Text Categorization Using Multi Label Text And Pyramid clustering Mechanism

Kiran Kumar Thanniru
Assoc. Prof.
Department of CSE
Siddhartha Institute of Technology and Sciences
Narapally, Hyderabad, Telangana, India

Dr.Arun Prasath Raveendran
Professor
Department of ECE
Siddhartha Institute of Technology and Sciences
Narapally, Hyderabad, Telangana, India

***Abstract* - In the recent days, text categorization is gaining popularity in the fields of machine learning, information retrieval, and natural language processing (NLP). Text classification challenges have sparked widespread interest in a variety of fields, including a. news selection and classification, b. document classification in digital libraries, social networks, and websites, and c. e-mail categorization, which includes spam filtering. A multi-label text classification issue is a variation of this problem in which each document can be connected with a number of labels or classes. A hierarchical text classification issue is an extension of such an issue in which the classes are interconnected by a specified hierarchy. We evaluate the performance of multi label text classification studies using a real-time dataset in this study. The algorithm is used to take advantage of class hierarchy and investigates the impact of various algorithmic techniques and dataset attributes on classification performance.**

Key words: Pyramid clustering, Multi label, NLP

1. INTRODUCTION

Multi label classification (MLC) is a topic that has gotten a lot of attention because it is applicable to a wide range of areas such as biology and music categorization, etc. Nonetheless, situations in which individual instances are linked to many classes remain difficult. The tasks of MLC are treated as multiple binary classification problems by the majority of task classification algorithms. Furthermore, this method may overlook potential relationships between characteristics and classes. A better MLC solution should be both efficient and effective; nevertheless, a large number of irrelevant and redundant variables may raise the cost of communication and the time required to learn and analyze multi-label classifiers, lowering classification performance. Feature selection is an important operation in data mining and machine learning approaches, and it has been widely used in classification frameworks to improve performance. Choosing features before applying classification techniques to unique datasets provides a number of advantages, including filtering information, lowering computing costs, and improving classification precision [2, 3]. In this way, we employ a feature selection strategy to improve MLC's standard.

2.RELATED WORK

2.1Classifications

Data mining classification strategies are capable of managing large amounts of data. It can be used to predict specific class labels and categorize data according to the training set and class labels, which can be used to characterize recently available data. This word can be applied to any situation in which a few decisions or forecasts are made based on newly available data. The classification technique is regarded as a strategy for making such decisions in novel situations on a regular basis. If it is assumed that the issue is a concern with the creation of a system that will be connected to a series of scenarios in which each new scenario must be assigned to one of a set of pre-defined class sets based on recognized information features, Supervised learning or pattern recognition is the process of creating a classification method using a set of data for which the right classes are already known. For example, allocating people to credit status based on money-related and other individual data, and the underlying determination of a patient's disease to choose speedy treatment while anticipating acceptable test outcomes are examples of contexts where a classification operation is important. Decision or categorization issues are probably the most basic issues that are emerging in business, industry, and research. There are three distinct authentic research strands: neural networks, machine learning, and statistical systems. In general, all classes have a few objectives. They've all tried to come up with a methodology that could deal with the situation.

2.2 Classification algorithms

Classification is one of numerous data mining approaches that is mostly used for reviewing the provided datasets and accepting every instance of them and allocating such instance to a certain class so that classification errors are minimized. It's used to create

distinct models that accurately characterized key information types within given datasets. There are two stages to classification. The model is created in the beginning by using classification algorithms on training datasets. The extracted model is then compared to a preset experimental dataset to determine the model's trained execution and precision in the second step. As a result, classification is the process of assigning class labels to datasets with ambiguous class labels.

2.3J48 Algorithm

J48 is the extended form of ID3. Missing measure representation, decision tree shortening, unbroken attribute estimate ranges, rule derivation, and so on are some of J48's secondary features. WEKA is one of the data mining tools, and J48 is a Java implementation of the C4.5 algorithm that is also open source. The WEKA tool provides a number of tree-pruning options. If there is an occurrence of suspected overfitting, pruning could be used as a precision technique. The classification of the information is done in many approaches till each and every leaf is ideal, i.e. the classification of the information must be as perfect as possible. This technique establishes the principles around which the information's unique personality is based. The goal is to simulate a decision tree in real time until it achieves a balance of adaptability and precision.

