Multi Label Prediction Using Association Rule Generation and Simple k-Means

H Haripriya[1] , Prathibhamol Cp[2], Yashwant RPai[2], M Sai Sandeep[2],
Arya M Sankar[2], Srinivas Nag Veerla [2], Prema Nedungadi[1]

1 Amrita CREATE

2 Department of Computer Science and Engineering,

Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, India

haripriya@am.amrita.edu,prathibhamolcp@am.amrita.edu,yashwantrpai @gmail.com,m.saisandeep93@gmail.com,
aryamsankar479@gmail.com,veerlasrinivasnag @gmail.com,prema@am.amrita.edu

***Abstract*—Lately, modern applications like information retrieval, semantic scene classification, music categorization and functional genomics classification highly require multi label classification. A rule mining algorithm apriori is widely used for rule generation. But Apriori is used many times on categorical data, it is seldom used for numerical data. This leads to an idea that with proper data pre-processing, a lot of intangible rules can be derived from such numerical datasets. Since the algorithm will check each and every datasets, we used a simple k-means clustering approach for dividing the processing space of Apriori and thus rules are generated for each cluster. The accuracy of the algorithm is calculated using hamming loss and is presented in the paper. This hybrid algorithm directly aims to find out hidden patterns in huge numerical datasets and make reliable label prediction easier.**

Keywords—Multi label classification; Apriori; rule mining; k-means clustering; mixed data; feature space; label space

I. INTRODUCTION

Multi label classification is a fundamental issue in data mining. It is the deviation of classification problems where unique instance is assigned to multiple target labels. In machine learning, classification can be termed as those observations in the training data whose category is known and the recognition of categories for the new observations. Assigning an email into spam or non-spam classes can be taken as an example.

Many effective researches and efficient studies about multi-label classifications are going on and the two leading procedures for approaching multi-label classification problems are Algorithm adaptation and Problem transformation methods. In the initial approach these problems are transformed as binary classification problems, which can then be handled using single class classifiers. Whereas, binary classification is the strategy used for constructing classifiers for unique labels into positive and negative groups. In Algorithm adaptation methods they redesign the algorithms to directly perform multi-label classification rather than converting the problem to simpler one.

In our paper, we proposed a method on the basis of the popularly used pattern mining algorithm like Apriori by the usage of association rules. Association rule mining is one among the chief functionalities of data mining. The benefits of these rules lie in decision making and prediction, they help in detecting unknown relationships and producing results [1]. The development of association rules is split up into two phases [5] detecting frequent item sets and rule generation. Initially, every set of items is termed as item set, if their combined occurrence is much more than the minimum support threshold [4], these item sets can be termed as frequent item set. Recognizing the reiterative items and extending them to larger sets such that they occur frequently frequent item set identification is comparably easy, but very costly hence this is the most important phase.

In the following phase, it generates n2-1 rules from each item set, where n can be considered as the number of items. Minimum support and confidence is to be pre-defined which represents the threshold of the rules so as to eliminate the rules which lay under it.

The major issue while dealing with association mining is identifying the parallelism among different items within a large set of training dataset [3].Both apriori and frequent pattern growth algorithms are used for frequent item set mining and rule generation over transactional databases.

The remaining sections in the paper comprise of section II Related Works. Section III covers System Architecture of the proposed method. Section IV and V deal with Experimental Evaluations and Conclusion.

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

II. Related Works

Most of the existing classifiers in machine learning [9] are dealing with single label classification. Some methods also under research convert multi-label datasets into multiple sets of single label dataset to fit with existing labels.

In a Comparative Study of Problem Transformation [16], Binary relevance is used for relatively fast binary classification, it doesn't consider label correlation ship. But label correlation has to be considered because there is a possibility of label dependencies. Also, labels and their characteristics can be overlapping, so Ranking via single label is not viable. While CLR is good for considering label relationship, it can't be used for unlabeled data.

