Information Gain Based Term Weighting Method for Multi-label Text Classification Task

Ahmad Mazyad, Fabien Teytaud, Cyril Fonlupt

Information Gain Based Term Weighting Method for Multi-label Text Classification Task

Ahmad Mazyad, Fabien Teytaud and Cyril Fonlupt

LISIC, Université du Littoral Côte d'Opale,
50 Rue Ferdinand Buisson, 62100 Calais - France

August 22, 2018

Abstract

In text classification, terms are given weights using Term Weighting Scheme (TWS) in order to improve classification performance. Multi-label classification task are generaly simplified into several single-label binary task. Thus, the term distribution are considered only in terms of positive and negative categories. In this paper, we propose a new TWS based on the information gain measure for multi-label classification task. This TWS try to overcome this shortness without affecting the complexity of the problem. In this paper, we examine our proposed TWS with eight well-known TWS on two popular problems using 5 learning algorithms. From our experimental results, our new proposed method outperforms other methods specialy regarding the macro-averaging measure.

1 Introduction

Text Categorization (TC) goal is to classify a text document into one or more categories. Generally, this approach is to learn an inductive classifier from a set of predefined categories. This approach requires that documents are represented in a suitable format such as the Vector Space Model (VSM) representation [10]. In a VSM, a document d_j is defined by a term vector $d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$ in which each term is associated with a weight $w_{k,j}$.

The weight represents the quantity of information a term contributes to the semantics of a document. The method which assigns a weight to a term is called TWS.

TC belongs to the family of supervised learning. Thus, TWS could be either unsupervised or supervised depending on whether it makes use of class information. The unsupervised methods include the famous $tf.idf$ proposed in [11] by Jones. $tf.idf$ stands for Term Frequency-Inverse Document Frequency, and it is borrowed from the information retrieval field.

The Supervised Term Weighting (STW) methods incorporate documents membership information when computing each term weight. These methods include the feature selection metrics such as χ^2, information gain ig, gain ratio gr, odds ratio or in [3], [4]. Recently, some authors proposed multiple intuition based methods. Wang *et al.* in [14] proposed the *inverse category frequency* icf proposed. In [6], Lan *et al.* presented *relevance frequency* rf.
A general formula for the different TWSs experimented in this paper could be defined as:

$$w_{t,d} = tf_{t,d} \times CF_t \ .$$

Where $w_{t,d}$ is the weight of a term t in a document d, $tf_{t,d}$ stands for the term frequency of t in d and CF_t is the collection frequency factor of t.

Different works performed on term weighting methods have shown different results and contradictory conclusions [6]. For instance, by comparing $tf.idf$ to three STW, Debole *et al.* in [3] showed the superiority of $tf.gr$ over $tf.ig$ and $tf.\chi^2$ while finding no consistent superiority over $tf.idf$. Another study done by Lan *et al.* in [6] confirmed the superiority of $tf.idf$ over $tf.chi$. A recent and fair comparaison beween state of the art TWS [7] have shown similar results as shown in [3]. However, In [4], Deng *et al.* concluded unlike Debole the superiority $tf.chi$ over $tf.idf$.

For this work, we seek to find an efficient TWS for multi-labeled classification task.

The paper is organized as follows: Section 2 presents the standard TWSs alongside with our proposed approach. In Sect. 3 we present 5 learning algorithms used in order assess the performance of TWSs. In Sect. 4, we compare the TWSs applied to two well-known data sets. Lastly, we consider future works in Sect. 5.

2 Term Weighting Methods

In this section, first, we present well-known TWS and second, our method.

