Hybrid Classifiers for Improved Semantic Subspace Learning of News Documents

Nandita Tripathi, Michael Oakes, and Stefan Wermter

Abstract—**The volume and diversity of documents available in today's world is increasing daily. It is therefore difficult for a single classifier to efficiently handle multi-level categorization of such a varied document space. In this paper we analyse methods to enhance the efficiency of a single classifier for two-level classification by combining it with classifiers of other types. We use the maximum significance value as an indicator for the subspace of a test document. We represent the documents using the conditional significance vector which increases the distinction between classes within a subspace. Our experiments show that dividing a document space into different semantic subspaces increases the efficiency of such hybrid classifier combinations. Applying different types of classifiers on different subspaces substantially improves overall learning.**

Index Terms — **Hybrid Classifiers, Text Classification, News Categorization, Semantic Subspace Learning, Maximum Significance Value.**

I. Introduction

The high complexity of today's data spaces often leads to documents being represented by a high number of dimensions. This leads to the curse of dimensionality [1] which degrades the performance of many classifiers. The vast data space is also divided into many subspaces which are quite different from each other, e.g. medicine and politics. These subspaces are often subdivided into further categories. Therefore, we need methods which can detect categories within these subspaces [2]. Hybrid classifiers can be effectively applied to subspace analysis. Since each subspace can be viewed as an independent dataset, different classifiers can be used to process different subspaces. Instead of using the complete set of full space feature dimensions, classifier performances 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]. In the Random Subspace Method (RSM) [4], classifiers were trained on randomly chosen subspaces of the original input space and the outputs of the models were then combined. However, a random selection of features does not guarantee that the selected inputs have necessary distinguishing information. To address this problem, several variations of RSM have been proposed by various researchers such as Relevant Random Feature Subspaces for Co-training (Rel-RASCO) [5], Not-so-Random Subspace Method (NsRSM) [6] and Local Random Subspace Method [7].

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 this division into account. In this paper, we describe a hybrid parallel architecture (Fig. 1) which takes advantage of the different semantic subspaces existing in the data. We use this architecture to show how the performance of various basic classifiers can be improved by combining them with classifiers of other types. We test the hybrid combinations of classifiers using the conditional significance vector representation [8] which is a variation of the semantic significance vector [9, 10] to incorporate semantic information in the document vectors. The conditional significance vector enhances the distinction between subtopics within a given main topic. The region of the test data is determined by the maximum significance value [8] which is evaluated in O(k) time where k is the number of level 1 topics and thus can be very effective where time is critical for returning search results.

II. Methodology and Architecture

We used two different text datasets for our experiments - the Reuters Headlines dataset and the Reuters Full Text dataset – both drawn from the Reuters Corpus [11]. The Reuters Corpus is a well-known test bench for text categorization experiments. It has a hierarchical organization with four major groups which is well suited to test our hybrid architecture. Ten

Manuscript received October 18, 2011.

Nandita Tripathi is with the Department of Computing, Engineering and Technology, University of Sunderland, St Peters Way, Sunderland SR6 0DD, United Kingdom
(email: Nandita.Tripathi@research.sunderland.ac.uk)

Michael Oakes is with the Department of Computing, Engineering and Technology, University of Sunderland, St Peters Way, Sunderland SR6 0DD, United Kingdom
(email: Michael.Oakes@sunderland.ac.uk).

