CLASSIFICATION WITH MULTI-IMPRECISE LABELS

By

SHENG ZOU

A DISSERTATION PROPOSAL PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2018

To my Lord Jesus Christ.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure page

LIST OF SYMBOLS, NOMENCLATURE, OR ABBREVIATIONS

ALC	Ambiguous label classification
APR	Axis-parallel rectangle
BCM	Beta compositional model
BS	Bag-space
DD	Diverse density
EDC	Evidential distance-based classifier
eFUMI	Extended functions of multiple instances
EM	Expectation maximization
EMDD	Expectation maximization diverse density
ES	Embedded-space
GMM	Gaussian mixture model
HSI	Hyperspectral image
IS	Instance-space
KNN	K-nearest neighbor
$\mathcal{L}$	Label matrix where each element $\ell_{i,j}^{m}$ denotes the labeling information (varies among different types of uncertain labels) on *instance level* or *bag level* for ith instance over mth class by jth annotator.
LDA	Latent Dirichlet allocation
LDL	Label distribution learning
LMM	Linear mixture model
MAP	Maximum a posteriori
MCMC	Markov Chain Mote Carlo

MESMA	Multiple endmember spectral mixture models
MH	Metropolis-Hasting
MI	Multiple instance
MI-ACE	Multiple instance adaptive cosine estimator
MI-HE	Multiple instance hybrid estimator
MIL	Multiple instance learning
MILES	Multiple instance learning via embedded instance selection
MIML	Multiple instance multi label
MI-SMF	Multiple instance spectral matched filter
MI-SVM	Multiple instance support vector machine
MLL	Multi-label learning
MRF	Markov random field
NCM	Normal Compositional model
NEON	National Ecological Observatory Network
PM-LDA	Partial membership latent Dirichlet allocation
PSO-EM	Particle swarm optimization expectation maximization
QP	Quadratic programming
SCM	Spatial compositional model
SDP	Semidefinite programming
SEM	Stochastic expectation maximization
SLL	Single-instance learning
S-PCUE	Sampling piecewise convex unmixing and endmember estimation
SVM	Support vector machine

USGS United States Geological Survey

VCA Vertex component analysis

$\mathcal{W}$ Labeling reliability matrix where each element $w_{i,j}^{m}$ denotes the labeling reliability of jth annotator for labeling the ith instance as mth class

$\mathbf{X}$ Training set

Abstract of Dissertation Proposal Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

CLASSIFICATION WITH MULTI-IMPRECISE LABELS

By

Sheng Zou

Nov 2018

Chair: Alina Zare
Cochair: Paul Gader
Major: Electrical and Computer Engineering

Imprecise labels and label uncertainty are common problems in many real supervised and semi-supervised learning problems. However, most of the state-of-the-art supervised learning methods in the literature rely on accurate labels. Accurate labels are often either expensive, time consuming or even impossible to be obtained in many real applications. There are many approaches in the literature that address imprecise and uncertain labels of various types. However, all types of label uncertainty and imprecision are not addressed. Furthermore, the overwhelming majority of methods that address label imprecision and uncertainty generally address only one form of imprecision/uncertainty even when many problems have more than one type of imprecise label coexisting in a problem.

Multiple instance (MI) label is one type of imprecise labels. MI label is a shared label for a set of instances, or called "bag", instead of one instance. For example, in tree species classification from remote sensed hyperspectral imagery, precise species class label for each pixel is difficult or expensive to obtain, as stated above. In comparison, an imprecise species label can be much easier to be assigned to a region (a bag) in the hyperspectral image, which can be a tree polygon, generated by an ecologist or tree delineation algorithm. The imprecise species label indicates the existence of the labeled species in the region as a form of subpixel, a pixel, or several pixels. However, this MI assumption, assuming the existence of labeled class in the bag, may not hold in some applications. For instance, the tree species labeling requires expertise in the tree species knowledge and sometimes relatives of a tree species are

a difficult to tell apart, the MI labels (species class label) for the bags (tree polygon) may be incorrect. In other words, there can be no labeled instance in the bag at all. In this case, another imprecise label, probabilistic label, can be introduced to incorporate the imprecision of MI labels. Probabilistic label is the confidence degree assigned to the original label, a way to show how probable the labeling is correct by the ecologist. In this case, a mixture of two types of imprecise labels, probabilistic labels over the multiple instance labels are in the training set.

In this thesis, different types of imprecise labels are defined and reviewed. An approach will be investigated to address classification problems with probabilistic labels over the multiple instance bag-level labels. In addition, the class in the multiple instance problem will be modeled by distribution instead of single instance to account for any variability in a class being modeled. The classification approach to be proposed will be applied to a variety of data sets with label uncertainty including, for example, hyperspectral data sets and medical imagery.

CHAPTER 1
INTRODUCTION

The imprecise labels, compared with accurate labels, are more realistic labels and cheaper to obtain in many machine learning applications, such as classification, regression and unmixing. For example, it will be much easier to label image patches instead of hundreds of thousands of pixels in an training image, even though there are some pixels in the image patch are mislabeled as the patch labels. A multiple instance learning based approach can often address this type of imprecise labels and learn a pixel level classifier, similar to or slightly worse than the classifier learned with accurate pixel level labels. Sometimes, imprecise labels offer a relatively suboptimal labeling mechanism, when the absolute accurate labeling for the training data is infeasible or require many experts (which is expensive!). For instance, for the classification of different tree species, a very precise species labeling is difficult; for the classification of different topics in the text, it is generally a hard task to choose only one topic class for a document since there are usually several topics involved with some degrees. However, it is easier and more reasonable to associate a probability/confidence value to all the candidate species labels. Also, it is more accurate to use multi-labels with some memberships for all the candidate topics in the document. In the both two examples, the imprecise labels, allow a more real and feasible labeling mechanism, compared to ideal, infeasible accurate labeling approaches.

Training data with imprecise labels have a variety of definitions based on the cause of the imprecision. Many types of imprecise labels have been tackled in the literature. However, there is few study on labels where multiple types of imprecise labels are involved. Multiple types of imprecise labels are widely observed in the classification of many different research areas, especially when crowdsourcing or labeling with a committee is used. For instance, in medical images, a committee of doctors can individually label the possible lesion regions of a medical image. In this case, an average confidence vector is assigned to the candidate labels of the labeled region, by fusion the labeling of all doctors. For the tree species labeling application

discussed above, a probability/confidence value can be queried from expert or by exploiting the neighboring species information and assigned to the labeled region. For both two examples, a probabilistic label over multiple instance label case is in the training label. Compared with only one type of imprecise labels, this case is more real and allows for a possibility of incorporating as much information as possible to help the training procedure.

Most state-of-the-art multiple instance learning based approaches generally estimate a "concept" instance for each class. However, one instance is usually not adequate to cover the variability inside the class, especially for remote sensing applications. For instance, each tree species class usually have a large spectral variability. Motivated by this issue, a multiple instance learning with distribution will be proposed, on the basis of the modeling probabilistic labels over multiple instance labels. The goal is to model each class as distribution to model the intra-class variability.

1.1 Imprecise label

The imprecise label is common problem in supervised and semi-supervised applications and, in particular, remote sensing applications. Imprecise labels (also known as label noise, or label uncertainty) are often caused by the following reasons. First, the labeled samples is composed of several classes and can be described using a fuzzy membership, for instance, the emotion classification of a face can be classified to be both happy and surprised, with a membership of 70% on class "happy" and 30% on "surprised". Second, the annotator is uncertain about the true class of the labeled pixel or object so that the annotator give a label probability value about how confidence the label/object belongs to each class. Third, in some applications, the existence of a target class is certain in a region of the scene, but the accurate pixel-level spatial location of the target in this region is uncertain. For instance, the polygons in Figure 1-1 are the regions all contain some wiregrass determined by the annotator, but the accurate pixel-level wiregrass labels are infeasible. Fourth, some real-value labels are imprecise. For instance, the age of a person can be labeled as 26, or between 25 and 27. However, the real age is an irrational number and impossible to obtain. Lastly, the labels can simply

be wrong because of the less reliable, easy-to-get labeling from nonexperts (e.g. Amazon Mechanical Turk). These are some examples of some types of imprecise labels. More complete list of imprecise labels are reviewed in Chapter 2.

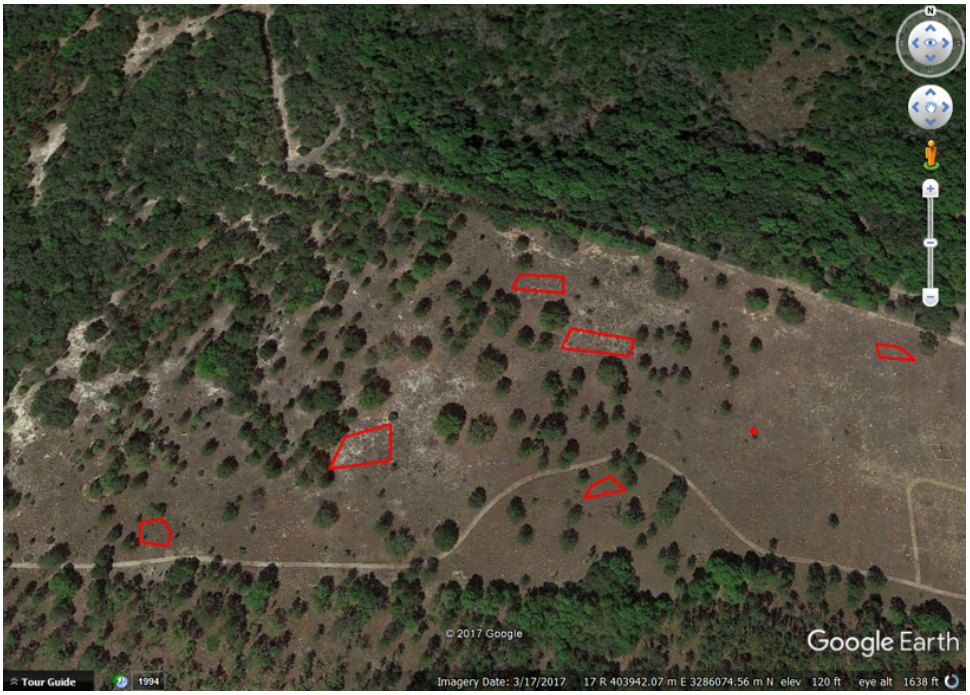

Figure 1-1. Wiregrass polygons in Ordway-Swisher Biological Station (OSBS)

There are many works in the literature addressing imprecise labels, such as looking for and correcting the imprecise labels, looking for and only using the precise labels, or building a noise-robust model. However, there are few works in the literature addressing more than one type of imprecise label, although multiple types of imprecise labels are commonly seen in many real applications. Two types of imprecise labels, the probability label over the multiple instance bag-level label, will be studied in this thesis, since the state-of-the-art methods in the literature can only classify on data with either multiple instance bag-level label or probabilistic label.

1.2 Applications of classification with imprecise labels

Imprecise labels widely exist for remote sensing applications. The first application that will be studied in this thesis is tree species classification of hyperspectral image over Ordway-Swisher Biological Station (OSBS). The hyperspectral images are collected and the

tree species labels are annotated by National Ecological Observatory Network (NEON). A species label is assigned to the center of tree truck. After some pre-processing steps, for instance, the center can be expanded to a region (*e.g.* a square shown in Figure 1-2 or circle or polygon), based on the measured size of the tree crown or some other factors. The regions are be modeled using bags in multiple instance learning (MIL). Then, the tree crown region is assigned with the species class. This is not accurate since there are many pixels in the region that are not the assigned labels, such as sand, soil, neighboring species and underneath grass. In addition, these species labels are not 100% accurate since some species have some minor difference and are difficult to tell. Fortunately, some species are more probable to grow together with the knowledge of ecology. So a model can be proposed to infer the probability of the species label based on its neighboring species. In this application, there are probabilistic labels over the bag-level labels.

