Anomaly Detection based Multi Label Classification using Association Rule Mining (ADMLCAR)

Prathibhamol C. P
Department of
Computer Science and Engineering
Amrita School of Engineering
Amritapuri,
Amrita Vishwa Vidyapeetham
Amrita University
India
prathibhamolcp@am.amrita.edu

Amala G. S
Department of
Computer Science and Applications
Amrita School of Engineering
Amritapuri,
Amrita Vishwa Vidyapeetham
Amrita University
India
amalagopalakrishnan@gmail.com

Malavika Kapadia
Department of
Computer Science and Applications
Amrita School of Engineering
Amritapuri,
Amrita Vishwa Vidyapeetham
Amrita University
India
malavikamcamalu@gmail.com

***Abstract*— Multi label classification contains multiple labels in the label space. Any Multi label classification problem (MLC) deals with numerous class labels associated with data instances. Due to this, correct prediction of labels for a test data remains as a challenge in this field. In this paper an Anomaly Detection based Multi Label Classification using Association Rule Mining (ADMLCAR) is used for solving MLC problem. Conventionally, most of the multi label classification problem is solved by either of the two methods: Problem transformation, Algorithm adaptation. But the method discussed in this paper aims at a novel method different from traditional solution to multi label classification problem. For clustering, ADMLCAR uses k-means algorithm and for association rule mining purpose it uses vertical data format. To predict the test data instances, this method locates the nearest cluster. Once the clusters are identified it uses oversampling principal component analysis (PCA) within the nearest cluster with respect to test instances. Oversampling PCA is used to emphasize the need for confirming the fact that test instance's label set will not only be confined to its nearest cluster label set. This is because, anyways the test instance will be identified to a nearest cluster by means of Euclidean distance measure but as clustering is unsupervised the nearest cluster may contain many objects entities of different label sets.**

Keywords- Multi label classification, k-means, Vertical data format, Oversampling PCA

I. INTRODUCTION

Multi label classification(MLC) problem is more challenging as compared to Single label classification(SLC) as any data instance may related with different labels. So correct and true prediction labels for any data instance becomes a tedious task.

Single label classification contains single label in the label space. Classification tasks involve the separation of data into training and testing sets. The objective of classification is to process a method that predicts labels of the given test data correctly[4].

Fig 1 shows the example of multi label problem with three class labels. Labels = ['run,swim,jump']. This table shows that one instance can have multiple labels at a time. The tick mark indicates that the labels are present for the instance.i.e, A can run,swim,jump at a time. In the label format it indicates like A=[L1, L2, L3].

Earlier research work was done on single label data sets. But, now a days a lot of research is continuing in the area of multi label classification. MLC can be used in many domains such as scene classification, bioinformatics and text categorization. Text categorization can be classified into the categories of privacy, copyright and software and bioinformatics can be classified into functions of transcription,metabolism and protein synthesis.

Any Multi label classification problem may be solved in either of the two methods.

1) Problem transformation.
2) Algorithm Adaption

In Problem transformation, this method is converted into binary classification problem and then treated same as single label classification. But in Algorithm adaption, this method contains specifics learning algorithms for managing multi label directly.

Instances	Class Labels		
	Run	Swim	Jump
A	✓	✓	✓
B	✓		
C		✓	✓
D	✓		✓

Fig. 1. Sample model of Multilabel dataset .

In this work, a new MLC solution concept for predicting the class labels correctly. For this purpose yeast and scene data sets are used. Initially k-means clustering algorithm [7] is used in feature space. We are doing clustering only in the feature space and not in the label space. The advantage of clustering is in the testing phase where any new test instance

978-1-5090-2029-4/16/$31.00 @2016 IEEE

may be compared only with the nearest cluster comprising instances rather than scanning the whole data base. As k-means is a popular clustering algorithm with good performance, we have used it in our training phase. Using this algorithm similar behavior items are automatically assigned to same cluster. Next for finding the rule between the labels we use vertical data format[1]. Vertical data format is mainly used for generating rules between labels. This is because it uses the format class label:TID whichever transactions that label is present. By differentiating the unique label sets in the first pass itself this algorithm reduces the computation for next time to form candidate 2 label sets or 3 label sets as repeated database scanning is not required.

