IJCTA, 9(33), 2016, pp. 183-189

An Improved Multi-Label Classifier using Data Gravitation Method

Rose S* Krishnadas S* and Ranjidha Rajan*

Abstract : Multi-label classification is a supervised classification task where each data instance may be associated with more than one class labels. Data gravitation based classification (DGC) is a technique used for classifying class labels based on the theory of gravitation. In the classifier the force between two data instance is computed and the same is used to find the best rules for a multi label classification. Here in the paper we focus on applying the data gravitation principle to select the most relevant label or labels from a set of transformed data set. The gravitation between the labels is calculated considering there given weights and the best rules are formed using the labels having highest force. A multi label classifier is learned based on the gravitation force of labels. The final rules are used to predict a new instance. Here in this paper unlike other methods we use a new method including both problem transformation and an algorithm adaptation method name CL_DGC for a better prediction results. The experiments were done using different transformation methods and compared with using the transformed data set in a normal classifier and data gravitation force method for four different benchmark data set from UCI Solar-flare, Music, Enron, and Scene. The results were verified using evaluation measures like precision, recall and accuracy of the classifiers.

Keywords : Machine learning, Multi label classification (MLC), data gravitation based classification (DGC), Label correlation.

1. INTRODUCTION

In machine learning task, single label classification is a common learning method where the classifier learns from a set of instances, in which each instance is associated with a single class label from a set of different class labels. If the class label is just one it is said to be single label classification and if there are multiple class label it is said to be multi label classification. The main aim in multi-label classification is to learn from a set of instances where each instance may be belonging to one or more classes.

There are mainly two approaches used for multi-label learning namely Problem transformation approach and Algorithm adaptation method. The first approach produces single label dataset from multi label dataset and later applies any single label classifier to the problem. Second approach classifies multi-label datasets by adapting certain algorithms directly. Two main methods used in problem transformation are namely Binary Relevance (BR), Label Powerset (LP)[11]. In Binary classification, it decomposes a multi-label classification problem into several distinct single-label binary classification problems. In Label Powerset method, it finds each unique class labels in a training dataset as one class in the new dataset called transformed dataset. So that the new dataset is a single label one.

Ranking of labels can be done using LP method. In this paper we are utilizing LP method to convert multi-label datasets into single label. It considers label correlations so we can assign weights for each label. To that can apply different algorithm adoption methods. Examples are boosting methods, k-nearest neighbors that uses ML-KNN algorithm, and Decision trees that uses C4.5 algorithm which is an improved version of Id3 algorithm.

* Department of CS & IT Amrita School of Arts and Sciences Kochi, Amrita Vishwa Vidyapeetham, Amrita University, India rosesivan@gmail.com, krishnadasofficial92@gmail.com, ranjidha@yahoo.com

In the proposed approach we are using LP method [1]to convert the multi-label problem into single label and data gravitation based classification (DGC)[7] method is applied to that transformed dataset. First we transform multi-label dataset into single label dataset and finds correlations between the labels. Data gravitation is defined as the relationship between data elements. The basic principle behind DGC classification method is to classify samples by comparing the computed data gravitation among the different classes. If the value of gravitation is greater for a sample then the sample belongs to that particular class. The main advantage of this method is that the procedure is comparatively simple and performance is much higher than other methods.

In 1687 Newton proposed the universal law of gravitation in his paper. According to the law the mutual force (F) between two bodies is given by the equation:

$$F = G\frac{m1m2}{r^2}$$

Where '$m1$' and '$m2$' are mass of objects under consideration and 'r' is the distance between them. The constant of proportionality, G, is the gravitational constant.

According to the law of gravitation the strength of gravitation between to data particle in the data space is directly proportional to the data masses and inversely proportional to the distance between two data particles. There are certain methodologies used in this proposed work like; data particle, it's a data unit with a data mass. Data mass is the sum of data elements inside a data particle; these data elements have an attraction among themselves.

The main feature used in this method is the distance between the data particle .It can be calculated using following equation given

$$r = \sqrt{\sum_{i=1}^{n} w_i (x_{1i} - x_{2i})^2}$$

Classification is done using the procedure, suppose a1 and a2 are two classes in a training data set. For given test data component *k*, gravitation force that data particles in '$a1$'and '$a2$' acts on *k* is Force1 and Force2 consecutively. If Force1 > Force2, then k can be influenced by '$a1$' than '$a2$'.