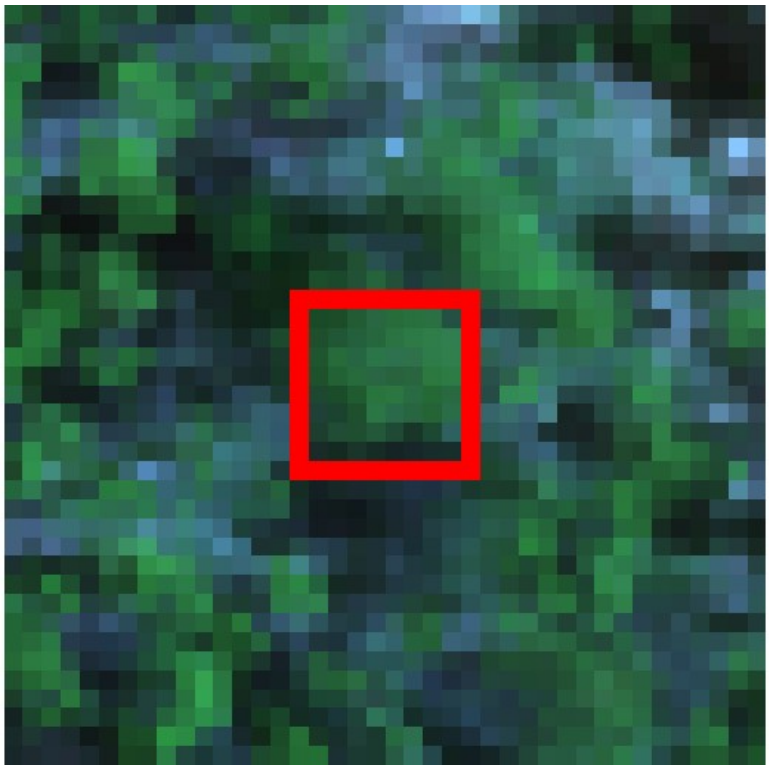

Figure 1-2. NEON tree crown image (RGB extracted from HSI) and its MIL bag

The second application that will be studied is very similar to the first application. It is also tree polgyons outlined and labeled by the experts from University of Santa Barbara (UCSB). Each polygon can be viewed as a bag in MIL. The bag label indicates that there are more than 70% of a certain class pixels inside the polygon. Similar to NEON datasets, the probabilistic label of the MIL label can be inferred by the neighboring tree species.

The third application that will be investigated is classification of a disease, diabetic retinopathy, using RGB images. The task using the DIARETDB1 dataset is to classify each color fundus image to a binary class of with or without diabetic retinopathy. The diabetic retinopathy and the level of the disease can be diagnosed by checking its signs (features), which are microaneurysms, soft exduates, hard exudates and hemorrhages. Four medical experts labeled same color fundus images using polygons, circles or ellipses, indicating the possible regions of each of the four signs, shown in Figure 1-3. Each polygon, circle or ellipse can be viewed as a bag. Since some regions are labeled from more than one expert as one of the four signs, these regions have higher probability of bag labels.

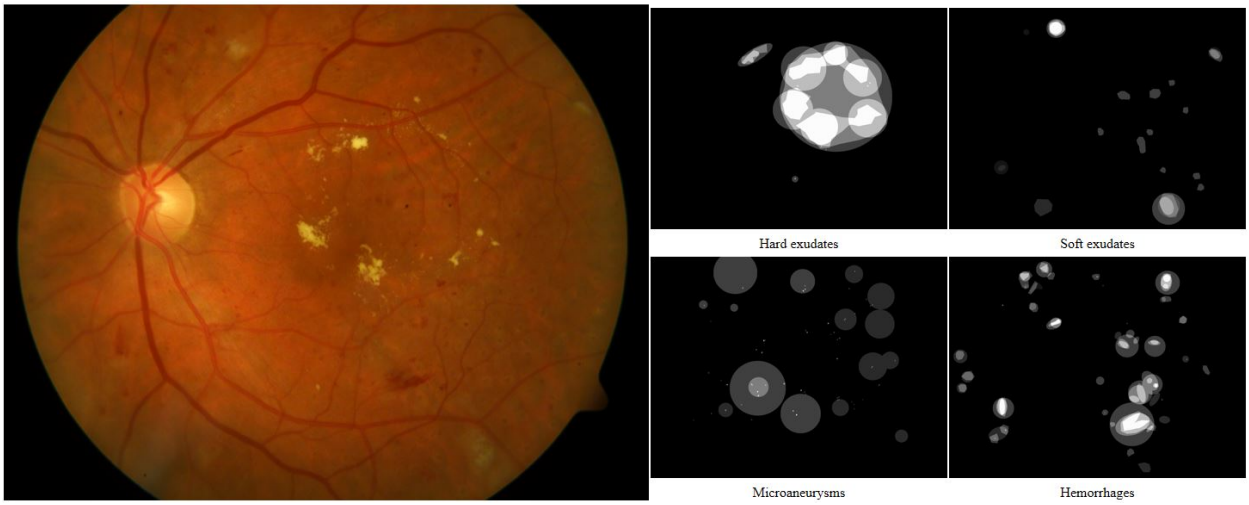

Figure 1-3. Fundus image and its MIL bags

CHAPTER 2
LITERATURE REVIEW

As stated in the chapter 1, imprecise labels have many definitions based on the cause of the imprecision. Different types of imprecise labels have different mathematical form and there are various approaches to be tackle them. In this chapter, a complete review of different types of imprecise labels and how they are addressed by the state-of-the-art. More importantly, it is found that there is few study on imprecise labels with multiple types of imprecision involved. Hyperspectral applications, such as unmixing and classification, are some of the applications where multiple types of imprecise labels exist. After the review of imprecise labels, the spectral variability of hyperspectral image is reviewed. Here, spectral variability are grouped into two categories in the literature, sets based and distribution based, is the motivation for proposing the distribution based multiple instance learning approach.

2.1 Existing Label Uncertainty Algorithm

Uncertain labels, or called noisy labels, imprecise labels in literature, refers to the observed training labels that are not accurate or reliable in classification. For instance, incorrectly setting a negative label on a positive instance in binary classification(Frénay and Verleysen, 2014).

The main reason leading to the uncertain labels is that the accurate training labels are usually expensive or time-consuming or impossible to obtain. For instance, in remote sensing applications, assuming we run a SVM algorithm on an image to classify different objects in a scene. A standard approach needs a set of pixel-level training labels. This training set is usually hundreds of thousands of pixels that need to be labeled, which is extremely time-consuming and infeasible for many real applications. Therefore, many training labels are obtained by some cheap, easy-to-get non-expect labeling framework. These non-expert frameworks are usually less reliable or even randomly assign labels when they have no knowledge of the labeling problems, resulting in imprecise labels. The accuracy of labeling also relies on the information provided by labeled source. For example, in the diagnosis (labeling) process on a patient, if the patient provides imprecise answers related to the symptoms, the diagnosis result is not

reliable. In other words, the information provided to the expert may not be sufficient for reliable labeling.

Not only the non-expert framework labeling process has labeling errors, but the expert labeling is subjective in many applications. In medical applications, for example, the labeling of medical images to determine if a specific disease exists may vary among doctors with their subjective understanding. The disagreement among annotators can be characterized as a confidence value or probability value by voting on the labeling results from different annotators. The disagreement can not only exist among annotators, but also exist in a single annotator. One annotator can provide a confidence vector for all possible labels when labeling an training data.

Label uncertainty can also stem from the multi-label characteristic of the training data itself. In other words, compared with traditional supervised learning where one single instance is associated with one (and only one) true single label, a single instance is associated with multiple true labels. To name a few, in hyperspectral classification, a single pixel can be composed by several different materials; in text categorization, a document can contain several topics, such as *policy*, *education* simultaneously. Hence, a label set and an associated fuzzy membership set is obtained for this type of uncertain label. If only a crisp label is provided to this kind of training data, the labels are imprecise.

In some applications, only an accurate high-level label information can be obtained because of the limited labeling source or difficulty of low-level labeling. For instance, in some semi-supervised hyperspectral unmixing or classification applications, only a high level training label can be accurately obtained. To be more specific, the annotator is capable of giving an identical label to a group of pixels in the hyperspectral image indicating the possible existence of target pixel(s) in this group of pixels. However, the precise pixel-level labels are not provided, due to the reasons such as the high cost of pixel level labeling and GPS error for a single pixel.

The uncertain labels can grouped into four cases according to the six scenarios above, which are *noisy labels*, *probabilistic labels*, *fuzzy and multi- labels*, *multiple instance labels*, *multiple instance multi-labels* and *regression labels*.

2.1.1 The framework of label uncertainty

Let us assume that there are N training instances $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathbb{R}^D$ is a D-dimensional real-valued feature vector. Let $\mathcal{W} = \{w_{i,j}^m\}_{i=1:N,j=1:K}^{m=1:M}$ be the labeling reliability matrix where $w_{i,j}^m \in [0,1]$ denotes the how much we trust ($w_{i,j}^m = 0$ denotes not trust; $w_{i,j}^m = 1$ denotes completely trust) on jth annotator for labeling the ith instance as mth class where K is the number of annotators and M is the number of classes. In literature, It is assumed that $w_{i,j}^m$ does not depend on the instance $\mathbf{x}$ to simplify the problem such that $w_{i,j}^m$ can be simplified to be w_j^m. However, the original form of labeling reliability, $w_{i,j}^m$, is more general and suitable for real applications. Note that even thought the simplified form, w_j^m, is assumed, the original form can also assumed for all types of uncertain labels in this section for future work. There are several common assumptions on $\mathcal{W}$ shown in Equation 2–1, 2–2, 2–3 and 2–4:

$$w_j^m = 1, \forall j, m \tag{2–1}$$

where Equation 2–1 denotes that all annotators are fully trusted. Even though the annotators are assumed to 100% accurately on labeling, it does not mean there is no label uncertainty. Since the label information can be not only a label indicator but label confidence, fuzzy membership, label interval or a high-level label provided by the annotators, containing a certain level of uncertainty for the true underlying label(s).

$$w_1^m = w_2^m =, \ldots, w_K^m, \forall m \quad \text{and} \quad w_j^1 \neq w_j^2 \neq, \ldots, w_j^M, \forall j \tag{2–2}$$

where Equation 2–2 represents the case that annotators have the same labeling reliability but for an individual annotator, he or she has different labeling reliabilities on different classes.

$$w_1^m \neq w_2^m \neq, \ldots, w_K^m, \forall m \quad \text{and} \quad w_j^1 = w_j^2 =, \ldots, w_j^M, \forall j \tag{2–3}$$

where Equation 2–3 represents the case that different annotators have different labeling reliabilities but for an individual annotator, he or she has the same labeling reliability on all classes.

$$w_1^1 \neq \ldots w_j^m \neq \ldots w_K^M, \forall j, m \tag{2–4}$$

where Equation 2–4 denotes the most realistic case that different annotators have different labeling reliabilities on different classes.

