Semantic Channel and Shannon's Channel Mutually Match for Multi-Label Classification

Chenguang Lu [0000-0002-8669-0094]

College of Intelligence Engineering and Mathematics,
Liaoning Engineering and Technology University, Fuxin, Liaoning, 123000, China
lcguang@foxmail.com

Abstract. A semantic channel consists of a set of membership functions or truth functions which indicate the denotations of a set of labels. In the multi-label learning, we obtain a semantic channel from a sampling distribution or Shannon's channel. If samples are huge, we can directly convert a Shannon's channel into a semantic channel by the third kind of Bayes' theorem; otherwise, we can optimize the membership functions by a generalized Kullback-Leibler formula. In the multi-label classification, we partition an instance space with the maximum semantic information criterion, which is a special Regularized Least Squares (RLS) criterion and is equivalent to the maximum likelihood criterion. To simplify the learning, we may only obtain the truth functions of some atomic labels to construct the truth functions of compound labels. In a label's learning, instances are divided into three kinds (positive, negative, and unclear) instead of two kinds as in the One-vs-Rest or Binary Relevance (BR) method. Every label's learning is independent as in the BR method. However, it is allowed to train a label without negative examples and a number of binary classifications are not used. In the label selection, for an instance, the classifier selects a compound label with the most semantic information. This classifier has taken into the consideration the correlation between labels already. For example, it will not add label "Adult" or "Non-youth" to an example that already has the label "Old person". As a predictive model, the semantic channel does not change with the prior probability distribution (source) of instances. It still works when the source is changed. The classifier does change with the source and hence can overcome the class-imbalance problem. It is shown that the old population's increase will change the classifier for label "Old person" and has been impelling the evolution of the semantic meaning of "Old". The CM iteration algorithm for unseen instance classification is introduced.

Keywords: Shannon's channel, Bayes' theorem, Natural language processing, Semantic information, Multi-label classification, Membership function, Semi-supervised learning.

1 Introduction

Multi-label classification generally includes two steps: multi-label learning and multi-label selection. In multi-label learning, for every label y_j, we need to train its

posterior probability estimation $P(y_j|X;\theta)$ by a sample, where X is a random variable denoting an instance, and θ is a predictive model with parameters. In multi-label selection, we need to partition the instance space into different classes with specific criteria and to label every class.

There have been many valuable studies about multi-label classifications [1-2]. Information, cross-entropy, and uncertainty criteria also have been used [3-5]. In label learning, if there are more than two labels to be learned, it is hard to obtain the posterior probability estimation $P(y_j|X;\theta)$ (j=1,2,…n) because of the need of normalization ($\sum_j P(y_j|X;\theta)$ =1). Therefore, most researchers convert multi-label learning into multiple single label learnings [1]. The One-vs-Rest is a famous method [1]. However, in some cases, the conversion is improper. For instance, a sample has two examples (age 25, "Youth") and (age 24, "Adult"); it is unreasonable to regard (age 24, "Adult") as a negative example of "Youth". Therefore, the Binary Relevance (BR) method [2] requires that every instance is related to n labels with n "Yes" or "No". However, it demands too much of samples. It has another problem because of labels' correlation. In natural language, many negative labels, such as "Non-youth" and "Non-old person", are rarely used; it is unnecessary to add label "Adult" or "Non-youth" to an instance with label "Old person". So, the canonical BR relevance method needs improvements.

For the above reasons, the author develops a new method. It is similar to the BR method but has the following distinct features:

- It uses membership functions instead of posterior probability estimations so that the normalization is unnecessary; the learned membership functions can be used to classify other samples with different distribution $P(X)$.
- In the label learning, we directly obtain membership functions from sampling distribution $P(X, Y)$ (Y is a label), by a new Bayes' formula or a generalized Kullback-Leibler (KL) formula, without binary classifications. This method allows us to train a label with only positive examples. It is not necessary to prepare n data sets for n labels.
- In label selection or classification, we use the Maximum Semantic Information (MSI) criterion to partition the instance space and to label classes. The classifier changes with $P(X)$.

