Solving Multi Label Problems with Clustering and Nearest Neighbor by Consideration of Labels

C. P. Prathibhamol and Asha Ashok

Abstract In any Multi label classification problem, each instance is associated with multiple class labels. In this paper, we aim to predict the class labels of the test data accurately, using an improved multi label classification approach. This method is based on a framework that comprises an initial clustering phase followed by rule extraction using FP-Growth algorithm in label space. To predict the label of a new test data instance, this technique searches for the nearest cluster, thereby locating k-Nearest Neighbors within the corresponding cluster. The labels for the test instance are estimated by prior probabilities of the already predicted labels. Hence, by doing so, this scheme utilizes the advantages of the hybrid approach of both clustering and association rule mining.The proposed algorithm was tested on standard multi label datasets like yeast and scene. It achieved an overall accuracy of 81% when compared with scene dataset and a 68% in yeast dataset.

1 Introduction

Multi Label Classification (MLC) is one of the most challenging problems in the field of data mining. It fundamentally revolves around the fact that each data instance may be associated with numerous target labels. If the test instance within the dataset is coupled with only one target class, then the classification problem becomes single-label. The relevant application areas where multi label classification can be effectively applied include semantic scene classification, protein function classification, music categorization etc.

A lot of recent studies and proficient researches depicts the reality that a fine solution of this problem results in condensed human effort as well as minimal time

C.P. Prathibhamol(✉) · A. Ashok
Department of CSE, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, Kerala, India
e-mail: {prathibhamolcp,ashaashok}@am.amrita.edu

S.M. Thampi et al. (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*,
Advances in Intelligent Systems and Computing 425,
DOI: 10.1007/978-3-319-28658-7_43

consumption. Two well known and traditional approaches exist for solving MLC problems. They are commonly known as Algorithm adaptation and Problem transformation method [9]. In the preliminary approach the key notion is converting the multi label problem into numerous single label problems. Whereas, in the latter approach, any existing single label classification algorithm is adapted to handle multi label problems.

In this paper, we have implemented an Algorithm adaptation approach on fundamental k-Nearest Neighbor algorithm, which is a popular classification algorithm. In this proposed method, there are mainly two stages involved: Training phase followed by a Testing phase. Initially, for the Training of the multi label dataset, we cluster the data pertaining to feature space using a well known clustering algorithm, k-means. After the execution of k-means algorithm, the normal clusters are obtained. The main advantage associated with clustering phase is that it is able to break the entire dataset into dis-joint clusters. In each of the clusters, Frequent FP-growth mining algorithm is then applied corresponding to label space. At the end of this mining, dependency rules among the labels of each cluster is generated. On the completion of training stage, the testing stage is commenced by inputting a new test data instance and checking for the appropriate cluster to which it belongs to. This is achieved by comparing the test data with the centers of all clusters using Euclidean distance. This improves the overall execution time due to the fact that the test data is then later checked only for the subsequent rules of that particular cluster only. As a result, label cardinality of the multi label dataset is significantly reduced by inferring dependency between the extracted labels. Towards the completion of the testing stage, the antecedent part of the mined rules is ultimately passed to instance based algorithm for the purpose of effective classification. An alteration done in the conventional kNN algorithm is that instead of taking into account the majority labels of k nearest neighbors, the prior probabilities of those labels are examined. If the estimated probabilities are greater than a certain threshold, those labels are predicted as that of test data instance. The rest of this paper is prepared as follows: Literature Survey and related papers are discussed in Section 2. System Architecture of our novel method is explained in Section 3. Section 4 is devoted to experiments with several standard datasets. The paper concludes with a summary and future works in Section 5.

2 Related Works

Considerable amount of work have been made in the field of single label classification. At the same time, efforts have also been directed to convert multi label datasets into multiple sets of single label dataset to match the existing labels.

In [1], the authors have developed two probabilistic approaches to solve multi label classification. The first approach is based on logistic regression and nearest neighbor classifiers. The second approach deals with notion of grouping related labels. The former approach is known as Method using Partial Information (MPI) and the latter

approach is known as Method using Association Rules (MAR). In comparison with MAR, MPI takes large amount of time but provides accurate results.

The authors in [2] have contributed to solve MLC problems by using the frame work of Improved Conditional Dependency Networks (ICDN).This method is based on double layer based classifier chain (DCC) to make use of the label correlations in training stage and modifiers the conditional dependency networks (CDN) by initializing the entered values of the second layer with the fore casted values from the first layer during the testing stage.The experimental results from the work confirms that it reduces randomization of input for the conditional dependency networks and convergence rate is considerably improved.

