Efficacy of Feature Selectors for Classifiers in Opinion Mining

J. Isabella [1] *R.M.Suresh*[2]

Abstract

Emphasis placed on feature selection depends upon the learning algorithms. Feature selection identifies and removes irrelevant and redundant information and usually involves combining search and attribute utility estimation. It also evaluates the specific learning structures leading to many permutations. Data engineering is thought to be central in the mining applications development. Feature selection is an important and regularly used techniques in pre-processing the data for data mining as it reduces features number, removes irrelevant and noisy data and expedites data mining algorithms thereby improving mining performance parameters like accuracy and clarity. The aim of applying attribute evaluators for feature selection is the reduction of computational complexity and selected feature subsets improved classification accuracy. This paper aims to analyse different attribute evaluators/feature selectors and their classification using the K-Nearest Neighbour classifier following which evaluator performance is compared.

Keywords: *Feature selectors, K-Nearest Neighbour classifier, Mining*

I. Introduction

Feature selection is a research and development area in statistical pattern recognition, machine learning and data mining [1, 2] from the 1970s, and currently applicable to varied fields like text categorization [3, 4] and image retrieval [5, 6]. Feature selection is used to select a subset of original features. A feature subset is optimality measured by an evaluation criterion and this feature selection process includes 4 basic steps like subset generation, subset evaluation, stopping criterion, and result validation.

The search process of subset generation [7, 8] results in candidate feature subsets for evaluation based on a specific search strategy. Every candidate subset being evaluated and compared with the earlier best set based on specific evaluation criteria. An improved new subset replaces an earlier best subset with this process of subset generation and evaluation continuing till it satisfies a stopping criterion. Now the newly selected subset must be validated by earlier knowledge or through various tests via synthetic and/or real world datasets. Feature selection is widely applied in the field of data mining in areas like classification, clustering, association rules, and regression. The feature selection is also titled subset or variable selection in statistics. A simple feature selection algorithm is ad hoc in nature, but that is not to say that other methodical approaches do not exist. From a theoretical viewpoint, it is a fundamental requirement for a total search of all feature subsets as optimal feature selection for supervised learning problems improves the performance. But in the real world problems, this is impractical when many feature sets are available, so the search for a satisfactory set is done instead of an optimal set.

[1] Research Scholar, Sathyabama university, Chennai,India isabellajones71@gmail.com

[2] *Jerusalem.Engineering College, Chennai,India*

Feature selection algorithms typically are of two types; feature ranking and subset selection. The former ranks features through a metric eliminating those features without an adequate score, while the latter tries to locate features for an optimal subset. The major problem is to decide when to stop the algorithm, which a cross validation does in the machine learning. Subset selection evaluates features as a group for suitability. Subset selection algorithms are further broken down into Wrappers, Filters and Embedded. The first uses search algorithms to search through features space, evaluating subsets by running a model on them. The wrapper model requires a predetermined mining algorithm to evaluate performance and attempts feature location suiting mining algorithm which improve performance. But it is expensive computationally when compared to filter models [9, 10]. Wrappers are also expensive computationally and risk over fitting models. Filters and Wrappers are same in a search approach, but a simple filer is evaluated instead of model evaluation. Data characteristics form the core on which the filter model relies to evaluate and select feature subsets sans the use of mining algorithms. Embedded techniques are embedded in and model specific.

Though search approaches use greedy hill climbing, for evaluating a candidate subset of features iteratively, before modifying a subset and evaluating whether the new subset is an improvement over the old. Subset evaluation needs metric scoring which grades subsets features. As a voluminous search is not practical, a features subset with the highest score is selected as the satisfactory feature subset. The stopping criterion varies according to algorithm with possible criteria including a subset score exceeding a threshold, a program's maximum allowed run time being surpassed, etc. Two popular classification problem filter metrics are correlation and mutual information. Both are true metrics/'distance measures' in a mathematical sense as they do not obey the triangle inequality and hence do not compute actual 'distance' – and hence can only be regarded as 'scores', which are computed between a candidate feature and needed output category. But there are also true metrics which are a simple function of mutual information.

In this paper, different attribute evaluators/feature selectors such as Correlation based feature selector (CFS), SVM attribute evaluator, Principal components evaluator is analysed and their classification using the K-Nearest Neighbour classifier following which evaluator performance is compared. The opinions mined from unstructured text documents are classified to evaluate the efficiency of the feature selectors.

