Analysis and evaluation of Feature selectors in opinion mining

J. ISABELLA
Research Scholar, Sathyabama university,
Chennai,India
isabellajones71@gmail.com

Dr. R.M.SURESH
Jerusalem.Engineering College,
Chennai,India

ABSTRACT

Computational performance is improved by use of basic feature selection in most of the research works. Sentiment analysis identifies whether opinion expressed on a topic in a document is positive or negative. But many potential sentiment analysis applications are not feasible because of the voluminous amount of features present in the corpus. This paper evaluates a range of feature selectors systematically with respect to their efficiency in improving the performance of the classifiers for sentiment analysis. Movie reviews are used for sentiment analysis in this study.

KEYWORDS: Sentiment analysis, Movie reviews, Feature Selection, Correlation based feature selector (CFS), Information Gain, Support Vector Machine (SVM), Principal component analysis (PCA)

INTRODUCTION

Machine learning studies algorithms which automatically improves performance with experience. Prediction is the main outcome of performance. An algorithm is said to have learned when it improves its ability to predict key elements of a task when presented with proper data. Machine learning algorithms are characterized by languages which represent knowledge. Research reveals that no single learning approach is superior and usually different learning algorithms produce the same results [1]. A factor which has much impact on learning algorithm's success is the nature of data that characterizes the task to be learned. Learning fails in machine learning algorithms when data does not exhibit statistical regularity. While new data can be constructed from old to exhibit statistical regularity and facilitate learning, it is a complex task that a fully automatic method is difficult [2].

If data suits machine learning, then discovering regularities is easy and less time bound through removal of data features irrelevant/redundant regarding to the task to be learned. This procedure termed is feature selection. Unlike constructing new input data, feature selection is fully automatic, computationally tractable process. Feature selection benefits learning by reduction in data required to achieve learning, improved predictive accuracy, compact learned knowledge and easily understood, and lower execution time.

Generally, sentiment analysis approaches focus on classifying documents based on the sentiment of individual features [3]. While these did not require domain-specific training data, accuracy achieved is not high. Later research focused on supervised learning techniques like Support Vector Machine (SVM) for text categorisation work [4]. Though these supervised learning techniques were proven to be more accurate than text-based approaches, the huge number of features made the process computationally expensive.

When classifying a document, a number of words which are used as features are considered, though only a few words in the corpus actually express sentiment. These extra features has to be eliminated as they slow down document classification as there are more words than really needed and secondly it reduces accuracy as the classifier should consider such words when a document is under classification. Using fewer features is advantageous and hence to remove those unnecessary features, selection is resorted to. As the name suggests, feature selection is the process wherein a corpus is run through prior to the classifier being trained to remove any unnecessary features. This enables a classifier to fit a model to the problem set expeditiously as there is less information to consider leading to better accuracy.

Existing machine learning feature selection methods fall into two divisions — those evaluating features worth using learning algorithm that could be applied to data, and those evaluating feature worth through use of heuristics based on data's general characteristics. The former are called wrappers and the latter filters [5, 6]. Wrappers provide better results as regards final predictive learning algorithm accuracy than filters as feature selection is optimized for a particular learning algorithm. But as a learning algorithm evaluates every features set considered, wrappers are very costly to run, and are intractable for large databases having many features.

Further, as feature selection is combined with a learning algorithm, wrappers are not as common as filters. They should also be re-run when moving from one learning algorithm to another.

The process of filters and wrappers are shown in Figure 1.