Two factors are often considered when DGC method is used, one is the distance, and other is the number of samples in the group. The shorter the distance between the samples, the more the single data is related to the group; to measure the performance of the system various evaluation measures Precision, Recall, Accuracy are calculated. Precision is the amount of correct predictions made, and recall is the amount of correct samples that are predicted correct.

$$\text{Precision} = \frac{TP}{TP + FP} \text{ and Recall} = \frac{TP}{TP + FN}$$

Where TP is the amount of true positives that is correctly predicted positive samples, FP is the amount of false positives that is positive predictions that are incorrect, and FN the amount of false negatives that is positive examples that are incorrectly predicted negative.

2. LITERATURE REVIEW

Multi-label classification [11] is considered an efficient way to classify datasets. Because instances can be assigned to multiple labels it has more real world applications than ordinary single label classification. There are a number of classifiers used to classify every kind of datasets. Many researchers provide new classifiers and improvements to existing classifiers which made them more efficient.

This paper presents an overview on various multi-label classification techniques and various measures used. Most of the classification problems correlate a single class to each instance. However, there are many real life classification tasks as well where each instance should be related with one or more classes. This problem can be solved by using multi-label classification. The article explains various frequently used classification methods, similarities and differences between them.[2]

Raed Alazaidahet et al. proposes a multi-label classification method based on correlations among labels that makes use of both problem transformation approach and algorithm adaptation approach. The approach transforms multi-label dataset into a single label dataset using least frequent label, and also finds correlations among labels using predictive Apriori algorithm, then applies the PART algorithm on the dataset. The output of this approach is multi-labels rules.[1]

In this paper they propose Data Gravitation Based classification (DGC) in which DGC algorithm is used to classify data by comparing the data gravitation between different classes. Greater the gravitation value from a class the sample is belonging more to that particular class. Feature selection plays an important role in this approach. They apply this method in different data sets from UCI machine learning repository. [7]

Oscar Reyeset.et al. have focus on multi-label lazy algorithm based on the data gravitation model, known as ML-DGC. ML-DGC handles the multi-label data and it considers each instance as a single data particle. They used 34 multi-label datasets in the method and the results of the experiments showed that the performance is much higher than other lazy learning methods. [8]

In their paper Hiteshri Modiet.et al. provides an experimental comparison of different problem transformation methods for multi-label classification using MEKA software. They performed experiments on a number of multi-label datasets and various evaluation measures like hamming loss, exact match, accuracy etc of various problem transformation methods are compared. As a result of these experiments they found that LC and PS give better results than BR method. The reason behind that is that LC and PS methods consider label correlation while converting multi-label to single label while BR method does not.[6]

3. MATERIALS AND METHODOLOGIES

The proposed method combines two classification methods to make it more efficient namely correlation based classification technique using LP method and data gravitation based classification. The major advantage of LP method over BR is that it considers label dependency. For example, if a sample is related with three labels L_2, L_4, L_6 then the new single-label class will be $L_{2, 4, 6}$. So that the new dataset formed is a single-label classification task and then any base classifier can be applied to the transformed dataset. For this classification purpose we are considering different datasets from UCI datasets repository namely Solar-flare, Music, Enron and Scene. In solar flare dataset we considered 323 instances and 13 nominal attributes.

We transform multi-label dataset into single label dataset and finds correlations between the labels. Thus we can find weights of each label. Then we apply DGC method to the dataset in which the concept of data gravitation is applied. Data gravitation [7] is defined as the relationship between data elements. Classification is done using the multi label k- nearest neighbor algorithm. Suppose a1 and a2 are two classes in a training data set. For a given test data element k, let the gravitation that data particles in a1 acts on k and a2 act on k is Force1 and Force2 respectively. If Force1 > Force2, then the data element k belongs more to a1 than a2.

For classification and preprocessing purposes we are using MEKA tool [9]. In which we initially preprocess the dataset. LP method is used to convert to single label task. For that we used methods, supervised “attribute_selection” filter and “information gain” evaluator and “Ranker” search method. Then we make use of data gravitation method and finds relationship among the labels then it can be classified based on the gravitation value using multi-label KNN classifier. Different rules are formed using the results tool is used to find evaluation measures like accuracy, precision, and recall for which confusion matrix is made.

