A Review on Multi-label Classification

Mariyam Ansari[1] & Prof. N. M. Shahane[2]
[1]M.E. Student, [2]Associate Professor
[1]Department of Computer Engineering,
[1]K.K.Wagh Institute of Engineering Education and Research, Nashik,
Savitribai Phule Pune University, Maharashtra, India.

Received: December 09, 2018 Accepted: January 14, 2019

ABSTRACT: *Multi-label Classification is a form of supervised learning where the classification algorithm is learned from the set of training instances and based on this learned model the instances are classified into a label-set. In multi-label classification an instance is associated with more than one label at the same time which is inherently different from the traditional classification where an instance is classified into a single label. This paper includes the basic introduction of multi-label classification and existing methods for multi-label classification are studied along with their evaluation metrics.*

Key Words: *Machine Learning, Multi-label Classification, Supervised learning, nearest neighbor*

I. INTRODUCTION

Traditional *single-label* classification is concerned with learning from a set of instances that are associated with a single label l from a set of disjoint labels L, $|L| > 1$. If $|L| = 2$, then the learning problem is called a *binary* classification problem else if $|L| > 2$, then it is called a *multi-class* classification problem.

In a multi-label dataset, every instance x is described by a number of input features and associated with a labelset. This labelset is represented as a binary vector $L_x = <l_1(x), l_2(x), \ldots, l_m(x)>$, with m the total number of possible class labels in the dataset. The value $l_i(x)$ indicates whether or not x belongs to class l_i. The task of a multi-label classifier is to predict the complete labelset of a target instance. This is inherently different from single-label classification, where only one outcome label needs to be predicted.

Multi-label data are ubiquitous in real-world applications. There exists a wide range of applications which use multi-label classification, such as text categorization, semantic image labeling, gene functionality classification, bioinformatics etc. and the scope and interest is increasing with modern applications. .

Several approaches to multi-label classification have been proposed in the literature[1]. There are two strategies for multi-label classification the data transformation methods and method adaptation algorithms which are explained in this review paper.

II. Data transformation Methods

The first method transforms the multi-label dataset into one or more easier-to-handle single-label dataset, on which a single-label classifier can be applied. Two well-known methods of this family are the Binary Relevance (BR, [2]) and Label Powerset (LP, [3]) transformations. In BR transformation method m binary single-label datasets are created, one for each class. Each dataset contains the same instances as the original multi-label dataset, but with a single label. If for a particular dataset of label l_i, an instance is related then the instance is labeled as positive(1) else it is negative(0).Whereas in LP transformation method creates only one single-label dataset. Each possible labelset receives an identifier, which is used as the single new class label, such that labelsets that entirely coincide are associated with the same identifier.

III. Method adaptation Algorithms

The second family of multi-label classification algorithms, method adaptation algorithm handle the multi-label dataset directly and are often based on modifications or generalizations of existing single-label classification schemes. There are several approaches which can be used and are listed below.

A. Lazy Learning:

These algorithms are very similar in the sense that they all use K nearest neighbor as a lazy learning approach, but what differentiates them is the aggregation of the label sets of the given instances. Two basic method were proposed by combining the LP and BR transformation methods with k-NN which are explained in detail in section V.

B. Neural Networks and Multi-layer perceptron based algorithms:

The Neural Networks and Multi-layer perceptron based algorithms that has been extended for multi-label data. In BP-MLL [4], the error function for the very common neural network learning algorithm, back-

propagation has been modified to account for multi-label data. Multi-layer perceptron is easy to extend for multi-label data where one output node is maintained for each class label. A family of online algorithms for Multi-class Multi-layer Perceptron (MMP) has been proposed in [5] where the perceptron algorithms weight update is performed in such a way that it leads to correct label ranking.

C. Tree Based Boosting:

AdaBoost.MH and AdaBoost.MR [6] are two simple extensions of AdaBoost for multi-label data where the former tries to minimize hamming loss and the latter tries to find a hypothesis with optimal ranking. In AdaBoost.MH, examples are presented as example-label pairs and in each iterations increases the weights of misclassified example-label pairs. In contrast, AdaBoost.MR works on pairs of labels for any instance and in each iteration increases the weights of the example with mis-ordered label pairs.

D. Discriminative SVM Based Methods:

One important problem with tree based boosting [6] is that, they are likely to overfit with relatively smaller (< 1000) training set. Elisseeff et. al. in [7] propose an SVM algorithm that has an intuitive way of controlling such complexity while having a small empirical error. Godbole et. al.[16] present three improvements for BR with SVM to exploit label correlations that improves the margin. predictions of each binary classifier at the first round. The k new binary classifiers are trained on this extended dataset. In this way the extended BR takes into account potential label dependencies. The second idea, ConfMat, based on a confusion matrix, removes negative training examples of a complete label if it is very similar to the positive label. The third idea is called BandSVM, removes very similar negative examples that are within a threshold distance from the learned decision hyperplane, and this helps building better models especially in the presence of overlapping classes.

