Mining Classes by Multi-label Classification

Yuichiro KASE ∗, Takao MIURA ∗∗

∗ Dept. of Advanced Sciences, yuichiro.kase.7n@stu.hosei.ac.jp
∗∗ Dept. of Elect. & Elect. Eng., miurat@hosei.ac.jp

HOSEI University
3-7-2 KajinoCho, Koganei, Tokyo, 184–8584 Japan

Résumé. We propose a new approach to mine potential classes in news documents by examining close relationship between new classes and probability vectors of multiple labeling of the documents. Using EM algorithm to obtain the distribution over linear mixture models, we make clustering and mine classes.

1 Motivation

Recently cloud systems through internet have been spread widely so that we can get to huge amount of complex information easily and quickly. However we can hardly catch up with the changes inside and most of the information disappear immediately whatever valuable they are. Very often we like to classify information into classes which come from classes given in advance. A class can be obtained through human recognition by which we can imagine what’s going on by using classes. Since every class corresponds to certain concept, we may see what a word does mean once we know the word belongs to the class.

In this work, we discuss multi-label classification problem and how to find potential classes. Multi-class classification means a process to put information into one of multiple categories. Any information in one category share common aspects which characterize the category given in advance, called a class and its name a label. Automatic classification allows us to extract the rules by inductive learning. We examine a collection of histories (attribute values with labels, called training data) and then extract features specific to classes (Han et Kamber, 2011).

Research of multi-label classification has been initially motivated by the difficulty of concept ambiguity encountered in text categorization. In fact, every document may belong to several themes (labels) simultaneously and few document contains single label. One of the typical approaches is probabilistic classification (Kita, 1995), since the traditional classification results depend heavily on training data. More important is that there is few corpora, although we see huge amount of information with no label (raw data). Here in this work, we take a semi-supervised approach within a framework of probability.

Here we focus our attention on a fact that how classes are constituted. Any news article about international dispute of "Trading Vessels" in China may come from several labels of politics, economy as well as history and culture. Every category carries its own meaning, although it contains weighted combination of labels’ concepts as a one of the features (Han et Kamber, 2011). This means we can define new classes for new categories by giving weight
vectors (0.1, 0.2, 0.3, 0.4) over these labels. To mine new classes by combining labels over the
probability spaces, we could have infinite number of label combinations because of infinite
number of weights.

Main contribution of this work is summarized as follows:
1. We can mine potential classes based on multi-label classification.
2. By means of a linear mixture model, we obtain membership probabilities over given labels.
   Then we make clustering to label probabilities.
3. Our experiments shows that new classes identify new aspects of potential classes which
differ from the constituent labels.

The rest of the paper is organized as follows. In section 2 we describe multi-label clas-
sification for documents and probabilistic approach as well as some related works. Section
3 contains a framework of our approach including EM algorithm and clustering. Section 4
contains some experimental results. In section 5 we conclude this investigation.

2 Multi-label Classification

There have been many multi-label classification algorithms proposed so far. They consist
of two kinds of approach, classification algorithms and probabilistic estimation(Han et Kam-
ber, 2011). The former constitutes classification and label-set combination(Tsoumakas et Ka-
takis, 2007). Classification approach contains clustering, ranking(Elisseeff et Weston, 2002),
entropy (Decision Tree) and translation of binary results into multi-label(Rifkin et Klaatou,
2004). One difficulty arises about dependency among labels as shown in (Zhang et Zhang,
2010).Probabilistic estimation concerns about how to estimate parameters of some probability
distribution functions. Basically we count frequencies from documents and make up classifiers
based on them. Since we estimate a label c which makes \( P(c|x) \) the maximum, we must have
\( P(x|c) \times P(c) \) by Bayes theorem. The simplest one is a Naive Bayesian (NB) classifier. where
both \( P(x|c) \) and \( P(c) \) can be obtained quickly by frequencies. Naive classifiers should depend
on training data and we assume probability model and semi-supervised learning.