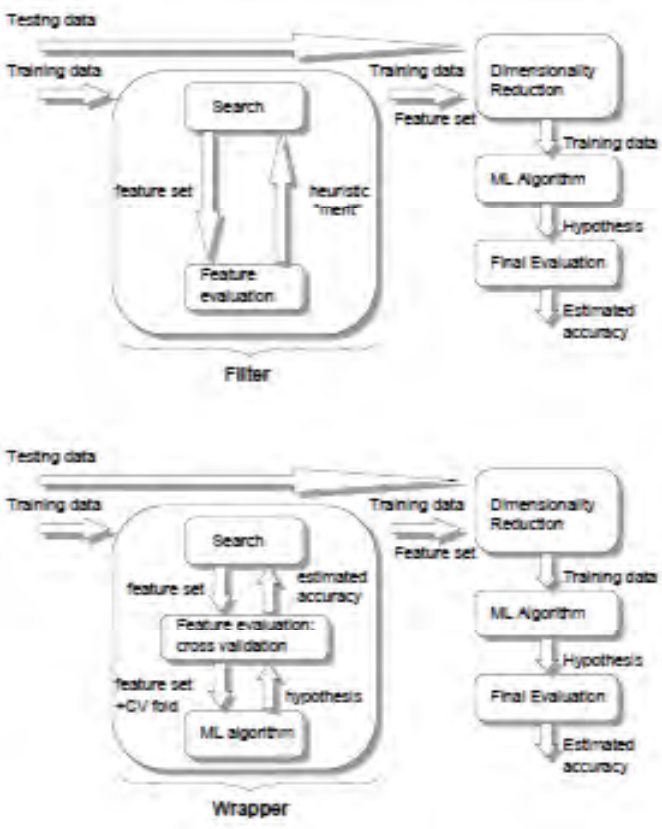

Figure 1: Filter and Wrapper Feature selection process

Data is represented as a table of examples/instances in a supervised machine learning task with each instance being described by a fixed number of measurements/features, with a class denoting label. Features are of two types: nominal where values are members of an unordered set, or numeric where values are real numbers. A machine learning algorithm application needs two sets of examples: training and test examples. Training examples produce learned concept descriptions for each class and a separate test sample set is required to evaluate accuracy. Class labels are not presented to the algorithm when under test. The algorithm inputs a test example and outputs a class label.

Feature selection algorithms with some exceptions, perform a search through feature subsets and consequently addresses four basic issues affecting search [7]:

1. Starting point: Starting point of the search in the feature subset space affects the search direction. The search starts with no features; it proceeds in forward direction when features are added successively. If the search starts with all features, the search proceeds backwards by successively removing features [8].

2. Search organisation: A total feature sub space search is computationally expensive except for limited features as there are 2^N possible subsets for N initial features. Thus, heuristic search strategies are feasible and commonly used which give good results, though they do not warrant locating the optimal subset.

3. Evaluation strategy: Filters [5, 6] operate independently of learning algorithms by removing undesirable features through a filter in the data before learning starts. Such algorithms use heuristics based on general data characteristics to evaluate feature sub set merits. Wrapper uses an induction algorithm with a statistical re-sampling technique like cross-validation to estimate feature sub sets last accuracy.

4. Stopping criterion: A feature selector decides when to stop searching through feature subsets space. Based on evaluation, a feature selector can stop adding or removing features when no alternative improves current feature subset's merits. Searching feature subsets space within reasonable time constraints is required if a feature selection algorithm should operate on data with many features.

Pang et al. [4] movie review dataset which is considered as the benchmark for sentiment analysis of movie reviews contains around 51,000 unique words and symbols. As only a few features provide useful information to the classifier, feature selection can reduce feature number. This paper studies the effects of various feature

selection methods in sentiment analysis. Four feature selection methods, Correlation based feature selector (CFS), Information Gain, Support Vector Machine (SVM), Principal component analysis (PCA) are studied in this paper. The efficiency of the feature selectors are evaluated on the basis of the classification accuracy achieved for sentiment in movie review.

RELATED WORKS

Pang et al. [4] used supervised learning in sentiment analysis with the aim of determining whether it could be treated as a special case of topic-based categorisation with positive and negative topics. Naıve Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) classifiers were tested to achieve this, with all performing well in topic-based categorisation. Document words and symbols were used for features as either a unigram or a bigram bag-of-features, with the former performing better. Feature Frequency (FF) and Feature Presence (FP) when tested revealed that by using a SVM with unigram FP better accuracy could be achieved (82.9%) in a 3-fold cross validation.

