International Journal of Computer Science and Informatics

Volume 1 | Issue 3 Article 15

January 2012

Issues in Large-Scale Hierarchical Classifications

M.Balaji Prasath
Department of Computer Science and Engineering, Anna University Chennai, Chennai, India,
itprasath@gmail.com

D. Manjula
Department of Computer Science and Engineering, Anna University Chennai, Chennai, India,
manju@annauniv.edu

Issues in Large-Scale Hierarchical Classifications

M.Balaji Prasath & D.Manjula

Department of Computer Science and Engineering, Anna University Chennai, Chennai, India
E-mail : itprasath@gmail.com, manju@annauniv.edu

Abstract **-** Text documents in the web are in hierarchy, increase in the content, information grows over the years. To classify those text documents, need a class labels. But documents in the corpus belong to more than one class or category. Most of the corpus is large in size example. Wikipedia, Yahoo ODP directory. To classify those large-Scale dataset need a multi-label to categorize those datasets. More number of document added to the hierarchy, it create very high imbalance between classes at the different levels of hierarchy. Difficult to assign the documents to the actual class, so that relevance measure is used to calculate, relevance of text document to the class label, to maintain stable hierarchy. Another issue is if number of unique label is increase, it create instability in a classification, and also slow the classification process, so that try to limit the unique label in the classification, it improves the classification performance.

Keywords **-** *Classification, Hierarchical SVM, text mining, Large – Scale Dataset.*

I. INTRODUCTION

Multi class classification involves assigning an object to one of several classes. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers. Hierarchical classification mostly used in the classification of web text documents, growth of the documents in the corpus increases more classes to the corpus hierarchy, it creates imbalance between the classes at the different levels of the hierarchy. Suppose consider a classification, given a description d ε X of a document, where X is a document space, and fixed set of classes C = {c1, c2… c_j }. Classes also called as categories or labels. Typically a document space X is some type of high-dimensional space, and the classes are human defined for the need of an application, as in the examples India and the documents that talk about a 2G scam above. Given a training set D of labeled documents <d,c>, where <d,c>εX * c. for example,

<d,c> = < 2G Scam Bail order, India >

For one sentence document 2G Scam and the class India.

Using a learning method, wish to learn a classifier or classifier function ϒ that maps the documents to the classes.

$$\gamma: X \rightarrow C$$

This type of learning is called supervised learning. Denoted by Γ and written as Γ (D) = ϒ. The learning method Γ takes the training set D as input and returns the learned classification function ϒ. If it grows, when adding a text document to this collection, it increase classes within class or create new class label. Several issues in the large scale hierarchical classification are[2],

- Learning to classify against many categories
- Data Sparseness in the presence of large datasets.
- Use of statistical dependence of hierarchically organized classes.
- The roles of shrinkage methods in large hierarchy.
- Ensemble methods for hierarchical classification.
- Extending exiting large – scale classifier to hierarchies.
- Computational issues in large-scale classification.

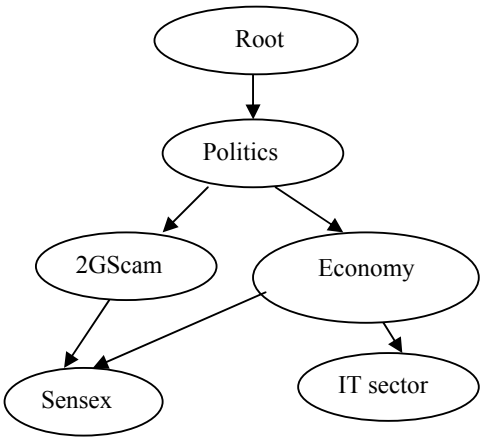
Consider a massive dataset [2] like Wikipedia, Yahoo ODP Directory, and DMOZ corpus. Those corpuses contain large document collections. In order to process that corpus, hierarchical classification is better to perform. There is a wide range of machine learning methods like support vector machine or KNN or Naïve Bayes models may be used in this scenario. But they not tried for large – scale dataset like Wikipedia, because it not like a tree structure, but grow in future, when more

documents added to a corpus. Wikipedia dataset contains more than 300,000 categories and 2M documents for training. The largest categorical system used ever in the past for evaluation process. Data pre-processing technique like stemming, stop word removal, feature extraction is very difficult to do in large scale datasets.

