Feature subset selection in text-learning

Dunja Mladenić

Department of Intelligent Systems, J.Stefan Institute,
Jamova 39, 1100 Ljubljana, Slovenia
Phone: (+38)(61) 1773 272, Fax: (+38)(61) 1258-158
E-mail: Dunja.Mladenic@ijs.si
http://www-ai.ijs.si/DunjaMladenic

Abstract. This paper describes several known and some new methods for feature subset selection on large text data. Experimental comparison given on real-world data collected from Web users shows that characteristics of the problem domain and machine learning algorithm should be considered when feature scoring measure is selected. Our problem domain consists of hyperlinks given in a form of small-documents represented with word vectors. In our learning experiments naive Bayesian classifier was used on text data. The best performance was achieved by the feature selection methods based on the feature scoring measure called *Odds ratio* that is known from information retrieval.

1 Introduction

In propositional learning problem domain is given by a set of examples, where each example is described with a class value and a vector of feature values. Features used to describe examples are not necessary all relevant and beneficial for the inductive learning and may reduce quality of induced model. Additionally, a high number of features may slow down the induction process while giving similar results as obtained with much smaller feature subset.

Section 2 describes approach commonly used for feature subset selection in learning on text data (text-learning). In Section 4 we experimentally compare different feature scoring measures on real-world data collected from Web users. Section 3 describes our problem domain and naive Bayesian classifier for text that we used in experiments. Discussion is given in Section 5.

2 Feature subset selection approaches

Different methods have been developed and used for feature subset selection in statistics, pattern recognition and machine learning, using different search strategies and evaluation functions. John et al. [4] pointed out the difference between the two main approaches used in machine learning to feature subset selection: filtering approach where the feature subset is selected independent of the learning method and wrapper approach where the feature subset is selected using the same learning algorithm that will be used for learning on domain represented with the selected feature subset.

The usual way of learning on text defines a feature for each word that occurred in training documents. This can easily result with several tens of thousands of features. Most methods for feature subset selection that are used information retrieval and text-learning (eg. [1], [3], [11]) are very simple compared to the methods developed in machine learning. Basically, some scoring measure that is used on a single feature is selected, a score is assigned to each feature independently, features are sorted according to the assigned score and a predefined number of the best features is taken to form the solution feature subset. Scoring of individual features can be performed using some of the measure used in machine learning for feature selection during the learning process, for example, $Information\ gain$ used in decision tree induction. In information retrieval a commonly used feature scoring is by $Odds\ ratio$ [12], where the problem is to rank out documents according to their relevance for the positive class value C_1 using occurrence of different words as features.

$$OddsRatio(F) = \log \frac{odds(W|C_1)}{odds(W|C_2)} = \log \frac{P(W|C_1)(1-P(W|C_2))}{(1-P(W|C_1))P(W|C_2)}$$

Where $P(W|C_i)$ is the conditional probability of word W occurring given the i-th class value. We handle singularities as proposed in [13]. Notice that Odds ratio is not averaged over all feature values, like Information gain, but only the value recording that word occurred is considered. We experimented also with three measures inspired by the original Odds ratio formula: $FreqOddsRatio(A) = Freq(W) \times OddsRatio(W)$, $FreqLogP(A) = Freq(W) \times \log \frac{P(W|C_1)}{P(W|C_2)}$, $ExpP(A) = e^{P(W|C_1)-P(W|C_2)}$. In our experimental comparison, we also include a very simple measure reported to work well in text classification domains [14] that scores each word by $Freq(W)$ frequency of word W. As a baseline measure we used Random score assignment.

In some text-classification experiments some other measures similar to Information gain are used like expected $Cross\ entropy$, $Keyword\ checker$ and $Mutual\ information$ used in [6], [5], [14] respectively. These are not included in our experiments, since Information gain they are based on performed worst on our data than Odds ratio (see Section 4 for more details).

