International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

# Survey of Large-scale Hierarchical Classification

Ms. Ankita A. Burungale[1], Prof. Dinesh A. Zende[2]

[1]Student, M. E. Computer Engineering, VPCOE, Baramati, Savitribai Phule Pune University, India

ankitaburungale@gmail.com

[2]Assistant Professor, Information Technology Department, VPCOE, Baramati, Savitribai Phule Pune University , India

dineshzende@gmail.com

**Abstract**— Large-scale classification taxonomies have thousands of classes, deep hierarchies and skewed category distribution over documents. Hierarchical classification can speed up the classification process because problem is sub divided into smaller sub problems, and each of which can be efficiently and effectively managed. Most commonly used method for multiclass classification is one versus rest method. It is inflexible due to computational complexity. The top down method is usually accepted, but it has an error propagation problem. The metaclassification method solves error propagation problem. In this paper, several challenges for hierarchical document classification such as scalability, complexity, and misclassification are reviewed. The questions concerning about the learning and the classification processes are reviewed.

**Keywords**— Large-scale hierarchical classification, Top-down method, Metalearning, Metaclassification, Text classification, Ensemble learning, Feature selection.

## INTRODUCTION

Multiclass classification is one of the most important tasks in data mining. It is applicable in various areas like text categorization, patent classification, protein function prediction, music genre classification and so on. Information on Internet is growing day by day. Thus, it is very difficult to search required information and utilize this large information. The solution to this problem is to classify the information into topic, where this topic is arranged hierarchically.

There are three techniques used for multiclassification are (1) One versus rest method, (2) Top down method, (3) Metaclassification method.

In a one versus rest method, a single classifier is trained per class to distinguish that class from all other classes. It does not consider structural relationships among them. The decision about assigning document to a category based on score of only one classifier. This method has very high time complexity.

A top down method builds classifiers for every level of the category tree, where every classifier acts as a flat classifier at that level [1-4]. At the root level in the hierarchy, first a document is classified into one or more sub-categories. The classification process can be repeated for the document in each of the subcategories until it reaches to leaf categories or cannot be further classified into any categories.

A metaclassification method solves error propagation problem of the top down method [6-8]. To solve this problem, it uses the predication of all base classifier for training of the metaclassifier, and then metaclassifier reclassifies the sample based on score of all base classifiers.

## LITERATURE SURVEY

In this section, top-down methods and metaclassification method are reviewed.

## 1.TOP DOWN METHOD

A] Dumais and Chen [1] work on hierarchies of classes for classifying web content. The proposed paper uses support vector machine for learning and classification. This technique is applicable for huge dynamic collections. The hierarchical structure is used for two purposes. First, second-level category models are trained using different contrast sets (either by categories in the flat non-hierarchical structure or by the same top-level category in the hierarchical structure). Then scores from the top-level and second-level models are combined using different combination rules. These techniques combine the probabilities of first and second level by using Boolean and Multiplicative rule. In Boolean rule, first a threshold is set at the highest level and only match next level categories that pass this test, i.e., calculate by *P (L1) && P (L2)*. These both constraints must be satisfied in order to classify a test instance. This method is very efficient, meanwhile huge numbers of next level categories do not need to be tested. The Multiplicative rule is calculates by *P (L1)* × *P (L2)*. This rule allows to match next level category even their scores are lower than threshold.

**Advantages**

1. Many search results is confused at top level. To tackle this issue, these methods concentrate only on the top levels of the hierarchy.
2. The performance is improved for automatically categorised search result by using an interface that strongly couples search results and category structure.
3. The negative sample is smaller at top level, because it includes the item only from same top level. As a result of that, training is faster at top level.
4. This classification method organizes test sample into existing hierarchical structure.
5. This classifiers are trained offline by using human label training set of document and web categories. Thus, run time classification is very efficient and human label categories are easy to understand.
6. This technique is theoretically modest and scalable for hierarchical training and classification. This method is applicable for large text categorization.
7. This methods uses taxonomy structure, because of that the classification efficiency and accuracy is improved.

