## 1. Effectiveness Analysis

### *1.1. Feasibility analysis*

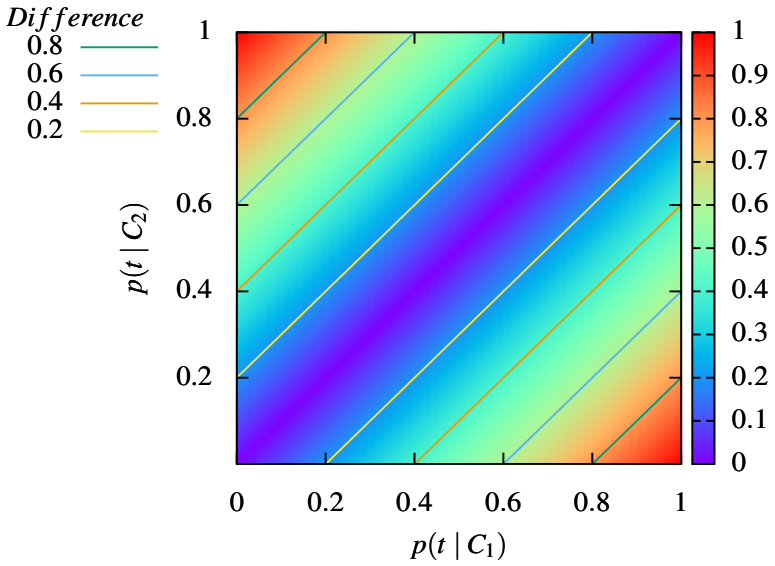



**Fig. 1** Contours for $Difference$.

As mentioned above, MDMC is related to the document rate difference. Here, we first show the feasibility of document rate difference. Let $Differnce = |p(t \mid C_1) - p(t \mid C_2)|$. Discrimination of terms could be measured using $Difference$. When $Difference = 0$ for a term, the corresponding term does not have discrimination because it occurs in both classes at the same frequency. If a term has $Difference = 1$, it has the strongest discrimination because it appears in all documents of only one class.

Fig. 1 shows contour lines and color distribution for $Difference$, where contour lines go parallel to the diagonal, and terms located on the same contour line have the same $Difference$. Because $Difference$ of terms on the diagonal line is zero, terms located around the diagonal line have a weak discrimination. From contour lines, we can observe that terms would have a strong discriminant ability if they are far away from the diagonal and close to two coordinate axes. Moreover, terms near the bottom-left are also known as sparse ones, terms near the top-right are common ones, and terms located in top-left and bottom-right corners are rare terms with a strong discrimination. On the basis of above observations, it is reasonable to measure the relevance of terms with classes using $Difference$. However, terms located on the same contour line can be of different discrimination. On the line of $Difference = 0.4$, for example, terms on the horizontal axis (rare terms) have a higher relevance than that far away from the horizontal axis even if they are on the same contour line. To solve this issue, existing methods (NDM, MMR and TCM) incorporate weights with document rate difference $Difference$. We follow this idea and design MDMC by taking the category information occupancy ratio as a component of weight.

On the basis of $Difference$, MDMC assigns the smallest weight of $\frac{1}{2}$ to terms with the same frequency in two classes, which are also on the diagonal line in Fig. 1 with $Difference = 0$. Terms located around the diagonal line are given weights that are close to $\frac{1}{2}$ by MDMC. Moreover, terms are of weights that gradually increase and get closer to 1 when they are far away from the diagonal and close to two coordinate axes. Additionally, $Difference$ can distinguish between sparse terms and rare ones. MDMC does not overestimate the importance of sparse terms while increasing the weight of terms that only appear in one class. In consequence, MDMC can enhance scores of rare terms and simultaneously depress the scores of sparse terms and common terms, which is beneficial for selecting terms with high discrimination.