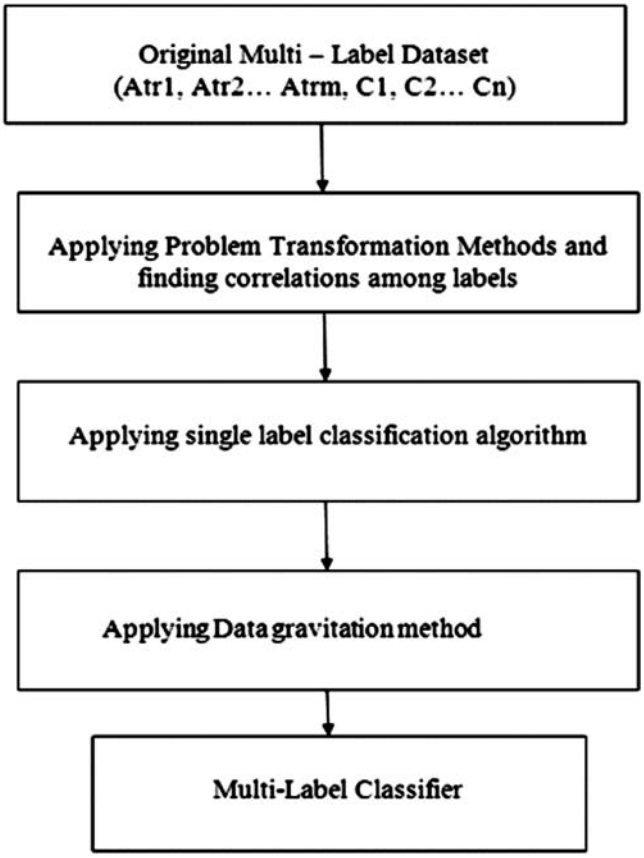


Figure 1: The work flow of the proposed approach

Algorithm 1: An improved Multi-Label Classification approach based on Label Correlation using Data Gravitation method.

Input: Multi-label dataset as training data.
Output: A set of Multi-Label rules.

Step 1: *Dataset Transformation:* Transforming multi-label dataset into a single label dataset by applying Label Powerset method.

Step 2: *Learning:* For every label in the label set of the dataset, find the correlation and weight, then applying the data gravitation method and finding the gravitational force for each label.

Step 3: *Classification:* Generating the multi-label rules set, by combining the single rules set produced by the classifier in Step 2.