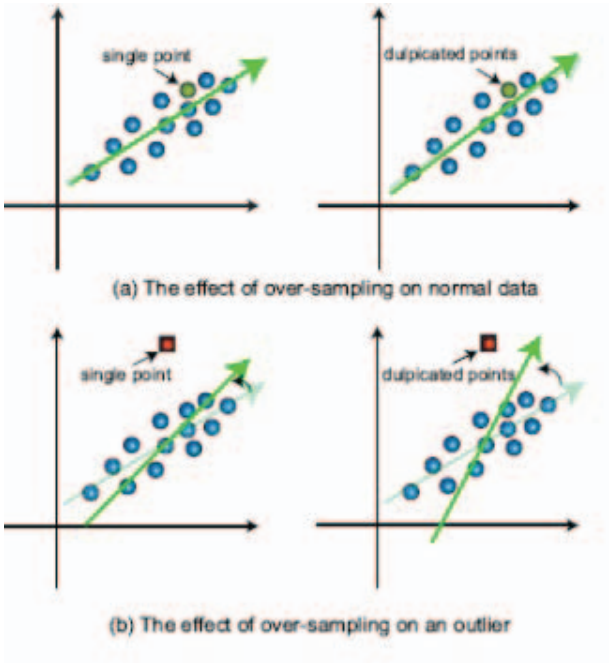

Fig. 2. The outcome of oversampling on a outlier and normal instance.

We have employed oversampling PCA for ensuring or confirming whether test instance belongs to the nearest cluster since clustering is an unsupervised method. The test instance may fall in other nearby clusters also. So, we are checking whether the test instance is an anomalous data in the nearest cluster on oversampling the test instance within that cluster. If it is found to be anomalous within the nearest cluster, then the next nearby cluster is identified and the whole procedure is repeated again.In oversampling PCA(osPCA), the target instance is duplicated many times to enlarge the result of outlier instance relatively that of a normal data. Thus, we are amplifying the target instance. With this oversampling method the principal directions and mean of the information will only be speculated more or less than the target will be a normal information point (Fig 2(a)). On the adverse, the fluctuation will be magnify if we replicate an outlier (Fig 2(b)) [8].

By oversampling, we need to recalculate the principal directions which make complex computation. For avoiding this huge computation, fast updating for the co-variance matrix and determining the Eigen measure problem via the power method can be used.

II. RELATED WORK

This section of the paper represents the existing works and researches done on the various multi label classification method and association rule mining techniques.

The problem transformation and algorithm adaptations are the two methods for doing multi label classification. And the leading process was adaptation method when compared to problem transformation method. Here a novel method differs from traditional solution to a multi label classification problem is taken. The authors in [4] presented a survey of problem transformation and algorithm adaptation methods for multi label classifications. They have used data sets like Gene base,Yeast, Medical and Scene.

Anomaly detection based on clustering algorithm and osPCA was presented by the authors in [2]. In [5], a multi label classification problem was solved by a combination framework of clustering and association rule mining followed by a classification strategy.Susan Tony.et.al [6] reviewed two ways of clustering methods; i.e.k-Means and Hierarchical cluster. The strength and weakness of both the clustering techniques are compared. And also their methodology and process are listed in the paper. The authors in [7] concluded that ML-KNN is perfect in classifying the labels for discrete data.

Feng Qin et.al [3] presented an approach based on multi label classification using apriori algorithm. In the paper they said that multi label learning is the set of training composed of instances. They researched on proposing Apriori algorithm for searching the relationship between all the labels. The compound labels are then replaced by the existing single label according to frequent item sets. GUO Yi-ming et.al [1] proposed a method of processing the record frequently in a particular order using the vertical data format of association rule mining technique [10]. They compared the strength and weakness of apriori method and vertical data format. But in vertical data format the database scans only once. They concluded that Vertical data format needs only shorter storing range and increased the effectiveness process. From these existing works, vertical data format is selected for our proposed method to find the label relationships.

Yi-RenYeh.et.al [8] presented that outlier detection is a significant topic in data mining and has been analyzed in different research fields. They used Leave One Out procedure to verify each individual point the with or without oversampling effect on the fluctuation of principal directions.Based on this approximation, an over-sampling principal component analysis outlier detection method is proposed for underlining the work of an abnormal outlier. Also, they concluded the numerical experiments that the proposed method is effective in computation time and anomaly detection.The oversampling PCA[2] was a novel method of identification of exception by

clustering methods and build as oversampling PCA. Here, the data are clustered into groups by means of k-means and k-medoids clustering algorithms. Also, the anomalous cluster is given into online oversampling PCA phase to predict and rank the outliers. Later, working by oversampling the target data points and evoking the principal direction of the data, the suggested method permits determining the irregularity of a particular data point on the different outputs of dominant eigen vector.So for anomaly detection and thus by predicting whether the cluster are predicted correctly, oversampling PCA will be of great help.

