IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

A Review of Label Dependency and Feature Similarity for Multi-Label Classification

Rohit Tiwari[1], Shivank Kumar Soni[2]

M.Tech Scholar, Department of CSE RITS, Bhopal, India[1]

Professor Department of CSE RITS, Bhopal, India[2]

Abstract: The increasing rate of data diversity in current decade faced a problem of data categorization. Data categorization used a classification technique such as KNN, decision tree and support vector machine. The process of classification divided into two sections one is trained model and other is test model. The assign class measured the similarity between trained and test. The dependency of feature bound the limitation of accuracy of classifier. The process of classification mapped data into labels and labels categorized in the different predefined class for the classification purpose. This paper concludes multi label classification technique removal of similarity of dependency.

Keywords: Multi-Class Classification, Label Dependency, Data Mining.

I. INTRODUCTION

In today's world we are mostly dealing with multi-label data, data assigned to more than one class or label. A lot of ongoing research is focusing on multi-label data sets. Let us consider some real world examples for multi-labeled data. A text document about scientific contributions in medical science can belong to both science and health categories. An images that capture a field and fall colored trees can belong to both field and fall foliage categories [1]. In a conventional classification task, given a set of m possible disjoint classes, each instance is associated with one and only one class. Different kinds of machine learning algorithms, such as the k-nearest neighbor [12], support vector machine, and logistic regression methods, teaching learning based optimization [16] have been proposed to resolve such classification problem, and have achieved a satisfactory level of success.

Web search result visualization is a process that layouts search results in a more clear and coherent way according to the content of each result. It aims at increasing the search efficiency and accuracy and ameliorating users' browsing experiences. Prevalent techniques for this task is text clustering. It regards the visualization task as an unsupervised pattern classifications problem. Following the methodology of pattern classification, text features are first extracted from text to represent the document, and then the document is assigned to a cluster in which documents have high similarity. Existing multi-label classification efforts for networked data focus on designing effective and yet scalable algorithms [13].

Although most of the algorithm differ from one another in the concrete approaches to mining the linkage structure, to the best of our knowledge, they all suffer from two weaknesses: (1) None of previous studies separate different types of activity graphs from the heterogeneous information networks and exploit the correlations among the set of class labels within each activity graph and across multiple activity graphs; and (2) None of previous works combine both the vertex-centric multi-label classification and the edge-centric multi-label classification to boost the effectiveness and efficiency [14]. The discovery of incident-related information is a complex task, requiring the separation of valuable information from daily chatter in the vast amount of information created on social platforms. This can be realized based on techniques from data mining and machine learning. Classification is one method which can be utilized to extract relevnt information from social networks [8].

In a classification task, a system learns to label messages with exactly one label out of a predefined label set (e.g. “fire” or “crash”). Some reviewer proposed a novel bi-directional model for multi-label classification, which introduces a compact mid-level representation layers between the input features and the output labels to capture the common prediction representations shared across multiple labels [9]. The mid-level representation layer is constructed from both input and output space, and it has two complementary parts, one of which captures the predictive low-dimensional semantic representation of the input feature and the other capture the predictable low-dimensional intrinsic representation of the output labels.

These two part augment each other to integrate information encoded in both the feature and label spaces and enhance the overall multi-label classification. Canonical Correlation Analysis (CCA) finds the correlations between two sets of variables and projects are two views of the same objects onto a low-dimensional data space where their correlations are maximized [6]. For multi-label problems, one view comes from the data features and the other view is generated by the group of multiple labels. CCA has found widespread applications in

DOI 10.17148/IJARCCE.2016.58127

multi-label learning, e.g., CCA was utilized to enhance multi-label image observation by structural grouping sparsity. Section II gives the information related work. In section III discuss the problem state and formulation. In section IV discuss comparative study of different methods and finally in section-V conclusion and future scope.