Generally, to review the quality of product the most accepted way is online reviews. A decision is made by the customer using the other customer's feedback on blog, websites. Hence, to improve marketing, upgrade product and services the businesses should essentially observe the feedback. Various forums like blogs, review websites, and discussion forums contain the necessary feedback. In order to obtain and process reviews and give a synopsis of essential information automatically and efficiently is achieved by opinion mining. Isabella et al., [9] proposed the feature set extraction from the movie opinions. Implementing the proposed correlation based feature reduction technique the feature set is minimized by computing the inverse document frequency. Employing Naive Bayes and Support Vector Machine classifier the proposed preprocessing technique efficacy is tested. Experiments conducted showed promising results.

Simeon et al., [10] proposed categorical proportional difference (CPD), a novel feature selection technique. CPD is a measure for the degree of the word's contribution to distinguish a specific category from other categories. In a text corpus, for a word in a specific category the CPD is a ratio that relates the word present in the number of documents of a category and the word present in number of documents in other categories. To evaluate CPD a set of experiments are performed using OHSUMED, 20 Newsgroups, and Reuters-21578 text corpora. SVM and Naive Bayes text classifiers were used to evaluate performance. The measures used to estimate the performance are precision, recall and the F-measure. The results accomplished by CPD are compared with the six popular feature selection techniques and CPD outperformed four out of six text categorization tasks.

O'Keefe et al., [11] evaluated a variety of feature selectors and feature weights using both Naïve Bayes and Support Vector machine classifiers in a systematic manner. Two novel feature selection techniques and three feature weighting techniques were used. Experiment results showed that when only 29% of the features were used, a classification accuracy of about 87.15% was achieved.

Yessenov et al., [12] proposed an empirical study of efficiency of machine learning methods to classify text messages using semantic meaning. The movie review comments obtained from the common popular social network Digg was used as the data set. By the subjectivity/objectivity and attitudes negative/positive the text are classified. In order to extract the text features, bag-of-words model was used. Their effectiveness on the accuracy of four machine learning techniques- Decision Trees, Naive Bayes, K-Means clustering and Maximum-Entropy are estimated. To conclude the summary is provided regarding the accuracy rates achieved. The results reveal that it can be further be refined implementing the selection of features on the basis of syntactic and semantic information obtained from the text.

METHODOLOGY

IMDb Dataset

The study uses online movie reviews as data for evaluating the feature selectors. Internet Movie Database (IMDb) is an online database of information regarding movies, television and fictional visual entertainment media. Pang et al [4] provide a collection of movie-review documents from IMDb archives that are categorized on overall sentiment polarity as positive/negative or subjective rating (e.g., two stars).

Inverse Document Frequency (IDF)

The term document frequency is computed as follows for a set of documents x and a set of terms a. Each document is modeled as a vector v in the a dimensional space R^a. When the term frequency denotes by $freq(x, a)$, it expresses the number of occurrence of the term a in document x. The term-frequency matrix $TF(x, a)$ measures term association a with regard to a given document x. $TF(x, a)$ is assigned zero when the document has no term and $TF(x, a)$ = 1when term a occurs in the document x or uses a relative term

frequency which is **term frequency** as against the total occurrences of all document terms. Frequency is generally normalized by [13]:

$$TF(x,a) = \begin{cases} 0 & freq(x,a) = 0 \\ 1 + \log\left(1 + \log\left(freq(x,a)\right)\right) & otherwise \end{cases}$$

Inverse Document Frequency **(IDF)**, represents scaling factor. When a term a occurs frequently in many documents, its importance is then scaled down because of its lowered discriminative power. The $IDF(a)$ is defined as follows:

$$IDF(a) = \log \frac{1+|x|}{x_a}$$

x_a is the set of documents containing term a.

Though TF-IDF is commonly used metric in text categorisation [14], its use in sentiment analysis is less known and is also not known to have been used as a unigram feature weight. TF-IDF is has two scores, term frequency and inverse document frequency. Term frequency is just counting the number of times a term occurs in a specific document, whereas inverse document-frequency is got by dividing total documents by those documents in which a specific word appears repeatedly. Multiplication of these values leads to a high score for words occurring repeatedly in few documents. The score is low for terms that appear frequently in all documents.

