International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Time Conserving Multi-Label Classification System by Incorporating Pyramid Data Structure

Y. Jeyasheela, S.H. Krishnaveni

Abstract*: Data classification is one of the evergreen research areas of data analysis. Numerous data classification approaches exist in the literature and most of the classification systems are based on binary and multi-class classification. Multi-label classification system attempts to suggest multiple labels for a single entity. However, it is complex to attain a better multi-label classification system. Taking this as a challenge, this work proposes a multi-label classification system, which extracts the features of both entities and labels. The relationship between them are organised in the pyramid data structure. As the features are organized effectively, the interrelated labels are present in the same tier. This feature makes it simple for suggesting multiple labels for a single entity. The performance of this work is analysed over three different datasets and compared against existing approaches in terms of precision, recall, accuracy and time consumption.

***Keywords*: About Data Classification, multi-label classification, pyramid data structure.**

I. INTRODUCTION

Data analysis involves two important areas, which are data clustering and data classification. Data clustering is an unsupervised way of data analysis, which groups the related data together to form data clusters. Data that shares more similarity are placed in the same cluster. This kind of clustering operation does not require any prior knowledge about the dataset and can work without training. On the other hand classification is another important technique, which requires the process of training for gaining knowledge about the dataset. A classification system involves a classifier, which is trained by the data with the associated labels, such that the classifier gains knowledge. With this knowledge, the classifier is equipped to classify between the data with respect to the label.

As far as data classification is concerned, it can be either binary-class or multi-class classification. Binary-class classification involves two classes and hence, a data item can either belong to one or the other class. Multi-class classification problem involves many classes and a data item can belong to one of more classes. In this case, any number of classes can be involved but, a data item can belong to only one class.

Due to the skyrocketing increase of data, it is highly challenging to perform data analysis effectively. For instance, to ensure perfect classification, a data item can be in one or more classes and this type of classification is called multi-label classification. Multi-label classification is so popular now-a-days, as the classification is realistic and reliable. Though multi-label classification is appreciated by almost all the domains, the text processing and medical diagnostic systems have reaped most of the benefits of multi-label classification. To justify this statement, a patient who is suffering from cancer may also get diabetes and the news about a temple may have its place under history or literature [1-4]. Hence, today's world looks forward for more multi-label classification system.

However, it is not simple to map the same data item with multiple labels and it can be attained only with a well-trained classifier. A multi-label classification system may fall into two types, which are problem transformation and algorithm adaptation [5]. The problem transformation techniques break a multi-label classification problem into several single-label classification problems. Yet, this approach is not effective and it consumes so much of time and computational resources. On the other hand, the algorithm adaptation techniques utilize a purposeful learning algorithm to attain multi-label classification. The multi-label classification techniques based on algorithm adaptation techniques are efficient, provided the algorithm is properly utilized.

With this knowledge, this article intends to propose a multi-label classification system based on pyramid data structure [6]. The proposed multi-label classification approach is broken down into three important phases, which are data pre-processing, feature extraction and classification. The data pre-processing phase prepares the data to make it suitable for the forthcoming processes. The feature extraction phase extracts the useful features from the data and exploits the extracted features for training the classifier. Finally, the classification is carried out by applying the gained knowledge over the test data. The noteworthy points about this work are listed as follows.

- The proposed multi-label classification is proven to be faster, as the data is properly organised in the pyramid data structure.
- The multi-label classification system allows a single data item to be under multiple labels, which increases the efficiency of the classification.
- The multi-label classification provides reliable and promising results, which makes it more suitable for real-time applications.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Y. Jeyasheela*, Department of Information Technology, Noorul Islam Centre for Higher Education, Kumaracoil, Nagercoil, India. Email: sheelaniuphd@gmail.com
Dr. S.H. Krishnaveni, Department of CSE, Baselios Mathew II College of Engineering, Kerala, India.. Email: shkrishnaveni@gmail.com

- The incorporation of layer based pyramid data structure provides the highest degree of data organization.
- The employment of pyramid data structure introduces multiple benefits to the system, which are layered data organization, efficient multi-label classification, reduced time, space and resource overhead.
- This work is not confined to a particular domain, which makes sense that it is applicable across multiple domains such as healthcare, retail, marketing and so on.