2.1 Preliminary

Traditional classification algorithms are well suited for single label data sets. Thus, it can not learn from multi-labeled data sets. Several approaches exist to handle the multilabel classification task [13] such as problem transformation methods, and algorithm adaptation methods. Binary relevance transformation strategy is the most widely used strategy that simplifies the multi-labeled data set into several distinct single-label binary data set. That is, given the list of labels $L = \{l_1, l_2, ..., l_m\}$, the original data set is transformed into m different data sets $D = \{D_1, D_2, ..., D_m\}$. For each data set D_i, documents having the label l_i will be tagged as the positive category c_i, and the rest as the negative category $\overline{c_i}$. Weights are then computed independently for each binary data set. Based on the binary transformation, given a term t_k and a category c_i, STW could be expressed using statistical information a, b, c and d obtained from the

training data: a, b, c, d represents the number of documents that contain/do not contain t_k and belong/do not belong to the positive category c_i

These statistical information are used in all TWS included in our work, except for $tf.icf$.
In this paper, we logarithmically normalized the Term Frequency (tf) formula :

$$tf_{t,d} = \log(f_{t \in d}) + 1,$$

with $f_{t \in d}$ the number of occurrences of the term t in the document d.

2.2 Collection Frequency Factors

A Collection Frequency (CF) factor is a combination of statistical information. It is intended to measure the discriminative power of a term, i.e. it tells how much a term is related to a certain category. χ^2 corresponds to a test of independence between two variables (a term and a category). χ^2 is a popular feature selection method. χ^2 and other supervised feature selection schemes have been tested in several papers, as a term weighting methods for text categorization. For example, Deng *et al.* in [4], replaced the idf component with χ^2 component, claiming that $tf.\chi^2$ is more efficient than $tf.idf$. In contrast, in a similar test, Debole *et al.* in [3], compare tf.idf with three supervised term weightings, namely, χ^2, information gain and gain ratio. The authors have found no consistent superiority of these new term weighting methods over tf.idf.

Information Gain (ig) [1] is a measure of dependence between two random variables. In the context of text classification, it can be expressed as a measure of dependency between one random term and one random class. Mutual information is widely used in feature selection for text classification [3], [4].

Debole *et al.* used Gain Ratio (gr) applied to a feature selection method [3]. The authors claim that $tf.gr$ is a better term evaluation functions than the $tf.ig$. In their text categorization test, they confirmed the superiority of $tf.gr$ over $tf.ig$ and $tf.\chi^2$.

Odds Ratio (or) is a measure that describes the strength of association between two random variables. It was first used as a feature selection methods by Mladeni'c *et al.* [8] who found that odds ratio outperforms 5 other scoring methods studied in text classification experiments. Another comparative study on feature weight in text categorization is done by Deng *et al.* in [4]. The study shows a good performance of $tf.or$ but still outperformed by $tf.\chi^2$.

Relevance frequency (rf) is a supervised weight scheme proposed in [6]. rf measures the distribution of term t_k between positive and negative category and favors those terms that are more concentrated in the positive category than in negative category.

Inverse Category Frequency (icf) is another supervised term weighting method proposed by Wang *et al.* in [14]. icf stands for inverse category frequency and aims to favor terms that appear in fewer categories.

Main known CF factors are presented in tab. 1. We present some state of the art TWS in the next section.

Table 1: Seven traditional CF factors. N is the total number of docs, a the number of docs in the positive category *cat* that contain the term t_k, b defines the number of docs in *cat* with no occurrences of t_k, c is the number of docs not in *cat* in which t_k occurs at least once, d the number of docs that don't belong to *cat* and have no occurrences of t_k. $|C|$ reprensents the number of categories and $|C_{t_k}|$ is the number of categories that contain t_k.

CF	Formula
idf	$\log(\frac{N}{a+c})$
χ^2	$N \times \frac{(a\times d-b\times c)^2}{(a+c)(b+d)(a+b)(c+d)}$
ig	$(\frac{a}{N} \times \log \frac{a\times N}{(a+b)(a+c)}) + (\frac{c}{N} \times \log \frac{c\times N}{(c+d)(a+c)})$ $+(\frac{b}{N} \times \log \frac{b\times N}{(a+b)(b+d)}) + (\frac{d}{N} \times \log \frac{d\times N}{(c+d)(b+d)})$
gr	$ig/(-\frac{a+c}{N} \times \log \frac{a+c}{N} - \frac{b+d}{N} \log \frac{b+d}{N})$
or	$\log(2 + \frac{a\times d}{b\times c})$
rf	$\log\left(2 + \frac{a}{\max(1,c)}\right)$
icf	$\log_2\left(\frac{\lvert C\rvert}{\lvert C_{t_k}\rvert}\right)$

The TWS presented has proved to be efficient in text classification through a huge number of experimental studies. However, all these methods, except for $tf.icf$ has a common shortness: they consider the distribution only in terms of positive and negative categories.

