IJE TRANSACTIONS B: Applications Vol. 35, No. 08, (August 2022) 1516-1523

International Journal of Engineering

Journal Homepage: www.ije.ir

Multi-label Text Categorization using Error-correcting Output Coding with Weighted Probability

V. Balamurugan[a*], V. Vedanarayanan[a], A. Sahaya Anselin Nisha[a], R. Narmadha[a], T. M. Amirthalakshmi[b]

[a] *Department of ECE, Sathyabama Institute of Science and Technology, Oldmamallapuram Road, Chennai, India*
[b] *Department of Electronics and Communication Engineering, SRM Institute of Technology, Ramapuram, Chennai, India*

PAPER INFO

Paper history:
Received 07 September 2021
Received in revised form 25 March 2022
Accepted 27 March 2022

Keywords:
Text Categorization
Multi-label Classification
Multi-label Text Categorization
Error Correcting Output Coding
Posterior Probability

ABSTRACT

In several real-world categorization problems, labeled data is generally hard to acquire when there is a huge number of unlabeled data. Hence, it is very important to devise a novel approaches to solve these problems, thereby choosing the most valuable instances for labeling and creating a superior classifier. Several existing techniques are devised for the binary categorization issues, only a limited number of algorithms are designed for handling the multi-label cases. The multi-label classification problem turns out to be more complex when the sample belongs to multiple labels from the group of accessible classes. In World Wide Web, text data is generally present nowadays, and is an obvious example for such type of tasks. This paper develops a novel technique to perform the multi-label text categorization by modifying the Error-Correcting Output Coding (ECOC) approach. Here, a cluster of binary complimentary classifiers are employed to facilitate the ECOC more effective for the multi-class problems. In addition, a weighted posterior probability is computed to enhance the multi-label text classification performance more effectively. Moreover, the performance of the proposed ECOC with weighted probability is analyzed using the performance metrics, like precision, recall, and f-measure with maximal precision of 0.897, higher recall value of 0.896, and maximum f-measure of 0.895.

doi: *10.5829/ije.2022.35.08b.08*

1. INTRODUCTION

Text categorization, also named as document categorization, play a vital part in various applications based on the Natural Language Processing (NLP) and Information Retrieval (IR)-based systems. Text categorization is fundamental and classical problem in NLP areas in which it has been used for a variety of applications, like pattern recognition, statistics, and machine learning. In text classification applications are grouped into automatic indexing for Boolean IR systems, hierarchical classification of web pages, word sense disambiguation, document filtering and arrangement [1-3]. The internet contains infinite number of text documents. This massive quantity of data poses a key challenge yet for simpler jobs, like IR. A feasible way is to organize textual data into various groups. Automatic document categorization is applied for discovering the fundamental document information in an automatic manner [4, 5], thereby saving human attempts and computational time. In addition, the documents are assigned to the pre-determined category groups with respect to the contents [6]. Text categorization divides the texts into various categories with respect to its content, topic, and attributes, and has been remains as a major issue. Moreover, multi-label text categorization is to assign multiple category labels to every text, and this technique is commonly employed in applications, such as information retrieval, sentiment analysis, news subject organization, and spam recognition [7-9].

Multi-class categorization is the process of classifying the unknown objects into numerous pre-determined classes. In general, multi-class categorization techniques are categorized into two different groups. The

*Corresponding Author Institutional Email: balamurugan.ece@sathyabama.ac.in (V. Balamurugan)

Please cite this article as: V. Balamurugan, V. Vedanarayanan, A. Sahaya Anselin Nisha, R. Narmadha, T. M. Amirthalakshmi, Multi-label Text Categorization using Error-Correcting Output Coding with Weighted Probability, ***International Journal of Engineering, Transactions B: Applications***, Vol. 35, No. 08, (2022) 1516-1523

first group is the direct multi-class categorization technique, which contains approaches, namely Multi-class Support Vector Machine (SVM), Neural Network, k-Nearest Neighbors (kNN), Naive Bayes (NB), Classification and Regression Trees (CART), Convolutional Neural Network (CNN) [9-12], and so on. On the other hand, the second one is the indirect technique, where the multi-class problem is decomposed into a cluster of binary sub-problems [13-16]. There are two different techniques utilized to perform the multi-label categorization. The first one is the problem transformation techniques where the multi-label problems are transformed into several single-label problems, whereas the second one is the algorithm adaptation techniques for handling the multi-label processing directly. Nowadays, multi-class classification remains as a major challenge because of the class imbalance problems and class overlapping problems [17]. ECOC has been utilized in the multi-label learning, and has shown an efficient performance [18]. ECOC algorithm includes two main phases, such as encoding and decoding. The encoding procedure decomposes the multiclass problem into a cluster of binary problems such that decomposition mechanisms are recorded as a code matrix with every column indicating a binary problem [19, 20]. Using ECOC, multiclass problem is partitioned into multiple binary class problems in such way that the binary class problem re-labels the original classes to either positive or negative groups, signifies by a column [21].