**Disadvantages**

1. This is a tree-based method. Therefore, there is problem of multiple taxonomies, evolving taxonomies or unnecessary intermediate categories on the path from the root to deeper categories.
2. It is applicable only for three levels of hierarchy.
3. There is problem of an information organization, because huge collection of heterogeneous web content is considered for training.

B] Sun et al [2] provide solution to the blocking problem of the top down method by using restricted voting, threshold reduction, and extended multiplicative methods. The threshold reduction method is based on the principle of lower threshold for sub-tree classifier. Hence, more documents can be passed to the classifiers at lower-levels. All classifier at same level use same threshold to minimize the number of threshold combinations. Even though the threshold reduction is able to pass more documents to the classifiers at the lower levels, there is still possibility that documents mistakenly rejected by the higher-level sub-tree classifiers.

The restricted voting method solves the error propagation problem, by giving a chance to low-level classifiers to access documents before the sub-tree classifiers of their parent nodes forbid them. This method generates secondary channels, so that the local or sub-tree classifier of a node can able to receive documents from the sub-tree classifier of its grandparent node. Here the hierarchy is modified for passing sample down to grandchild nodes, so it results in increased computational complexity.

The extended multiplicative method as its name suggests, is resulting from the multiplicative method proposed by Dumais and Chen [1]. While the multiplicative method works only for three level of hierarchy. The extended multiplicative method handles the category trees with more than three levels. It passes the sample down to next level if products of two-classifier probability are accepted by the threshold strategy.

**Advantages**

1. It solves the error propagation problem.
2. All sub-tree classifiers at the same level use same threshold value. Thus, this technique reduces the number of threshold combination.
3. The restricted voting method gives chance to low-level classifiers to access documents, before the sub-tree classifiers of their grandparent nodes reject them.
4. The restricted voting method works tremendously well for classes with a minor number of positive test documents (and these classes have a smaller number of training documents as well).
5. The restricted voting method provides good classification performance.

**Disadvantages**

1. The use of threshold reduction method can result in the blocking problem.

2. If the thresholds of all ancestors sub-tree classifiers are zero, then threshold reduction method will degenerate into a flat classification method.
3. In the threshold reduction method, the challenge is to determine the thresholds for sub-tree classifiers.
4. In the threshold reduction method, documents erroneously rejected by the higher-level sub-tree classifiers.
5. In the restricted voting method, second level channel is added so complexity is increased.

C] Bennett and Nguyen proposed a technique called expert refinements [3]. The tree-based approach has two main problems. First, documents are wrongly rejected at higher level (false negative), and second, documents are wrongly come at lower level (false positive). To pass correct document to lower level node, stronger indicator is required. The Refined expert method uses the predictions from the lower nodes and cousins as meta attribute for the higher levels. To do this, refined expert method trains classifiers at the leaf nodes using cross-validation on the training data, and then uses the predictions from the training data, that are collected during cross-validation as meta features to next higher level. This method uses bottom-up training followed by top down training. The bottom up training solves false negative problem, and top down training halts the transmission of false positives document to an incorrect branch. The predication from cousins are included as meta attribute, because a high probability at a cousin node denotes document are belonging to a sibling, so it cannot pass down to next level.

**Advantages**

1. Accuracy is improved as it uses complete set of feature for training.
2. It utilizes additional features that are precise to a particular domain, thus classification performance is improved.
3. It enhances the top down method.

**Disadvantages**

1. It has deficiency in classification accuracy, i.e. its accuracy is lower than the one-versus-rest method. It is caused by the error propagation in deep levels of the hierarchy.
2. The training of the root classifier is performed on training set, which is very time consuming.
3. It requires complex decision at the top level of the hierarchy.
4. It does not use structure of hierarchy for the task of feature extraction.