2.4 K-NN Classification

K-nearest neighbors (k-NN) is a non-parametric technique and a form of classification algorithm used for pattern recognition and regression [1]. The input in both cases consists of the k closest training instances in the feature spaces. The outcome is determined on whether k-NN is used for classification or regression. The predicted output of this k-NN classification is class membership. Objects are classified by a majority vote of their neighbors, with the objects being assigned to the most basic class among their k closest neighbors, where k is a positive whole number that is usually small. Let's use k = 1 as an example. The items are then simply assigned to the class with the smallest number of nearest neighbors. The property estimation for the objects is the outcome of the k-NN regression. This metric is the average of their k closest neighbors' estimates.

2.5Random forests Algorithm

The random forest algorithm is a group learning technique for classification, regression, and other operations. It works by constructing a large number of decision trees during training times and determining the class that is the mean prediction (for regression) or classes (for classification) of the individual trees.

2.5 Multilayer perceptron

Multilayer perceptron (MLP), which has at least three layers of nodes, is one sort of feed-forward artificial neural network. With the exception of the input nodes, each node functions as a neuron with a nonlinear activation operation. MLP performs training using back propagation, a supervised learning paradigm [1] [2]. MLP is distinguished from linear perceptron's by its non-linear activation and numerous layers. It can recognize information that isn't easily distinguishable [3]. Multilayer perceptron are also referred to as "vanilla" neural networks, particularly when they have a single hidden layer [4].

2.6 Data mining Tools

WEKA, like SAS Enterprise Miner, is a data mining suite, but unlike SAS Enterprise Miner, it is an open source code that is available for free. WEKA is a better tool to use if someone needs to change the algorithm's source code. WEKA, which includes C4.5, also known as J48, allows for the re-implementation of various traditional data mining algorithms. WEKA has a significant advantage over SAS Enterprise Miner in that the Enterprise Miner is only accessible via a graphical user interface (GUI), making it difficult to automate testing, which is common in research when multiple variations of an analysis are required. WEKA, on the other hand, offers a unique operation mode that makes experimentation simple.

3. Methodology

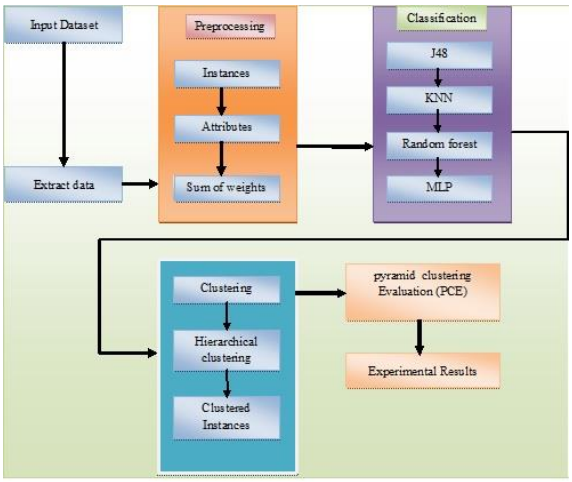

Fig-1 Architecture Diagram

3.1 Data set

A dataset is a collection of data organized in a specific format. In general, data sets refer to the content of a single database table or a single statistical data matrix in which each portion of the table represents a single variable and each row corresponds to a specific dataset in query. The dataset executes a listing procedure of values for each dataset member for each variable, such as item height and weight. A datum is a term that refers to any value. The number of rows in a dataset determines whether it contains information for a single or several people. The term dataset can also be used

more loosely to refer to the information contained in a collection of closely linked tables pertaining to a certain test or event. The datasets obtained by space research agencies conducting experiments with various devices on board space tests are an example of this type. Big data [1] refers to datasets that are so large that traditional data processing applications are unable to handle them. Datasets are the unit for analyzing data issued in an open information repository in the open information discipline. The European Open Data portal [2] contains almost five million datasets. Different definitions have been proposed in this area [3], but there is currently no official one. Different issues, such as real-time information sources, non-social datasets, and so on, make it difficult to reach an agreement.