This paper is based on the author's studies on the semantic information theory [6-8], maximum likelihood estimations (semi-supervised learning) [9], and mixture models (unsupervised learning) [10]. In the recent two decades, the cross-entropy method has become popular [11]. The author's above studies and this paper use not only cross-entropy but also mutual cross-entropy.

The main contributions of this paper are:

- Providing a new Bayes' formula, which can directly derive labels' membership functions from continuous sampling distributions $P(X, Y)$.
- Proving that the Maximum Semantic Information (MSI) criterion is a special Regularized Least Squares (RLS) criterion.
- Simplifying multi-label classification by the mutual matching of the semantic channel and Shannon's channel, without a number of binary classifications and labels' correlation problem.

- Overcoming the class-imbalance problem and explaining the classification of "Old people" changes with the age population distribution in natural language.

The rest of this paper is organized as follows. Section 2 provides mathematical methods. Section 3 discusses multi-label classifications and relevant problems. Section 4 introduces an iterative algorithm for unseen instance classifications. Section 5 is the summary.

2 Mathematical Methods

2.1 Distinguishing Statistical Probability and Logical Probability

Definition 1 Let U denote the instance set, and X denote a discrete random variable taking a value from $U=\{x_1, x_2, \ldots\}$. For the convenience of theoretical analyses, we assume that U is one-dimensional. Let L denote the set of selectable labels, including some atomic labels and compound labels and let $Y\in L=\{y_1, y_2, \ldots\}$. Similarly, let L_a denote the set of some atomic labels and let $a\in L_a=\{a_1, a_2, \ldots\}$.

Definition 2 A label y_j is also a predicate $y_j(X)=$ "$X\in A_j$." For each y_j, U has a subset of A_j, every instance of which makes y_j true. Let $P(Y=y_j)$ denote the statistical probability of y_j, and $P(X\in A_j)$ denote the Logical Probability (LP) of y_j. For simplicity, let $P(y_j)= P(Y=y_j)$ and $T(y_j)=T(A_j)= P(X\in A_j)$.

We call $P(X\in A_j)$ the logical probability because according to Tarski's theory of truth [12], $P(X\in A_j)=P($"$X\in A_j$" is true$)=P(y_j$ is true). Hence the conditional LP of y_j for given X is the feature function of A_j and the truth function of y_j. We denote it with $T(A_j|X)$. There is

$$T(A_j)=\sum_i P(x_i)T(A_j \mid x_i) \tag{2.1}$$

According to Davidson's truth-conditional semantics [13], $T(A_j|X)$ ascertains the semantic meaning of y_j. Note that statistical probability distribution, such as $P(Y)$, $P(Y|x_i)$, $P(X)$, and $P(X|y_j)$, are normalized whereas the LP distribution is not normalized. For example, in general, $T(A_1|x_i)+T(A_2|x_i)+\ldots+T(A_n|x_i)>1$.

For fuzzy sets [14], we use θ_j as a fuzzy set to replace A_j. Then $T(\theta_j|X)$ becomes the membership function of θ_j. We can also treat θ_j as a sub-model of a predictive model θ. In this paper, likelihood function $P(X|\theta_j)$ is equal to $P(X|y_j; \theta)$ in the popular method.

2.2 Three Kinds of Bayes' Theorems

There are three kinds of Bayes' theorem, which are used by Bayes [15], Shannon [16], and the author respectively.

Bayes' Theorem I (used by Bayes): Assume that sets A, $B\in 2^U$, A^c is the complementary set of A, $T(A)=P(X\in A)$, and $T(B)= P(X\in B)$. Then

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)=T(A|B)T(B)+T(A|B^c)T(B^c) \tag{2.2}$$

There is also a symmetrical formula for $T(A|B)$. Note there are only one random variable X and two logical probabilities.

