Hybrid Parallel Classifiers for Semantic Subspace Learning

Nandita Tripathi[1], Michael Oakes[1], and Stefan Wermter[2]

[1]Department of Computing, Engineering and Technology, University of Sunderland, St Peters Way, Sunderland, SR6 0DD, United Kingdom
(Nandita.Tripathi@hotmail.com, Michael.Oakes@sunderland.ac.uk)

[2]Institute for Knowledge Technology, Department of Computer Science, University of Hamburg, Vogt Koelln, Str. 30, 22527 Hamburg , Germany
(wermter@informatik.uni-hamburg.de)

Abstract. Subspace learning is very important in today's world of information overload. Distinguishing between categories within a subset of a large data repository such as the web and the ability to do so in real time is critical for a successful search technique. The characteristics of data belonging to different domains are also varying widely. This merits the need for an architecture which caters to the differing characteristics of different data domains. In this paper we present a novel hybrid parallel architecture using different types of classifiers trained on different subspaces. We further compare the performance of our hybrid architecture with a single classifier and show that it outperforms the single classifier system by a large margin when tested with a variety of hybrid combinations. Our results show that subspace classification accuracy is boosted and learning time reduced significantly with this new hybrid architecture.

Keywords: parallel classifiers, hybrid classifiers, subspace learning, significance vectors, maximum significance

1. Introduction

The *curse of dimensionality* [1] degrades the performance of many learning algorithms. Therefore, we need ways to discover clusters in different subspaces of datasets which are represented with a high number of dimensions [2]. Subspace analysis lends itself naturally to the idea of hybrid classifiers. Each subspace can be processed by a classifier best suited to the characteristics of that subspace. Instead of using the complete set of full space dimensions, classifier performance can be boosted by using only a subset of the dimensions. The method of choosing an appropriate

reduced set of dimensions is an active research area [3]. The use of Random Projections [4] in dimensionality reduction has also been studied. In the Random Subspace Method (RSM) [5], classifiers were trained on randomly chosen subspaces of the original input space and the outputs of the models were then combined. However random selection of features does not guarantee that the selected inputs have necessary discriminating information. Several variations of RSM have been proposed such as Relevant random feature subspaces for co-training (Rel-RASCO) [6], the Not-so-Random Subspace Method (NsRSM) [7] and the Local Random Subspace Method [8].

In the real world, documents can be divided into major semantic subspaces with each subspace having its own unique characteristics. The above research does not take into account this division of documents into semantic categories. We present here a novel hybrid parallel architecture (Fig. 1) which takes advantage of the different semantic subspaces existing in the data. We further show that this new hybrid parallel architecture improves subspace classification accuracy as well as significantly reduces training time. In this architecture, we tested various hybrid combinations of classifiers using the conditional significance vector representation [9] which is a variation of the semantic significance vector [10], [11] to incorporate semantic information in the document vectors. We compare the performance of this hybrid parallel classifier against that of single Multilayer Perceptron (MLP) classifiers using the significance vector as well as the Term Frequency – Inverse Document Frequency (tf-idf) vector representation. Our experiments were performed on the Reuters corpus (RCV1) [12] using the first two levels of the topic hierarchy.

2. Methodology Overview and Overall Architecture

Ten thousand Reuters headlines along with their topic codes were extracted from the Reuters Corpus. These headlines were chosen so that there was no overlap at the first level categorization. At the second level, since most headlines had multiple level 2 subtopic categorizations, the first subtopic was taken as the assigned subtopic. A total of fifty subtopics were included in these experiments. Headlines were then preprocessed to separate hyphenated words. Dictionaries with term frequencies were generated based on the TMG toolbox [13]. These were then used to generate the Full Significance Vector [9] and the Conditional Significance Vector [9]. The tf-idf [14] representation for each document was also generated by the TMG toolbox.

