An Empirical Study of Category Skew on Feature Selection for Text Categorization

Mondelle Simeon and Robert Hilderman

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{simeon2m, hilder}@cs.uregina.ca

Abstract. In this paper, we present an empirical comparison of the effects of category skew on six feature selection methods. The methods were evaluated on 36 datasets generated from the 20 Newsgroups, OHSUMED, and Reuters-21578 text corpora. The datasets were generated to possess particular category skew characteristics (i.e., the number of documents assigned to each category). Our objective was to determine the best performance of the six feature selection methods, as measured by F-measure and Precision, regardless of the number of features needed to produce the best performance. We found the highest F-measure values were obtained by bi-normal separation and information gain and the highest Precision values were obtained by categorical proportional difference and chi-squared.

1 Introduction

Due to the consistent and rapid growth of unstructured textual data that is available online, text categorization is essential for handling and organizing this data. For a survey of well studied text categorization methods and other automated text categorization methods and applications, see [1] [2] [3] [4]. Feature selection methods [5] [6] are used to address the efficiency and accuracy of text categorization by extracting from a document a subset of the features that are considered most relevant.

In the literature, it is common for a feature selection method to be evaluated using a particular dataset or group of datasets. However, the performance of a method on different datasets where well-defined and well-understood characteristics are varied, such as the category skew (i.e., the number of documents assigned to each category), could prove to be useful in determining which feature selection method to use in certain situations. For example, given some text repository, where the number of documents assigned to each category possesses some level of skewness, which feature selection method would be most appropriate for a small training set? A large training set? A near uniform distribution of documents to categories? A highly skewed distribution? Some initial steps were made in addressing these important issues in [5]. However, this study has a fundamental problem in that it assumes the best performance regardless of the situation is obtained with no more than 2000 features.

1.1 Our Contribution

In this paper, we present an extensive empirical comparison of six feature selection methods. The methods were evaluated on 36 multiple category datasets generated so that the selected documents possessed particular category skew characteristics based upon the standard deviation of the number of documents per category. In this study, we use an exhaustive search on each of the datasets to determine the set of features that produces the maximum F-measure and Precision values regardless of the number of features needed to achieve the maximum. We also contrast the best performance against the most "efficient" performance, the smallest set of features that produces F-measure and Precision values higher than those obtained when using all of the features. Finally, we study the performance of the feature selection methods using predetermined fixed percentages of the feature space.

2 Methodological Overview

In this work, we used the SVM classifier provided in the Weka collection of machine learning algorithms, stratified 10-fold cross-validation, and macro-averaged F-measure and Precision values. Six feature selection methods were used in conjunction with the SVM classifier. Five are the "best" performing methods reported in [5]: information gain(IG), chi-squared(χ^2), document frequency(DF), bi-normal separation(BNS), and odds ratio(OR). The sixth method, categorical proportional difference(CPD), is a method described in [7].

We utilize the OHSUMED [8], 20 Newsgroups [8], and Reuters-21578 [8] text corpora as repositories from which to randomly generate datasets which contain documents of uniform, low, medium, and high category skew. Clearly, in a dataset where documents are uniformly distributed across categories, the standard deviation is 0%. In datasets, where documents are not uniformly distributed across categories, the standard deviation will be non-zero. In this work, the low, medium, and high category skew datasets have a standard deviation of 16%, 64%, and 256%, respectively. Given these datasets and the set of feature selection methods, our approach to generating the experimental results consists of two phases, as described in [7].

3 Experimental Results

In this section, we present the results of our experimental evaluation of each of the feature selection methods on the 36 datasets.

3.1 Average Maximum Precision and F-measure

In Table 1, the P and F columns for the *Base*, CPD, OR, BNS, DF, χ^2, and IG describe the Precision and F-measure values respectively, obtained when using the complete feature space (*Base*) and each of the respective methods. The highest values are highlighted. Table 1 shows that all the feature selection methods obtained higher average maximum Precision values than the *Base* on all the datasets. The highest average maximum Precision values on the uniform,

Table 1. Average Maximum Precision and F-Measure

	Base		*CPD*		*OR*		*BNS*		*DF*		χ^2		*IG*	
Skew	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*
Uniform	0.670	0.662	**0.901**	0.736	0.810	0.719	0.843	**0.774**	0.667	0.661	0.755	0.744	0.737	0.733
Low	0.669	0.646	**0.920**	0.727	0.850	0.698	0.845	**0.749**	0.678	0.666	0.762	0.735	0.756	0.726
Medium	0.670	0.576	**0.918**	0.652	0.829	0.610	0.865	0.660	0.672	0.623	0.786	0.692	0.765	**0.695**
High	0.379	0.273	0.565	0.322	0.697	0.476	0.731	0.524	0.565	0.408	**0.734**	0.522	0.722	**0.565**

low, and medium category skew datasets is obtained by CPD, representing an increase of 34.4%, 37.5%, and 37.0%, respectively, over the *Base* values. On the high category skew datasets, although χ^2 obtained the highest maximum Precision value, the values obtained by χ^2, BNS, IG, and OR are not significantly different. The highest average maximum F-measure values on the uniform and low category skew datasets is obtained by BNS, representing an increase of 16.9% and 15.9% over the *Base* values, respectively. On the medium and high category skew datasets, the highest average maximum F-measure value is obtained by IG, representing an increase of 20.7% and 107% over the *Base* values, respectively.