The major objective of this research is to devise an ECOC framework with the weighted posterior probability function. Here, the ECOC is utilized to generate the soft labels for each data, which can either categorize the data through the threshold label directly or to rank every label for each data. With respect to the soft labels after categorization, the diversity and the uncertainty of every data is obtained. In addition, the posterior probability computation makes the multi-label classification tasks more reliable. Proposed ECOC with weighted probability: An effective multi-label text categorization approach is devised using proposed ECOC with weighted probability. Here, the ECOC is used for incorporating multiple binary classifiers, thereby handling the multi-class classification problems. Furthermore, weighted posterior probability is computed for every class to enhance the categorization performance.

The structure of the research paper is designed as follows. The literature survey of several existing ECOC techniques is reviewed in section 2. Section 3 portrays the developed ECOC with weighted probability for the multi-label text categorization. The results and discussion of developed method is portrayed in section 4, and the paper is concluded in section 5

2. MOTIVATION

The existing ECOC techniques are illustrated in this section along with the merits and demerits. In addition, the challenges faced by the ECOC techniques during multi-label text categorization are also described as follows.

2. 1. Literature Review This section reviews the various existing ECOC approaches along with their drawbacks. Kajdanowicz et al. [1] developed an Extension of the ECOC algorithm named ML-ECOC for the multi-label text categorization. This method utilized a set of binary complimentary classifiers in order to handle the multi-class issues. This technique effectively reduced the computational cost and complexity, but this method failed to obtain accurate text categorization. Shan et al. [2] developed a Randomized Multi-Label Sub problems Concatenation (RMSC) technique for the multi-label classification problems. Using this method, the imbalance issues were tackled in such a way that the diversity between the classifiers was enhanced. In this technique, time complexity was very much reduced, but the data subsets utilized in automatic web page classification tasks are more challenging to learn.

Jin et al. [3] introduced an Image-driven wafer map defect pattern classification method (WMDPC). This technique was comprised with two various phases, such as feature extraction and classification. This method obtained better generalization ability. However, this method failed to improve the performance of the system. Gu et al. [4] devised a Multi-class active learning algorithm for the multi-class classification problems. Here, a codeword was created for every class, and then a test code was created for the labeled instances for addressing the multi-class cases. This algorithm effectively reduced the computation complexities, thereby enhanced the classification accuracy, but the overall training time was high in this technique.

2. 2. Challenges The challenges faced by different existing ECOC techniques for the text categorization are illustrated as follows,

The RMSC model was developed by Shan et al. [2] for the multi-label classification problems, but this method failed to learn the binary classifiers with their relationships together for enhancing the performance diversity.

WMDPC method was introduced by Bui et al. [11] for multi-class classification. However, the major challenge lies in utilizing various decomposition mechanisms, and other types of binary classifiers for improving the classification accuracy.

Nazari et al. [21], Multi-class active learning algorithm is devised for addressing multi-class classification problems. However, this approach failed to

consider batch mode active learning for incorporating the sample diversity standards, and the classifier uncertainty for enhancing the performance.

Multi-label active learning was introduced by Qin et al. [22] for determining the label instances on every class by hybridizing the multiple classifiers, but the main challenge lies in incorporating the Revisiting ECOC (RECOC) classification with other active learning mechanisms for improved performance results.

3. PROPOSED ECOC WITH WEIGHTED PROBABILITY

This section illustrates the proposed ECOC with weighted probability for the multi-label categorization tasks. The major steps involved in the process of text categorization are elucidated as follows: Here, the modification is carried out in the coding and the decoding steps of the standard ECOC algorithm in such a way that the process is more appropriate for multi-label classification issues. This modification involves creating novel rules in the steps of coding and decoding order to eliminate the inconsistency issues, while managing the multi-label data. Moreover, ECOC is the classifier-driven ensemble approach [1], which is encouraged by the transmission of signals based on information theory. It is utilized for transmitting and receiving the data in a safe and effective way. A set of binary complimentary classifiers were used in such a way that ECOC applications are considered to be more effective for solving the multi-class problems. Furthermore, a posterior probability of every class is computed with respect to the weighted function for the efficient labeling process.

3. 1. Error-Correcting Output Coding ECOC is a classifier ensemble technique motivated by the transmission of the signals in the information theory, which is utilized for transmitting and receiving the data. The error-correcting ability is used for recovering the errors occurred in every categorization level of sub problems. In addition, ECOC has the advantage of decomposing the multi-class problems into binary sub-problems, called dichotomies in the concept-based on machine learning. Here, all sub-problem is solved by the dichotomizer, such that the final solution necessary for multi-class problem is obtained by incorporating the outcomes obtained by the dichotomizers based on the divide-and-conquer rule. Moreover, ECOC executes well mainly on the inconvenience with several classes, whereas the other kinds of classifiers generally have complications.