### *1.2. Visualization for comparison*

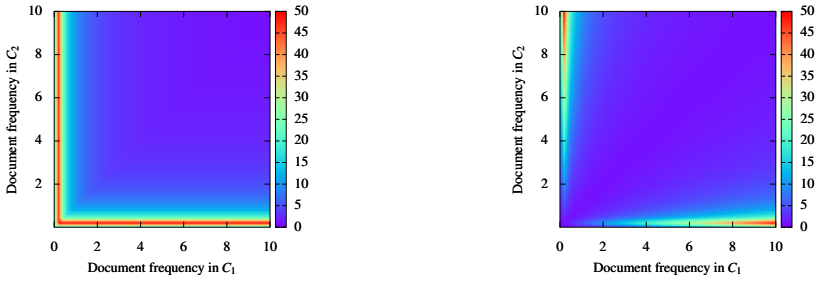
**Weight visualization** To vividly observe the feasibility of weight in MDMC, we consider the distribution of a term. Assume that there is two-class document data, and each class contains 10 documents. According to the distribution of this term, we can show its weight distribution obtained by difference-based metrics and then illustrate the characteristics of weights generated by these metrics. Here, let $k = 100$ for TCM, and the denominator is replaced with 0.05 for NDM and MMR when encountering zero, which follows the general setting in [1–3].

Fig. 2 shows the weight distribution of the term. Overall, we can find that the weights of MDMC and TCM fall into the interval $[0, 1]$, while those of NDM and MMR are not. From Fig. 2(a), it is observed that NDM does not give high weights to all possible rare ones but assigns high weights to some sparse terms. As an improved version of NDM, MMR solves the issue partially (Fig. 2(b)), but there still exists underestimation over rare terms compared with sparse ones. Similar to MMR, TCM cannot treat all possible rare terms correctly, as shown in Fig. 2(c). From Fig. 2(d), we can see that MDMC assigns relatively higher weights to all possible rare terms that are located above the bottom-right and around the top-left.

**Importance visualization** On the basis of the two-class document data above, we give the importance distribution of terms in Fig.3 after combining the different weights deduced by four metrics and $Difference$. Note that the importance distribution of a term is reflected by the color distribution. We can see that the important score of a term is dependent on its distribution between class $C_1$ and $C_2$. It is obvious that the working mechanism of four metrics in assigning important scores is distinct according to Fig. 3. By combining the category information occupancies with $Difference$, MDMC characterizes the importance of different term types in corpus more reasonably when compared with NDM, MMR and TCM. Specifically, MDMC can depress the scores of sparse terms and improve those of rare ones that are located above the bottom-right and around the top-left. Moreover, it is found that the importance scores given by MDMC and TCM fall into the interval $[0, 1]$, while those of NDM and MMR are not on account of their weights shown in Fig. 2(a) and Fig. 2(b). Thus, it is possible for MDMC to compare the discrimination of terms in different corpora.

### *1.3. An example*

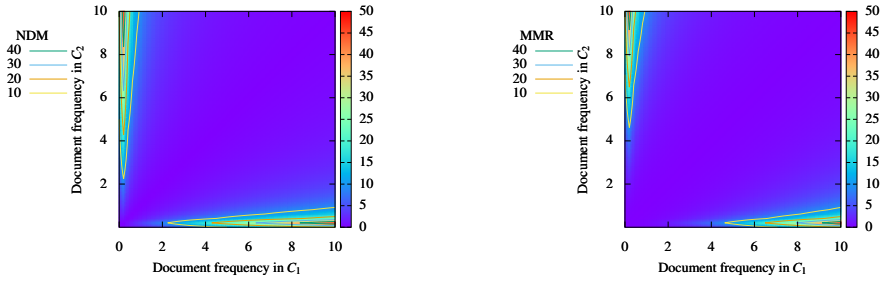
In this subsection, we give an intuitive understanding of MDMC through an example. Assume that we are given a text dataset, which contains two classes, where class $C_1$ has 1000




(a) NDM (b) MMR

(c) TCM (d) MDMC

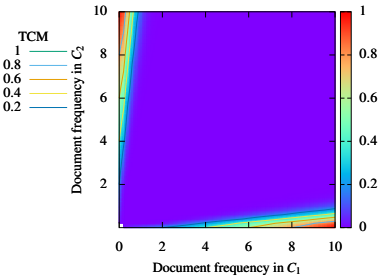
**Fig. 2** Weight distribution of a term obtained by four metrics.