Bayes' Theorem II (used by Shannon): Assume that $X\in U$, $Y\in L$, $P(x_i)=P(X=x_i)$, and $P(y_j)=P(Y=y_j)$. Then

$$P(x_i \mid y_j)=P(y_j \mid x_i)P(x_i)/P(y_j),\ P(y_j)=\sum_i P(x_i)P(y_j \mid x_i) \tag{2.3}$$

There is also a symmetrical formula for $P(y_j|x_i)$. Note there are two random variables and two statistical probabilities.

Bayes' Theorem III: Assume that $P(X)=P(X=\text{any in } U)$ and $T(\theta_j)=P(X\in\theta_j)$. Then

$$P(X \mid \theta_j)=T(\theta_j \mid X)P(X)/T(\theta_j),\ T(\theta_j)=\sum_i P(x_i)T(\theta_j \mid x_i) \tag{2.4}$$

$$T(\theta_j \mid X)=P(X \mid \theta_j)T(\theta_j)/P(X),\ T(\theta_j)=1/\max(P(X \mid \theta_j)/P(X)) \tag{2.5}$$

The two formulas are asymmetrical because there is a statistical probability and a logical probability. $T(\theta_j)$ in (2.5) may be called longitudinally normalizing constant.

The Proof of Bayes' Theorem III: Assume the joint probability $P(X,\ \theta_j)=P(X=\text{any}, X\in\theta_j)$, then $P(X|\theta_j)T(\theta_j)=P(X=\text{any}, X\in\theta_j)=T(\theta_j|X)P(X)$. Hence there is

$$P(X \mid \theta_j)=P(X)T(\theta_j \mid X)/T(\theta_j),\ T(\theta_j|X)=T(\theta_j)P(X \mid \theta_j)/P(X)$$

Since $P(X|\theta_j)$ is horizontally normalized, $T(\theta_j)=\sum_i P(x_i)\ T(\theta_j|x_i)$. Since $T(\theta_j|X)$ is longitudinally normalized and has the maximum 1, we have

$$1=\max[T(\theta_j)P(X|\theta_j)/P(X)]=T(\theta_j)\max[P(X|\theta_j)/P(X)]$$

Hence $T(\theta_j)=1/\max[P(X|\theta_j)/P(X)]$. **QED.**

2.3 From Shannon's Channel to Semantic Channel

In Shannon's information theory [16], $P(X)$ is called the source, $P(Y)$ is called the destination, and the transition probability matrix $P(Y|X)$ is called the channel. So, a channel is formed by a set of transition probability function: $P(Y|X)$: $P(y_j|X)$, $j=1, 2, \ldots, n$.

Note that $P(y_j|X)$ (y_j is constant and X is variable) is different from $P(Y|x_i)$ and also not normalized. It can be used for Bayes' prediction to get $P(X|y_j)$. When $P(X)$ becomes $P'(X)$, $P(y_j|X)$ still works. $P(y_j|X)$ by a constant k can make the same prediction because

$$\frac{P'(X)kP(y_j \mid X)}{\sum_i P'(x_i)kP(y_j \mid x_i)}=\frac{P'(X)P(y_j \mid X)}{\sum_i P'(x_i)P(y_j \mid x_i)}=P'(X \mid y_j) \tag{2.6}$$

Similarly, a set of truth functions forms a semantic channel: $T(\theta|X)$ or $T(\theta_j|X)$, j=1, 2, …, n. According to (2.6), if $T(\theta_j|X) \propto P(y_j|X)$, there is $P(X|\theta_j)=P(X|y_j)$. Conversely, let $P(X|\theta_j)=P(X|y_j)$, then we have $T(\theta_j|X)=P(y_j|X)/\max(P(y_j|X))$.