Let us consider the categorization problem with M_k classes where the ECOC is utilized for generating a binary and ternary code for every class. The code matrix N is used for organizing the code words in the form of matrix rows, where $\mathrm{N} \in \{-1,0,+1\}^{\mathrm{M_k} \times C}$, and C represents the length of the code (coding step). Based on the learning perspective, N signifies M_k classes for training C dichotomizers, $g_1 \cdots g_C$. The training of the classifier g_c is performed with respect to the column $N(.,c)$. If $N(j,c)=+1$, then the instances of the class j are positive, whereas if $N(j,c)=-1$, then every instances are negative supper-class. If $N(j,c)=0$, then no instances of class j participate in order to train g_c.

Let $\bar{x}=[x_1 \ldots x_C], xc \in \{-1,+1\}$ be considered as the vector output of the C classifier ensemble for the input y. In the step of decoding, the output class maximizes similarity measure S among $\bar{x}$ and row $N(i,.)$ is chosen,

$$ClassLabel = ArgMax S(\bar{x}, N(i,.)) \quad (1)$$

The matrix of ECOC codifies the class labels for obtaining different class partitions measured by every dichotomizer. The strategy for coding can be partitioned into problem-dependent and problem-independent. The most common pre-defined constructions of the problem-independent codeword meet the requirements based on high separabaility among the columns and rows for improving the error-correcting ability, and the diversity among the dichotomies.

3. 2. Multi-label ECOC for Text ECOC decomposes multi-class problems into some binary sub-class problems [1]. The ECOC approach generally consists of three main steps, namely coding, binary classifier learning, and decoding. The encoding step maps all the class to the codeword, which includes the results of the decomposed binary problems on that class. After that, the set of binary classifiers are trained using the several partitions of the original data with respect to every column of coding matrix [23]. Once binary classifier learning is performed, a new instance is assigned for the classes using decoding on the basis of the trained binary classifier outputs, and the code matrix rows. In order to perform the encoding process, various binary coding designs, and the ternary coding designs are devised. In the binary coding design, the code words are $+1$ and -1, whereas in ternary coding design, the code words are +1, 0, and -1. A classifier is defined based on every column of the multi-label matrix, which is utilized for computing the membership degree of *d* into super-class, which consists of numerous categories. The dichotomizing procedure with some inconsistencies is eliminated by defining only neutral set and positive class, which does not containsany type of area for overlapping.

3. 3. Proposed ECOC with Weighted Posterior Probability Let us consider a predicted codeword $\bar{x}_d = [\bar{x}_1 ... \bar{x}_C], 0 \le \bar{x}_c \le +1\}$, and it is a string assigned for the document d where each bit signifies the output of the classifier, that is $P_c(+1|d)$. The posterior probability of every class is computed as follows:

$$P(k_M|d) = \frac{1}{|N(M,.)|} \sum_{c=1}^{C} P_c(+1|d) N(M,c) \frac{E_c}{n} \quad (2)$$

where, E_c signifies mean square error of training algorithm of classifier, n represents normalizing factor. For every document, ECOC arranges class by score and assigns YES to every t top-ranking categories. t is an integer ranging from 1 to the number of categories.

4. RESULTS AND DISCUSSION

The results and discussion of developed ECOC with weighted probability by considering the various performance metrics is illustrated in this subsection.

4. 1. Experimental Setup The implementation of developed ECOC with weighted probability is done in MATLAB tool using Reuters database [24] and rcv1data[2].

4. 2. Dataset Description Reuters database [20] is a text Categorization collection data set donated by David D. Lewis. Here, the documents are structured and indexed based on their categories. In this database,the total number of instances is 21578 with five attributes. The total number of webhits -obtained by the dataset is 163417.

4. 3. Performance Evaluation Metrics This section illustrates the various evaluation metrics, namely as precision, recall, and f-measure utilized for performing the assessment of the technique

Precision: It is a measure used for defining the fraction of appropriately classified texts, and the equation is formulated as follows:

$$P = \frac{|T_r \cap T_c|}{T_c} \quad (3)$$

Here, T_r signifies relevant texts, and T_c signifies the categorized texts.

Recall: It is measure of relevant texts present in classified texts that is related to text data.

$$R = \frac{|T_r \cap T_c|}{T_r} \quad (4)$$

F-measure: It computes the mean difference among the precision and the recall measure and is given as follows:

$$FM = 2 * \left(\frac{P * R}{P + R} \right) \quad (5)$$

4. 4. Comparative Techniques The various techniques, such as ML-ECOC [1], RMSC [19], WMDPC [25], dynamic semantic representation model and deep neural network (DSRM-DNN) [8], and Multi-relation Message Passing (MrMP) [24] are utilized for performing the comparative analysis of the developed ECOC with weighted probability.

