ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Meta-Top-Down Method for Large-Scale Hierarchical Classification

Kolawale Abhijeet, Surase Rohit, Sule Krishna, Lambhate Dhiraj

B.E. Student, Dept of Information Technology, DYPIET, Pimpri, India

ABSTRACT-Recently large-scale hierarchical classification tasks typically have tens of thousands of classes on which the most widely used approach to multi-class classification one-versus-rest—becomes intractable due to computational complexity. The top-down methods are usually adopted instead, but they are less accurate because of the so-called error-propagation problem in their classifying phase. To address this problem, this paper proposes a meta-top-down method that employs metaclassification to enhance the normal top-down classifying procedure. The proposed method is first analyzed theoretically on complexity and accuracy, and then applied to five real-world large-scale data sets.

KEYWORDS:-Meta-Data,Data-mining,Top-Down Classification,Hierarchical Classification.

I. INTRODUCTION

Classification classifying samples into multiple predefined classes is a fundamental task in both machine learning and data mining domains . For example, each document in the Reuters-215781 corpus is assigned one or more labels from 120 predefined classes such as business, sports, and military. The ensemble method of oneversusrest is the most widely adopted solution for multiclass classification. First a binary-class classifier fi i 1; . . . ; n, named base classifier, is trained for each class ci to predict whether an input sample x belongs to this class; then thresholding strategies are employed to decide the predicted labels according to the confidence scores of the base classifiers. Two commonly used thresholding strategies are score-cut (S-cut) that accepts the classes whose scores are larger than a predefined threshold, and rank-cut (R-cut) that accepts the classes whose scores are among the top-r (r is a predefined integer). In recent years, two types of classification strategies have been developed for multiclass classification problems. One is to reduce the computational complexity on the tasks that have large numbers of predefined classes, such as tens of thousands; usually hierarchies are used to organize the classes, so these tasks are called large-scale hierarchical classification. The examples include categorizing patent documents into the taxonomy of the International Patent Classification and categorizing web pages into the directories of the Open Directory Project or Yahoo!

II. RELATED WORK

Experimental results in this section show that the proposed methods outperform the state-of-the-art algorithms almost in all cases on both real and synthetic data sets. It will apply Meta TD to more large scale hierarchical classification tasks.

It will apply Meta TD to more large scale hierarchical classification tasks. We expect that developing a flexible method of selecting label candidates for Meta TD will be a promising solution.

A Survey on..

Hierarchical levelled Classification alludes to allocating of one or more suitable classes from a progressive class space to a report. While past work in progressive classification concentrated on virtual classification trees where reports are doled out just to the leaf classifications, Author propose a top-down level-based classification technique that can group records to both leaf and inside classes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

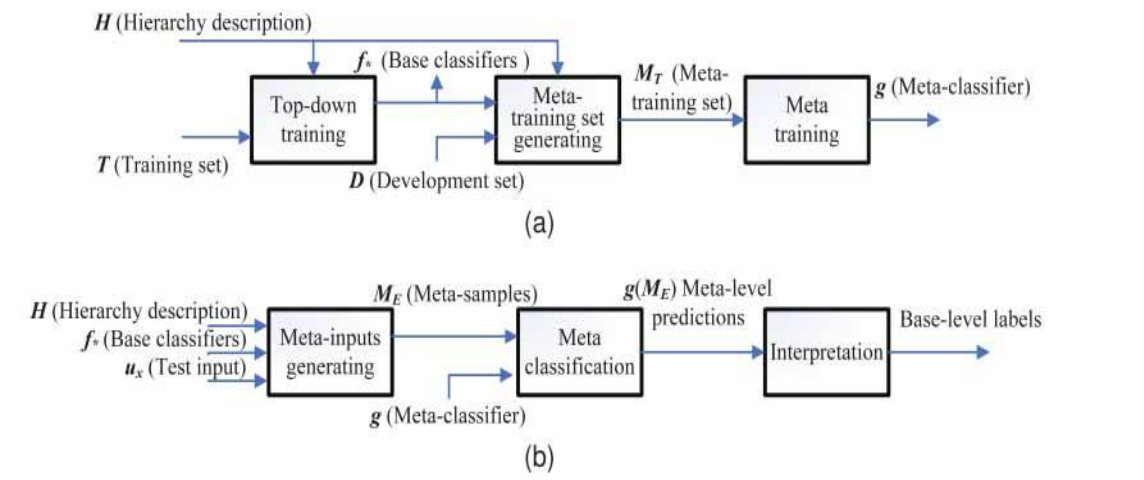
The proliferation of topic hierarchies for text documents has resulted in a need for tools that automatically classify new documents within such hierarchies.

