Survey of User Interested Optimized Rule Mining with Artificial Bee Colony In Document Classification

Drashti Panchal¹, Mohit Patel²
¹Student, Swaminarayan College Of Engineering And Technology, Gujarat Technological University Ahmedabad, Gujarat, India
²A/Prof. Computer Swaminarayan College Of Engineering And Technology, Gujarat Technological University, Ahmedabad, Gujarat, India

Abstract—Document classification is problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. In my base paper they have used "Association Rule Mining" for Document Categorization. So I am going to used Optimized Rule Mining with Artificial Bee Colony (ABC) algorithm recently introduced optimization algorithms to simulate intelligent foraging behavior of honey bee swarm was proposed by Karboga and Ozturk. We will customize the ABC Algorithm to improve accuracy of Document Classification.

Index Terms—ABC Algorithm, Document classification, Rule Mining

I. INTRODUCTION

Data mining is to extract or mine knowledge from a lot of data called Knowledge Discovery in Databases (KDD), which is the result of information technology natural which is the result of information technology natural evolution. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to efficiently conclude the target class for each case in the data. For example a classification model could be used to identify loan applicants as low, medium or high credit risks.[1]

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label \( l \) from a set of disjoint labels \( L \), \( | L | > 1 \). If \( | L | = 2 \), then the learning problem is called a binary classification problem (or filtering in the case of textual and web data), while if \( | L | > 2 \), then it is called a multi-class classification problem.[1]

In multi-label classification, the examples are associated with a set of labels \( Y \subseteq L \). In the past, multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Text documents usually belong to more than one conceptual class. For example, a newspaper article concerning the reactions of the Christian church to the release of the Da Vinci Code film can be classified into both of the categories Society\Religion and Arts\Movies. Similarly in medical diagnosis, a patient may be suffering for example from diabetes and prostate cancer at the same time.[1]

II. TYPES OF DOCUMENT CLASSIFICATION

• Content-based Classification: is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. It is, for example, a common rule for classification in libraries that at least 20% of the content of a book should be about the class to which the book is assigned. In automatic classification it could be the number of times given words appears in a document.[2]

• Request-oriented classification: (or-indexing) is classification in which the anticipated request from users is affect how documents are being classified. The classifier asks himself Under which description should this entity be found and think of all the possible queries and choose for which particular the entity at hand is applicable.[2]
Association Rule Mining

- As shown in Figure this algorithm finds rules that are associated with a class label.
- Here each set of documents that belong to one class is considered to generate association rules for that particular class.
- If a document belongs to more than one class than it will be present in each set of documents associated with the class label that the document falls into.[3]

III. THE PROPOSED MULTI-LABEL META-LEARNING FRAMEWORK

This section outlines the new multi-label meta learning framework proposed for recommending classification algorithms for educational data. The approach can be split into two phases, as shown in Figure 1.

In the training or offline phase, the final goal is to generate a multi-label data set from educational data sets. To this end, several steps need to be addressed. In Step 1, a set of classification algorithms are executed over the original single-label educational data set so that several classification measures are calculated. Note that the algorithms selected at this point are those that will be recommended at the end of the process.

In Step 2, the algorithms that perform best must be found for each data set. The multiple-comparison Friedman or Iman & Davenport statistical tests can be employed for this purpose. Both tests compare the mean ranks of k algorithms over N evaluation measures. These ranks indicate which algorithm obtains the best results considering all the measures studied. To calculate them, a rank of 1 is assigned to the algorithm with the highest value in the first Measure studied, the algorithm with the next highest Value in this measure is given a rank of 2, and so on. The same procedure is then carried out for the other Evaluation metrics involved in the study. According to the test performed, it is possible to find out if the algorithms present significant differences in performance among themselves, according to the classification evaluation measures studied. If there are significant differences, a post-hoc test must then be performed to reveal such performance differences. Several statistical tests can be used at this point, such as Bonferroni-Dunn, Holm’s and Hochberg’s methods. As result, we will know the subset of algorithms recommended for each particular data set. Note that the algorithms recommended will not present significant differences among themselves regarding the classification metrics evaluated.

On the other hand, the meta-features of each original data set in the repository are extracted in Step 3, such as statistical, complexity and domain features. Then, in Step 4, the meta-features extracted for a given data set will become part of an instance of the multi-label data set. Specifically, the meta-features extracted will correspond to the predictive attributes, while the value of the labels will come from the subset of algorithms recommended for the same data set. The number of instances in the multi-label data set is equal to the number of single-label educational data sets in the repository, and the number of labels is equal to the number of algorithms employed in Step 1. Note that, for a given instance, a value of 1 will be set in a label if the algorithm associated belonged to the subset of algorithm recommended in Step 2, and 0 otherwise.

