School of Informatics

Informatics Research Review
Evaluation of Feature Selection Methods for Text Classification

[redacted]

January 2021

Abstract

With the constant and exponential increase in the data available on the web in the recent past, text classification has attracted abundant research studies, particularly in the field of feature selection. Feature selection is the process of extracting a set of features which add most value to a system's projection, enabling access to useful information. Selecting a feature selection method which fits best with our text classification model is vital as it enhances accuracy and efficiency of the system. In this review, we will survey some of the evaluation techniques for feature selection methods.

Date: Thursday 21st January, 2021

Supervisor: [redacted]

1 Introduction

With the onset of the information age, information and data have become the driving force in the advancement of the civilization. Most of the data available around us is in the form of text documents, ranging from social media, news, literature, academic research etc. Searching through this plethora of textual data to retrieve useful information is paramount in numerous applications. This is facilitated by text classification: the process of assigning a predefined category to a textual script/document, enabling efficient search and indexing. Text classification is applied in various fields ranging from topic detection, e-mail filtering, web page classification, etc. Textual documents are transformed into a set of features, where a feature refers to a part of the document that is used to represent its content. This set is used to create data structures that enable fast searching.

Classification problems often focus on classifier speed and effectiveness. Given the volume of information available in the current digital world, creation of the set of features must be efficient, in terms of both time and space. Addressing the large number of terms/features and feature sparseness is an active research topic. Feature selection is the process of extracting a set of important features from the whole, which adds the most value to a system's projection, enabling access to useful information. The extraction is done so to exclude features which are unrelated or unnecessary without causing information loss, to attain reduction in dimensionality of feature space, increment in processing performance, improvement in accuracy of the classification model and to avoid over-fitting.

However, selecting a feature selection method for a classification task is not very straightforward. The ideal feature selection method for a classification task depends on various factors: the application, classifier being used, data sets, data volume, etc. Hence, evaluation of a feature selection method against all the other methods available becomes essential for the task in question. For this review, we will present a survey of techniques used to evaluate the superiority of a feature selection method. We will mainly focus on the novel MCDM (Multiple Criteria Decision-Making) technique for evaluation of feature selection methods for text classification published in Kou et al. 2020 and present a comparative commentary against the traditional single criteria evaluation techniques based on classification performance. However, we will not examine each of the single criteria techniques presented in our review. Instead, we will only comment on how the multi-criteria decision making method differs from it and its effectiveness. We will not talk about the differences in the evaluation techniques between the various single criteria techniques either, as our primary focus is multi-criteria decision making method's evaluation of feature selection methods. Also, we will only maintain our focus on filter feature selection methods.

This literature review first describes the popular evaluation metrics used by various single criteria (classification success) evaluation techniques in section 2.1. Section 2.2 will discuss some filter feature selection methods that our review will analyse. Subsequently, we will consider a couple of single criteria evaluation studies, multi-criteria decision making experiments, and draw comparisons between them in sections 2.3, 2.4 and 2.5, respectively. Section 3 presents a discussion of the entire review and deliberates on future work.

2 Literature Review

2.1 Evaluation Metrics

The most common criteria used to evaluate a feature selection method is classification performance. This is a quantity that can be measured using metrics like Accuracy, Recall, Precision, F-score etc.

2.1.1 Accuracy

Accuracy represents the fraction of documents that were rightly classified in the dataset, i.e. the number of correct predictions per total number of predictions. More formally, it is the degree of closeness of measurements of a quantity to that quantity's true value. It is calculated as:

$$\frac{TN + TP}{TP + FP + TN + FN}$$

where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

2.1.2 Precision

Precision is defined as the fraction of true positives that were correctly predicted, i.e the number of true positives correctly predicted per total number of elements labeled as belonging to the true class. Given by:

$$\frac{TP}{TP + FP}$$

2.1.3 Recall

Recall is defined as the fraction of relevant elements which were retrieved to the total relevant elements. Mathematically, it is the total number of true positives correctly predicted per total number of elements actually belonging to the true class:

$$\frac{TP}{TP + FN}$$

2.1.4 F-measure

Traditional f-measure or f1-score is the harmonic mean of precision and recall. Having two measures (precision and recall) to determine the effectiveness of a classification can be complex, hence, f1-score. It gives a combined score of precision and recall.

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Intuitively, the f-measure represents how precise and sensitive (recall) the performance of the classification is. F-measures can be extended to measure multi class classification problems. Macro averaging calculates the f-measure of each class in the collection and averages them. Alternately, micro f-measure is obtained by globally calculating the total number of true positives, false negatives and false negatives, and then calculating the f-measure.