A decision tree algorithm C4.5 [12] is used for analysis of phenotype data is discussed in [13]. It is simple and easy to learn. More informative attributes are used for tree splitting. It introduces a new global error function that captures the characteristics of multi-label learning.

Multi label classification using Apriori to generate rules and predict labels for new instances is improved by reducing the need to check through all the rules [1]. But we have considered all the instances and clustered them. Although there are improved algorithms for association rule mining in large databases [3], in this algorithm, combinations of clustering and association is used simultaneously to improve the efficiency.

In most of the multi label classification problems, clustering is commonly used to group the data based on similarities. Whereas ML-KNN[7] is mostly used for dividing the data based on majority of k nearest neighbors class label. But in our method, k-means algorithm has been used for clustering data, henceforth applying the Apriori algorithm in feature space of the generated clusters. Both ML-KNN and in our proposed method, Euclidean distance used as a measure to cluster the data instances. The ultimate aim of k-means reduce the need of checking an instance with every other instance, thereby reducing the time complexity of the problem.

III. System Architecture

In our paper, we proposed a method for algorithm adaptation. Here we make use of both k-means clustering and Apriori algorithm. k-means algorithm divides any given number of instances into k clusters, such that any single instance of a cluster will resemble other instances in the same cluster. The k-means method used for clustering uses Euclidean distance. Euclidean distance is used to measure and cluster together the most similar instances in this case. We use the concept of normalization to bring all the elements of every attribute in between the ranges 0 and 1.

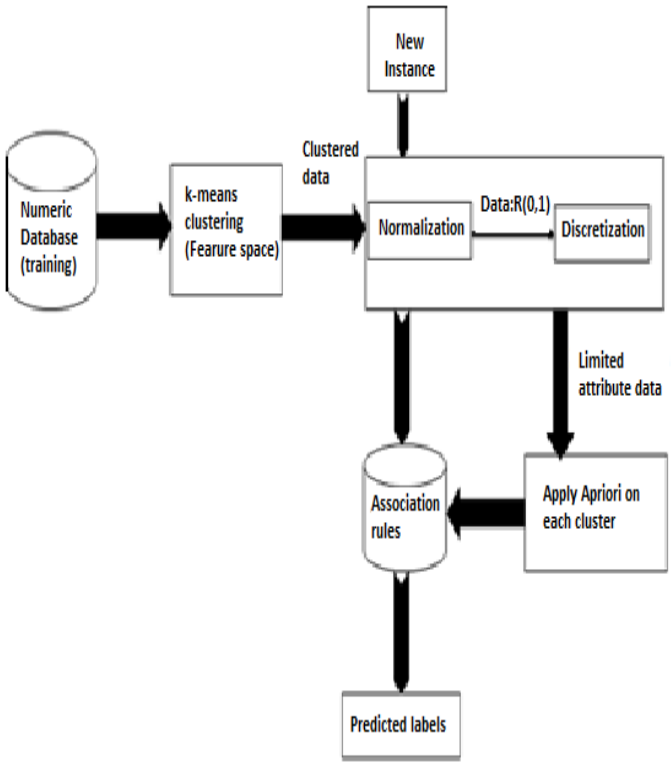

Fig. 1. Multi-label Classifier using k-Means and Apriori.

We employed some preprocessing techniques for fitting the data well for our multi label classifier. One of the first steps concern the normalization of the data. In this method, normalizing the data is very important, especially because various parameters of different units and scales are dealt with. k-means clustering can be subjected only onto numerical values. The importance of an attribute should not be lost due to its low range values. At the same time, an attribute which has higher ranges of values should not affect the cluster. So all the values in the dataset are normalized, ranging from 0 to 1.