3.2 Data Preprocessing

Data pre-processing is considered a necessary step in the data mining process. The phrase "garbage in, trash out" is especially apt when it comes to data mining and machine learning applications. Data collection procedures are frequently approximated, resulting in out-of-range measurements (such as Income: $200), unworkable data integrations (such as Sex: Male, Pregnant: Yes), missing measures, and so on. Investigating data that hasn't been purposefully filtered for these concerns can lead to misleading results. In this way, characterization and quality of data are prioritized before running an analysis [1]. Data pre-processing is, in general, one of the most important stages of a machine learning process, especially in computational biology [2]. Knowledge discovery during the training stage becomes more difficult if there is more needless and repeated data or noisy and irregular information. The processes of data production and filtering might consume a significant amount of data processing time. Filtering, instance determination, standardization, feature selection, feature extraction and modification, and so on are all examples of data pre-processing. The last training set is the result of data preparation. For each stage of data pre-processing, Kotsiantis et al. (2006) offer a popular technique [3].

3.3 Pyramid Clustering

A tree-structured hierarchy is used to organize the labels in numerous text categorization systems. A case is linked to a given label only if it is also linked to the label's parent in the hierarchy. Furthermore, the structure of the labels is never considered in most classic multi-label categorization algorithms. Rather, the labels are treated as separate entities, necessitating the training of a large number of classifiers, one for each label. Furthermore, because a few leaf labels may have a lower number of positive images, the training data may be biased, causing problems in many classifiers. In addition, the inconsistent labelling between parent and kid complicates interpretation. Finally, because structural conditions among labels are not exploited in the learning process, prediction performance is harmed. The following are some current methodologies: if a parent of a label is predicted, only the positive prediction can perform for that label; generate training drawings for every node from testing of the parent node; use structured predictors with huge margins, or alter decision trees.

4. Experimental Results

We used four datasets in our analysis, including diabetes, labor, segmentation, and soybeans. J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer Perceptron are four distinct classifiers used for classification (MLP). Precision, recall, F-measure, MCC, ROC, and PRC are some of the characteristics that are measured and the results are summarized.

Table 1: Overall Comparison of better classification using and its Pyramid cluster evaluation

Evaluation parameter	J48	KNN	Random Forest	Multilayer perceptron
Specificity (%)	0.63	0.5	0.67	0.66
Sensitivity (%)	0.79	0.76	0.8	0.80
Accuracy (%)	0.74	0.69	0.76	0.75

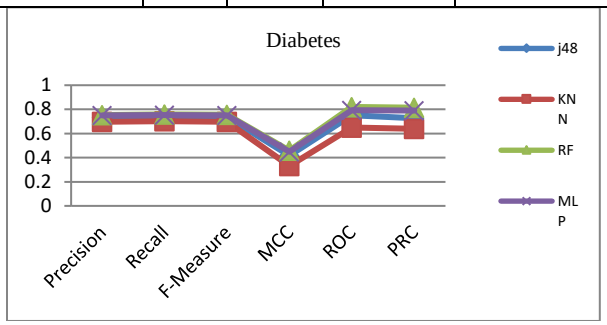

Fig-2 Diabetes data set on various classification

Figure 2 shows a diabetic dataset with four different classifiers (J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer Perceptron (MLP) applied and the results plotted. The MLP classifier definitely gives higher performance interns with increased precision, as seen in the figure.

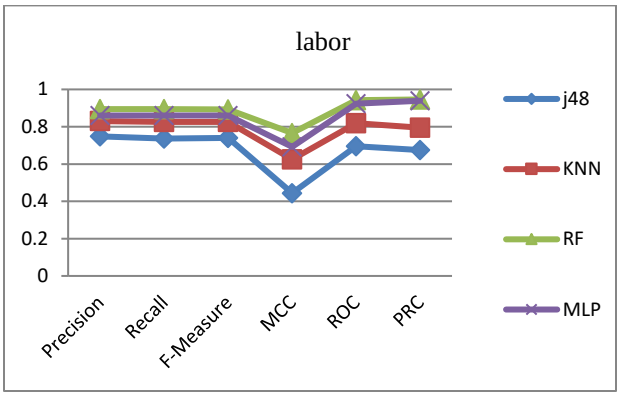

Fig-3 labor data set on various classifiers

In Figure 3, four different classifiers, including J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer Perceptron (MLP), are applied to a labor dataset, and the results are presented. The RF classifier gives superior performance interns with increased precision, as seen in the figure.