The remainder of this article is organized as follows. Section 2 presents the related literature with respect to multi-label classification. The proposed multi-label classification system is elaborated in section 3 and the performance of the proposed work is analysed in section 4. The conclusions of this work are drawn in section 5.

II. REVIEW OF LITERATURE

This section reviews the state-of-the-art literature with respect to multi-label classification.

A multi-label classification system based on joint and feature concept correlation is proposed in [7]. The correlations between the concepts are detected by means of hypergraph. Additionally, the feature-concept relevance is also measured by employing a sparsity constraint. However, this work involves more complexity in terms of computation. In [8], a multi-label classification system for Unmanned Aerial Vehicle (UAV) images is proposed. This work is based on Conditional Random Field (CRF), which exploits both the spatial information and the cross-correlation between the labels. Initially, this work divides the entire image into several blocks and multilayer perceptron is utilized to ensure block based multi-label prediction. This is followed by the application of CRF to combine the spatial information and the correlation between the labels. However, this work is applicable for images alone.

A multi-label learning scheme based on joint feature selection and classification is presented in [9]. This technique extracts the features based on label correlation and the multi-label classifier is built. However, this work does not organise the features. In [10], a land classification scheme based on multi-label classification is proposed for remotely sensed data. This technique learns the spectral relationship between the satellite images and different profiles of surface materials. Yet, this technique is meant for satellite images. In [11], a multi-label classification system is proposed, which is based on instance correlation functions. This work maps the training and testing instances based on coefficients, which is based on the correlation between the instances and the relationship among the labels are not considered.

A hierarchical multi-label classification system based on Bayesian decision theory is presented in [12]. Initially, a learning model is developed and then Bayesian optimal predictions are developed. Greedy algorithm is employed to solve the optimization problem. This work involves so much of complex computations and consumes more resources. A hierarchical multi-label classification system is developed on the basis of Mandatory Leaf Node Prediction (MLNP) method in [13]. The MLNP algorithms follow the global label hierarchical structure. The work may minimize the symmetric loss. This work consumes more time to process the data.

In [14], an image classification technique based on two-dimensional multi-label active learning model is presented. This work considers both the samples and their corresponding labels, instead of considering only the sample. Additionally, the proposed approach is online adaptable and this work is completely meant for images. The performance of the feature selection methods in solving multi-label classification problem is evaluated in [15]. This work evaluates the feature selection techniques in three different multi-label classification problem transformations such as binary relevance, pairwise and label power set.

In [16], a multi-label classification system that is based on neighbourhood preservation is proposed. This work verifies every feature subset and computes the ability of the subset, such that the neighbourhood relationship is preserved. The feature selection process of this work is carried out by employing a ranking and greedy algorithm. A hierarchical tree model is proposed for multi-label learning in [17]. The tree is constructed by considering the data hierarchy by using the Support Vector Machine (SVM) classifier. For each node of the tree, a predictive label vector is represented for predicting the multi-label and to identify the relationship between the data. The repeatedly occurring labels at leaf nodes are considered to be relevant to each other. However, the formation of tree based on employing multiple SVMs consumes more resource.

In [18], Label Partitioning by Sub-linear Ranking (LPSR) that trains the system with the labels is proposed. The LPSR constructs a label hierarchy by using a classifier, however this work requires increased training cost. Multi-Label Random Forest (MLRF) is proposed in [19] does not involve the process of explicit learning. Yet, the ensemble of random trees is learnt and the drawback of this work is its inaccuracy.