Let $\mathcal{L} = \{\ell_{ij}^m\}_{m=1}^M$ denote the complete set of label information, where ℓ_{ij}^m represents the labeling information (varies among different types of uncertain labels) on *instance level* or *bag level* for ith instance over mth class by jth annotator. The label information assigned to the instance can be in instance level or bag level. Label assigned in instance level is based on the observation of the single instance only, but label assigned in bag level is based on the observation of a set of instances together. In other words, all instances in a bag have the same bag level label(s). Thus, the instance level label is generally more precise than bag level label, in terms of each instance. In some scenarios, bag level labels are preferred, for example, when the instance level labels are expensive or difficult to obtain. For any uncertain labels with labeling reliability the satisfying Equation 2–1 and 2–2 can be simplified to single annotator labeling using majority voting on the label information, i.e. we use $\mathcal{L} = \{\ell_i^m\}_{m=1}^M$ to denote the label information set.

There are some common assumptions on $\mathcal{L}$ shown in Equation 2–5, 2–6, 2–7 2–9 and 2–10.

$$\ell_{ij}^m \in \{0, 1\} \quad \text{s.t.} \sum_{m=1:M} \ell_{ij}^m = 1, \forall i, j \tag{2–5}$$

where ℓ_{ij}^m denotes a single label indicator, indicating there is a single true label for each instance where the labeled class is the class with $\ell_{ij}^m = 1$.

$$\ell_{ij}^m \in \{0, 1\} \tag{2–6}$$

where ℓ_{ij}^m denotes a multi-label indicator, indicating there are multiple true labels for each instance where the labeled classes are the classes with $\ell_{ij}^m = 1$.

$$\ell_{ij}^m \in [0,1]_{Pro} \quad \text{s.t.} \sum_{m=1:M} \ell_{ij}^m = 1, \forall i, j \tag{2–7}$$

where ℓ_{ij}^m denotes a probability value, indicating the probability that the true label is mth class for ith instance provided by jth annotator.

$$\ell_{ij}^m \in [0,1]_{Pos} \tag{2–8}$$

where ℓ_{ij}^m denotes a possibilistic value, indicating the possibility that the true label is mth class for ith instance provided by jth annotator.

$$\ell_{ij}^m \in [0,1]_{Fuz} \quad \text{s.t} \sum_{m=1:M} \ell_{ij}^m = 1, \forall i, j \tag{2–9}$$

where ℓ_{ij}^m denotes a fuzzy membership value, indicating the *degree* that the mth class can fully describe the ith instance provided by jth annotator.

$$\ell_{ij}^m \in \mathbb{R} \tag{2–10}$$

where ℓ_{ij}^m denotes a real value, indicating an approximate label in regression problems.

$$\ell_{ij}^m \in [r_1, r_2], r_1, r_2 \in \mathbb{R}, \quad r_1 < r_2 \tag{2–11}$$

where ℓ_{ij}^m denotes a real-value set, a range of real values, indicating the possible range of label in regression problems.

2.1.2 Noisy labels

Noisy labels represent the case that there is no other information associated with the inaccurate labels except for the label indicators (Frénay and Verleysen, 2014). Also, a traditional supervised learning framework is assumed that only one label is associated with each instance satisfying Equation 2–5.

2.1.2.1 Classification with equal weighted annotators

In this framework, the labeling reliability is assumed to consistent over different annotators satisfying Equation 2–2. If $\mathcal{L}_i$ doesn't match the true label indicator of instance x_i, it is called a noisy label. For instance, incorrectly labeling an apple as orange in an apple-vs-orange binary classification problem. There are three main categories of approaches in literature in terms of how to address noisy labels, which are label noise robust methods, data cleansing methods and label noise tolerant methods.

Label noise robust methods are the approaches that naturally less sensitive to label noise than others. These methods have been shown to remain satisfactory performances, though the training data are corrupted by a certain level of wrong training labels. Some techniques embedding with a regularization term used for avoiding overfitting can address label noise more effectively (Teng, 2000, 2001, 2005). Data cleansing methods are the approaches that are capable of detecting the noisy labels followed by a correction step. In the correction step, the noisy labels can be removed, relabeled or the training data with noisy labels can be removed. Generally, the data cleansing method is applied, and a cleansed training dataset is generated before the learning algorithm. Similar to outlier detection or anomaly detection, some techniques are based on *ad hoc* measures where the training data are removed when the confidence of anomaly detection is over a certain threshold. For instance, Brodley et al. (1996) presented a method that can detect the mislabeled training data, similar to outlier detection, without any assumption on the learning approaches. The idea is to train a set of classifiers using a part of training data and then test on the remaining part of the training data. The instances for which the classifier disagrees most are relabeled using the predicted labels.
Label noise tolerant methods are the methods that naturally learn the label noise during the learning procedure. Most of the label noise tolerant methods are probabilistic models, including Bayesian (Pérez et al., 2007) and frequentist methods (Eskin, 2000). For Bayesian approaches, Bayesian priors are widely used on the mislabeling probabilities (Joseph et al., 1995; Gaba and Winkler, 1992). The choice for the Bayesian priors can be Beta priors (Zhang et al., 2005) and

Dirichlet priors (Ruiz et al., 2008). An indicator variable α_i is defined by Rekaya et al. (2001), indicating the switched label for the associated instance x_i if $\alpha_i = 1$. Therefore, each indicator follows a Bernoulli distribution based on mislabeling rate. The mislabeling rate is assumed to follow a Beta prior. The frequentist methods consider the label noise as a stochastic process (Eskin, 2000).

2.1.2.2 Classification by multiple weighted annotators

In previous binary or multi-class classification examples, if there are multiple annotators, the reliability of each annotator is treated equally. However, this is too ideal and not always true in real applications. For example, the annotators in crowdsourcing usually have different knowledge backgrounds and thus, have different error rates when labeling. Even for the labeling by a committee of experts, the labeling accuracy may vary based on the level of expertise of each expert. In this section, the performance of annotator is regarded as a variable and characterized using weights in Equation 2–4.

Raykar et al. (2010) presents a Bayesian model to characterize the performance of each annotator. Without loss of generality, a binary classification is considered as an example and Equation 2–5 is satisfied where $M = 2$. The class indicator values of class 1 and 2 assigned to an instance $\mathbf{x}$ are ℓ_j^1 and ℓ_j^2, respectively. The true class indicator values are defined as $\hat{\ell}_j^1$ and $\hat{\ell}_j^2$. Therefore, *sensitivity* and *specificity* are used to represent the probability that the annotator correctly labels the instance with true label of 1 and 0, respectively. More formally, the *sensitivity* α^j for jth annotator is defined in Equation 2–12. *Sensitivity* is also called true positive rate.

$$\alpha^j := Pr\left(\ell_j^1 = 1 | \hat{\ell}_j^1 = 1\right) \tag{2–12}$$

The *specificity* β^j for jth annotator is defined in Equation 2–13. *Specificity* is also called false positive rate.

$$\beta^j := Pr\left(\ell_j^2 = 1 | \hat{\ell}_j^2 = 1\right) \tag{2–13}$$

It is assumed that α^j and β^j are both consistent over all instances for jth annotator. For multiclass classification, a more general form of weights, w_j^m can be used to replace α^j (can be viewed as w_j^1) and β^j (can be viewed as w_j^2). Therefore, under a Bayesian framework, priors can be imposed on the *sensitivity* and *specificity* to incorporate the labeling performance of each annotator. Since α^j and β^j represent a binary classification problem, beta priors are suggested by Raykar et al. (2010). Thus, the beta priors for *sensitivity* and *specificity* can be represented as

$$Pr\left(\alpha^j | a_1^j, a_2^j\right) = Beta\left(\alpha^j | a_1^j, a_2^j\right) \tag{2–14}$$

$$Pr\left(\beta^j | b_1^j, b_2^j\right) = Beta\left(\beta^j | b_1^j, b_2^j\right) \tag{2–15}$$

where a_1^j, a_2^j, b_1^j, b_2^j are the hyperparameters for beta distributions. Then Raykar et al. (2010) proposed a model that sets the true labels as latent variables and jointly estimates the true labels and a classifier by maximum a *posteriori* (MAP) estimator optimized by EM algorithm.

2.1.3 Probabilistic labels

Probabilistic labels are also called ambiguous labels. In this setting, an instance may be labeled in a non-unique way by a subset of classes, similar to multi-label classification (Hüllermeier and Beringer, 2006). A confidence or probability set is also provided by the annotator, which characterizes how confidence or probable that the training data belongs to each class in the subset. But the existence of a unique correct classification is assumed, and the labels are regarded as candidates. The general process in the literature to address ambiguous labels refers to *ambiguous label classification* (ALC). There are many works in literature addressing the probabilistic labels and can be roughly categorized into the following three scenarios: binary classification by equal weighted annotators, multiclass classification by equal weighted annotators.

2.1.3.1 Binary classification

Nguyen et al. (2014) proposed to associate *soft-label* information by annotator with training labels for binary classification. *Soft-label* information is the additional confidence value/level reflecting how strongly the annotator feels about the class labels. In this scenario, the additional confidence value/level is assumed to be provided by a single annotator or averaged by multiple annotators who are considered to have same labeling ability and be fully trusted. More formally, Equation 2–1 is satisfied. Assuming the binary classification task is to label the fruit between apple (class 1) and orange (class 2). Generally, *soft-label* information can be represented by (1) a probability in Equation 2–7 where $M = 2$ for binary classification, for instance, the probability that the ith fruit is an apple is 0.7 ($\ell_i^1 = 0.7$ and $\ell_i^2 = 0.3$), or (2) an ordinal category, for example, the annotator 'strongly agree' that the fruit is an apple where the possible ordinal categories including 'strongly agree' , 'weakly agree', 'weakly disagree' and 'strongly disagree'.

Soft-label information, proved in (Nguyen et al., 2014), can help to learn a classification model more efficiently than with binary labels only, when the number of training labels is limited. There algorithms leveraging *soft-label* information vary based on different types of *soft-label* information, i.e., probabilistic labels or categorical labels.

For probabilistic labels, *soft-label* information is defined as the probability value of the most probable class. More formally, $c_i = \max(\ell_i^1, \ell_i^2)$ for $\mathbf{x}_i$. One approach considers the learning procedure as discriminative linear regression which regresses the features directly to probabilities. Thus, assuming a linear regression mapping, $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, where $\mathbf{w}$ are the weights of the model. The learning procedure is to minimize the following objective function:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i - c_i\right)^2 + Q(\mathbf{w}) \tag{2–16}$$

where $Q(\mathbf{w})$ is the regularization term preventing overfitting.

Another approach regards the learning procedure as logistic regression. Unlike the linear regression where the output is unbounded and inconsistent with probability, the logistic

regression naturally outputs the values in the range between 0 and 1. Therefore, similarly, the learning procedure is to minimize the following objective function:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{w}^T \mathbf{x}_i - t\left(c_i\right)\right)^2 + Q\left(\mathbf{w}\right) \tag{2–17}$$

where $Q\left(\mathbf{w}\right)$ is the regularization term preventing overfitting and $t\left(c_i\right) = \ln \frac{c_i}{1-c_i}$ is the inverse of logistic function.

For categorical labels, the *soft label* information is no longer probabilities but several crisp levels. The authors suggested to use ordinal regression based on SVM approach. The main idea is to construct a regression function $f\left(\mathbf{x}\right) = \mathbf{w}^T\mathbf{x}$ mapping each $\mathbf{x}$ onto a real line such that instances in each categories are projected to a compact and well separated region. In SVM, in order to apply the categorical labels, a set of inequality constraints are encoded. More specifically, let $\mathbf{b} = b_1, b_2, \ldots, b_{r-1}$ represent the boundaries that separate r categories. For one instance $\mathbf{x}_i$ assigned to category c_j, it should also satisfy $b_1 < \ldots < b_{j-1} < f\left(\mathbf{x}\right) < b_j < b_{j+1} < \ldots < b_{r-1}$.

