Towards Higher Order Complexity Measures for Text Classification

KVS Dileep[1] Sutanu Chakraborti[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Madras, Chennai-36, India

Problem Statement

How to come up with reliable measures of quantifying the complexity of classifying a dataset?

1. Introduction

- Economics vs Geography - easy to classify and IBM Hardware vs Mac Hardware - difficult to classify.
- Capturing the difference in difference between these tasks with a quantitative measure - a non-trivial problem.
- Helps in deciding the right set of features and motivate to search for richer features, while classifying the given dataset.
- Try to come up with a reliable complexity estimator that overcomes the shortcomings of previous measures.

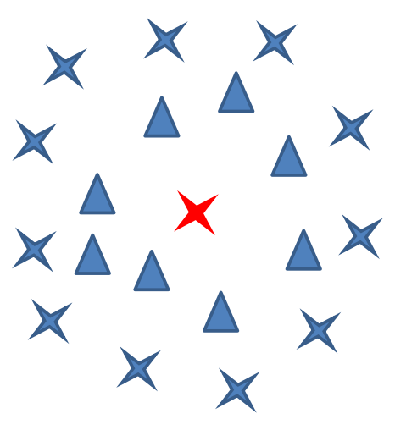

Figure: Example to motivate higher orders. Similar shapes correspond to similar labels

2. Background

- [Chakraborti et al., 2008][Vinay et al., 2006] take the dataset as a whole and measure the clustering tendency between document clusters and label clusters.
- How reliable the neighbors are in predicting complexity?
- Extending the neighborhood by including the neighbors of neighbors would give a more reliable estimate.
- Higher order neighbors refer to neighbors obtained through expansion of neighbors from the query. First order neighbors - immediate neighbors of the query document, second order neighbors - neighbors of the neighbors of query and so on.
- Alignment is a complementary notion of complexity. More the alignment, lesser the complexity.