Motivated by these works, the proposed work intends to propose a multi-label classification system that considers the correlation between the features and the labels as well. Additionally, the gained knowledge is properly organized in the pyramid data structure that helps in reducing the time consumption. The proposed multi-label classification system is elaborated in the following section.

III. PROPOSED MULTI-LABEL CLASSIFICATION SYSTEM

This section elaborates the proposed multi-label classification system along with the overview of the proposed work.

A. Overview of the Proposed Work

The aim of this work is to present a multi-label classification system for the data. Mostly, the image annotation and text mining based applications exploit the merits of multi-label classification. The reason is that the textual words and image regions may belong to one or more class, which enhances the reliability of classification. The multi-label classification system provides one or more labels for the entities, which depends on the correlation between the entities.

However, most of the existing multi-label classification systems utilize the correlation between the labels and do not involve any process with the data samples. This work conceives the idea of enhancing the efficiency of multi-label classification by considering the relationship between the labels and the entities. The overview of the proposed approach is depicted in figure 1.

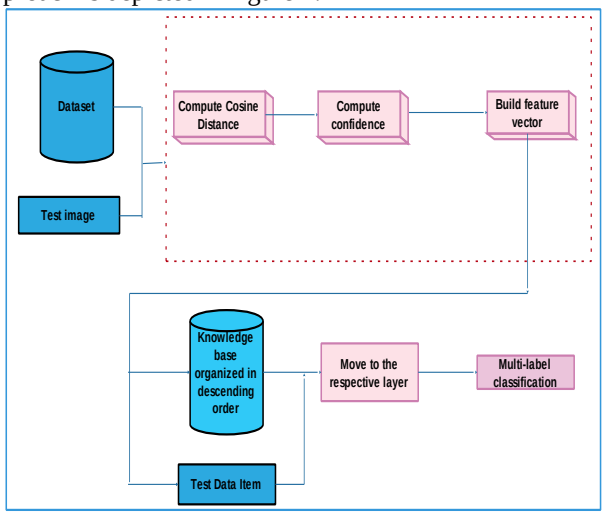

Fig. 1.Overview of the proposed multi-label classification system

The objective of this work is attained by subdividing the work into two phases and they are knowledge feeding and testing. In the knowledge feeding phase, the correlation between the data and data labels are computed. The relationship between the data and data labels are computed by means of similarity measure and confidence values. The feature vector is formed and is loaded into the pyramid data structure with respect to the range of values. Any task can be accomplished faster, when the data is organized properly. As this work organizes the features along with its labels in the pyramid data structure, the multi-label classification is performed on the streak. The following section presents the detailed description of the proposed multi-label classification system.

B. Proposed Approach Based on Pyramid Data Structure

The main phases involved in the proposed multi-label classification system are data pre-processing, feature extraction and multi-label classification. Each and every phase is concerned with a specific task and all the phases are interlinked with each other, such that the goal of this work is attained. The following subsections describe the functionality of each and every phase.

- Data Pre-processing

Data pre-processing is the most basic step of any important process. This phase weeds out any irrelevant or redundant data in the dataset. At the same time, it adds value to the dataset, if the dataset seems to be incomplete. In short, the data pre-processing activity cleanses the dataset, such that the pre-processed data is perfect to be processed by the next step. All the research activities involve the process of data pre-processing and the techniques involved in data pre-processing vary with respect to the application. The proposed multi-label classification system removes the redundant or duplicate data from the dataset. In addition to this, some columns of the dataset may be left empty and are filled with zeroes. By this way, the dataset is pre-processed and the pre-processed data is passed to the next phase.

- Feature Extraction and Classification

As soon as the dataset is pre-processed, the features of the entities and the labels are extracted. The multi-label classification system aims to suggest multiple relevant labels to the data items. This is possible only when the relationship between the entities and the labels are well studied. During the knowledge feeding phase, the entities of every particular label must be analysed, such that the relationship between the entity and the label can be understood better. Hence, this work computes the similarity between the entities by means of popular distance measures such as cosine (C_{dis}), manhattan (M_{dis}) and Euclidean (E_{dis}) distances for checking the performance of the measures.