3 Domain description

Machine learning problem is here defined as predicting clicked hyperlinks from the set of Web documents visited by the user. This is performed on-line while user is sitting behind some Web browser and waiting for the requested document. Our prototype system called Personal WebWatcher [10] uses text-learning on this problem, learning separate model for each user and highlighting hyperlinks on the requested Web documents. All hyperlinks from the visited documents are used as machine learning examples. Each is assigned one of the two class values: positive (user clicked on the hyperlink) or negative. We use machine learning to model the function $User_{HL} : HyperLink \to \{pos, neg\}$. Our hope is that this function is also some approximation of interesting hyperlinks (user clicked on hyperlinks that she/he is interested in and skipped all other hyperlinks, that is of course not always true!). We represented each hyperlink as a small document containing underlined words, words in a window around them and words in all the headings above the hyperlink. Our documents are represented as word vectors (using so called bag-of-words representation commonly used in information retrieval) and learning was performed using Naive (simple) Bayesian classifier the same way as in [3] or [10]. For each word position in the document, a feature is defined having a word as its value [9].

Our experiments are performed on data collected for users participating in the HomeNet project [2]. Results for two users are described in Section 4 with the data characteristics are given in Table 1. For each user approximately 4000 different words occurred in documents resulting here with 4000 features.

Domain (user id.)	Positive class probability	Number of examples	data entropy
usr150101	0.104	2 528	0.480
usr150211	0.044	2 221	0.259

Table 1. Domain description for data collected from two HomeNet users. It can be seen that we are dealing with unbalanced class distribution, since 10 % or less examples are positive, all other are negative.

4 Experiments

We used hold-out testing with 10 repetitions using 30% randomly selected examples as testing examples and reported average value and standard error. Feature selection and learning was performed on training examples only. For each data set we observed the influence of the number of the best features selected for learning (vector size) to the system performance. Since we have unbalanced class distribution (see Table 1), Classification accuracy can give misleading results. For such domains more appropriate measure is Information score [7] or Geometric mean of accuracy [8]. In the experimental results presented in Figure 1 Classification accuracy and Information score are used to estimate model quality. For both domains the highest Classification accuracy and the highest Information score are achieved by the measures based on Odds ratio: $ExpP, FreqLogP, OddsRatio$ (see Table 2 and Figure 1). For these measures the best vector size is approximately between 60 and 200 best features. This means that the selected feature subset includes just 2% - 5% of all features. The similar reduction (up to 90%) in the number of features used in text-learning was observed in [14]. The other three measures ($InfGain, Freq, FreqOddsRatio$) for most vector sizes achieved worst results than *Random*. Closer look to words sorted by Information gain showed that most best words are characteristic for negative class value (their probability for positive documents is 0). This means that in classification a new positive hyperlink is represented with a word vector almost full of zeros, since it contains very few of the selected best words. In our experiments we didn't remove any common or frequent words. That resulted with html-tags and other common words to be the most frequent, contributing to the poor performance of the Frequency measure *Freq* (the lowest line on the all four graphs). Our explanation for bad results achieved by the combination of Frequency and Odds ratio *FreqOddsRatio* is that the value of Frequency is standing out in this combination. Odds ratio has most values between 1 and 20, Frequency between 1 and 1000 and their combination between 10 and 1000. In case of the combination of Frequency with logarithm of probability quotient *FreqLogP*, the logarithmic part is standing out and measure achieves better results.

Scoring measure	Accuracy					Information score				
best features	10	100	200	500	1000	10	100	200	500	1000
user150101										
ExpP	**94.35**	**94.49**	**94.52**	**94.49**	**94.19**	**0.037**	**0.042**	**0.043**	**0.038**	**0.021**
FreqLogP	**94.32**	**94.49**	**94.52**	**94.50**	**94.18**	**0.036**	**0.041**	**0.042**	**0.037**	**0.019**
OddsRatio	94.08	94.27	94.04	93.68	93.26	0.030	0.036	0.027	0.016	-0.002
InfGain	94.24	92.62	92.27	92.16	92.09	0.011	-0.064	-0.073	-0.074	-0.071
FreqOddsRatio	92.44	92.35	92.33	92.15	91.87	0.048	-0.045	-0.044	-0.047	-0.06
Freq	91.72	90.85	90.75	90.73	91.74	0.264	-0.242	-0.227	-0.210	-0.197
Random	93.39	93.37	93.37	93.23	92.73	0.005	-0.012	-0.020	-0.035	-0.059
user150211										
ExpP	**96.61**	**96.63**	**96.60**	**96.42**	**95.97**	**0**	**0.001**	**-0.003**	**-0.014**	**-0.039**
FreqLogP	**96.60**	**96.62**	**96.56**	**96.41**	**95.97**	**-0.001**	**0**	**-0.005**	**-0.017**	**-0.042**
OddsRatio	**96.61**	**96.64**	**96.51**	**96.22**	**95.76**	**0.002**	**0.005**	**-0.008**	**-0.027**	**-0.055**
InfGain	94.31	94.01	93.66	93.29	93.09	-0.226	-0.294	-0.320	-0.334	-0.337
FreqOddsRatio	95.76	95.52	95.43	95.25	94.86	0.130	-0.136	-0.138	-0.145	-0.16
Freq	94.49	92.60	92.05	91.74	91.62	-0.374	-0.40	-0.420	-0.428	-0.427
Random	96.57	996.46	96.38	96.18	95.77	**0.003**	-0.015	-0.035	-0.066	-0.103