3.2 Efficiency

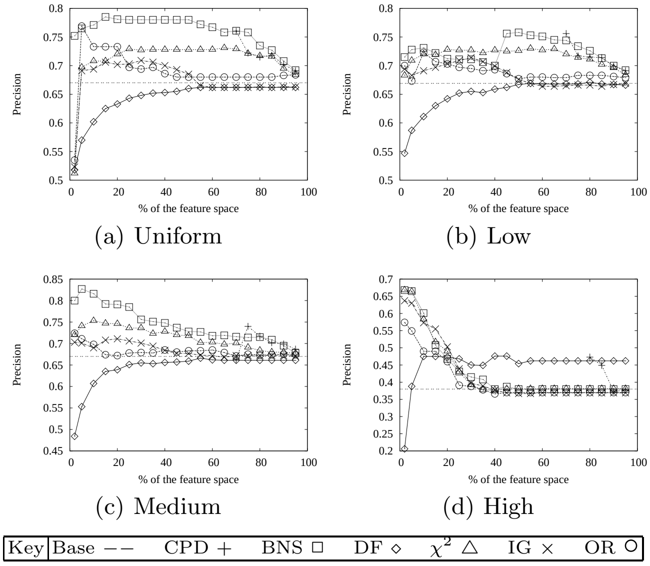
Efficiency is defined as the smallest percentage of the feature space required to obtain F-measure or Precision values that are higher the *Base*. To obtain the efficiency values for each of CPD, OR, BNS, DF, χ^2, and IG, the smallest percentages of the feature space (satisfying the definition) from each of the uniform, low, medium, and high category skew datasets for each text corpora were averaged for each respective method. Efficiency values for CPD, OR, BNS, DF, χ^2, and IG were 71.5%, 13.9%, 7.0%, 13.3%, 4.1%, and 2.8% respectively for F-measure, and 63.5%, 2.8%, 2.0%, 17.3%, 3.8%, and 5.9%, respectively, for Precision.

3.3 Pre-determined Threshold

Fig. 1 shows the average Precision values (along the vertical axis) obtained for each feature selection method using a fixed percentage of the feature space, varying according to the sequence 2%, 5%, 10%, 15%, ..., 95% (along the horizontal axis). Figs. 1(a), 1(b), 1(c) shows that BNS obtains the highest Precision values at most of the sampled cutpoints. In Fig. 1(d), at 2% of the feature space, the Precision values obtained by BNS, χ^2, OR, and IG represent a 76%, 76%, 51% and 67% improvement, respectively, over the *Base* values. Above 20%, the Precision values obtained by DF are higher than those obtained by any of the other methods, except at 80%, where CPD is highest. Similar results were also obtained for F-measure (not shown due to space constraints).

4 Conclusion

This paper presented a comparative study of six feature selection methods for text categorization using SVM on multiple category datasets having uniform, low, medium, and high category skew. We found that the highest F-measure



(a) Uniform (b) Low

(c) Medium (d) High

Fig. 1. Precision vs. % of the feature space

values were obtained by bi-normal separation (uniform and low) and information gain (medium and high), while the highest Precision values were obtained by categorical proportional difference (uniform, low, and medium) and χ^2 (high). We also found the most efficient methods were bi-normal separation (F-measure) and information gain (Precision).

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1) (2002) 1–47
2. Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some effective techniques for naive bayes text classification. IEEE Trans. on Knowl. and Data Eng. **18**(11) (2006) 1457–1466
3. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features, Springer Verlag (1998) 137–142
4. Han, E.H., Karypis, G., Kumar, V.: Text categorization using weight adjusted k-nearest neighbor classification. In: PAKDD '01, London, UK, Springer-Verlag (2001) 53–65
5. Forman, G.: Feature selection for text classification. In H.Liu, Motoda, H., eds.: Computational Methods of Feature Selection, Chapman and Hall/CRC (2008) 257–276
6. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization, Morgan Kaufmann Publishers (1997) 412–420
7. Simeon, M., Hilderman, R.J.: Categorical proportional difference: A feature selection method for text categorization. In: AusDM. (2008) 201–208
8. Asuncion, A., Newman, D.: UCI machine learning repository (2007)