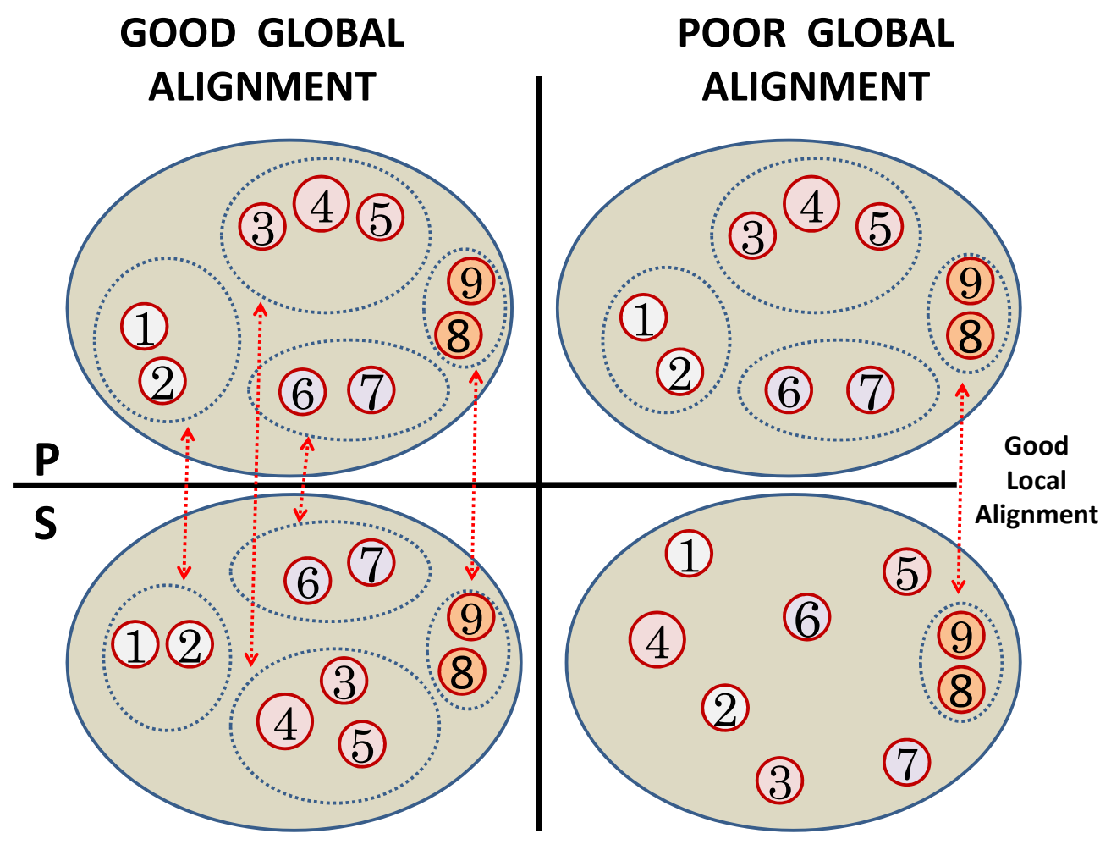

Figure: Illustration to explain complexity in a dataset

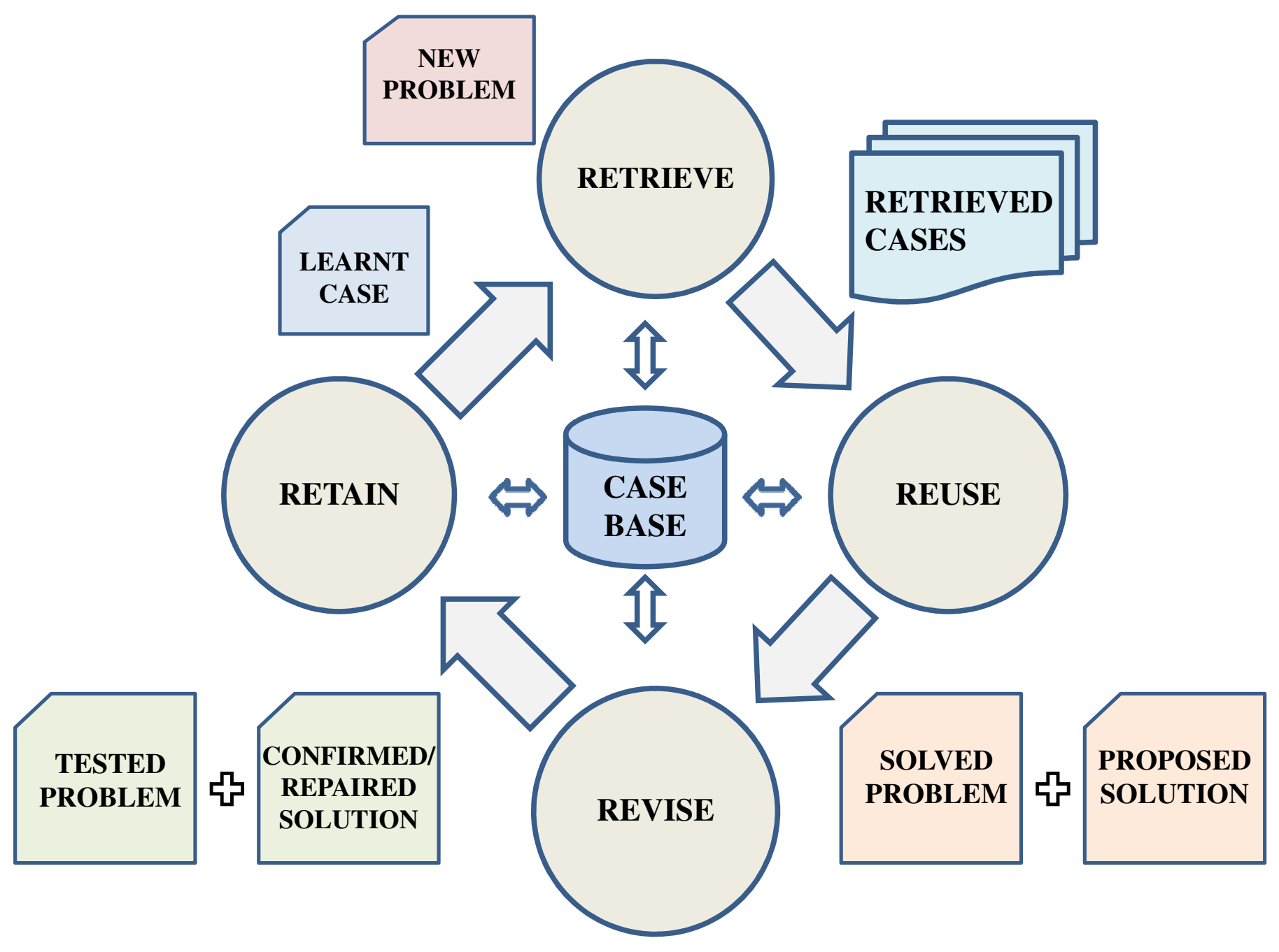

Figure: Case Based Reasoning cycle

- Case Based Reasoning(CBR) - solve new problems by reusing solutions to existing problems. Relates to instance based learning.
- Retain part - decide whether a solved problem goes into existing database or not. Also called Case Base Maintenance.
- Alignment an important input for maintenance algorithms.