Ying Yu et al. [3] proposed MLRS(Multi Label classification using Rough Sets) which contained the effect of association between the labels and the ambiguity that subsist in the mapping between the feature space and label space. A chain of experiments demonstrated the results that, for seven multi label datasets MLRS obtained better accuracy when compared with basic multi label learning algorithms such as MLkNN, BR, RAkEL etc.

Jiayang Li and Jianhua Xu [4] proposed OVODLSVM(one-versus-one decomposition strategy with double label support vector machine) is a MLC solution by using binary Support Vector Machine(SVM) to build a model of double label SVM by searching for double label instances in the margin between positive and negative instances.Hence by using a voting criteria. This method worked quite well on computational aspects and proved as an efficient solution.

In [5], the authors have conducted a research study of clustering based multi label classification (CBMLC) for the MLC problem. They have tried three clustering algorithms namely simple K-means, Expectation Maximization and Sequential Information Bottleneck algorithm. One main disadvantage of this study was that it didn't consider the dependency between labels.

In the work proposed by the authors in [6], the Apriori algorithm was used to generate all frequent item sets, compound labels with strong associations are replaced by existing single labels. In the case of classification ML-KNN was widely used. But one main limitation concerned with this methodology is that it is not suitable for weak relationship between labels.

An Improved method of Multi label Classification Performance by Label Constraints (IMCP-LC) is proposed by authors in [7]. This method mainly deals with label ranking strategy and label constraints.By using one-against-all decomposition technique, MLC problem is broken down in to numerous binary classification sub-problems. Then for each label training is done by applying binary SVM classifier.After that association rule learning method is applied to mine label constraints. In the last phase in order to correct the output from SVM classifiers, a correction model dealing with label constraints is utilized.

As discussed earlier, the main categorization of MLC problems is done in two ways including Problem Transformation (PT) and Algorithm Adaptation (AA). The authors in [8] have presented a detailed comparison between these two techniques. They have confirmed that AA based methodology is better than PT based methods. It is analyzed with respect to the results of experiments they have carried out in

multi label datasets. Apart from this outcome, they have studied about the different methods of PT and AA techniques. In PT based methods, Binary relevance is suitable for fast binary classification, but a limitation is it doesn't take into account label co-relationship. But label correlation is vital as there is a possibility of label dependencies. Also, labels and their characteristics can be over lapping, so Ranking via single label is not viable. While CLR (Calibrated Label Ranking) [9] is of good quality for considering label relationship, it can't be used for unlabeled data.

Tsoumakas G and Vlahavas I [10] proposed an ensemble method for MLC known as RAKEL (Random k-Label sets). In this method each member of the ensemble is created by selecting random subset of labels. It is followed by using single label classifier for prediction of each member of the power set.

An instance based approach to multi label classification has guided to ML-KNN [11] method. This method combines the idea of the conventional kNN algorithm and Bayesian inference. Based on statistical information derived from the label sets of an unseen instance's neighboring instances, ML-KNN utilizes maximum a posteriori principle to determine the label set for the unseen instance.

In another approach as stated in [12] the authors have worked to solve the MLC problem by association rule mining. They have tried to decompose multi label datasets to extract single label rules and then combine labels with the same attributes to generate multi label rules. This experiment achieves good performance in an application to scene classification.

In [13], the authors combined multi label methods by ensemble techniques. They solved by using a combination of various multi label learners to avoid discrepancy in training sets and correlation difficulties.They have accomplished and tried to use two methods namely EML_M, EML_T in this regard.

3 System Architecture

In this paper we proposed a method for algorithm adaptation. As in `Fig.1`, we have utilized simple k-means for clustering application. This is quite effective as it divides the massive dataset into disjoint clusters which considerably reduces the training time for any classification stage. By doing so we have overcome the main limitation of the work proposed by the authors in [4]. As mentioned above, in their work, they have used Apriori algorithm for mining of rules. Since FP-Growth is far better than Apriori algorithm, in terms of efficiency and resolving of label dependencies, we have utilized the former algorithm in this work. By our method, we are aiming at feature space reduction through clustering algorithm. Furthermore, label space reduction is also achieved by FP-growth algorithm.

In the testing phase as in `Fig.2`, initially for any test data instance the nearest cluster to it is identified. For this purpose we have measured Euclidean distance of the test data instance with all the other generated cluster centers. Once the test instance is known to lie in which cluster, then it is checked for the already generated label dependencies of that particular cluster.