Large-scale order scientific classifications ordinarily have a huge number of classes, profound progressive systems, and skewed classification dispersion over archives. Be that as it may, it is still an open inquiry whether the cutting edge advancements in mechanized content order can scale to (and perform well on) such huge scientific categorizations.

III. PROPOSED SYSTEM

We will apply Meta TD to more large scale hierarchical classification tasks, particularly then on mandatory leaf classification tasks such as Yahoo! categories. We expect that developing a flexible method of selecting label candidates for Meta TD will be a promising solution.
It will apply Meta TD to more large scale hierarchical classification tasks, particularly then on mandatory leaf classification tasks such as Yahoo! categories. We expect that developing a flexible method of selecting label candidates for Meta TD will be a promising solution

IV. SIMULATION



V. ARCHITECTURE/SIMULATION THEORY

After careful analysis the system has been identified to have the following modules:

ALL PASSIVE MODULES

1 Large-Scale Hierarchical Classification

2 Meta classification

3 Top-Down Method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

1.Large-scale hierarchical classification(LSHTC) -

The LSHTC Challenge is a hierarchical text classification competition, using very large datasets. This year's challenge focuses on interesting learning problems like multi-task and refinement learning.

Hierarchies are becoming ever more popular for the organization of text documents, particularly on the Web. Web directories and Wikipedia are two examples of such hierarchies. Along with their widespread use, comes the need for automated classification of new documents to the categories in the hierarchy. As the size of the hierarchy grows and the number of documents to be classified increases, a number of interesting machine learning problems arise. In particular, it is one of the rare situations where data sparsity remains an issue, despite the vastness of available data: as more documents become available, more classes are also added to the hierarchy, and there is a very high imbalance between the classes at different levels of the hierarchy. Additionally, the statistical dependence of the classes poses challenges and opportunities for new learning methods.

2.Metaclassification-

Meta learning is a subfield of Machine learning where automatic learning algorithms are applied on meta-data about machine learning experiments. Although different researchers hold different views as to what the term exactly means (see below), the main goal is to use such meta-data to understand how automatic learning can become flexible in solving different kinds of learning problems, hence to improve the performance of existing learning algorithms.

Flexibility is very important because each learning algorithm is based on a set of assumptions about the data, its inductive bias. This means that it will only learn well if the bias matches the data in the learning problem. A learning algorithm may perform very well on one learning problem, but very badly on the next. From a non-expert point of view, this poses strong restrictions on the use of machine learning or data mining techniques, since the relationship between the learning problem (often some kind of database) and the effectiveness of different learning algorithms is not yet understood.

2(A) Meta-Top-Down Method Algorithm

The proposed meta-top-down method employs metaclassification to reclassify samples based on the output of the normal top-down methods. MetaTD takes the confidence scores of the base classifiers along a root-to-leaf path as the metalevel input, and takes whether the leaf node is a correct label as a metalevel target. This metaclassification task is formulated as follows:

plexity on large-scale hierarchical classification tasks as it MetaTD is based on the above settings, and its workflow presented in Fig The training phase consists of three steps as follows:

1. Train base classifiers f on the training set T, which is the same as the normal top-down methods.
2. Construct a metatraining set with the base classifiers and the development set D through the pruning method
3. Train a metaclassifier g on M The whole training phase requires the base-level training set T, the development set D, and the description of the

hierarchy H. It produces a base classifier f per child node and a met classifier g.

3.Top-Down Method -

Top-down and bottom-up are both strategies of information processing and knowledge ordering, used in a variety of fields including software, humanistic and scientific theories (see systemic), and management and organization. In practice, they can be seen as a style of thinking and teaching. A top-down approach (also known as stepwise design and in some cases used as a synonym of decomposition) is essentially the breaking down of a system to gain insight into its compositional sub-systems. In a top-down approach an overview of the system is formulated, specifying but not detailing any first-level subsystems. Each subsystem is then refined in yet greater detail, sometimes in many additional subsystem levels, until the entire specification is reduced to base elements. A top-down model is often specified with the assistance of "black boxes", these make it easier to manipulate. However, black boxes may fail to elucidate elementary mechanisms or be detailed enough to realistically validate the model. Top down approach starts with the big picture. It breaks down from there into smaller segments.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

