Multi Label Text Classification and the Effect of Label Pyramid Clustering

Stalin Jose J
Research Scholar, Bharathiar University, Department of Computer Science, Coimbatore, Tamil Nadu, India.

Dr. P. Suresh
Head of the Department, Department of Computer Science, Salem Sowdeswari College, Salem, Tamil Nadu, India.

Abstract – Now-a-days, text classification has gained more attention in the field of machine learning, information retrieval, and natural language processing (NLP). The text classification issues discover wide interest in different areas to perform tasks such as a. selection and classification of News, b. Classification of documents in digital libraries, social networks and websites, etc, c. E-mail categorization comprising filtering of spam. A deviation of this issue in which every document can be associated with some number of labels or classes that is known as a multi- label text classification issue. An expansion to such an issue in which the classes are inter-related through a definite hierarchy is known as a hierarchical text classification issue. In this research work, we use real-time dataset to perform multi label text classification experiments and investigate its performance. The algorithm is implemented for leveraging the hierarchy in the classes and examines the impact of various algorithmic methodologies and properties of the dataset over the classification performance.

Index Terms – NLP, Multilabel, Pyramid Clustering.

1. INTRODUCTION

Multi label classification (MLC) is taken as an issue which is suitable to a broad diversity of domains like bioinformatics and music categorization, which has gained more attention. Nevertheless, circumstances where individual instances are related with multiple classes remain challenging one. Most of the task classification algorithms consider the tasks of MLC as multiple binary classification tasks. Besides, the potential correlations between features and classes might not be considered by this technique. A better solution of MLC should be efficient as well as effective; but, a huge number of irrelevant and redundant attributes might increase the cost of communication and the time needed for learning and analyze multi-label classifiers that degrade the performance of classification. In data mining and machine learning techniques, feature selection is considered as an essential operation, which has been extensively utilized in classification frameworks for improving the performance. Choosing features before implementing classification techniques to unique datasets has numerous benefits, for example, filtering the information, decreasing the costs of computation, and enhancing the precision of classification [2, 3]. In this manner, we use a feature selection technique for enhancing the standard of MLC.

2. METHODOLOGY

1. Data set

A collection of data arranged in a format is known as a dataset. Generally, the data sets relate to the substance of an individual database table, or an individual statistical data matrix in which each section of the table demonstrates a specific variable, and every row relates to a provided number of the dataset in query. For every variable, the dataset performs listing process of values, for example, object height and its weight, for every dataset member. Every value is called as a datum. A dataset may contain information of single or multiple members, based on the number of rows. The word dataset may likewise be utilized more freely, for referring to the information in a gathering of firmly related tables, related to a specific test or occasion. A case of such kind is the datasets gathered by the agencies of space research executing experiments with different instruments on board space tests. Datasets which are extremely huge that conventional applications of data processing are deficient to manage them are called as big data [1]. In the discipline of open information, datasets are the unit for evaluating the data discharged in an open information repository. More than five lacs datasets is comprised by the European Open Data portal [2]. In such area different definitions have been presented [3] however at present there is not an authorized one. Some different problems such as real-time information sources,[4] non-social datasets, and so on, builds the trouble to achieve an agreement about it.

2. Data Preprocessing

In the process of data mining, data pre-processing is considered as an essential step. The expression “garbage in, garbage out” is especially appropriate for the projects of data mining and machine learning. Data collection strategies are often approximately controlled, results in out-of-extend measures (for example, Income: −200), impractical information integrations (for example, Sex: Male, Pregnant: Yes), missing measures, and so forth. Investigating information that has not been deliberately screened for these issues can create deceiving outcomes. In this manner, the characterization and quality of information is primarily before executing an analysis

Generally, data pre-processing is one of the most critical stage of a machine learning process, particularly in computational biology [2]. If there is more unnecessary and repetitive data exist or noisy and irregular information, at that point knowledge discovery throughout the training stage is more troublesome. Data creation and filtering stages can utilize considerable measure of data processing time. Data pre-processing incorporates filtering, Instance determination, standardization, feature selection, feature extraction and transformation, and so forth. The result of data preprocessing is the last training set. Kotsiantis et al. (2006) introduce an popular technique for every stage of data pre-processing [3].