Finally, Step 5 consists in training a multi-label classifier using the multi-label data set generated as
training data. Any kind of multi-label classification algorithm can be employed to generate the classifier. In the prediction or online phase, given a new educational data set, its meta-features must first be extracted in Step 6. The same meta-features used to generate the multi-label data set should be extracted. Then, in Step 7, the values of these meta-features can be used as input for the multi-label classifier, which will generate the algorithms recommended for the new data set by prediction.[4]

IV. ABC ALGORITHM

The ABC algorithm is based on bee’s behavior in finding the food source positions without the benefit of visual information (Karaboga and Ozturk, 2011). The information exchange from bees is assimilation knowledge about which path to follow and quality of food through a waggle dance. Bees calculate their food source using probabilistic selection and abounding source by sharing their information through waggle dance and food source with less anticipation of producing new food source in neighborhood of old source in relation to their useful. The ABC has three necessary components: food source, employed bee, scout bee and onlooker bee, and the behaviors are: selection and rejection of the food source.[4]

- Employed Bee: The employed bees store the food source information which includes the gap the guidance and share with others according to a certain probability and shares with other bees waiting in the colony abundance and extraction of energy nectar taste and strength of the solution.
- Onlooker Bee: It takes the information from selected numbers of employed bee and decides the probability of higher nectar amount information of the food source is selected according to beneficial of food source.
- Scout Bee: If the position of food source is not better through best number of cycles food source will be removed from the community employed bee becomes a scout bee and names a new random food source. Based on the performance of strength value if the named new food source is better than dismiss one then scout bee becomes employee bee. This process is repeated until the maximum number of cycles to determine the optimal solution of food source.[5]

ABC algorithm is a new swarm intelligent algorithm and consists of three essential components:
1. Food Sources: It represents a position of solution of the problem.
2. Employed Foragers: The number of employed bees is equal to the number of food sources. The employed bees store the food source information and share with others according to certain probability.
3. Unemployed Foragers: Their main task is exploring and exploiting food source. There are two choices for the unemployed foragers (i) It becomes an onlooker and determines the nectar amount of food source after watching the waggle dances of employed bee and collection food source according to beneficial (ii) It becomes a scout and randomly searches new food sources around the nest.[5]

V. METHODOLOGY

In this section, we are going to describe proposed methodology as shown in Fig. This includes the Document used, pre-processing phase, Rule Generation, Prediction phase and Testing Phase.[1]
1) Tokenization: The input documents from the training set are divided into tokens. We have used the java in built classes for tokenization purpose.[1]

2) Dimensionality Reduction: Dimensionality reduction for training documents is done by calculating the document frequency (DF) i.e the number of documents that involves a particular term DF is the essential process to dimensionality reduction.[1]

DF removes un-repeated words. For all those terms or words that appears in less than “N” documents of all text documents are not treated as features, where “N” is a pre-assumed. The value of N is different for different sets of documents for example we have taken the value of N as 50 for Reuters documents set.[8]

3) Term Weighting: Each term has its own importance in a document. The two main factors are term frequency and inverse document frequency. Term frequency is the ratio of the count of a term in a document to the entire number of terms in that document.

\[ TF = \frac{\text{term count in } d}{\text{total term in } d} \]

Where d is particular document. Inverse document frequency can be described as the ratio of log of whole number of documents to the number of documents in which that term appears.

\[ IDF = \log \frac{N}{\text{no. of document in which } t \text{ appears}} \]

Where, N indicates the entire number of documents in the training documents set.

We have removed all that terms from each document that have very high tf-idf and very low tf-idf because that term will not be so much important for classification purpose. So we have only taken the mid-range value of tf-idf and only that terms will be retained in our document that have the values of tf-idf lying in that range and others are deleted from the documents.[1]

B. Rule generation phase:

Generation of Association rules: Association rule mining is the technique of data mining that finds out relationships between items in a huge transactional database. Suppose we have a set of items I= \{i1, i2,…im\}, in a transactional database D. Association rules are those in which we find out relationships among these item sets. An association rule is an implication of the form A→B, where A, B are subset of I and A∩B=null.[1]

C. Classifier Construction Phase:

We will consider every rule generated for each class and check every rule one by one. If a rules antecedent part matches the training document than we will keep that rule. If a rule doesn’t matches any training document it will not be considered for building classifier. Once all those rules which matches training documents are sorted according to support and confidence values. and confidence values. Finally only those rules which have support and confidence greater than or equal to minimum support and confidence are kept for building the classifier finally.[1]

D. Prediction Phase:

The procedure for predicting the class label of the test documents and to evaluate the accuracy of our classifier, so we ensure if the rule actually matches test data in our model when a test data to be predicted.[1]

E. Testing Phase:

About the various performance evaluation measures for documents classification algorithms like K-NN[6] Naive Bayes[7] MCAR[8] and we will compare the results of these algorithms with our proposed algorithm. The various evaluation measures are precision, recall, f1 measure. The dataset used in the experiments is Reuters TC corpus and 20 Newsgroup. The main purpose of choosing these algorithms is that they employ different strategies for rule finding.

We will also compare our model with other AC technique like MCAR [8]. All these algorithms have different features regarding rule discovery like Naïve Bayes [7] is a probabilistic classification algorithm. Also KNN [6] depends on the similarity among the training documents and number of training documents, then determining the class label of document. Finally, MCAR [8] is an AC algorithm that is based on the discovery of attribute values using hidden correlations among the class a
VI. CONCLUSION

As document classification is key point to data mining process, it is very important to assign proper label to particular document for proper usage. Generally single Label does not satisfy the overall requirement of labeling the document. We will try to set the multiple label to the documents using Clustering the document and Applying the Association Rule Mining Technique. It will assign proper multi labels to the Document.

REFERENCES

[11] (11) Bo Li, Hong Li, Min Wu, PingLi, “Multi-label Classification based on Association Rules with Application to Scene Classification”, 2008,IEEE, pp. 36-41