represented as  $S_w(X)$ , and increasing the between class scatter of  $X$ , which is represented as  $S_B(X)$  and defined as,

$$S_w(X) = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - m_i)(x_k - m_i)^T$$

$$S_B(X) = \sum_{i=1}^c n_i(m_i - m)(m_i - m)^T$$

The mean vector of  $X_i$  and  $X$  are represented as  $m_i$  and  $m$  and the total number of examples in  $A_i$  is represented as  $n_i$ .

Now discriminative coefficient term be defined as  $f(X)$  as  $tr(S_w(X)) - tr(S_B(X))$ . But this function  $f(X)$  is unstable and non-convex. This problem can be solved by adding the elastic term  $\|X\|_F^2$ . So the discriminative coefficient term can be denoted as

$$f(X) = tr(S_w(X)) - tr(S_B(X)) + \eta \|X\|_F^2 \quad (3)$$

Where  $\eta$  is a parameter.

### 3.6.2 Model of FDDL

Complete FDDL model is obtain by combining Eqs. (2) and (3) into Eq. (1) as given below:



$$J_{(D,X)} = \underset{(D,X)}{\operatorname{argmin}} \left\{ \begin{array}{l} \sum_{i=1}^c r(A_i, D, X_i) + \lambda_1 \|X\|_1 + \\ \lambda_2 (\operatorname{tr}(S_w(X)) - \operatorname{tr}(S_B(X)) + \eta \|X\|_F^2) \end{array} \right\} \quad (4)$$

The objective function defined above in Eq. (4) is convex with respect to each of  $D$  and  $X$  when the other is fixed but is not mutually convex to  $(D, X)$ . Therefore, an algorithm can be design to alternatively optimizing  $D$  and  $X$ .

### 3.6.3 Optimization of FDDL

The objective function of FDDL derived in Eq. (4) can be categorized into two sub-problems. First is, updating  $X$  by fixing  $D$ , and second is updating  $D$  by fixing  $X$ . The algorithm is iteratively implemented for the desired dictionary  $D$ , and the coding coefficients  $X$ .

First, the objective function  $J_{(D,X)}$  in Eq. (4) is reduced to sparse coding problem to find  $X = [X_1, X_2, \dots, X_c]$  by fixing  $D$ . We find  $X_i$  class by class. When we are computing  $X_i$ , all  $X_j, j \neq i$ , are fixed. So the objective function is further reduced to:

$$J_{(X_i)} = \underset{(X_i)}{\operatorname{argmin}} \{r(A_i, D, X_i) + \lambda_1 \|X_i\|_1 + \lambda_2 f_i(X_i)\} \quad (5)$$

With

$$f_i(X_i) = \|X_i - M_i\|_F^2 - \sum_{k=1}^c \|M_k - M\|_F^2 + \eta \|X_i\|_F^2$$

Where mean vector matrix of class  $k$  is denoted by  $M_k$  and mean vector matrix of all classes is denoted by  $M$ . If we want that  $f_i(X_i)$  not only convex but also must have enough discrimination, we set  $\eta = 1$ . Eq. (5) can be solved by employing the Iterative Projection Method.

Now dictionary  $D_i$  is updated class by class by fixing coding coefficients . When updating  $D_i$ , all  $D_j, j \neq i$ , are fixed. Now the objective function in Eq. (4) is reduced to:

$$J_{(D_i)} = \left\{ \begin{array}{l} \|A - D_i X^i - \sum_{j=1, j \neq i}^c D_j X^j\|_F^2 + \\ \|A_i - D_i X_i^i\|_F^2 + \sum_{j=1, j \neq i}^c \|D_i X_j^i\|_F^2 \end{array} \right\} \quad (6)$$

Where  $X^i$  denotes the coding coefficients of  $A$  over  $D_i$ . Eq. (6) is a quadratic programming problem and it can be solved by using the technique described in [28], which updates  $D_i$  atom by atom.