II Literature Survey

Machine learning survey review on feature selection is seen in [1], [11]. Particularly, Liu et. al. [11] used small artificial data sets to discover strengths and weaknesses of various attribute selection methods in connection with noise, different attribute types, multi-class data sets and computational complexity.

Some very successful learning algorithms include Quinlans iterative dichotomiser 3 (ID3) [12] and classifier 4.5 (C4.5) [13], classification and regression trees (CART) proposed by Breiman et al [14]. All of them use greedy search through decision trees space using an evaluation function at each stage to select an attribute with the best ability to discriminate classes.

Kira and Rendell [15] suggested another approach for feature selection with the filter based feature ranking algorithm (RELIEF), which was another proposal of theirs, assigning a weight to every feature on its capability to differentiate classes. It then chooses features whose weights are in excess of the defined threshold as the correct feature. Weight computation here is computed on the basis of the probability of nearest neighbours from two different classes with differing values for an attribute and probability of two nearest neighbours from the same class having the same attribute value. The bigger the difference between two probabilities, more significant is the attribute. Thus measure for a two class problem is defined and later extended to handle multiple classes. This is done through splitting the problem into many two-class problems.

Kononenko [16] suggested the use of k-nearest neighbours to increase probability approximation reliability. He also proposed that RELIEF could be extended to work efficiently with multiple sets. This is due to the fact that weighting schemes are easy to implement and hence are preferred for their efficiency.

Bo Pang et al., 2004 [17] investigated the effectiveness of classification of documents by overall sentiment using machine learning techniques. Experiments showed that the machine learning techniques give a better result than human produced baseline for sentiment analysis on movie review data. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews.

Peter Turney 2002 [18] proposes an unsupervised learning algorithm, using semantic orientation of the phrases containing adjectives and adverbs, to classify reviews. The approach initially extracts phrases containing adjectives and adverbs; the semantic orientation of the phrase is estimated using PMI-IR; based on the average semantic orientation the phrases the review is classified as recommended (Thumbs up) or not recommended (Thumbs down). Experiment was conducted using 410 reviews on various topics; average accuracy of 74% was achieved.

Xiaowen Ding et al., 2008 [19] proposes a holistic lexicon-based approach which uses external indications and linguistic conventions of natural language expressions to determine the semantic orientations of opinions. The proposed algorithm uses linguistic patterns to deal with special words, phrases.

Wiebe J et al., 2004 [20] proposed a learning method for creation of subjective classifiers, which can be used on unannotated text. The method developed is superior to other previously used supervised learning approaches. In an attempt to build classifiers which can distinguish subjective and objective sentences, a new objective classifier was created using new objective clues which achieved higher recall than previous works. Their approach began with seeding process which uses known subjective words to automatically create training data.

Pang et al., 2002 [21] proposed a machine-learning method to find subjective portions in a document. Extracting of the subjective portions can be done using techniques for finding the minimum cuts in graphs. This makes it easy to incorporate the cross sentence related constraints. Pang et al., studied the relationship between polarity classification and subjectivity discovery, showing that shorter extracts got from compressed reviews retain polarity information as that of the full review.

III Methodology

In this paper it is proposed to use online movie reviews as data due to the availability of a large number of reviews online. Internet Movie Database (IMDb) is an online database of information related to movies, television shows, and fictional visual entertainment media. Bo Pang and Lillian Lee [6] provide collections of movie-review documents collected from the IMDb archives which are categorized based on its overall sentiment polarity as positive/negative or subjective rating (e.g., two stars). Features are extracted by using list of stop words for commonly occurring words and stemming words with similar context. The terms document frequency is computed. In a set of documents x and a set of terms a, each document can be modeled as a vector v in the dimensional space, this is a vector space model. Let the term frequency be denoted by, this expresses the number of occurrence of the term in the document. The term-frequency matrix measures the association of a term with respect to the given document. is assigned zero if the document does not contain the term, and a number otherwise. The number could be set as $= 1$when term occurs in the document or uses the relative term frequency. The **relative term frequency** is the term frequency as opposed to the total number of occurrences of all the terms in the document. The term frequency is

generally normalized by :

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

Another measure used is the **inverse document frequency (IDF)**, it represents the scaling factor. If term a occurs frequently in many documents, then its importance is scaled down due to its reduced 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.

