Applying Machine Learning Diversity Metrics to Data Fusion in Information Retrieval

David Leonard, David Lillis, Lusheng Zhang,
Fergus Toolan, Rem W. Collier, and John Dunnion

School of Computer Science and Informatics
University College Dublin, Ireland
{david.leonard, david.lillis, lu-sheng.zhang,
fergus.toolan, rem.collier, john.dunnion}@ucd.ie

Abstract. The Supervised Machine Learning task of classification has parallels with Information Retrieval (IR): in each case, items (documents in the case of IR) are required to be categorised into discrete classes (relevant or non-relevant). Thus a parallel can also be drawn between classifier ensembles, where evidence from multiple classifiers are combined to achieve a superior result, and the IR data fusion task.

This paper presents preliminary experimental results on the applicability of classifier ensemble diversity metrics in data fusion. Initial results indicate a relationship between the quality of the fused result set (as measured by MAP) and the diversity of its inputs.

1 Introduction

Data fusion is a technique for combining the ranked lists of documents returned by multiple Information Retrieval (IR) systems in an attempt to improve performance. One rationale for the success of data fusion is similarity between the relevant documents and diversity among the non-relevant documents returned by the component systems [1]. Similarly, in the area of Supervised Machine Learning (SML), diversity with respect to the errors committed by component classifiers in ensembles has received much attention. In particular, it has been proposed that the accuracy of and diversity between these systems are necessary and sufficient conditions to improve the performance of the combined system on classification learning tasks [2]. In an attempt to formalise such a relationship, Kuncheva devised 10 metrics to characterise diversity among classifiers [3]. This paper presents initial work in adapting one of these metrics to data fusion in order to explore the question of an accuracy/diversity trade-off in IR.

2 Background

In Machine Learning, classification is the problem of selecting the correct class for a data point from a discrete set of class labels. Multi-classifier systems combine the outputs of an ensemble of classifiers in an attempt to yield more accurate classification performance. A question that arises across all the different incarnations of multi-classifier techniques is whether there are certain characteristics of the individual component classifiers that guarantee improved performance of the combined system.

The primary experimental work in the field of diversity and its relationship to accuracy in multi-classifier systems has been carried out by Kuncheva [3, 4]. In an attempt to quantify diversity, ten metrics were formalised, which operate by studying the relationship between classifier outputs at the so-called "oracle" level (i.e. for each data point a classifier scores 1 if it classifies it correctly and 0 otherwise). These scores are then aggregated across the entire dataset of training points to obtain a measure of diversity. An example of such a metric is the entropy measure, E, which is given by:

$$E = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{(L - \lfloor L/2 \rfloor - 1)} min \left\{ \sum_{i=1}^{L} y_{j,i}, L - \sum_{i=1}^{L} y_{j,i} \right\} \tag{1}$$

where L is the number of classifiers in the ensemble, N is the total number of data points in the training dataset and $y_{j,i}$ is the value (zero or one), that the ith classifier received on the jth data point. E varies between 0 and 1 where 0 indicates no difference and 1 represents the highest possible diversity.

3 Mapping

At the heart of the current work is the mapping of the metrics between the domains of SML and IR. As mentioned above, in the Machine Learning context the metrics operate at the oracle level (i.e. the output of each classifier is either correct or incorrect), classifiers may agree or disagree and be right or wrong in this respect. Analogously, in the IR context a document may be relevant or non-relevant and similarly IR systems may agree or disagree about this. If, in response to a query, an IR system returns a document then this may be considered as evidence that it considers it to be relevant and likewise the absence of a document in a result set is an affirmation that the system views it to be non-relevant.

One key difference between these scenarios is the notion of ordering. As designed, the metrics operate on unordered sets of outputs. However, the ranked lists returned by IR systems impose an ordering relation between the documents. The strategy outlined above takes a global measurement of diversity without taking this into consideration. It can be argued that this is not necessary because the accuracy(as measured by IR evaluation metrics such as MAP), of the individual IR systems have already taken ranking information into account. A second, related, point is the implicit assumption of an arbitrary cut-off point in the ranked lists, beyond which the systems no longer consider documents to be relevant.

4 Experimental Work

At this early investigative stage of the research the emphasis is on a) the ability of the metrics to capture diversity between document result sets and b) whether a relationship between diversity, accuracy and combined performance may be postulated. A number of decisions were made with respect to the parameters and scope of the experiment to investigate this.