When the similarity measure is computed between the entities, the relationship between the entities and their associated labels is studied. The computation of similarity measure alone cannot help in determining multiple labels for a test entity. Hence, a technique to enhance the available knowledge is required and is achieved by computing the confidence value of the entity and the label. The confidence value is computed with the entity's distance and the label itself. The confidence value is computed by the calculating the ratio of the support of the distance and label to the support of the distance. By computing the confidence value, the occurrence frequency of an entity's distance in association with the label is detected. The formulae for computing the distance and the confidence are presented as follows.

$$C_{dis} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

$$E_{dis} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \quad (2)$$

$$M_{dis} = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

$$Con = \frac{S(x_{dis} \cup L_i)}{S(x_{dis})} \quad (4)$$

In the above equations, x_i and y_i are the data entities. x_{dis} is the distance of an entity x with the label L_i. After computing the similarity measure and the confidence value, the feature vector is formed. The feature vector of this work contains two parameters. On computing the feature vector, the features are loaded onto the pyramid data structure. The feature vectors are arranged in descending order with respect to the confidence value. The features are loaded with respect to this range in the data structure. The main reason for loading the feature vector into a data structure is to enhance the speed of the classification process. The time consumption is reduced as much as possible, as the data is highly organized and hence, the search process is not tougher anymore.

During the process of classification, a test entity is passed to the classification system. The feature of the test entity is computed and compared with the tiers of the pyramid data structure. The value that matches with the tier alone is processed and the multiple labels are suggested for the data entity. This idea conserves time and utilizes the memory space in a full-fledged manner.

The search process is performed in a streak and multiple labels are suggested. The multiple labels suggested by the proposed classification system are reliable and accurate, owing to the better analysis of both the entities and labels. The relationship between the entities and labels are computed and well utilised for multi-label suggestion. The following section evaluates the performance of the proposed approach.

IV. RESULTS AND DISCUSSION

The performance of this work is analysed on a stand-alone computer with 8 GB RAM and intel i7 processor. The capability of the proposed approach is tested by simulating the work using Java in Netbeans platform. The performance of the proposed approach is tested in terms of accuracy, precision, recall, time consumption and misclassification rates. The results attained by the proposed approach are compared against LPSR and MLRF. The datasets being utilized for analysing the performance of the proposed approach are Bibtex, bookmarks and delicious [20].

The bibtex dataset contains 7395 instances with 159 labels. The bookmarks dataset possesses 87856 instances with 208 labels and the delicious dataset consists of 161705 instances with 983 labels.

Table- I: Dataset Description

Datasets / Details	Bibtex	Bookmarks	Delicious
File size (Mb)	6.65	69.8	7.27
Total instances	7395	87856	16105
Total labels	159	208	983

The experimental results of the proposed approach by varying the distance measure are presented in table 2.

Table- II: Performance Analysis by Varying Similarity Measure

Dataset	Techniques/ Perf.measures	Accuracy (%)	Precision (%)	Recall (%)	Time (ms)	Misclassification
Bibtex	Cosine	93.1	92.4	94.3	28	6.9
	Euclidean	82.2	76.3	72.3	43	17.8
	Manhattan	84.4	79.7	78.4	44	15.6
Bookmarks	Cosine	94.2	93.1	94.7	62	5.8
	Euclidean	82.4	79.4	80.04	78	17.6
	Manhattan	86.9	82.7	80.3	81	13.1
Delicious	Cosine	94.8	95.4	96.8	53	5.2
	Euclidean	89.9	86.7	83.2	62	10.1
	Manhattan	90.4	86.9	87.2	68	9.6

