ISSN 2278-7763

On Machine Learning Techniques For Multi-class Classification

G. Malik[1], M. Tarique [2]

[1]Department of Mathematics, RLA(Eve) College, University of Delhi, Delhi, INDIA; [2]Department of Mathematics, Dyal Singh College, University of Delhi, Delhi, INDIA .
Email: gufranmalik@gmail.com[1], tarique.1984@gmail.com[2]

ABSTRACT

Most of the practical applications involve multi-class classification, especially in remote sensing land cover classification, protein function classification, music categorization and semantic scene classification. In this paper we introduces multi-label classification, organization of sparse vectors and related literature and performs comparative experimental results of various multi-class classification methods. This paper also discusses the quantification concepts of the multi-class nature of data set.
Keywords : *Classification, Conversion, Machine Learning, Neural, Networks*

1 INTRODUCTION

In single label classification each data point are associated with a single label l. When the data set has two label then the problem is called a binary classification problem, while if the dataset has more than two classes , then it is called multi-class classification problem.
In multi-class classification, given a set of labeled examples with labels selected from a finite set, an inductive procedure builds a function that (hopefully) is able to map unseen instances to their appropriate classes. In the past, multi- label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Text documents usually belong to more than one conceptual class. For example, a newspaper article concerning the reactions of the Christian church to the release of the Da Vinci Code film can be classified into both of the categories Society\Religion and Arts\Movies. Similarly in medical diagnosis, a patient may be suffering for example from diabetes and prostate cancer at the same time.

This paper aims to serve as a starting point and reference for researchers interested in multi-label classification. The main contributions are
1.A structured presentation of the sparse literature on multi-label classification methods with comments on their relative strengths and weaknesses and when possible the abstraction of specific methods to more general and thus more useful schemata,
2.The introduction of an undocumented multi-label method, c) the definition of a concept for the quantification of the multi-label nature of a data set, d) preliminary comparative experimental results about the performance of certain multi-label methods.
The next section discusses tasks that are related to multi- label classification and then introduces the concept of label density and presents the metrics that have been proposed in the past for the evaluation of multi-label classifiers.

2 RELATED WORK

Several well-known methods for binary classification, including neural networks (Rumelhart et al., 1986), decision trees (Quinlan, 1993), k-NN (see for example (Mitchell, 1997)), can be naturally extended to the multi-class domain.
One of the method that can be used for multi-label classification is called ranking. In ranking the task is to order a set of labels, so that the topmost labels are more related with the new instance. But the ranking method requires post processing of data to give a set of labels.
Jin and Ghahramani (2002) call multiple-label problems, the semi-supervised classification problems where each example is associated with more than one classes, but only one of those classes is the true class of the example. This task is not that common in real-world applications as the one we are studying.
In certain classification problems the labels belong to a hierarchical structure. When the labels in a data set belong to a hierarchical structure then we call the task hierarchical classification. If each example is labeled with more than one node of the hierarchical structure, then the task is called hierarchical multi-label classification. In this paper we focus on flat (non-hierarchical) multi-label classification methods.

Multiple-instance learning is a variation of supervised learning, where the task is to learn a concept given positive and negative bags of instances (Maron & Lozano-Perez, 1997). Each bag may contain many instances, but a bag is labeled positive even if only one of the instances in it falls within the concept. A bag is labeled negative only if all the instances in it are negative.

ISSN 2278-7763

3 MULTI -CLASS CLASSIFICATION METHODS

Basically we can divide multi-class classifications into two main categories,

(I) Conversion methods, and

(II) Algorithmic adaption methods.

In conversion method, we convert multi-class classification problem either into one or more single-class classification problem.

3.1CONVERSION METHOD

One of the conversion method is called one vs all (OvA) (or one-vs.-rest OvR) strategy, where a single classifier is trained per class to distinguish that class from all other classes. Prediction is then performed by predicting using each binary classifier, and choosing the prediction with the highest confidence score.

In pseudo-code, the training algorithm for an OvA learner constructed from a binary classification learner L is as follows:

Inputs:

1 L, a learner (training algorithm for binary classifiers)

2 samples X

3 labels y where $y \in \{1, \cdots K\}$ is the label for the sample X

Output:

a list of classifiers f_k for $k \in \{1, \cdots K\}$

Procedure:

step 1 For each k in $\{1 \ldots K\}$:

step 2 Construct a new label vector $y' = 1$ where $y = k$, 0 (or -1) elsewhere

step 3 Apply L to X, y' to obtain f_k

Making decisions proceeds by applying all classifiers to an unseen sample x and predicting the label k for which the corresponding classifier reports the highest confidence score.