2.1.3.2 Multiclass classification

Multiclass classification is a general case of binary classification where there are more than two possible classes in the classification. In this case, the same scenario is assumed as a binary classification that the labels are provided by a single annotator or averaged by multiple equally treated annotators. Similar to binary classification, it also satisfies Equation 2–1 and Equation 2–7 where $M > 2$ for multiclass classification. Note that multiclass classification is different from multi-label classification where the former allows a unique class (label) for each instance but the latter allows multiple classes (labels) associated with each instance. Assuming the multiclass classification task is to label the fruit among apple, orange and banana. An instance x_i can be labeled as $\mathcal{L}_i = \{0.7, 0.2, 0.1\}$, denoting the probability values that this fruit is an apple, orange, or banana, respectively. Since there are more than two classes, a more general case can be assumed that no probability/confidence values are provided for each training instance. Instead, a subset of possible labels, where each candidate label has the same

probability, is provided by the annotator. For instance, an instance x_i can be labeled as either the apple or orange but not banana. In other words, $\mathcal{L}_i = \{0.5, 0.5, 0\}$.

Jin and Ghahramani (2003) presents an approach that models the multiclass classification problems. It is capable of inferring the correct label among the set of candidate labels and achieve a performance close to the case of training with true labels. They proposed the EM and EM+prior methods to address the two cases defined above in terms of with or without probability/confidence values. The general idea is to assume a parameterized classifier with some unknown parameters to be estimated. A maximum likelihood criterion is used to estimate the parameters such that the target label has a high probability of being a member of the label set. To incorporate the prior information on the class labels, they generalize the likelihood function so that the estimated label distribution has low relative entropy with the prior on the class labels. The authors also prove that the EM+prior model can not only take advantage of the probability information provided to the candidate label set but is robust to a small amount of noise on the prior distribution over class labels.

2.1.4 Possibilistic labels

Possibilistic labels are more relaxed imprecise labels, compared with probabilistic labels. In probabilistic label, there is an underlying assumption that there must be at least one class is completely possible, while for possibilistic labels, the instance may actually belong to none of the enumerate categories. Equation 2–8 is satisfied. The possibilistic labels can be obtained from a single annotator, who is asked to provide a real number between 0 and 1 as the degree of possibility that ith instance belongs to mth class, or obtained by multiple annotators, who is asked to provide a binary number for the possibility that ith instance belongs to mth class (1 for possible, 0 for impossible) and the degree of possibility is aggregated by voting. For either case, Equation 2–1 is satisfied. Denœux and Zouhal (2001) suggested to use model and tackle the possibilistic labels within the framework of Evidence Theory. They introduced evidential distance-based classifier (EDC) and generalized it to address the more general possibilistic labels, where the degree of possibility (a real value between 0 and 1) instead of

binary possibility (0 or 1) is associated to the class labels of training instances. The EDC method use a belief function to model the class possibility of the unseen testing instance. Two approaches were proposed based on belief function. One is to transform each possibility distribution into a consonant belief function. The other one is to use generalized belief structures with fuzzy focal elements.

2.1.5 Fuzzy and multi labels

Fuzzy and multi-labels represent the scenario that the label for each instance is not unique. In other words, the candidate labels from the label set for each instance are not mutually exclusive. *Multi-label learning* (MLL) (Zhang and Zhou, 2014) assumes indiscriminate importance within the irrelevant label set while *label distribution learning* (LDL) (Geng, 2016) allows direct modeling of different importance of each label to the instance. The labels corresponding to MLL and LDL are called *multi labels* and *fuzzy labels*, respectively.

2.1.5.1 Multi-labels

For the case of *multi-labels*, the true underlying labels associated with a training instance are usually more than one. *Multi-labels* denotes the possible labels that can represent a single instance together (Zhang and Zhou, 2014). It can handle some real applications that traditional supervised learning method can't. For example, real-world instances may have multiple semantic meanings simultaneously. For an image labeling, the labels for the natural scene can both have mountain and ocean as the labels. For the topic modeling of an article, the labels (topics) for the article can have politics, economic and student at the same time. Therefore, a single label associated with one instance does not fit perfectly for these applications. *Multi-label learning* (MLL) is proposed by Zhang and Zhou (2014), allowing multiple labels for a single instance. In MLL, annotators are assumed to have the same labeling reliability, satisfying Equation 2–2. To label an instance $\mathbf{x}_i$, assigning a binary indicator number $\ell_{ij}^m \in \{0, 1\}$ to mth label, referring to the *existence* of mth label for instance x_i. However, the binary label indicator doesn't sum to one over the whole label set for an instance since there is more than one class for each instance. More formally, Equation 2–6 is satisfied.

2.1.5.2 Fuzzy labels

For *fuzzy labels*, the labels associated with a training instance is measured using a *fuzzy membership* set. Compared with *Ambiguous labels* where this is only one true label for each instance, *fuzzy labels* allow for multiple labels simultaneously. The *fuzzy membership*, or called *description degree* (Geng, 2016), denotes the *degree* that each label describe one instance. For instance, assuming a training instance is a face emotion. The fuzzy labels associated with the face emotion can be 70% happy and 30% surprised. Fuzzy labels can also be used for film rating. For example, the scores for a film can range from 0 to 10, labeled by audiences. Thus, there is no crisp score for this film but can be modeled using the score distribution by fuzzy labels. The general process in literature to address fuzzy labels refers to *label distribution learning* (LDL) (Geng, 2016). In LDL, annotators are assumed to have the same labeling reliability and fully trusted, satisfying Equation 2–1. To label an instance $\mathbf{x}_i$, assigning a real number ℓ_{ij}^m to mth label, referring to the *degree* to which jth label describe x_i. Let $\mathcal{L} = \{\ell_{ij}^m\}_{m=1}^M$ be the label distribution for instance x_i. It is assumed that the label set is complete, so there is a sum-to-one constraint $\sum_{m=1:M} \ell_{ij}^m = 1$ and non-negative constraint $\ell_{ij}^m \geqslant 0$ on the *degree*. In other words, Equation 2–9 is satisfied. Notice that the *degree* ℓ_{ij}^m is not the probability that the label of instance x_i is jth label, but the proportion that label accounts for a full description of instance x (Geng, 2016).

There are several approaches proposed in the literature to address the LDL problem. The first category is the problem transformation. A simple and straightforward method to transform the LDL problem to single-instance learning (SLL) problem. SLL denotes the traditional method that each instance is associated with only one label. To be more specific, assume that there are c labels in total. Each training instance $(\mathbf{x}_i, \mathcal{L}_i)$ can be transformed to M single label instance $(\mathbf{x}_i, \ell_i^m)$ using the *degree* as the weight. Then the training dataset is resampled to the same size according to the weights, resulting in a $M \times N$ training set. At last, any SLL approaches can be applied to the resampled training dataset. The second category is algorithm adaptation. Algorithm adaptation represents that some existing algorithm can be naturally

adapted to fit the LDL problem, for example, k-nearest neighbor (kNN) method. In details, the label distribution for mth label y_m of testing instance $\mathbf{x}$ can be represented as the mean value of the corresponding *degree* of its k nearest neighbors:

$$p\left(y_m|\mathbf{x}\right) = \frac{1}{k}\sum_{i\in N_k(\mathrm{x})} \ell_i^{y_m} \tag{2–18}$$

where $N_k(\mathbf{x})$ is the index set of the k nearest neighbors of $\mathbf{x}$ in the training set.

2.1.6 Multiple Instance (MI) labels

Multiple Instance (MI) labels refer to some high level labels, for example, only bag level labels, instead of *instance* level labels are obtained. A bag represents a group of *instances*. In *Multiple Instance Learning (MIL)* (Maron and Lozano-Pérez, 1998), there are usually two types of bag labels, positive bag and negative bag. $\mathbf{X}_k = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \subseteq \mathbf{X}$ is denoted as a kth bag, with a set of instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Each bag is a subset of the entire training set. If one given bag label is negative, $\mathcal{L}_k = -$, then we know that all instances in a bag are negative, $\mathcal{L}_k = [\ell_1^-, \ldots, \ell_i^-, \ldots, \ell_n^-]$ where ℓ_i^- denotes an instance level label that truly belongs to the negative class. If it's positive, $\mathcal{L}_k = +$, then we know that at least one instance in the bag is labeled as positive, $\mathcal{L}_k = [\ell_1^+, \ldots, \ell_i^+, \ell_{i+1}^-, \ldots, \ell_n^-]$ where ℓ_i^+ denotes an instance level label that truly belongs to the positive class. Let the positive and negative class be class 1 and 2, respectively. For instances in positive bags, the underlying true instance level labels are unknown. Only bag level labels for an instance is obtained where $\ell_i^1 = 1$ denotes the bag level label is positive and $\ell_i^2 = 1$ denotes the bag level label is negative for $\mathbf{x}_i$. Generally, the bag level labeling reliability of each annotator is considered as fully trusted so that Equation 2–1 is satisfied. The bag level labels ℓ_i^m satisfies Equation 2–5 where $M = 2$.

MIL is proposed motivated by some real applications where the precise instance level labeling is expensive or infeasible. For example, in the target characterization of hyperspectral image, the target size is usually a couple of pixels or even sub-pixel, and the target location is not precise because of the GPS error. Thus, the accurate pixel (instance) level labeling of the training target objects is usually infeasible. But a halo covering the target can be easily

labeled as a positive bag, indicating the existence of the target inside the halo. Therefore, MIL based approaches can address the label uncertainty and infer the accurate pixel locations of the targets. Many MIL methods have been proposed, such as learning axis-parallel concepts (Dietterich et al., 1997), diverse density (Maron and Lozano-Pérez, 1998), extended Citation *kNN* (Wang and Zucker, 2000).

The goal of Multiple instance learning based classification is to train a model which can predict the bag-level class labels or instance-level labels of a testing bag. The algorithms in the literature can be categorized as three paradigms, Instance-Space (IS) paradigm, Bag-Space (BS) paradigm and Embedded-Space (ES) paradigm (Amores, 2013). The IS paradigm considers the instance-level, local discriminative information. It learns a discriminative classifier on the instance space that separates the underlying true positive instances from the true negative instances. For a testing bag, the bag-level classification is obtained by aggregating the instance-level classification information. The BS paradigm considers the bag-level, global discriminative information. It learns a discriminative classifier on the bag space and uses the information from the whole bags and classify the whole bags. The ES paradigm is also another type of paradigm that considers the bag-level, global information. For ES based MIL algorithm, each bag is mapped to a feature vector capturing the information in the whole bag. Therefore, a discriminative classifier can be trained on the mapped embedded space. In other words, the MIL problem is transformed to a traditional supervised learning problem.