Similar documents have similar relative term frequencies. The similarity can be measured among a set of documents or between a document and a query. Cosine measure is generally used to find similarity between documents; the cosine measure is got 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 After performing IDF for the text documents, four different feature selectors namely, Correlation based feature selector, Info Gain attribute evaluator, SVM attribute evaluator and Principal component evaluator are applied to the documents.

A. Correlation based feature selector (CFS)

Feature selection plays an important role in machine learning, as machine learning algorithms uses features for analysis and prediction. Feature selection is a process of identifying most relevant features for learning; it eliminates irrelevant and redundant features of the data thus improving the performance of the learning algorithm. Feature selection is accomplished by filters, wrappers or CFS, a correlation-based feature selector algorithm. A good feature subset is one that contains features which are highly correlated to the class and uncorrelated with other features.

Correlation based feature selector (CFS) is a simple filter algorithm which ranks feature subsets according to correlation based heuristic evaluation function [10]. Irrelevant features get eliminated as it will have low correlation with the class and redundant features are removed as they are highly correlated with one or more feature. The CFS's feature subset evaluation function is 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.

B. Info Gain Attribute Evaluator

The Info gain procedure calculates an instance's probability as it is a segment border comparing this to the segment border probability in that a feature has a specific value. The higher the probability change the more useful the feature. This is a simple and quick attribute ranking process used regularly in text categorisation applications where a huge volume of data precludes use of highly sophisticated attribute selection techniques. The amount by which class entropy decreases reflects additional class information provided by the attribute and is known as information gain.

C. SVM attribute evaluator

A support vector machine (SVM) evaluator uses nonlinear mapping to transform the original training data into a higher dimension. With nonlinear mapping data from two classes is separated by a hyperplane. The SVM uses support vectors and margins to find hyperplane. The disadvantages of this method is that its time consuming. The advantages are that it is very accurate and less prone to overfitting. The margin is the distance between the hyperplane and the entity. The output for a 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:

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

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}

The optimization problem is converted to quadratic programming where the objective function ψ is dependent on Lagrange multipliers α_i,

$$\min_{\alpha} \psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \left(x_i . x_j\right)\alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i$$

Subject to constraints,

$$\alpha_i \geq 0 \text{ and}$$

$$\sum_{i=1}^{N} y_i \alpha_i = 0$$

D. Principal components evaluator

Principal component analysis is a statistical technique which reduces data dimensionality as a result of transforming original attribute space. Computing original attributes covariance matrix and extracting eigenvectors results in the formation of transformed attributes. The eigenvectors (principal components) define an original attribute space from a linear transformation to a new space where attributes are not correlated. Eigenvectors are ranked according to variations in the original data they account for. Generally, the first few transformed attributes account for most of the retained data variation with the rest being discarded. It should be noted that when comparing all attribute selection techniques, principal components proves to be the lone method which needs no supervision. In other words, it makes no use of the class attribute.

The PCA feature extraction method obtains new attributes through a linear combination of original attributes. Holding on to highest variance components achieves dimensionality reduction. Principal components number less than or are equal to original variables in numbers. This transformation is so defined that the first principal component has the biggest possible variance , accounting for as much variability as possible in the data. Thus each succeeding component in turn has the highest variance possible under an orthogonal constraint - uncorrelated to - preceding components. Principal components are independent only when data sets are jointly and normally distributed.

IV. Results and Discussion

A total of two hundred reviews with 100 positive opinions and 100 negative opinions are chosen in this work and their IDF computed. The proposed feature selectors are applied to perform the feature extraction. K Nearest Neighbour classifier is used for calculating the classification accuracy and the results are compared and shown in the table below Table1.

Table1.

Attribute Evaluator	% of Efficacy
Cfs Subset attribute evaluator	68
InfoGain attribute evaluator	75
SVM sttribute evaluator	70
Principalcomponents evaluator	95

The results are then depicted in the following figure.1.