### 3.6.4 Classification scheme

After finding the dictionary, the testing examples can be classified through coding it over the learned dictionary. We code the testing examples  $y$  over the whole dictionary  $D$ . Sparse coding coefficients could be got by solving  $\hat{\alpha} = \{\|y - D\alpha\|_2^2 + \gamma \|\alpha\|_1\}$  where  $\gamma$  is a constant, and  $\hat{\alpha} = [\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_c]$ , where  $\hat{\alpha}_i$  is the coefficient vector associated with each sub dictionary  $D_i$ . Final classification model is defined as:

$$e_i = \|y - D_i \hat{\alpha}_i\|_2^2 + w \cdot \|\hat{\alpha} - m_i\|_2^2$$
$$\text{identify}(y) = \text{argmin}\{e_i\}$$

First term represents the reconstruction error for class  $i$ , the second term represents the distance between the coefficient vector  $\hat{\alpha}$  and the mean vector  $m_i$  of class  $i$  and  $w$  is the weight to balance the two terms.

### 3.6.5 K-Singular value Decomposition (K-SVD)

It is used as a dictionary learning algorithm for making a dictionary for sparse representations. Singular value decomposition technique is used by K-SVD. It is a simplification of k-means which is a clustering method. It works by iteratively interchanging between sparse coding the input data based on the current dictionary, and updating dictionary atoms to better fit data. K-SVD is used in many applications such as audio processing, image processing, document analysis and biology [29]. The objective function is given by:

$$\min_{D, X} \{\|Y - DX\|_F^2\} \text{ subject to } \forall i, \|x_i\|_0 \leq S_0$$

$D$  represent the dictionary,  $S_0$  shows the sparsity level and  $X$  represent the sparse coefficients.

K-SVD has two stages. First is sparse coding stage in which we find best coefficient matrix and the second stage is to find a better dictionary.

- Sparse coding stage:

In this stage, the objective is to find the optimum sparse coefficients while dictionary  $D$  is fixed. So above objective function is reduced to sparse coding problem and is given by

$$\min_{x_i} \{ \| y_i - Dx_i \|_2^2 \} \text{ subject to } \| x_i \|_0 \leq S_0$$

So by using any pursuit algorithm, we can find the sparse coefficients  $x_i$  for each example  $y_i$ . Where  $i = 1, 2, 3 \dots N$  and the total number of examples is represented as  $N$ .

- Dictionary update stage:

After finding the sparse coefficients in the first stage, now we find better dictionary. In this stage coefficients  $X$  and dictionary  $D$  is fixed and at a time only one column of dictionary i.e.  $d_k$  and its corresponding coefficients  $x_T^k$  i.e.  $K^{th}$  row in  $X$  is updated. So the objective function can be rewritten as,

$$\begin{aligned} \| Y - DX \|_F^2 &= \| Y - \sum_{j=1}^k d_j x_T^j \|_F^2 \\ \| Y - DX \|_F^2 &= \| (Y - \sum_{j \neq k} d_j x_T^j) - d_k x_T^k \|_F^2 \\ \| Y - DX \|_F^2 &= \| E_k - d_k x_T^k \|_F^2 \end{aligned}$$

Where  $E_k$  is the representation error matrix. Now apply singular value decomposition on  $E_k$  to find dictionary column and its corresponding coefficient.

$$E_k = U \Delta V^T$$

After applying SVD decomposition, choose the first column of  $U$  as an updated dictionary column  $d_k$  and first column of  $V$  multiplied by  $\Delta(1,1)$  is selected as an updated coefficient vector  $x_T^k$ .

