ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 1, January 2016

Generalized k-labelset Ensemble for multi label classification and cost-sensitive classification: Review

Vaishali Bansode[1], Prof. Dr. S. S. Sane.[2]

PG Student, Dept. Of Computer Engg. K. K. Wagh College of Engineering, Nashik, Maharashtra, India[1]
Head of Department, Computer Engg, K. K. Wagh College of Engineering, Nashik, Maharashtra, India[2]

ABSTRACT- In multi-label classification, set of labels are associated with each example. An algorithm called Random k-labelsets (RAkEL) is an algorithm for multi-label classification that follows problem transformation approach and uses Label powerset (LP) classifier. RAkEL assumes equal weightage for each label set. To overcome this drawback, a new approach as reported in the literature that is Generalized k-labelset ensemble (GLE) advocates the basis expansion model to train LP classifier on random k label set. To reduce the global error between the estimate and ground truth, the expansion coefficients are learned. This model is further extended to solve the multi label misclassification problems. It is capable of handling noisy data sets such as social tagging by treating tag count as misclassification cost. This paper gives the review of GLE method.

KEYWORDS: multi-label classification, RAkEL, GLE, cost-sensitive

I. INTRODUCTION

Classification is a method in which a problem associates a single label or set of labels to each example or instance. There are many classification tasks where each instance can be associated with one or more labels. When a single label is assigned to an instance then this type of classification is single label classification, whereas when multiple labels are simultaneously assigned to an instance then this type of classification is multi label classification. Label powerset (LP) [1], problem transformation method, considers every different set of labels in the training dataset as a new class. The limitation of LP method is that the number of classes increases as the number of labels in the labelset increases, where each class may be associated with very less training data. To overcome this limitation, a new method Random k-Labelsets (RAkEL) [2] is proposed, where k is the parameter which specifies the size of the labelsets. RAkEL method randomly breaks the original labelset into different k-sized subsets and then applies LP method to train each subset. For the final prediction of RAkEL method voting of the LP classifier is done in ensemble. This method reduces the number of classes, as well as, allows each class to have more training instances.

The limitation of RAkEL method is that, it gives equal importance to the every base classifier in the ensemble, which is problematic as every LP classifiers are trained on different randomly selected k-labelsets where some may give worst performance than others or can be even redundant.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 1, January 2016

II. RELATED WORK

Multi labels classification is widely used in data mining. Predicting the appropriate label to the given instance is very importance for the effective performance. GLE for multi label classification helps to minimize the global error between the actual truth and the prediction, for random k-labelset algorithm and GLE for cost sensitive MLC also helps to minimize the misclassification cost.

Multilabel classification (MLC) and label ranking (LR) are two major tasks in supervised learning from multi-label data. MLC model outputs a bipartition of the set of labels into irrelevant and relevant with respect to input query instance. LR learns a model which outputs an ordering of the class labels with respect to their relevance to a query instance. The LR and MLR methods [1] are categorised as: i) problem transformation approach, that transforms the multi label classification into single label classification; and ii) algorithm adaptation approach, adapts or extends the existing algorithms.

Supervised and un-supervised data can be represented using the high order relations. For the representation of these high order relations hypergraphs and tensors are proposed. Hypergraphs are a generalization of graphs in which the edges are arbitrary non-empty subsets of the vertex set [3]. Hypergraphs have edges that connect sets of two or more vertices rather than having edges between pairs of vertices. S. Agarwal, K. Branson, S. Belongie have discussed different hypergraph learning algorithms such as Clique Expansion, Star Expansion, etc.

Hung-Yi Lo, Shou-De Lin, Hsin-Min Wang proposed a new Generalized k-labelset ensemble [4] method which learns makes use of expansion model for multi label classification. The expansion coefficients are learned which reduces the global error between the ground truth and prediction. The output of the experiment conducted by author shows that the performance of LP-based ensemble method significantly improved by assigning different weights to the classifiers in the ensemble.

In cost-sensitive classification, for each instance a misclassification cost is coupled with each label associated with the label of that instance. The main aim of the cost-sensitive multi label classification is to train the classifier which may help to minimize the misclassification cost of a new instance.

GLE method is also extended for cost-sensitive multi-label classification [8] and uses social tagging by considering tag counts information as the misclassification costs. Social tags are the text labels which may be assigned by the domain experts or the users. These social tags may consist of errors or noise. Tag count is the number of users, who have assigned tag to the given resource. This tag count shows the confidence degree associated with the tag. The researchers have observed that, if the same tag is annotated by many users to the visually same images, then these tags may represent the semantic concept of the image. The information of tag count is used for training the cost sensitive classifier which may help to minimize the training error related with tag counts.

A web-based game, MajorMiner, is proposed in [8], which collects the music tags. The participants have given ten seconds audio clip, they have to describe the audio clip with the relevant words, and score the points when their

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 1, January 2016

relevant words matches with relevant words of other participants. These tags are useful to train automatic music description algorithms, compared to social tags from a popular website.