3.2Algorithm Adaptation Methods

Clare and King (2001) adapted the C4.5 algorithm for multi-label data. They modified the formula of entropy calculation as follows:

$$entropy(S) = -\sum_{i=1(1)N} (p(c_i) \log p(c_i) + q(c_i) \log q(c_i))$$

where $p(c_i)$ = relative frequency of class c_i and $q(c_i) = 1-p(c_i)$. They also allowed multiple labels in the leaves of the tree.

Adaboost.MH and Adaboost.MR (Schapire & Singer, 2000) are two extensions of AdaBoost (Freund & Schapire, 1997) for multi-label classification. They both apply AdaBoost on weak classifiers of the form H:X ×L → R. In AdaBoost.MH if the sign of the output of the weak classifiers is positive for a new example x and a label l then we consider that this example can be labeled with l, while if it's negative then this example is not labeled with l. In AdaBoost. MR the output of the weak classifiers is considered for ranking each of the labels in L.

Although these two algorithms are adaptations of a specific learning approach, we notice that at their core, they actually use a Conversion method.

ML-kNN (Zhang & Zhou, 2005) is an adaptation of the kNN lazy learning algorithm for multi-label data. In essence, ML-kNN uses the kNN algorithm independently for each label l: It finds the k nearest examples to the test instance and considers those that are labeled at least with l as positive and the rest as negative. What mainly differentiates this method from the application of the original kNN algorithm to the transformed problem. ML-kNN has also the capability of producing a ranking of the labels as an output.

ISSN 2278-7763

Luo and Zincir-Heywood (2005) present two systems for multi-label document classification, which are also based on the kNN classifier. The main contribution of their work is on the preprocessing stage for the effective representation of documents. For the classification of a new instance, the systems initially find the k nearest examples. Then for every appearance of each label in each of these examples, they increase a corresponding counter for that label. Finally they output the N labels with the largest counts. N is chosen based on the number of labels of the instance. This is an inappropriate strategy for real-world use, where the number of labels of a new instance is unknown.

McCallum (1999) defines a probabilistic generative model according to which, each label generates different words. Based on this model a multi-label document is produced by a mixture of the word distributions of its labels. The parameters of the model are learned by maximum a posteriori estimation from labeled training documents, using Expectation Maximization to calculate which labels were both the mixture weights and the word distributions for each label. Given a new document the label set that is most likely is selected with Bayes rule.

Elisseeff and Weston (2002) present a ranking algorithm for multi-label classification. Their algorithm follows the philosophy of SVMs: it is a linear model that tries to minimize a cost function while maintaining a large margin. The cost function they use is ranking loss, which is defined as the average fraction of pairs of labels that are ordered incorrectly. However, as stated earlier, the disadvantage of a ranking algorithm is that it does not output a set of labels.

Godbole and Sarawagi (2004) present two improvements for the Support Vector Machine (SVM) classifier in conjunction with the conversion method for multi-label classification. The first improvement could easily be abstracted in order to be used with any classification algorithm and could thus be considered an extension to conversation. The main idea is to extend the original data set with |L| extra features containing the predictions of each binary classifier. Then a second round of training |L| new binary classifiers takes place, this time using the extended data sets. For the classification of a new example, the binary classifiers of the first round are initially used and their output is appended to the features of the example to form a meta-example. This meta-example is then classified by the binary classifiers of the second round.

The second improvement of (Godbole & Sarawagi, 2004) is SVM specific and concerns the margin of SVMs in multi-label classification problems. They improve the margin by a) removing very similar negative training instances which are within a threshold distance from the learnt hyperplane, and b) removing negative training instances of a complete class if it is very similar to the positive class, based on a confusion matrix that is estimated using any fast and moderately accurate classifier on a held out validation set. Note here that the second approach for margin improvement is actually SVM independent.

4 EXPERIMENTAL RESULTS

We experimented on the following multi-label data sets: genbase (Diplaris et al., 2005) and yeast (Elisseeff & Weston, 2002) are biological data sets that are concerned with protein function classification and gene function classification respectively. The scene data set (Boutell et al., 2004) contains data related to a scene classification problem. These data sets were retrieved from the site of the Support Vector Classification library LIBSVM (Chang & Lin, 2001), and transformed to a specific format that is suitable for our software, based on the ARFF file format of the WEKA library. The transformed data sets are also available at the aforementioned URL.
The details of the data sets, such as the number of examples, the number of numeric and discrete attributes the number of classes and their label density are given in table We notice that genbase (LD=0.05) and scene (LD=0.18) are quite sparse multi-label data sets with less than 1.5 labels per example on average. The yeast dataset on the other hand is denser (LD=0.30) with more than 4 labels per example on average.