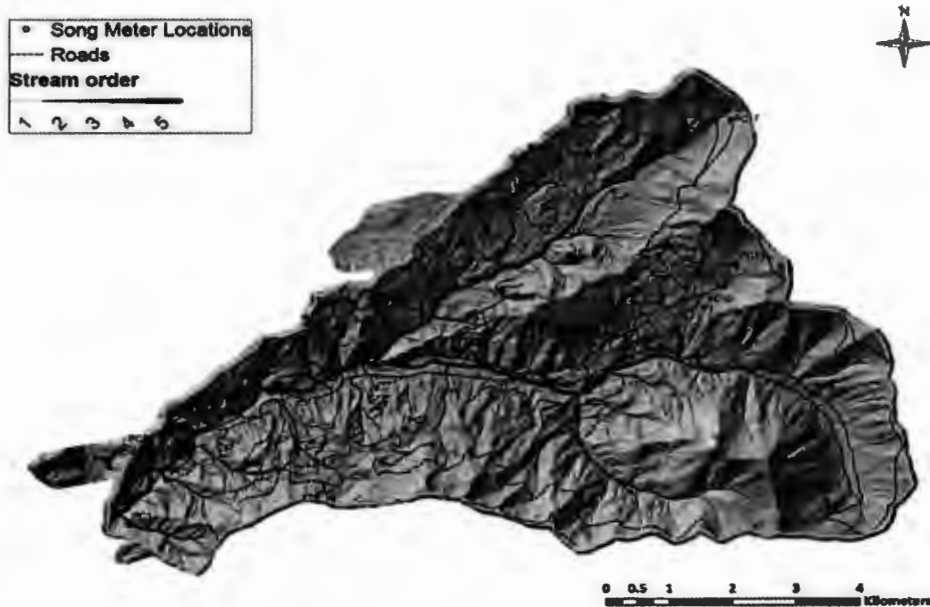
## Chapter 4 - Experiment

### 4.1 Data Set

The dataset is in the form of audio recordings which were collected from Long-Term Experimental Research Forest, by H. J. Andrews (HJA) that is situated in the Oregon's Cascade mountain range. The data is collected by using song meter which contains microphone which records the sounds and there is almost 10TB of audio data. This data was recorded in .wav format using song meter, in flash memory. This song meter has two omnidirectional microphones that can be place in the field.

HJA is basically a site that contains number of experiments and the sources to collect the data can be ecological, geological, and meteorological. Information about the environment such as vegetation composition, elevation, weather etc., can be obtained from this data which is collected using song meter. This data has wide application in research work, discovery and analysis.

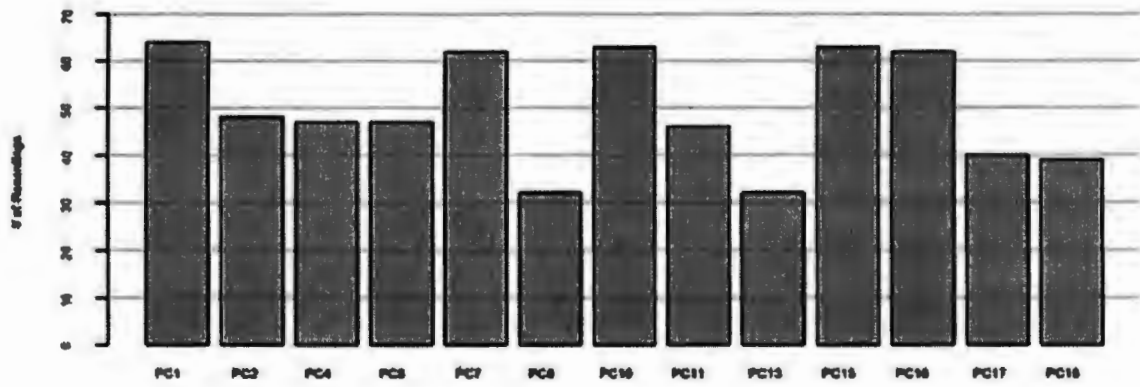
In this study, we consider the database from 2009 and 2010 that contains representative samples of HJA and have 13 sites illustrated in Figure 4.1. This data set surrounds the wind, rain or recordings in which no bird exists. The whole data set contains the 645 recordings of length 10 second in WAV format. Table 1 represents 10 different bird's species that are present in data set. The division of 645 soundtracks would be reasonable enough for providing thorough information to all sites as illustrated in Figure 4.2.