3. Algorithm

Alignment of Higher Orders

Base Alignment[Massie et al., 2006]

1. Given a collection of documents C, and the number of neighbors k, and a query q.
2. Find the k nearest neighbors for the query.
3. Calculate the alignment as

$$AlignVal(q) = \frac{\sum_{c \in NN(q)} DocSim(c,q) * LabelSim(c,q)}{\sum_{c \in NN(q)} DocSim(c,q)} \quad (1)$$

Higher Orders

1. Given a collection of documents C, neighborhood order n and the number of neighbors k.
2. For each query q do,
   1. Find the k^n neighbors for the query, and calculate their alignments using 1.
   2. At every level, keep the parent point i.e the point on level $n-1$, from where the k neighbors on the next level were obtained.
   3. From the outer level, propagate the alignment score to the next inner level, weighted by the similarities to the parent, till the query point is reached.

4. Experimental Setup

- Six datasets - Relpol, Hardware, Recreation, Science, Lingspam and Usremail. First four datasets were constructed by dividing the 20 Newsgroups dataset into 4 categories based on the topics.
- The reason to divide such a way - get datasets with a varying levels of difficulty. Our notion of difficulty - accuracy on test set by various classifiers.
- Disjoint sets created, - each set contains 20 % of the original corpus randomly chosen. 15 such splits were created each containing 6 datasets of varying difficulty. The measure must be able to estimate the relative difficulty along all the 15 trials.
- We keep some documents aside as queries and use the rest of the corpus for calculating the higher order neighbors.
- Aim: To show the proposed measure can predict the complexity of the dataset. This is shown by looking at the correlation between the calculated alignment and the accuracy of various standard classifiers.
- Bayes error more intrinsic to the dataset and relates to the original distribution of data. More global in nature, while maintenance algorithms require local alignment as input.
- Thus, checking for correlations with accuracy of classifier conducted as part of the experiment.

5. Results

- Different variations of the algorithms are tested on the dataset. unwt1, unwt2 correspond to the unweighted versions of the first order and second order alignment respectively.
- wt1, wt2 correspond to the weighted version of the first and second order alignment. comb corresponds to the combination of first and second order alignments, while prop is a propagating version of the second order alignment.
- The weights used to combine the alignment of the neighbors is the distance between the neighbor and query.

	svm	nb	knn	crn	randforest
unwt1	0.95	**0.96**	**0.96**	0.65	0.80
wt1	0.95	0.96	0.96	0.64	0.80
unwt2	0.94	0.95	0.94	0.65	0.78
wt2	0.93	0.95	0.93	0.62	0.76
comb	0.95	0.96	0.95	0.65	0.79
prop	**0.95**	0.89	0.94	**0.75**	**0.91**

Table: Correlation values of the alignments scores with the classifiers - Support Vector Machine(svm) with linear kernel, Naive Bayes Classifier(nb),k-Nearest Neighbor(knn)with $k = 3$,Case Retrieval Net(crn),a spreading activation based classifier; Random Forest(randforest), an ensemble classifier for different datasets for one of the trials

6. Results (contd...)

We plot the alignment value generated for each of datasets across all the trials to see if the difference in difficulty among datasets in all the 15 sets is estimated consistently.

