Multilabel Classification Evaluation using Ontology Information

Stefanie Nowak, Hanna Lukashevich

Fraunhofer Institute for Digital Media Technology IDMT
Ehrenbergstrasse 31, 98693 Ilmenau, Germany,
{nwk,lkh}@idmt.fraunhofer.de *

Abstract. Multilabel classification using ontology information is an emerging research area that combines machine learning methods with knowledge models. The performance assessment of such classification systems poses new challenges. We propose an evaluation measure that considers the mapping of label sets to their groundtruth and allows for the incorporation of real world knowledge. A distance-based measure from the area of hierarchical unilabel classification evaluation is extended to the case of multilabel classification and enriched with additional ontology information. The evaluation measure considers structure information, relationships and the agreement between annotators.

1 Introduction

Traditional classification approaches classify multimedia documents in one of several categories. Applied to automatic image annotation that means that a photo depicts e.g., either a `landscape`, a `city` or `persons`. In reality, it is difficult to judge to which category a multimedia document exclusively belongs to. Mostly the documents contain aspects of different categories and should be labelled with all relevant items. This task is performed in multilabel classification.

Figure 1 shows a simple hierarchical organization of concepts that can be assigned to photos in a multilabel annotation scenario. The hierarchy allows to make assumptions about the assignment of concepts to documents. E.g., if a photo is classified to contain `trees`, it also contains `plants`. Further semantic knowledge is provided if the concepts are organized in an ontology. Then, next to the *is-a* relationship of the hierarchical organization of concepts, additionally other relationships between concepts determine possible label assignments. The ontology restricts e.g., that for a certain sub-node only one concept can be assigned at a time (disjoint items) or that a special concept (like `portrait`) postulates other concepts like `persons` or `animals`.

This paper discusses several approaches to multilabel classification evaluation and highlights the information that is used to assess the quality. It proposes a hierarchical multiannotation evaluation measure that incorporates ontology knowledge and can be used in a benchmarking scenario.

* This work has been supported by grant No. 01MQ07017 of the German THESEUS program, founded by the Federal Ministry of Economics and Technology.

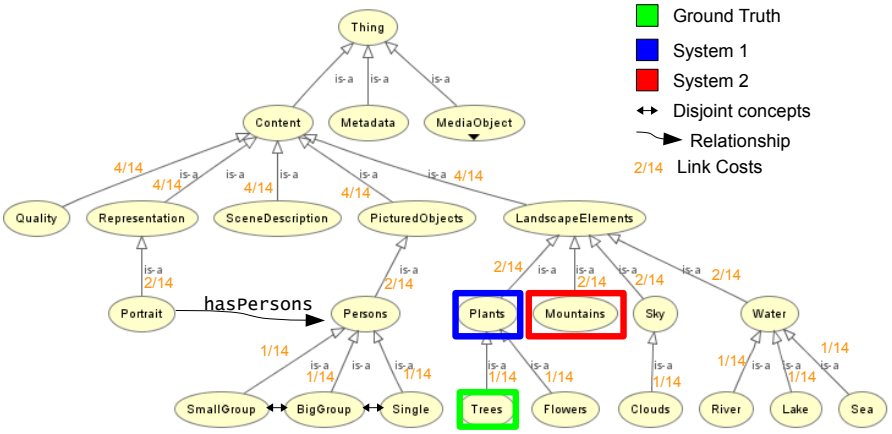

Fig. 1. Visualization of an ontology fragment for image annotation. The concepts are hierarchical structured and different types of relationships are exemplarily highlighted.

2 Related work

Many researchers regard the results of a multilabel classification system similar to the unilabel classification approach. The prediction is evaluated for each concept in isolation. This allows to use the well-known evaluation measures like precision, recall, f-measure or accuracy. E.g., Fan et al. use the accuracy in [1] to determine the quality of the classifier for each concept. The opposite way is to start with the multimedia document and evaluate if all concepts are assigned correctly. Then instead of comparing a single predicted label to a single ground-truth label, one needs to compare two sets of labels. As a result, the predicted labels can be *fully correct* (label sets are identical), *fully wrong* (the intersection of the sets is empty), or *partly correct* (the sets have common labels, but are not fully identical). The Accuracy that is often used as measure in traditional classification evaluation cannot be used to judge partial matches, because it only regards one instance as correctly classified if all associated labels were correctly predicted. Partial matches are e.g., considered by utilizing the macro-average and micro-average f-measures, proposed by Tague in [2]. Shen et al. [3] proposed an α-evaluation and multilabel class recall and precision. α-evaluation generates a score while taking the ground-truth, predicted labels, missed labels and false positive labels into account. Moreover, false positives and missed labels can be penalized differently as it is more suitable for the particular application. The parameter α introduces the so-called *forgiveness rate* as a trade-off between the fully correct and partly correct prediction. Sun et al. propose in [4] semantic-based misclassification costs. Each class of documents is represented by a feature vector of all documents belonging to a certain category. The cosine distance between feature vectors of two categories is used as similarity measure and defines the *average category similarity*.