3. Classifications

Classification strategies of data mining are equipped for handling a huge amount of information. It can be utilized to anticipate definite class labels and categorizes information as per the training set and the class labels which can be utilized for characterizing recently accessible information. This term can cover any setting in which a few choices or predictions are made based on recently accessible data. Classification technique is perceived strategy for frequently creating such choices in novel circumstances. If an assumption is made that issue is a worry with the development of a system which will be connected to a proceeding sequence of scenarios in which every novel scenario must be allocated to one of a pre-defined class sets based on recognized features of information. Formation of a classification method from a group of information for which the correct classes are familiar in prior is named as supervised learning or pattern recognition. Contexts where a classification operation is principal incorporate, for instance, allocating people to credit status based on money related and other individual data, and the underlying determination of a disease of patient to choose quick treatment while anticipating appropriate test outcomes. Probably the most basic issues emerging in business, industry, and science can be known as decision or classification issues. Three fundamental authentic strands of research can be distinguished: neural network, machine learning and statistical system. All classes have a few goals in general. They possess all endeavored to create methodology which would have the capacity to deal with.

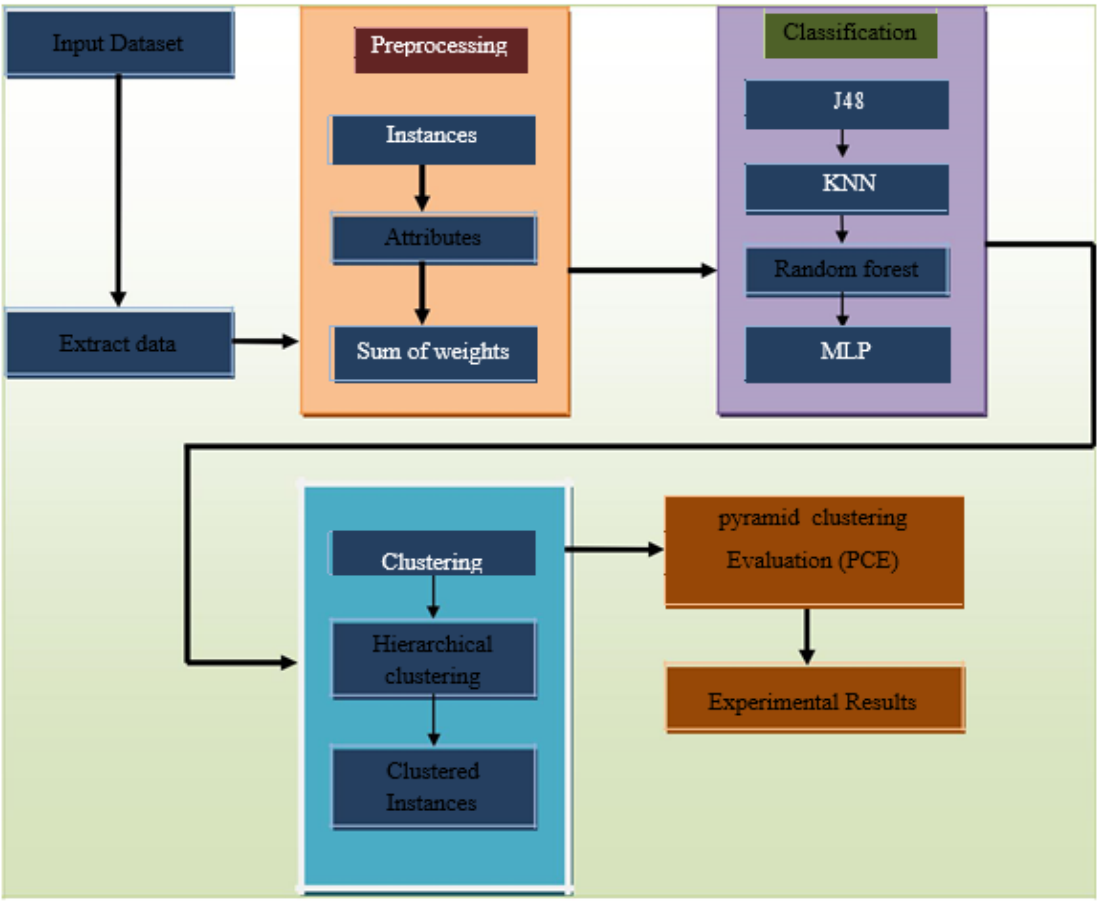


Fig-1 Architecture Diagram

4. Classification algorithms

Among several data mining techniques classification is one of them which is primarily utilized for examining given datasets and accepts every instance of them and allocates such instance to a specific class so that classification errors will be less. It is utilized to separate models which precisely characterize vital information classes inside the given datasets. Classification can be done in two phases. At initial phase the model is made by implementing classification algorithms over training datasets at that point in second phase the extracted model is analyzed against a predefined experimental dataset for measuring the model trained execution and precision. Thus, classification is the operation for allocating class labels from datasets whose class label is obscure.

4.1 J48 Algorithm

Using the classification, a model of classes is built from a group of records which comprises class labels. Decision Tree technique is for discovering the ways the attributes-vector performs for various occasions. Additionally on the basis of the training instance the classes for the recently created cases are being discovered [15]. This technique creates the principles for the prediction of the objective variable. With the assistance of tree classification technique the basic dissemination of the information is effectively comprehensible [5].

The extended version of ID3 is J48. The secondary features of J48 are representing missing measures, decision trees shortening, uninterrupted attribute estimation ranges, rules derivation, and so on. One of the data mining tools is WEKA in which J48 is a Java execution of the C4.5 algorithm and is open source as well. The WEKA tool renders various choices related with tree pruning. If there should arise an occurrence of possible over fitting pruning could be utilized as a tool to precise. In different techniques the classification is done continuously till each and every leaf is perfect, i.e. the classification of the information must be as flawless as possible. This technique it creates the principles from which specific personality of that information is produced. The goal is dynamically speculation of a decision tree until it obtains balance of adaptability and precision.