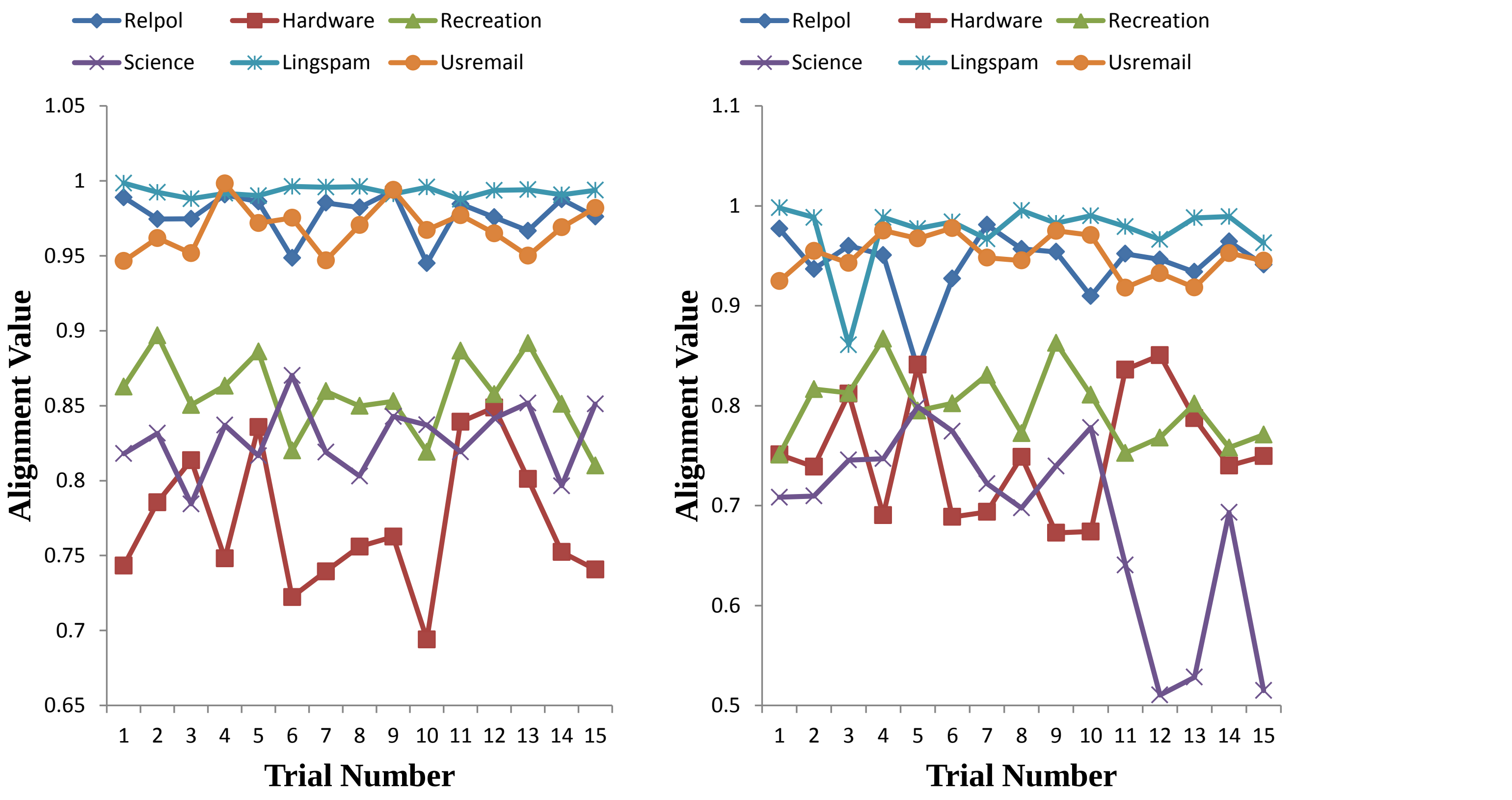

Figure: Comparing First Order(left) and Propagating Second Order(right) Alignment results across different datasets over 15 trials

7. Conclusions

- We have proposed the alignment of higher orders as an attempt to achieve a more reliable complexity estimator. The results look promising, and we wish to investigate further into much higher orders of alignment.
- The influence of higher order neighbors on lazy learning algorithms like knn is another interesting direction. We want to see if we can improve the performance of knn through intelligent choices of k derived from alignment.
- Adaptive knn - where different values of k chosen for each query. We want to see if we can make a strategy based on alignment to choose k.

References

[Chakraborti et al., 2008] Chakraborti, S., Beresi, U. C., Wiratunga, N., Massie, S., Lothian, R., and Khemani, D. (2008).
Visualizing and evaluating complexity of textual case bases.
In *Advances in Case-Based Reasoning*, volume 5239, pages 104–119.

[Massie et al., 2006] Massie, S., Craw, S., and Wiratunga, N. (2006).
Complexity profiling for informed case-base editing.
In *Proc of the 8th European Conf. on Case-Based Reasoning*, pages 325–329.

[Vinay et al., 2006] Vinay, V., Cox, I. J., Milic-Frayling, N., and Wood, K. (2006).
Measuring the complexity of a collection of documents.
In *Proceedings of the 28th European conference on Advances in Information Retrieval*, ECIR'06, pages 107–118.

Contact: kvsdilip@cse.iitm.ac.in