A bag-to-class divergence approach to multiple-instance learning

Abstract

A concise and factual abstract is required (maximum length 150 words). The abstract should state briefly the purpose of the research, the methods used, the principal results and major conclusions. An abstract is often presented separate from the article, so it must be able to stand alone.

Key words:
Multi-instance, Multiple-instance, Query-to-model, Object-to-class, Divergence, Dissimilarity, Classification, Uncertain object, Image analysis, Prototype, Template matching, Multiple-instance learning, bag-to-class

1. Introduction

1.1. History

Multiple-instance learning (MIL) is a form of supervised learning where each object consists of several observations. The observations themselves are not labelled, but the objects are. An object is referred to as *bag* in the MIL context, and the individual observations within the bag are referred to as *instances*. In binary MIL, the classes are referred to as *positive* and *negative*. Image classification is a typical example of MIL. Each image (bag) consists of a number of pixels (instances). An image in the training set is labelled 'positive' if it contains a certain material or body of interest, e.g. tumour tissue, and 'negative' if the material is absent. However, the exact location of the material within the image is unknown, and hence, the individual pixels are not labelled. A positive image contains pixels both from the tumour and from normal tissue, whereas a negative image contains only pixels from normal tissue. The task is to train a classifier using the pixels values with only information on image level.

The term multiple instance learning was introduced by Dietterich et at (1997). The musk data was introduced, where some molecules will give a musk odour (positive class), others will not (negative class). Each molecule has different shapes, and only certain shapes will give the musky smell. Hence, each molecule is a bag, and the various shapes are its instances. The assumption is that if a bag contains at least one positive instance, the bag is positive, whereas negative bags contains only negative instances. This is referred to as the standard assumption. The main task is to identify the positive instances. If this is done successfully, classification is straightforward.

Since then, the MIL field has developed and expanded. Most notably, the strict assumption of Dietterich has been relaxed and replaced by other assumptions. MIL was introduced as a binary classification problem, but is applicable also in multiple classification, regression and clustering. This article will stick to the binary classification, but will go beyond the standard assumption.

In the image example, the pixel intensity x_i measured in tumour tissue might typically be above a certain threshold T_{tissue} (positive instance). In normal tissue, the intensity might also be above the threshold, but it will be fewer pixels. Hence, both positive and negative bags contain positive instances, but the assumption is that positive bags contains a larger number of positive instances than negative bags.

Weidmann et al (2003) introduced a hierarchy of assumptions, with the standard assumption as the least general. In the review article of Foulds and Frank (2010), the Weidmann hierarchy is part of a more extensive taxonomy for MIL assumptions, where also the data representations and similarity measures are taken into account.

Amores (2013) provides a different viewpoint and taxonomy, focusing on the type of information; instance-level or bag-level, and the representation of the information; explicitly or implicitly. Both Amores and Foulds and Frank categorise a range of previously proposed algorithm within their taxonomy, many of them whose assumption or information have not been explicitly stated by the authors. Amores and Foulds and Frank offer different and complementary analyses of the MI problem and its proposed algorithms, and touch upon many of the same obstacles and choices:

- Assumption
- Type of information (instance vs bag)
- Information representation (explicit vs implicit)
- Dissimilarity measures
- Prototypes/concepts

As the MI methodology has developed over the last few decades, the need for relaxed assumptions has arisen and has been incorporated in new methods and algorithms. As in all fields of research, there is no universal MI methodology that solves all problems the best. Using stricter assumptions is beneficial if the data meet them, but can be devastating if they are violated. The problem to be solved can also be of various characteristics, and the methodology must be chosen accordingly.

This paper offers an alternative viewpoint for MI classification: Hierarchical distribution and bag-to-class divergence. The hierarchical distribution describes the assumption of the instances in each bag, whereas the bag-to-class divergence offers a dissimilarity measure. The alternative viewpoint can be fitted into both Foulds and Frank's and Amores taxonomy, by adding and specifying properties.

2. Probabilistic framework for MI problems

The instances in a bag are most often treated as fixed points, but can also be viewed as observations of random variables with an (unknown) probability distribution. The probabilistic viewpoint is not an MI assumption by itself, but can be used as a tool to analyse the problem and choose the appropriate concept.

For ease of notation, we will not use bold characters to denote vectors, as there is no need to distinguish between vectors and scalars in the following, and we will refer to random vectors as random variables. The superscript $^+$ or $^-$ refers to the instance class, the subscript $_{pos}$ or $_{neg}$ refers to the class label of the bag.

Let X^+ denote the random variable of a positive instance, and let $f^+(x|\theta_b^+)$ denote the pdf of the positive instances in bag b. Similarly, Let X^- denote the random variable of a negative instance, and let $f^-(x|\theta_b^-)$ denote the pdf of the negative instances in bag b. Then, the pdf of bag b is

$$f_b(x) = \pi_b^+ f^+(x|\theta_b^+) + (1 - \pi_b^+) f^-(x|\theta_b^-), \tag{1}$$

where $0 \leq \pi_b^+ \leq 1$ denotes the proportion of positive instances in bag b.