Expectation Maximization (EM) algorithm has been discussed for document classifica-
tion based on multinomial probability. During each step, we apply MAP estimation to obtain
new parameters but they examined multi-class classification. Then, for multi-label classifica-
tion, there have been some label-set approach in a probabilistic way. McCallum(McCallum,
1999) has discussed multi-label classification using EM algorithm based on Gauss probability
\( N(\mu, \sigma) \). They have examined the mixtures of normal distributions over all the combination
of labels and estimated the labels of the maximum likelihood. Clearly it takes much times because
of exponential number of the combination.

Ueda(Ueda et Saito, 2003) has proposed a new approach to describe documents by multiple
labels, considering every label as mixtures of topics and every topic as multinomial probability
distribution over words. They have estimated the probability distributions using EM algorithm
and proposed 2 models of label relationship, PMM1 and PMM2 but there still remain some
issues of label dependencies. Using topic model, Wang has proposed some model of inter-
relationship among labels(Wang et al., 2008). Although Latent Dirichlet Allocation approach
can’t model the situation directly, they have introduced label-vectors which can be generated
in a multinomial manner and examined the performance.
3 Estimating Multi-labels

To classify document $d$ for multi-label classification over labels $L = \{L_1, .., L_n\}$, we introduce a probability vector $\vec{d} = (c_1, ..., c_n)$, $\sum c_k = 1, c_j \geq 0$ to describe the probability $c_j$ of $d$ over a label $L_j, j = 1, ..., n$. We like to obtain probability vector $\vec{d}$ over multiple labels over $L$ by means of semi-supervised learning. Since $P(d) = \sum_j P(L_j)P(d|L_j)$ by marginalization, we like to estimate $P_{L_j}(d) = P(d|L_j)$ with a weight $\lambda_{L_j} = P(L_j)$. Note $c_j = P(L_j|d) = P(L_j)P(d|L_j)/P(d)$.

Formally our classification works well by a linear mixture model over labels (Kita, 1995). Let $X$ be a random variable which corresponds to a document and the event probability is generated by a random mixture: $P(X) = \sum\lambda c P_c(X)$ where $P_c(X)$ means a probability of $X$ coming from multinomial probability distribution of a label $c$ and $\lambda$, a probability of choice of $c$ independent of $X$. We assume a word $w$ happens $x_w$ times in a document $d$ of $c$ according to the multinomial probability distribution $P_c(d)$ with a word probability $p_w$ in a naive Bayes manner: $P_c(d) = \frac{n_d!}{\prod_w x_w! \prod_w p_w} \cdot n_d = \sum_{w\in d} x_w$.

To estimate the probabilities $P_c(X)$ and the coefficients $\lambda_c$, we improve several parameters $\theta$ (of probability distribution functions) during EM algorithm until the convergence in such a way that $P_c(X) = p(X|c, \theta_c)$ is a multinomial function with parameters $\theta_c$ (Han et Kamber, 2011). By applying maximum likelihood estimation many times, we get to the stable state because each EM iteration will not decrease the likelihood. Eventually we must have a collection of membership probabilities in a consistent manner. To estimate them at any iteration, we obtain new parameters $\bar{\theta}$ from $\theta$ by maximizing a posterior (MAP) probabilities of multinomial distribution functions.

The story goes with probabilistic process, a linear mixture of multinomial distributions over words depending on labels given in advance. Each document keeps likelihood according to each label, and we assume word arises depending on mixture probabilities. We estimate all the posterior probabilities of the constituent labels $P(d|L_j)$ as well as prior probabilities $P(L_j)$ by means of EM algorithm. For more detail go to (Han et Kamber, 2011).

Let us note that multi-label classification works well with identifying concept correctly in documents and one document may contain several themes over several. This leads us to clustering in label space and we extract clusters of documents according to the probability vectors. That is, every document carries certain probability to every label, which describes distribution of themes it contains to some degree. Considering a set of probabilities as a new aspect, we give a (possibly new) class to the document.

