Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

K-Nearest Neighbor with EM Algorithms for Multi-label Classification

K. Gayathri, Dr. A. Marimuthu

Abstract: Labeling text data is quite time consuming but essential for automatic text classification. Especially manually creating multiple labels for each document may become impractical when a very large amount of data is needed for training multi-label classifiers. Text categorization is the process of sorting text documents into one or more pre-defined categories or classes of similar documents. In this paper we are interested in examining whether automatic classification of news texts can be improved by a pre-filtering the vocabulary to reduce the feature set used in the computations.

Keywords: Multi-label, Text Categorization, TF/IDF, Knn, EM algorithms.

I. Introduction:

As Text data becomes a major information source in our daily life, many research efforts have been conducted in text classification to better organize text data in applications like document filtering email classification web search, etc. in particular multi-label text classification problems have received considerable attention, since many text classification tasks are multi-labeled (i.e) each document can belong to more than one category. Take news classification as an example one news article talking about the effect to Olympic games on tourism industry might belong to the following topic categories: sports, economy and travel. In the literature supervised learning algorithms are widely used in text classification. It requires a sufficient amount of labeled data for training a high quality model. However, labeling is usually a time consuming and expensive is an approach to reduce the labeling cost. The active learner iteratively selects a sample of data to be labeled based on some selection strategies suggesting that the data most deserves to be labeled. Thus it can achieve comparable performance with supervised learners while using much less labeled data. Active learning is particular important for the multi-label text classification task. The reason is that, in the single label case, a human judge can stop labeling an instance once its category is identified. But in the multi-label case, human judges need to decide all possible categories for each instance. Thus the effort of assigning labels for multi-label data is much larger than for the single-label data [1]. Automatic text categorization is one particular tool to retrieve and make use of the text information efficiently. There are several applications where text categorization plays an important role like technical professional business and web based areas. Also the classification is considered to be an important research field used to identify the data and classify it based on several theoretical approaches. Using automatic text categorization the stories can be categorized based on subject categories, academic papers are often classified by technical domains and sub-domains. Automatic text categorization where it needs more number of people to manually label or categorize the data. several methods can be implemented for categorizing the text that varies in their accuracy and computation efficiency.[2]

Feature selection can reduce the dimensionality of feature space, decrease the computing complexity and improve the accuracy rate of classification. A number of feature selection methods has been applied to text classification including document frequency information gain mutual information etc. among the different machine learning techniques we found that clustering can interpret the mulit-label of data in a more meaningful way. In fact, we found that the notion of clustering matches that of text data [3]

K-Nearest neighbor is used broadly in text classification. It is one of the most popular algorithms for text categorization. To classify a new document the system finds the k-nearest neighbors among the training documents and uses the categories of the K-nearest neighbors to weight the category candidates [4].

The EM algorithm is a general method of finding the maximum likelihood estimate of the parameters of an distribution from a given data set when the data is incomplete or has missing values. There are two main application of the EM algorithm. The first occurs when the data indeed has missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing parameters. The latter application is more common in the computational pattern recognition community.

II. Document Representation:

The document representation is the preprocessing process that is used to reduce the complexity of the documents and make them easier to handle which needles to be transformed from the full text version to a document vector.

A. Feature Extraction

The process of feature extraction is to make clear the border of each language depends factors, tokenization, stop words removal and stemming.

B. Feature selection

After feature extraction the important step in preprocessing of text classification, is feature selection to construct vector space or bag of words. Which improve the scalability, efficiency and accuracy of a text classifier. The main idea of FS is to select subset of feature from the original documents. Hence feature selection is commonly used in text classification to reduced the dimensionality of feature space and improve the efficiency and accuracy of classifiers. The Term Word Frequency/Inverse Document Frequency (TF/IDF) approach is commonly used to weight each word in the text document according to how unique it is in other word the TF/IDF approach captures the relevancy among words text document and particular categories [5].

III. K- Nearest Neighbor algorithms:

The K-Nearest Neighbor algorithm (K-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its K nearest neighbors. K is a positive integer, typically small. If K=1, then the object is simply assigned to the class of its nearest neighbor. In binary classification problems, it is helpful to choose K to be an odd number as this avoids tied votes. K-NN classifier is an instance-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or cosine similarity measure. The cosine similarity is commonly used in information retrieval and we used as the basic similarity measure in our algorithms.[6].

IV. EM Algorithm

E-step (Expectation): the step that compute the probabilities that each object belongs to clusters. This corresponds to their likelihoods based on the current parameters characterizing the clusters.

M-step (Maximization): the step that updates the parameters of the clusters by replacing the current parameters by new parameters based on the probabilities computed from the E-step.

In the E-step parameters characterizing clusters should be initialized. The number of clusters should be determined initially as an input parameter and initial clusters are built at random. Then, mean vectors and covariance matrices can be easily computed as initial parameters for entering the E-step from the numerical vectors in cluster.