III. METHODOLOGY

The main objective is to accurately predict the class labels of the test instances. In this paper, yeast and scene dataset are considered for experimentation purpose.

In Training phase the feature space of multi label numerical dataset is clustered using k-means (algorithm 1). The main idea is to calculate the k-means by defining the k as the number of desired clusters (assume k clusters). After getting the desired cluster, for finding the rules of the labels we have used vertical data format mining algorithm (algorithm 2)[1].

Algorithm 1: k-means Algorithm

Input: A matrix
Output: Clustered data

1) Assign the value of clusters(k)
2) Select clusters(k) points randomly and fix them initial centroids or center.
3) Calculate the distance from each data points to all centroid using Euclidean distance method.
4) Assign each data points to the minimum centroid.
5) Recompute of the centroid value.
6) Perform steps 3-5 till the centroid merge.

The training phase is initiated by taking the multi label data set without the class label and giving it as input to k-means clustering algorithm. Similarly once similar behaviour tuples are clustered, then it is followed by vertical data mining by which association rules between class labels are obtained. So the output of training phase consists of distinct clusters and for each clusters, the rules between the labels are also identified.

Next the testing phase is performed. In this any test instance will be considered for correct prediction of labels. So, first the near by cluster to which this test instance belongs to is located. The nearby cluster is located by taking Euclidean distance between test instance and cluster centroid using k-means clustering method in the feature space. After that nearby cluster is located, its rules between labels are considered. But at the very same time it cannot be guaranteed that this test instances will contain only labels pertaining to

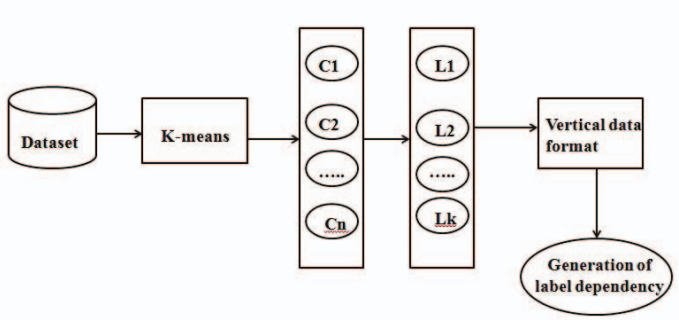

Fig. 3. Training Phase of our proposed method.

that nearby cluster. So as to validate the result correctly, Oversampling PCA is performed by duplicating that target data instance within the nearby cluster a number of times. If there is a variation in the already computed principal component with the resulting one, then the test instance is to be checked with the next nearby cluster. But on oversampling if the already computed principal component and resulting principal component did match together, then it would have been correctly predicted that the label set would only match with that particular cluster's label set.

Fig.3 demonstrates the training phase of our proposed method whereas Fig. 4 depicts the testing phase involved in our work.

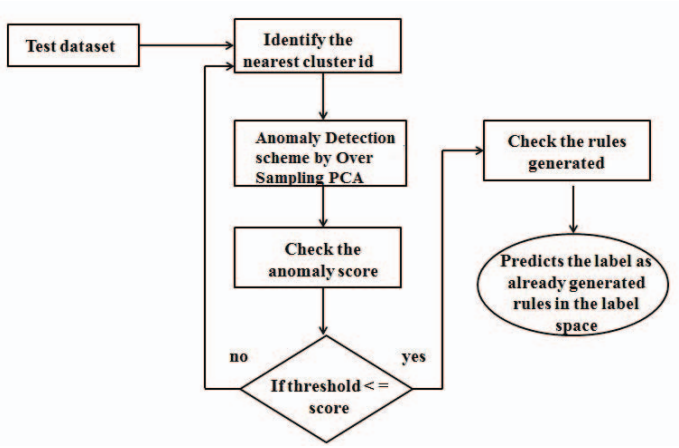

Fig. 4. Testing Phase of our proposed method.

IV. EXPERIMENTAL EVALUATIONS

A. Experimental Data Set

For verifying the accuracy of our proposed method, we conducted experiments on actual datasets. The data sets are taken from http://mulan.sourceforge.net/datasets-mlc.html

Here yeast dataset obtained from mulan repository contains various variety of gene information of one specific organism . It carries 103 integer attributes and 14 class labels with 1500 instances.

Algorithm 2: Vertical Data Format Algorithm

Input: A matrix
Output: Frequent itemset

1) Convert the horizontally formatted data to the vertical format by examining the data set once.
2) Set the support count as 2.
3) Go through the dataset to get frequent 1-item set.
4) Then perform 'and operation' among every element of frequent item set and store the outcome.
5) If the outcome is above minimum support then perform the next 'and operation'.
6) Repeat the process until there is no frequent itemset to perform' and operation 'or the outcome is less than minimum support.