2.3 Our Information Gain Based Method

The basic idea of our proposed ig based method comes in form of a question: how much information gain a term t_k have about a category after subtracting the information gain of the same term t_k of the other categories. It is to say that the higher the difference between a term information gain of one category and the average of the other categories, the more the term helps in separating positive and negative categories. As explained in sect. 2.1, a multi-label classification task is transformed into multiple binary single-label classification task, therefore, a term has multiple collection frequency weights, one for each binary task. Each weight only considers the distribution of a feature/term in terms of the positive category and the negative category (all documents that do not belong to the positive category). We think that using these weights could be helpful for more effective TWS.
Considering this idea, we propose a new TWS based on information gain. Its formula is defined by:

$$w'_{t,c} = w_{t,c} - (\mu_{c'\in C} w_{t,c'} + \sigma_{c'\in C} w_{t,c'}) \ .$$

Table 2: Comparison of the weighting values of ig and the proposed method. $\mu + \sigma$ is the average plus the standard deviation of scores of categories which are not the positive category. The values were hand-chosen.

Feature	ig	$\mu + \sigma$	New
t_1	0.3	0.5	-0.2
t_2	0.2	0.2	0
t_3	0.3	0.1	0.2

Where $w'_{t,c}$ is the new weight of a term t and a category c, $w'_{t,c}$ is the information gain score of a term t and a category c, $\mu_{c' \in C} w_{t,c'}$ is the mean of weights on all other categories, and $\sigma_{c' \in C} w_{t,c'}$ is the standard deviation of weights on all other categories.
To evaluate the differences between the information gain measure and our proposed method, let us consider the weights for the three terms in tab. 2.
First, let us clarify some points:

- When $\mu + \sigma > ig$, the term contributes more to the negative categories than to the positive category.
- When $\mu + \sigma < ig$, the term contributes more information to the positive category.
- When $\mu + \sigma = ig$, the term has about the same amount of information about both positive and negative categories.

First, considering the term t_1 in tab. 2, $\mu + \sigma$ (0.5) is higher than ig value (0.3), which means that the negative categories have higher weights than the positive category, however the ig value of t_1 is a positive value, in contrary to our new method. That said, the difference doesn't have a big impact on scores especially when the number of categories in the corpus is big, as $\mu + \sigma$ will have about the same value.
Now, if we consider terms t_1 and t_3, they both have the same ig value (0.3), which means that they both contribute the same amount of information to the positive category, however by looking at the values of $\mu + \sigma$, t_1 has a value of $0.5 > 0.3$ and t_3 has a value of $0.1 < 0.3$. In this case, we think that t_3 should have a higher value than t_1 as it has the same information gain in the positive category but smaller information gain in the negative categories.
Finally, t_2 has same information gain value both in the positive category and the negative categories $ig = \mu + \sigma = 0.2$, thus, the ig-based value is equal to 0.

3 Classifiers

Generally, the performance of a TWS is assessed on known benchmarks by evaluating a classification model on VSM representation of this TWS. In order

to build the classification models, we experiment 5 different algorithms, namely : Passive-Aggressive, C4.5, Support Vector Machine, Stochastic Gradient Descent and Nearest Centroid.

Support Vector Machine (SVM)s are a set of supervised machine learning methods introduced by Boser *et al.*. Developed from statistical learning theory, SVMs have shown good performance in many fields. In text classification, Joachims in [5] used SVM in which he demonstrates the better efficiency of SVM over other learning algorithms. Passive-Aggressive (PA) proposed by Crammer *et al.* in [2] is a learning algorithm focused on online learning and large scale data set. The method treats a flow of documents, and outputs a prediction once a document is received. Later at any time a document true label is discovered, the method redefines its prediction function . Stochastic Gradient Descent (SGD) classifier [15] is a linear classifier such as linear SVM, PA that uses SGD for training. This classifier is also used for large scale categorization problem. Nearest Centroid (NC) [12] is a neighborhood-based classification algorithm, and C4.5 (C4.5) [9] is a state of the art supervised learning algorithm based on decision tree.