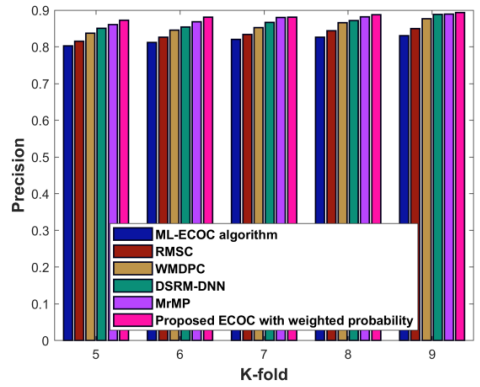
4. 5. Comparative Analysis The comparative assessment of developed ECOC with weighted probability based on the evaluation metrics, like precision, recall, and f-measure with respect to Reuters database and rcv1 data is explained in this section.

4. 5. 1. Analysis using Reuter Database

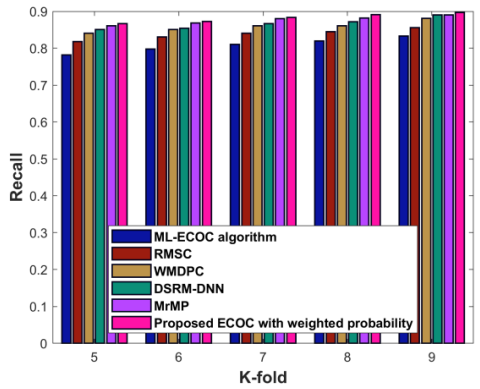
4. 5. 1. 1. Analysis using Training Data Figure 1 illustrates the analysis using training da data with respect to precision, recall, and f-measure. The assessment using precision measure is portrayed in Figure 1(a). By considering the training data as 50%, the precision value achieved by the developed ECOC with weighted probability is 0.832, whereas the existing techniques, such as ML-ECOC, RMSC, WMDPC, DSRM-DNN, and MrMP achieved a precision value of 0.742, 0.716, 0.755, 0.760, and 0.792, respectively. Figure 1(b) presents the assessment using recall metric. The recall value obtained by the ML-ECOC is 0.742, RMSC is 0.775, WMDPC is 0.795, DSRM-DNN is 0.810, MrMP is 0.822, and the developed ECOC with weighted probability is 0.845 for the training data 70%. Figure 1(c) shows the analysis using f-measure. When the training data is 60%, the developed ECOC with weighted probability measured an f-measure value of 0.832, whereas the f-measure value obtained by the existing techniques ML-ECOC is 0.737, RMSC is 0.734, WMDPC is 0.761, DSRM-DNN is 0.783, and MrMP is 0.799.

4. 5. 1. 2. Analysis using K-fold The k-fold analysis for the developed approach with respect to precision, recall, and f-measure is presented in Figure 2. Figure 2(a) presents the analysis using precision metric. For the k-fold value 5, the developed ECOC with

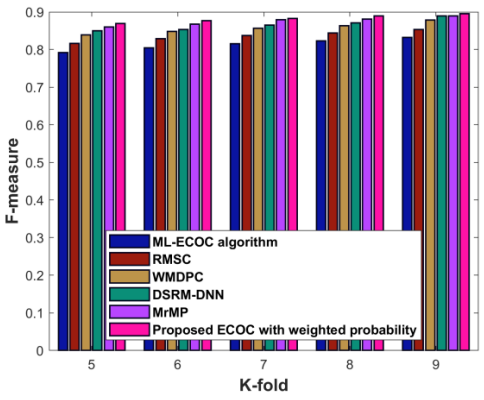
[2] https://www.kaggle.com/kerneler/starter-rcv1data-1c5d94f9-d/data



(a)

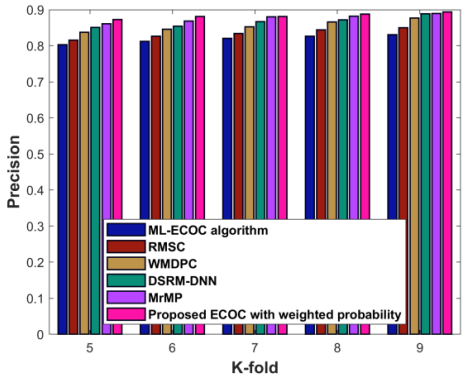


(b)