II RELATED WORK

In this section discuss the related work of dependency of label in classification technique also discuss the removal of label dependency in classification technique. In 2011, Zhihua Wei et. al. described a Naive Bayesian (NB) multi-label classification algorithm is proposed by incorporating a two-step feature selection strategy in 2011 which aims to satisfy the assumptions of conditional independency in NB classification theory. The experiments over public data set demonstrate that the proposed methods has highly competitive performance with several well-established multi-label classification algorithms. They implement a prototype system named TJ-MLWC based on the proposed algorithm, which acts as an intermediate layer between user and a commercial Internet Search Engine, allowing the search results of a query displaying by one or multiple categories. Testing results indicate that our prototype improves search experience by adding the function of browsing search results by category [3]. In 2011, Hang Li et. al. gives an introduction to learning to rank, and it specifically explains the fundamental problems, existing approaches, and coming work of learning to rank. Several learning to rank methods using SVM techniques are described in details. Learning to rank can be employed in a wide variety of applications in Information Retrieval (IR), Natural Language Processing (NLP), and Data Mining (DM). Typical applications are document retrieval, expert search, definitions search, collaborative filtering, questions answering, key-phrase extraction, document summarization, and machine translation [4]. In 2011, Xin Li et. al. described a novel supervised bi-directional model that learns a low-dimensional mid-level representation for multi-label classification. Unlike the traditional multi-label learning methods which identify intermediate representations against either the input space or the output space but not both, the mid-level representation in our model has two equivalent parts that capture intrinsic information of the input data and the output labels respectively under the auto encoder principle while augmenting one and all for the target output label prediction. The resulting optimization problem can be solved efficiently applying an iterative procedure with alternating steps, while closed-form solutions exist for one major step [9].

In 2012, Tsung-Hsien Chiang et. al. has been presented an interesting finding, namely, being able to identify neighbors with trustable labels can significantly improve the classifications accuracy in 2012. Based on this finding, we propose a k-nearest-neighbor-based ranking approaches to solve the multi-label classification problem. The approach exploits a ranking model to learn which neighbor's labels are more credible candidates for a weighted KNN-based strategy, and then assigns higher weights to those candidate when making weighted-voting decisions. The weights can then be determined by using a generalized pattern search technique [2] .

In 2012, Purvi Prajapati et. al. introduce the task of multi-label classification, methods for multi-label classification and evolution measure for multi-label classifications. Also done comparative analysis of multi label classification methods on the basis of theoretical study and then on the basis of simulation done on various data sets. Multi-label classification methods are increasingly compulsory by modern applications, such as text classification, gene functionality, music categorization and semantic scenes classification. The number of class labels is predicted for each instance [7] . In 2014, Prema Nedungadi et. al. described combination of both k-Nearest Neighbor (KNN) algorithm and multiple regressions in an efficient way for multi-label classification. KNN works well in feature space and multiple regression works well for preserving label dependent information with generated models for labels. Their classifier incorporates feature similarity in the feature space and label dependency in the label space for prediction. It has a wide range of applications in various domains such as in information retrieval, query categorization, medical diagnosis and marketing. The results obtained with various multi-labeled dataset justify our method. Future work would involve experimentation with more datasets and would focus on increasing processing and calculation speed [1].

In 2014, Yang Zhou et. al. introduced a novel concept of vertex-edge homophile in terms of both vertex labels and edge labels and transform a general collaboration graphs into an activity-based collaboration multi graph by augmenting its edges with class labels from each activity graphs through activity-based edge classification. Second, we utilize the label vicinity to pick up the pair-wise vertex closeness based on the labeling on the activity-based collaboration multi graph. They incorporate both the structure affinity and the label vicinity into a unified classifier to speed up the classifications convergence. Third, we design an iterative learning algorithm, AEC class, to dynamically refine the classification results by continuously adjusting the weights on different activity-based edge classification schemes from different activity graphs, while constantly learning the contribution of the structure affinity and the label vicinity in the unified classifier [5]. In 2014, Yaqing Wang et. al. described to construct a hyper graph for exploiting the high-order label relations and in current edge a novel framework for multi-label classification named Hyper graph Canonical Correlation Analysis (HCCA). This idea is based on canonical correlation analysis, and it further takes into account the high-order label structure information via hyper graph regularization. Thus, the label relations can be better respected both globally by the normalized similarity matrix of CCA and locally by the normalized hyper graph Laplacian in a unified framework [15]. In concrete, the objective method can be develop by solving a generalized