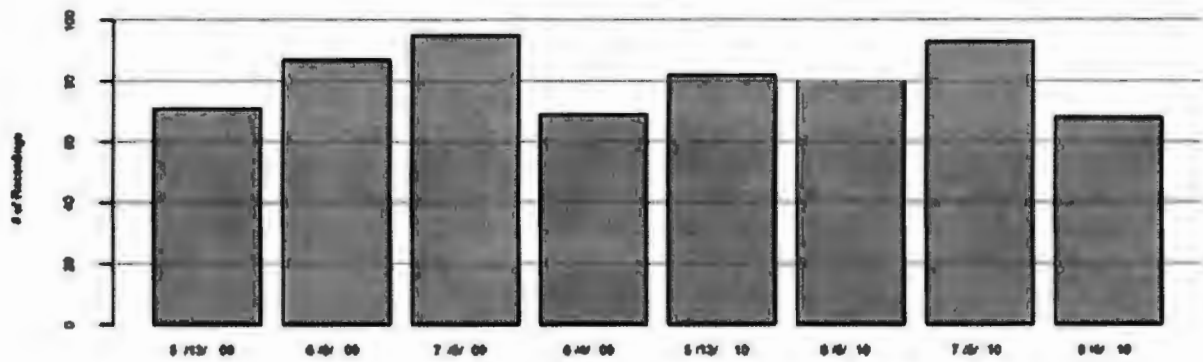
*Figure 4.1: Song meter data collection locations*

In the recording, every soundtrack is categorized with some specie set. The assigning of label is after the observing of sound by spectrogram and sound listening. Experts listen the sound give some label to it as per their own knowledge keeping in view the confidence level corresponding to it. For the final labeling of recording the corresponding confidence level and expert labeling is kept in mind. In WAV file info is encoded that contains the time, date and location (1/13 sites) of recording [24].

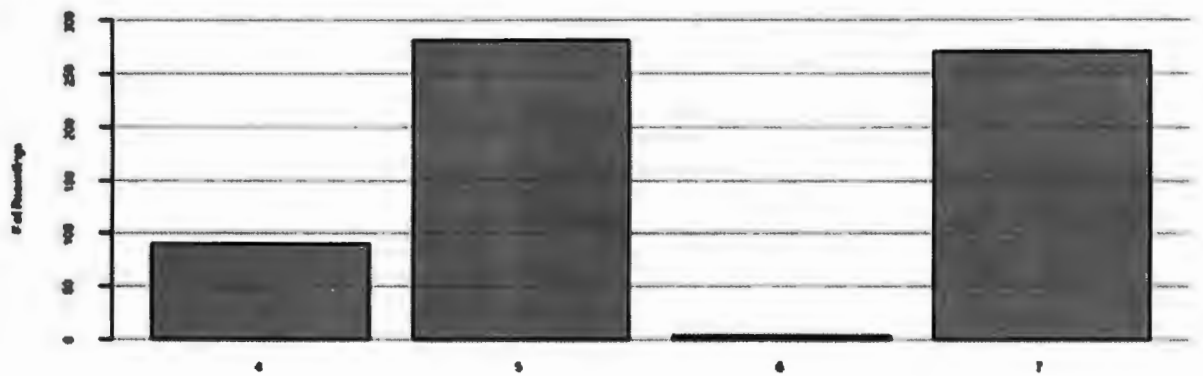
The 16 KHz sampling rate provided for audio recording. For spectrogram obtaining FFT is applied, the range of frequencies is 0-8KHz. In HJA for the recording of sound this frequency range is enough. It is likely that some birds sound mislaid due to this sampling frequency, but with limitation the model that we have proposed still work for the species of birds that is recognized.