As an illustrative example, consider images of tissue with class label *tumour* (positive) or *normal* (negative). The image is the bag, and the pixels or image segments are its instances. The value of each pixel is then considered to be an observation from the underlying bag distribution. A pixel will belong to either tumour, with pdf $f^+(x|\theta_b^+)$, or normal tissue, $f^-(x|\theta_b^-)$, but the individual pixels are not labelled. In image b, a certain proportion are tumour pixels (π_b^+), the rest $(1 - \pi_b^+)$ are normal pixels. The pixels from tumour image b has pdf

$$f_{b,pos}(x) = \pi_{b,pos}^+ f^+(x|\theta_b^+) + (1 - \pi_{b,pos}^+) f^-(x|\theta_b^-). \tag{2}$$

Similarly, the pixels from a normal image b' has pdf

$$f_{b',neg}(x) = \pi_{b',neg}^+ f^+(x|\theta_{b'}^+) + (1 - \pi_{b,neg}^+) f^-(x|\theta_{b'}^-). \tag{3}$$

Since both tumour and normal tissue vary from image to image, we assume θ_b^+ being an observation from the distribution $P(\Theta|\tau^+)$ and θ_b^- being an observation from the distribution $P(\Theta|\tau^-)$, where τ^+ and τ^- are parameters.

And since the tumour size varies, we also assume that $\pi^+_{b,pos}$ is an observation from $P(\Pi^+_{pos})$. Written as a hierarchical distribution:

$$X_{pos}|\Theta \sim P(X_{pos}|\Theta) \tag{4}$$

$$\Theta|\mathcal{T} \sim P(\Theta|\mathcal{T}) \tag{5}$$

$$\mathcal{T} \sim \begin{cases} P(\mathcal{T} = \tau^+) = \Pi^+_{pos} \\ P(\mathcal{T} = \tau^-) = 1 - \Pi^+_{pos} \end{cases} \tag{6}$$

$$\Pi^+_{pos} \sim P(\Pi^+_{pos}) \tag{7}$$

Similarly, for the normal images we have

$$X_{neg}|\Theta \sim P(X_{neg}|\Theta) \tag{8}$$

$$\Theta|\mathcal{T} \sim P(\Theta|\mathcal{T}) \tag{9}$$

$$\mathcal{T} \sim \begin{cases} P(\mathcal{T} = \tau^+) = \Pi^+_{neg} \\ P(\mathcal{T} = \tau^-) = 1 - \Pi^+_{neg} \end{cases} \tag{10}$$

$$\Pi^+_{neg} \sim P(\Pi^+_{neg}) \tag{11}$$

2.1. MI assumption in the probabilistic framework

This is a general description of the origin of the instances, and does not imply any MI assumptions. Restrictions on the different levels of the hierarchical distribution will correlate to MI assumptions. The standard MI assumption is then
(1) $P(\Pi^+_{neg} = 0) = 1$: No positive instances in a negative bag
(2) $0 < \pi^+_{pos} \leq 1$: At least one positive instance in a positive bag
(3) $\mathcal{X}^+ \cap \mathcal{X}^- = \emptyset$: No overlap between the set of possible positive values and possible negative values
The last assumption is generally not explicitly stated, but many algorithms are dedicated to finding $\mathcal{X}^+$, and thereafter use a concept that assumes (3).

The count-based assumption is
(1) $t_{low,pos} < \pi^+_{pos} < t_{up,pos}$
(2) $\mathcal{X}^+ \cap \mathcal{X}^- = \emptyset$

The count-based GMIL can be defined in the probabilistic setting by letting π^+_{pos} be a vector with k elements, and $\pi^+_{b,pos} f^+(x|\theta^+_b) = \sum_{k=1}^K \pi^+_{b,pos,k} f^+_k(x|\theta^+_{b,k})$, with the restrictions
(1) $\max\{\pi^+_{pos,k} == 0\} = r$
(2) $\mathcal{X}^+_k$: defines *sufficiently close.*
(3) $\mathcal{X}^+_k \cap \mathcal{X}^- = \emptyset$ for all k

2.2. Level of information in the probabilistic framework

In the review article of Amores, MIL algorithms are categorised according to whether concepts are based on instance-level information or bag-level information. For the instance-level based concepts, the classifier is an aggregation of instance level scores, and any characteristic of the bags themselves are ignored. An example of this is the *collective* assumption in Foulds and Frank, where the concept is based on the (estimated) posterior class label of the instances: $P(c|b) = \frac{1}{n_b}\sum_{i=1}^{n_b} Pr(c|x_i)$.

The bag-level information is categorised into bag space and embedded space, where embedded space means that each bag is mapped to a single feature vector, and a single instance classifier is used. In the bag space, the classifiers are based on a dissimilarity measure (distances or kernel functions) comparing bags.

As with the MI assumptions, the probabilistic framework does not directly imply instance-level, bag-space or embedded-space. Since each bag is described as a pdf, it fits into the bag space. The probabilistic framework is applicable to instance-level information, as in the collective assumption. Embedded space can easily be use by estimating the parameters of the pdfs.

3. An MI classifier

As with all classification problems, if we knew the exact nature of the data, an optimal classifier could be chosen. Instead, we have to choose between very general assumptions that lead to sub-optimal classifiers, or stronger assumptions that ensure higher performance, but only if the assumptions are correct. There are three fundamental levels for MI problems:
(1) Choice of MI assumption
(2) Estimation
(3) Choice of classifier These three steps are present both for deterministic and probabilistic viewpoint, and step (3) is closely linked to step (1) and (2). In the probabilistic framework, the choice of MI assumption can be stated as restrictions on the underlying pdfs. The pdfs that are used in classification of new bags have to be estimated based on the available bags with labels. Finally, the classifier can be chosen and trained based on the previous assumptions and choices.