4 Results and Discussion

In this study, we compare eight term weighting methods alongside with our approach on two popular data sets, i.e. Reuters-21578[1] and Oshumed[1] using 5 classification algorithms in terms of micro- and macro-averaged F_1 measure.

4.1 Data Corpora

Two widely-used datasets are used to compare the performance of our proposed method with the performance of eight well-known TWS: Reuters-21578 and Oshumed. Binary relevance transformation strategy is applied on the two multi-label classification task as explained in sect. 2.1. A default list of stop words, numbers and punctuation are removed. Lower case transformation is applied, and the Porter's stemming is performed.

4.1.1 Reuters-21578 Benchmark Corpus

This data set is a well-known benchmark for TC research. We use the "Apte-Mod" split [5]. The Apte split includes 10788 documents from the financial service, divided into a training set (7769 documents) and a test set (3019 documents). The data set is highly skewed, the smallest catgeory contains only 2 documents and the biggest contains 3964 documents.

[1] `http://disi.unitn.it/moschitti/corpora.htm`

Table 3: Statistics on the selected data sets used for our experiments (training/test).

	Reuters	Oshumed
number of documents	7769/3019	6286/7643
number of terms	26000	30198
number of categories	90	23
the smallest category	1/1	65/70
the largest category	2877/1087	1799/2153

4.1.2 Oshumed Benchmark Corpus

The second dataset is another well-known benchmark from the Oshumed[1] collection created by W. Hersh. the corpus is includes a total number of 13,929 medical abstracts splitted into a training subset of 6,286 abstracts and a test subset of 7,643 abstracts from the MeSH categories of the year 1991. Each document in this data set belongs to one or more categories from 23 cardiovascular diseases categories.

Table 3 presents statistics about the two datasets.

4.2 Evaluation

Numerous evaluation metrics exist to evaluate the classification models such as F_1 measure. The F_1 measure can be considered as a weighted average of the precision (the fraction of positive predicictions that is correct) and recall (the fraction of actual positives that have been correctly classified) and can be formally defined as:

$$F_1 = \frac{2 * recall * precision}{recall + precision} \ .$$

Generally, the F_1 measure is computed in two ways, micro-averaged and macro-averaged. In micro-averaged, big categories are emphasized while in macro-averaged, all categories have the same importance.

In the two tables, underlined results represent the highest score over a column, and the bolded results is the best pair of micro-/macro-averaged F_1 scores when all the classifiers and all the TWSs are considered. The pair having the highest mean is chosen as the best.

4.3 Results

Table 4 and Table 5 show the micro-/macro-averaged F_1 performances of different TWSs. using linear SVM for the two data sets Reuters and Oshumed, respectively.

Table 4: micro-/macro- averaged F_1 results (%) on Reuters-21578 corpus using eight standard TWSs and the proposed method.

	PA	SVM	SGD	NC	C4.5
tf	86.6/48.5	85.9/39.7	86.4/41.1	54.6/34.7	81.9/53.6
tf.χ^2	86.3/48.5	84.8/43.9	86.4/40.8	54.6/34.7	81.8/53.2
tf.idf	87.2/48.2	85.7/40.3	86.6/42.7	73.5/47.0	81.3/53.4
tf.gr	86.5/47.1	86.5/42.4	86.4/41.1	54.6/34.7	82.0/51.8
tf.or	86.6/48.5	87.3/48.6	86.3/40.8	54.6/34.7	81.9/52.8
tf.ig	86.9/47.7	86.5/42.4	86.3/41.0	54.6/34.7	82.1/54.2
tf.icf	85.9/46.4	84.0/37.8	85.0/40.3	62.5/46.4	80.9/52.0
tf.rf	86.5/46.7	87.8/45.3	86.4/40.8	54.6/34.7	82.0/52.8
New	**87.3/57.7**	88.7/51.7	88.4/49.0	66.5/54.7	82.2/51.3

Table 5: micro-/macro- averaged F_1 results (%) on Oshumed using eight standard TWSs and the proposed method.