(a) Location code



(b) Date



(c) Hour of day (AM)

Figure 4.2: Recordings in the dataset, counted in different ways [24].

For training and evaluating the proposed algorithm to forecast the samples of species that is present in the soundtrack we require some training data set which contain their corresponding labels. We obtain the data from the site which contain data for months therefore it is not possible to give label to all of data. All soundtracks are recorded in time range from 5:00 am to 5:20 am because birds are active at this time of day. Most of the soundtrack contain the multiple articulation of different species of birds.

Table 1 represent the recorded sound tracks of 10 species of birds. Every sound track contains sounds of 1-3 species. So, in average 2.14444 species are there in each sound track. Those samples which don't contains any voice of the bird have been removed in the segmentation process so at the end we have only those files that comprise at least single voice.

*Table 4.1: The 10 bird species in the data set*

Code	Name
0	Brown Creeper
1	Dark-eyed Junco
2	Hermit Thrush
3	Chestnut-backed Chickadee
4	Varied Thrush
5	Hermit Warbler
6	Swainson's Thrush
7	Hammond's Flycatcher
8	Stellar's Jay
9	Common Nighthawk

Durations of audio tracks can be changed by our choice. The syllable present in the audio track will get reduced if we increase the syllable period of audio tracks. The boundary of recording of data set is cutoff which we consider for training the segmentation algorithm. Although, if the duration period of all the audio tracks is increased then there are far more chances for the bag to identify all the sound of birds at any particular site. If we consider an audio track and every specie present in that track is labeled with a bag and pass this data for learning then, it would be very difficult to for the learning technique to learn [2].

Noise, sound of insects, air and other distortions were present in the audio track because it was recorded from a distance of 1Km.

## 4.2 Experimental setup

We apply the proposed method on bird species dataset. We use total 10 species of birds so we have total 10 number of classes. We divide the data set in to training and testing data set which is used by our proposed classifier. All the training and test data is converted in to feature vectors using pre-processing steps which includes, first finding of spectrogram of each recording, 2-D segmentation is applied on these spectrograms using random forest classifier, noise is removed from these segmented spectrogram using wiener filter and then 38 dimensional feature vector is find out of each segment. The detail of each step is given in chapter 3. As a result, the training data set contains total 187 examples that is used for training the classifier and test data contains total 87 examples for which we predict the class label. Each example has 38 dimensional feature vector. The important parameter in Fisher discrimination dictionary learning (FDDL) is the number of atoms in  $D_i$ . The number of dictionary atoms in FDDL on each class is set as the total number of training examples. So the size of dictionary  $D$  is  $187 \times 38$ . The other parameters of FDDL are lambda 1 which is the scalar parameter of coefficients and lambda 2 is the scalar parameter of fisher discriminative coefficient term and they are set to  $\lambda_1 = 0.15, \lambda_2 = 0.015, \gamma = 0.005$  and  $w = 0.05$ . Where  $w$  shows the weight to balance the equation. The dictionary is found by optimizing the objective function i.e.

$$J_{(D,X)} = \underset{(D,X)}{\operatorname{argmin}} \{r(A, D, X) + \lambda_1 \|X\|_1 + \lambda_2 f(X)\}$$



After finding the dictionary, the testing examples can be classified through coding it over the learned dictionary. Sparse coding coefficients are found by solving  $\hat{\alpha} = \{\|y - D\alpha\|_2^2 + \gamma \|\alpha\|_1\}$  where  $\gamma$  is a constant, and  $\hat{\alpha} = [\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_c]$ , where  $\hat{\alpha}_i$  is the coefficient vector associated with each sub dictionary  $D_i$ . Final classification model is defined as:

$$e_i = \|y - D_i \hat{\alpha}_i\|_2^2 + w \cdot \|\hat{\alpha} - m_i\|_2^2$$

$$\text{identify}(y) = \text{argmin}\{e_i\}$$

First term represents the reconstruction error for class  $i$ , the second term represents the distance between the coefficient vector  $\hat{\alpha}$  and the mean vector  $m_i$  of class  $i$  and  $w$  is the weight to balance the two terms.