As seen above, accuracy considers both true positives and true negatives equally. Thus, when working with extremely skewed data, accuracy measures may show high scores even if the number of irrelevant documents is marginally high, i.e., low positive rate sample. On the other hand, f1-score strikes a balance between precision (associated with positives) and recall (associated with false negatives) and does not consider true negatives.

2.2 Feature Selection Methods

Some of the popular filter feature selection methods like Chi-square (CHI), Gini Index (GI), Distinguished Feature Selector (DFS), Information Gain (IG) and Document Frequency (DF) are discussed in this section.

2.2.1 Chi-square-χ^2 (CHI)

χ^2 measure for classification indicates the co-relation between a word and a class. It gives a relative measure of the number of co-existing entries when they are dependant and independent, and normalizes this measure by the expected number. Formally, it is a measure of the difference between the observed and expected frequencies of the terms. Although, χ^2 is widely used, notably erroneous results are drawn when working with small datasets or with low frequency terms.

$$\chi^2(T,C) = \frac{N \times (TC \times NTNC - CNT \times TNC)^2}{(TC + CNT) \times (TNC + NTNC) \times (TC + TNC) \times (CNT + NTNC)}$$

where N = Total number of documents, TC = Number of times T and C occur simultaneously, TNC = number of times T appears without C, CNT = number of times C occurs without T and $NTNC$ = number of neither C nor T occurs.

2.2.2 GINI Index (GI)

The GINI Index is an improved form of the original metric used to measure the association (impurity) of an attribute for class. The improved form measures the purity of an attribute towards a category. It is given by:

$$GI(t) = \sum_{n=1}^{M} P(t|C_n)^2 P(C_n|t)^2$$

where $P(t|Ci)$ = Probability of term t for presence of class C_i and $P(Ci|t)$ = Probability of class C_i for presence of term t.

2.2.3 Distinguished Feature Selector (DFS)

DFS assesses the contributions of terms to the class discrimination in a probabilistic approach and assigns certain importance scores to them. In comparison to the other mentioned feature selection methods, DFS focuses on a feature's extent to distinctively distinguish a class instead of its distribution across the collection.

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(!t|C_i) + P(t|!C_i) + 1}$$

Uysal and Gunal 2012 present a detailed term analysis of the top-N features for DFS in comparison to IG, GI and CHI. For IG, GI and CHI many non informative terms were ranked quite high which DFS did not even consider as mid-ranked.

2.2.4 Information Gain (IG)

Information gain is one of the popular measures used in language processing. IG indicates how much information is gained about a class when a feature is observed, based on entropy. Hence, if a feature has high IG value with respect to a category, it can be concluded that the feature is more suitable for classification task.

$$IG(w) = H(C) - H(C|w)$$

where $H(C)$ = entropy of class C and $H(C|w)$ = Entropy of class C after observing word w.

2.2.5 Document Frequency (DF)

Document frequency measures the number of documents in the collection that have the occurrence of a term. It is one of the most simple and cost effective feature selection methods available. However, it often discards rare terms or terms with low document frequency, which may prove to be very informative. It is to be noted that document frequency only considers the presence of the term in a document and not its frequency. Further, document frequency may consider features which have no informative contribution over previously observed features.

2.3 Single Criteria Evaluation

According to the study conducted by Yiming and Jan O. 1997, CHI and IG far outperform other filter feature selection methods such as MI, DF and term strength (TS) in terms of classification accuracy, measured by precision. The study was performed using a kNN classifier and two corpora: the Reuters-22173 collection and OHSUMED collection. SMART system Salton 1989 was applied for unified pre-processing. They observe that CHI and IG display a similar performance in term selection. The identical performance between CHI and IG measures can be attributed to their emphasis on common and rare terms, although they share a bias towards common terms over rare terms, and also their use of categories. Moreover, Yiming and Jan O. 1997 note that DF displayed a high accuracy even though it does not rely on categories and could be considered as a replacement for CHI or IG in time expensive applications. The Rogati and Yang 2002 study presents a similar conclusion. Macro F1 and Micro F1 scores were used as metrics to measure accuracy. They used a kNN classifier as well and the Reuters-21578 collection as data set. Further, removal of low document frequency terms yielded improved results. As recognized in section 2.2.5, DF is a greedy method. It may select features with disregard to previously selected features and with no potential incremental information. Therefore, the above observation may not hold true for different applications and datasets.

