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 muli-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 classification 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 Kamber, 2011). The former constitutes classification and label-set combination(Tsoumakas et Katakis, 2007). Classification approach contains clustering, ranking(Elisseeff et Weston, 2002), entropy (Decision Tree) and translation of binary results into multi-label(Rifkin et Klautau, 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 classification 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 classification, 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(μ, σ). 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 : $P(X) = \sum_c \lambda_c P_c(X)$ where $P_c(X)$ means a probability of X coming from *multinomial* probability distribution of a label c and λ_c 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 : $P_c(d) = \dfrac{n_d!}{\Pi_w x_w!}\Pi_w p_w^{x_w}, n_d = \sum_{w \in d} x_w$.

To estimate the probabilities $P_c(X)$ and the coefficients λ_c, we improve several parameters θ (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 θ_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 θ 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 Λ and $\vec{d}$ as $||\Lambda \cdot \vec{d}|| = \Lambda \cdot \vec{d}/|\Lambda \cdot \vec{d}|$.

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 $\sum_j^K \sum_i ||\Lambda \cdot (\vec{d_{ij}} - t_j)||$, each cluster is labeled with centers $t_1, .., t_K$. In this investigation $P(\vec{d}|c)$ 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 $\{c_1, ..., c_C\}$, training documents $\mathcal{L}$ and test documents $\mathcal{T}$, we generate probability vectors over $D = \mathcal{L} \cup \mathcal{T}$ 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, $(\lambda_1, .., \lambda_C$ and $P(d|c_1), ..., P(d|c_C))$.
(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 repeatition 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, Correcness 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.

Articles	Label	Articles	Label
507	earn	13	interest
226	acq	11	grain, wheat, corn
55	crude	5	earn,acq
41	trade	5	crude, ship
34	grain,wheat	2	earn, crude
29	money-fx,interest	2	grain, wheat, ship
20	money-fx	2	money-fx, trade
19	grain,corn	1	acq, ship
13	ship	1	grain, ship
13	grain	1	wheat, corn
1000			

TAB. 1 – *Labels in Articles*

Label	Articles	Matched	Precision
(Ours)			
earn	468	453	0.968
acq	172	153	0.890
earn/acq	162	2	0.012
trade	30	17	0.567
ship	21	8	0.381
grain	17	3	0.176
crude	28	25	0.893
interest	31	4	0.129
money-fx	19	4	0.211
wheat	24	0	0.0
corn	28	0	0.0
(total)	1000	669	0.669
(NB)			
trade	30	8	0.267
(total)	1000	8	0.008

TAB. 2 – *Multilabels Fully-Matched*

Label	Ans	Corr	Recall	Ans	Corr	Prec	Ans	Corr	Recall	Ans	Corr	Prec
(Ours)							(NB)					
acq	232	198	0.853	334	232	0.695	232	165	0.711	708	165	0.233
corn	31	15	0.484	28	15	0.536	31	21	0.677	750	21	0.028
crude	62	26	0.419	28	26	0.929	62	43	0.694	821	43	0.052
earn	514	508	0.988	630	530	0.841	514	350	0.681	466	350	0.751
grain	80	12	0.15	17	12	0.706	80	50	0.625	760	50	0.066
interest	42	22	0.523	31	22	0.710	42	28	0.667	759	28	0.037
money-fx	51	7	0.137	19	7	0.368	51	33	0.647	772	33	0.043
ship	22	16	0.727	21	16	0.762	22	14	0.636	713	14	0.0196
trade	43	18	0.419	30	18	0.6	43	26	0.605	866	26	0.030
wheat	48	16	0.333	24	17	0.708	48	28	0.583	760	28	0.037
(total)	1125	838	0.745	1162	895	0.770	1125	758	0.674	7375	758	0.103

TAB. 3 – *Recall/Precision per Label*

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

Elisseeff, A. et J. Weston (2002). A kernel method for multi-labeled classification. *Advances NIPS 14*, 681–687.

Han, J. et M. Kamber (2011). Data mining: Concepts and techniques. *Morgan Kauffman*.