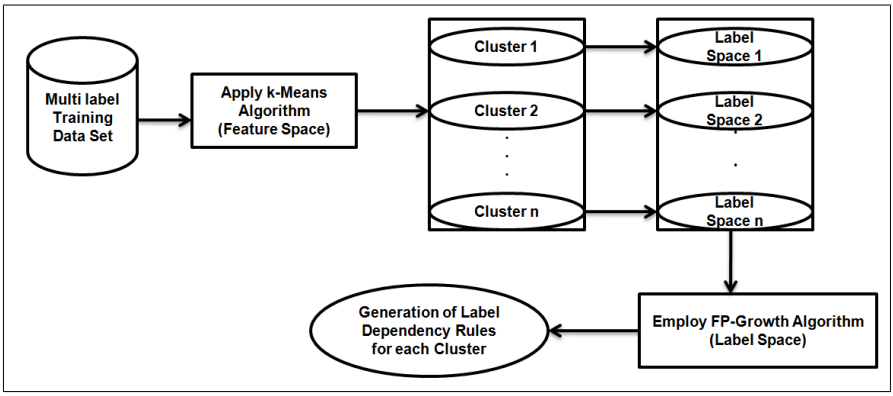

Fig. 1 Training Phase of our proposed method. (SMLP-CNN-CL)

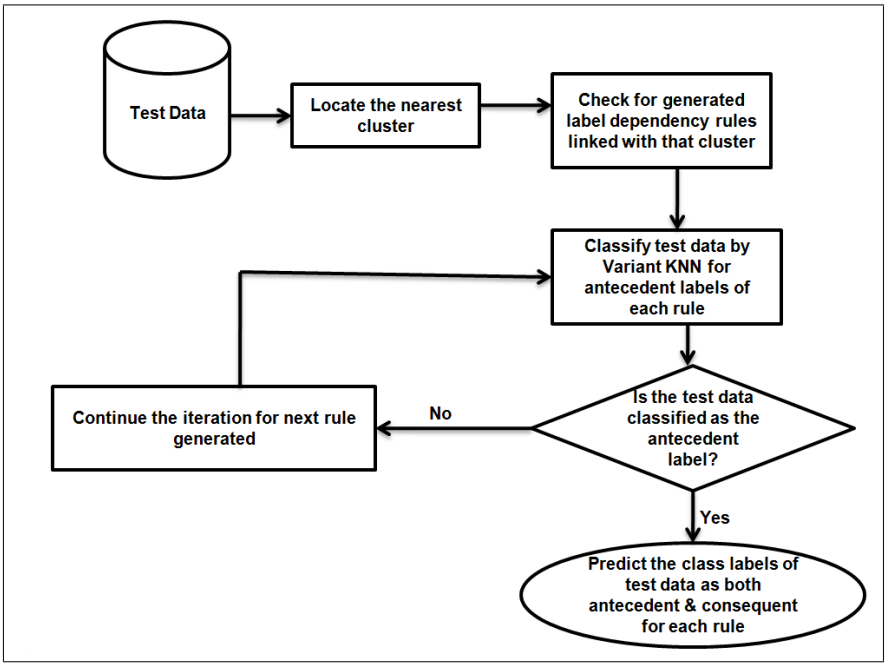

Fig. 2 Testing Phase of our proposed method. (SMLP-CNN-CL)

Any rule of the form

$$L_1 \Rightarrow L_2 \wedge L_3 \,, \tag{1}$$

implies that if the antecedent label L1 is present, then the consequent labels or compound labels both L2 and L3 are also sure to appear along with it at the same time. A rule is considered to be strong if it satisfies a minimum support and confidence.

Table 1 Characteristics of datasets

Datasets	Total Labels	Total Attributes	Label Cardinality	Label Density
Yeast	14	103	4.237	0.303
Scene	6	294	1.074	0.179

In the classification stage, we adopt Variant k-Nearest Neighbors (V-kNN) within the identified cluster. kNN is the algorithm by which k nearest neighbors is detected with the test instance. k is a value which is specified according to the users choice. In a conventional kNN, the test instance is assigned a class label which is the majority of k nearest neighbors class labels. The modification done to the nave kNN algorithm is that here we have considered the prior probability of antecedent labels. If the probability is greater than a particular threshold, then the antecedent label along with its consequents are considered as predicted test instances labels. For the above mentioned rule, antecedent label i.e. L_1 ś probability is greater than an assumed value, then L_1 along with L_2 and L_3 are the estimated class labels.

4 Experimental Evaluations

The remaining portion of the paper gives a brief insight in to the various multi label datasets used for evaluation of our proposed method. Also, the various evaluation metrics are calculated on these datasets so as to confirm the efficiency of SMLP-CNN-CL when compared with other multi label approaches.