The most general MI assumption in the probabilistic framework is that $\tau^+ \neq \tau^-$ (the pdf of positive instances is different from the pdf of negative instances) and $\Pi_{pos}^+ \neq \Pi_{neg}^+$ (the distribution of the proportion of positive instances in a positive bag is different form that of positive instances in a negative bag).

Accurate estimation the four levels of the Baye's hierarchy is out of reach for most problems. Therefore, we suggest to directly estimate the pdf of a bag with unknown label based on its instances, $\hat{f}_{bag}(x)$. A dissimilarity measure can now be chosen to constitute the classifier.

In Amores, all bag level classifiers are based on bag-to-bag dissimilarity measures. As pointed out by the author, this has the drawback that the number of comparisons that has to be made increases exponentially, and so the computational burden. However, instance level algorithms suffer from inaccurate information (the true instance labels are not known), and the embedded space algorithms suffer from possible oversimplification.

We therefore suggest a bag level classifier based on the bag-to-class dissimilarity. Hence, there are only two comparisons for each classification.

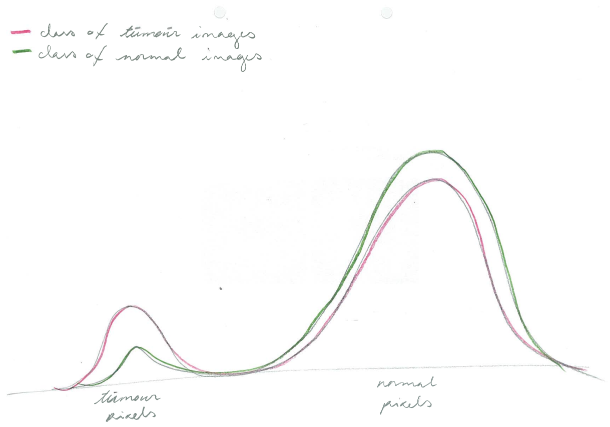
Let $f_{POS}(x)$ be the pdf corresponding to the random variabel X_{pos}, i.e. instances in the positive labelled bags. Similarly, let $f_{NEG}(x)$ be the pdf corresponding to the random variabel X_{neg}. Let $D(f_{bag}(x), f_{POS}(x)$ denote the dissimilarity measure from $f_{bag}(x)$ to $f_{POS}(x)$.

3.1. Divergences

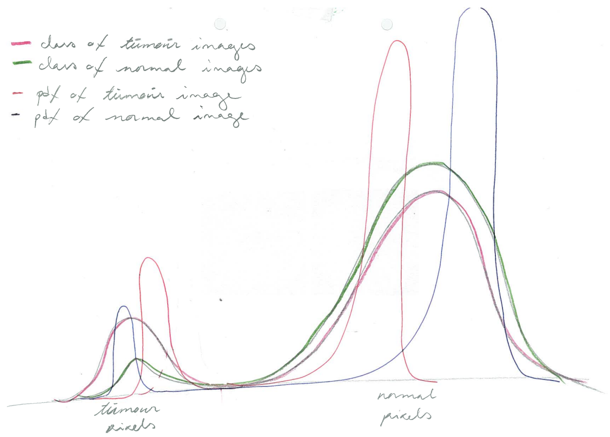
Dissimilarity measures between two probability distributions are referred to as divergence functions, or simply divergences, and measure the distance from one probability distribution to another. Divergences are not distance functions in the mathematical definition, as they don't necessarily fulfil the symmetry or the triangle inequality properties. Because distance between probability distributions is not uniquely defined, but very much depends on the problem at hand, there is a huge range of divergences. Common divergences are the Kullback-Leibler information, the Bhatacharrya distance, etc. There are no common properties that a function has to fulfil to be called a divergence, but many divergences have known properties, and are categorised accordingly.

In the MI classification, we suggest the following:
Because we measure the divergence from a bag to a class, we do not expect the bag's distribution to be equal to the class' distribution. A bag's distribution is one of the possible distributions of that class.

Back to the example of tissue images, we can illustrate it as follows. The class of tumour images and the class of normal images overlap, but there is a larger difference between the two classes for pixels values in the tumour



range than in the normal range. The pdf of a bag will not resemble the pdf of either classes of images.



If we could estimate $f^+(x)$, the distribution of the tumour pixels, and then π^+_{bag}, the proportion of tumour pixels in an unlabelled image, we could treat this as a count-based assumption and classify accordingly. However, by using the pdfs directly, without explicitly estimating the parameters in the different levels of the Bayes' hierarchy, we can possibly make more accurate estimates. In this example, we see that the overlapping area between bag and class is fairly similar for both classes in the normal pixel range. But for the tumour pixel range, the difference becomes larger. However, since the proportion of normal pixels is larger than that of tumour pixels, there is a risk of ignoring this if we measure dissimilarity by overlapping (or non-overlapping) areas. A different way of measuring dissimilarity is by distribution ratio $f_{bag}(x)/f_{POS}(x)$ and $f_{bag}(x)/f_{NEG}(x)$.