2.4 To Define Semantic Information with Log (Normalized Likelihood)

The (amount of) semantic information conveyed by y_j about x_i is defined with log-normalized-likelihood [8, 9]:

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{2.7}$$

For an unbiased estimation y_j, its truth function may be a Gaussian function:

$$T(\theta_j|X)=\exp[-(X-x_j)^2/(2d^2)] \tag{2.8}$$

Then $I(x_i;\theta_j)=\log[1/T(\theta j)]-(X-x_j)^2/(2d^2)$. It clearly shows that this information criterion reflects Popper's thought [17]. It tells that the larger the deviation is, the less information there is; the less the logical probability is, the more information there is; and, a wrong estimation may convey negative information. To average $I(x_i;\theta_j)$, we have

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{2.9}$$

$$\begin{aligned} I(X;\theta)&=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}\\ &=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}=H(\theta)-H(\theta\mid X)\end{aligned} \tag{2.10}$$

$$H(\theta)=-\sum_j P(y_j)\log T(\theta_j),\ H(\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i)$$

where $I(X;\theta_j)$ is the generalized Kullback-Leibler (KL) information, and $I(X;\theta)$ is the semantic mutual information (a mutual cross-entropy). When $P(x_i|\theta_j)=P(x_i|y_j)$ for all i, j, $I(X;\theta)$ reaches its upper limit: Shannon mutual information $I(X;Y)$. To bring (2.8) into (2.10), we have

$$\begin{aligned} I(X;\theta)&=H(\theta)-H(\theta\mid X)\\ &=-\sum_j P(y_j)\log T(\theta_j)-\sum_j\sum_i P(x_i,y_j)(x_i-x_j)^2/\left(2d_j^{\,2}\right)\end{aligned} \tag{2.11}$$

It is easy to find that the maximum semantic mutual information criterion is a special Regularized Least Squares (RLS) criterion [18]. $H(\theta|X)$ is similar to mean squared error and $H(\theta)$ is similar to negative regularization term.

Assume that a sample is $D=\{(x(t);\ y(t)|t=1,\ 2,\ \ldots,\ N;\ x(t)\in U;\ y(t)\in L\}$, a conditional sample is $D_j=\{x(1), x(2), \ldots, x(N_j)\}$ for given y_j, and the sample points

come from independent and identically distributed random variables. If N_j is big enough, then $P(x_i|y_j)= N_{ij}/N_j$, where N_{ij} is the number of x_i in D_j. Then we have the log normalized likelihood:

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_j I(X;\theta_j) \qquad (2.12)$$

3 Multi-Label Classification for Visibal Instances

3.1 Multi-Label Learning (the Receiver's Logical Classification) for Truth Functions without Parameters

From the viewpoint of semantic communication, the sender's classification and the receiver's logical classification are different. The receiver learns from a sample to obtain labels' denotations, e. g., truth functions or membership functions whereas the sender needs, for a given instance, to select a label with the most information. We may say that the learning is letting a semantic channel match a Shannon's channel and the sender's classification is letting a Shannon's channel match a semantic channel

We use an example to show the two kinds of classifications. Assume that U is a set of different ages. There are subsets of U: A_1={young people}={X|15≤X≤35}, A_2={adults}={X|X≥18}, A_3={juveniles}={X|X<18}=A_2^c (^{c} means complementary set), which form a cover of U. Three truth functions $T(A_1|X)$, $T(A_2|X)$, and $T(A_3|X)$ represent the denotations of y_1, y_2, and y_3 respectively, as shown in Fig. 2.

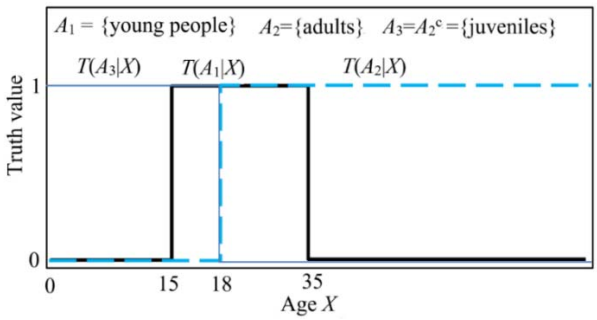

Fig. 1. Three sets form a cover of U, indicating the semantic meanings of y_1, y_2, and y_3.