IV. Multi-label Learning and Fuzzy Logic Based Learning

It is important to distinguish between multi-label classification and fuzzy logic based classification. While they both define membership functions to deal with multiple classes, but the goal and the problems addressed are quite different. Fuzzy logic often deals with the ambiguity in the feature space and used as an additional block before classification to help distinguishing between multiple classes. In contrast, multi-label classification is about labelling an instance with one or more classes. Fuzzy based classification is often followed by a de-fuzzification step that makes the final classification decision. Fuzzy membership values when normalized usually sum up to 1 where there is no such requirement for multi-label classification; even each class in multi-label classification can have a membership value of 1 (ideally). Usually, fuzzy membership values over different classes are correlated but membership values for multi-label classification could be just coincidence.

V. Nearest neighbor based multi-label classification

Several multi-label classifiers based on or extending KNN have been proposed. Two basic approaches were proposed in [] , by [8] combining the LP and BR transformations with the single-label kNN classifier.

A. LPKNN Method

This is the basic approach in which an instance is classified by first locating its k nearest neighbors and then predicting the most prevalent labelset among these elements. LP transformation is performed on the multi-label dataset then classification with the single-label KNN classifier [8].

B. BRKNN Methods

The BRKNN method is equivalent to using the single-label KNN method within a binary relevance setup, but runs considerably faster, since it computes the nearest neighbors of an instance only once instead of m times. The classifier assigns x to a class l when this label is present in at least half of its nearest neighbors. Using this heuristic, it is possible that x is not assigned to any class at all. To address this issue, the authors of [8] developed two extensions, called BRKNN-a and BRKNN-b. For each class l , these methods set the label confidence score to the percentage of the nearest neighbors of x that belong to class l . When none of the classes is present in at least half of its neighbors, BRKNN-a assigns x to the class with the highest label confidence score, while BRKNN-b assigns x to the s classes with the highest label confidence score, where s is the average size of the labelsets of the neighbors of x .

C. MLKNN and related methods

The MLKNN method[9] makes label predictions based on the maximum a posteriori principle and the k nearest neighbors of a target instance. Simply put, it counts the occurrences of all classes among the neighbors and evaluates how likely the presence of a class label is based on these counts. It has been pointed out in the literature that a limitation of the MLKNN method is that it does not take label correlations into account. The FSKNN method [10] uses a fuzzy similarity measure to first group the training elements into

clusters. It reduces the computational cost of the neighbor search of MLKNN, since it only uses a subset of the clusters to locate them. The authors of [11] proposed the IBLR method that combines nearest neighbor based learning and logistic regression. The class labels of the neighbors of an element are considered as supplementary features. One classifier, a logistic regression model, is trained for each class, but dependencies between labels are taken into account. The IBLR+ generalization takes additional features of the target instance into account, aside from its neighborhood information. In IBLR, the bias term of the logistic regression model is constant, while it depends on the target instance in IBLR+.

D. Multi-label classification using k-NN and Fuzzy rough set theory:

A new member of this family is proposed in this paper which aggregates the labelsets of the k nearest neighbors to a prediction based on fuzzy rough set theory. Fuzzy rough set theory [13] is an alternative to traditional set theory and models uncertainty in data. It covers two complimentary aspects of uncertainty, namely vagueness and indiscernibility. The vagueness is because of unclear descriptions of concepts, to which elements can belong to a certain degree. As an example, the set of elements that are similar to a given element x is necessarily fuzzy, since some elements are intrinsically more similar to x than others and making a strict division between similar and non-similar is difficult. The membership degree of an element to a fuzzy set is denoted by a real number between 0 and 1. Roughness in a dataset concerns the issue when observations that have indistinguishable descriptive features but have distinct outcomes. In this case, it is challenging to sharply predict the outcome concept based on the input features. Instead, a lower and an upper approximation are provided. Fuzzy rough set theory was developed as a hybridization of fuzzy set theory [14] and rough set theory [15] and has been used successfully in a variety of machine learning techniques.

IV. EVALUATION METIRCS

In traditional classification such as multi-class problems, accuracy is the most common evaluation criteria. Additionally, there exists a set of standard evaluation metrics that includes precision, recall, f-measure defined for single label multi-class classification problems. However, in multi-label classification, predictions for an instance is a set of labels and, therefore, the prediction can be fully correct, partially correct (with different levels of correctness) or fully incorrect[17]. None of these existing evaluation metrics capture such notion in their original form. This makes evaluation of a multi-label classifier more challenging than evaluation of a single label classifier. Following are some of evaluation metrics.