A hierarchical organization of concepts enables a hierarchical evaluation. Different hierarchical measures for unilabel classification are summarized in [5]. Intuitively the concepts, that are located near in a hierarchy are more similar than the ones that are located far. The idea is to judge an annotation from the predictor that does not match exactly to the groundtruth by their distance in the hierarchy. The most important measures are the *depth independent distance-based misclassification costs* and the *depth dependent distance-based misclassification costs*. In the former case, the predicted concept is compared to the correct one and the number of edges of the shortest path in the hierarchy between both are counted. In the latter case, an additional weight is assigned to each edge in the hierarchy. So, misclassifications in deeper levels of the hierarchy have lower costs than at an upper level. An evaluation measure for hierarchical multiclassification evaluation, is proposed by Blockeel et al. [6]. They extrapolate distances between individual labels to distances between sets of labels by mapping the feature vectors of the sets into Euclidean space where the individual labels form the base vectors. In [7] a hierarchical loss function is proposed that considers classification into a hierarchy with multiple and partial paths. The first wrongly classified node is regarded as mistake and adds to the loss while sub-nodes after a mistake are not considered. Underlying is the assumption that for each classification a path from root to leaf or from root to an internal node is present. They compare their work to the zero-one loss and symmetric-difference loss.

3 Evaluation Measure

3.1 Requirements on the Evaluation Measure

The proposed evaluation measure is utilized in a benchmark for multilabel annotation of photos.[1] The participant's task is to annotate photos with 53 concepts in a multilabel annotation scenario. A small ontology of the concepts is provided that may be used for training the classifiers. To objectively compare the approaches of the participants, different conditions have to be assured.

First, the evaluation measure should consider partial matches and deliver an annotation score for each image. Second, it has to be assured that a system that annotates `plants` instead of `trees` is judged better than a system that annotates `mountains` (see Fig. 1). This issue is considered in the depth dependent distance-based misclassification costs, introduced in Sec. 2 for the unilabel annotation task. In a multilabel annotation scenario, the challenge is how to map the predicted label set to the groundtruth set. Third, the groundtruthing of the concepts depicted in an image, is no easy task, also not for humans. Some concept assignments are not objective and last in long discussions between the annotators. Therefore a study about the agreement on concepts between different annotators was conducted. Altogether 10 annotators annotated the same 100 images on their own. The degree of agreement for each concept over all photos was computed and should serve as a probability if the groundtruthing

[1] http://www.imageclef.org/2009/PhotoAnnotation

was correct. E.g., for the concept `clouds` in 96% the annotators agreed on annotating the concept. In case of `Aesthetic Image` the annotators agreed only at 75%. These empirically obtained values are used as weighting factors for the calculated costs in case of misclassification. So, the more subjective concepts are weighted less than the more objective ones. Fourth, the relationships defined in the ontology should be kept. There should be an option to penalize a system, if it simultaneously annotates concepts that are defined as disjoint or ignores preconditions for relationships.

None of the described evaluation measures in Sec. 2 fulfils these requirements. E.g., the hierarchical loss measure assumes that every node of the hierarchy possibly has annotation instances and that a continuous path exists through the hierarchy. In the ontology for image annotation, *abstract nodes* are defined that represent real-world knowledge but no visual concept. An example is the node `representation`. A `portrait` is a visual subconcept of `representation` and can be annotated by a classification system, but the concept `representation` itself is no visual concept.