IS paradigm learns an instance-level classifier. One of the early works of IS paradigm is Axis-Parallel Rectangle (APR) (Dietterich et al., 1997). The goal of APR is to find an axis-parallel hyper-rectangle in the feature space. The APR is the minimum size hyper-rectangle of all possible hyper-rectangles that covers at least one instance from each positive bag and not include any instance from negative bags. The most effective approach of the three solutions proposed by Dietterich et al. (1997) finding the optimal APR is called 'inside-out' approach. This approach estimates the smallest APR by growing the APR from a seed point in the feature space which covers at least one instance from each positive bag and no instances

from negative bags. Similar to APR, Diverse Density (DD) (Maron and Lozano-Pérez, 1998) algorithm also follows IS paradigm. DD approach learns a *concept* point in feature space such that the concept point is as close to at least one instance from each positive bag as possible and far away from all instances from negative bags as possible. Diverse Density is defined as a measure about how close the positive bags and how far the negative bags are from the concept point. Therefore, the MI problem is to find such as concept point that maximizes the DD values. Gradient ascend approach is suggest by Maron and Lozano-Pérez (1998) to find the concept point with a strategy of starting with instances from each positive bag repetitively. EM-DD (Zhang and Goldman, 2002) is a EM version of the DD method. EM-DD assumes that there is a 'most representative' point in each bag capturing the label information of the bag. Since these 'most representative' points are unknown, hidden variable , they are estimated using EM based approach, leading to EM-DD, a combination of EM based approach with DD method. EM-DD has a similar framework as $k-$means clustering. It starts with an initial concept with estimated by DD algorithm. In the E step, the current concept point is used to find the 'most representative' point for each bag. In the M step, a new concept point is estimated such that the DD is maximized and used to replace the current concept. EM-DD iterates the E and M steps until convergence. By finding the 'most representative' points, EM-DD converts the multiple instance problem to single instance problem such that the computationally complexity is reduced. Extended Functions of Multiple Instances (eFUMI) (Jiao and Zare, 2015) is another EM style MI algorithm that learns positive and negative *concept* points. Each instance is considered as a convex combination of positive and negative concepts. Multiple Instance Spectral Matched Filter (MI-SMF) and Multiple Instance Adaptive Cosine Estimator (MI-ACE) (Zare et al., 2017) maximize the ACE and SMF responses respectively by learning a discriminative positive *concept*. Instead of learning only one 'concept' point for positive or negative bag, Multiple Instance Hybrid Estimator (MI-HE) (Jiao et al., 2017) learns a set of 'concept' points to capture the variability within the bags. mi-SVM and MI-SVM are two SVM based approaches aim for instance-level classification and

bag-level classification, respectively (Andrews et al., 2003). For mi-SVM, all instances are accounted for estimating the margin. The margin is maximized with the constraint that at least one instance from each positive bag are in one halfspace and all instances in the negative bags are in the other halfspace. For MI-SVM, only the 'most representative' instances from all bags are used. The margin is defined by the 'most positive' instance from each positive bag and 'least negative' instance from each negative bag. For all the IS paradigm based algorithm discussed above (ARP, DD, EM-DD and MI-SVM), the bag-level classifier can be the max rule, one of the aggregation rules used by many IS methods:

$$F(\mathbf{X}) = \max_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x}) \tag{2–19}$$

where $\mathbf{x}$ denotes every instance in a bag $\mathbf{X}$ and $f(\mathbf{x})$ is a selected instance level classifier with an binary output (1 for positive and 0 for negative) for each instance $\mathbf{x}$.

The standard multiple instance problem assumes that the positive bag is defined if containing a class of instances that negative bag do not contain. However, there is a different multiple instance problem definition. For the alternative definition, instances from all classes can exist in both positive and negative bags. Each positive bag contains instances from more than one class while each negative bag contains instances from only a single class. Note that different negative bags may contain instances from different class. For the alternative MI definition, the IS paradigm based methods have bad performances since it learns an instance-level classifier and both positive and negative bags may contain instances from every class. Thus, global, bag-level information becomes necessary for the alternative multiple instance definition.

BS paradigm learns a bag-level classifier. The most common approaches in the literature learn a distance function or a kernel function for pairwise bags. Since a bag is a set of instances in the feature space, distance or kernel metrics that compare two sets can be applied. Related BS based MIL classification methods use minimal Hausdorff distance (Wang and Zucker, 2000), Earth Movers Distance (Zhang et al., 2007), the Chamfer distance (Belongie

et al., 2002) and the kernel (Gärtner et al., 2002). Chamfer distance (Belongie et al., 2002) is one of the most widely used BS based approach. Let $\mathbf{X}$ and $\mathbf{Y}$ be two bags. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ be the corresponding instances in each bag. The Chamfer distance between bag $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$D(\mathbf{X}, \mathbf{Y}) = \frac{1}{|\mathbf{X}|} \sum_{x \in X} \min_{y \in Y} \|(\mathbf{x} - \mathbf{y})\| + \frac{1}{|\mathbf{Y}|} \sum_{y \in Y} \min_{x \in X} \|(\mathbf{x} - \mathbf{y})\| \tag{2–20}$$

ES paradigm also considers the bag level information for multiple instance problem as BS paradigm. ES paradigm based approaches learn a mapping function which maps each bag to a feature vector. Then the bag-level classification is performed on the space of mapped feature vectors. The ES paradigm based methods can be categorized into two groups, non-vocabulary and vocabulary-based methods. For the non-vocabulary based methods, the instances in a bag are equally treated for mapping the bag to a feature vector. For the vocabulary based methods, the instances in a bag are unequally treated, for example, some prototypes are learned from the instances in the training set. Then the mapping of the bag is performed by comparing how similar between its instances and each of these prototypes.

The simplest non-vocabulary method is the Simple MI method proposed by Dong (2006). Simple MI maps each bag to the average of the instances inside the bag. Gärtner et al. (2002) proposed a max-min operator as the mapping function. Each bag is mapped to a 1-by-2d feature vector which is the concatenation of max value of each dimension and min value of each dimension, where d is the dimensionality of the instances.

The idea of vocabulary methods is to discover what classes of instance present in the bag. An general vocabulary based approach contains three major steps: 1) building a vocabulary, i.e., by clustering into K clusters; 2) proposing a mapping function, which maps the bag into a 1-by-K feature vector, by comparing the instances in a bag with the concept (prototype) from each cluster; 3) classifying the feature vector in the embedded $K-$dimensional space. Different vocabulary based approaches varies mainly on vocabulary generation and/or mapping function design.

Histogram-based bag-of-words method is a vocabulary based method (Nowak et al., 2006). First, the vocabulary is formed as K concepts by running a clustering method, i.e., k-means. These concepts are the corresponding average value of each cluster, a prototype for each class. Second, each instance in the testing bag is associated with its nearest concept from the K concepts. At last, each element of the mapped feature vector is the count of how many instances are assigned to each concept. Distance-based bag-of-words method is another vocabulary based method. For previous histogram-based methods, the Euclidean distance is used to find the nearest concept for each instance, i.e. the Multiple-Instance Learning via Embedded Instance Selection (MILES) method (Chen et al., 2006). However, some authors use Mahalanobis distance (Opelt et al., 2006) or Gaussian kernel (Serre et al., 2007). Another major difference between distance-based and histogram based methods is the mapping function. For distance-based methods, each element of the mapped feature vector is the minimum distance between instances in the bag and each concept. To be more specific, histogram based methods map the bag by checking how many instances fall into each class (a histogram), while distance based method map the bag by checking how far between the instances in the bag and each class (a distance).

2.1.7 Multiple Instance Multi-label (MIML) labels

Multiple Instance Multi-label (MIML) is proposed by Zhou et al. (2012), which can be regarded as the combination of Multi-label learning and Multiple instance learning, aiming to solve the real complicated applications where the instance level label is not feasible, and the object has multiple semantic meanings. For instance, an image object can belong to classes of ocean and mountain simultaneously. Previously for image retrieval, the image is considered as an instance with multiple labels. But the user may only interested in the concept ocean instead of mountain. To choose the right semantic meaning, MIML considers the image patches as instances of the image object, where each patch can be classified as one of the classes. In other words, the ocean patches and mountains patches can be learned from the image. Similar to MIL, annotators are fully trusted such that Equation 2–1 is satisfied. In MIL, each instance

in a bag sharing the same one bag level label. However, in MIML, each instance in a bag sharing the same subset of bag level labels. More formally, the bag level labels for an instance in MIML satisfy Equation 2–6.

2.1.8 Regression labels

For previous different types of uncertain labels, the classes for an instance are usually pre-defined fixed categories. However, the labels can also be a continuous value in some applications. For instance, in the application of age estimation or pose estimation (Yan et al., 2008), the age is usually a real number, e.g. 26.5 years old. Thus, these applications are to predict a regression value for new data. The corresponding training labels are defined as *regression labels*. However, there is still label uncertainty for regression labels, for example, the 26.5 years old is an approximate number, leading to some noise in the training set. Motivated by this type of label uncertainty, Yan et al. (2008) proposed to use a label interval as the *regression labels*. In other words, using label interval $(26, 27)$ years old instead of 26.5 years old. Thus, the label information ℓ_{ij}^{m} is defined as a continues real-value number or a real-value interval, shown in Equation 2–10 and Equation 2–11, respectively. If the label interval is used as the *regression labels*, the annotator's labeling reliability is assumed to be fully trusted, satisfying Equation 2–1. If the continuous value is used as the *regression labels*, the labeling reliability satisfies Equation 2–2. In literature, the label interval can be modeled as two inequality constraints of a non-linear regression solved by semidefinite programming (SDP) (Yan et al., 2008).

2.2 Hyperspectral unmixing with endmember variability

Label imprecision is a common issue in remote sensed imagery applications, such as hyperspectral unmixing and classification. There are a sort of imprecise labels in hyperspectral imagery, including 1) multiple instance labels, the high level label assigned to a group of pixels; 2) probabilistic labels, the confidence value for the labeling accuracy; 3) multi-labels, a couple of co-existing labels assigned for each mixed pixel or each image patch, depending on the applications; and 4) noisy labels, wrong labels assigned becase of bad imaging condition or less

reliable annotators. In particular, the MI labels and probabilistic labels are studied in the thesis. Most conventional MI classification methods tend to estimate a "concept" instance (Maron and Lozano-Pérez, 1998; Zhang and Goldman, 2002) for the positive class, where one instance is generally not enough to capture the variability of feature vectors in the positive class. The most recent state-of-the-art approach estimate multiple "concept" instances for the positive class, which can alleviate the variability problem in MIL (Jiao et al., 2017). The drawback of learning a single "concept" instance (or a few) is more obvious in remote sensed images. For instance, each tree species class in hyperspectral image usually have a large spectral variation. In addition, the spectral difference among some species classes are very small. Thus, learning one "concept" instance for each species class may not be able to capture the overall features of each class, resulting in a weak classification/unmixing performance. In the literature, a number of approaches have been proposed to address spectral variability in hyperspectral unmixing. There are two categories for representing these approaches: *endmembers as sets* and *endmembers as statistical distributions* (Zare and Ho, 2014). The former category is based on the linear mixture model, and the latter category can be regarded as the stochastic mixing model. In the following subsections, methods that account for these two categories are reviewed, respectively. In this thesis, the variability will be incorporated in MIL with the point of view of *endmembers as statistical distributions*.

2.2.1 Hyperspectral Classification and Unmixing

Hyperspectral image (HSI), collected using a hyperspectral camera, is a stack of image planes, where each plane corresponds to radiances at a specific electromagnetic wavelength acquired over all pixels in a scene (Bioucas-Dias et al., 2012). The wavelength for each image plane can range from visible to near infrared (e.g. 0.4 μm to 2.5 μm). Since HSI is a three-dimensional data cube, each pixel location in HSI is a two-dimensional radiance spectrum vector (or called spectral signature). Due to the spatial resolution limit of the hyperspectral camera, The spatial resolution of HSI is often low, resulting in mixed pixels. Mixed pixels are pixels composed of more than one pure material. Each pure material in a hyperspectral scene

is called an *endmember*. Thus, the spectral signature of each mixed pixel can be regarded as a mixture of a number of endmembers in the scene. However, different mixed pixels may be consisted by a number of endmembers with different weights. The weight is called *proportion*, representing the fractional proportion of each endmember in the mixed pixel.