The Kullback-Leibler information from $f_{bag}(x)$ to $f_{POS}(x)$ is defined as

$$D_{KL}(f_{bag}, f_{POS}) = \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{POS}(x)} dx. \tag{12}$$

It is non-symmetric, and it is an f-divergence.

The Kullback-Leibler information is suited for the tissue image example, where the discriminative power lies with the tumour pixels, since the ratio of positive bag to negative class is large for tumour pixels.

If the proportion of tumour pixels in the normal image class is zero (standard MI assumption), the Kullback-Leibler information attains its maximum value (infinity) between a tumour image and the normal image class.

4. Distribution

Let $x_{b,i}$ be the observed instance i in bag b, an observation from the underlying distribution $P(X|\theta)$. An instance can either be positive $x^{+}_{b,i}$, or negative, $x^{-}_{b',i}$. Both positive and negative instances occur in both positive and negative bags. An instance randomly drawn from a positive bag will be positive with probability π^{+}_{pos} (and negative with probability $1 - \pi^{+}_{pos}$). An instance randomly drawn from a negative bag will be positive with probability π^{+}_{neg} (and negative with probability $1 - \pi^{+}_{neg}$). The distribution of positive instances is $P(X|\theta^{+}_{b})$, where θ^{+} is distributed according to $P(\theta^{+}|\tau^{+})$, and τ^{+} is a fixed parameter. The distribution of negative instances is $P(X|\theta^{-}_{b})$, where θ^{-} is distributed according to $P(\theta^{-}|\tau^{-})$, and τ^{-} is a fixed parameter. The pdf of a positive bag b can be written as

$$p_{bag(pos),b}(x) = \pi^{+}_{pos} p^{+}(x|\theta^{+}_{b}) + (1 - \pi^{+}_{pos}) p^{-}(x|\theta^{-}_{b}), \tag{13}$$

and that of a negative bag b'

$$p_{bag(neg),b'}(x) = \pi^{+}_{neg} p^{+}(x|\theta^{+}_{b'}) + (1 - \pi^{+}_{neg}) p^{-}(x|\theta^{-}_{b'}), \tag{14}$$

This is a Bayesian hierarchical distribution where the hyperparameters π^{+}_{pos} and π^{+}_{neg} differ between the positive and the negative class, whereas the hyperparameters θ^{+} and θ^{-} differ from bag to bag.

For $x_{pos,i}$, an observed instance in the positive class, the empirical pdf of the positive class can be written as

$$p_{pos}(x) = \frac{1}{B} \sum_{b=1}^{B} p_{bag(pos),b}(x), \tag{15}$$

and likewise for the negative class

$$p_{neg}(x) = \frac{1}{B'} \sum_{b'=1}^{B'} p_{bag(neg),b'}(x). \tag{16}$$

B and B' are the number of all possible bag distributions.

We assume that $\pi^{+}_{pos} > \pi^{+}_{neg}$, that is, the probability of a randomly drawn instance from a randomly drawn positive bag being positive is greater than if the instance were drawn from a negative bag.

We make no assumptions regarding the distributions, besides continuity, and we will not attempt to estimate the parameters.

Instead, the distributions will be approximated by a mixture of Gaussian distributions.

5. Dissimilarity measure

A common approach for MIL is to use some form of dissimilarity measure in the classification of new bags. For instance-level information, the dissimilarity is measured between the instances in a bag and instances in labelled bags or prototypes. Examples are the Harrington distance, and its variants. For bag-level information, the dissimilarity is measured between the bag as a whole and labelled bags or prototypes. For explicit information, the dissimilarity is calculated based on the instance values. For implicit information, the instance values are transformed to a single vector (e.g. the median), and a single-instance dissimilarity measure can be used. Hybrid versions also exist, see e.g. Chepuliga (2016).

We here introduce two new approaches that has not been studied in the MIL context: (1) Divergence-based dissimilarity measures. (2) Bag-to-class dissimilarities.

A divergence function (or simply divergence) is a measure of dissimilarity between two probability distributions. A divergence does not necessarily fulfil the mathematical requirements of a distance, especially the triangle inequality. In the MIL context, the instances of a bag form the basis of a probability distribution estimate, and then a divergence can be applied.

$$D(P_{bag}, P_{ref})\,, \tag{17}$$

where P_{bag} is the probability distribution of the bag that we want to classify, and P_{ref} is the reference distribution. P_{ref} can be any distribution, but would typically be that of labelled bags, classes, or prototype distributions.

In practice, the distributions must be estimated from the instances of the bag, using the assumption that they are independent observations from a common underlying distribution. Commonly used in MIL is the EM-algorithm. Which method to apply for distribution estimation depends on the data, especially the sparsity, previous knowledge, and requirements regarding time consumption. We will not go into the details, but simply use the EM-algorithm. Assuming that the instances are observations from an underlying distribution is the collective assumption.

The bag-to-class dissimilarity has not been used in MIL, although there are obvious advantages: For one, the computation time decreases when the dissimilarity is calculated only between a bag and the classes, compared to pairwise dissimilarities between a bag and all labelled bags or prototypes. A class can be seen as a prototype, and in that case, bag-to-class will be the minimum number of prototypes. We propose a dissimilarity measure fro bag-to-class comparison, based on the divergence between two probability distributions. We will also shortly discuss divergence for bag-to-bag comparison. A bag-to-class approach can be used without a divergence.