4.1 Experimental Datasets

In order to confirm the feasibility of our proposed method, we have conducted experiments on real datasets. The datasets are available at http://mulan.sourceforge.net/datasets-mlc.html. The details of datasets are listed in Table 1.

The yeast dataset in essence contain information about several types of genes of one particular organism. It contains 1500 instances which consist of 103 numerical valued attributes and 14 labels.

The scene dataset contains several types of scene environmental information such as mountain, beach, sunset, fall foliage, urban and field. It comprises 1211 instances with 294 numerical valued attributes and 6 labels.Label Cardinality is more for Yeast dataset and is less for Scene. So Yeast is having more dimensionality in the label space and Scene is having less dimensionality.

4.2 Evaluation Metrics

In this paper, 4 measures have been selected for comparison of the proposed method with previously existing multi label classification algorithms.

In all the definitions given below, x_i denotes the actual labels of the i^{th} test instance. Also, y_i represents the set of predicted labels for the corresponding test instance. If L is the total number of labels associated with the data set and D is the number of instances to be tested. All the evaluation measures are taken from [14]. The evaluation measures are listed as below:

4.2.1 Hamming Loss

This measure indicates the number of times misclassification of example-label pair occurs. It is estimated as the number of wrong labels to the total number of labels. The predicted labels are checked with respect to the original labels and are added up as 1 if they are wrong and 0 if they are correct labels associated with the dataset and D is the number of instances to be tested, the hamming loss is calculated as:

$$HL(x, y) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|x_i \oplus y_i|}{|L|} \tag{2}$$

4.2.2 Accuracy

It is measured as the number of correct labels divided by the union of predicted and true labels.

$$accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|x_i \cap y_i|}{|x_i \cup y_i|} \tag{3}$$

4.2.3 Precision

This ratio estimates the number of correct matches obtained between the true and predicted label to the number of total predicted labels.

$$precision = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|x_i \cap y_i|}{|y_i|} \tag{4}$$

Table 2 Experimental results of yeast dataset

Algorithms	Hamming Loss	Accuracy	Precision	Recall
IMCP-LC	0.190	0.552	0.678	0.715
EML_M	0.193	0.500	0.738	0.553
EML_T	0.197	0.553	0.682	0.690
ML-KNN	0.198	0.492	0.732	0.549
C4.5	0.259	0.423	0.561	0.593
Naive-Bayes	0.301	0.421	0.610	0.531
Binary-SVM	0.202	0.530	0.586	0.633
CLR	0.210	0.497	0.674	0.596
RAKEL	0.244	0.465	0.601	0.618
I-BLR	0.199	0.506	0.712	0.581
SMLP-CNN-CL	**0.111**	**0.686**	**0.921**	**0.750**

Table 3 Experimental results of scene dataset

Algorithms	Hamming Loss	Accuracy	Precision	Recall
IMCP-LC	0.102	0.705	0.722	0.728
EML_M	0.084	0.699	0.730	0.716
EML_T	0.095	0.694	0.725	0.754
ML-KNN	0.099	0.629	0.661	0.655
C4.5	0.148	0.576	0.579	0.588
Naive-Bayes	0.139	0.605	0.615	0.624
Binary-SVM	0.103	0.702	0.715	0.720
CLR	0.122	0.577	0.600	0.669
RAKEL	0.112	0.571	0.598	0.612
I-BLR	0.091	0.647	0.676	0.655
SMLP-CNN-CL	**0.062**	**0.815**	**0.830**	**0.894**

4.2.4 Recall

This ratio estimates the number of correct matches obtained between the true and predicted label to the number of true labels.

$$recall = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|x_i \cap y_i|}{|x_i|} \tag{5}$$

4.3 Results and Discussions

We have compared the above stated measures for our proposed method with variant multi label classification algorithms. Table 2 and 3 demonstrate the facts and figures pertaining to this results. Table 4 again confirms that, when compared with many variant solutions of MLC problem, Hamming Loss is at its minimum for the proposed approach(SMLP-CNN-CL).