Figure 2: The main steps of the proposed method Algorithm

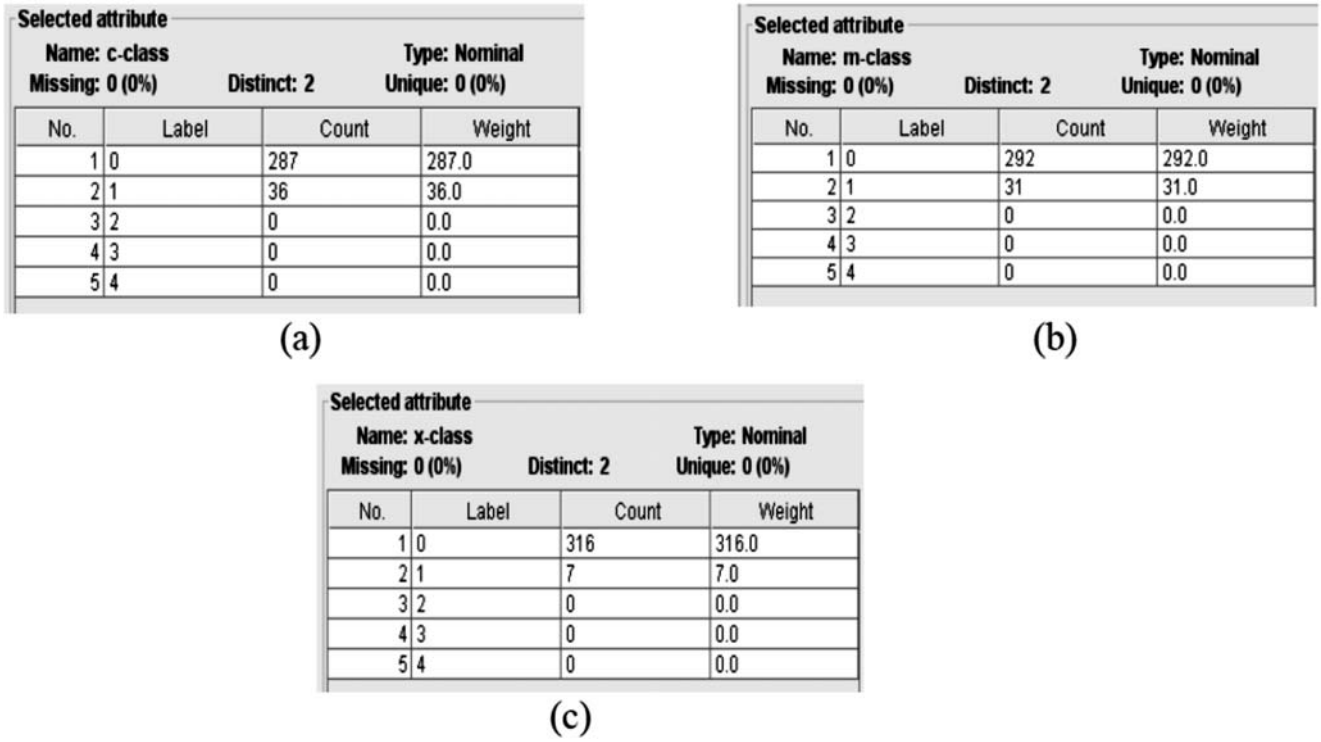
4. EXPERIMENTAL RESULT

4.1. Preprocessing

We applied “Attribute_Selection” filter which is a supervised attribute filter that can be used to select attributes. It is flexible method that allows various search and evaluation methods to be combined. And

"InfoGainAttributeEval" evaluator method which evaluates the importance of an attribute by measuring the information gain with respect to the class and Ranker search method which ranks attributes by their individual evaluations. It can be used in combination with attribute evaluators.

We will get a set of preprocessed classes from which we can find count of each attributes. An example of attribute evaluation is given below. The attribute "spot_distribution" has the labels c-class, m-class, x-class, and each label count is given.



Selected attribute
Name: c-class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	287	287.0
2	1	36	36.0
3	2	0	0.0
4	3	0	0.0
5	4	0	0.0

(a)

Selected attribute
Name: m-class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	292	292.0
2	1	31	31.0
3	2	0	0.0
4	3	0	0.0
5	4	0	0.0

(b)

Selected attribute
Name: x-class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	316	316.0
2	1	7	7.0
3	2	0	0.0
4	3	0	0.0
5	4	0	0.0

(c)

Figure 3 : (*a*), (*b*), (*c*) represents the summery of 3 labels in the Solar-flare dataset

Table 1

Frequency of labels

Labels	***c-class***	***m-class***	***x-class***
Frequency	36	31	7

4.2. Calculating weights for each label

The next step is converting the dataset into single label dataset. For that LP method is used which finds each unique class labels in a training dataset as one class in the new transformed dataset. So that the new transformed dataset is a single label classification task. Ranking of instances can be done using label powerset method. Each label is assigned with a weight. The distance between each labels are calculated using the equation and from that data gravitation force for label is found. The instanced in the dataset are classified using this gravitation value. Instances are assigned to the label with highest value of gravitation.

This proposed model is applied in different multi label datasets and derived results. The above described model is able to obtain high classification performances during many classification cases, especially for the balanced data sets such as Solar flare and other data sets. For these data sets, the numbers of instances of different classes are almost the same and the data gravitation is also balanced. Our experiment results also show that if there is a data class which contains unbalanced data i.e. data with highly varying values the DGC approach may fail because data gravitation may also become unbalanced.