Hyperspectral classification is the process to assign a class label or multi class labels to each pixel in the hyperspectral image, according to the applications and data. Single-label classification is usually assumed when the spatial resolution of the hyperspectral image is high, resulting in a majority of pure pixels in the scene. Multi-label classification is often assumed when spatial resolution is low, such that a majority of pixels are mixed pixels in the scene. Thus, one class label is not enough for mixed pixel. However, single-label classification can still applied to mixed pixels for an approximated and simple classification.

Hyperspectral unmixing aims to decompose the mixed pixel signature into a set of endmembers with associated proportion vector. Hyperspectral unmixing algorithms can be categorized into two groups based on the expected types of mixing, which are linear mixing models and nonlinear mixing models (Bioucas-Dias et al., 2012). Linear mixing models assume that pure materials are uniformly partitioned on the surface of the mixed pixels, and there is only macroscopic scattering on the surface. But nonlinear mixing assumes a more complicated case that distribution of materials can be nonuniform and there is microscopic, multiple scattering among materials in a mixed pixel.

2.2.2 Linear Mixture Model

Linear Mixture Model (LMM) is a series of algorithms assuming the linear mixing case in a hyperspectral image. Therefore, each pixel is well represented by a convex combination of several endmember signatures weighted by the associated proportions (and additive random noise in some scenarios). Suppose there are M endmembers in a hyperspectral image. The kth endmember signature is denoted by e_k and the proportion vector of the kth endmember for this ith pixel is p_{ik}. The observed pixel signature $\mathbf{x}_i$ is represented by

$$\mathbf{x}_i = \sum_{k=1}^{M} \mathbf{e}_k p_{ik} + \epsilon_i \tag{2–21}$$

where ϵ_i is the error term accounting for noise.

2.2.3 Endmember variability

Endmember variability models each endmember as a set or distribution instead of a single signature because endmember has variability. There are many reasons resulting in endmember variability such as different compositions or different lighting conditions. For instance, the left image in Figure 2-1 is the RGB image from the *MUUFL* hyperspectral dataset (Gader et al., 2013). The highlighted region contains a red roof building, which can be considered as an endmember called *red roof*. It can be assumed most of the building roof is pure pixels. However, in fact, different locations of *red roof* shows different spectral signatures mainly because of the various illumination intensities (i.e. shadow). There are two manually picked red roof pixels as well as their associated spectral signatures shown on the right part of Figure 2-1. The spectral signature of the red roof pixel towards the sunlight has higher reflectance magnitude than that of the pixel on the shadow location. In addition, the shapes of the two pixel signatures are slightly different. It is because of other factors resulting in endmember variability such as chemical composition ratio variability on the same material or different textures on the material surface.

Endmember has variability could simply because of how it is defined, depending on the applications. For example, the endmembers in the *MUUFL* dataset are normally regarded as *red roof*, *grey roof*, *vegetation*, *asphalt* and *soil*. However, for *vegetation* in the scene, it is composed of trees and grasses. In addition, there could be more than one tree species for trees and more than one grass species for grasses. Therefore, it is oversimplified to represent the endmember *vegetation* with only one endmember signature. In the thesis, endmembers are considered as statistical distributions to model the endmember variability. A more accurate and faster endmember variability algorithm will be proposed.

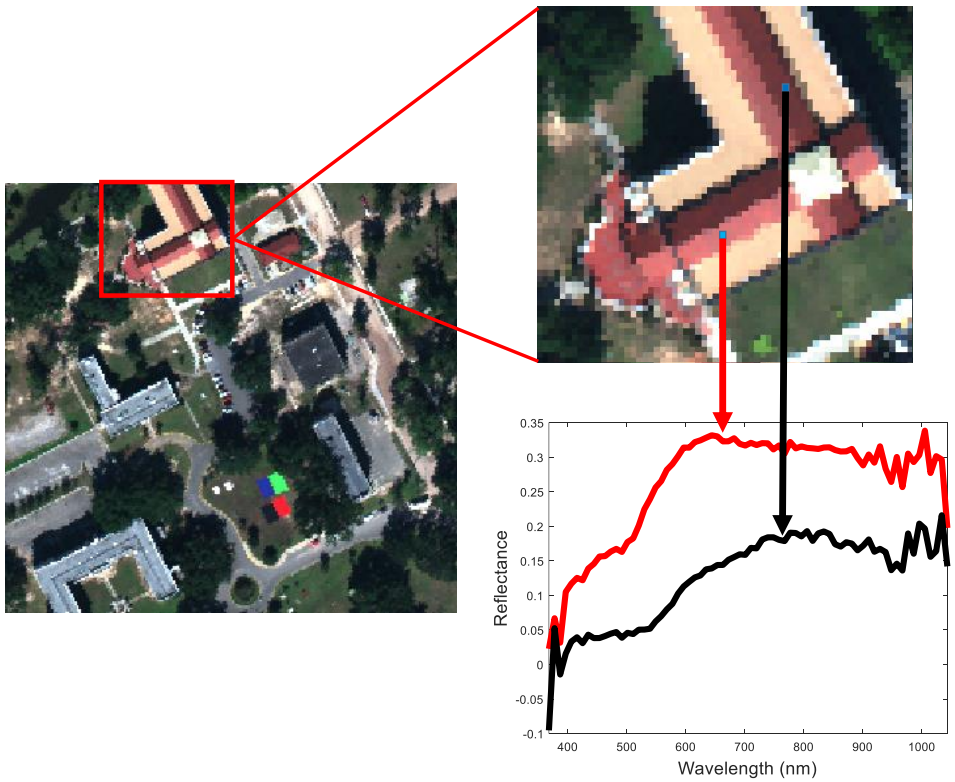

Figure 2-1. Spectral variability by illumination (Dataset: *MUUFL hyperspectral dataset*)

2.2.4 Endmembers as sets

Some algorithms address spectral variability by constructing a set of spectral signatures for each endmember instead of a single spectral signature under linear mixture model. The pixel signature can then be viewed as the convex combination of elements of endmember sets (one element from one set in most algorithms). From the perspective of the spectral library, the algorithms are further divided into two types, with or without a spectral library.

For algorithms given the spectral library, Multiple Endmember Spectral Mixture Models (MESMA), proposed by Roberts et al. (1998), and its extensions (Combe et al., 2008; Song, 2005; Asner et al., 2003) exhaustively search for one or more signatures for each endmember from the spectral library such that the estimated proportion values satisfy some predefined criteria. The signatures found, corresponding to each endmember, are grouped to form the endmember sets. Also, some other algorithms rely on a given spectral library, such as

endmember bundles (Bateson et al., 2000), band selection or weighting (Somers et al., 2011), and SVM unmixing (Mianji and Zhang, 2011; Bovolo et al., 2010)

For algorithms without the spectral library, the endmember sets are directly obtained from the pixels in the hyperspectral image. Automated endmember bundles (Somers et al., 2012) automatically selects a portion of pixels and applies an endmember extraction algorithm, for example, Vertex Component Analysis (VCA) (Nascimento and Dias, 2005), to estimate endmembers. This process is repeated for several times and generates a number of endmembers. Then the endmember sets are obtained by applying a clustering algorithm, for instance, K-means (Lloyd, 1982), to these endmembers. After estimating the endmember sets, the proportion values can be estimated using any previous unmixing algorithms. Additionally, other algorithms, such as sparse unmixing (Castrodad et al., 2011) and local unmixing (Canham et al., 2011; Goenaga et al., 2013), can also learn the endmember sets from the hyperspectral image.

2.2.5 Endmembers as distributions

Another type of method is based on modeling the endmember as a statistical distribution. Under the *endmembers as statistical distributions* approach, each endmember is considered as a statistical distribution rather than a set or a single value. Each sample from the endmember distribution can be regarded as a variant of the endmember signature. The variant is shown in equation 2–22.

$$\mathbf{e}_k \sim \mathscr{F}\left(\cdot|\theta_k\right) \tag{2–22}$$

where $\mathscr{F}$ is the statistical distribution for the kth endmembers $\mathbf{e}_k$ and θ_k represents the unknown parameters for this distribution.

Therefore, for endmembers following the statistical distribution, each pixel is regarded as the convex combination of these distribution-based endmembers. To be more specific, the pixel signature $\mathbf{x}_i$ can be written as

$$\mathbf{x}_i = \sum_{k=1}^{M} p_{ik}\mathbf{e}_k \tag{2–23}$$

where p_{ik} is the proportion value of one variant of the endmember distribution $\mathbf{e}_k$ for pixel $\mathbf{x}_i$. Thus, compared with the standard linear mixture model, this category of models assumes that each pixel signature can be represented by a linear combination of variants from endmember distributions with associated proportion values.

From the perspective of the assumed distributions, the algorithms are further divided into several types including Normal Composition Model (NCM), Gaussian Mixture Model (GMM) which can be viewed as a variation of NCM, Beta Compositional Model (BCM) and Spatial Compositional Model (SCM).

2.2.5.1 Normal Compositional Model based algorithm

A large number of endmember-distribution based, unmixing algorithms are based on a Bayesian framework. Once the distributions are indicated, both the distribution parameters and proportion values can be estimated simultaneously. The most commonly used distribution to represent endmembers is the normal distribution, that is,

$$\mathscr{F}\left(\mathbf{e}_k|\theta_k\right) = \mathscr{N}\left(\mathbf{e}_k|\mu_k, \mathbf{\Sigma}_k\right) \tag{2–24}$$

where μ_k is the mean parameter and $\mathbf{\Sigma}_k$ is the covariance parameter for endmember $\mathbf{e}_k$. The corresponding model assuming normal distributions for endmembers is named the Normal Compositional Model (Stein, 2003). We assume the variants from endmember distributions are mutually independent normal distribution variables. According to equation 2–22, 2–23 and 2–24, the pixel signature $\mathbf{x}_i$ under the NCM is represented as

$$\mathbf{x}_i \sim \mathscr{N}\left(\cdot \middle| \sum_{k=1}^{M} p_{ik}\mu_k, \sum_{k=1}^{M} p_{ik}^2\mathbf{\Sigma}_k\right) \tag{2–25}$$

A number of techniques for addressing spectral variability by assuming a normal compositional model have been developed.

Parameter estimation was addressed initially by Stein (2003) with a method based on the nested stochastic expectation maximization (SEM) algorithm (Diebolt and Ip, 1996). The proportion values p_{ik} are regarded as latent, hidden variables. The complete likelihood function is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, p_{11}, \ldots, p_{1M}, \ldots, p_{N1}, \ldots, p_{NM} | \{\mu_k, \mathbf{\Sigma}_k\}) \tag{2–26}$$

$$= \prod_{i=1}^{N} \mathcal{N}(\mathbf{p}_i; \mu(\mathbf{p}_i), \mathbf{\Sigma}(\mathbf{p}_i)) p(p_{i1}, \ldots, p_{iM}) \tag{2–27}$$

where N is the number of pixels and M is the number of endmembers in the data set.

$$\mu(\mathbf{p}_i) = c\mu_0 + \sum_{k=1}^{M} p_{ik}\mu_k \tag{2–28}$$

$$\mathbf{\Sigma}(\mathbf{p}_i) = c\mathbf{\Sigma}_0 + \sum_{k=1}^{M} p_{ik}^2 \mathbf{\Sigma}_k \tag{2–29}$$

It first chooses initial distribution parameters and latent proportion values. There are two types of distribution parameters, Gaussian means and Gaussian covariance matrices. The Gaussian means are initialized by applying an endmember extraction method, for example, VCA. The covariance matrices are initialized as the sample covariance matrices of clusters of points that are nearest the endmember. It then iterates between E and M steps, sampling latent proportions values and maximizing likelihood function until convergence. Specifically, in E step, the likelihood function value is (re-)calculated with the updated distribution parameters and proportion values. In M step, the proportion values are sampled such that the likelihood functions are maximized. Besides, the distribution parameters are updated with a gradient descent method.

