Categorization and Keyword Identification of Unlabeled Documents

Ning Kang Carlotta Domeniconi Daniel Barbará
ISE Department George Mason University
nkang@gmu.edu *carlotta@ise.gmu.edu* *dbarbara@gmu.edu*

Abstract

In this paper we first propose a global unsupervised feature selection approach for text, based on frequent itemset mining. As a result, each document is represented as a set of words that co-occur frequently in the given corpus of documents. We then introduce a locally adaptive clustering algorithm, designed to estimate (local) word relevance and, simultaneously, to group the documents. We present experimental results to demonstrate the feasibility of our approach. Furthermore, the analysis of the weights credited to terms provides evidence that the identified keywords can guide the process of label assignment to clusters. We take into consideration both spam email filtering and general classification datasets. Our analysis of the distribution of weights in the two cases provides insights on how the spam problem distinguishes from the general classification case.

1 Introduction

The most commonly used representation for documents is the so called Vector Space Model (VSM), or Bag of Words (BOWs). Such a word level representation of documents easily leads to a 30000 or more dimensions. In this high dimensionality, the effectiveness of any distance function that equally uses all input features is severely compromised. Furthermore, one would expect that different words might have different degrees of relevance for a given category of documents, and, at the same time, a single word might have a different importance across different categories. In addition, each word in a selected dictionary might be relevant for at least one of the categories. Thus, it may not always be feasible to prune off too many dimensions without incurring a loss of crucial information. A proper feature selection procedure should operate locally in input space.

In this paper we first propose a global unsupervised feature selection approach for text, based on frequent itemset mining. As a result, each document is represented as a *bag of frequent items*, that is a set of words that co-occur frequently in the given corpus of documents (each selected word, or item, corresponds to a feature). This step is applied initially to documents to reduce the number of features to a feasible dimensionality for clustering and local weighting of keywords. We then introduce a locally adaptive clustering algorithm, designed to estimate (local) word relevance and, simultaneously, to group the documents. Thus, this method achieves not only a clustering of the documents, but also the identification of cluster-dependent keywords. The analysis of such keywords allows to assign labels to clusters, and therefore to use the groups as a model for prediction.

The contributions of this paper are as follows: (1) We introduce a new unsupervised feature (word) selection approach to handle multi-class classification of documents in absence of labels. The approach is based on the mining of frequent itemsets. (2) We apply a locally adaptive clustering algorithm for documents. The output of our method is twofold: it achieves not only a clustering of the documents, but also the identification of cluster-dependent keywords via a continuous term-weighting mechanism. (3) The experimental results we present demonstrate the feasibility of our approach in terms of achieved accuracy measured against the ground truth. Furthermore, the analysis of the weights credited to terms provide evidence that the identified keywords can guide the process of label assignment to clusters. Thus, the resulting groups can be used as a model for prediction.

2 Related Work

Local dimensionality reduction approaches for the purpose of efficiently indexing high dimensional spaces have been recently discussed in the database literature [7, 4, 9]. In general, the efficacy of these methods depends on how the clustering problem is addressed in the first place in the original feature space.

The problem of finding different clusters in different subspaces of the original input space has been addressed in [2, 8, 1]. In [2, 8], the authors use a density based approach to identify clusters. The algorithm (PROjected CLUStering) proposed in [1] seeks subsets of dimensions such that the points are closely clustered in the corresponding spanned subspaces. Both the number of clusters and the average number of dimensions per cluster are user-defined parameters. In contrast to the PROCLUS algorithm, our method (LAC) does not require to specify the average number of dimensions to be kept per cluster. For each cluster, in fact, *all* features are taken into consideration, but properly weighted. The PROCLUS algorithm is more prone to loss of information if the number of dimensions is not properly chosen.

3 Feature Selection Based on Frequent Itemsets Mining

In [3] we introduced a feature selection algorithm for text, based on frequent itemsets mining. Our method (DocMine) addresses the categorization of documents (without labels) with an unknown number of classes, with the user interested in only one of them.