D] H. Malik combines the benefits of the flat and hierarchical schemes [4]. This technique flattens the original hierarchy to k$^{th}$ level, earlier to training hierarchical classifiers (where k is a user-defined parameter). Flattening replaces some categories by their descendant nodes. Flatten hierarchy is similar to flat structure having less levels, thus an error propagation problem is solved. The flattening is pre-processing step. The novel lazy classification approach is used for selecting the most promising classes for test sample. It uses primary and secondary classifiers. It does not depend on confusion graph to find the classes used for secondary classifier. Instead of training secondary and primary classifier in earlier fashion, it defers the training of secondary classifier to the classification phase. In the training phase, this method trains a top-down hierarchical classifier in normal way. To classify document, this method first identify the most positive classes using hierarchical classifier, and then trains a multi-class classifier based on only the selected classes. Then new classifier is used to make the final estimate.

**Advantages**

1. The hierarchical structure does not always provide better classification quality than the flat structure, because of computing errors at each level. To tackle this issue, this method flattens the hierarchy up to kth level.
2. It uses kth level hierarchy. Hierarchical classifiers are required to work with considerably fewer classes. As a result, the hierarchical scheme uses substantially fewer computational resources.
3. This method combines the benefits of flat and hierarchical schemes.
4. This method organizes set of categories hierarchically, so reasonable classification time is required.
5. It does not use few levels. Hence, it reduces the risk of making an error in the top-down classification process.

**Disadvantages**

1. Its complexity is higher because this technique flattens the hierarchy.
2. It is difficult to decide which levels in the hierarchy should be flattened.
3. Excessive flattening increases the time for training and prediction.

E] Koller and Sahami proposed an approach for classification [5], that uses structured hierarchy of topics, instead of ignoring categorical structure and constructing a single large classifier for the entire task. This method breaks the classification problem into manageable sub problem by using structure of hierarchy. The basic perception supporting this approach is the subjects that are close to each other in the hierarchy; logically have a lot more in mutual with each other than subjects that are far at a distance.

Each sub-problem is simpler than the original problem, as the classifier at a node in the hierarchy need to differentiate between a small numbers of categories. Therefore, it is possible to make decision based only on a small set of features. This feature set avoids the overfitting problem. For feature selection, probabilistic framework is used.

**Advantages**

1. It reduces the computational complexity, because it uses reduced feature set.
2. It arranges predefine category into hierarchy.
3. The vocabulary of category is built for each node, so it permits to use probabilistic model.
4. The accuracy is improved as feature extraction removes irrelevant features.
5. It provides few advantages when focus is made on single classifier.

**Disadvantages**

1. It assigns document to only leaf node.
2. This technique works effectively for small features.
3. It uses greedy method for selecting branches. Thus, this method is error prone.
4. It has blocking problem.

## 2. METACLASSIFICATION METHOD

A] Todorovski and Dzeroski developed a meta decision tree [6]. This method used to combine the predication of all base classifier prompted from different learning algorithm. This method uses the probability distributions of classes predicted by the base-level classifiers. It uses the predicted class values for identification of the set of meta-level attributes. The meta decision tree (MDT) is used to decide which base classifier should be used to classify a test sample. The arrangement of a meta decision tree is similar to the arrangement of an ordinary decision tree. A decision (inner) node states a test to be performed on a single attribute value. Each test result has its own branch leads to the suitable sub-tree. The leaf node of a meta decision tree specify which classifier should be used for classification, instead of guessing the class value directly. The Meta decision tree is domain independent, because it uses meta-level attributes as set of class distribution properties and does not use class attribute at internal nodes.

**Advantages**

1. It uses only ordinary attribute at internal nodes.
2. The spilt goodness for internal nodes is differently calculated.
3. MDT gives better performance than ordinary decision tree.
4. It reduces the size of tree, so it improves comprehensibility of meta decision trees.
5. MDTs are more accurate than ordinary decision tree because of expressive influence of meta decision tree leaves.
6. MDTs are generally too small so it is easy to understand.
7. It is useful when data include instances from heterogeneous subdomain.