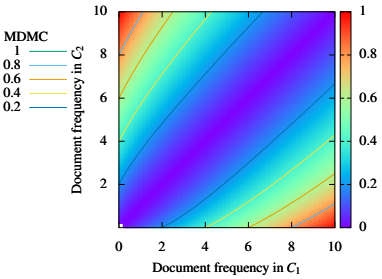

(a) NDM




(b) MMR




(c) TCM




(d) MDMC

**Fig. 3** Importance distribution of a term obtained by four metrics.

**Table 1**
Text data and feature scores obtained by difference-based methods.

| Terms | Document frequency | | $p(t \mid C_1)$ | $p(t \mid C_2)$ | Feature score | | | |
|---|---|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | | | NDM | MMR | TCM | MDMC |
| $t_1$ | 1000 | 0 | 1.00 | 0.00 | 1010.00 | 1010.00 | 1.00 | 1.00 |
| $t_2$ | 0 | 10 | 0.00 | 1.00 | 1010.00 | 1010.00 | 1.00 | 1.00 |
| $t_3$ | 1 | 10 | 0.00 | 1.00 | 999.00 | 999.00 | 1.00 | 1.00 |
| $t_4$ | 900 | 1 | 0.90 | 0.10 | 8.00 | 7.20 | 0.07 | 0.72 |
| $t_5$ | 900 | 2 | 0.90 | 0.20 | 3.50 | 3.15 | $3.57 \times 10^{-5}$ | 0.57 |
| $t_6$ | 900 | 8 | 0.90 | 0.80 | 0.13 | 0.11 | $8.09 \times 10^{-96}$ | 0.05 |
| $t_7$ | 800 | 7 | 0.80 | 0.70 | 0.14 | 0.11 | $2.00 \times 10^{-89}$ | 0.05 |
| $t_8$ | 1000 | 10 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $t_9$ | 1 | 1 | 0.00 | 0.10 | 99.00 | 9.90 | 0.10 | 0.10 |
| $t_{10}$ | 0 | 2 | 0.00 | 0.20 | 202.00 | 40.40 | 0.20 | 0.20 |
| $t_{11}$ | 4 | 1 | 0.00 | 0.10 | 24.00 | 2.40 | 0.07 | 0.09 |

documents, and class $C_2$ has only 10 ones. As shown in Table 1, there are 11 terms. According to the document frequency of these terms (the second and third columns of Table 1), they can be grouped into three categories, namely rare, common and sparse. Specifically, $t_1$, $t_2$, $t_3$, $t_4$ and $t_5$ are rare terms, while $t_6$, $t_7$ and $t_8$ are common terms. The rest three terms are sparse ones.

Table 1 also lists feature scores obtained by four difference-based methods: NDM, MMR, TCM, and MDMC. The goal of filter feature selection methods is to assign high scores to rare terms than irrelevant ones (sparse and common terms). It is observed from Table 1 that all methods can give high scores to rare terms than common terms. Both NDM and TCM overestimate the relevance of sparse terms because they assign higher ranks to sparse terms ($t_9$, $t_{10}$ and $t_{11}$) than rare terms ($t_4$ and $t_5$). Although MMR can address this issue to some extent, there still exists overestimation over sparse terms due to the choice of parameters, which is manifested in the fact that scores of sparse terms ($t_9$ and $t_{10}$) are higher than those of rare ones ($t_4$ and $t_5$).

Specifically, NDM and MMR give incorrectly high scores for $t_{10}$ because the weights of them are dependent on setting an appropriate value for the denominator. TCM also gives a high score to the sparse term $t_9$ on account of challenge setting for the parameter $k$. Compared with the three methods (NDM, TCM and MMR), MDMC correctly ranks rare terms ($t_4$ and $t_5$) whose scores are higher than those of sparse terms ($t_9$, $t_{10}$ and $t_{11}$) as shown in Table 1. Unlike the other three methods, MDMC without parameter can evaluate the true relevance for sparse and rare terms by combining the category information occupancy ratio and the document rate difference well.

## References


[1] A. Rehman, K. Javed, H. A. Babri, N. Asim, Selection of the most relevant terms based on a max-min ratio metric for text classification, Expert Systems with Applications 114 (2018) 78–96. doi:10.1016/j.eswa.2018.07.028.

[2] A. Rehman, K. Javed, H. A. Babri, Feature selection based on a normalized difference measure for text classification, Information Processing and Management 53 (2) (2017) 473–489. doi:10.1016/j.ipm.2016.12.004.

[3] K. Kim, S. Y. Zzang, Trigonometric comparison measure: A feature selection method for text categorization, Data & Knowledge Engineering 119 (2019) 1–21. doi:10.1016/j.datak.2018.10.003.