If 'D' is a dataset with numerical instances containing m attributes and 'n' labels, and if $x_{\min}$ is the minimum value of the attribute and $x_{\max}$ is the maximum value of the i-th attribute, the value of element I is:

$$\text{Normalized value (V)} = \frac{I - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Since the Apriori algorithm has restrictions to undertake actions on decimal values, we use the concept of discretization to transform previously normalized numerical values categorical ones. Again, followed by the application of Apriori algorithm leads to the rule generation mandatory for prediction of label space over. The equal-width interval approach of

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

discretization worked for some datasets, but for some we faced challenges. It produced a high value of hamming loss because of improper discretization. Since we used equal-width intervals, the distribution of values tends to be much more in some interval and almost zero in other intervals. This causes unequal distribution. So we analysed the dataset first and set the intervals in accordance with the distribution of values.

Apriori algorithm is used in transactional databases containing categorical data to generate frequent patterns and thereby generating association rules. For each cluster, we apply Association-rule mining. It generates rules of the form $X \rightarrow Y$, where X and Y are sets of one or more attributes in the feature space of a cluster.

Algorithm A is the main algorithm for multi label classification. There are mainly two phases for our algorithm, training phase and testing phase. Initially the training dataset consisting of numeric values undergoes k-means clustering on feature space to cluster the input dataset into k clusters. In our algorithm, the value of k is decided by the number of labels in the training set. This means, if there are n labels, n clusters are created allotting one unique label to each cluster.

The purpose of mapping a unique label for each cluster is to show that if a particular label occurs the most frequently (has the highest probability of occurrence) in a cluster, then that cluster points to that particular label. This concept is used to predict the label space of instances of the testing dataset.

If an instance of the testing dataset satisfies a threshold value (say 80%) of the total number of Apriori association rules generated for a particular cluster, then the unique label assigned for that cluster is said to occur for that instance, or the particular label is set as '1', for the instance. If it doesn't satisfy the threshold number of rules for the cluster, then the label assigned to the cluster does not occur for the instance, or is set as '0'. This process is done for all the clusters, and since only one unique label is assigned for a cluster (n labels assigned to n clusters in 1:1 relation), all the 'n' unpredicted labels of the testing instance get predicted. We perform the same procedure for every instance in the testing dataset thereby predicting the label space of all the testing instances.

A. Algorithm for k-Means Apriori Multi-label Classification

1) **Phase 1: Input { Training dataset with feature space and label space}**

Step 1: Cluster data using k-means clustering algorithm (Set k value as number of labels in label space)

Step 2: For each cluster:

2(a) Perform normalization using (1)
2(b) Perform discretization on the normalized data

Step 3: For each label li: ($0 \leq i < n$, n is the total number of labels)

Set lc(i) as number of instances where li=1
Set r(i) as ratio lc(i)/inc, inc is the total number of instances

Step 4: Assign label with highest r(i) value as label for the cluster.

*Step 5:*Apply Apriori algorithm to obtain association rules.

2) **Phase 2 : Input { Testing dataset feature space }**

Step 1: Set threshold (threshold number of rules)
Step 2: For each instance in testing dataset:

Step 2(a): For each cluster:
Set rule_count as 0

Step 2(b): For each association rule:
if (rule is satisfied)
Increment rule_count by 1

if (rule_count $\geq$ threshold)
Set corresponding label as 1.

else
Set corresponding label as 0.

Output { Label Set }

The allocation of a unique label to a cluster is determined by the highest occurring label in that cluster, i.e. for each label, we calculate the ratio of the number of instances for which the label occurs, to the total number of instances in the cluster. The label with the highest ratio is considered for assignment to that cluster. But there might be cases where the same label has the highest ratio for more than one cluster. To avoid such confusion, we developed an algorithm for this problem based on a mapping which is Algorithm B.

B. Algorithm for assigning unique label to each cluster

1) **Initialize data structres**

Step 1: unassigned_labels = Labels not assigned to any cluster (Initially all the labels)

Step 2: unassigned_clusters = Clusters not assigned any label (Initially all the clusters)

Step 3: label_support= (No. of instances for which value of label is 1/ Total no of instances)

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

Step 4: Array A = Two dimensional array with n rows and n columns. (n = number of labels). Here, rows signify clusters and columns signify labels.