	PA	SVM	SGD	NC	C4.5
tf	60.7/54.0	58.2/47.0	59.4/48.8	49.8/44.5	56.6/52.4
tf.χ^2	60.3/55.5	62.5/55.3	59.4/52.0	54.7/51.8	57.4/53.9
tf.idf	62.7/56.4	59.3/49.1	61.8/53.4	60.3/57.4	56.9/52.9
tf.gr	63.8/58.1	60.8/51.7	63.3/56.0	62.4/60.2	56.8/52.6
tf.or	65.2/62.4	64.9/58.8	66.0/60.6	59.2/57.4	56.6/52.8
tf.ig	63.7/58.1	60.8/51.7	63.2/56.0	62.4/60.2	57.1/53.5
tf.icf	56.5/51.3	49.3/41.9	54.6/48.3	59.1/55.2	56.5/52.7
tf.rf	64.0/60.1	63.4/55.5	64.4/57.2	58.4/56.0	56.6/53.0
New	64.4/60.8	67.0/60.8	**67.4/61.2**	59.4/57.0	56.7/52.5

Considering the Reuters data set, the best micro-averaged F_1 score 88.68% is achieved by using our method using SVM classifier. In terms of macro-averaged F_1 score, using our method gives the best score 57.70%. The best micro-/macro-averaged F_1 pair 87.29%/57.70% is also achieved by our information gain based method. Compared to the second best pair (87.27%/48.59%) achieved by $tf.or$, the proposed method records a boost of over 9% in terms of macro-averaged F_1. In terms of learning algorithms, in this experiment, PA, SVM and SGD show comparable performances. NC records the lowest results.
Considering the Oshumed data set, The highest micro-averaged F_1 (67.45%) is achieved using our proposed method. The highest macro-averaged F_1 (62.37%) achieved by using $tf.or$. As a pair of micro- and macro- averaged F_1, the

proposed method has a slightly higher average.
In terms of learning algorithms, SGD and SVM perform the best followed closely by PA and finally, NC and C4.5 show the lowest results.
Overall, in our study, we find that the proposed method give good results, better than the standard TWSs. $tf.or$, $tf.rf$, $tf.idf$ and $tf.ig$ have also shown good results. $tf.\chi^2$ and $tf.icf$ give the worst results.

5 Conclusion

For this work, we study a new term weighting scheme applied to multi-label text classification based on the information gain measure. The basic idea is that the information gain weight of a feature in negative categories should affect the importance of this term in the postive category.

We studied the effectiness of our method in comparaison to eight well-known TWSs applied to text classification tasks.

Experimental results show that our method outperformed all other methods tested in this study, specially in regard to the macro-averaged measure.

References

[1] Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2012)

[2] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research 7(Mar), 551–585 (2006)

[3] Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: Text mining and its applications, pp. 81–97. Springer (2004)

[4] Deng, Z.H., Tang, S.W., Yang, D.Q., Li, M.Z.L.Y., Xie, K.Q.: A comparative study on feature weight in text categorization. In: Advanced Web Technologies and Applications, pp. 588–597. Springer (2004)

[5] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. pp. 137–142. Springer (1998)

[6] Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(4), 721–735 (2009)

[7] Mazyad, A., Teytaud, F., Fonlupt, C.: A comparative study on term weighting schemes for text classification (2017)

[8] Mladeni'c, D., Grobelnik, M.: Feature selection for classification based on text hierarchy. In: Text and the Web, Conference on Automated Learning and Discovery CONALD-98. Citeseer (1998)

[9] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)

[10] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)

[11] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28(1), 11–21 (1972)

[12] Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences 99(10), 6567–6572 (2002)

[13] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview

[14] Wang, D., Zhang, H.: Inverse category frequency based supervised term weighting scheme for text categorization. preprint arXiv:1012.2609v4 (2013)

[15] Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: ICML 2004. pp. 919–926 (2004)