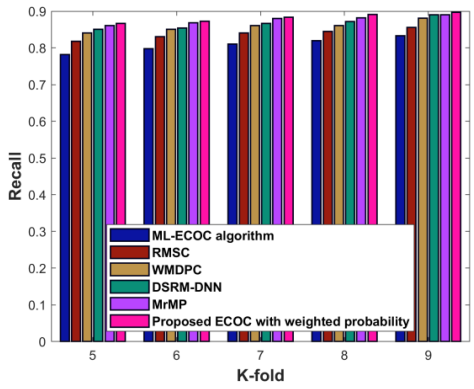


(c)

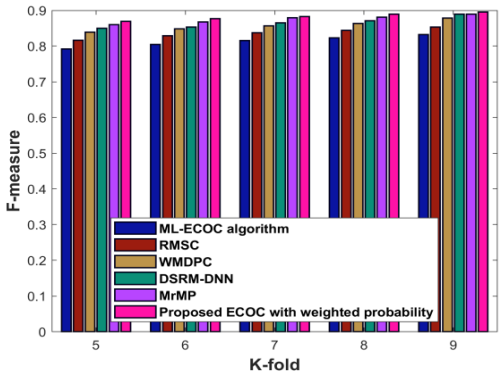
Figure 1. Analysis of the developed technique using a) Precision b) Recall c) F-measure



(a)



(b)



(c)

Figure 2. Analysis of the developed technique using a) Precision b) Recall c) F-measure

weighted probability obtained a precision value of 0.872, while the existing techniques such as ML-ECOC, RMSC, WMDPC, DSRM-DNN, MrMP achieved a precision value of 0.803, 0.816, 0.837, 0.851, and 0.861, respectively. The assessment using recall metric is portrayed in Figure 2(b). The recall value obtained by the ML-ECOC is 0.811, RMSC is 0.841, WMDPC is 0.861, DSRM-DNN is 0.866, MrMP is 0.880, and the developed ECOC with weighted probability is 0.884 for the k-fold value 7. The f-measure analysis is depicted in Figure 2(c).

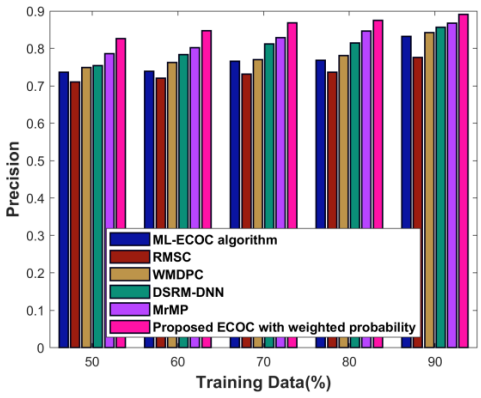
By considering the k-fold value as 8, the f-measure value achieved by the developed ECOC with weighted probability is 0.889, whereas the f-measure value measured by the existing ML-ECOC is 0.823, RMSC is 0.844, WMDPC is 0.863, DSRM-DNN is 0.871, and MrMP is 0.881.

4. 5. 2. Analysis using Rcv1data

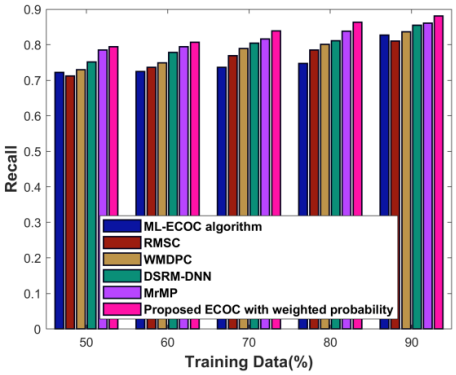
4. 5. 2. 1. Analysis using Training Data Figure 3 illustrates the analysis using training data using rcv1

data. The assessment using precision measure is portrayed in Figure 3(a). By considering the training data as 50%, the precision value achieved by the developed ECOC with weighted probability is 0.827, whereas the existing techniques, such as ML-ECOC, RMSC, WMDPC, DSRM-DNN, and MrMP achieved a precision value of 0.736, 0.710, 0.749, 0.754, and 0.786, respectively. Figure 3(b) presents the assessment using recall metric. The recall value obtained by the ML-ECOC is 0.737, RMSC is 0.770, WMDPC is 0.790, DSRM-DNN is 0.804, MrMP is 0.817, and the developed ECOC with weighted probability is 0.839 for the training data 70%. Figure 3(c) shows the analysis using f-measure. When the training data is 60%, the developed ECOC with weighted probability measured an f-measure value of 0.827, whereas the f-measure value obtained by the existing techniques ML-ECOC is 0.732, RMSC is 0.729, WMDPC is 0.756, DSRM-DNN is 0.777, and MrMP is 0.794.