The method presented in [3] requires multiple sets of documents to be available (e.g., collections of documents retrieved by several search engines as result of a given query), and makes the assumption that relevant documents are more frequent in the majority of the sets. By computing the itemsets (words) that are frequent in the majority of the collections, it identifies positive features. The documents that contain the identified words are labeled as positive documents.

In this work we extend our unsupervised feature selection approach to handle multi-class classification problems in absence of labels. We no longer require the existence of multiple sets of documents.

Given a document, it is possible to associate with it a *bag of words* [6]. Specifically, we represent a document as a binary vector $\mathbf{d} \in \Re^N$, in which each entry records if a particular word stem occurs in the text. The dimensionality N of $\mathbf{d}$ is determined by the number of different terms in the corpus of documents (size of the *dictionary*), and each entry is indexed by a specific term.

Given a sample of unlabeled documents $\{\mathbf{d}_i\}$ of different categories, we mine them to find the frequent itemsets that satisfy a given support level. In principle, the support level is driven by the target dimensionality of the data (to make the subsequent clustering step suitable). Each resulting itemset is a set of words that co-occur *frequently* in the given corpus of documents. We consider the union of such frequent items, and represent each document as a *bag of frequent items*. The actual value of the entry is the frequency of the corresponding word in the document. This provides a suitable representation since it is *compact* (the level of compactness being driven by the support), and captures keywords that co-occur frequently within each category. We observe that additional spurious (non discriminant) features may be selected by this process (e.g., words that are frequent in documents across classes). The subsequent locally adaptive clustering algorithm is designed to estimate word relevance and, simultaneously, to group the documents. Thus, it achieves not only a clustering of the documents, but also the identification of cluster-dependent keywords. The analysis of such keywords can assist the assignment of labels to clusters, and therefore the use of groups as a model for prediction.

4 Locally Adaptive Clustering

Here we briefly describe our locally adaptive clustering algorithm [5]. Consider a set of points in some space of dimensionality n. A *weighted cluster* C is a subset of data points, together with a vector of weights $\mathbf{w} = (w_1, \ldots, w_n)$, such that the points in C are closely clustered according to the L_2 norm distance weighted using $\mathbf{w}$. The component w_j measures the degree of correlation of points in C along feature j. The problem becomes now how to estimate the weight vector $\mathbf{w}$ for each cluster in the data set.

Our approach progressively improves the quality of initial centroids and weights, by investigating the space near the centers to estimate the dimensions that matter the most. We start with *well-scattered* points in a dataset S as the k centroids: we choose the first centroid at random, and select the others so that they are far from one another, and from the first chosen center. We initially set all weights to $1/n$. Given the initial centroids $\mathbf{c}_j$, for $j = 1, \ldots, k$, we compute the corresponding sets $S_j = \{\mathbf{x} | (\sum_{i=1}^n w_{ji}(x_i - c_{ji})^2)^{1/2} < (\sum_{i=1}^n w_{li}(x_i - c_{li})^2)^{1/2}, \forall l \neq j\}$, where w_{ji} and c_{ji} represent the ith components of vectors $\mathbf{w}_j$ and $\mathbf{c}_j$ respectively (ties are broken randomly). We then compute the average distance along each dimension from the points in S_j to $\mathbf{c}_j$: $X_{ji} = \frac{1}{|S_j|} \sum_{\mathbf{x} \in S_j} (c_{ji} - x_i)^2$, where $|S_j|$ is the cardinality of set S_j. The smaller X_{ji} is, the larger is the correlation of points along dimension i. We use the value X_{ji} in an exponential weighting scheme to credit weights to features (and to clusters), as given in $w_{ji}^* = \frac{exp(-X_{ji}/h)}{\sum_{i=1}^n exp(-X_{ji}/h)}$ (the parameter h controls the strength of the incentive for clustering on more features). The computed weights are used to update the sets S_j, and therefore the centroids' coordinates as given in $c_{ji}^* = \frac{1}{|S_j|} \sum_{\mathbf{x} \in S_j} x_i$. The procedure is iterated until convergence is reached. The resulting algorithm is called LAC.