3.2 Evaluation Measure for Multiannotation Scenarios

Let us consider the predicted set of labels as P and the groundtruth set of labels as G. Each set contains labels l_i respectively l_j that are assigned to a multimedia document X. First, the false positive labels $P' = P \setminus (P \cap G)$ and the missed labels $G' = G \setminus (P \cap G)$ are computed. Please note that $|P'| + |G'| \leq |P \cup G|$ is always valid, because the number of false positive and missed labels can never be greater than the number of the union of labels in both sets. Next, for each label l_i from P' a match to a label l_j from G is calculated and for each label l_j from G' a mapping to a label l_i from P is performed in an optimization procedure (see Eq. 1). The costs between two labels l_i and l_j depend on the shortest path in the hierarchy between both concepts. Each link is associated with a cost that is cut in halves for each deeper level of the tree and is maximal equal to 1 for a path between two leaf nodes of the deepest level (see Fig. 1). The costs c_i for a link are calculated as $c_i = \frac{2^{(i-1)}}{2 \cdot \sum_{i=1}^{N} 2^{(i-1)}}$ with N as the number of links from the deepest node to the root. If $P = \emptyset$, the matching costs for all labels l_j of $G' = G$ are set to the maximum. The matching costs are computed as follows:

$$match(P, G) = \sum_{l_i \in P'} \left((\min_{l_j \in G} cost(l_i, l_j)) \cdot a(l_j{}^*) \right) + \sum_{l_j \in G'} \left((\min_{l_i \in P} cost(l_i, l_j)) \cdot a(l_j) \right) \quad (1)$$

with $l_j{}^* = \text{argmin}_{l_j \in G}(cost(l_i, l_j))$.
$a(l_j)$ determines the *annotation agreement factor* for a concept l_j and ranges between $[0, 1]$. It is empirically determined through the agreement of different annotators on a concept, as described in 3.1. Optionally, a crosscheck on the predicted label set P is performed during computing $match(P, G)$. If labels in P violate relationships from the ontology, these labels get the maximum costs of 1 as penalty assigned instead of calculating the minimal costs to a label of

G. Referring to the example in Fig. 1, a penalty of 1 is assigned if the system simultaneously annotates *single* and *small group* or if *portrait* is annotated without annotating one of the person concepts. Assuming that *single* is correct in the first example, the costs between *small group* from P' and *single* from G are equal 1 instead of only equal $2/14 \cdot a(single)$ because of the hierarchy distances.

$$score(X) = \left(1 - \frac{match(P, G)}{|P \cup G|}\right)^{\alpha} \tag{2}$$

The final score for each multimedia document X is based on the matching costs between P and G divided by the number of different concepts in both label sets (see Eq. 2). The score is 1 if all concepts are correctly annotated and goes to 0 if no concept was found. Additionally, Shens α-factor, ($\alpha \geq 0$), introduced in Sec. 2, is incorporated to weight the strictness of the score regarding fully and partly correct annotations, depending on the application demands.

4 Conclusion and Future Work

In this paper, we propose an evaluation measure for the performance assessment of multiannotation classification systems incorporating ontology knowledge. A distance-based misclassification cost was extended from the unilabel to the multilabel case and further enriched with ontology information like its hierarchy, an annotation agreement factor and penalties for ignoring relationships. Next, an extensive evaluation of the behaviour of the measure in a real benchmarking scenario will be conducted. In future work, we would like to investigate how relationships in ontologies can be incorporated more differentiated in dependence from the evaluation scenario. Another point is how to base the evaluation measure on semantic similarity of concepts instead of distances in hierarchies, as the structure of the ontology is rather subjective and may change during time.

References

1. Fan, J., Gao, Y., Luo, H., Jain, R.: Mining multilevel image semantics via hierarchical classification. IEEE Trans. on Multimedia **10**(2) (2008) 167
2. Tague, J.M.: Information retrieval experiment. In Jones, K.S., ed.: The pragmatics of information retrieval experimentation, Butterworths, London. (1981) 59–102
3. Shen, X., Boutell, M., Luo, J., Brown, C.: Multi-label machine learning and its application to semantic scene classification. In: Intern. Symp. on Elec. Imag. (2004)
4. Sun, A., Lim, E.: Hierarchical text classification and evaluation. In: Proc. of the IEEE Intern. Conf. on Data Mining. Volume 528., California, USA (2001)
5. Freitas, A., de Carvalho, A.: A tutorial on hierarchical classification with applications in bioinformatics. Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications (2007) 114–140
6. Blockeel, H., Bruynooghe, M., Džeroski, S., Ramon, J., Struyf, J.: Hierarchical multi-classification. In: Proc. of Workshop on Multi-Relational Data Mining. (2002)
7. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Incremental algorithms for hierarchical classification. The Journal of Machine Learning Research **7** (2006) 31–54