In this example, $T(A_2)+T(A_3)=1$. If $T(A_1)=0.3$, then the sum of the three logical probabilities is 1.3>1. However, the sum of three statistical probabilities $P(y_1)+P(y_2)+P(y_3)$ must be 1. $P(y_1)$ may change from 0 to 0.3.

Theorem 1 If $P(X)$ and $P(X|y_j)$ come from the same sample D that is big enough so that every possible example appears at least one time, then we can directly obtain the numerical solution of feature function of A_j (as shown in Fig. 2 (a)) according to Bayes' Theorem III and II:

$$T^*(A_j \mid X)=\frac{P(X \mid y_j)}{P(X)} \Big/ \max(\frac{P(X \mid y_j)}{P(X)})=P(y_j|X) / \max(P(y_j \mid X) \tag{3.1}$$

It is easy to prove that changing $P(X)$ and $P(Y)$ does not affect $T^*(A_j|X)$ because $T^*(A_j|X)$ (j=1, 2, …) reflect the property of Shannon's channel or the semantic channel. This formula is also tenable to a fuzzy set θ_j and compatible with Wang's random set falling shadow theory about fuzzy sets [20]. The compatibility will be discussed elsewhere. If $P(X|y_j)$ is from another sample instead of the sample with $P(X)$, then $T^*(A_j|X)$ will not be smooth as shown in Fig. 2 (a) and (b). The larger the size of D is, the smoother the membership function is.

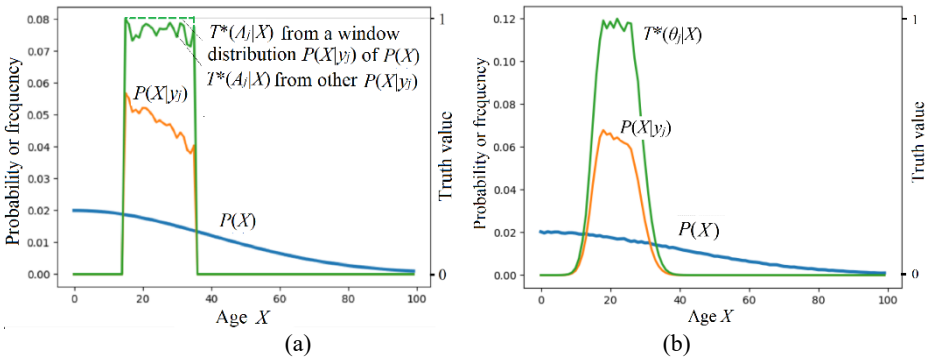

Fig. 2. The numerical solution of the membership function according to (3.1); (a) for a set and (b) for a fuzzy set.

3.2 Selecting Examples for Atomic Labels' Learning

According to mathematical logic, k atomic propositions may produce 2^k independent clauses. The logical add of some of them has $2^{2^{**}k}$ results. So, there are $2^{2^{**}k}$ possible compound labels. To simplify the learning, we may filter examples in a multi-label sample to form a new sample D_a with k atomic labels and k corresponding negative labels. We may use First-Order-Strategy [1] to split examples in D with multi-labels or multi-instances into simple examples, such as, to split (x_1; a_1, a_2) into (x_1; a_1) and (x_1; a_2), and to split (x_1, x_2; a_1) into (x_1; a_1) and (x_2; a_1). Let Y_a denote one of the $2k$ labels, i. e. $Y_a \in \{a_1, a_1', a_2, a_2', \ldots, a_k, a_k'\}$. Consider that some a_j' does not appear in D_a, $|D_a|$ may be less than $2k$. From D_a, we can obtain $P(X, Y_a)$ and corresponding semantic channel $T^*(\theta_a|X)$ or $T^*(\theta_{aj}|X)$ (j=1,2, …, k+k').