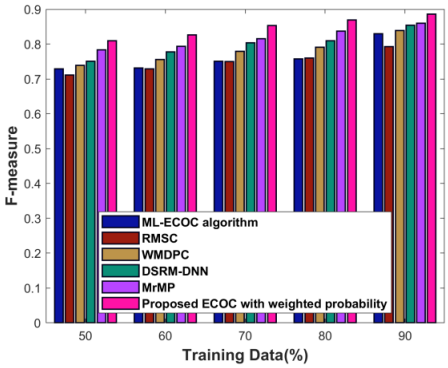
4. 5. 2. 2. Analysis using K-fold The k-fold analysis for the developed approach using rcv1 data is presented in Figure 4. Figure 4(a) presents the analysis using precision metric. For the k-fold value 5, the developed ECOC with weighted probability obtained a precision value of 0.867, while the existing techniques such as ML-ECOC, RMSC, WMDPC, DSRM-DNN,



(a)

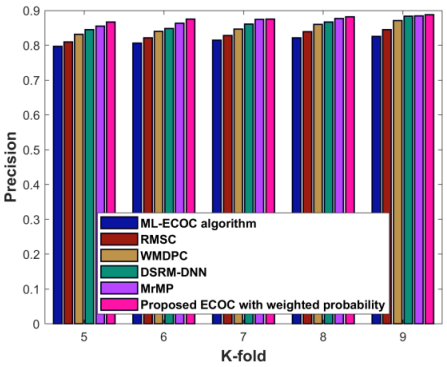


(b)

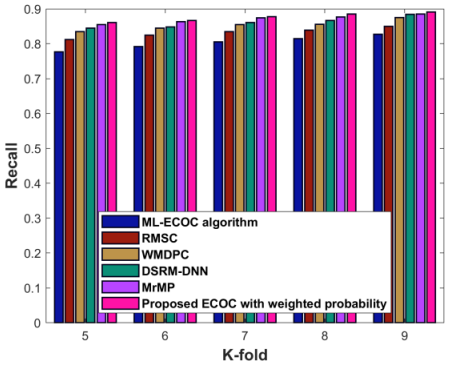


(c)

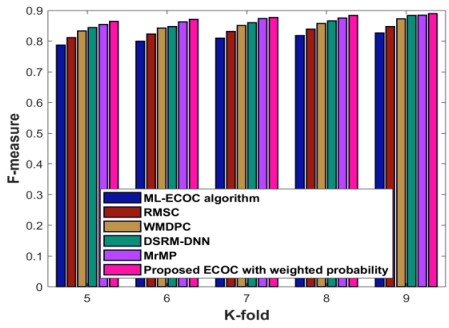
Figure 3. Analysis of the developed technique using a) Precision b) Recall c) F-measure



(a)



(b)



(c)

Figure 4. Analysis of the developed technique using a) Precision b) Recall c) F-measure

MrMP achieved a precision value of 0.797, 0.810, 0.832, 0.845, and 0.855, respectively. The assessment using recall metric is portrayed in Figure 4(b). The recall value obtained by the ML-ECOC is 0.805, RMSC is 0.835, WMDPC is 0.855, DSRM-DNN is 0.861, MrMP is 0.875, and the developed ECOC with weighted probability is 0.878 for the k-fold value 7. The f-measure analysis is depicted in Figure 4(c). By considering the k-fold value as 8, the f-measure value achieved by the developed ECOC with weighted probability is 0.884, whereas the f-measure value measured by the existing ML-ECOC is 0.818, RMSC is 0.839, WMDPC is 0.858, DSRM-DNN is 0.866, and MrMP is 0.876.

4. 6. Comparative Discussion This section illustrates comparative discussion of developed ECOC with weighted probability in comparison with various existing techniques, such as ML-ECOC, RMSC, and WMDPC with respect to the performance metrics, namely precision, recall, and f-measure. Table 1 portrays the assessment of developed approach with respect to the training value 90%, and k-fold value 9. The precision value achieved by the developed ECOC with weighted probability is 0.897, whereas the existing techniques, such as ML-ECOC, RMSC, WMDPC, DSRM-DNN, and MrMP achieved a precision value of 0.837, 0.781, 0.847, 0.863, and 0.873. The recall value measured by the ML-ECOC is 0.833, RMSC is 0.855, WMDPC is 0.880, DSRM-DNN is 0.860, and MrMP is 0.866, while the developed techniques ECOC with weighted probability measured a recall value of 0.896. The developed ECOC with weighted probability measured an f-measure value of 0.895, whereas the existing techniques, such as ML-ECOC, RMSC, WMDPC, DSRM-DNN, and MrMP achieved an f-measure value of 0.832, 0.852, 0.878, 0.889, and 0.89. From the table it is clearly shown that, the developed ECOC with weighted probability achieved a maximal precision of 0.897 using training data, higher recall of 0.896 using k-fold, and maximum f-measure of 0.895 using k-fold.