4.2 K-NN Classification

K-nearest neighbors (k-NN) is a type of classification algorithm which applied for pattern recognition and regression and it is a type of non-parametric strategy [1]. In the two cases, the input comprises of the k nearest training instances at the feature spaces. The result relies upon whether k-NN is utilized to perform classification or regression. In this k-NN classification, class membership is the expected output. Objects are categorized by a greater part vote of their neighbors, with the objects being allocated for the class most basic among their k nearest neighbors, where k is a positive whole number, ordinarily small. Let us take k = 1. Then the objects are just allocated to the class that possesses only one nearest neighbor. In the regression of k-NN, the result is the property estimation for the objects. This measure is the average of the estimations of their k nearest neighbors.

k-NN is one of the kinds of instance-based learning technique, which is otherwise known as lazy learning, where the operation is just approximated locally and overall computations are conceded until classification. Among all the machine learning algorithms, the k-NN algorithm is the simplest one. For both classification and regression, a valuable method can be for allocating weight to the commitments of the neighbors, such that the closer neighbors can contribute much more to the average than the far ones. For instance, a typical weighting technique comprises in providing every neighbor a weight of 1/D, where D is the neighbor distance [2]. From a group of objects the neighbors are selected for which the k-NN object property measure (regression) or the k-NN class (classification) is well known. It can be considered as the training sets for k-NN, however no direct training phase is needed. An idiosyncrasy of the k-NN technique is that it is much sensitive to the local patterns of the information. The technique is not to be mistaken for k-means, other famous machine learning method.

4.3 Random forests Algorithm

Random forest algorithm is a group learning technique to perform classification, regression, and some other operations, and it is operated by building a huge number of decision trees during training times and results the class which is the role of the mean prediction (for regression) or classes (for classification) of the separate trees [1][2]. This random forest algorithm is otherwise known as random decision forests. In general, decision trees possess the manner of over-fitting for its training set which can be corrected by the random decision forests [3]. Tin Kam Ho[1] has developed the primary algorithm for random decision forests by employing the random subspace technique [2] that is in Ho's definition, is an approach for establishing the "stochastic segregation" technique to deal with classification presented by Eugene Kleinberg.[4][5][6]. Leo Breiman [7] and Adele Cutler,[8] have extended this random decision forest algorithm and they use “Random Forests” as their logo [9]. This extension consolidates the concept of “bagging” (proposed by Breiman) and random feature selection, initially presented by Ho [1] and afterwards freely by Amit and Geman [10] to develop an accumulation of decision trees along with controlled difference.