The performance of the proposed approach is analysed in two rounds. The initial round of performance analysis is carried out by varying the similarity measure such as cosine, Euclidean and manhattan distance. From the experimental results, it is evident that the performance of cosine similarity measure is better than Euclidean and manhattan similarity measure. The formulae for computing the performance measures are presented as follows.

$$Acc_{rate} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Pr_{rate} = \frac{TP}{TP+FP} \quad (6)$$

$$R_{rate} = \frac{TP}{TP+FN} \quad (7)$$

In the equations, TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative rates. In this case, TP is the correctly classified entities and TN is the correctly rejected entities. FP is the incorrectly classified entities by suggesting a wrong label and FN is the incorrectly classified entities which should be placed under a specific label.

The second round of performance analysis is carried out by varying the techniques and comparing the performance of the proposed approach with the existing approaches LPSR and MLRF. The experimental results of the proposed multilabel classification system are presented as follows.

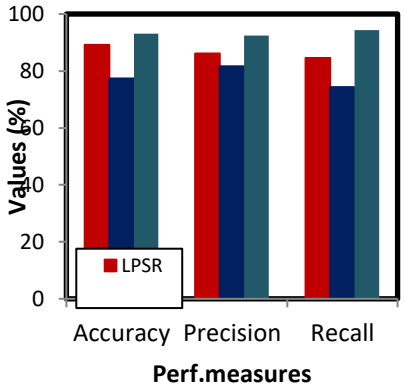

Fig. 2.Experimental results on Bibtex Dataset

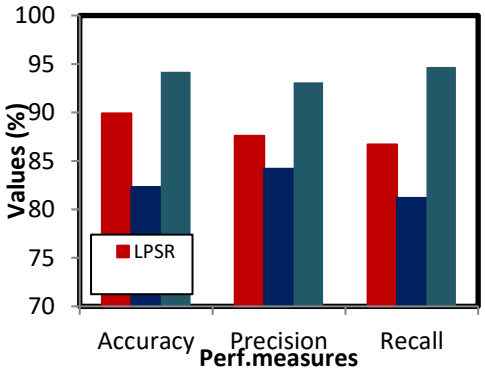

Fig. 3.Experimental results on Bookmarks Dataset

Retrieval Number F7627088619/2019©BEIESP
DOI: 10.35940/ijeat.F7627.088619
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

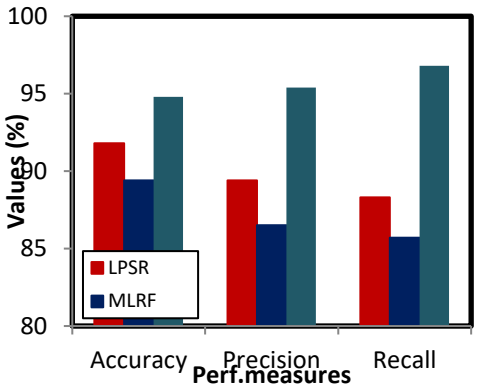

Fig. 4.Experimental results on Delicious Dataset

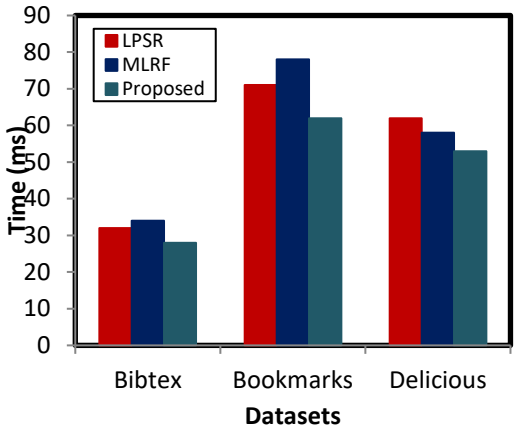

Fig. 5.Time Consumption Analysis

Based on the experimental results, the performance of the proposed approach is proven with maximum accuracy, precision and recall rates. On the other hand, the time consumption of the proposed classification approach is minimal, when compared to the existing approaches. The reason for the maximum accuracy rates is the better learning of the entities and the labels. The relationship between the entities and labels are represented by the feature vector. Additionally, the feature vectors are properly organised in the pyramid data structure, which makes the entire process of classification easier. The effective data organization helps in reducing the search time and reliable labelling process.