DOI 10.17148/IJARCCE.2016.58127

Eigen value problem, but this requests heavy computational overheads for huge amount of data. Therefore, we show a more efficient method that approximates the original problem by the minimum squares formulation under a mild condition [6].

In 2014, Axel Schule et. al. contributed the first in-depth analysis of multi-label classification of incident-related tweets. They present an approach assigning multiple labels to these messages, providing additional information about the position at-hand. An evaluation shows that multi-label classification is applicable for detecting multiple labels with an exact match of 84.35%. Thus, it is an important means for classifying incident-related tweets. Furthermore, we show that interrelationship between labels can be taken into account for these kinds of classification tasks [8].

III PROBLEM FORMULATION

In this section discuss the problem of multi-class classification technique and their dependency of similarity. The most of data mining classification technique is single classifier, but the extension of classifier work as multi-class classification. The extension of classifier comes along with ensemble technique and multi-class support vector machine [11].

In multi-class classification when dealing with different classes, as in object recognition and image classification, one needs an appropriate multi-class methods. Different possibilities include: Modify the design of the SVM, as in order to incorporate the multi-class learning precisely in the quadratic solving algorithm. Combine several binary classifiers: “One-against- One” (OAO) used for pair wise analogy between classes, while “One-against-All” (OAA) compares a given class with all others put together. In the “One-against-All” algorithm, n hyper planes are constructed, where n is the number of classes. In the process of review we found that some performance affected problem related to the imbalance data classification. These problem are affected the performance and accuracy of multi-class classifier and generate unclassified region. The unclassified region increase, decrease the accuracy and performance of classifier. Some problems are mentioned here [4, 6, 9, and 10].

1. Infinite population of data.
2. Feature selection of data
3. Voting of class
4. New class generation.
5. imbalanced data problem
6. Error Correcting Code
7. Label dependency

IV COMPARATIVE STUDY

In this section discuss the comparative study of different algorithm work in label dependency removal during the process of classification.

S.N	Methods	Merits	Demerits
1	KNN	Its simple classification algorithm, complexity is very low as compared to another classification algorithm.	Lazy classifier, the classification rate is low as compared to another classifier approximate 78%
2	NB	Probability based classification technique. Very efficient for small dataset.	Repetitive probability value conflict the decision of classification. The maximization of probability value decreases the classification rate.
3	SVM	Regression based classification technique. The classification rate is very high as compared to another classifier.	Boundary and outlier related problem. For the distribution of data used kernel function. Change the function of kernel according to the data.
4	Graph theory	Supported new feature evaluation concept for the classification. Remove the problem of data drift.	It’s very complex algorithm. The complexity and execution time is very high.
5	CCA	Reduces the gap of feature attribute. Increase the classification rate incorporation to another algorithm.	Not supported different feature attribute of same data. Used only case of neural network based classifier.
6	TLBO	TLBO algorithm remove the dependency of label. It is increase the accuracy and performance of classifier.	Convergence rate is the main problem of TLBO algorithm.

We concluded from this comparative study for removal of label dependency during the classification the support vector machine (SVM) is very good as compared to others. The TLBO algorithm improved the accuracy of minority class of classifier and reduces the unclassified region in multi-label classification. The increasing of multi-label classification region improved the accuracy and performance of classifier. Increase the accuracy of classifier; Remove the dependency of label, Reduces size of data, Decrease the feature dissimilarity and used real time data for the classification.