In case of KSVD the dictionary and sparse coefficients are found by iteratively optimizing the objective function,

$$\min_{D, X} \{\|Y - DX\|_F^2\} \text{ subject to } \forall i, \|x_i\|_0 \leq S_0$$

Where  $S_0$  shows the sparsity level. After finding the dictionary and sparse coefficients, the classification is performed on the basis of sparse representation of test signals.

## Chapter 5 -

## Results

### Results

In the proposed work we have done the classification on bird's species audio data using two dictionary learning techniques. We apply two techniques on these data set namely Fisher discrimination dictionary learning (FDDL) and K-Singular value decomposition (K-SVD). In case of Fisher discrimination dictionary learning, we calculated the recognition rates on different values of sparsity levels. The sparsity level of sparse coefficients of data set is tuned by the lambda parameter. So by changing the lambda we get different recognition rates which is shown in Table 5.1. From Table 5.1 and Figure 5.1 we can observe that achieved recognition rate at maximum is 63 %. In case of FDDL, classification is done by using the reconstruction error for each class. In case of K-SVD we also take classification rates on different values of sparsity levels. Table 5.2 shows recognition rates of K-SVD. In Table 5.2, columns shows the classification rates on different sparsity level of dictionary denoted as  $D$  and rows shows the classification rates on different sparsity level of coefficients denoted as  $L$ . As we can see in first column of Table 5.2 in which sparsity level of dictionary is  $D=3$ , we get maximum classification rates than other values of  $D$ . And also at sparsity of coefficient i.e.  $L=3$  we get maximum recognition rate i.e. 57%.

*Table 5.1: Recognition rates for FDDL*

Lambda	Recognition rate
0.005	0.54023

0.006	0.551724
0.007	0.586207
0.008	0.563218
0.009	0.586207
0.010	0.574713
0.011	0.597701
0.012	0.609195
0.013	0.586207
0.014	0.574713
0.015	0.632184

