Chapter 5
Effective Multi-Label Classification Using Data Preprocessing

Vaishali S. Tidake
https://orcid.org/0000-0003-4543-6361
MVPS's KBT College of Engineering, Nashik, India

Shirish S. Sane
K. K. Wagh Institute of Engineering Education and Research, Nashik, India

ABSTRACT

Usage of feature similarity is expected when the nearest neighbors are to be explored. Examples in multi-label datasets are associated with multiple labels. Hence, the use of label dissimilarity accompanied by feature similarity may reveal better neighbors. Information extracted from such neighbors is explored by devised MLFLD and MLFLD-MAXP algorithms. Among three distance metrics used for computation of label dissimilarity, Hamming distance has shown the most improved performance and hence used for further evaluation. The performance of implemented algorithms is compared with the state-of-the-art MLkNN algorithm. They showed an improvement for some datasets only. This chapter introduces parameters MLE and skew. MLE, skew, along with outlier parameter help to analyze multi-label and imbalanced nature of datasets. Investigation of datasets for various parameters and experimentation explored the need for data preprocessing for removing outliers. It revealed an improvement in the performance of implemented algorithms for all measures, and effectiveness is empirically validated.

INTRODUCTION

Many scenarios in the real-life today depict applications of multi-label data. A document may be related to health as well as education, according to its text. A piece of news may focus on new technology that is helpful for safety as well. An image may contain several objects like roads, shops, buildings, etc. Contents of a paper may be relevant to multiple domains. A video may focus on topics of networking

DOI: 10.4018/978-1-7998-7371-6.ch005

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

along with virtualization. Thus many objects reveal multiple semantic meanings. Many researchers are working for the last few decades on multi-label classification. It is a task that assigns with a thing a set of predefined labels as per its properties.

BACKGROUND

The related work about multi-label classification and label imbalance is presented here. For multi-label classification, there exist methods that use the *transformation* approach. It changes multi-label data such that methods for single-label classification can be used. Sometimes multi-label data is not modified. Thus *adaptation* methods modify the process of dealing with such data. There also exists an approach that ensembles multiple existing methods. CC (Read, 2009), MLkNN (Zhang & Zhou, 2007) and RAkEL (Tsoumakas et al., 2011) are examples of these three approaches respectively.

For few decades, many researchers have worked in the field of multi-label classification (Tsoumakas & Katakis, 2007) (Tsoumakas et al., 2009) (Trohidis et al., 2008) (Tsoumakas et al., 2010) (Madjarov et al., 2012) (Zhang & Zhou, 2014) (Tidake & Sane, 2018). K nearest neighbor has also been the choice of many researchers for multi-label classification. From the study, it is noticed that neighbors are obtained using only features always. In contrast, the scenario is different for data that is multi-label. Each instance belongs to a predefined set of labels. Hence it is possible to consider labels along with features for obtaining neighbors.

Zhang and Zhou discuss an approach in (Zhang & Zhou, 2007). It follows an *algorithm adaptation* approach. It is an improved version of the well-known nearest neighbor algorithm. Several researchers use it to perform multi-label classification. It utilizes feature similarity to determine nearest neighbors (Zhang & Zhou, 2005) (Zhang & Zhou, 2007) (Spyromitros-Xioufis et al., 2008). In the case of multi-label classification, since the instances are associated with multiple labels, label dissimilarity may also help determine a set of nearest neighbors.

Class imbalance also poses problems to multi-label classifiers and may lower their performance. According to Spyromitros-Xioufis (2011), label skew is considered a class imbalance when considering each class individually. Francisco et al. (2013) have proposed how to measure the level of imbalance in a multi-label scenario. They have also presented two dataset preprocessing methods specially designed for multi-label datasets. They used sampling and LP for preprocessing. Those label sets that occur in a majority (minority) were reduced (increased). A method was suggested by Huang et al. (2015) for the improvement of multi-label classifier involving several binary classifiers. It can be used for feature selection also. SOSHF was extended from structured forests (Zachary et al., 2017). At each node, it has used transformation followed by split action to tackle class imbalance. An imbalance ratio was defined using positive and negative samples (Zhang et al., 2018). This ratio and label correlation was considered to improve BR models. Liu and Tsoumakas (2018) have handled the imbalance faced by ECC. They used an ensemble of CC with random under-sampling that helps to balance the distribution of each class. COCOA method explored joint label correlation and imbalance ratio from skewness between positive and negative samples (Zhang et al., 2020). It induced an imbalanced multi-class classifier per label.

18 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/effective-multi-label-classification-using-data-preprocessing/280912

Related Content

AI-Based Automated Disease Detection Based on Symptoms Healthcare for Pets and Birds
Kathirvel Ayyaswamy, V. M. Gobinath, Naren Kathirveland C. P. Maheswaran (2025). *AI and the Revival of Big Data (pp. 311-346).*
www.irma-international.org/chapter/ai-based-automated-disease-detection-based-on-symptoms-healthcare-for-pets-and-birds/369506

Insights From Big Data Analytics With Machine Learning-Driven Predictive Maintenance in the Automotive Industry
Anil Kumar Adikeand S. Silvia Priscila (2026). *Machine Learning, Predictive Analytics, and Optimization in Complex Systems (pp. 103-122).*
www.irma-international.org/chapter/insights-from-big-data-analytics-with-machine-learning-driven-predictive-maintenance-in-the-automotive-industry/384450

Unit-Selection Speech Synthesis Method Using Words as Search Units
Hiroyuki Segi (2016). *International Journal of Multimedia Data Engineering and Management (pp. 1-15).*
www.irma-international.org/article/unit-selection-speech-synthesis-method-using-words-as-search-units/152868

Semi-Supervised Multimodal Fusion Model for Social Event Detection on Web Image Collections
Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, Haiwei Panand Yangbin Chen (2015). *International Journal of Multimedia Data Engineering and Management (pp. 1-22).*
www.irma-international.org/article/semi-supervised-multimodal-fusion-model-for-social-event-detection-on-web-image-collections/135514

Adaptive Security at the Edge With Real-Time Access Control and Monitoring Beyond Zero Trust
Sanjay Poddar (2025). *Data Governance, DevSecOps, and Advancements in Modern Software (pp. 321-334).*
www.irma-international.org/chapter/adaptive-security-at-the-edge-with-real-time-access-control-and-monitoring-beyond-zero-trust/377005