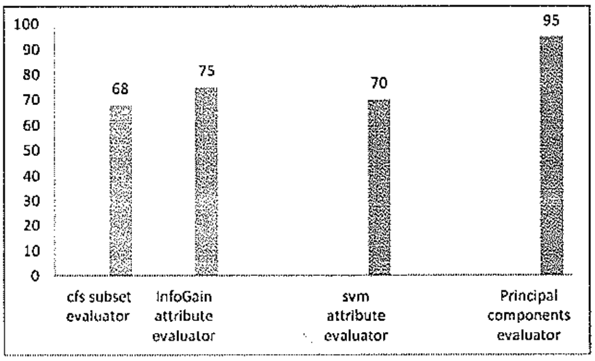

Figure1. Efficacy of Feature Selectors

V. Conclusion

In this paper, it is proposed to extract words from reviews and select words based on their importance using IDF. The feature set is reduced using the different types of feature selectors .K-Nearest Neighbour classifier is used and the classification accuracy is calculated. The work is more encouraging and as opinion mining is the fertile field of research, feature extractors on the large number of attributes and large scale multivariate data pertaining to opinion mining can be performed.

References

[1] A. L. Blum, P. Langley.Selection Election Of Relevant Features And Examples In Machine Learning. Artificial Intelligence, Vol.97, No. 1-2, pp. 245{271, 1997.

[2] G. H. John, R. Kohavi, K. P°eger. Irrelevant Feature and the Subset Selection Problem. In *Proceedings of the 11th International Conference on Machine Learning*, Morgan Kaufmann, New Brunswick, New Jersey, USA, pp. 121-129,1994.

[3] K. Kira, L. A. Rendell. The Feature Selection Problem:Traditional Methods and a New Algorithm. In *Proceedings of the 10th National conference on Artificial Intelligence*, MIT Press, San Jose, California, USA, pp. 129-134, 1992.

[4] R. Kohavi, G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence,* vol. 97, no. 1{2, pp. 273-324,1997.

[5] E. Leopold, J. Kindermann. Text Categorization with Support Vector Machines: How to Represent Texts in InputSpace? *Machine Learning*, vol. 46, no. 1, pp. 423{444, 2002.

[6] Y. Yang, J. O. Pederson. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the14th International onference on Machine Learning*, Morgan Kaufmann, Nashville, Tennessee, USA, pp. 412-420,1997.

[7] Y. Rui, T. S. F. Huang, S. Chang. Image Retrieval: CurrentTechniques, Promising Directions and Open

Issues. *Journal of VisualCommunication and Image Representation*, vol.10, no. 1, pp. 39{62, 1999.

[8] D. L. Swets, J. J. Weng. Efficient Content-based ImageRetrieval Using Automatic Feature Selection. In *Proceedings of IEEE International Symposium on Computer Vision*, IEEE Computer Society Press, pp. 85{90, 1995.

[9] H. Liu, H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic, Boston, USA,1998.

[10] P. Langley. Selection of Relevant Features in MachineLearning. In *Proceedings of AAAI Fall Symposium on Relevance*, AAAI Press, Menlo Park, California, USA, pp. 140-144, 1994.

[11] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, vol. 1, no. 3, 1997.

[12] J. R. Quinlan. Induction of Decision Trees. *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.

[13] J. R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1993.

[14] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. *Clasification and Regression Trees*, Wadsworth, Belmont, CA,1984.

[15] K. Kira, L. A. Rendell. A Practical Approach to Feature Selection. In *Proceedings of the 9th International Conference on Machine Learning*, Morgan Kaufmann, Aberdeen,Scotland, pp. 249-256, 1992.

[16] I. Kononenko. Estimating Attributes: Analysis and Extensions of RELIEF. In *Proceedings of Europe International Conference on Machine Learning*, Springer-Verlag,New York, USA, pp. 171{182, 1994.

[17] Pang B, Lee L. (2004) "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", *Proceedings of the ACL, 2004.*

[18] Turney P. (2002) "Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews", *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics – ACL*, 2002.

[19] Xiaowen Ding, Bing Liu, Philip S. Yu(2008). A holistic lexicon-based approach to opinion mining. WSDM '08 Proceedings of the international conference on Web search and web data mining.

[20] Wiebe J, Bruce R, Martin M, Wilson T, Bell M. (2004) "Learning Subjective Language", *Computational Linguistics*, Vol. 30, No. 3, pp. 277-308, January 2004.

[21] Pang B, Lee L. (2002) "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP,* 2002.

Author's Biography

J.Isabella is a Research Scholar from Sathyabamauniversity doing research in the field of Web mining and specializes in opinion mining. she received her Master of Computer Applications from Bharathidasan University,Trichy in 1995.She also received Master of Philosophy in Computer Science in 2008 from Bharathidasan University. She received **"Best Research Student Paper award"** in the 2nd annual International Conference on Software Engineering and Applications held at Singapore.