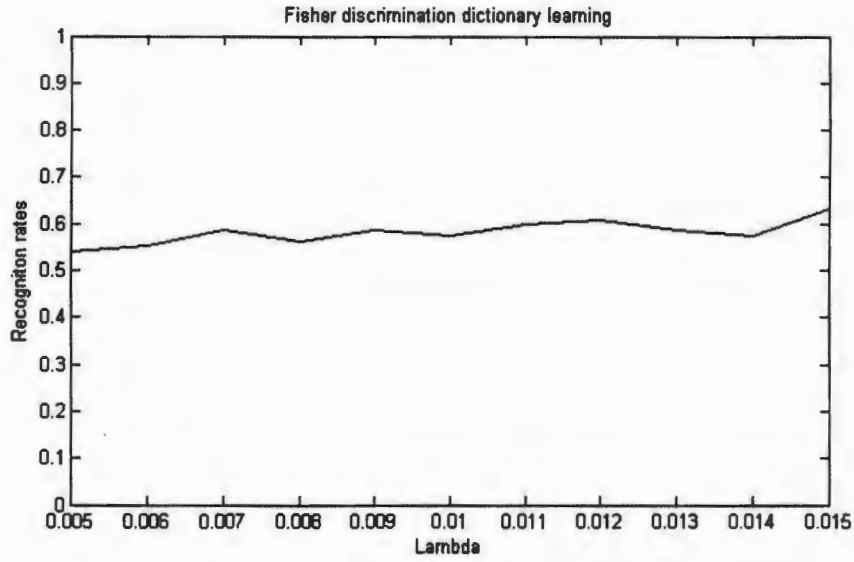


Figure 5.1: Recognition rates of FDDL

Table 5.2: Classification rates of K-SVD

Sparsity L	Sparsity D=3	Sparsity D=4	Sparsity D=5	Sparsity D=6	Sparsity D=7	Sparsity D=8	Sparsity D=9
	Reco rate	Reco rate	Reco rate	Reco rate	Reco rate	Reco rate	Reco rate
3	<b>0.5747</b>	0.4828	0.4713	0.4598	0.4713	0.4828	0.4713
4	<b>0.5172</b>	0.4483	0.4483	0.4253	0.4253	0.4713	0.4368
5	0.4943	<b>0.5287</b>	0.4828	0.4828	0.4713	0.4598	0.4368
6	<b>0.4943</b>	0.4368	0.4943	0.4368	0.4598	0.3908	0.4483
7	<b>0.5057</b>	0.4598	0.4943	0.4483	0.4598	0.4253	0.4138
8	0.4598	0.4483	0.4713	0.4483	<b>0.4943</b>	0.4943	0.4828
9	<b>0.5057</b>	0.4598	0.4943	0.4368	0.4483	0.4828	0.4138

## Chapter 6 - Conclusion

### Conclusion

In this proposed work, we use dictionary learning approaches to classifying the data set of bird's species, which is given in the form of audio recording. We start from raw audio recording and make them in to suitable features which is further used by dictionary learning methods. The data is collected from the field using omnidirectional microphones and we show that proposed work achieve high accuracy.

For the learning of dictionaries we use two techniques i.e. Fisher discrimination dictionary learning and K-singular value decomposition. In case of Fisher criteria of dictionary learning, a structured dictionary is learned whose sub dictionary has explicit class labels. The discrimination capability of Fisher discrimination dictionary learning has two folds. That is, in the whole dictionary each sub dictionary present has good representation capability to the examples that belongs to analogous class but its representation capability to the examples that belongs to other class is poor. Second, by using FDDL, sparse coefficients are also discriminative. These discriminative coefficients are achieved by reducing the scattering of coefficients with in the same class and increasing the scattering of coefficients between different classes. So the classification scheme used here in the work, used discriminative sparse coefficients and discriminative reconstruction error for the classification of test data set of bird species. So this discrimination property of FDDL demonstrate its dominance over K-SVD which is shown in the experimental part of the work. The dictionary learned in K-SVD method is simple and has no discriminative

behavior. So due to lack of discrimination property, K-SVD classification results lags from the Fisher discrimination dictionary learning approach.

The labels that are present in the training data are only predicted by the classifiers. When something appears that does not belong to the training classes are not predicted by the classifier. So classification of unexpected sounds present in the recording is not classify well. There is also difficulty for human labeler to find all the species that are existing in the recording because environment from where data is collected is highly noisy and also when microphones are very far from the vocalizing birds. So it is possible that sounds may also come from the bird species which are not existing in the training data set. Moreover, microphone also capture noise that comes from stream, waterfalls, rain and storm etc. rather than bird sound. So due to this some of the instances give segmentation error which reduces our classification accuracy. So future work is needed on noise instances, incomplete label sets and the classes which are not present in the training data.

Although, we consider the example of birds, this work can also be applicable to other acoustic signals for analysis. Including other animals or insects sounds like frogs, crickets, grasshoppers etc. Aside from the ecological application of this work, it also extends the scope of multi instance multi label domain from images and texts data set to audio data set.