Using inputs from the TREC 2004 Web Track, 51 queries were chosen for which there were between 8 and 147 relevant documents. Many of the available queries only had one relevant document associated with them. In this case, all systems would frequently return that relevant document somewhere in the 1000 documents they returned, resulting in no diversity being found between them. Avoiding queries for which there were very few relevant documents available avoids causing bias in the metrics in this way. In order to reduce the impact of differing-length result sets, 35 inputs were chosen that tended to return 1000 documents (the TREC maximum). The performance of these systems (measured by MAP) varied widely. Teams of 5 systems were fused using SlideFuse [5]. The performance (or 'accuracy' in SML terms) was measured by MAP. Diversity was quantified using the entropy measure presented in Section 2.

The entropy measure was calculated on a per query basis. For each of the N judged relevant documents, if a system included it in its result set it was given a value of 1 and otherwise it was given a value of 0 for that system. These intermediate statistics were then averaged across all queries resulting in a single value for the entropy of the fused system. The average MAP score for the inputs and the combined MAP score (of the fused output) were similarly aggregated across the query set.

4.1 Results

The first 250,000 combinations of the 35 candidate systems into ensembles of size 5, were fused using SlideFuse. Statistics pertaining to performance and diversity were gathered, from which the compound metrics were derived. A subsequent plot of these results was not promising, failing to reveal a pattern between the 3 variables at a global level across the entire spectrum of possibilities for accuracy and diversity. The original hypothesis proposed that the accuracy of and diversity between the component systems are necessary and sufficient conditions to improve the performance of the combined system. To investigate this relationship, the results were sorted with respect to the average MAP score and the top 5000 highest performing or most accurate combinations set aside. Correlation between each set of variables was measured using Pearson's correlation coefficient, r. Again, there did not appear to be a clear relationship between either the average and combined MAP, (r = 0.19), or between entropy and average MAP (r = -0.12). There does, however, appear to be a pattern with respect to entropy and the combined MAP score (r = 0.80). A plot of this relationship is shown in fig. 1. With reference to this figure it is clear that the combined MAP scores are higher for systems with a larger, hence more diverse, entropy value.

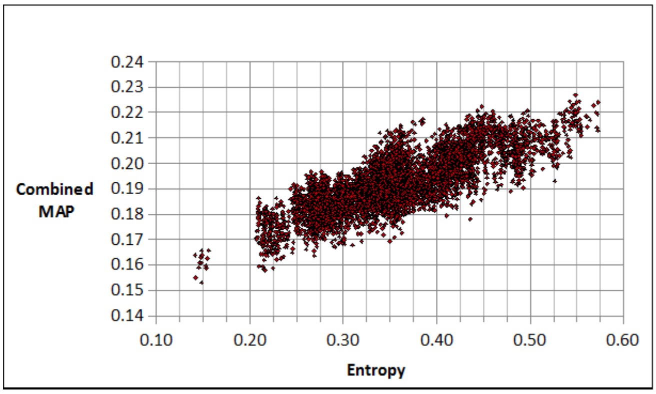

Fig. 1. A plot of Entropy versus Combined MAP for the top 5000 systems

5 Conclusions and Future Work

Based on the preliminary experimental work carried out to date there is evidence of a possible relationship between the combined performance of fused result sets, the average performance of the input result sets and the diversity between the relevant documents returned in the component result sets, as measured using the entropy metric.

In future work, it is proposed to investigate the suitability of metrics other than entropy for capturing diversity between document result sets, the role of diversity between non-relevant or unjudged documents and whether there is a particular cut-off point where the metrics should be applied (e.g. the top 100 rather than 1000 documents). It will also be necessary to ascertain whether such a result generalises to fusion techniques other than Slidefuse and, if so, determine the characteristics of the family of fusion techniques to which the diversity metrics are applicable.

References

1. Lee, J.H.: Analyses of multiple evidence combination. SIGIR Forum **31** (1997) 267–276
2. Dietterich, T.: Ensemble methods in machine learning. In Kittler, J., Roli, F., eds.: Multiple classifier systems, LNCS Vol. 1857 (2000) 1–15
3. Kuncheva, L., Whitaker, C.: Ten measures of diversity in classifier ensembles: limits for two classifiers. In: IEEE Workshop on Intelligent Sensor Processing, Birmingham, UK (2001)
4. Shipp, C., Kuncheva, L.: Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion **3**(2) (2002) 135–148
5. Lillis, D., Toolan, F., Collier, R., Dunnion, J.: Extending Probabilistic Data Fusion Using Sliding Windows. In: Proceedings of the 30th European Conference on Information Retrieval (ECIR '08). Volume 4956 of Lecture Notes in Computer Science., Berlin, Springer (2008) 358–369