In the M-step where the new parameters of the clusters are computed with respect to the maximum likelihood of objects with respect to the clusters. [7]

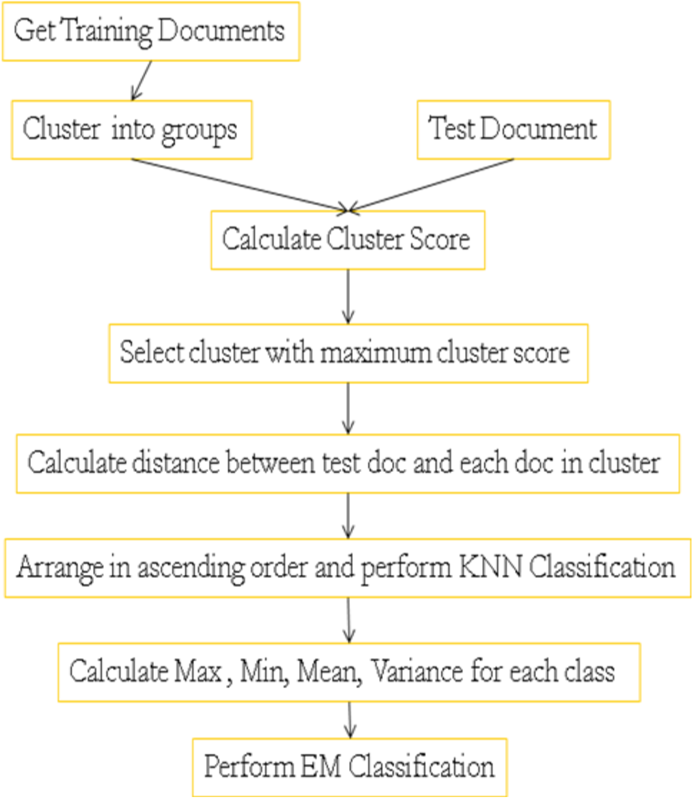
Get Training Documents
Cluster into groups
Test Document
Calculate Cluster Score
Select cluster with maximum cluster score
Calculate distance between test doc and each doc in cluster
Arrange in ascending order and perform KNN Classification
Calculate Max , Min, Mean, Variance for each class
Perform EM Classification

Figure 1

V. Experiments in Text categorization

5.1 News Page.Com

This test bed consists of 1,200 news articles in the format of plain texts built by copying and pasting news articles in the web site, www.newspage.com . in the table 1 shows the predefined categories, the number of documents of each category and the partition of the test bed into training set ans test set. As shown in table 1, the ratio of training set to test set is set as 7:3 Here, test bed is called Newspage.com based on the web site, given as its source.

Table-1 Training set and Test set of Newspage.com

Category Name	Training Set	Test Set	#Document
Business	280	120	400
Health	140	60	200
Law	70	30	100
Internet	210	90	300
Sports	140	60	200
Total	840	360	1200

The task of text categorization on this test bed is decomposed into five binary classification problems, category by category. In each binary classification problem, a classifier answers whether an unseen document belongs to its corresponding category, or not Table 2 shows the definition of training sets of the predefined categories. In table 2 “positive” indicates that documents belong to the corresponding category and such document will called positive documents, while ”negative” indicates that documents do not and such documents will be called negative documents. For each training set, all of documents not belongs to its corresponding category are allocated as negative documents. For each test set, negative documents are allocated as many as positive documents define in the third column of table 1.

Table-2 The Allocation of Positive and Negative Class in Training Set of each Category:

Category Name	Positive	Negative	Total
Business	280	560	840
Health	140	700	840
Law	70	770	840
Internet	210	630	480
Sports	140	700	840

VI. CONCLUSION

Analyzed the text classification using the KNN & EM with the Feature selection techniques. . The advantage of the proposed approach is, the classification algorithm learns importance of attributes and utilizes them in the similarity measure. In future the classification model can be build, which analyzes terms on the concept sentence in document.

REFERENCE:

1. Bishan yang, jian-Tao Sun, Tengjiao Wang, Zhergcher “Effective Multi-label active learning for text classification” Ministry of Education, china.
2. Suneetha Manne, Sita Kumari Kotha, Dr.S.Sameen Fatima “A Query based Text Categorization using K- Nearest Neighbor Approach”. Internaional Journal of computer application.
3. Mohammad salim Ahmed, Latifurkhan, mandava Rajeswari “Using correlation Based Subspace Clustering for Multi-lable Text Data Classification” 2010, 22nd international conference on Tools with Artifical Intelligence. IEEEm computer Society.
4. Chuanyao Yang, Yuqin Li, Chenghang Zhang, YufaHu, “ A Fast Knn Algorithm Based on simulated Annealing”
5. Aurangzeb khan, Baharum B.Bahurdin, Khairullah khan “An Overview of E-Documents classification”.2009, International Conference on machine learning and computing Vol.3(2011)IACSIT, Singapore
6. Muhammed Miah “Improved K-NN algorithm for text classification”.
7. Taeho Jo “The implementation of Dynamic Document Organization using the Integration of Text Clustering And Text Categorization.