Similar documents have same relative term frequencies measured among a document set or between a document and a query. Cosine measure aids in locating similarity between documents and the cosine measure is given by:

$$sim(v_1, v_2) = \frac{v_1 . v_2}{|v_1| \ |v_2|}$$

where v_1 and v_2 are two document vectors, $v_1 . v_2$ defined as $\sum_{i=1}^{a} v_{1i} v_{2i}$ and $|v_1| = \sqrt{v_1 . v_1}$.

After completion of IDF for text documents, four different feature selectors including Correlation based feature selector, Info Gain attribute evaluator, SVM attribute evaluator and Principal component evaluator are applied to them.

Correlation based feature selector (CFS)

Correlation based feature selector (CFS) is a filter algorithm which ranks feature subsets according to correlation based heuristic evaluation [14]. Its bias is toward subsets having features highly correlated to the class and uncorrelated with others. Irrelevant features are ignored due to the low correlation they have with the class. Redundant features are screened out as they are highly correlated with one or many features. Feature acceptance depends on the extent to which it predicts classes where space is not already predicted by other features. CFS's feature subset evaluation function given by:

$$M_S = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}}$$

where M_S is heuristic merit of a feature subset

S is feature subset

$\overline{r}_{cf}$ is mean feature-class correlation $(f \in S)$

$\overline{r}_{ff}$ is average feature-feature inter correlation

Information Gain

The Info gain procedure calculates an instance's probability because it is a segment border and compares it to a segment border probability where a feature has a specific value [15]. The higher the probability change, the more useful is the feature. This simple ranking process is regularly used regularly in text categorisation applications where voluminous data prevents the use sophisticated attribute selection techniques. Decreasing class entropy reveals additional class information provided by the attribute and is called information gain

Support Vector Machine (SVM)

Original training data is transformed into a higher dimension through nonlinear mapping used by a support vector machine (SVM) [16]. This is done through nonlinear mapping data from two classes which are separated by a hyperplane. The hyperplane is located by SVM which uses support vectors and margins. The distance between the hyperplane and the entity is the margin. The advantage is of SVM is that it is highly accurate and less liable to overfitting but the drawback being, it is time consuming. The SVM with input vector $\vec{x}$, and $\vec{w}$ the normal vector to hyperplane, the output u is given by:

$$u = \vec{w}.\vec{x} - b$$

The separating hyperplane is the plane $u = 0$. The margin is given by:

$$m = \frac{1}{\|w\|_2}$$

then maximizing the margin is equivalent to solving the following optimization problem:

$$\min_{w,b} \quad \frac{1}{2} \| w \|^2$$

subject to $y_i = \left(\vec{w}.\vec{x} - b \right) \geq 1$

b is a bias variable, and N is the number of training example. It follows that the margin corresponds to the quantity $1/\| w \|$ and the maximization of margin is achieved by minimizing $\| w \|^2$

Principal component analysis (PCA)

A statistical technique, Principal component analysis (PCA) reduces data dimensionality by transforming original attribute space. Computing original attributes covariance matrix and extracting eigenvectors leads to formation of transformed attributes. The former (principal components) defines original attribute space from linear transformation to a new space where attributes are un-correlated. Eigenvectors are ranked based on original data variations they account for. Usually, the first few transformed attributes account for most retained data variation while the remainder are rejected. PCA requires no supervision as it does not use class attribute.

PCA feature extraction gets new attributes using a linear combination of original attributes and achieves dimensionality reduction by holding on to highest variance components. Principal components are less than or equal to original variables in numbers and the transformation is so delineated that a first principal component has the biggest variance, leading to great data variability. Hence, each successive component has highest variance possible in turn under an orthogonal constraint - uncorrelated to - preceding components.

RESULTS AND DISCUSSION

A subset of the dataset containing two hundred reviews with 40 positive opinions and 40 negative opinions is used for evaluation of the feature selectors. A list of stop words for common words and stemming of words is done and then the IDF is computed. The various feature selectors Correlation based feature selector, Info Gain attribute evaluator, SVM attribute evaluator and Principal component evaluator are applied to perform the feature extraction. The features selected are classified using k Nearest Neighbour classifier and the results obtained are compared. Figure 2 shows the classification accuracy obtained from the various techniques.