**Disadvantages**

1. It is having high complexity.
2. It cannot applicable to large-scale data set.

B] Kong, Zhao and Luan proposed an adaptive ensemble learning strategy using an assistant classifier [7]. The proposed scheme divides imbalanced and large binary classification problem into independent balanced binary sub-problems. In the training phase, a large imbalanced training data set is segmented into many balanced training subsets and processed in parallel. Then base classifiers are trained on all these subsets separately. For every known sample in the original training set, the outputs of these base classifiers are prepared into vectors and an assistant classifier will learn from these vectors to discover an effective ensemble way to output a class label for each specified sample. For the classification phase, an unknown sample is given to all the base classifiers, and the outputs of all the base classifiers are integrated to make a solution to the original problem according to the assistant classifier.

**Advantages**

1. An imbalanced complex classification problem is divided into several smaller independent binary classification problems. Hence, it will improve efficiency and performance of patent classification.
2. This technique takes advantages of outputs of all base classifier. So an error propagation problem is solved.

3. This method uses an assistant classifier with base classifier. Therefore, accuracy is improved.
4. An assistant classifier based on module selection strategies for better adaptive ability and strong generalization, so any classifier algorithms can be used as the assistant classifier.

**Disadvantages**

1. Its time complexity is too high.

C] Kittler et al. [8], the dissimilar classifiers within a combination would never support to a misclassification i.e., same incorrect class must not assigned to a test instance by two or more voter classifiers. The dissimilar classifiers can be trained either by using different input representations for the information, or using different parameters for the similar type of classifier (e.g. different value of k for KNN classifier; different value of weights for an MLP classifier), or using different classifiers totally (e.g. Naïve Bayes and Decision Trees). There are set of rules for combining the outputs of dissimilar classifiers within a combination. The most common rule is the majority vote rule, which assign category to test instances if category receives the most votes. Other product, sum, min, max and median are based on mathematical functions.

**Advantages**

1. The combined classifiers improve an efficiency and accuracy.
2. This method can integrate different types of features. Hence, it can work with heterogeneous classifier.

**Disadvantages**

1. One problem with fixed set of rule is challenging to predict which rule would accomplish best result.
2. The rule considers reliable confidence and noise free estimates. It will fail if these estimates are accidentally zero or very small.

## CONCLUSION

In this survey, we have studied hierarchical classification method with their pros and cons. We came to conclusion that by combining top down and metaclassification method, problem of hierarchical classification can be solved very effectively.

## REFERENCES:

[1] S. Dumais and H. Chen, "Hierarchical Classification of Web Content," Proc. 23rd Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 00), pp. 256-263, 2000.

[2] A. Sun, E.P. Lim, W.K. Ng, and J. Srivastava, "Blocking Reduction Strategies in Hierarchical Text Classification," IEEE Trans. Knowledge and Data Engineering, vol. 16, no. 10, pp. 1305-1308, Oct. 2004.

[3] P.N. Bennett and N. Nguyen, "Refined Experts: Improving Classification in Large Taxonomies," Proc. 32nd Intl ACM (SIGIR09), pp. 11-18, 2009.

[4] H. Malik, "Improving Hierarchical SVMS by Hierarchy Flattening and Lazy Classification," Proc. ECIR Large-Scale Hierarchical Classification Workshop, 2010.

[5] D. Koller and M. Sahami, "Hierarchically Classifying Documents Using Very Few Words,'' Proc. Intl Conf. Machine Learning (ICML 97), pp. 170-178, 1997.

[6] Todorovski and S. Dzeroski, "Combining Classifiers with Meta Decision Trees," Machine Learning, vol. 50, no. 3, pp. 223-249, 2003.

[7] Q. Kong, H. Zhao, and B.L. Lu, "Adaptive Ensemble Learning Strategy Using an Assistant Classifier for Large-Scale Imbalanced Patent Categorization," Proc. 17th Int'l Conf. Neural Information Processing: Theory and Algorithms, pp. 601-608, 2010.

[8] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226-239, Mar. 1998