A huge variation among divergences exists. Popular functions are the Kullback-Leibler information (non-symmetric), the KL divergence, the Shannon entropy, and many more. Choice of divergence must reflect the problem at hand. See e.g. Mollersen et al for properties of some common divergences.

Let there be two classes, positive and negative. An instance, x_i is a sample from an underlying distribution, P_{pos} or P_{neg}. A bag contains instances from a hierarchical distribution

$$X_{ij} \sim \pi_j P_+ + (1-\pi)P_-\,, \tag{18}$$

where π_j is a parameter of the class. This means that we can allow both positive and negative instances to appear in both positive and negative bags.

Ultimately, $P_{bag,pos} \neq P_{bag,neg}$, or else they cannot be distinguished, so we require $\pi_{pos} \neq \pi_{neg}$.

In the standard assumptions, we have

$$X_{i,pos} \sim \pi_{pos} P_+ + (1 - \pi_{pos}) P_- \,, \tag{19}$$

where $\pi_{pos} > 0$, and

$$X_{i,neg} \sim P_- \tag{20}$$

This can be relaxed to

$$\begin{aligned} X_{i,pos} &\sim \pi_{pos} P_+ + (1 - \pi_{pos}) P_- && (21)\\ X_{i,neg} &\sim \pi_{neg} P_+ + (1 - \pi_{neg}) P_- \,, && (22) \end{aligned}$$

where $\pi_{pos} \sim P_{\pi_{pos}}$ and $\pi_{neg} \sim P_{\pi_{neg}}$, where $P_{\pi_{pos}} \neq P_{\pi_{neg}}$. This corresponds to the top level of Weidmanns hierarchy.

We assume that P_+ and P_- are parametrised distributions, where θ_+ and θ_- follow distributions on their own. This means that within a bag, θ_+ and θ_- are constants, but varies between bags. We believe this is a more realistic approach than assuming that all positive instances come from the same distribution.

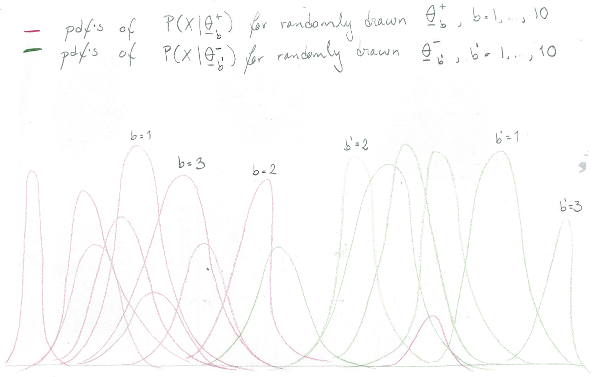
Example:

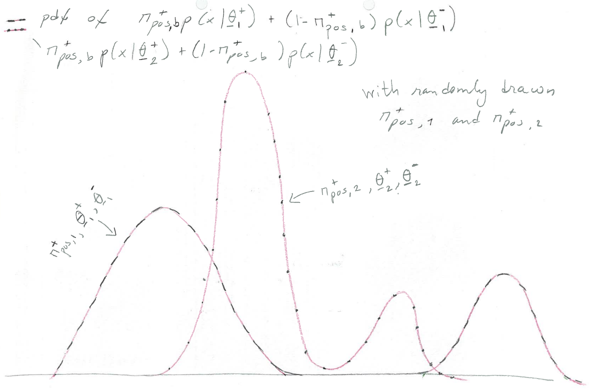
$$\begin{aligned} P_+ :& \mathcal{N}(\mu_+, \sigma_+^2) && (23)\\ \mu_+ \sim& \mathcal{N}(\nu_\mu, \tau_\mu^2) && (24)\\ \sigma_+^2 \sim& \mathcal{N}(\nu_\sigma, \tau_\sigma^2), && (25) \end{aligned}$$

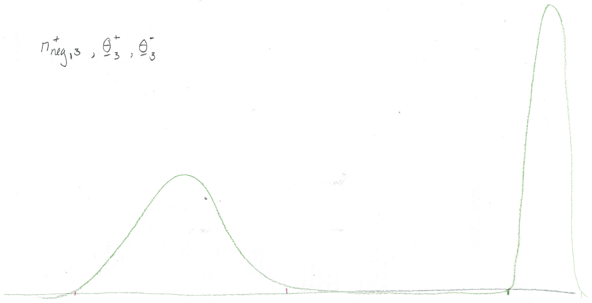
where ν and τ are constants, and μ_+ and σ_+ are drawn once for each bag.

The two class distributions are estimated based on the bag labels. The goal is not to distinguish between P_+ and P_-. We will get P_{pos} and P_{neg}. For a new bag, we want to measure $D(P_{bag}, P_{pos})$ and/or $D(P_{bag}, P_{neg}$. It is important to notice that P_{pos} is not equal to any $P_{bag,pos}$, but rather a collection of all possible $P_{bag,pos}$s with their probability of occurrence taken in mind. Therefore, symmetry of the divergence is not a requirement.