III. GENERALIZRD K-LABELSET ENSEMBLE METHOD

Figure 1 shows the basic block diagram for the GLE method:

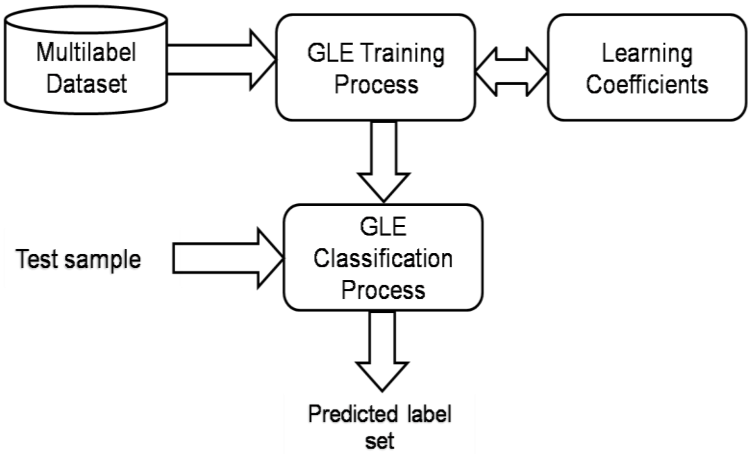

Figure 1. GLE method block diagram

GLE Training Process

For training process the input will be number of models M, size of labelset k, set of labels L, and training set D. Initially the L^k subsets will be assigned to the S. Then for each model m, the k-labelset will be selected randomly from S and assigned to the R_m. The LP classifier gm will be trained based on training set D and R_m. Transformed prediction of LP classifier g_m will be calculated using following formula, where g_m is the LP classifier, x is the instance,

$$q(g_m(x_i), j) = \begin{cases} 1, & \text{if } j \in \mathcal{R}_m \\ & \text{and } j \text{ is positive in } g_m(x_i), \\ -1, & \text{if } j \in \mathcal{R}_m \\ & \text{and } j \text{ is negative in } g_m(x_i), \\ 0, & \text{if } j \notin \mathcal{R}_m. \end{cases}$$

After this the R_m will be removed from S. Finally the coefficients are learned.

Output of training process is ensemble of LP classifier g_m, corresponding k-labelsets R_m, coefficients β_m.

GLE Classification Process

For classification process the input is the number of models M, a test sample x, an ensemble of LP classifier g_m,, corresponding k-labelsets R_m, coefficients β_m. For each label in the label space, its value is initialized to 0. Then for each LP classifier, if label belongs to corresponding labelset R_m, then the value for that label is calculated as below, consider r_j is the j^{th} label. Output of this process is multi-label classification vector r = (r_1, r_2, .., r_3).

Learning Coefficients

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 1, January 2016

1. GLE for multi-label classification

The first term in the objective function aims to minimize the global error between the prediction of LP classifier and the multi-label ground truth Y. The second term is a two-norm regularization term of the coefficients B. The third term is a hypergraph regularization term.

$$\min_{\beta} \frac{1}{2}||Y - \sum_{m=1}^{M} \beta_m Q_m||_F^2 + \frac{\gamma}{2}||\beta||_2^2 + \frac{\nu}{2}\mathrm{trace}\left(\left(\sum_{m=1}^{M} \beta_m Q_m\right)^T L\left(\sum_{m=1}^{M} \beta_m Q_m\right)\right)$$

2. GLE for cost-sensitive multi-label classification

The objective function is same as GLE for multi label classification, only one difference is there. The first term is modified to a cost-weighted global error by multiplying the global error with the multi-label misclassification cost matrix C.

$$\min_{\beta} \frac{1}{2}||C \circ \left(Y - \sum_{m=1}^{M} \beta_m Q_m\right)||_F^2 + \frac{\gamma}{2}||\beta||_2^2 + \frac{\nu}{2}\mathrm{trace}\left(\left(\sum_{m=1}^{M} \beta_m Q_m\right)^T L\left(\sum_{m=1}^{M} \beta_m Q_m\right)\right)$$

IV. CONCLUSION

The Generalized k-labelset ensemble method which classification has been reported in literature improves the performance of multi label classification compared with Label powerset and Random k-labelset methods. The Generalized k-labelset ensemble method for multi label classification and cost-sensitive classification minimizes the global error and misclassification cost, respectively.

REFERENCES

[1] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multilabel data," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. New York, NY, USA: Springer, 2010.

[2] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," IEEE Trans. Knowl. Data Eng., vol. 23,no. 7, pp. 1079–1089, Jul. 2011.

[3] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in Proc. Int. Conf. Mach. Learn., Pittsburgh, PA, USA, 2006.

[4] H.-Y. Lo, S.-D. Lin, and H.-M. Wang, "Generalized k-labelset ensemble for multi-label classification," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2012.

[5] M. I. Mandel and D. P. W. Ellis, "Multiple-instance learning for music information retrieval," in Proc. Int. Conf. Music Inform. Retrieval, 2007.

[6] Prati, R.C., "Fuzzy rule classifiers for multi-label classification," in *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on* , vol., no., pp.1-8, 2-5 Aug. 2015.

[7] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," J. New Music Res., vol. 37, no. 2, pp. 151–165, 2008.

[8] Hung-Yi Lo; Shou-De Lin; Hsin-Min Wang, "Generalized k-Labelsets Ensemble for Multi-Label and Cost-Sensitive Classification," in *Knowledge and Data Engineering, IEEE Transactions on* , vol.26, no.7, pp.1679-1691, July 2014