The WEKA machine learning workbench [15] was used to examine this architecture with various learning algorithms. Seven algorithms were compared for our representations to examine the different categories of classification algorithms. Classification Accuracy, which is a comparison of the predicted class to the actual class, and the Training Time were recorded for each experiment run.

3. Data Vector Sets Generation

Three different vector representations (Full Significance Vector, Category-wise separated Conditional Significance Vector and tf-idf) were generated for our data.

Full Significance Vector: For each Reuters Full Significance document vector the first four columns, representing the four main topics CCAT, ECAT, GCAT & MCAT, were ignored leaving a vector with 50 columns representing 50 subtopics. The order

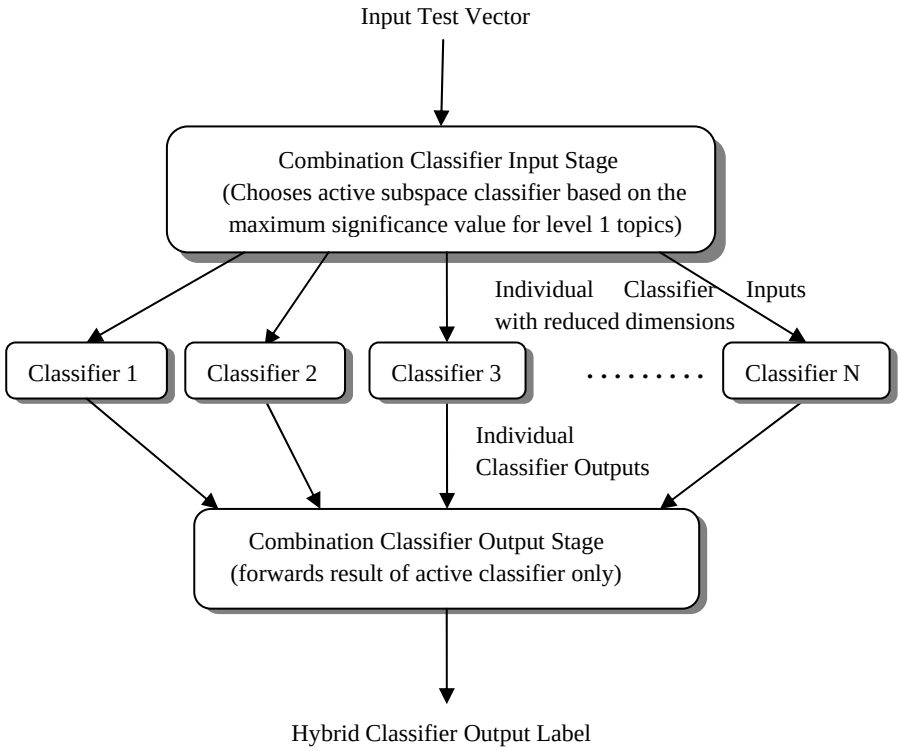


Fig. 1. Hybrid Parallel Classifier Architecture for Subspace Learning

of the data vectors was then randomised and divided into two sets – training set of 9000 vectors and a test set of 1000 vectors.

Category-wise separated Conditional Significance Vector: The order of the Reuters Conditional Significance document vectors was randomised and divided into two sets – a training set of 9000 vectors and a test set of 1000 vectors. The training set was then divided into 4 sets according to the main topic labels. For each of these for sets, only the relevant subtopic vector entries (e.g. C11, C12, etc for CCAT; E11, E12, etc

for ECAT; etc) for each main topic were retained. These 4 training sets were then used to train the 4 parallel classifiers of the Reuters hybrid classifier. The main category of a test data vector was determined by the maximum significance vector entry for the first four columns representing the four main categories. After this, the entries corresponding to the subtopics of this predicted main topic were extracted along with the *actual* subtopic label and given to the classifier trained for this predicted main category. The accuracy of choosing the correct main topic by selecting the maximum significance level 1 entry has been measured to be 96.80% for the 1000 test vectors i.e. 968 vectors were assigned the correct trained classifiers whereas 3.20% or 32 vectors were assigned to a wrong classifier – resulting in a wrong classification decision for all these 32 vectors. Hence the upper limit for classification accuracy is 96.80% for our hybrid parallel classifier.