Table 2. Subset of classification results plotted in Figure 1 for two domains.

5 Discussion

We experimentally compared seven attribute scoring measures on our data: Information gain used in most machine learning experiments on text data (eg. [3], [11], [14]), four variants of Odds ratio as the most promising for our problem, Frequency and Random that we used as a baseline measure.

The results of experiments suggest that in feature subset selection for text-learning the best 2 % to 5 % of features should be selected. This finding is not in contradiction with the results reported on other text-learning problem domains eg. [11], [14]. The experimental results further suggest that for our problem domain, where one class value is the target class value, features should be scored using some measure based on Odds ratio. In our problem, we would like to identify as many positive examples as possible and we don't really care about identifying negative examples. The prior probability of positive class value is rather low (0.1 or lower). Naive Bayesian formula that we used in learning, considers only words that occurred in document. When we use rapid feature reduction, many examples are represented with word vectors almost full of zeros that are

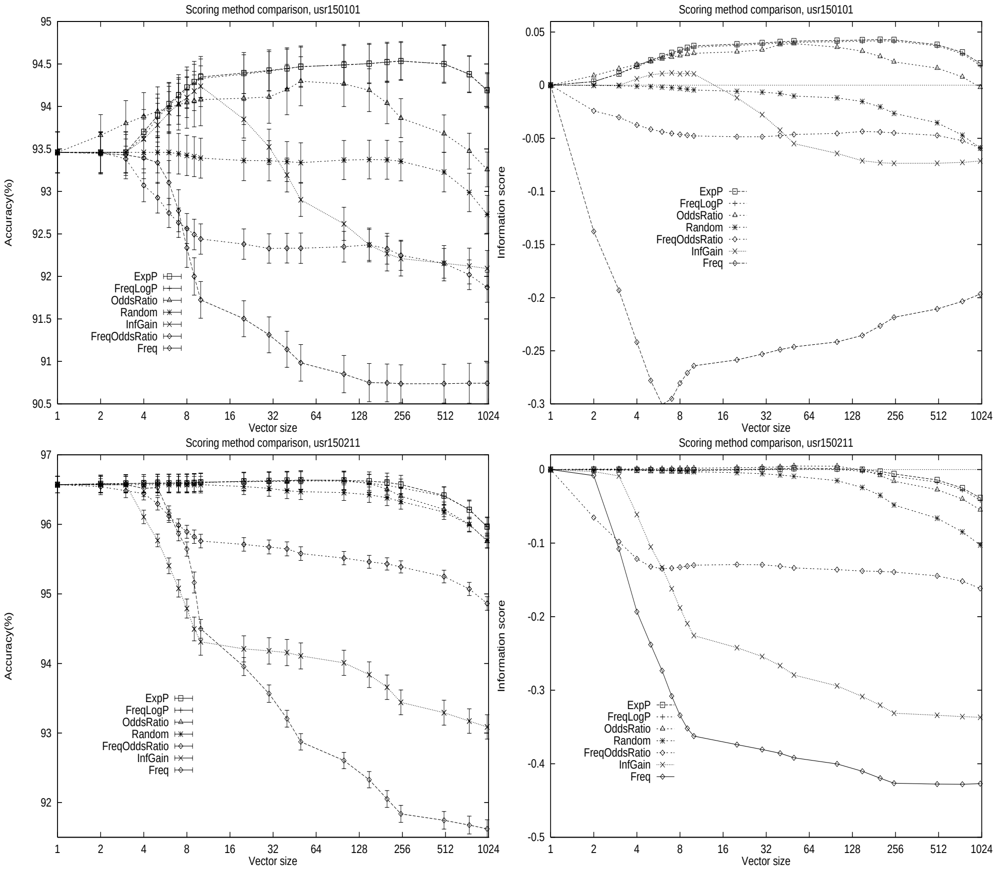

Fig. 1. Influence of vector size to Classification accuracy and Information score on data for HomeNet usr150101 (upper) and usr150211 (lower). Notice that curve names are sorted according to the values at the end (vector size=1000).