Through EM algorithm, we obtain membership probability $P(d|c)$ of a document $d$ and a label $c$, as well as the choice probabilities $P(c)$ of $c_1, .., c_C$, $\Lambda = (\lambda_1, .., \lambda_C), \sum \lambda_i = 1, \lambda_i \geq 0$. Since any document may belong to several labels at the same time, let us define $\vec{d} = (P(d|c_1), ..., P(d|c_C))$. Note $P(d) = \sum \lambda_i P(d|c_i) = \Lambda \cdot \vec{d}$ holds. Let us define the norm of $\Lambda$ and $\vec{d}$ as $||\Lambda \cdot \vec{d}|| = \Lambda \cdot ||\vec{d}||$. Given a collection of documents $d_1, ..., d_N$, we make clustering all the documents according to probability vectors $\vec{d}_1, ..., \vec{d}_N$ in our label space. Into $K$ exclusive sets in such way that we have the minimum $\sum_j \sum_i ||\Lambda \cdot (\vec{d}_{ij} - t_j)||$, each cluster is labeled with centers $t_1, ..., t_K$. In this investigation $P(d|c)$ is generated by means of multinomial probability distribution over words, all the clusters describe the maximum likelihood of document memberships.
Let us summarize our approach: Assume multinomial probability distribution function to each label. Given labels \( \{c_1, \ldots, c_C\} \), training documents \( L \) and test documents \( T \), we generate probability vectors over \( D = L \cup T \) and make clusters over weighted labels. We preprocess \( D \) in advance such as stopword removal and stemming.

1. By using EM algorithm, we estimate the choice probabilities and the probability vectors, \( (\lambda_1, \ldots, \lambda_C, P(d|c_1), \ldots, P(d|c_C)) \).
2. We make clustering all the documents exclusively according to the choice probabilities and the probability vectors.
3. We put a label \( c \), the center, to each cluster with some cluster-labeling technique.

4 Experiments

Let us show some experimental results in two aspects, multi-label classification and class mining. We examine ModApte subcorpora of Reuter-21578 Version 1.0 by selecting the top 10 frequent labels, and the first 1000 article as test purpose. A table 1 contains the labels. After preprocessing the articles like stopwords and stemming, we replaced each digit number by a special word "*d". We have selected 200 articles randomly as training with 10 times repetition of EM process.

As evaluation measure, we examine precision and recall to multi-label matching (full matching) and to single-label matching (single match). The former means we say correct only if all labels of an article are exactly estimated, while the latter means we say correct if any of the labels are estimated. We apply Naive Bayesian (NB) classification as baseline. We examine 200 articles and extract word frequencies per label as training data. We make binary decision by some threshold. Note the result doesn’t vary very much with several thresholds.

In a table 2, we get 669 articles in total and the precision is 0.669, while NB shows only 8 articles (precision 0.008). In NB, the 8 articles belong to one label "trade" no article matches other labels. A table 3 shows the results of recall and precision to each label obtained by our approach and the baseline (NB) where Ans, Corr, Prec means Answers, Correctness and Precision respectively. We get the averages 0.745 and 0.770 of the recall and the precision respectively by our approach, better than the averages 0.674 and 0.103 (NB). Although the recall values by NB are better than our approach, all the precision values of our approach outperform NB dramatically, say 748% improved. The recall result shows 1.12 times better and the precision 7.24 times better. In NB, we get rather high recall in each label which causes the rather worse precision.

In a table 4, we show all the clusters (centers) and all the number of the articles in each cluster. We get 11 non-empty clusters where (...) means the dominant parts. There are 10 clusters with single dominant label and 1 multi-label cluster ("earn,acq" new class).