Kita, K. (1995). Probabilistic language models (in japanese). *University of Tokyo Press*.

McCallum, A. (1999). Multi-label text classification with a mixture model trained by em. *Workshop on Text Learning, AAAI*.

No/Articles	earn	acq	money-fx	crude	grain	trade	interest	wheat	ship	corn
1/0	0.253	0.093	0.010	0.015	0.100	0.016	0.017	0.286	0.011	0.197
2/468	(1.000)	1.90E-05	2.05E-08	6.68E-10	2.78E-09	1.31E-10	1.92E-07	1.99E-10	9.11E-10	4.65E-09
3/172	3.27E-08	(1.000)	1.07E-12	1.87E-13	2.16E-15	2.02E-14	1.38E-10	2.11E-13	2.25E-11	1.45E-14
4/162	(0.558)	(0.205)	0.023	0.033	0.020	0.036	0.037	0.029	0.025	0.033
5/24	1.84E-30	2.99E-47	1.35E-48	1.95E-48	1.27E-34	2.10E-48	2.47E-46	(1.0)	1.47E-48	4.97E-12
6/0	1.58E-58	5.82E-59	6.43E-60	9.48E-60	5.75E-60	1.02E-59	1.05E-59	8.12E-60	1	9.48E-60
7/0	0.313	0.115	0.013	0.019	0.011	0.435	0.021	0.016	0.038	0.019
8/0	0.215	0.079	0.009	0.013	0.008	0.014	0.014	0.011	0.625	0.013
9/0	0.088	0.085	0.004	0.005	0.056	0.006	0.006	0.110	0.004	0.637
10/21	7.54E-60	2.77E-60	3.06E-61	4.51E-61	2.74E-61	4.83E-61	5.00E-61	3.87E-61	(1.0)	4.51E-61
11/0	0.129	0.047	0.005	0.008	0.235	0.008	0.009	0.314	0.006	0.238
12/0	0.201	0.122	0.007	0.465	0.043	0.048	0.048	0.009	0.008	0.047
13/31	4.56E-13	1.11E-14	2.45E-16	7.16E-10	1.67E-12	4.48E-17	(1.000)	1.37E-08	8.37E-16	7.40E-16
14/19	4.71E-50	1.73E-50	(1.0)	2.82E-51	1.71E-51	3.02E-51	3.12E-51	2.41E-51	2.11E-51	2.82E-51
15/0	0.168	0.112	0.207	0.010	0.006	0.261	0.161	0.009	0.008	0.060
16/0	1.58E-58	5.82E-59	6.43E-60	9.48E-60	5.75E-60	1.02E-59	1.05E-59	8.12E-60	1	9.48E-60
17/28	5.07E-11	1.20E-36	1.10E-39	(1.0)	1.22E-36	6.38E-34	1.35E-38	1.40E-36	5.19E-35	1.87E-37
18/17	2.58E-59	9.48E-60	1.05E-60	1.54E-60	(1.0)	1.65E-60	1.71E-60	1.32E-60	1.16E-60	6.65E-27
19/28	1.93E-46	7.10E-47	7.84E-48	1.16E-47	1.55E-23	1.24E-47	1.28E-47	1.21E-21	8.66E-48	(1.0)
20/30	4.85E-41	1.78E-41	1.97E-42	2.90E-42	1.76E-42	(1.0)	3.21E-42	2.49E-42	2.18E-42	2.90E-42

TAB. 4 – *Cluster Constituents*

Rifkin, R. et A. Klautau (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research 5*, 101–141.

Tsoumakas, G. et I. Katakis (2007). Multi-label classification: An overview. *J. of Data Warehouse and Mining 3-3*.

Ueda, N. et K. Saito (2003). Parametric mixture models for multi-label text. *Advances in NIPS 15*, 721–728.

Wang, H., M. H. X., et Zhu (2008). A generative probabilistic model for multi-label classification. *Intn'l Conf. on Data Mining (ICDM)*, 628–637.

Zhang, M. et K. Zhang (2010). Multi-label learning by exploiting label dependency. *Knowledge Discovery in Databases (KDD)*.

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.