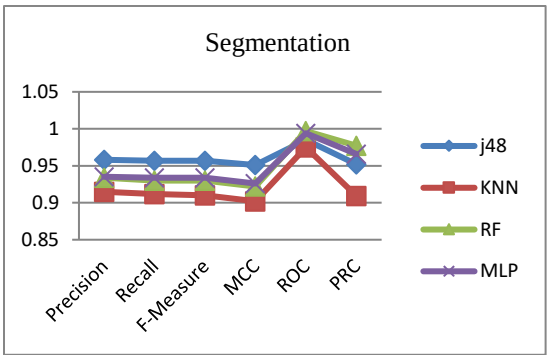

Fig-4 labor data set on various classifiers

Figure 4 shows a segmentation dataset in which four different classifiers are applied, including J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer Perceptron (MLP), and the results are presented. The J48 classifier, as seen in the figure, gives higher performance interns with increased precision.

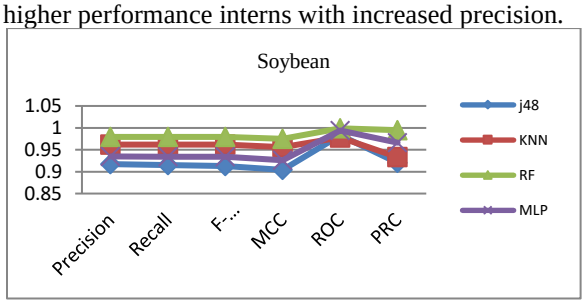

Figure 5 shows a diabetic dataset with four different classifiers (J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer Perceptron (MLP) applied and the results plotted. The J48 classifier clearly gives higher performance interns with increased precision, as seen on the graph.

5. CONCLUSION

We addressed a variety of topics related to multi-label text categorization in this work. The experiment is carried out on a variety of datasets employing classifications, feature transformations, and pyramid label space information consolidation. We reviewed the features of the outputs we acquired and attempted to elucidate them by doing a thorough analysis of the core algorithm, particularly the part of the hierarchical algorithm that we implemented from scratch. The perception we gained would also lead us to select specific algorithms and their settings that are relevant to the dataset elements we'd be managing. We have done a thorough analysis into how to choose different parameters of the hierarchical multi label prediction algorithm and which portions of the algorithm to use.

REFERENCE

[1]Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.
[2] "European open data portal". European open data portal. European Commission. Retrieved 2016-09-23.
[3] "Dataset definition – MELODA". www.meloda.org. Retrieved 2016-08-17.
[4] Atz, U (2014). "The tau of data: A new metric to assess the timeliness of data in catalogues" (PDF). CEDEM 2014 Proceedings. Retrieved 2016-08-01. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
[5] Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601
[6] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.
[7] Kleinberg E (1990). "Stochastic Discrimination" (PDF). Annals of Mathematics and Artificial Intelligence. 1 (1-4): 207–239. doi:10.1007/BF01531079
[8] Kleinberg E (1996). "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition". Annals of Statistics. 24 (6): 2319–2349. doi:10.1214/aos/1032181157. MR 1425956
[9] Kleinberg E (2000). "On the Algorithmic Implementation of Stochastic Discrimination" (PDF). IEEE Transactions on PAMI. 22 (5).
[10] Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324.
[11] Liaw A (16 October 2012). "Documentation for R package randomForest" (PDF). Retrieved 15 March 2013.
[12]U.S. trademark registration number 3185828, registered 2006/12/19.

[13] Amit Y, Geman D (1997). "Shape quantization and recognition with randomized trees" (PDF). Neural Computation.9 (7): 1545–1588. doi:10.1162/neco.1997.9.7.1545.