4.4 Multilayer perceptron

One of the types of feed-forward artificial neural network is multilayer perceptron (MLP) which comprises of not less than three layers of nodes. Aside from the input nodes, every node acts as a neuron which utilizes a nonlinear activation operation.

MLP uses a supervised learning methodology known as back propagation to perform training [1] [2]. The non-linear activation and multiple layers of MLP distinguish it from linear perceptrons. It can recognize information which is not directly separable [3]. Sometimes, the multilayer perceptrons are informally called as "vanilla" neural networks, particularly when they possess an individual concealed layer [4].

4.5 Data mining Tools

Alike with SAS Enterprise Miner, WEKA is known as a data mining suite, however it is an open source code which is accessible completely free of cost. If anyone needs to modify the source code of the algorithm, then WEKA is a better tool to utilize. The re-implementations of several traditional data mining algorithms can also be done in WEKA, comprising C4.5 that is known as J48. WEKA has a major benefit over SAS Enterprise Miner is which the Enterprise Miner is utilized just by means of a graphical user interface (GUI) and therefore it is difficult to robotize tests that is frequently important for researches when you need to execute possibly many variations of an analysis. On the other hand, WEKA has a different operation mode which creates experimentation simple.

5. Pyramid Clustering

In several text categorization applications, a tree-structured hierarchy is mostly followed to organize the labels. A case is related with a specific label just in the event that it is additionally connected with the parent of the label in such hierarchy. Besides, most of the traditional multi-label categorization techniques never consider the structure of the labels into account. Rather, the labels are basically treated independently, prompting the necessity for training a huge number of classifiers which takes one for every label. Moreover, as few leaf labels might possess less number of positive illustrations, the trained information turn out to be profoundly skewed, making issues in numerous classifiers. Also, the irregular labeling amongst parent and child creates trouble in interpretation. At last, the performance of prediction is weakened as structural conditions among the labels are not used in the process of learning. Some current methodologies do the accompanying: if a parent of a label is predicted then only the positive prediction can perform for that label; build the training illustrations for every node from tests of the parent node; utilize the structured predictors with large-margin, or by adjusting decision trees.

3. EXPERIMENTAL RESULTS

In this analysis, we have taken four datasets such as diabetes, labor, Segmentation and Soybean datasets. Four different classifiers are taken for classification such as J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer perceptron (MLP). Some of the parameters like Precision, Recall, F-Measure, MCC, ROC and PRC are taken and their obtained values are tabulated.

Datasets	Precision	Recall	F-Measure	MCC	ROC	PRC	Classification
Diabetes	0.735	00.738	0.736	0.417	0.751	0.727	J48
Diabetes	0.696	0.702	0.698	0.331	0.650	0.640	K-NN
Diabetes	0.754	0.758	0.755	0.458	0.820	0.814	Random Forest
Diabetes	0.750	0.754	0.751	0.449	0.793	0.786	MLP

Tabel-1 Diabetes data set performance in various classifications result

Relation	Instances	Attributes	Classifier	Time (s)	incorrectly clustered instances	
pima_diabetes	768	9	J48	4.47	267.0	34.7656 %
pima_diabetes	768	9	K-NN	4.5	267.0	35.0%
pima_diabetes	768	9	RF	4.5	267.0	35.0%
pima_diabetes	768	9	MLP	4.5	267.0	35.0%

Tabel-1(a) Diabetes data set and its Pyramid cluster evaluation

Datasets	Precision	Recall	F-Measure	MCC	ROC	PRC	Classification
labor	0.748	0.737	0.740	0.444	0.695	0.675	J48

labor	0.830	0.825	0.826	0.625	0.818	0.794	K-NN
labor	0.894	0.895	0.893	0.766	0.943	0.946	Random Forest
labor	0.860	0.860	0.860	0.692	0.923	0.939	MLP

Tabel-2 labor data set performance in various classifications result

Relation	Instances	Attributes	Classifier	Time (s)	incorrectly clustered instances
labor	57	17	J48	0.1	20.0 35.0877 %
labor	57	17	K-NN	0.1	20.0 35.0877 %
labor	57	17	RF	0.1	20.0 35.0877 %
labor	57	17	MLP	0.1	20.0 35.0877 %