Only one class of "earn,acq" arises where multiple labels ("earn" and "acq") are dominant among 11 clusters, but we get the worse precision of full matching. Note we have 76 correct articles among 162 articles assigned to the class "earn,acq". In fact, we get 33 articles of the label "earn", 41 articles of "acq" and 2 articles of "earn,acq". By hands we see the articles of "earn,acq" (containing multiple dominant labels) differ from the articles in the classes "earn" and "acq". These articles of a single label class have an aspect of economic analysis while the articles of the multilabel have a different aspect of financial trends. It seems better to define new class.
5 Conclusion

In this work we have proposed a new approach to mine potential classes. We introduced a linear mixture model of multinomial distribution functions to obtain membership probabilities over labels, then we made clustering to label probabilities. The approach outperforms 1.12 times better in recall and 7.48 times better in precision for single label matching. We obtained a new class which identifies new aspects compared to the constituent labels.

Références


<table>
<thead>
<tr>
<th>No/Articles</th>
<th>earn</th>
<th>acq</th>
<th>money-</th>
<th>crude</th>
<th>grain</th>
<th>trade</th>
<th>interest</th>
<th>wheat</th>
<th>ship</th>
<th>corn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/10</td>
<td>0.233</td>
<td>0.093</td>
<td>0.010</td>
<td>0.015</td>
<td>0.100</td>
<td>0.016</td>
<td>0.017</td>
<td>0.286</td>
<td>0.011</td>
<td>0.197</td>
</tr>
<tr>
<td>2/468</td>
<td>(1.000)</td>
<td>1.90E-05</td>
<td>2.05E-08</td>
<td>6.68E-10</td>
<td>2.78E-09</td>
<td>1.31E-10</td>
<td>1.92E-09</td>
<td>1.99E-09</td>
<td>9.11E-10</td>
<td>4.65E-10</td>
</tr>
<tr>
<td>3/172</td>
<td>3.27E-1 (1.000)</td>
<td>1.07E-12</td>
<td>1.87E-13</td>
<td>2.16E-14</td>
<td>2.02E-15</td>
<td>1.38E-14</td>
<td>2.11E-14</td>
<td>2.25E-14</td>
<td>1.45E-14</td>
<td></td>
</tr>
<tr>
<td>4/162</td>
<td>0.558</td>
<td>(0.205)</td>
<td>0.023</td>
<td>0.033</td>
<td>0.020</td>
<td>0.036</td>
<td>0.037</td>
<td>0.029</td>
<td>0.025</td>
<td>0.033</td>
</tr>
<tr>
<td>5/24</td>
<td>1.84E-30</td>
<td>2.99E-47</td>
<td>1.35E-48</td>
<td>1.95E-48</td>
<td>1.27E-34</td>
<td>2.10E-48</td>
<td>2.47E-48</td>
<td>(1.0)</td>
<td>1.47E-48</td>
<td>4.97E-49</td>
</tr>
<tr>
<td>6/0</td>
<td>0.258</td>
<td>0.428</td>
<td>0.643</td>
<td>9.48E-10</td>
<td>5.75E-10</td>
<td>1.02E-09</td>
<td>1.05E-09</td>
<td>8.12E-09</td>
<td>1</td>
<td>9.48E-10</td>
</tr>
<tr>
<td>7/0</td>
<td>0.313</td>
<td>0.115</td>
<td>0.013</td>
<td>0.019</td>
<td>0.011</td>
<td>0.435</td>
<td>0.021</td>
<td>0.016</td>
<td>0.038</td>
<td>0.019</td>
</tr>
<tr>
<td>8/0</td>
<td>0.215</td>