TFIDF Vector: The TMG toolbox [13] was used to generate tf-idf vectors for the ten thousand Reuters headlines used in these experiments. The tf-idf vector dataset was then randomized and divided into a training set of 9000 vectors and a test set of 1000 vectors.

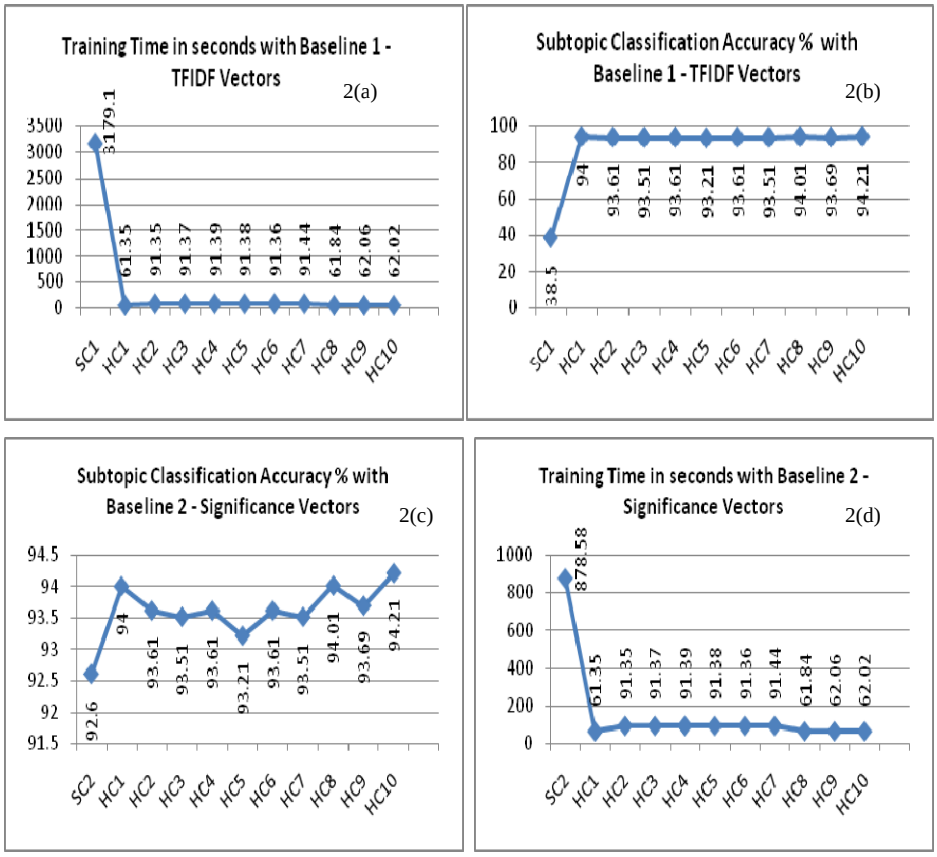
4. Classification Algorithms

Seven Classification algorithms were tested with our datasets namely Random Forest, C4.5, Multilayer Perceptron, Naïve Bayes, BayesNet, NNge and PART. Random Forests [16] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently. C4.5 [17] is an inductive tree algorithm with two pruning methods: subtree replacement and subtree raising. Multilayer Perceptron [18] is a neural network which uses backpropagation for training. Naive Bayes [19] is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. BayesNet [20] implements Bayes Network learning using various search algorithms and quality measures. NNge [21] is a nearest neighbor-like algorithm using non-nested generalized exemplars. A PART [22] decision list uses C4.5 decision trees to generate rules.