Tabel-2(a) Diabetes data set and its Pyramid cluster evaluation

Datasets	Precision	Recall	F-Measure	MCC	ROC	PRC	Classification
Segment	0.958	0.957	0.957	0.951	0.985	0.952	J48
Segment	0.915	0.912	0.910	0.902	0.975	0.909	K-NN
Segment	0.934	0.930	0.930	0.922	0.997	0.977	Random Forest
Segment	0.935	0.934	0.934	0.926	0.994	0.966	MLP

Tabel-3 Segement data set performance in various classifications result

Relation	Instances	Attributes	Classifier	Time (s)	incorrectly clustered instances
Segment	1500	20	J48	7.2	1263.0 84.2 %
Segment	1500	20	K-NN	1.36	1263.0 84.2 %
Segment	1500	20	RF	1.36	1263.0 84.2 %
Segment	1500	20	MLP	7.2	1263.0 84.2 %

Tabel-3(a) Segment data set and its Pyramid cluster evaluation

Datasets	Precision	Recall	F-Measure	MCC	ROC	PRC	Classification
Soybean	0.917	0.915	0.913	0.904	0.983	0.920	J48
Soybean	0.962	0.962	0.962	0.956	0.978	0.933	K-NN

Soybean	0.979	0.979	0.979	0.975	0.999	0.995	Random Forest
Soybean	0.935	0.934	0.934	0.926	0.994	0.966	MLP

Tabel-4 Soybean data set performance in various classifications result

Relation	Instances	Attributes	Classifier	Time (s)	incorrectly clustered instances	
Soybean	683	36	J48	1.38	590.0	86.3836 %
Soybean	683	36	K-NN	1.38	590.0	86.3836 %
Soybean	683	36	RF	1.38	590.0	86.3836 %
Soybean	683	36	MLP	1.38	590.0	86.3836 %

Tabel-4(a) Soybean data set performance and its Pyramid cluster evaluation

Comparative parameters

A. Accuracy

The number of correct assessment to the total assessments ratio can provide the measure of accuracy. Initially, from the entire dataset the extraction of appropriate images is done which is then contrasted with the entire dataset by applying the given expression where data quality and errors are the important factors which are estimated in terms of percentage (%).

Accuracy=(TN+TP)/(TN+TP+FN+FP)

Where, TN-True Negative, TP-True Positive, FP-False positive and FN-False Negative.

B. Sensitivity

In order to estimate sensitivity, the true positives and false negatives are extracted from the dataset which are then added. The count of true positive to the added estimation of true positive and false negative ratio provides the sensitivity. The precisely perceived information declare the quantity of positive measures. It is estimated by implementing the given expression and it is estimated in terms of percentage (%).

Sensitivity=TP/(TP+FN) (2)

C. Specificity

Specificity is utilized for predicting the impact of modifications at the result because of its progressions at the given input datasets. It is estimated from the correctly perceived negative measures and the specificity is estimated by percentage (%).It is characterized as the proportion of the quantity of negative assessments to the summation of the quantity of true negative and false positive assessments. The accompanying expression demonstrates the specificity.

Specificity= TN/(TN+FP) (3)

Evaluation parameter	J48	K N N	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

Table 5: Overall Comparison of better classification using and its Pyramid cluster evaluation

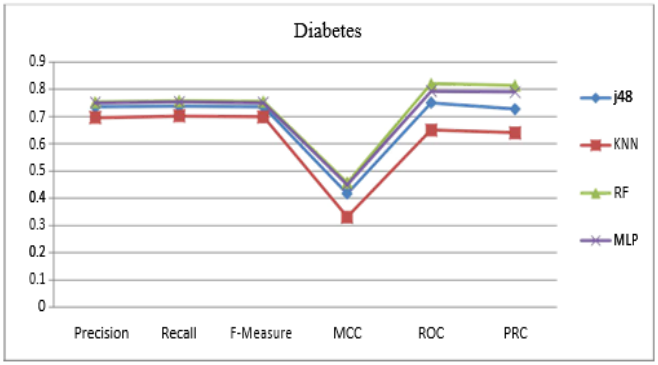


Fig-2 Diabetes data set on various classification

In Figure 2, diabetes dataset is taken on which four different classifiers like J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer perceptron (MLP) are applied and the obtained results are plotted. From the figure it is clearly known that the MLP classifier provides better performance interns of increased precision.