Uysal and Gunal 2012 proposed the filter based probabilistic feature selection method, DFS, as having competitive if not better scores of classification accuracy in comparison to CHI, IG and GI. The study was done using three different classification algorithms: Decision Tree (DT), Support Vector Machine (SVM) and neural network (NN). Four distinct datasets with varying characteristics were used for the assessment: Reuters-21578, 20 Newsgroups A. Asuncion 2017, Short message service (SMS) message collection Almeida, Hidalgo, and Yamakami 2011 and

Enron1. F-measure is employed to measure classification accuracy. In contrast with CHI and IG which focus on common terms, they considered dissimilar terms, terms with high occurrence rate in only one or few classes, along with the common terms to calculate the DFS measure. These dissimilar terms act as a discriminating feature for a specific class and their consideration if these terms in the top-N features can be deemed apt. Further, Uysal and Gunal 2012 presents a dimension reduction and timing analysis. Dimension reduction analysis was done using a scoring scheme given in Gunal and Edizkan 2008. For timing analysis, computation time of importance score for a single term was considered. Both these analysis independently conclude DFS to be better in comparison with the other feature selection methods mentioned above.

2.4 Multiple Criteria Decision-Making (MCDM) Evaluation

Evaluation functions commonly face the problem of multiple contradicting criteria. Assessing these conflicting criteria in decision making forms a sub-discipline of operations research, namely, Multiple-Criteria Decision-Making (MCDM). This has become a popular research topic and is employed in applications to solve problems in the area of construction, energy, quality management, production management, algorithm evaluation etc.

Feature selection is a delicate task, having a great impact on the classification results. Different application configurations, such as different classifiers, data sets, parameters, and different algorithms dictate feature selection's performance. Thus, it becomes very hard to evaluate a feature selection method for a particular application. There is no conclusive evidence claiming the effectiveness of a feature selection method without strict constraints. All the research reviewed (Yiming and Jan O. 1997, Rogati and Yang 2002, Uysal and Gunal 2012) has a fixed configuration setup as shown in section 2.3.

Undeniably, text classification problems must be addressed with maximum classification success, measured with various metrics discussed in section 2.1. As important as classification performance is for classification problems, criteria like stability and efficiency become significant when real world classification applications are considered. For time and data sensitive applications in the field of trading, business analytics, bioinformatics, etc., effectiveness of classification is not sufficient. Criteria other than classification success is needed to fairly evaluate a feature selection method's success as text classification solution. Kou et al. 2020 present such a study, where parameters like stability and efficiency are presented in juncture with classification performance. Stability of a feature selection method refers to its robustness in feature choices. One of the earliest work found by us to have considered the stability of feature selection procedures was Kalousis, Prados, and Hilario 2005. They defined stability as follows:

> "We define the stability of a feature selection algorithm as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution $P(X, C)$. Stability quantifies how different training sets affect the feature preferences."

Typically, instability of a system's predictions does not invoke confidence for a system's utility in an application. Thus, ensuring a stable feature selection method widens the classification solution's utility and enhances productivity. Therefore, Kou et al. 2020 rightly note stability as a benefit criterion.

Further, efficiency of the feature selection method is taken into consideration by Kou et al. 2020 for evaluating feature selection. Kou et al. 2020 apply runtime, training time and test time to evaluate the efficiency of the feature selection methods. They present 5 MCDM measures

by categorizing the aforementioned criteria (classification performance, stability and efficiency) into benefit criteria and cost criteria and conclude a compromise method. Notably, Uysal and Gunal 2012 do provide a timing analysis in their study, but they fail to consider classification performance and efficiency (timing) criteria together. Moreover, Uysal and Gunal 2012 evaluate the computation time of importance score of a single term. In contrast, Kou et al. 2020 present a more practically operational (runtime, training time and test time) approach. Therefore, combining stability and efficiency with classification performance offers a more balanced evaluation.

2.5 MCDM vs Single Criteria Evaluation

Although abundant research has been done in single criteria feature selection methods, different studies have different configurations. Hence, in this section we present the data comparison between the study performed in the literature Uysal and Gunal 2012, discussed in section 2.3, against the MCDM literature Kou et al. 2020, discussed in section 2.4. Both of which use the Support Vector Machine (SVM) classifier and Reuters-21578 or 20 Newsgroups datasets.