Let $x \in X$ be a test instance, L_x its true class vector and $\hat{L}_x$ the predicted class vector. The above evaluation metrics are defined as follows:

i)**Hamming Loss:**

$$hloss = \frac{1}{|X|}\frac{1}{m}\sum |Lx \,\Delta\, \hat{L}x\,|$$

where the Δ operator constructs the symmetric difference between its two arguments and the $|\cdot|$ operation measures the cardinality of the resulting set. The total number of prediction errors is divided by both the number of instances $|X|$ and number of labels m.

ii)***F*-measure:**

$$F = \frac{2.p.r}{p+r}$$

where p and r are the precision and recall measures calculated by

$$p = \frac{1}{|X|}\sum_{x \in X}\frac{|Lx \cap \hat{L}x|}{|\hat{L}x|}$$

and

$$r = \frac{1}{|X|}\sum_{x \in X}\frac{|Lx \cap \hat{L}x|}{|Lx|}$$

iii)**Subset Accuracy:**

$$SubAcc = \frac{1}{|X|}\sum_{x \in X} I(Lx = \hat{L}x)$$

where the indicator function $I(\cdot)$ evaluates to 1 if its argument is true and to 0 otherwise. This is the most stringent metric of the three, because it evaluates full equality of L_x and $\hat{L}_x$.

VI. CONCLUSION

In this paper, study of different multi-label algorithms, their applications and evaluation metrics has been presented. A sparse set of existing algorithms has been explained. The broad classification strategies i.e. data transformation methods and method adaptation algorithms have been studied. We have focused on the lazy learner based algorithms which use k-NN as the learning algorithm. In conclusion multi-label classification techniques have been reviewed.

VII. Acknowledgment

I would like to express my gratitude to Prof. N. M. Shahane, Associate Professor, Computer Engineering Department, K.K.W.I.E.E.R., Nashik for giving me moral support, valuable guidance and encouragement in making this survey paper. A special thanks to Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R, Nashik for their kind support and suggestions.

References

1. F. Herrera , F. Charte , A. Rivera , M. Del Jesus , Multilabel Classification: Problem Analysis, Metrics and Techniques, Springer, 2016
2. S. Godbole , S. Sarawagi , Discriminative methods for multi-labeled classification, in: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2004, pp. 22–30 .
3. P. Bhowmick , A. Basu , P. Mitra , A. Prasad , Sentence level news emotion analysis in fuzzy multi-label classification framework, Special issue: Natural Lang. Process. Appl. (2010) 143.
4. M. L. Zhang and Z. H. Zhou. Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering, 18(10):1338–1351, 2006.
5. Koby Crammer and Yoram Singer. A family of additive online algorithms for category ranking Journal of Machine Learning Research, 3:1025–1058, 200.
6. Schapire. Boostexter: a boosting-based system for text categorization. Machine Learning, 39(2/3):135–168, 2000.
7. A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In Advances in Neural Information Processing Systems 14, 2002.
8. E. Spyromitros , G. Tsoumakas , I. Vlahavas ,An empirical study of lazy multi-label classification algorithms, in: Proceedings of the Hellenic conference on Artificial Intelligence, Springer, 2008, pp. 401–406
9. M. Zhang , Z. Zhou , ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048 .
10. J. Jiang , S. Tsai , S. Lee , FSKNN: multi-label text categorization based on fuzzy similarity and k nearest neighbors, Expert Syst. Appl. 39 (3) (2012) 2813–2821 .
11. W. Cheng , E. Hüllermeier , Combining instance-based learning and logistic regression for multi-label classification, Mach. Learn. 76 (2–3) (2009) 211–225 .
12. A. Radzikowska , E. Kerre , A comparative study of fuzzy rough sets, Fuzzy Sets Syst. 126 (2) (2002) 137–155 .
13. D. Dubois , H. Prade , Rough fuzzy sets and fuzzy rough sets, Int. J. General Syst. 17 (2–3) (1990) 191–209 .
14. L. Zadeh , Fuzzy sets, Inf. Control 8 (3) (1965) 338–353 .
15. Z. Pawlak , Rough sets, Int. J. Parallel Program 11 (5) (1982) 341–356 .
16. S. Godbole and S. Sarawagi. Discriminative Methods for Multi-labeled Classification. In Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004), pages 22–30, 2004.
17. Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. O. Maimon, L. Rokach (Ed.), Springer, 2nd edition, 2010.