Table 4 Hamming loss results of SMLP-CNN-CL and various other MLC algorithms

Datasets	Tested algorithms	Hamming Loss
yeast	MPI	0.3065
scene	MPI	0.0992
yeast	MAR	0.3335
scene	MAR	0.1219
yeast	ICDN	0.197
scene	ICDN	0.096
yeast	MLRS	0.2004
scene	MLRS	0.0927
yeast	OVODLSVM	0.181
scene	OVODLSVM	0.098
yeast	**SMLP-CNN-CL**	**0.111**
scene	**SMLP-CNN-CL**	**0.062**

It is evident from the results obtained that our proposed method SMLP-CNN-CL outperforms most of the existing multi label classification algorithms. In all the experiments, the parameters assumed for K-means algorithm is k=4. The minimum support threshold parameters of FP-Growth are fixed as minsup=2 and minimum confidence=75. For the efficient working of KNN algorithm, k is assumed at a value N/2. Here, N denotes the number of instances within that cluster in which test data belongs to. We have used k=N/2 as for this estimate, we got better results of classification. When k is very small, correct identification was not done as only few neighbors were considered. But as k is made very high, then variation in the results happen due to noise as numerous neighbors are being taken into account. Similarly, in the variant kNN approach, we have considered probability threshold at 0.5. This is because at this threshold, hamming loss was obtained at the minimum value. Beyond this value or at a value less than this, then the results are affected as correct identification of labels is not done. But at 0.5, we are able to attain satisfactory results. The evaluation results of C4.5, Naive Bayes and Binary-SVM algorithms on scene and yeast datasets were taken from [7] and the evaluation results of and ML-KNN, CLR, RAKEL, I-BLR were taken from [13].

As clear from Table 4, we have mainly measured Hamming Loss as the evaluation criteria for comparison of variant MLC approaches with SMLP-CNN-CL on yeast and scene datasets.

5 Conclusion

In this paper, we have aimed at improving multi label classification method based on SMLP-CNN-CL. As evident from the results, this proposed approach works well to give satisfactory results when compared with many other multi label classification

approaches. We intend to do future work in this approach by applying other well known clustering algorithms and also apply the proposed method to various other datasets.

References

1. Kommu, G.R., Trupthi, M., Pabboju, S.: A Novel approach for multi-label classification using probabilistic classifiers. In: IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), pp. 1–8 (2014)
2. Tao, G., Guiyang, L.: Improved conditional dependency networks for multi-label classification. In: Proceedings of the Seventh IEEE International Conference on Measuring Technology and Mechatronics Automation, pp. 561–565 (2015)
3. Yu, Y., Pedrycz, W., Miao, D., et al.: Neighborhood rough sets based multi-label classification for automatic image annotation. International Journal of Approximate Reasoning **54**, 1373–1387 (2013). Elseiver
4. Li, J., Xu, J.: A fast multi-label classification algorithm based on double label support vector machine. In: IEEE International Conference on Computational Intelligence and Security (CIS 2009), vol. 2, pp. 30–35 (2009)
5. Nasierding, G., Sajjanhar, A.: Multi-label classification with clustering for image and text categorization. In: Proceedings of the Sixth IEEE International Conference on Image and Signal Processing (CISP), vol. 2, pp. 869–874 (2013)
6. Qin, F., Tang, X.-J., Cheng, Z.-K.: Application of apriori algorithm in multi-label classification. In: Proceedings of the Fifth IEEE International Conference on Computational and Information Sciences (ICCIS), pp. 717–720 (2013)
7. Chen, B., Hong, X., Duan, L., et al.: Improving multi-label classification performance by label constraints. In: Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1–5 (2013)
8. Prajapati, P., Thakkar, A., Ganatra, A.: A Survey and Current Research Challenges in Multi-Label Classification Methods. International Journal of Soft Computing and Engineering (IJSCE) **2** (2012)
9. Fürnkranz, J., Hüllermeier, E., Mencía, E.L., et al.: Multilabel classification via calibrated label ranking. The Journal of Machine Learning **73**, 133–153 (2008). Springer
10. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: an ensemble method for multilabel classification. Proceedings of IEEE Transactions on Knowledge and Data Engineering, 406–417 (2007)
11. Zhang, M.-L., Zhou, Z.-H.: ML-KNN: A lazy learning approach to multi-label learning. The Journal of Pattern Recognition Society **40**, 2038–2048 (2007). Elseiver
12. Li, B., Li, H., Wu, M., et al.: Multi-label classification based on association rules with application to scene classification. In: Proceedings of the Ninth IEEE International Conference for Young Computer Scientists (ICYCS), pp. 36–41 (2008)
13. Tahir, M.A., Kittler, J., Mikolajczyk, K., et al.: Improving Multilabel Classification Performance by Using Ensemble of Multi-label Classifiers. The Journal of Machine Learning **10**, 11–21 (2010). Springer
14. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Advances in Knowledge Discovery and Data Mining, pp. 22–30. Springer (2004)