For each row i:

Calculate label_support for each label j of cluster i. Store values in array.

Sort row in descending order of label_support.

2) Assignment of labels

For each columnj:

clusters = array of i clusters ($0 \leq i < n$).

labels = array of labels in column j for cluster i.

Sort both arrays simultaneously in the decreasing order of label support.

For each row i:

k = label at column j for cluster i.

if cluster i € unassigned_clusters and
label k € unassigned_labels:

assign label k to cluster i.

This algorithm helps to resolve an ambiguity which can be encountered while assigning a unique label to a cluster. It assigns the most relevant label to each cluster.

IV. Experimental Evaluations

We have done our multi-label classifier experimentations on various datasets. These datasets were obtained from KEEL repository and MULAN repository. We used hamming loss for evaluating our results.

A. *Yeast dataset*

These datasets basically comprise of information about several types of genes of one particular organism. It contains 2417 instances which consist of 103 numerical valued attributes and 14 labels. These datasets were for physiological data modelling contests, 2001 kdd cup data mining competition and so on.

B. *Emotions dataset*

The facial expression evaluation has been universally used in different research areas, such as emotional analysis. In the area of sign language, special importance is attached to facial expressions since they play a critical part in developing the grammatical structure of language, hence are called Grammatical Facial Expressions. Videos recorded using Microsoft Kinect sensor is added in dataset they are in total eighteen. By using Microsoft Kinect, we can obtain images of each frame and also a text file containing one hundred coordinates (x, y, z) of the points from eyes, nose, eyebrows, face contour and iris; each line in the file corresponds to points extracted from one frame.

C. *Scene dataset*

These dataset contains several types of scene environmental information such as mountain, beach, sunset, fall foliage, urban and field. It contains of 294 numerical valued attributes and 6 labels which mentioned above.

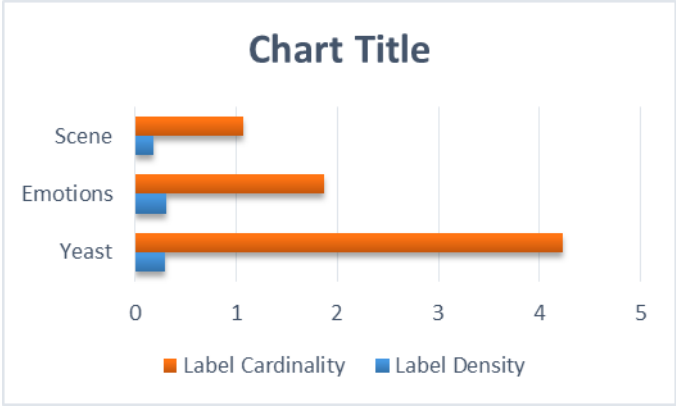

Fig. 2. Properties of multi label datasets used for experimentation

The Figure 2 shows the properties of the datasets.

Hamming Loss : Hamming loss [7] is the fraction of the 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. If 'L is the number of labels to be predicted and 'D' is the number of instances, the hamming loss is calculated as:

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

Table I shows the result obtained after testing our algorithm with various other algorithms on Emotion dataset. In comparison, it is observed that the proposed method discussed in this paper performs better than mere application of ML-KNN algorithm [17] for solving multi-label classification problem. Also, by using Apriori algorithm prior to ML-KNN as a solution to the multi label classification problem displayed a weak result when compared to our proposed method.

TABLE I. Comparison Of Various Algorithms For Emotion Dataset

Dataset	Proposed method	Direct ML-KNN algorithm	Apriori+ML-KNN algorithm
Hamming Loss	**0.38**	0.87	0.83

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

Table II depicts the hamming loss computed by our proposed method on Yeast and Scene datasets. Hamming loss was lowest for Yeast data set when compared with Scene and Emotions.