Eches et al. (2010a,b) and Kazianka (2012) suggested to use a Markov Chain Mote Carlo (MCMC) sampler to estimate proportion values and endmember covariances under NCM given endmember mean signatures or a spectral library.

An example of NCM based MCMC method with given endmember mean signatures (Eches et al., 2010a) is presented below. It generates samples distributed according to the

joint posterior of proportion values, endmember variance values and one hyperparameter. The endmember extraction algorithm such as VCA or N-FINDR, is applied first to estimate the endmember means. Then, the endmember covariances and proportion values are estimated using MCMC. Several parameter priors are imposed. The Dirichlet prior is assigned to the proportions to enforce the non-negative and sum-to-one constraints, that is, $\mathbf{p}_n \sim \mathscr{D}(\cdot|\mathbf{1})$. The covariance matrix of each endmember can be written as $\sigma^2\mathbf{I}_L$, where $\mathbf{I}_L$ is the $L \times L$ identity matrix and σ^2 is the endmember variance in any spectral band. A conjugate inverse Gamma distribution is imposed on the variance σ^2 as the prior, given by $\sigma^2|\delta \sim \mathscr{IG}(\nu, \delta)$, where ν and δ are two user-defined hyperparameters (shape and scale parameters). A non-informative Jeffreys' prior is assigned to the δ, which is $f(\delta) \propto \frac{1}{\delta}\mathbf{1}_{R^+}(\delta)$.

The parameter estimation of this method contains two major steps, initialization and sampling using a hybrid Gibbs sampler. In the initialization step, the proportion values are initialized from a uniform distribution and normalized to be sum-to-one. The variance values are initialized from the pdf of an inverse Gamma distribution, $\sigma^2|\delta \sim \mathscr{IG}(\nu, \delta)$. Then the scale parameters are initialized from the pdf of non-informative Jeffreys' prior, $f(\delta) \propto \frac{1}{\delta}\mathbf{1}_{R^+}(\delta)$. In the sampling step, the proportion values are suggested to be sampled using a Metropolis-within-Gibbs algorithm such that the non-negative and sum-to-one constraints are both satisfied. The variance values (of the covariance matrices) are sampled following an inverse-Gamma distribution, which is

$$\sigma^2|\mathbf{x}, \mathbf{p}, \delta \sim \mathscr{IG}(\frac{L}{2} + 1, \frac{\|\mathbf{x} - \mu(\mathbf{p})\|^2}{2c(\mathbf{p})} + \delta) \tag{2–30}$$

For scale parameter δ, it is sampled from a Gamma distribution, which is

$$\delta|\sigma^2 \sim \mathscr{G}(1, \frac{1}{\sigma^2}) \tag{2–31}$$

where $\mathscr{G}(a, b)$ is the Gamma distribution with shape parameter a and scale parameter b.

Zare and Gader (2010) and Zare et al. (2013) presented MCMC sampler approach to estimate endmember spectral means and proportion values given endmember covariances under an NCM model.

In terms of hyperspectral data which are usually nonconvex, Zare et al. (2013) proposed to use several convex regions instead of a single convex region to represent the whole data set, motivated by the observation that hyperspectral data is usually nonconvex. The algorithm can automatically determine the number of endmember distribution sets by sampling from a Dirichlet process. Each set is viewed as a random simplex, where each vertex is modeled as an endmember distribution under the normal compositional model. The pixels are then divided into different sets according to the convex regions using a Dirichlet process prior. The Metropolis-within-Gibbs sampler is applied to divide the data set into convex regions with the learned number of regions and estimate the endmember distributions and proportion values for each convex region.

The endmember distribution covariances are assumed to be known in advance and data means are assumed to be drawn from the normal distributions $\mu_k \sim \mathscr{N}(\cdot|\mathbf{m}, \mathbf{C})$. The proportion vector for pixel $\mathbf{x}_i$ is modeled as $\mathbf{p}_i \sim \mathscr{D}(\cdot|\mathbf{1})$. The $\mathbf{m}$ and $\mathbf{C}$ are hyperparameters defining the prior distribution on the endmember means.

The Sampling Piecewise Convex Unmixing and Endmember Estimation (S-PCUE) starts with the initialization of the endmember means (the covariance matrices are given) of each convex set using VCA and proportion values of each pixel by drawing from a uniform Dirichlet distribution. Then S-PCUE iterates to sample the proportions for each pixel for each set of endmember using a Metropolis-Hasting step, sample each endmember mean in each set using another Metropolis-Hasting step and sample the hyperparameters for endmember means. Additionally, in each iteration, K new potential partitions, consisting of K new sets of endmember distributions and proportion values, are sampled. After sampling the new partitions, the DP partition probabilities for each pixel is calculated, which determines if the

convex set the pixel belongs to should be changed to a new partition or existing partition. The S-PCUE iterates to sample all these parameters until converging.

Some other approaches also address the NCM based hyperspectral unmixing problem. Zhang et al. (2014) introduced a particle swarm optimization expectation maximization (PSO-EM) method to estimate the endmember spectral means, endmembers covariances and proportion values. Zou and Zare (2017) introduced the Partial Membership Latent Dirichlet Allocation (PM-LDA) unmixing approach to estimate all endmember distributions and proportion values under the NCM while leveraging spatial information. Additionally, Gaussian Mixture Model (GMM) can be viewed as an extension of NCM (Zhou et al., 2018). GMM uses Gaussian Mixtures to model the underlying endmember distributions, motivated by the observation that the distribution of spectra from a material may be multi-modal.

2.2.5.2 Spatial Compositional Model based algorithm

Zhou et al. (2016) relaxes the assumption in NCM that the pixels are independent random variables and proposed Spatial Compositional Model (SCM). The authors defined a new concept named *endmember uncertainty*, similar but different than *endmember variability*, to model the error of endmembers. Since the pixels are not assumed to be independence, compared to standard NCM, the full likelihood of the pixels are estimated to obtain the endmember uncertainty. The author also applied a smoothness term which works locally to promote the spatial similarity. To be more specific, a Markov Random Field (MRF) prior is assumed on the proportion values to drive the neighboring pixels to be similar. The parameter estimation of SCM is accomplished by maximizing the posteriori using block coordinate descent method. However, the SCM results are sensitive to the weighting parameters on the spatial similarity term, i.e., a badly tuned weighting parameter may result in over-smoothing or under-smoothing.

2.2.5.3 Beta Compositional Model based algorithm

Under Beta Compositional Model (BCM), proposed by Du et al. (2014), each endmember is a random variable distributed according to a beta distribution, motivated by the observation

that the underlying endmember distribution may be skewed. The authors found that the beta distributions, for some hyperspectral dataset, have the better fit than normal distribution by comparing the quantile-quantile plots between the assumed the distributions and hyperspectral data. There are two types of BCM algorithm, which are BCM-spectral and BCM-spatial. Both methods require an initial step to identify the pixels that have similar proportion values. The BCM-spectral method only considers the spectral information in the initial step while the BCM-spatial utilize both the spatial and spectral information. The BCM-spectral is solved by quadratic programming (QP). The proportion values are estimated by minimizing the difference between the original and reconstructed data means. BCM-spatial uses a Metropolis-Hasting (MH) sampler to estimate the unknown unmixing parameters. The proportion values are estimated by minimizing the difference between both the original and reconstructed data means and variances. For both algorithms, the endmember distributions are estimated using the Maximum Likelihood (ML) method.

CHAPTER 3
RESEARCH QUESTIONS

The goal of this dissertation is to address more complicated but more realistic imprecise training label problems. These problems are classifications with probabilistic labels over multiple instance labels. Multiple instance imprecise labels relax the label precision from the instance level to bag level, while probabilistic labels over multiple instance labels further add more imprecision, by allowing the bag level MI labels to be imprecise as well by introducing a confidence/probability value on MI labels. In addition, the classes in MIL will be modeled using probabilistic distributions, instead of one or several "concept" instances. The class distributions are able to capture the intra-class variability of feature vector, for example, the spectral variability of classes in hyperspectral image. A hybrid classification method will be proposed to both address the probabilistic labels over MI labels problem and model classes using distributions in MIL. More specifically, the research questions of the study are the following:

(a) What types of data classification favors using multiple instance learning with distributions rather than conventional non-distribution methods? What scenarios does MIL with distributions fails?

(b) If the imprecise classification problem with distributions is addressed in the Bayesian framework, a likelihood function could be proposed to be optimized. If so, what is the probable form of the likelihood function?

(c) How to evaluate the classification results? For classification with multiple types of imprecise labels, how can we know the degree of classification error, corresponding to each type of imprecise label? In other words, how is the different imprecision accumulated leading to the final classification error?

Datasets:

(a) NEON dataset (b) UCSB dataset (c) DIARETDB1 dataset

Task table:

Propose a distribution based multiple instance classification model;
Make synthetic datasets with unimodal and multi-modal positive class distribution;
Implement the proposed method;
Run experiment on datasets above (assuming no probabilistic labels);
Evaluate the results by calculating the instance classification and bag classification on testing dataset;
Compare the results with MI-ACE, MI-HE, DD, EM-DD, mi-SVM, etc.
Propose a term to incorporate the probability to the bag labels (e.g. weighting)
Repeat the above tasks but with probabilistic labels.

Table 3-1. Task table

REFERENCES

Amores, Jaume. "Multiple instance classification: Review, taxonomy and comparative study." *Artificial Intelligence* 201 (2013): 81–105.

Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. "Support vector machines for multiple-instance learning." *Advances in neural information processing systems*. 2003, 577–584.

Asner, Gregory P, Bustamante, Mercedes MC, and Townsend, Alan R. "Scale dependence of biophysical structure in deforested areas bordering the Tapajos National Forest, Central Amazon." *Remote Sensing of Environment* 87 (2003).4: 507–520.

Bateson, C Ann, Asner, Gregory P, and Wessman, Carol A. "Endmember bundles: A new approach to incorporating endmember variability into spectral mixture analysis." *IEEE transactions on geoscience and remote sensing* 38 (2000).2: 1083–1094.

Belongie, Serge, Malik, Jitendra, and Puzicha, Jan. "Shape matching and object recognition using shape contexts." Tech. rep., CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING, 2002.

Bioucas-Dias, José M, Plaza, Antonio, Dobigeon, Nicolas, Parente, Mario, Du, Qian, Gader, Paul, and Chanussot, Jocelyn. "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches." *IEEE journal of selected topics in applied earth observations and remote sensing* 5 (2012).2: 354–379.

Bovolo, Francesca, Bruzzone, Lorenzo, and Carlin, Lorenzo. "A novel technique for subpixel image classification based on support vector machine." *IEEE Transactions on Image Processing* 19 (2010).11: 2983–2999.

Brodley, Carla E, Friedl, Mark A, et al. "Identifying and eliminating mislabeled training instances." *Proceedings of the National Conference on Artificial Intelligence*. 1996, 799–805.

Canham, Kelly, Schlamm, Ariel, Ziemann, Amanda, Basener, Bill, and Messinger, David. "Spatially adaptive hyperspectral unmixing." *IEEE Transactions on Geoscience and Remote Sensing* 49 (2011).11: 4248–4262.