4.1 Subspace Clustering for Text

Our overall approach consists of the following steps: (1) Preprocess the documents by eliminating stop and rare words, and by stemming words to their root source; (2) Apply our global unsupervised feature selection approach based on frequent itemset mining. As a result, we obtain documents represented as bag of frequent items; (3) Apply our locally adaptive clustering algorithm to estimate local word relevance and, simultaneously, to group documents. As a result, we obtain a clustering of the documents, and the identification of cluster-dependent keywords.

5 Experimental Evaluation

In our experiments we used several datasets. Due to lack of space, below we report the results obtained on one spam email problem and on one general classification case. **Email-1431**. This email dataset consists of 1431 emails, falling into three categories: conference (370), jobs (272), and spam (789). The original size of the dictionary is 38713. We consider a 2-class classification problem by merging the conference and jobs mails into one group (non-spam). **20 Newsgroups**. This dataset is a collection of 20,000 messages collected from 20 different netnews newsgroups. One thousand messages from each of the twenty newsgroups were chosen at random and partitioned by newsgroup name. In our experiments we consider the two categories Medical (990) and Electronics (981) (the original size of the dictionary in this case is 22820). The documents in each dataset were preprocessed by eliminating stop words (based on a stop words list), and stemming words to their root source. In addition, rare words that appeared in less than four documents were also removed. After the initial global feature selection step, we use as feature values for the vector space model the relative frequency of the selected words (frequent itemsets) in the corresponding document.

Tables 1-2 report the results. Each table includes: the support values tested (S), the dimensionality of the data after the preprocessing step (N), the dimensionality of the data after feature selection based on frequent itemset mining (n), the total number of documents (D) (as well as the number of documents per class), the average error rate computed over nine runs of LAC for $1/h = 1, \dots, 9$ (along with the standard deviations), the minimum error rate over such nine runs, and (as baseline comparison) the error rate of K-means. Error rates are computed according to the confusion matrices based on the ground truth labels.

Table 1. Results for Email-1431 (Spam (789) - Non Spam (642))

S	N	n	D	Ave Err	Min Err	K-means
5%	9210	791	1431	2.0 ± 0.3	1.7	45.0
7%	9210	519	1431	2.0 ± 0.5	1.3	45.0
10%	9210	285	1431	2.0 ± 0.4	1.5	45.0

Table 2. Results for NewsGroups (Electronic (981) - Medical (990))

S	N	n	D	Ave Err	Min Err	K-means
1%	6217	1359	1971	11.5 ± 2.4	9.5	49.6
2%	6217	583	1971	18.1 ± 11.8	13.5	49.7
3%	6217	321	1971	21.0 ± 9.5	16.8	49.6
4%	6217	201	1971	21.8 ± 0.4	20.8	49.7
5%	6217	134	1971	29.1 ± 7.5	23.3	49.6

For increasing support values, and therefore decreasing number of selected features, we can observe an increasing trend for the minimum error rates. In general, lower error rates were achieved for larger h values, which favor multi-dimensional clusters. As expected, the optimal dimensionality depends on the dataset. Particularly low error rates are achieved for the problem on spam emails, and for a wide range of dimensionalities. K-means often fails to detect any structure in the data, and provides error rates close or above 45%.

The analysis of the weights credited to words provides some insights on the nature of the spam email filtering problem and the general classification case. As Figures 1-2 show, the selected keywords (and in particular those that receive largest weight values) are representative of the underlying categories, which provides evidence that our global feature selection method successfully retains discriminant words. In addition, our subspace clustering technique is capable of further sifting the most relevant ones, while discarding the additional spurious words.

Let us consider the distribution of weights obtained for the Email-1431 dataset. Figure 1 shows the weight values and corresponding keywords for the two class case (the non-spam class corresponds to both conference and jobs emails). Here we plot the top words that received highest weight for each class. We observe that words reflecting the topic of a category receive a larger weight in the *other class*. For example, the words "*free*", "*money*", "*sales*", "*marketing*", and "*order*" get a larger weight in the non-spam class (their weights in the spam category are very close to zero, which cause the corresponding bar not to show up in the plot). Similarly, the words "*conference*", "*applications*", "*papers*", "*science*", "*committee*", "*institute*", "*neuroscience*", etc receive larger weights within the spam category. The weights for these words in the non-spam class are very close to zero. While surprising at first, this trend may be due to the nature of the spam and non-spam email dis-