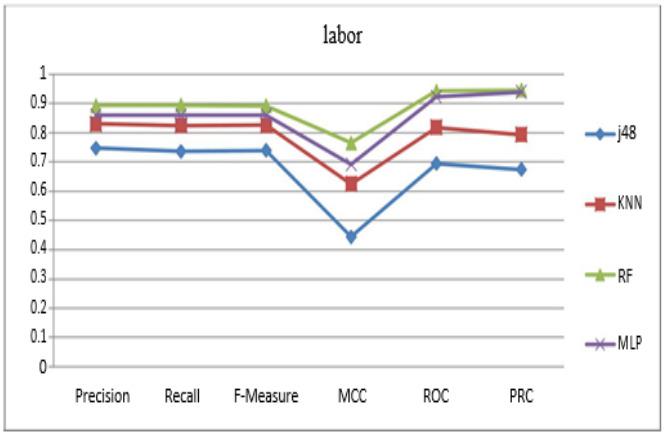


Fig-3 labor data set on various classifiers

In Figure 3, labor dataset is taken on which four different classifiers like J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer perceptron (MLP) are applied and the obtained results are plotted. From the figure it is obtained that the RF classifier provides better performance interns of increased precision.

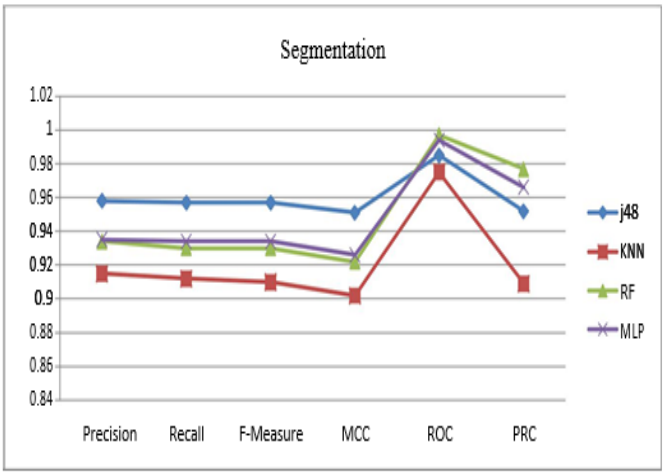
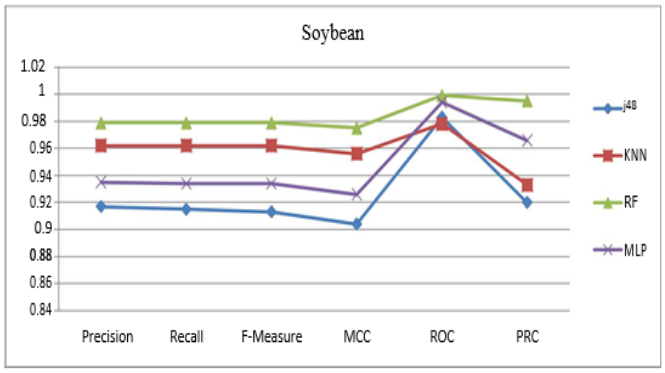


Fig-4 labor data set on various classifiers

In Figure 4, segmentation dataset is taken on which four different classifiers like J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer perceptron (MLP) are applied and the obtained results are plotted. From the figure it is observed that the J48 classifier provides better performance interns of increased precision.



In Figure 5, diabetes dataset is taken on which four different classifiers like J-48, K-Nearest Neighbor (K-NN), Random Forest (RF), and Multilayer perceptron (MLP) are applied and the obtained results are plotted. From the figure it is clear that the J48 classifier provides better performance interns of increased precision.

4. CONCLUSION

In this paper we handled different issues associated with multi label text classification. The experiment is done on different datasets using classifications, feature transformations, and consolidation of pyramid label space information. We discussed the characteristics in the outputs we have obtained and attempted to clarify them through a complete investigation of the fundamental algorithm, particularly in the part of hierarchical algorithm that we executed from scratch. The perception we obtained would additionally direct us towards selecting specific algorithms with their parameters which take to certain dataset aspects we would manage. In particular, we have given a comprehensive investigation of how to select different parameters of the hierarchical multi label prediction and which segments of such algorithm to utilize.

REFERENCES

[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.