3.3 Multi-label Learning for Truth Functions with Parameters

If $P(Y, X)$ is obtained from a not large enough sample, we can optimize the truth function with parameters of every compound label by

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \quad (3.2)$$

It is easy to prove that when $P(X|\theta_j)=P(X|y_j)$, $I(X; \theta_j)$ reaches the maximum and is equal to the KL information $I(X; y_j)$. So, the above formula is compatible with (3.1). Comparing two truth functions, we can find logical implication between two labels. If $T(\theta_j|X) \leq T(\theta_k|X)$ for every X, then y_j implies y_k, and θ_j is the subset of θ_k.

We may learn from the BR method [2] to optimize the truth function of an atomic label with both positive and negative instances by

$$\begin{aligned} T^*(\theta_{aj} \mid X) &= \arg\max_{T(\theta_j|X)} [I(X;\theta_{aj}) + I(X;\theta_{aj}{}^c)] \\ &= \arg\max_{T(\theta_{aj}|X)} \sum_i [P(x_i \mid a_j) \log \frac{T(\theta_{aj} \mid x_i)}{T(\theta_{aj})} + P(x_i \mid a_j{}') \log \frac{1\text{-}T(\theta_{aj} \mid x_i)}{1\text{-}T(\theta_{aj})}] \end{aligned} \quad (3.3)$$

$T^*(\theta_{aj}|x_i)$ is only affected by $P(a_j|X)$ and $P(a_j'|X)$. For a given label, this method divides all examples into three kinds (the positive, the negative, and the unclear) instead two kinds (the positive and the negative) as in One-vs-Rest and BR methods. $T^*(\theta_{aj}|x_i)$ is not affected by unclear instances or $P(X)$. The second part may be 0 because the new method allows that a negative label a_j' does not appear in D or D_a.

In many cases where we use three or more labels rather than two to tag some dimension of instance spaces, the formula (3.2) is still suitable. For example, the truth functions of "Child", "Youth", and "Adult" may be separately optimized by three conditional sampling distributions; "Non-youth" will not be used in general. A number of binary classifications [1, 2] are not necessary.

3.4 Multi-label Selection (the Sender's Selective Classification)

For the visible instance X, the label sender selects y_j^* by the classifier

$$y_j^* = h(x_i) = \arg\max_{y_j} \log I(\theta_j; x_i) = \arg\max_{y_j} \log[T(\theta_j \mid x_i) / T(\theta_j)] \quad (3.4)$$

Using $T(\theta_j)$ can overcome the class-imbalance problem. If $T(\theta_j|X) \in \{0, 1\}$, the information measure becomes Bar-Hillel and Carnap's information measure [19]; the classifier becomes

$$y_j^* = h(x_i) = \arg\max_{y_j \text{ with } T(A_j|x_i)=1} \log[1 / T(A_j)] = \arg\min_{y_j \text{ with } T(A_j|x_i)=1} T(A_j) \quad (3.5)$$

For $X=x_i$, if several labels are correct or approximatively correct, y_j^* will be one of 2^k independent clauses. When k=2, these clauses are $a_1 \wedge a_2$, $a_1 \wedge a_2'$, $a_1' \wedge a_2$, and $a_1' \wedge a_2'$. Therefore, this result is similar to what the BR method provides. When sets are fuzzy, we may use a slightly different fuzzy logic [6] from what Zadeh provides [14] so that a compound label is a Boolean function of some atomic labels. We use

$$T(\theta_1 \cap \theta_2{}^c|X)=\max(0,\quad T(\theta_1|X)-T(\theta_2|X)) \tag{3.6}$$

so that $T(\theta_1 \cap \theta_1{}^c|X)=0$ and $T(\theta_1 \cup \theta_1{}^c|X)=1$. Fig. 3 shows the truth functions of 2^2 independent clauses, which form a partition of plan U*[0,1].