<td>0.079</td>
<td>0.009</td>
<td>0.013</td>
<td>0.008</td>
<td>0.014</td>
<td>0.011</td>
<td>0.011</td>
<td>0.625</td>
<td>0.013</td>
</tr>
<tr>
<td>9/0</td>
<td>0.088</td>
<td>0.085</td>
<td>0.004</td>
<td>0.005</td>
<td>0.056</td>
<td>0.006</td>
<td>0.110</td>
<td>0.004</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>10/21</td>
<td>7.54E-10</td>
<td>6.0</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>11/0</td>
<td>0.129</td>
<td>0.047</td>
<td>0.005</td>
<td>0.008</td>
<td>0.235</td>
<td>0.008</td>
<td>0.314</td>
<td>0.006</td>
<td>0.238</td>
<td></td>
</tr>
<tr>
<td>12/0</td>
<td>0.201</td>
<td>0.122</td>
<td>0.007</td>
<td>0.465</td>
<td>0.043</td>
<td>0.048</td>
<td>0.048</td>
<td>0.009</td>
<td>0.008</td>
<td>0.047</td>
</tr>
<tr>
<td>13/11</td>
<td>4.56E-11</td>
<td>1.11E-12</td>
<td>2.45E-13</td>
<td>7.16E-16</td>
<td>1.67E-14</td>
<td>4.48E-14</td>
<td>(1.000)</td>
<td>1.37E-14</td>
<td>8.37E-14</td>
<td>7.40E-14</td>
</tr>
<tr>
<td>14/19</td>
<td>4.71E-12</td>
<td>1.73E-13</td>
<td>(1.0)</td>
<td>2.82E-13</td>
<td>1.71E-13</td>
<td>3.02E-13</td>
<td>3.12E-13</td>
<td>2.41E-13</td>
<td>2.11E-13</td>
<td>2.82E-13</td>
</tr>
<tr>
<td>15/0</td>
<td>0.168</td>
<td>0.112</td>
<td>0.207</td>
<td>0.010</td>
<td>0.006</td>
<td>0.261</td>
<td>0.161</td>
<td>0.009</td>
<td>0.008</td>
<td>0.060</td>
</tr>
<tr>
<td>16/0</td>
<td>1.58E-13</td>
<td>5.82E-15</td>
<td>6.43E-16</td>
<td>9.48E-17</td>
<td>5.75E-17</td>
<td>5.02E-17</td>
<td>3.42E-17</td>
<td>9.78E-17</td>
<td>9.48E-17</td>
<td></td>
</tr>
<tr>
<td>17/28</td>
<td>5.07E-11</td>
<td>1.20E-11</td>
<td>1.10E-12</td>
<td>(1.0)</td>
<td>1.22E-12</td>
<td>6.38E-12</td>
<td>1.35E-12</td>
<td>1.40E-12</td>
<td>5.19E-12</td>
<td>1.87E-12</td>
</tr>
<tr>
<td>18/17</td>
<td>2.58E-11</td>
<td>9.48E-12</td>
<td>1.05E-13</td>
<td>1.54E-13</td>
<td>(1.0)</td>
<td>1.65E-13</td>
<td>1.71E-13</td>
<td>1.32E-13</td>
<td>1.16E-13</td>
<td>6.65E-13</td>
</tr>
<tr>
<td>19/28</td>
<td>1.93E-11</td>
<td>7.10E-13</td>
<td>7.84E-14</td>
<td>1.16E-14</td>
<td>1.55E-14</td>
<td>1.24E-14</td>
<td>1.28E-14</td>
<td>1.21E-14</td>
<td>1.16E-14</td>
<td>6.65E-14</td>
</tr>
<tr>
<td>20/30</td>
<td>4.85E-11</td>
<td>1.78E-11</td>
<td>1.97E-12</td>
<td>2.90E-12</td>
<td>1.76E-12</td>
<td>(1.0)</td>
<td>3.21E-12</td>
<td>2.49E-12</td>
<td>2.18E-12</td>
<td>2.90E-12</td>
</tr>
</tbody>
</table>


### Summary

By examining close relationship between new classes and probability vectors of multiple labeling of the documents, we obtain probability distribution function to each label of documents. With the assumption of multinomial distribution over words, we apply EM algorithm to obtain the distribution. Then we apply clustering to label probabilities to mine classes.