TABLE 1. Strouhal number for different geometric cases

	Metrics	ML-ECOC	RMSC	WMDPC	Proposed ECOC with weighted probability
Using training data (%)	***Precision***	0.837	0.781	0.847	**0.897**
	Recall	0.832	0.816	0.842	**0.886**
	F-measure	0.835	0.798	0.845	**0.891**
Using K-fold	***Precision***	0.831	0.850	0.876	**0.893**
	Recall	0.833	0.855	0.880	**0.896**
	F-measure	0.832	0.852	0.878	**0.895**

5. CONCLUSION

This research work designs a multi-label text categorization technique, combining a robust ECOC classification technique with a posterior probability computation. Here, the inconsistency in the coding phase is eliminated by utilizing the proposed ECOC as this approach decomposes the various multi-class problems into a few balancing one-class sub problems. In addition, the multi-label association is considered in the phase of testing by utilizing the decoding mechanism implemented for the ECOC algorithm. Finally, the posterior probability is computed as a weighted function for improving the performance efficiently. The advanced performance of the developed technique is demonstrated by the experiments performed using the Reuters database. The proposed ECOC with weighted probability outperforms the various existing multi-label text classification techniques by considering the various evaluation metrics, namely precision, recall, and f-measure. However, the developed ECOC with weighted probability achieved efficient performance with the maximum precision value of 0.897, higher recall value of 0.896, and maximum f-measure value of 0.895. The future work would be the concern of developing more novel classifiers to enhance the performance effectiveness of multi-label text categorization.

6. REFERENCES

1. Kajdanowicz, T. and Kazienko, P., "Multi-label classification using error correcting output codes", ***International Journal of Applied Mathematics and Computer Science***, Vol. 22, No. 4, (2012), 829-840.
2. Shan, J., Hou, C., Tao, H., Zhuge, W. and Yi, D., "Randomized multi-label subproblems concatenation via error correcting output codes", ***Neurocomputing***, Vol. 410, (2020), 317-327, doi: 10.1016/j.neucom.2020.06.035.
3. Jin, C.H., Kim, H.-J., Piao, Y., Li, M. and Piao, M., "Wafer map defect pattern classification based on convolutional neural network features and error-correcting output codes", ***Journal of Intelligent Manufacturing***, Vol. 31, No. 8, (2020), 1861-1875, doi: 10.1007/s10845-020-01540-x.
4. Gu, S., Cai, Y., Shan, J. and Hou, C., "Active learning with error-correcting output codes", ***Neurocomputing***, Vol. 364, (2019), 182-191, doi: 10.1016/j.neucom.2019.06.064.
5. Sun, N., Shan, J. and Hou, C., "Multi-label active learning with error correcting output codes", in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer., (2019), 331-342.
6. Zhang, Y.-P., Ye, X.-N., Liu, K.-H. and Yao, J.-F., "A novel multi-objective genetic algorithm based error correcting output codes", ***Swarm and Evolutionary Computation***, Vol. 57, (2020), 100709, doi: 10.1016/j.swevo.2020.100709.
7. Almuzaini, H.A. and Azmi, A.M., "Impact of stemming and word embedding on deep learning-based arabic text categorization", ***IEEE Access***, Vol. 8, (2020), 127913-127928, doi: 10.1109/ACCESS.2020.3009217.