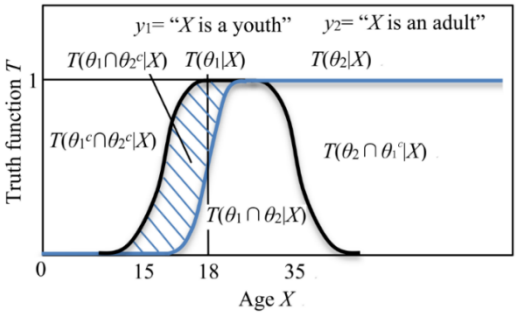

Fig. 3. The truth functions of 2^2 independent clauses

3.5 Classifier $h(X)$ Changes with $P(X)$ to Overcome Class-imbalance Problem

Although optimized truth function $T^*(\theta_j|X)$ does not change with $P(X)$, the classifier $h(X)$ changes with $P(X)$. Assume that y_4="Old person", $T^*(\theta_4|X)=1/[1+\exp(-0.2(X-75))]$, $P(X)=1-1/[1+\exp(-0.15(X-c\ ))]$. The $h(X)$ changes with c as shown in Table 1.

Table 1. The classifier $h(X)$ for y_4="Old person" changes with $P(X)$

c	Population density decreasing ages	Classifier x^* ($y_1=f(X\|X\geq x^*)$)
50	40-60	49
60	50-70	55
70	60-80	58

The dividing point x^* of $h(X)$ increases when old population increases because the semantic information criterion encourages us to reduce the failure of reporting small probability events. Macrobian population's increase will change $h(X)$ and Shannon's channel. Then, the new semantic channel will match new Shannon's channel, and so on. Therefore, the semantic meaning of "Old" should have been evolving with human lifetimes in this way. Meanwhile, the class-imbalance problem is overcome.

4 The CM Iteration Algorithm for the Multi-Label Classification of Unseen Instances

For unseen instances, assume that observed condition is $Z \in C=\{z_1, z_2, \ldots\}$; the classifier is $Y=f(Z)$; a true class or true label is $X \in U=\{x_1, x_2, \ldots\}$; a sample is $D=\{(x(t); z(t))|t=1, 2, \ldots, N; X(t) \in U; z(t) \in C\}$. From D, we can obtain $P(X, Z)$. If D is not big

enough, we may use the likelihood method to obtain $P(X, Z)$ with parameters. The problem is that Shannon's channel is not fixed and also needs optimization. Hence, we treat the unseen instance learning as semi-supersized learning. We can use the Channels' Matching (CM) iteration algorithm [9,10].

Let C_j be a subset of C and $y_j=f(Z|Z\in C_j)$. Hence $S=\{C_1, C_2, \ldots\}$ is a partition of C. Our aim is, for given $P(X, Z)$ from D, to find optimized S, which is

$$S^* = \arg\max_S I(X;\theta|S) = \arg\max_S \sum_j \sum_i P(C_j)P(x_i \mid C_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{4.1}$$

First, we obtain the Shannon channel for given S:

$$P(y_j \mid X) = \sum_{z_k \in C_j} P(z_k \mid X),\ j = 1, 2, ..., n \tag{4.2}$$

From this Shannon's channel, we can obtain the semantic channel $T(\theta|X)$ in numbers or with parameters. For given Z, we have the conditional semantic information

$$I(X_i;\theta_j|Z) = \sum_i P(X_i \mid Z)\log\frac{T(\theta_j \mid X_i)}{T(\theta_j)} \tag{4.3}$$

Then let the Shannon channel match the semantic channel by

$$y_j = f(Z) = \arg\max_{y_j} I(X;\theta_j \mid Z),\quad j=1, 2, \ldots, n \tag{4.4}$$

Repeat (4.2)-(4.4) until S does not change. The convergent S is the S^* we seek. Some iterative examples show that the above algorithm is fast and reliable [10].