Algorithm 3: Oversampling PCA Algorithm

Input: a data matrix and the oversampling ratio r
Output: outlier score

1) Determine the product, Q=$\frac{AA^T}{n}$ the mean μ,and the first principal direction u
2) Replicate the test instance x_t and calculate the modified mean vector $\overline{\mu}$ and co-variance matrix $\overline{\Sigma}$
   $\overline{\mu} = \frac{\mu + r.x_t}{1+r}$
   $\overline{\Sigma} = \frac{1}{1+r}Q + \frac{r}{1+r}x_t x_t^T - \overline{\mu}\,\overline{\mu}^T$
3) Find the modified first principal direction $\overline{u}$ and then calculate the cosine similarity of u and $\overline{u}$
4) Repeat the steps 2 to 3 till the scanning of each and every data in the clusters.
5) Classify all test data according to their outlier outcomes(1-|cosine similarity|).

The scene dataset has different kinds of scene environmental details. It carries 2407 instances, with 294 numerical attributes along with the 6 labels.

B. Evaluation Metrics

Here four examples based evaluation measures are calculated to find the accuracy of our suggested method. The measures are Hamming loss, accuracy, precision, recall.

In the below definitions we use a_j for actual labels of the jth test instances and b_j for the predicted labels of the checking test data. If L is the aggregate numbers of labels for data set, whereas I denotes the number of instances for testing the instances. The measures are listed as follows:

Hamming Loss: Hamming loss is used to find unclassified label.If the value of hamming loss is less (h), then better the performance of the system. The hamming loss is calculated as:

$$HL(a,b) = \frac{1}{|I|}\sum_{j=1}^{|I|}\frac{|a_j \oplus b_j|}{|L|}$$

Accuracy: Accuracy is used to find the percentage of correctly classified label. Accuracy can be find by:

$$accuracy = \frac{1}{|I|}\sum_{j=1}^{|I|}\frac{|a_j \cup b_j|}{|a_j \cap b_j|}$$

Precision: Precision is the ratio of retrieved or predicted label that are appropriated. The equation is :

$$precision = \frac{1}{|I|}\sum_{j=1}^{|I|}\frac{|a_j \cap b_j|}{|b_j|}$$

Recall: It is the ratio of the labels that are relevant and are successfully predicted. Recall can be measured by:

$$recall = \frac{1}{|I|}\sum_{j=1}^{|I|}\frac{|a_j \cap b_j|}{|a_j|}$$

C. Results and Discussions

We have evaluated the above mentioned values for this aimed approach with diverse multi label classification algorithms. Table I and Table II depicts the result. Fig 5 and 6 shows the accuracy rate and hamming loss of ADMLCAR with other different existing algorithms .

TABLE I
OBSERVATIONAL OUTCOMES OF YEAST DATASET

Algorithms	Hamming Loss	Accuracy	Precision	Recall
ML-KNN	0.198	0.4920	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	0.2021	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
EML_M	0.193	0.500	0.738	0.553
EML_T	0.197	0.553	0.682	0.690
ADMLCAR	0.121	0.748	0.845	0.810

TABLE II
OBSERVATIONAL OUTCOMES OF SCENE DATASET

Algorithms	Hamming Loss	Accuracy	Precision	Recall
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
EML_M	0.084	0.699	0.730	0.716
EML_T	0.095	0.694	0.725	0.754
ADMLCAR	0.068	0.823	0.883	0.892

It is evident from the outcomes that ADMLCAR outruns almost all the present multi label classification methods. In

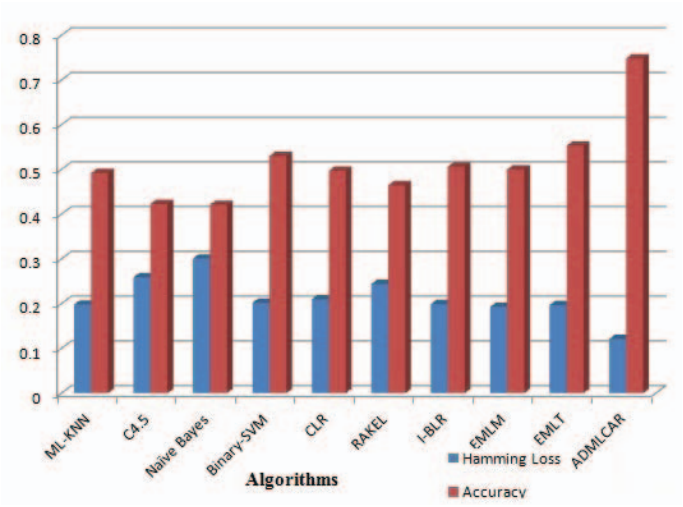

Fig. 5. Accuracy and Hamming Loss of our proposed method with different methods.