Section 2 Describes some of the issues, Section 3 describe about the merits and demerits of past classification used before, Section 4 describes the performance evaluation of hierarchical classification. If the dataset contains too many features, it's difficult to handle, so that reduce the number of features of the data using the algorithms like PCA or C4.5, and sometimes for problem having dimension >3 are useful using Kohonen maps.

II. ISSUES IN LARGE-SCALE CLASSIFICATION

- Classify against many categories:
- To Classify against the several categorize of hierarchy in the text-mining is difficult.



Suppose that article speaks about the petrol price hike, indicated in Business Class, and also Politics.

- Data Sparseness in the presence of large datasets:

Suppose Consider a dataset like Politics, that include records about the 2G Scam, that have a repeated words of "Raja", "Spectrum", "CBI" etc. it may cause many words have a large no of frequency counts. Several methods to identify the frequency of term is there, one is tf-idf (term frequency – inverse document frequency) method. Suppose consider a word *w* in a text document *d,* is follows [7],

$$\text{tf-idf}(w) = \text{tf}. \log (N / \text{df}(w)) \quad (1)$$

tf denotes term frequency of term *i* in the document *d*.(*no of word occurrences in the doc*)

N denotes the total no. of documents in the corpus.

df denotes the document frequency (*no of doc containing the word*)

tf-idf(*w*) denotes the relative importance of the word in a document.

Equation (1) denotes the tf-idf of the term *w* in the document *d.*

That is, due to the vastness of data, some term in the document *d*, occurred very often in the document.

- Statistical dependence of hierarchically organized classes.

Statistical dependence refers to statistical relationship between two of random variables or two set of data. Mostly the hierarchy consists of many dependence classes. Like 2G Scam and Politics, within 2G Scam, go for Election Class.

- Methods for hierarchical classification.

Support Vector Machine based on Hierarchical and Flat classification already done on different datasets. Some of them like, Hierarchical SVM based on max-cut off method, and hierarchical classification based on SVM with Naive Bayes. But hierarchical SVM outperforms the Flat SVM in the Large-Scale Datasets Problem.K-Nearest Neighbor Method (KNN) could be used in the classification of large datasets; it's sometimes outperformed the Hierarchical SVM.But KNN method very sensitive to noise [10], and also efficiency very low, when the size and dimension of training sample is high. So to avoid this difficulty use parallel KNN framework, in which training set is divided in to these parts, and nearest neighbor's are retrieved in parallel from these parts and a final results are synthesized from these parts[11] .Therefore, it is highly desired to reduce the dimensionality of data by selecting desired features and removing the noise. If u integrate the hierarchical information in the centroid based classifiers, it reduce the training and test timings [9].

Datasets	**Training**	**Test**
Small	Less 1 min	3 – 5 mins
Large	2 – 3 mins	2 -3 hours

- Extending exiting large – scale classifier to hierarchies.

Suppose that, add the new class to the existing Hierarchy is a complex task, it create an imbalance

between the classes at the different levels, when the hierarchy grows. Suppose that to add a document based on 2G Scam, it need a Separate Class, it may included in more than one class, it create a statically dependence between a class, so that to maintain a relationship between class hierarchy in the large-scale dataset is complex, but it may create a new classes in the hierarchy, when the no of document in the corpus grows.

- Computational issues in large-scale classification

To process over a million documents in the, it need a large storage space to accumulate those data, it's difficult to process those dataset in practical, but it should be done. Evaluation process involves computing the accuracy, example-based F-measure, label-based macro F-measure, label-based micro F-measure and multi-label graph-induced error.