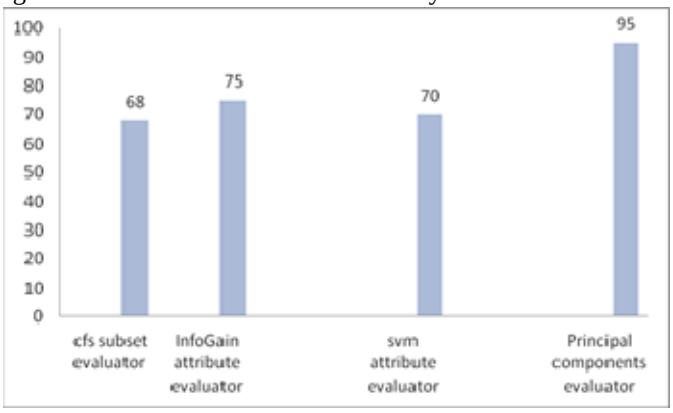

Figure 2: Efficacy of Feature Selectors

It is observed from Figure 2 that the PCA achieves the maximum accuracy of 95% and outperforms other selectors by 20 to 27%. Thus, for mining movie reviews, PCA is the most appropriate feature selector.

CONCLUSION

This paper proposes to extract words from movie reviews, select words based on importance through use of IDF. The feature set is reduced through use of various types of feature selectors like Correlation based feature selector, Info Gain attribute evaluator, SVM attribute evaluator and Principal component evaluator. K-Nearest Neighbour classifier is used to calculate classification accuracy to evaluate the efficiency of the feature extractors. Experimental results show that Principal component evaluator achieves the best feature subset for classification of sentiment for movie reviews. To conclude, Principal component evaluator as feature extractors is effective for multiple attributes and large scale multivariate data pertaining to opinion mining for movie reviews.

REFERENCES

[1] P. Langley and H. A. Simon. Applications of machine learning and rule induction. Communications of the ACM, 38(11):55–64, 1995.
[2] C. J. Thornton. Techniques in Computational Learning. Chapman and Hall, London, 1992.
[3] P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In ACL '02: Proc. of the 40th Annual Meeting on Association for Computational Linguistics, pages 417–424, Morristown, NJ, USA, 2002. ACL.
[4] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP '02: Proc. of the ACL-02 conf. on Empirical methods in natural language processing, pages 79–86, Morristown, NJ, USA, 2002. ACL.
[5] R. Kohavi. Wrappersfor Performance Enhancement and Oblivious Decision Graphs. PhD thesis, Stanford University, 1995
[6] R. Kohavi and G. John. Wrappers for feature subset selection. Artificial Intelligence, special issue on relevance, 97(1–2):273–324, 1996.
[7] P. Langley. Selection of relevant features in machine learning. In Proceedings of the AAAI Fall Symposium on Relevance. AAAI Press, 1994.
[8] A. J. Miller. Subset Selection in Regression. Chapman and Hall, New York, 1990.
[9] Isabella, J., and R. M. Suresh. "An SVM Classifier using Correlation based Feature Selection for Opinion Mining." (2011), 1st International Conference on Information System, Computer Engineering & Applications (ICISCEA 2011) Date : November 28 – 29, 2011.
[10] Simeon, M., & Hilderman, R. (2008, November). Categorical proportional difference: A feature selection method for text categorization. In Proceedings of the Seventh Australasian Data Mining Conference (AusDM 2008) (Vol. 87, pp. 201-208).
[11] O'Keefe, T., & Koprinska, I. Feature Selection and Weighting in Sentiment Analysis. Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.
[12] Yessenov, K., & Misailovic, S. (2009). Sentiment Analysis of Movie Review Comments. Methodology, 1-17.
[13] Liu, T. Y., Xu, J., Qin, T., Xiong, W., & Li, H. (2007, July). Letor: Benchmark dataset for research on learning to rank for information retrieval. In Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval (pp. 3-10).
[14] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, vol. 1, no. 3, 1997.
[15] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.
[16] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). Feature selection for SVMs. Advances in neural information processing systems, 668-674.
[17] Guo, Q., Wu, W., Massart, D. L., Boucon, C., & De Jong, S. (2002). Feature selection in principal component analysis of analytical data. Chemometrics and Intelligent Laboratory Systems, 61(1), 123-132.