5 Summary

This paper provides a ne multi-label learning method: using the third kind of Bayes' theorem (for larger samples) or the generalized Kullback-Leibler formula (for not big enough samples) to obtain the membership functions from sampling distributions, without the special requirement for samples. The multi-label classification is to partition instance space with the maximum semantic information criterion, which is a special regularized least squares criterion and is equivalent to the maximum likelihood criterion. To simplify multi-label learning, we discuss how to use some atomic labels' membership functions to form a compound label's membership function. We also discuss how the classifier changes with the prior distribution of instances and how the class-imbalance problem is overcome for better generalization performance. We treat unseen instance classification as semi-supervised learning and solve it by the Channel Matching (CM) iteration algorithm, which is fast and reliable [9].

From the third kind of Bayes' theorem, we can develop a new Bayesian inference: logical Bayesian inference [21], which needs further study.

References

1. Zhang, M. L., Zhou, Z. H.: A review on multi-label learning algorithm. IEEE Transactions on Knowledge and Data Engineering 26(8), 1819-1837(2014).
2. Zhang, M. L., Li, Y. K., Liu, X. Y., et al.: Binary Relevance for Multi-Label Learning: An Overview, Front. Comput. Sci. 12(2), 191–202(2018).
3. Gold, K., Petrosino, A.: Using information gain to build meaningful decision forests for multilabel classification. Proceedings of the 9th IEEE International Conference on Development and Learning, pp. 58–63. Ann Arbor, MI (2010).
4. Doquire, G., Verleysen, M.: Feature selection for multi-label classification problems, In: Cabestany et al. (Eds,) Lecture Notes in Computer Science 6691, pp. 9–16. Berlin: Springer (2011).
5. Reyes, O., Morell, C., Ventura, S.: Effective active learning strategy for multi-label learning, Neurocomputing 273(17), 494-508 (2018).
6. Lu., C.: B-fuzzy quasi-Boolean algebra and a generalize mutual entropy formula. Fuzzy Systems and Mathematics (in Chinese), 5(1), 76-80 (1991）.
7. Lu, C.: A Generalized Information Theory (in Chinese). China Science and Technology University Press, Hefei (1993).
8. Lu, C.: A generalization of Shannon's information theory. Int. J. of General Systems 28(6), 453-490 (1999).
9. Lu C.: Semantic Channel and Shannon Channel Mutually Match and Iterate for Tests and Estimations with Maximum Mutual Information and Maximum Likelihood. In: 2018 IEEE International Conference on Big Data and Smart Computing, pp. 227-234, IEEE Conference Publishing Services, Piscataway (2018).
10. Lu, C.: Channels' matching algorithm for mixture models. In: Shi et al. (Eds.) IFIP International Federation for Information Processing, pp. 321–332. Springer International Publishing, Switzerland (2017).
11. Anon, Cross entropy, Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Cross_entropy, edited on 13 January 2018.
12. Tarski, *A.:* The semantic conception of truth: *and* the foundations of semantics. Philosophy and Phenomenological Research 4(3)*:* 341–376 (1944).
13. Davidson D.: Truth and meaning. Synthese 17(1): 304-323 (1967).
14. Zadeh, L. A.: Fuzzy Sets. Information and Control 8(3), 338–53 (1965).
15. Bayes T, Price R. An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society of London 53(0), 370–418 (1763).
16. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–429 and 623–656 (1948).
17. Popper K. Conjectures and Refutations. Repr. Routledge, London and New York (1963/2005)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, J.: Generative Adversarial Networks. arXiv*:*1406.2661*[*cs.LG*]* (2014).
19. Bar-Hillel Y, Carnap R.: An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT (1952).
20. Wang, P. Z.: Fuzzy Sets and Random Sets Shadow (in Chinese), Beijing，Beijing Normal University Press（1985).
21. Lu, C.: From Bayesian inference to logical Bayesian inference: A new mathematical frame for semantic communication and machine learning. ICIS2018, Beijing, China (2018).