5. Results and their Analysis

Experiments were run using seven algorithms from Weka on the Reuters Corpus. In the first step, the algorithms were run using category-wise separated data from the training set to select the algorithm with the highest classification accuracy for each main category. In case of a tie between two algorithms, the one with the lower training time was chosen. Subsequently these selected algorithms were applied to the test data and the performance of the hybrid classifier was measured. The category-wise separated Conditional Significance Vectors were used here. Each of the algorithms was also run on the full (not category-wise separated) data set to provide a comparison for the hybrid classifier. Two vector representations were used for the comparison baseline – the Full Significance Vector and tf-idf. As the performances of many classifiers for each main category were quite close to each other, we also ran

some experiments using a predefined set of classifiers. The combination of MLP with different types of classifiers (Bayesian, rule-based and tree-based classifiers) was



Classifier Index

SC1- Single MLP over full data using tf-idf Vectors; **SC2**- Single MLP over full data using Significance Vectors; **HC1**- Hybrid Parallel Classifier with category-wise classifiers chosen from training data performance

Hybrid 2-Classifier Systems :

HC2 (MLP/NB), HC3 (MLP/BNet)
HC4 (MLP/NNge) , HC5 (MLP/PART)
HC6 (MLP/J48), HC7 (MLP/RF)

Hybrid 4-Classifier Systems :

HC8 (MLP/NB/ NNge/ J48)
HC9 (MLP/BNet/PART/RF)
HC10 (MLP/NNge/PART/NB)

Abbreviations: MLP - Multilayer Perceptron; NB - Naïve Bayes,; BNet – BayesNet; RF – Random Forest

Fig. 2. Hybrid Parallel Classifiers Performance Metrics

evaluated and the best combination was identified. For a two-classifier combination, MLP and the other classifier were used alternately on the main category topics while for a four-classifier system four different classifiers were used on the four main topics.

The charts in Fig 2 show a comparison of the performance of hybrid classifiers with that of MLP. The baseline single MLP classifier experiment is run with two different vector representations - Significance Vector and tf-idf. The accuracies of all the hybrid parallel classifiers are better than that of the single MLP classifier. The Hybrid 4-classifier system (HC10) shows the best result which is quite similar to that of the hybrid classifier with category-wise classifiers chosen from the training set (HC1).

Overall, it was observed that there was an improvement in subtopic classification accuracy along with a significant reduction in training time. The classification accuracies of all the hybrid classifiers were quite close to each other but all of them were much better than the classification accuracy of the single classifier with tf-idf baseline. The difference with the significance vector baseline was smaller but even there the classification accuracies of the hybrid systems were better. The training times showed a very steep reduction compared to both baselines. The average of 10 runs was taken for each experiment. In the hybrid classifier, even though we are using more classifiers, the training time is reduced. This is because each classifier now trains on a reduced set of data with a reduced set of vector components. This two-fold reduction translates to a significance decrease in training time.

6. Conclusion

In this work, we attempt to leverage the differences in the characteristics of different subspaces to improve semantic subspace learning. The main objective here is to improve document classification in a vast document space by combining various learning methods. Our experiments show that hybrid parallel combinations of classifiers trained on different subspaces offer a significant performance improvement over single classifier learning on full data space. Individual classifiers also perform better when presented with less data in lower dimensions. Our experiments also show that learning based on the semantic separation of data space is more efficient than full data space learning. Combining different types of classifiers has the advantage of integrating characteristics of different subspaces and hence improves classification performance. This technique can work well in other domains like pattern / image recognition where different classifiers can work on different parts of the image to improve overall recognition. Computational Biology too can benefit from this method to improve recognition within sub-domains.