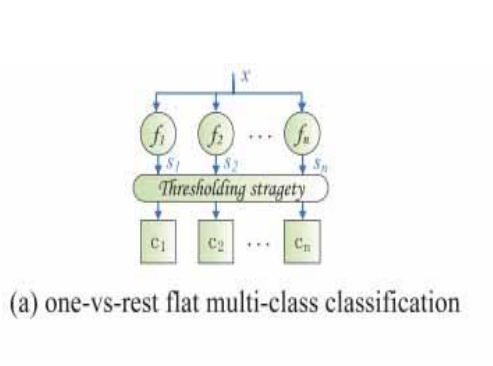
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

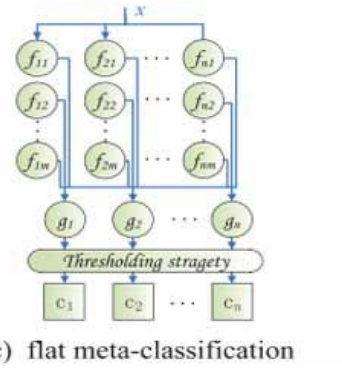
Vol. 3, Issue 12, December 2015

Application -
1)Authencation model
2)E-mail verification
3)Data upload module
4)Data storage and Download module.

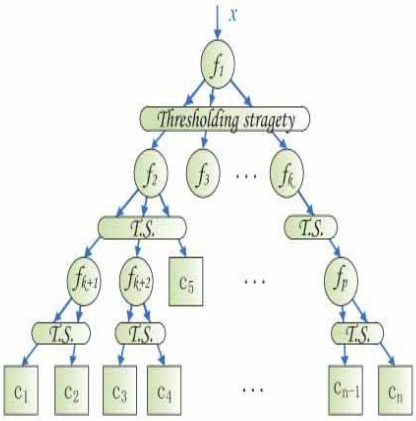
VI. FIGURE



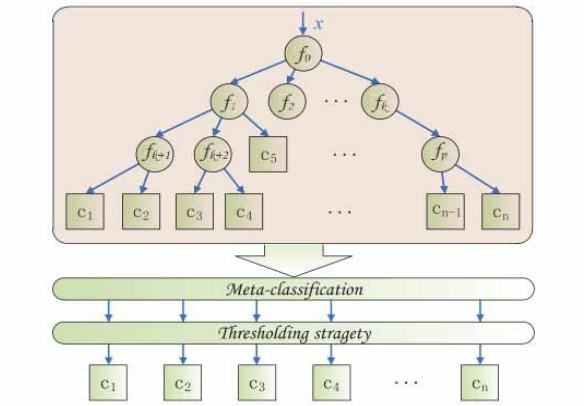
(a) one-vs-rest flat multi-class classification



(c) flat meta-classification



(b) top-down hierarchical classification



(d) meta-top-down hierarchical classification

VII. CONCLUSION

This paper proposes a meta-top-down method (MetaTD) to relieve the error-propagation problem of the normal topdown methods while retaining their capability for largescale hierarchical classification. In the accuracy analysis, MetaTD is proved to subsume the normal top-down methods, ensuring that it can provide higher classification accuracy. The experimental results show that, on the aspect of classification accuracy, MetaTD outperforms ScutTD on multilabeled data sets by 36.2-57.3 percent, and outperforms RcutTD on single-labeled data sets by 5.9 percent. The comparison with the results from LSHTC1-3 challenges indicates that MetaTD is among the state-of-the-art methods. On the aspect of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

computational complexity, MetaTD raises the training time costs of ScutTD and RcutTD by 4.37.5 percent and 70.0-82.6 percent, respectively. Such performance is competitive among the related work.

REFERENCES

[1] Hierarchical Text Classication and Evaluation, Aixin Sun and Ee-Peng LimCenter for Advanced Information Systems Nanyang Technological University.
[2] Hierarchically Classifying Documents Using Very Few Words, Proc. Intl Conf. Machine Learning (ICML 97), pp. 170-178, 1997.
[3] Support Vector Machines Classification with a Very Large-Scale Taxonomy.
[4] Producing a Test Collection for Patent Machine Translation in the Seventh NTCIR Workshop.
[5] Machine Learning in Automated Text Categorization, ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
[6] Introduction to the Special Issue on Patent Processing, Information Processing and Management, vol. 43, no. 5, pp. 1149-1153, 2007.
[7] Incorporating Prior Knowledge into Task Decomposition for Large-Scale Patent Classification, Proc. Sixth Intl Symp. Neural Networks: Advances in Neural Networks (ISNN 09), pp. 784-793, 2009.

BIOGRAPHY

Kolawale Abhijeet, **Surase Rohit**, **Sule Krishna**, **Lambhate Dhiraj are** pursuing the B.E. degree in Information Technology from the Dr.D Y Patil College,Pimpri, pune, India

.