Let $\mathcal{X}_{pos>bag,pos}$ be the region where $p_{pos}(x) > p_{bag,pos}(x)$. Then $\mathcal{X}_{pos>bag,pos} \geq \mathcal{X}_{pos\leq bag,pos}$, meaning that the region where p_{pos} is greater than p_{p_bag} is bigger than the opposite. This comes from the hierarchical nature of P_{pos}, where







$\theta_{pos} = [\pi_{pos}\ \theta_+\ \theta_-]$ has a distribution, and p_{p_bag} is uniquely defined by the

θ_{pos} sample. An illustration

The goal is to pick a divergence that is able to discriminate between P_{p_bag} and P_{n_bag} by measuring the dissimilarity to P_{pos} and/or P_{neg}. Let $\mathcal{X}_{pos>neg}$ be the region where $p_{pos}(x) > p_{neg}(x)$. Unless $p_{pos} = p_{neg}$ we have $\mathcal{X}_{pos>neg}$ nonempty, which gives $\mathcal{X}_{pos<neg}$ nonempty, whereas $\mathcal{X}_{pos=neg}$ might be empty or not.

The Kullback-Leibler information (KL inf),

$$d_{KL}(p_{bag}, p_{neg}) = \int_{\mathcal{X}} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx, \tag{26}$$

is a non-symmetric f-divergence. The log ratio function gives a positive contribution whenever $p_{bag} > p_{neg}$, and a negative contribution for $p_{bag} < p_{neg}$, and zero contribution for $p_{bag} = p_{neg}$. A large positive contribution for $p_{bag} >> p_{neg}$ and $p_{bag} >> 0$, which means that if p_{bag} is outside the range of p_{neg}, the dissimilarity approaches infinity. This is a suitable property, because if $p_{bag}/p_{neg} \to \infty$, the probability the parameters of p_{bag} are not sampled from the negative class. A simple straightforward measure is then

$$D(p_{bag}, p_{neg}) = d_{KL}(p_{bag}, p_{neg}), \tag{27}$$

and the ratio $d_{KL}(p_{bag}, p_{pos})/d_{KL}(p_{bag}, p_{neg})$ will give a classification rule.

We will propose a divergence-based dissimilarity function where the two classes are integrated

$$D(p_{bag}, p_{neg}|p_{pos}) = \int_{\mathcal{X}_{pos}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx \tag{28}$$

Have a look at this

$$\int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx \leq\geq a \int_{\mathcal{X}_{neg}} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx \tag{29}$$

$$D^*(p_{bag}, p_{neg}|p_{pos}) = \int_{\mathcal{X}_{pos}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx + \int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag}(x) \log \frac{p_{bag}(x)}{p_{neg}(x)} dx \tag{30}$$

Because we also assume that $\pi_{pos} > \pi_{neg}$ it follows that $p_{p\_bag} < p_{n\_bag}, X \in \mathcal{X}_{neg}$ and therefore

$$\int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{p\_bag}(x) \log \frac{p_{p\_bag}(x)}{p_{neg}(x)} dx < \int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{n\_bag}(x) \log \frac{p_{n\_bag}(x)}{p_{neg}(x)} dx \tag{31}$$

which is an unwanted property, and therefore we use only $\mathcal{X}_{pos}$.

Like with KL inf, the log ratio function ensures large positive contributions for $p_{bag} >> p_{neg}$ when also $p_{bag} >> 0$. In addition, we require that $p_{pos} > p_{neg}$ for this contribution to be large. This is because if $p_{bag} >> p_{neg}$ but $p_{pos} < p_{neg}$ we have a bag whose pdf cannot be explained by the negative class, but neither by the positive class, and therefore is uninformative for classification. If $p_{pos} > p_{neg}$, or even $p_{pos} >> p_{neg}$, then $D \to \infty$. How is this different from $d_{KL}(p_{bag}, p_{pos})/d_{KL}(p_{bag}, p_{neg})$? If $p_{bag}/p_{pos} \to \infty$ and $p_{bag}/p_{neg} \to \infty$, then the ratio will be one.

6. Conditional

$$p_{bag(pos)} = \pi^{+}_{pos} p^{+}(x|\theta^{+}_{b}) + (1 - \pi^{+}_{pos}) p^{-}(x|\theta^{-}_{b}) \tag{32}$$

$$p_{bag(neg)} = \pi^{+}_{neg} p^{+}(x|\theta^{+}_{b}) + (1 - \pi^{+}_{neg}) p^{-}(x|\theta^{-}_{b}) \tag{33}$$

$$B \to \infty : \sum_{b=1}^{B} p^{+}(x|\theta^{+}_{b}) = \sum_{b'=1}^{B'} p^{+}(x|\theta^{+}_{b'}) \tag{34}$$

$$\mathcal{X}_{neg} : p_{pos}(x) < p_{neg}(x) \tag{35}$$

$$\mathcal{X}_{neg} : E(p_{bag(pos)}) < E(p_{bag(neg)}) \tag{36}$$

$$\int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag(pos)}(x) \log \frac{p_{bag(pos)}(x)}{p_{neg}(x)} dx < \int_{\mathcal{X}_{neg}} \frac{p_{pos}(x)}{p_{neg}(x)} p_{bag(neg)}(x) \log \frac{p_{bag(neg)}(x)}{p_{neg}(x)} dx$$