TABLE II. PROPOSED METHOD'S HAMMING LOSS ON YEAST AND SCENE DATASETS

Dataset	Yeast	Scene
Hamming Loss	0.349	0.39

Table III demonstrates the accuracy measure on Emotions, Yeast and Scene datasets. In addition to this, we aim to check the performance of our proposed method by conducting more experiments with Direct ML-KNN and Apriori+ML-KNN algorithm on Yeast and Scene datasets.

TABLE III. PROPOSED METHOD'S ACCURACY MEASURE ON VARIOUS DATASETS

Dataset	Emotion	Yeast	Scene
Accuracy	0.62	0.65	0.61

V. CONCLUSION

Our paper is aimed at coming up with a hybrid algorithm to predict labels in the numerical dataset without losing accuracy. The label set of the new instances is predicted with the effective use of k-means algorithm and a rule mining Apriori algorithm. This method simplifies the rule generation in numerical datasets by taking advantage of the existing algorithms and modifying them. Future work can be done incorporating intelligent discretization and improved rule mining algorithms for higher accuracy.

ACKNOWLEDGMENT

We thank our mentors and faculty at the Computer Science Department at Amrita University for considering and supporting our project, conducting periodic reviews and providing feedback at every stage of the project, and helping us in its implementation.

REFERENCES

[1] Mohammed Al-Maolegi, Bassam Arkok, "An improved Apriori Algorithm For Association Rules," Vol. 3, No.1, February 2014

[2] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, Zheng Chen, "Effective Multi-Label Active Learning for Text Classification", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009

[3] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, "An Improved Algorithm for Mining Association Rules in Large Databases," Vol. 1, No. 7, 311-316, 2011.

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD Record*, vol. 22, pp. 207–216, 1993.

[5] Grigorios Tsoumakas, and Ioannis Vlahavas, "Random *k*-Label sets: An Ensemble Method for Multilabel Classification", Machine learning: ECML , Springer, 2007.

[6] Xingjian Li, "An Algorithm for Mining Frequent Itemsets from Library Big Data," VOL. 9, NO. 9, SEPTEMBER 2014.

[7] Min-Ling Zhang, Zhi-Hua Zhou, ML-KNN:A lazy learning approach to multi-label learning, 2007 Pattern Recognition Society. Published by Elsevier Ltd

[8] Tsung-Hsien Chiang, Hung-Yi Lo ,Shou-De Lin,A Ranking-based KNN Approach for Multi-Label Classification, 2012 Asian Conference on Machine Learning

[9] Charu C. Aggarwal , ChengXiang Zhai ,A Survey of Text Classification Algorithms, Chapter 6: MINING TEXT DATA.

[10] Hang LI,A Short Introduction to Learning to Rank , IEICE TRANS. INF. & SYST., VOL.E94{D, NO.10 OCTOBER 2011.

[11] Thorsten Joachims,Optimizing Search Engines using Clickthrough Data, *SIGKDD* 02 Edmonton, Alberta, Canada Copyright 2002 ACM.

[12] J.R. Quinlan, Induction of Decision Trees, Kluwer Academic Publisher, 1986.

[13] Amanda Clare, Ross D. King, Knowledge Discovery in Multi-Label Phenotype Data Springer-Verlag Berlin Heidelberg 2001.

[14] Robert E. Schapire, Yoram Singer,BoosTexter: A Boosting-based System for Text Categorization, Machine Learning, 39, 135–168, 2000.

[15] G. Ratsch, T. Onoda, K.R.Muller, Soft Margins for AdaBoost, Kluwer Academic Publishers, 2001.

[16] Purvi Prajapati, Amit Thakkar and Amit Ganatra, A Survey and Current Research Challenges in Multi-Label Classification Methods, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[17] Feng Qin, Xian-Juan Tang, Ze-Kai Cheng, "Application of Apriori Algorithm in Multi-Label Classification," IEEE Fifth International Conference on Computational and Information Sciences (ICCIS), June 2013.