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