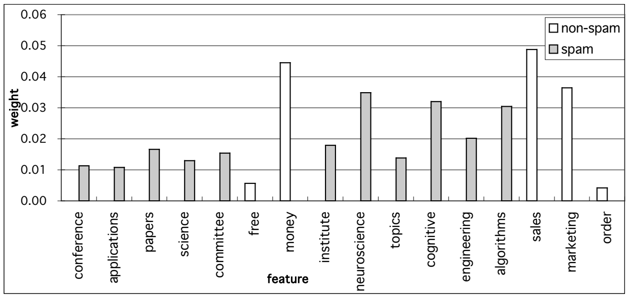

Figure 1. Email-1431: Keywords and corresponding weight values ($s = 10\%, h = 1/9$).

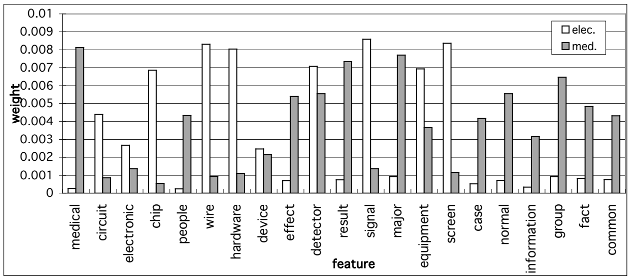

Figure 2. Newsgroup dataset (electronics-medical): Keywords and corresponding weight values ($s = 3\%, h = 1/9$).

tributions. Each of these two categories is actually a combination of subclasses. The non-spam class in this case is the union of conference and jobs emails (by construction). Likewise, the spam messages can be very different in nature, and therefore different in their word content. As a consequence, the dispersion of feature values for words reflecting the general topic of a category is larger within the same category than in the other one (e.g., the word *money* has a wider range of relative frequency values within the class spam than within the class non-spam). Since the weights computed by the LAC algorithm are inversely proportional to a measure of such spread of values (i.e., X_{ji}), we obtain the trend shown in Figure 1. This analysis can be interpreted as the fact that the absence of a certain term is a characteristic shared across the emails of a given category; whereas the presence of certain keywords shows a larger variability across emails of a given category.

Results for the Newsgroups dataset are shown in Figure 2. In this case, the collections of terms receiving largest feature relevance weights in each cluster reflect the topic of that category. This is indeed expected in a typical categorization problem.

6 Conclusions

We have introduced a new unsupervised feature selection approach, based on frequent itemset mining, to handle multi-class classification of documents in absence of labels. In addition, we have derived a locally adaptive clustering algorithm that provides a clustering of the documents and the identification of cluster-dependent keywords via a continuous term-weighting mechanism. Our experimental results demonstrate the feasibility of our approach in terms of achieved accuracy measured against the ground truth. We have shown that the selected keywords are representative of the underlying categories, which provides evidence that our global feature selection method successfully retains discriminant words. Moreover, our subspace clustering technique is capable of further sifting the most relevant ones, while discarding the additional spurious words.

Acknowledgements. C. Domeniconi is in part supported by the NSF CAREER Award No. 0447814.

References

[1] C. Aggarwal, C. Procopiuc, J. L. Wolf, P. Yu, and J. S. Park. Fast algorithms for projected clustering. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1999.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1998.

[3] D. Barbará, C. Domeniconi, and N. Kang. Classifying documents without labels. In *Proceedings of the SIAM International Conference on Data Mining*, 2004.

[4] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *Proceedings of the VLDB Conference*, 2000.

[5] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, 2004.

[6] S. T. Dumais, T. Letsche, M. L. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[7] E. Keogh, K. Chakrabarti, S. Mehrotra, and M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 2001.

[8] C. Procopiuc, M. Jones, P. Agarwal, and T. Murali. A monte carlo algorithm for fast projective clustering. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 2002.

[9] A. Thomasian, V. Castelli, and C. Li. Clustering and singular value decomposition for approximate indexing in high dimensional spaces. In *Proceedings of the CIKM Conference*, 1998.