V CONCLUSION & FUTURE WORK

In this paper we present the review of multi-class classification technique in terms of label dependency and removal of label dependency. The removal of label dependency is very critical task in multi-class classification. For the removal of label various authors are used optimization technique for the better prediction of class and classifier. In multi-class classification process the building of classifier depend upon the process of classification technique. Now ensemble process of classification built the concept of multi-class classification. In this paper basically we discuss various methods but the performance of TLBO algorithm for the optimization of data feature and remove the dependence of label is good. In future implementation researchers can use this concept and some standard data for the evaluation of his new research in this area.

REFERENCES

[1] Prema Nedungadi, H. Haripriya “Exploiting Label Dependency and Feature Similarity for Multi-Label Classification” IEEE, 2014. Pp 2196-2200.

[2] Tsung-Hsien Chiang, Hung-Yi Lo, Shou-De Lin “A Ranking-based KNN Approach for Multi-Label Classification” Workshop and Conference Proceedings, 2012. Pp 81-86.

[3] Zhihua Wei, ongyun Zhang, Zhifei Zhang, Wen Li, Duoqian Miao “A Naive Bayesian Multi-label Classification Algorithm With Application to Visualize Text Search Results” International Journal of Advanced Intelligence, Volume-3, 2011. Pp 173-188.

[4] Hang LI “A Short Introduction to Learning to Rank” IEICE TRANS. INF. & SYSTEM, Vol-94, 2011. Pp 1-9.

[5] Yang Zhou, Ling Liu “Activity-edge Centric Multi-label Classification for Mining Heterogeneous Information Networks” ACM, 2014. Pp 1-10.

[6] Yaqing Wang , Ping Li , Cheng Yao “Hypergraph canonical correlation analysis for multi-label classification” Elsevier ltd, signal processing, 2014. Pp 258-267.

[7] Purvi Prajapati, Amit Thakkar, Amit Ganatra “A Survey and Current Research Challenges in Multi-Label Classification Methods” International Journal of Soft Computing and Engineering, 2012. Pp 248-252.

[8] Axel Schule, Eneldo Loza Mencía, Thanh Tung Dang, Benedikt Schmidt “Evaluating Multi-label Classification of Incident-related Tweets” Micropost workshop proceeding, 2014. Pp 26-33.

[9] Xin Li, Yuhong Guo “Bi-Directional Representation Learning for Multi-label Classification” 2011. Pp 1-15.

[10] Francisco Charte, AntonioRivera, Marıa Josedel Jesus, Francisco Herrer “Improving Multi-label Classifiers via Label Reduction with Association Rules” Springer, 2012. Pp 188-199.

[11] Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang “Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval” In Multimedia and Expo, IEEE, 2010. Pp 304-309.

[12] Grigorios Tsoumakas, Ioannis Katakis, Ioannis Vlahavas “Random k-labelsets for multilabel classification” IEEE Transactions on Knowledge and Data Engineering, 2011. Pp 1079-1089.

[13] M.-L. Zhang, K. Zhang “Multi-label learning by exploiting label dependency” In Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD’10), 2010, Pp 999-1007.

[14] G. Chen, Y. Song, F. Wang, C. Zhang “Semi-supervised multi-label learning by solving a Sylvester equation” In Proceedings of the SIAM International Conference on Data Mining, 2008, Pp 410-419.

[15] Y. Liu, R. Jin, L. Yang “Semi-supervised multi-label learning by constrained non-negative matrix factorization” In Proceedings of the 21st National Conference on Artificial Intelligence, 2006, Pp 421-426.

[16] Suresh Chandra Satapathy, Anima Naik and K Parvathi” A teaching learning based optimization based on orthogonal design for solving global optimization problems” in Springer Open Journal, 2013.

DOI 10.17148/IJARCCE.2016.58127