Table 1 shows the juxtaposition of the data presented by single criteria evaluation study Uysal and Gunal 2012 against the MCDM evaluation of Kou et al. 2020 for Reuter21578 dataset. In Table 1 and Table 3, ranking for each evaluation technique/metric is mentioned. Kou et al. 2020 claim PROMETHEE multi-criteria decision making method to be most effective to their study. Therefore, PROMETHEE scores are presented in Table 1 and Table 3.

FS Method	Single Criteria				MCDM	
	Micro-F1	Rank	Macro-F1	Rank	PROMETHEE	Rank
CHI	0.8590	2	0.6426	4	0.2245	3
GI	0.8633	1	0.6591	1	-0.0928	4
DFS	0.8579	4	0.6493	2	0.2853	2
IG	0.8586	3	0.6492	3	0.3721	1

Table 1: FS method evaluation using SVM classifier and Reuter21578 dataset for 500 features.

Table 2 shows the comparison of the rankings given by 6 different MCDM methods for Reuter21578 dataset presented in Kou et al. 2020.

FS Method	TOPSIS ranking	VIKOR ranking	GRA ranking	WSM_N1 ranking	WSM_N2 ranking	PROMETHEE ranking
CHI	1	3	3	2	3	3
GI	4	4	4	4	4	4
DFS	2	2	2	3	2	2
IG	3	1	1	1	1	1

Table 2: MCDM rankings using SVM classifier and Reuter21578 dataset for 500 features.

Table 3 shows the juxtaposition of the data presented by single criteria evaluation study Uysal and Gunal 2012 against the MCDM evaluation of Kou et al. 2020 for 20-newsgroup dataset.

Table 4 shows the comparison of the rankings given by 6 different MCDM methods for 20-newsgroup dataset presented in Kou et al. 2020.

FS Method	Single Criteria				MCDM	
	Micro-F1	Rank	Macro-F1	Rank	PROMETHEE	Rank
CHI	0.9660	1	0.9623	3	0.0054	4
GI	0.9628	3	0.9628	2	0.1739	1
DFS	0.9618	4	0.9619	4	0.0749	3
IG	0.9632	2	0.9632	1	0.1176	2

Table 3: FS method evaluation using SVM classifier and 20-newsgroup dataset for 500 features.

FS Method	TOPSIS ranking	VIKOR ranking	GRA ranking	WSM_N1 ranking	WSM_N2 ranking	PROMETHEE ranking
CHI	2	4	3	2	4	4
GI	3	2	2	3	2	1
DFS	1	1	1	1	1	3
IG	4	3	4	4	3	2

Table 4: MCDM rankings using SVM classifier and 20-newsgroup dataset for 500 features.

3 Conclusion

3.1 Discussion

In Table 1 the f1-scores for single criteria evaluation on Reuter21578 dataset yields very similar scores. The delta between the scores for each FS method varies between 0.0004-0.0043 for micro f1-scores and between 0.0001-0.0098 for macro f1-scores, respectively. Similarly, in Table 2 for the 20-newsgroup dataset, the f1-score delta varies between 0.0004-0.0028 for micro f1-scores and between 0.0004-0.0005 for macro f1-scores. In contrast with the MCDM method, PROMETHEE scores produce higher values of delta between FS methods which vary between 0.0608-0.3173 for Reuter21578 and 0.0427-0.0563 for 20-newsgroup, respectively. Therefore, conclusion can be made that PROMETHEE metrics are more distinct and conclusive compared to the f1-scores for FS method comparisons. F1-scores are too similar to draw a clear result on the superiority of an FS method.

Moreover, Table 1 and Table 3 indicate that the ranking produced by the single criteria evaluation significantly differs from the MCDM rankings. For the Reuter21578 dataset shown in Table 1, single criteria evaluation considers GI superior to the other FS methods whereas PROMETHEE ranks IG at the top. According to Raileanu and Stoffel 2004, GI and IG have a very low disagreement frequency rate and share the same intuitive idea, i.e., calculation of impurity. Further, the accuracy metric in MCMD calculation for both GI and IG showed quite similar measures. Based on the experimental data presented in the literature Kou et al. 2020, the substantial difference between GI and IG's PROMETHEE score is attributed to the remarkable stability shown by IG. In fact, for the 20-newsgroup dataset in Table 3, similar rationale was seen behind CHI feature selection method's ranking. As discussed in section 2.4, MCDM evaluation considers multiple criteria with practical implications (stability and efficiency). Therefore, the feature selection methods which transcend classification performance and display a potential to enhance practical applications are ranked higher by the MCDM method.