III. CLASSIFICATION METHODS

Support vector machine classification mainly used for binary classification, but several methods proposed to extend the binary classifier to handle the multiclass problem. Support Vector Machine [1] [4] [5]: in the case of flat SVM, when the no. of trained nodes increases, it increases computational loads at the unacceptable levels compared to the hierarchical SVM. First, let us consider the complexity of flat SVM.With the one-against-rest strategy [8], when training the SVM model for any individual category, we always use the entire training set considering that the complexity of SVMs grows super-linearly with the number of training documents the overall complexity of flat SVMs can be represented by Complexity of flat SVM

$$Q_{train}^{flat} = M.\ O\ (N^{\varepsilon}),\ C > 1 \qquad (2)$$

Where N is the number of training documents; M is the number of categories in the training set, and O (N^{ε}) denotes the average training time per SVM model.

For testing, since we need to pass an instance onto all of the M SVM models to find the category with the highest confidence score, the time complexity is

$$Q_{test}^{flat} \sim M.O\ (1) \qquad (3)$$

Where O (1) denotes the average test time per document per SVM model.

In the case of Hierarchical classification [5] is little more complicated, because size of training set is varies at different levels of hierarchy. Accordingly, the complexity is,

$$Q_{train}^{hierarchical} \sim 1/M \Sigma^{n}_{i=0}\ b_i\ .\ Q_{test}^{flat} \qquad (4)$$

Where, N denotes the hierarchy levels in the dataset.

K-Nearest Neighbor Method : In the case of Multi Class classification KNN method is better than the SVM model [11].Multi-lable classification KNN method uses the Euclidean distance or Cosine Similarity measurement used to measure the distance between the test samples.

IV. PERFORMANCE EVALUATION

To evaluate the performance of the classification hierarchy involves two stages [2], (i) computational performance (ii) classification accuracy. Computational performance involves computational and memory requirements of each method or the method used in the classification. But due to the dimension of dataset computational performance may be low, but the classification accuracy should be improved upon the classification task, while choosing vital features from a dataset, reduce the dimension of dataset.

V. CONCLUSIONS

Size and the dimension of the corpus play a vital role in the classification of dataset and also accuracy. Machine learning methods like a KNN lacks in process the huge dimension of dataset, and SVM produce better classification results, and centroid based method with hierarchical information involves a less training and testing time, perceptron like a Neural Network involves a huge training time, train the network involves a lot of computational time, so that choose a learning algorithm based a classification task.

REFERENCES

[1] Akbani R., Kwek S., Japkowicz N. "Applying support vector machines to imbalanced datasets" ECML, 39-50, 2004

[2] Second Pascal Challenge on Large Scale Hierarchical Text classification, The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2011Athens, Greece on September 5th, 2011

[3] Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma," Support Vector Machines Classification with A Very Large-scale Taxonomy"SIGKDD Explorations, Volume 7, Issue 1 - Page 43.

[4] Susan Dumais , Hao Chen ," Hierarchical Classification of Web Content".

[5] iao-Lin Wang, Bao-Liang Lu," Improved Hierarchical SVMs for Large-scale Hierarchical Text Classification Challenge ".

[6] Tsoumakas, G., Katakis, I., Vlahavas, I.: "Random k-Labelsets for Multi-Label Classification ". IEEE Transactions on Knowledge Discovery and Data Engineering (2010)

[7] http://nlp.stanford.edu/IRbook/html/htmledition/tf-idf-weighting-1.html

[8] Chih-Wei Hsu and Chih-Jen Lin "A Comparison of Methods for Multiclass Support Vector Machines", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 2, MARCH 2002

[9] Youdong Miao and Xipeng Qiu "Hierarchical Centraiod Based Classifier for Large Scale Text Classification".

[10] Xipeng Qiu, Jinlong Zhou and Xuanjing Huang "An Efficient Feature Selection method for Text Categorization" Lecture Notes in Computer Science, 2011, Volume 6634/2011, 50-61, DOI: 10.1007/978-3-642-20841-6

[11] JIAN ZHANG, HAI ZHAO, BAO-LIANG LU1, "A Comparative Study On Two Large-Scale Hierarchical Text Classification Tasks" Solutions Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

[12] Dingquan Wang, Weinan Zhang, Gui-Rong Xue, and Yong Yu "Deep Classifier for Large Scale Hierarchical Text Classification".

◈◈◈