8. Wang, T., Liu, L., Liu, N., Zhang, H., Zhang, L. and Feng, S., "A multi-label text classification method via dynamic semantic representation model and deep neural network", ***Applied Intelligence***, Vol. 50, No. 8, (2020), 2339-2351, doi: 10.1007/s10489-020-01680-w.
9. Kimura, K., Kudo, M., Sun, L. and Koujaku, S., "Fast random k-labelsets for large-scale multi-label classification", in 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE., (2016), 438-443.
10. Sebastiani, F., "Machine learning in automated text categorization", ***ACM Computing Surveys (CSUR)***, Vol. 34, No. 1, (2002), 1-47.
11. Bui, D.D.A., Del Fiol, G. and Jonnalagadda, S., "Pdf text classification to leverage information extraction from publication reports", ***Journal of Biomedical Informatics***, Vol. 61, (2016), 141-148, doi: 10.1016/j.jbi.2016.03.026.
12. Yu, B. and Xu, Z.-b., "A comparative study for content-based dynamic spam classification using four machine learning algorithms", ***Knowledge-Based Systems***, Vol. 21, No. 4, (2008), 355-362, doi: 10.1016/j.knosys.2008.01.001.
13. Loh, W.Y., "Classification and regression trees", ***Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery***, Vol. 1, No. 1, (2011), 14-23, doi: 10.1002/widm.8.
14. Zhong, G., Huang, K. and Liu, C.-L., "Joint learning of error-correcting output codes and dichotomizers from data", ***Neural Computing and Applications***, Vol. 21, No. 4, (2012), 715-724, doi: 10.1007/s00521-011-0653-z.
15. Kyeong, K. and Kim, H., "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks", ***IEEE Transactions on Semiconductor Manufacturing***, Vol. 31, No. 3, (2018), 395-402, doi: 10.1109/TSM.2018.2841416.
16. Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks", ***Advances in Neural Information Processing Systems***, Vol. 25, (2012), doi: 10.1145/3065386.
17. Krawczyk, B., Galar, M., Woźniak, M., Bustince, H. and Herrera, F., "Dynamic ensemble selection for multi-class classification with one-class classifiers", ***Pattern Recognition***, Vol. 83, (2018), 34-51.
18. Feng, K.-J., Liong, S.-T. and Liu, K.-H., "The design of variable-length coding matrix for improving error correcting output codes", ***Information Sciences***, Vol. 534, (2020), 192-217, doi: 10.1016/j.ins.2020.04.021.
19. Li, K.-S., Wang, H.-R. and Liu, K.-H., "A novel error-correcting output codes algorithm based on genetic programming", ***Swarm and Evolutionary Computation***, Vol. 50, (2019), 100564, doi: 10.1016/j.swevo.2019.100564.
20. Baró, X., Escalera, S., Vitria, J., Pujol, O. and Radeva, P., "Traffic sign recognition using evolutionary adaboost detection and forest-ecoc classification", ***IEEE Transactions on Intelligent Transportation Systems***, Vol. 10, No. 1, (2009), 113-126, doi.
21. Nazari, S., Moin, M.-S. and Kanan, H.R., "Securing templates in a face recognition system using error-correcting output code and chaos theory", ***Computers & Electrical Engineering***, Vol. 72, (2018), 644-659, doi: 10.1016/j.compeleceng.2018.01.029.
22. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J. and Wang, Y., "Zero-shot action recognition with error-correcting output codes", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., (2017), 2833-2842.
23. Sadjadi, S., Mashayekhi, H. and Hassanpour, H., "A two-level semi-supervised clustering technique for news articles", ***International Journal of Engineering***, ***Transactions C: Aspects,*** Vol. 34, No. 12, (2021), 2648-2657, doi: 10.5829/IJE.2021.34.12C.10.
24. Vidyadhari, C., Sandhya, N. and Premchand, P., "A semantic word processing using enhanced cat swarm optimization algorithm for automatic text clustering", ***Multimedia Research***, Vol. 2, No. 4, (2019), 23-32, doi: 10.46253/j.mr.v2i4.a3.
25. Lee, Y., Kim, E., Kim, Y. and Seol, D., "Effective message authentication method for performing a swarm flight of drones", ***Emergency***, Vol. 3, No. 4, (2015), 95-97, doi: 10.2991/eers-15.2015.23.

Persian Abstract

چکیده

در چندین مشکل طبقه‌بندی در دنیای واقعی، وقتی تعداد زیادی داده بدون برچسب وجود دارد، معمولاً به‌دست آوردن داده‌های برچسب‌دار سخت است. از این رو، ابداع رویکردهای جدید برای حل این مشکلات بسیار مهم است و در نتیجه ارزشمندترین نمونه ها برای برچسب گذاری و ایجاد یک طبقه بندی برتر انتخاب می شود. چندین تکنیک موجود برای مسائل دسته‌بندی باینری ابداع شده‌اند، تنها تعداد محدودی الگوریتم برای رسیدگی به موارد چند برچسبی طراحی شده‌اند. مشکل طبقه‌بندی چند برچسبی زمانی پیچیده‌تر می‌شود که نمونه متعلق به چندین برچسب از گروه کلاس‌های قابل دسترس باشد. امروزه در وب جهانی، داده های متنی به طور کلی وجود دارد و نمونه بارز این نوع وظایف است. این مقاله یک تکنیک جدید را برای انجام دسته‌بندی متن چند برچسبی با اصلاح رویکرد کدگذاری خروجی تصحیح خطا (ECOC)ایجاد می‌کند. در اینجا، خوشه‌ای از طبقه‌بندی‌کننده‌های مکمل دودویی برای تسهیل ECOC برای مشکلات چند کلاسه مؤثرتر استفاده می‌شوند. علاوه بر این، یک احتمال پسین وزنی محاسبه می‌شود تا عملکرد طبقه‌بندی متن چند برچسبی را به طور مؤثرتری افزایش دهد. علاوه بر این، عملکرد ECOC پیشنهادی با احتمال وزن‌دار با استفاده از معیارهای عملکرد، مانند دقت، فراخوان و اندازه‌گیری f با حداکثر دقت ۰.۸۹۷، ارزش فراخوان بالاتر ۰.۸۹۶ و حداکثر اندازه‌گیری 0.895 f تجزیه و تحلیل می‌شود.