In our experiments, subspace detection is done by processing a single document vector. This method is independent of the total number of data samples and only compares the level 1 topic entries. The time complexity of the combining classifier is thus O(k) where k is the number of level 1 topics. The novelty of our approach is in

the use of a maximum significance based method of input vector projection for a hybrid parallel classifier along with the use of Conditional Significance Vectors (which increase the distinction between categories within a subspace) for improved subspace learning. Combining MLP in parallel with a basic classifier (Bayesian, tree based or rule based) improves the classification accuracy and significantly reduces the training time. The experiments also show that using maximum significance value is very effective in detecting the relevant subspace of a test vector.

References

1. Friedman J.H.: On Bias, Variance, 0/1—Loss, and the Curse-of- Dimensionality. Data Mining and Knowledge Discovery, Volume 1, Issue 1, pp 55 – 77(1997)
2. Parsons L., Haque E., Liu H.: Subspace Clustering for High Dimensional Data : A Review. In: ACM SIGKDD Explorations Newsletter**,** Vol 6, Issue 1, 2004, pp 90 – 105(2004)
3. Varshney K.R., Willsky A.S.: Learning dimensionality-reduced classifiers for information fusion. In: Proceedings of the 12th International Conference on Information Fusion, pp 1881–1888, Seattle, Washington, (2009)
4. Fradkin, D., Madigan, D.: Experiments with Random Projections for Machine Learning. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 517-522(2003)
5. Ho, Tin Kam: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 20, Issue 8, pp 832-844(1998)
6. Yaslan Y., Cataltepe Z.: Co-training with relevant random subspaces. Neurocomputing 73, pp 1652-1661. Elsevier (2010)
7. Garcia-Pedrajas N., Ortiz-Boyer D.: Boosting Random Subspace Method. Neural Networks 21, pp 1344-1362(2008)
8. Kotsiantis S.B.: Local Random Subspace Method for constructing multiple decision stumps. In: International Conference on Information and Financial Engineering, pp 125-129 (2009)
9. Tripathi N., Wermter S., Hung C., Oakes M.: Semantic Subspace Learning with Conditional Significance Vectors. In: Proceedings of the IEEE International Joint Conference on Neural Networks, pp 3670-3677, Barcelona (2010)
10. Wermter S., Panchev C. , Arevian G.: Hybrid Neural Plausibility Networks for News Agents. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp 93-98(1999)
11. Wermter S.: Hybrid Connectionist Natural Language Processing. Chapman and Hall. 1995
12. Rose T., Stevenson M.,Whitehead M.: The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC- 02),pp 827–833(2002)
13. Zeimpekis D. , Gallopoulos E.: TMG : A MATLAB Toolbox for Generating Term Document Matrices from Text Collections. Book Chapter in Grouping Multidimensional Data: Recent Advances in Clustering, J. Kogan and C. Nicholas, eds., Springer (2005)

14. Manning C., Raghavan P., Schutze H.: Introduction to Information Retrieval. Cambridge University Press. (2008)
15. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P.,Witten I.: The WEKA Data Mining Software: An Update. In: ACM SIGKDD Explorations Newsletter, Volume 11, Issue 1,pp 10-18(2009)
16. Breiman L.: Random Forests. Machine Learning 45(1),Oct 2001, pp 5-32(2001)
17. Quinlan J.R.: C4.5 : Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. (1993)
18. Verma B.: Fast training of multilayer perceptrons. In: IEEE Transactions on Neural Networks, Vol 8, Issue 6, pp 1314-1320 (1997)
19. Zhang H.: The optimality of Naïve Bayes. American Association for Artificial Intelligence, www.aaai.org (2004)
20. Pernkopf F.: Discriminative learning of Bayesian network classifiers. In: Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications, pp 422-427(2007)
21. Martin B.: Instance-Based learning : Nearest Neighbor With Generalization. Master Thesis, University of Waikato, Hamilton, New Zealand (1995)
22. Frank E.,Witten I.H.: Generating Accurate Rule Sets Without Global Optimization. In: Shavlik, J., ed., Machine Learning: Proceedings of the Fifteenth International Conference, Morgan Kaufmann Publishers (1998)