Castrodad, Alexey, Xing, Zhengming, Greer, John B, Bosch, Edward, Carin, Lawrence, and Sapiro, Guillermo. "Learning discriminative sparse representations for modeling, source separation, and mapping of hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing* 49 (2011).11: 4263–4281.

Chen, Yixin, Bi, Jinbo, and Wang, James Ze. "MILES: Multiple-instance learning via embedded instance selection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006).12: 1931–1947.

Combe, J-Ph, Le Mouélic, S, Sotin, C, Gendrin, A, Mustard, JF, Le Deit, L, Launeau, P, Bibring, J-P, Gondet, B, Langevin, Y, et al. "Analysis of OMEGA/Mars express data

hyperspectral data using a multiple-endmember linear spectral unmixing model (MELSUM): Methodology and first results." *Planetary and Space Science* 56 (2008).7: 951–975.

Denœux, Thierry and Zouhal, Lalla Meriem. "Handling possibilistic labels in pattern classification using evidential reasoning." *Fuzzy sets and systems* 122 (2001).3: 409–424.

Diebolt, Jean and Ip, Eddie HS. "Stochastic EM: method and application." *Markov chain Monte Carlo in practice*. Springer, 1996. 259–273.

Dietterich, Thomas G, Lathrop, Richard H, and Lozano-Pérez, Tomás. "Solving the multiple instance problem with axis-parallel rectangles." *Artificial intelligence* 89 (1997).1-2: 31–71.

Dong, Lin. *A comparison of multi-instance learning algorithms*. Ph.D. thesis, The University of Waikato, 2006.

Du, Xiaoxiao, Zare, Alina, Gader, Paul, and Dranishnikov, Dmitri. "Spatial and spectral unmixing using the beta compositional model." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (2014).6: 1994–2003.

Eches, O., Dobigeon, N., Mailhes, C., and Tourneret, J. Y. "Bayesian estimation of linear mixtures using the normal compositional model. Application to hyperspectral imagery." *IEEE Transactions on Image Processing* 19 (2010a).6: 1403–1413.

Eches, O., Dobigeon, N., and Tourneret, J. Y. "Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical Bayesian algorithm." *IEEE Journal of Selected Topics in Signal Processing* 4 (2010b).3: 582–591.

Eskin, Eleazar. "Detecting errors within a corpus using anomaly detection." *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, 148–153.

Frénay, Benoît and Verleysen, Michel. "Classification in the presence of label noise: a survey." *IEEE transactions on neural networks and learning systems* 25 (2014).5: 845–869.

Gaba, Anil and Winkler, Robert L. "Implications of errors in survey data: a Bayesian model." *Management Science* 38 (1992).7: 913–925.

Gader, Paul, Zare, Alina, Close, Ryan, Aitken, Jen, and Tuell, Grady. "Muufl gulfport hyperspectral and lidar airborne data set." *Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570* (2013).

Gärtner, Thomas, Flach, Peter A, Kowalczyk, Adam, and Smola, Alexander J. "Multi-instance kernels." *ICML*. vol. 2. 2002, 179–186.

Geng, Xin. "Label distribution learning." *IEEE Transactions on Knowledge and Data Engineering* 28 (2016).7: 1734–1748.

Goenaga, Miguel A, Torres-Madronero, Maria C, Velez-Reyes, Miguel, Van Bloem, Skip J, and Chinea, Jesus D. "Unmixing analysis of a time series of Hyperion images over the Guánica

dry forest in Puerto Rico." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6 (2013).2: 329–338.

Hüllermeier, Eyke and Beringer, Jürgen. "Learning from ambiguously labeled examples." *Intelligent Data Analysis* 10 (2006).5: 419–439.

Jiao, Changzhe and Zare, Alina. "Functions of multiple instances for learning target signatures." *IEEE Transactions on Geoscience and Remote Sensing* 53 (2015).8: 4670–4686.

Jiao, Changzhe, Zare, Alina, and McGarvey, Ronald G. "Multiple Instance Hybrid Estimator for Hyperspectral Target Characterization and Sub-pixel Target Detection." *arXiv preprint arXiv:1710.11599* (2017).

Jin, Rong and Ghahramani, Zoubin. "Learning with multiple labels." *Advances in neural information processing systems*. 2003, 921–928.

Joseph, Lawrence, Gyorkos, Theresa W, and Coupal, Louis. "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard." *American journal of epidemiology* 141 (1995).3: 263–272.

Kazianka, H. "Objective Bayesian analysis for the normal compositional model." *Computational Statistics & Data Analysis* 56 (2012).6: 1528–1544.

Lloyd, Stuart. "Least squares quantization in PCM." *IEEE transactions on information theory* 28 (1982).2: 129–137.

Maron, Oded and Lozano-Pérez, Tomás. "A framework for multiple-instance learning." *Advances in neural information processing systems*. 1998, 570–576.

Mianji, Fereidoun A and Zhang, Ye. "SVM-based unmixing-to-classification conversion for hyperspectral abundance quantification." *IEEE Transactions on Geoscience and Remote Sensing* 49 (2011).11: 4318–4327.

Nascimento, José MP and Dias, José MB. "Vertex component analysis: A fast algorithm to unmix hyperspectral data." *IEEE transactions on Geoscience and Remote Sensing* 43 (2005).4: 898–910.

Nguyen, Quang, Valizadegan, Hamed, and Hauskrecht, Milos. "Learning classification models with soft-label information." *Journal of the American Medical Informatics Association* 21 (2014).3: 501–508.

Nowak, Eric, Jurie, Frédéric, and Triggs, Bill. "Sampling strategies for bag-of-features image classification." *European conference on computer vision*. Springer, 2006, 490–503.

Opelt, Andreas, Pinz, Axel, Fussenegger, Michael, and Auer, Peter. "Generic object recognition with boosting." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006).3: 416–431.

Pérez, Carlos Javier, Girón, F Javier, Martín, Jacinto, Ruiz, Manuel, and Rojano, Carlos. "Misclassified multinomial data: a Bayesian approach." *RACSAM* 101 (2007).1: 71–80.

Raykar, Vikas C, Yu, Shipeng, Zhao, Linda H, Valadez, Gerardo Hermosillo, Florin, Charles, Bogoni, Luca, and Moy, Linda. "Learning from crowds." *Journal of Machine Learning Research* 11 (2010).Apr: 1297–1322.

Rekaya, R, Weigel, KA, and Gianola, D. "Threshold model for misclassified binary responses with applications to animal breeding." *Biometrics* 57 (2001).4: 1123–1129.

Roberts, Dar A, Gardner, M, Church, R, Ustin, S, Scheer, G, and Green, RO. "Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models." *Remote Sensing of Environment* 65 (1998).3: 267–279.

Ruiz, M, Girón, FJ, Pérez, CJ, Martín, J, and Rojano, C. "A Bayesian model for multinomial sampling with misclassified data." *Journal of Applied Statistics* 35 (2008).4: 369–382.

Serre, Thomas, Wolf, Lior, Bileschi, Stanley, Riesenhuber, Maximilian, and Poggio, Tomaso. "Robust object recognition with cortex-like mechanisms." *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2007).3: 411–426.

Somers, Ben, Asner, Gregory P, Tits, Laurent, and Coppin, Pol. "Endmember variability in spectral mixture analysis: A review." *Remote Sensing of Environment* 115 (2011).7: 1603–1616.

Somers, Ben, Zortea, Maciel, Plaza, Antonio, and Asner, Gregory P. "Automated extraction of image-based endmember bundles for improved spectral unmixing." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (2012).2: 396–408.

Song, Conghe. "Spectral mixture analysis for subpixel vegetation fractions in the urban environment: How to incorporate endmember variability?" *Remote Sensing of Environment* 95 (2005).2: 248–263.

Stein, David. "Application of the normal compositional model to the analysis of hyperspectral imagery." *Advances in techniques for analysis of remotely sensed data, 2003 IEEE Workshop on*. IEEE, 2003, 44–51.

Teng, Choh Man. "Evaluating noise correction." *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2000, 188–198.

Teng, Choh-Man. "A Comparison of Noise Handling Techniques." *FLAIRS Conference*. 2001, 269–273.

Teng, Choh Man. "Dealing with data corruption in remote sensing." *International Symposium on Intelligent Data Analysis*. Springer, 2005, 452–463.

Wang, Jun and Zucker, Jean-Daniel. "Solving multiple-instance problem: A lazy learning approach." (2000).

Yan, Shuicheng, Wang, Huan, Tang, Xiaoou, Liu, Jianzhuang, and Huang, Thomas S. "Regression from uncertain labels and its applications to soft biometrics." *IEEE Transactions on Information Forensics and Security* 3 (2008).4: 698–708.

Zare, A. and Gader, P. "PCE: Piecewise convex endmember detection." *IEEE Transactions on Geoscience and Remote Sensing* 48 (2010).6: 2620–2632.

Zare, A., Gader, P., and Casella, G. "Sampling piecewise convex unmixing and endmember extraction." *IEEE Transactions on Geoscience and Remote Sensing* 51 (2013).3: 1655–1665.

Zare, Alina and Ho, KC. "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing." *IEEE Signal Processing Magazine* 31 (2014).1: 95–104.

Zare, Alina, Jiao, Changzhe, and Glenn, Taylor. "Discriminative multiple instance hyperspectral target characterization." *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2017).1: 1–1.

Zhang, B., Zhuang, L., Gao, L., Luo, W., Ran, Q., and Du, Q. "PSO-EM: a hyperspectral unmixing algorithm based on normal compositional model." *IEEE Transactions on Geoscience and Remote Sensing* 52 (2014).12: 7782–7792.

Zhang, Jianguo, Marszałek, Marcin, Lazebnik, Svetlana, and Schmid, Cordelia. "Local features and kernels for classification of texture and object categories: A comprehensive study." *International journal of computer vision* 73 (2007).2: 213–238.

Zhang, Min-Ling and Zhou, Zhi-Hua. "A review on multi-label learning algorithms." *IEEE transactions on knowledge and data engineering* 26 (2014).8: 1819–1837.

Zhang, Qi and Goldman, Sally A. "EM-DD: An improved multiple-instance learning technique." *Advances in neural information processing systems*. 2002, 1073–1080.

Zhang, Wensheng, Rekaya, Romdhane, and Bertrand, Keith. "A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer." *Bioinformatics* 22 (2005).3: 317–325.

Zhou, Yuan, Rangarajan, Anand, and Gader, Paul D. "A spatial compositional model for linear unmixing and endmember uncertainty estimation." *IEEE Transactions on Image Processing* 25 (2016).12: 5987–6002.

———. "A Gaussian mixture model representation of endmember variability in hyperspectral unmixing." *IEEE Transactions on Image Processing* 27 (2018).5: 2242–2256.

Zhou, Zhi-Hua, Zhang, Min-Ling, Huang, Sheng-Jun, and Li, Yu-Feng. "Multi-instance multi-label learning." *Artificial Intelligence* 176 (2012).1: 2291–2320.

Zou, Sheng and Zare, Alina. "Hyperspectral unmixing with endmember variability using partial membership latent dirichlet allocation." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, 6200–6204.

BIOGRAPHICAL SKETCH

Sheng Zou received his Bachelor of Science degree in Applied Physics from the Northeastern University of China in 2013. From 2013 to 2016, he continued his studies at the University of Missouri-Columbia to graduate with his Master of Science degree in Computer Engineering from the department of Electrical and Computer Engineering in 2016. His research interests include machine learning, hyperspectral classification and unmixing, remote sensing and image processing.