7. Bag-to-class

Why not use bag-to-bag?

$$\frac{p_{bag,b}}{p_{bag,b'}} \to \infty \tag{37}$$

8. Non-hierarchical model

The positive instances are observations from $p^+(x|\theta^+)$. The negative instances are observations from $p^-(x|\theta^-)$. A positive bag contains the observations:

$$p_{bag(pos)}(x) = \pi^+_{pos} p^+(x|\theta^+) + (1 - \pi^+_{pos}) p^-(x|\theta^-) \tag{38}$$

A negative bag contains the observations:

$$p_{bag(neg)}(x) = \pi^+_{neg} p^+(x|\theta^+) + (1 - \pi^+_{neg}) p^-(x|\theta^-) \tag{39}$$

If all positive bags follow the same distribution and all negative bags follow the same distribution, then the best estimation of $p_{bag(pos)}(x)$ is to pool all instances from the positive bags, and we get $\hat{p}_{bag(pos)}(x)$. We can then look at $\hat{p}_{bag(pos)}(x)$ and $\hat{p}_{bag(neg)}(x)$ as prototypes to which the distance from $\hat{p}_{bag}(x)$ is measured.

However, more realistically, we assume that

$$p_{bag(pos),b}(x) = \pi^+_{pos,b} p^+(x|\theta^+_b) + (1 - \pi^+_{pos,b}) p^-(x|\theta^-_b) \tag{40}$$

and that

$$p_{pos}(x) = \sum_{b=1}^{B} p_{bag(pos),b}(x) = \sum_{b=1}^{B} \pi^+_{pos,b} p^+(x|\theta^+_b) + (1 - \pi^+_{pos,b}) p^-(x|\theta^-_b) \tag{41}$$

Assume that $\pi^+_{pos} p^+(x|\theta^+)$ to $\pi^+_{neg} p^+(x|\theta^+)$ is more discriminative than $(1 - \pi^+_{pos}) p^-(x|\theta^-)$ to $(1 - \pi^+_{neg}) p^-(x|\theta^-)$.

$$\frac{\pi^+_{pos}}{\pi^+_{neg}} > \frac{1 - \pi^+_{pos}}{1 - \pi^+_{neg}} \tag{42}$$

Therefore, non-symmetric divergence function. Kullback-Leibler meets the requirement.

9. Bayes hierarchy

A random variable X_{pos} from a positive bag can be seen as a three level Bayes hierarchy. X_{pos} is distributed with parameter θ, which is a random variable. θ is distributed with parameter τ, which takes value τ^+ with probability $p = \pi^+_{pos}$ and τ^- with probability $p = 1 - \pi^+_{pos}$.

$$X_{pos}|\theta \sim P(X_{pos}|\theta) \tag{43}$$

$$\theta|\tau \sim P(\theta|\tau) \tag{44}$$

$$\tau \sim \begin{cases} \tau^+, & \text{with probability } p = \pi^+_{pos} \\ \tau^-, & \text{with probability } p = 1 - \pi^+_{pos} \end{cases} \tag{45}$$

The pdf of the bth positive bag is then

$$f_{b,pos}(x) = \pi^+_{pos} f^+(x|\theta^+_b) + (1 - \pi^+_{pos}) f^-(x|\theta^-_b), \tag{46}$$

where θ^+_b is the bth observation of the random variable θ with parameter τ^+, and θ^-_b is the bth observation of the random variable θ with the parameter τ^-. The pdf of positive bags is then

$$f_{pos}(x) = \pi^+_{pos} \int_{\Theta|\tau^+} f^+(x|\theta) h(\theta|\tau^+) d\theta|\tau^+ + (1 - \pi^+_{pos}) \int_{\Theta|\tau^-} f^-(x|\theta) h(\theta|\tau^-) d\theta|\tau^-,$$

where θ^+_b is the bth observation of the random variable θ with parameter τ^+, and θ^-_b is the bth observation of the random variable θ with the parameter τ^-.

Similarly for negative bags we have

$$X_{neg}|\theta \sim P(X_{neg}|\Theta) \tag{47}$$

$$\Theta|\mathcal{T} \sim P(\theta|\tau) \tag{48}$$

$$\tau \sim \begin{cases} P(\mathcal{T} = \tau^+) = \pi^+_{neg} \\ P(\mathcal{T} = \tau^- = 1 - \pi^+_{neg} \end{cases} \tag{49}$$

and

$$f_{b',neg}(x) = \pi^+_{neg} f^+(x|\theta^+_{b'}) + (1 - \pi^+_{neg}) f^-(x|\theta^-_{b'}), \tag{50}$$

10. Equality and orthogonality

1. $f_{bag}(x) = f_{NEG}(x)$
2. $f_{bag}(x) = f_{POS}(x)$
3. $f_{POS}(x) = f_{NEG}(x)$
4. $f_{bag}(x) \perp f_{NEG}(x)$
5. $f_{bag}(x) \perp f_{POS}(x)$
6. $f_{POS}(x) \perp f_{NEG}(x)$

$condI_{KL}(bag, \textsc{neg}|\textsc{pos}) = \int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx$