Table 2
Label Ranking

Label Ranking Label Powerset Method : Calculation of weights				
Labels	***P(c/s)***	***L1***	***L2***	***L3***
L1	0.07121	1	0	0
L2	0.003096	0	1	0
L3	0.05573	0	0	1
L1, L2	0.006192	1	1	0
L1, L3	0.03096	1	0	1
L2, L3	0.009288	0	1	1
L1, L2, L3	0.003096	1	1	1
Weights =		0.111458	0.0217	0.099074

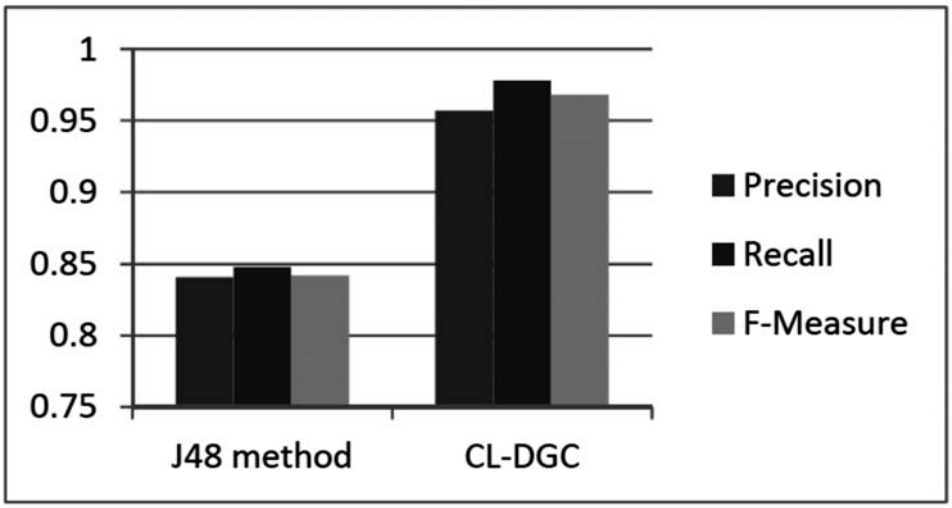


Figure 3: Methods J48, CL-DGC Vs Accuracy

5. CONCLUSION AND FUTURE WORK

In this paper, an improved label correlation based data gravitation based classification method is combined with classification method. An important feature of DGC is that it is a simple classification procedure which is also very easy to implement and with more accuracy.

In the process of multi-label classification where we transform multi-label dataset into a single-label dataset, loss in information might occur.CL-DGC method does not result any information loss because when correlation of labels is found, values are substituted. This method is much more flexible as any base classifier could be used in the process of classifying the transformed data set. Experimental results show that this method has got a better precision, recall and accuracy when compared to other approaches. The proposed methodology proves to be efficient for larger data-sets to get faster results.

Future work is to consider correlation among attributes to apply different base classifiers for each transformed dataset for better performance.

6. REFERENCES

1. Raed Alazaidah, Fadi Thabtah, Qasem Al-Radaideh .A Multi-Label Classification Approach Based on Correlations among Labels *(IJACSA) Vol. 6, No. 2, 2015*
2. André CPLF de Carvalho, Alex A Freitas. A Tutorial on Multi-Label ClassificationTechniques.
3. Alberto Cano, Amelia Zafra, Sebasti´an Ventura .Weighted Data Gravitation Classification for Standard and Imbalanced Data.
4. Tsung Hsien Chiang, Hung-Yi Lo, Shou De Lin. A Ranking-based KNN Approach for Multi Label Classification. JMLR: Workshop and Conference Proceedings 25:81{96, 2012.
5. Hiteshri Modi, Mahesh Pancha. Experimental Comparison of Different Problem Transformation Methods for Multi-Label Classification using MEKA. Volume 59 No.15, December 2012
6. Lizhi Peng , Bo Yang, Yuehui Chen, Ajith Abraham. Data gravitation based classification. Information Sciences, 2009 - Elsevier
7. Oscar Reyes, Carlos Morell, Sebasti´an Ventura. Effective lazy learning algorithm based in data gravitation model for multi-label learning.
8. Jesse Read.Tutorial. Meka 1.7.6, A Multi label/multi target Extension to WEKA. June 2015.
9. Mohammad S Sorower. A Literature Survey on Algorithms for Multi-label Learning.
10. Grigorios Tsoumakas, Ioannis Katakis. Multi-Label Classification: An Overview.
11. Celine Vens, Jan Struyf, Leander Schietgat, Sa˘soD˘zeroski, Hendrik Blockeel. Decision Trees for Hierarchical Multi label Classification.
12. Min-Ling Zhang, Zhi-Hua Zhou. Ml-knn: A Lazy Learning Approach to Multi-Label Learning.