## References

- [1] Z. H. Zhou, M. L. Zhang, S. J. Huang, and Y. F. Li, "Multi-Instance Multi-Label Learning," *Artif. Intell.*, vol. 176, no. 1, pp. 2291–2320, Aug. 2008.
- [2] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *J. Acoust. Soc. Am.*, vol. 131, no. 6, p. 4640, Jun. 2012.
- [3] S. Fagerlund, "Automatic Recognition of Bird Species by Their Sounds," *Dep. Electr. Commun. Eng. Acoust. Audio Signal Process.*, p. 56, Nov. 2004.
- [4] F. Briggs, "Multi-instance multi-label learning : algorithms and applications to bird bioacoustics," Oregon State University, 2013.
- [5] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *J. Acoust. Soc. Am.*, vol. 103, no. 4, pp. 2185–2196, 1998.
- [6] T. Dietterich, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1–2, pp. 31–71, 1997.
- [7] F. Briggs, R. Raich, Z. Lei, Y. Huang, and K. Eftaxias, "THE NINTH ANNUAL MLSP COMPETITION : OVERVIEW," *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, 2013.
- [8] J. Foulds and E. Frank, "A Review of Multi-Instance Learning Assumptions," *Knowl. Eng. Rev.*, pp. 1–24, Jun. 2014.
- [9] Z. H. Zhou and M. ling Zhang, "Multi-instance multilabel learning with application to scene classification," *Neural Inf. Process. Syst.*, vol. 40, no. 7, pp. 2038–2048, 2007.

- [10] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *Knowl. Data Eng. IEEE Trans.*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [11] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, 2006.
- [12] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 5, pp. V–701–4.
- [13] S. Fagerlund, "Bird Species Recognition Using Support Vector Machines," *EURASIP J. Adv. Signal Process.*, Mar. 2007.
- [14] T. S. Brandes, "Feature vector selection and use with hidden markov models to identify frequency-Modulated bioacoustic signals amidst noise," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 6, pp. 1173–1180, 2008.
- [15] M. L. Zhang, "A k-nearest neighbor based multi-instance multi-label learning algorithm," *Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2, pp. 207–212, 2010.
- [16] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. Acoust. Soc. Am.*, vol. 100, no. 2, pp. 1209–1219, Aug. 1996.
- [17] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 2003, vol. 5, pp. V–545–8.
- [18] P. Somervuo and A. Harma, "Bird song recognition based on syllable pair histograms," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, vol. 5, pp. V–825–8.



- [19] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 8, pp. 1541–1550, Nov. 2008.
- [20] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, 1980.
- [21] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. I, pp. 1–9, 2005.
- [22] F. Briggs, R. Raich, and X. Z. Fern, "Audio classification of bird species: A statistical manifold approach," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 51–60, 2009.
- [23] F. Briggs, X. Z. Fern, and J. Irvine, "Multi-Label Classifier Chains for Bird Sound," vol. 28, p. 6, Apr. 2013.
- [24] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern, J. Irvine, L. Neal, A. Thomas, G. Fodor, G. Tsoumakas, H. W. Ng, T. N. T. Nguyen, H. Huttunen, P. Ruusuvuori, T. Manninen, A. Diment, T. Virtanen, J. Marzat, J. Defretin, D. Callender, C. Hurlburt, K. Larrey, and M. Milakov, "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–8.
- [25] A. Ozerov, E. Vincent, and F. Bimbot, "A General Flexible Framework for the Handling of Prior Info. in Audio Source Separation," *Audio, Speech, Lang. Proc., IEEE Trans.*, vol. 19, no. 8, pp. 1118–1133, 2011.
- [26] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *2011 IEEE International Conference on Acoustics,*

*Speech and Signal Processing (ICASSP)*, 2011, pp. 2012–2015.

- [27] D. K. Mellinger and J. W. Bradbury, “Acoustic Measurement of Marine Mammal Sounds in Noisy Environments,” *Proc. Second Int. Conf. Underw. Acoust. Meas. Technol. Results, Heraklion, Greece, 25-29 June 2007*, p. 8, 2007.
- [28] M. Yang, L. Zhang, X. Feng, and D. Zhang, “Fisher Discrimination Dictionary Learning for sparse representation,” in *2011 International Conference on Computer Vision*, 2011, pp. 543–550.
- [29] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.