Further, even though IG shows a high PROMETHEE score for the configuration presented in our review, Kou et al. 2020 recommend DF to be the preferred overall FS method. This is

based on its consistently good performance across various datasets and criteria. Remarkably, Yiming and Jan O. 1997 (discussed in 2.3) posits that IG and DF scores of a term are strongly correlated. Although DF is a task-free measure, it displayed an identical performance to IG, which is task-sensitive. They recommend DF as an alternative to IG when computation is considered expensive, concluding DF to be efficient and effective.

Table 2 and Table 4 show that different MCMD methods output different rankings. This is because of the way the MCDM method calculates its score. PROMETHEE scores are calculated based on relative measures. Instead of finding the best decision, it focuses on finding the best alternative. PROMETHEE is mainly used on applications related to ranking and priority. On the contrary, TOPSIS, VIKOR and GRA try to obtain an alternative closest to the ideal solution. Whereas, WSM applies additive weighting, it can be considered when all the criteria are normalized to same scale and have an equivalent weighting. Thus, MCDM method consideration depends on the field of application, for example, methods like PROMETHEE are useful in resource management, business analytics whereas methods analogous to TOPSIS are more useful in the field of medicine or bioinformatics.

3.2 Future Work

Conclusively, in comparison to single criteria evaluation, multi criteria decision making evaluation technique in text classification is more informative and effective. It has wider relevance and should be considered as an active research interest.

From the literature reviewed, evidently, no feature selection method outperforms in all aspects (criteria or datasets) and thus, no distinct feature selection can be concluded to be best performing. Instead, an interesting research area for future work would be to analyse multi-criteria decision making techniques for feature selection specific to a particular field of study. Based on the field of study, select multiple criteria which would most effect the applications in that area, and consider them to obtain an aggregate multi-criteria decision making score. For example, stability becomes vital in the field of molecular biology and linguistics in retrospection to areas like sentiment analysis.

Additionally, the review was done on small datasets. Realistically, feature sets are much larger and more dimensional. Conducting studies with more reasonable configuration would have more applicable outcomes.

References

A. Asuncion, D.J. Newman (2017). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Almeida, Tiago A., José Maria G. Hidalgo, and Akebo Yamakami (2011). "Contributions to the Study of SMS Spam Filtering: New Collection and Results". In: *Proceedings of the 11th ACM Symposium on Document Engineering*. DocEng '11. Mountain View, California, USA: Association for Computing Machinery, pp. 259–262. ISBN: 9781450308632. DOI: `10.1145/2034691.2034742`. URL: `https://doi.org/10.1145/2034691.2034742`.

Gunal, Serkan and Rifat Edizkan (2008). "Subspace based feature selection for pattern recognition". In: *Information Sciences* 178.19, pp. 3716–3726. ISSN: 0020-0255. DOI: `https://doi.org/10.1016/j.ins.2008.06.001`. URL: `http://www.sciencedirect.com/science/article/pii/S0020025508001850`.

Kalousis, Alexandros, Julien Prados, and Melanie Hilario (2005). "Stability of Feature Selection Algorithms". In: ICDM '05, pp. 218–225. DOI: 10.1109/ICDM.2005.135. URL: https://doi.org/10.1109/ICDM.2005.135.

Kou, Gang et al. (Jan. 2020). "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods". In: *Appl. Soft Comput. J.* 86, p. 105836. ISSN: 15684946. DOI: 10.1016/j.asoc.2019.105836.

Raileanu, Laura Elena and Kilian Stoffel (2004). "Theoretical Comparison between the Gini Index and Information Gain Criteria". In: pp. 77–93. DOI: 10.1023/B:AMAI.0000018580.96245.c6. URL: https://doi.org/10.1023/B:AMAI.0000018580.96245.c6.

Rogati, Monica and Yiming Yang (2002). "High-performing feature selection for text classification". In: *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 659–661. DOI: 10.1145/584792.584911.

Salton, Gerard (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0201122278.

Uysal, Alper Kursat and Serkan Gunal (2012). "A novel probabilistic feature selection method for text classification". In: *Knowledge-Based Syst.* 36, pp. 226–235. ISSN: 09507051. DOI: 10.1016/j.knosys.2012.06.005. URL: http://dx.doi.org/10.1016/j.knosys.2012.06.005.

Yiming, Yang and Pedersen Jan O. (1997). "A Comparative Study on Feature Selection in Text Categorization". In: *Proceeding ICML '97 Proc. Fourteenth Int. Conf. Mach. Learn.* 53.9, pp. 412–420. ISSN: 1-55860-486-3. arXiv: arXiv:1011.1669v3.