classified mostly according to the prior probability. This means that in order to identify positive documents $P(‘pos'|D) > P(‘neg'|D)$, we have to select features that will raise the probability of positive class value. This is possible only if $P(W_j|‘pos') > P(W_j|‘neg')$ holds with sufficient difference for sufficient number of the product members used in naive Bayesian formula. Thus, we based our new feature scoring measures ($FreqLogP$, $ExpP$) on that condition.

Experimental results pointed out the need to consider problem domain and machine learning algorithm characteristics when selecting a feature scoring measure. This is especially important for such simple feature subset selection approach as used in text-learning, where the solution quality is traded for time complexity. The feature subset found in this way is an approximation that assumes feature independence. The same false assumption is used by naive Bayesian classifier that was used is our experiments.

In further experiments we plan to include more datasets, removal of infrequent features and common words (using 'stop-list'). With this last modification we would like to test on our data the hypothesis about good behavior of a simple scoring by Frequency set by Yang and Pedersen [14].

Acknowledgements This work was financially supported by the Slovenian Ministry for Science and Technology. Part of this work was performed during the authors stay at Carnegie Mellon University. Author is grateful to Tom Mitchell and his machine learning group at Carnegie Mellon University for generous support, suggestions and fruitful discussions.

References

1. Apté, C., Damerau, F., Weiss, S.M., Toward Language Independent Automated Learning of Text Categorization Models, *Proc. of the 7th Annual Int. ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.
2. Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J., Kiesler, S., The HomeNet Field Trial of Residential Internet Services, *Communications of the ACM* Vol. 39, No. 12, pp.55—63, December 1996.
3. Joachims, T., A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *Proc. of the 14th International Conference on Machine Learning ICML97*, pp. 143—151, 1997.
4. John, G.H., Kohavi, R., Pfleger, K., Irrelevant Features and the Subset Selection Problem, *Proc. of the 11th International Conference on Machine Learning ICML94*, pp. 121—129, 1994.
5. Kindo, T., Yoshida, H., Morimoto, T., Watanabe, T., Adaptive Personal Information Filtering System that Organizes Personal Profiles Automatically, *Proc. of the 15th Int. Joint Conference on Artificial Intelligence IJCAI-97*, pp. 716—721, 1997.
6. Koller, D., Sahami, M., Hierarchically classifying documents using very few words, *Proc. of the 14th International Conference on Machine Learning ICML97*, pp. 170—178, 1997.
7. Kononenko, I. and Bratko, I., Information-Based Evaluation Criterion for Classifier's Performance, *Machine Learning 6*, Kluwer Academic Publishers, 1991.
8. Kubat, M., Holte, R., Matwing, S., Learning When Negative Examples Abound, *9th European Conference on Machine Learning ECML97*, pp. 146—153, 1997.
9. Mitchell, T.M., Machine Learning, The McGraw-Hill Companies, Inc., 1997.
10. Mladenić, D., Personal WebWatcher: Implementation and Design, *Technical Report IJS-DP-7472*, October, 1996. http://www-ai.ijs.si/DunjaMladenic/papers/PWW/
11. Pazzani, M., Billsus, D., Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning 27*, Kluwer Academic Publishers, pp. 313—331, 1997.
12. van Rijsbergen, C.J,. Harper, D.J., Porter, M.F., The selection of good search terms, *Information Processing & Management*, 17, pp.77—91, 1981.
13. Shaw Jr, W.M., Term-relevance computations and perfect retrieval performance, *Information Processing & Management*, 31(4), pp.491—498, 1995.
14. Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization, *Proc. of the 14th International Conference on Machine Learning ICML97*, pp. 412—420, 1997.

This article was processed using the LaTeX macro package with LLNCS style