Table : Examples, numeric and discrete attributes, labels and LD of datasets

Data Set	Examples		Attributes		Labels	Label	Label
	Train	Test	Numeric	Discrete		Density	Cardinality

ISSN 2278-7763

gen-base	463	199	0	1185	27	0.05	1.35
yeast	1500	917	103	0	14	0.30	4.25
scene	1211	1196	294	0	6	0.18	1.08

5 CONCLUSION

This paper discussed multi-label classificstion methods and also introduced the problems. In this paper we gave some well known methods that exist in the literature and given comparative study results for some of these methods. It will be useful to the researchers those want to start work on multi-label classification. We also intend to perform a comparative experimental study of support vector machine and man other methods with different data sets.

REFERENCES

[1] Aha, D.W., Kibler, D., & Albert, M.K. (1991), 'Instance-based learning algorithms', Machine Learning, vol. 6, no. 1, pp. 37-66 .
[2] Boutell, M.R., Luo, J., Shen, X. & Brown, C.M. (2004), 'Learning multi-label scene classification', Pattern Recognition, vol. 37, no. 9, pp. 1757-71.
[3] Chang, C.-C., & Lin, C.-J. (2004), 'LIBSVM : a library for support vector machines', Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
[4] Clare, A. & King, R.D. (2001), 'Knowledge Discovery in Multi-Label Phenotype Data', paper presented to Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001), Freiburg, Germany.
[5] Diplaris, S., Tsoumakas, G., Mitkas, P. & Vlahavas, I. (2005), Jin, R. & Ghahramani, Z. (2002), 'Learning with Multiple Labels',
[6] paper presented to Proceedings of Neural Information Processing Systems 2002 (NIPS 2002), Vancoucer, Canada.
[7] John, G. & Langley, P. (1995), 'Estimating continuous distributions in bayesian classifiers', paper presented to Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Vancoucer, Canada
[8] Lauser, B. & Hotho, A. (2003), 'Automatic multi-label subject indexing in a multilingual environment', paper presented to Proceedings of the 7th European Conference in Research and Advanced Technology for Digital Libraries (ECDL 2003).
[9] Lewis, D.D., Tony, Y. Y., Rose, G. & Li, F. (2004). 'RCV1: A new benchmark collection for text categorization research', Journal of Machine Learning Research, Vol 5, pp 361-397.
[10] Li, T. & Ogihara, M. (2003), 'Detecting emotion in music', paper presented to Proceedings of the International Symposium on Music Information Retrieval, Washington D.C., USA.
[11] Luo, X. & Zincir-Heywood, A.N. (2005), 'Evaluation of Two Systems on Multi-class Multi-label Document Classification', paper presented to Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems.
[12] Maron, O. & Lozano-Perez, T. (1997), 'A framework for Multiple-Instance learning', paper presented to Proceedings of Neural Information Processing Systems 1997 (NIPS 1997).
[13] McCallum, A. (1999), 'Multi-label text classification with a mixture model trained by EM', paper presented to Proceedings of the AAAI' 99 Workshop on Text Learning.
[14] Platt, J. (1998), 'Fast training of support vector machines using sequential minimal optimization', In B. Scholkopf, B., Burges, C., & Smola, A., Advances in Kernel Methods - Support Vector Learning, MIT Press.
[15] Quinlan, J.R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann.
[16] Schapire, R.E. & Singer, Y. (2000), 'Boostexter: a boosting-based system for text categorization', Machine Learning, vol. 39, no. 2/3, pp. 135-68.
[17] Thabtah, F.A., Cowling, P. & Peng, Y. (2004), 'MMAC: A New Multi-class, Multi-label Associative Classification Approach', paper presented to Proceedings of the 4th IEEE International Conference on Data Mining, ICDM '04.
[18] Witten, I.H. & Frank, E. (1999), 'Data Mining: Practical machine learning tools with Java implementations', Morgan Kaufmann.
[19] Classification', paper presented to Proceedings of the 1st IEEE International Conference on Granular Computing.