As the features are computed by taking both the labels and entities into account, the features are sharper. In addition to this, the effective organization of data on the pyramid data structure paves way for easy access and the multiple labels are easily suggested by the proposed approach. The precision and recall rates of the proposed approach are greater, as they are based on the impact of the FP and FN rates respectively. The proposed approach shows lesser FP and FN rates, because of which the precision and recall rates are greater.

The reason for attaining low FP and FN rates is the sufficient feature utilization and effective organization of the extracted features with the associated labels in the pyramid data structure. Each tier of the pyramid data structure is loaded with the sorted feature sets with the associated labels. As the tier of the pyramid data structure contains features with respect to the range of the features, each tier contains interrelated labels. Hence, multiple labels can easily be suggested for the entities.

The underlying reason for the minimal time consumption of the proposed approach is the utilization of the simple features and effective organization of the features. Both these characteristics of the proposed approach reduces the overall time consumption. The proposed work can easily suggest the labels of the entities, as the interrelated labels are placed in the same tier. Hence, the objective of the proposed multi-label classification approach is attained with maximal accuracy, precision and recall rates over minimal time consumption.

V. CONCLUSION

This article presents a multi-label classification system for different datasets. This work is subdivided into three phases, which are data pre-processing, feature extraction and classification. The data pre-processing phase aims to eliminate the redundant data and to cleanse the data. The similarity between the entities and the labels are computed and the feature vector is formed. The feature vector is organised in the pyramid data structure, which is composed of several tiers. Each tier is loaded with the interrelated labels and is organised based on the range of values. Hence, an entity which falls under a label may also be a member of another label in the same tier. By this way, multiple labels can be suggested for a single data entity. The performance of the proposed multi-label classification system is analysed in terms of precision, recall, accuracy and time consumption. The proposed approach proves better results, when compared to the analogous approaches. In future, this work is planned to be extended by incorporating the multi-label classification system to a real-time application.

REFERENCES

1. Rousu, J., Saunders, C., Szedmak, S., and Shawe Taylor, J. “Kernel-based learning of hierarchical multilabel classification models”. *Journal of Machine Learning Research*, 7:1601–1626, 2006.
2. Barutcuoglu, Z. and Troyanskaya, O.G. “Hierarchical multi-label prediction of gene function”. *Bioinformatics*, 22(7), 2006.
3. Zhang, M.-L. and Zhou, Z.-H. “ML-KNN: A lazy learning approach to multi-label learning”. *Pattern Recognition*, 40(7), 2007.
4. Silla, C.N. and Freitas, A.A. “A survey of hierarchical classification across different application domains”. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2010.
5. Tsoumakas, G., Katakis, I., and Vlahavas, I. “Mining multi-label data”. In Maimon, O. and Rokach, L. (eds.), *Data Mining and Knowledge Discovery Handbook*. Springer, 2nd edition, 2010.
6. Jeya Sheela Y., Krishnaveni S.H., "A Novel Frequent Pattern Mining Approach with OTSP", *International Journal of Computer Technology and Applications*, Vol.8, No.5, pp.2275-2284, 2015.
7. Xiaowei Zhao ; Zhigang Ma ; Zhi Li ; Zhihui Li, "Joint Concept Correlation and Feature-Concept Relevance Learning for Multilabel Classification", *Neural Computation*, Vol.3, No.2, pp.526-545, 2018.
8. Abdallah Zeggada ; Souad Benbraika ; Farid Melgani ; Zouhir Mokhtari, "Multilabel Conditional Random Field Classification for UAV Images", *IEEE Geoscience and Remote Sensing Letters*, Vol.15, No. 3. pp.399-403, 2018.
9. Jun Huang ; Guorong Li ; Qingming Huang ; Xindong Wu, "Joint Feature Selection and Classification for Multilabel Learning", *IEEE Transactions on Cybernetics*, Vol.48, No.3, pp.876-889, 2018.
10. Konstantinos Karalas ; Grigorios Tsagkatakis ; Michael Zervakis ; Panagiotis Tsakalides, "Land Classification Using Remotely Sensed Data: Going Multilabel", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.54, No.6, pp.3548-3563, 2016.