1. $\int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{NEG}(x) \log \frac{f_{NEG}(x)}{f_{NEG}(x)} dx = 0$
2. $\int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{POS}(x) \log \frac{f_{POS}(x)}{f_{NEG}(x)} dx$
3. $\int_{\mathcal{X}} f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx$
4. $\int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx = \infty$
5. $\int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx = 0$
6. $\int_{\mathcal{X}_{POS}} \frac{f_{POS}(x)}{f_{NEG}(x)} f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx = \infty$

$D_{KL} = \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx / \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{POS}(x)} dx$

1. $\int f_{NEG}(x) \log \frac{f_{NEG}(x)}{f_{NEG}(x)} dx / \int f_{NEG}(x) \log \frac{f_{NEG}(x)}{f_{POS}(x)} dx = 0$
2. $\int f_{POS}(x) \log \frac{f_{POS}(x)}{f_{NEG}(x)} dx / \int f_{POS}(x) \log \frac{f_{POS}(x)}{f_{POS}(x)} dx = \infty$
3. $\int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx / \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx = 1$
4. $\int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx / \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{POS}(x)} dx = \infty$
5. $\int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx / \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{POS}(x)} dx = 0$
6. $\int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{NEG}(x)} dx / \int f_{bag}(x) \log \frac{f_{bag}(x)}{f_{POS}(x)} dx$

11. Regions

$condI_{KL}(bag, \textsc{neg}|\textsc{pos})$
$\mathcal{X}_{POS} : condI_{KL}(bag(\textsc{pos})) > condI_{KL}(bag(\textsc{neg}))$

$\mathcal{X}_{NEG} : condI_{KL}(bag(\textsc{pos})) < condI_{KL}(bag(\textsc{neg}))$

$$D_{KL}(\frac{bag, \textsc{pos}}{bag, \textsc{neg}}) \tag{51}$$
$$\mathcal{X}_{POS} : I_{KL}(bag(\textsc{pos}), \textsc{pos}) > I_{KL}(bag(\textsc{neg}), \textsc{pos}) \tag{52}$$
$$: I_{KL}(bag(\textsc{pos}), \textsc{neg}) > I_{KL}(bag(\textsc{neg}), \textsc{neg}) \tag{53}$$
$$D_{KL}? \tag{54}$$
$$\mathcal{X}_{NEG} : \text{similar} \tag{55}$$

12. Notes 1

(1) Method for density estimation
(2) Dissimilarity measure
(3) Classifier
(1) Bayesian framework
(2) Non-symmetric divergence function
(3) Conditional divergence
Concept: Automatically identifies the concept(s) as $x \in \mathcal{X} : f_{pos} > f_{neg}$
"Point set distance measures"
Strong class overlap
Assumptions?

$$\frac{\pi^+_{pos}}{\pi^+_{neg}} \text{ vs } \frac{a - \pi^+_{pos}}{1 - \pi^+_{neg}} \tag{56}$$
$$\theta^+ \text{ vs } \theta^- \tag{57}$$

Class level: Good density estimates
High dimensions: Assume independence

12.1. Kullback-Leibler

$f_{bag} \perp f_{neg} : D_{KL} = \infty$ contribution
Conditional: $f_{bag} \perp f_{neg}$ and in addition $f_{bag} \perp f_{pos} : D_{KL|pos} = 0$ contribution

13. Notes 2

- Drawback: Not for one-vs-all
- Multi-label (instance?) as semi-supervised or multi-class of fussy?
- Other: Multi-label learning, regression, clustering, semi-supervised
- Kernel density estimation
- Dimensionality
- Prototype = pdf from training set?
- No estimation of weight
- Drawback of classical CBIR

14. Nomenclature

$f(x)$: probability density function (pdf)
$f^{+}(x)$: pdf of positive instances
$f^{-}(x)$: pdf of negative instances
$f_{bag}(x)$: pdf of bag (unknown class)
$f_{pos}(x)$: pdf of positive bag
$f_{neg}(x)$: pdf of negative bag
$f_{POS}(x)$: pdf of positive class
$f_{NEG}(x)$: pdf of negative class
π^{+}_{pos} : probability of an instance in a positive bag being sampled for $f^{+}(x)$
π^{+}_{neg} : probability of an instance in a negative bag being sampled for $f^{+}(x)$
X_{pos} : random variabel in positive bag
$x_{pos,b,i}$: observation i in positive bag b
X_{neg} : random variabel in negative bag
$x_{neg,b',i}$: observation i in negative bag b'
Θ^{+} : random variable with distribution parameter τ^{+}
θ^{+}_{b} : bth distribution parameter (observation of Θ^{+})
Θ^{-} : random variable with distribution parameter τ^{-}
θ^{-}_{b} : bth distribution parameter (observation of Θ^{-})
$\mathcal{T}$: random variable
τ^{+}, τ^{-} : observations of $\mathcal{T}$

$I_{KL}(bag, \textsc{pos}) = I_{KL}(f_{bag}(x), f_{POS}(x))$
$D_{KL} = I_{KL}(bag, \textsc{neg})/I_{KL}(bag, \textsc{pos})$
$condI_{KL}(bag, \textsc{neg}|\textsc{pos}) = I_{KL}(f_{bag}(x), f_{NEG}(x)|f_{ALT}(x) = f_{POS}(x))$

References