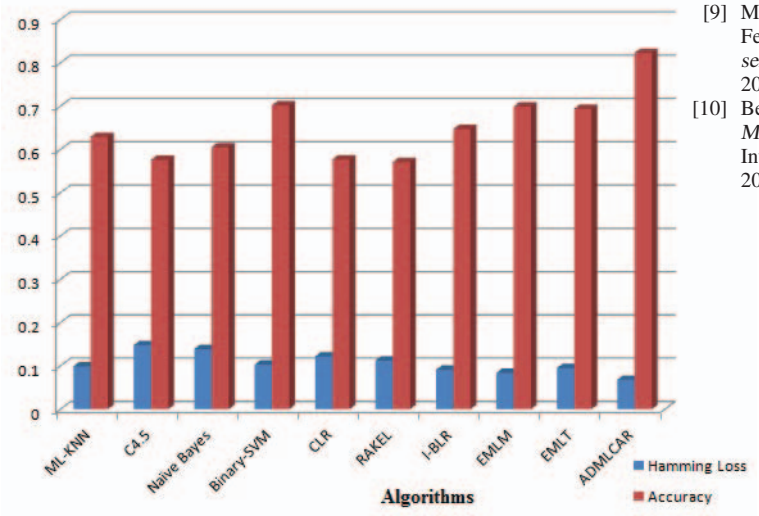

Fig. 6. Accuracy and Hamming Loss of our proposed method with different methods.

the experiment the k in k-means is fixed as 4. The minimal threshold of Vertical data format is fixed as 25% and minimal confidence=75%. The algorithms ML-KNN, CLR, RAKEL, I-BLR, EML_M, EML_T evaluation results on yeast and scene values were taken from [9] and the evaluation results of C4.5, Naive Bayes and Binary-SVM were taken from [10].

V. CONCLUSIONS

In this work we developed a better multi label classification approach other than algorithm adaption and problem transformation. From the obtained result, our proposed method works better and give better results when compared with several alternative multi label classification algorithms. In future we can use other clustering algorithms and association rule mining method instead of vertical data format. Also,we can use some other multilabel dataset for predicting.

REFERENCES

[1] GUO Yi-ming, WANG Zhi-jun,*A vertical format algorithm for mining frequent item sets*,International Journal on Advanced Computer Control(ICACC), 2010.

[2] Asha Ashok, Geethu U,*Outlier detection using a clustering based oversampling principal component analysis*, International Conference on ijret, 2015.

[3] Feng Qin,Xain Juan Tang, Ze-Kai Cheng. *Application of Apriori Algorithm in Multi label classification*, International Conference on Computational and Information Sciences, 2013.

[4] Purvi Prajapati, AmitThakkar, AmitGanatra,*A Survey and Current Research Challenges in Multi-Label Classification Methods*, International Journal of Soft Computing and Engineering (IJSCE),March 2012.

[5] Prathibhamol C.P, Asha Ashok, *Solving Multi Label Problems with Clustering and Nearest Neighbor by Consideration of Labels, Advances in Signal Processing and Intelligent Recognition Systems*, Springer International Publishing, pp. 511-520, 2016.

[6] Susan Tony, Ujjwal Harode,*A Comparative Study On k-means and Hierarchical Clustering*, International Journal of Electronics, Electrical and Computational System (IJEECS), February 2015.

[7] Bhupesh Akhand, V.Susheela Devi, *Multi Label Classification of Discrete Data*, IEEE International Conference on Fuzzy sets,2013.

[8] Y.-R. Yeh, Z.-Y. Lee, and Y .-J. Lee, *Anomaly Detection via Oversampling Principal Component Analysis*, Proc. First KES Intl Symp. Intelligent Decision Technologies, pp. 449-458,2009.

[9] Muhammad Atif Tahir, Josef Kittler, Krystian Mikolajczyk, and FeiYan,*Improving Multilabel Classification Performance by Using Ensemble of Multi-label Classifiers*, Springer-Verlag Berlin Heidelberg 2010.

[10] Benhui Chen, Xuefen Hong, Lihua Duan and Jinglu HU, *Improving Multi-label Classification Performance by Label Constraints*, IEEE International Joint Conference on Neural Networks (IJCNN), August 2013.