11. Huawen Liu ; Xuelong Li ; Shichao Zhang, "Learning Instance Correlation Functions for Multilabel Classification", *IEEE Transactions on Cybernetics*, Vol.47, No.2, pp. 499-510, 2017.
12. Wei Bi ; Jame T. Kwok, "Bayes-Optimal Hierarchical Multilabel Classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol.27, No.11, pp. 2907-2918, 2015.
13. Wei Bi ; James T. Kwok, "Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification", *IEEE Transactions on Neural Networks and Learning Systems*, Vol.25, No.12, pp. 2275-2287, 2014.
14. Guo-Jun Qi ; Xian-Sheng Hua ; Yong Rui ; Jinhui Tang ; Hong-Jiang Zhang, "Two-Dimensional Multilabel Active Learning with an Efficient Online Adaptation Model for Image Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.31, No.10, pp. 1880-1897, 2009.
15. Juan Manuel Rodriguez ; Daniela Godoy ; Alejandro Zunino, "An Empirical Comparison Of Feature Selection Methods In Problem Transformation Multi-label Classification", *IEEE Latin America Transactions*, Vol.14, No.8, pp.3784-3791, 2016.
16. Zhiling Cai ; William Zhu, "Feature selection for multi-label classification using neighborhood preservation", *IEEE/CAA Journal of Automatica Sinica*, Vol.5, No.1, pp. 320-330, 2018.
17. Qingyao Wu ; Yunming Ye ; Haijun Zhang ; Tommy W. S. Chow ; Shen-Shyang Ho, "ML-TREE: A Tree-Structure-Based Approach to Multilabel Learning", *IEEE Transactions on Neural Networks and Learning Systems*, Vol.26, No.3, pp. 430-443, 2015.
18. J. Weston, A. Makadia, and H. Yee." Label partitioning for sublinear ranking". *In ICML*, volume 28, pages 181-189, 2013.
19. R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. "Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages". In WWW, pages 13-24, 2013.
20. http://mulan.sourceforge.net/datasets-mlc.html

AUTHORS PROFILE

Jeyasheela.Y received her B.E.(2003) in Computer Science and Engineering from M.S. University, Tirunelveli and M.E. (2006) in Computer Science and Engineering from Anna University, Chennai. At present she is working as Assistant Professor in the department of Information Technology, Noorul Islam Centre for Higher Education,Kumaracoil,TamilNadu. Her research interests include Data mining ,Network Security,Image Processing and Data Structures.She has authored over 7 publications.She is currently pursuing the Ph.D in Computer Science and Engineering at Noorul Islam Centre for Higher Education,Kumaracoil, TamilNadu.

Dr.S.H. Krishna Veni received her B.E.(2003) in Electrical and Electronics Engineering from Noorul Islam College of Engineering,Kumaracoil, TamilNadu and M.E. (2005) in Computer Science and Engineering from Anna University , Chennai and Ph.D.(2010) in Computer Science and Engineering from M.S. University, Tirunelveli. Presently she is working as Associate professor in CSE department of Baselios Mathews II College of Engineering , Kollam, Kerala. She has 15years of diverse experience in teaching , Life member of ISTE, and IEEE Member. Her Research interests include Image processing, network Security , Data Mining and Soft Computing. She has authored or co-authored over 30 publications.