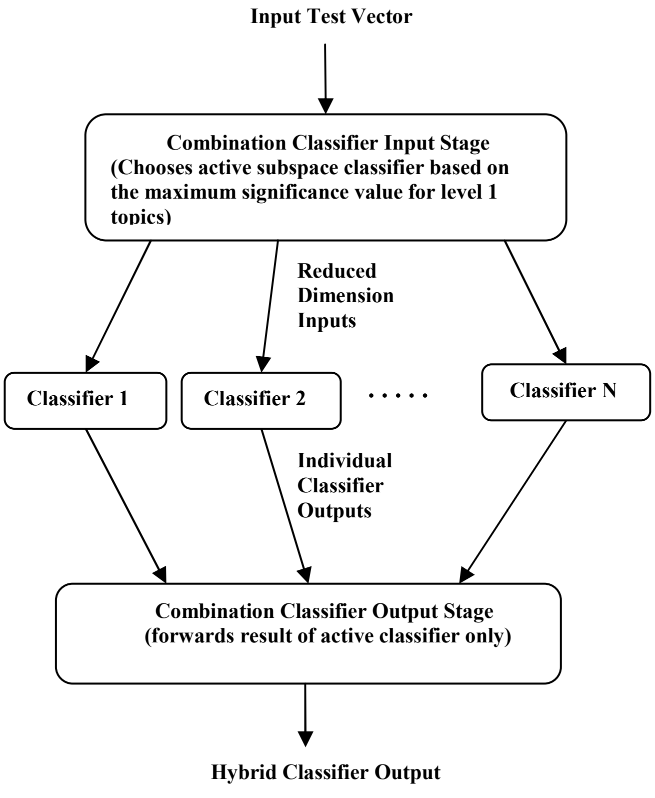
Stefan Wermter is with the Institute for Knowledge Technology, Department of Computer Science, University of Hamburg, Vogt Koelln, Str. 30, 22527 Hamburg , Germany (email: wermter@informatik.uni-hamburg.de)

thousand Reuters Headlines along with their topic codes were extracted from the Reuters corpus to constitute our first dataset. Some examples of Reuters Headlines are:

"Questar signs pact to buy oil, gas reserves."
"Ugandan rebels abduct 300 civilians, army says."
"Estonian president faces reelection challenge."

For the second dataset, we extracted ten thousand Reuters full text items which included both headline as well as body text for each news item. The news items in both cases were chosen so that there was no overlap at the first level categorization. Each news item belonged - to only one level 1 category. At the second level, since most news items had multiple level 2 subtopic categorizations, the first subtopic was taken as the assigned subtopic. Thus, each news item (for both the headlines as well as the full text datasets) had two labels associated with it – the main topic (Level 1) label and the subtopic (Level 2) label. The news items were then preprocessed to separate hyphenated words. Dictionaries with term frequencies were generated using the Text to Matrix Generator (TMG) [12] toolbox. These were then used to generate the Full Significance Vector [8] and the Conditional Significance Vector [8] for each document. The two datasets were then randomised and divided into training sets of 9000 documents and corresponding test sets of 1000 documents.

Fig. 1 Hybrid Classifier Architecture for Subspace Learning


Input Test Vector
Combination Classifier Input Stage (Chooses active subspace classifier based on the maximum significance value for level 1 topics)
Reduced Dimension Inputs
Classifier 1
Classifier 2
. . . . .
Classifier N
Individual Classifier Outputs
Combination Classifier Output Stage (forwards result of active classifier only)
Hybrid Classifier Output

The Waikato Environment for Knowledge Analysis (WEKA) [13] is a machine learning workbench with a number of learning algorithms of different types. We used two Bayesian algorithms (Naïve Bayes and BayesNet), two tree-based algorithms (J48 and Random Forest), one rule-based algorithm (PART) and one neural network (Multilayer Perceptron) as our test algorithms. These algorithms act as representatives of different classes of learning. We combined each algorithm with algorithms from other types in various new hybrid architectures in order to test a variety of learning algorithms. Classification accuracy, which is a comparison of the predicted class to the actual class, was recorded for each experiment run.

III. Experimental Data Generation

3.1 Text Data Preprocessing

Ten thousand Reuters Headlines and ten thousand Reuters Full Text items were used in these experiments. The Level 1 categorization of the Reuters corpus divides the data into four main topics, namely Corporate/ Industrial (CCAT), Economics (ECAT), Government/ Social (GCAT) and Markets (MCAT). Level 2 categorization further divides these into subtopics e.g. C11(Strategy), E21(Government Finance), GVIO(War) & M14 (Commodity Markets), which are subtopics of CCAT, ECAT, GCAT & MCAT respectively. A total of 50 subtopics were included in these experiments. Since all the news items had multiple subtopic assignments, e.g. C11/C15/C18 (Strategy/ Performance/ Ownership changes), only the first subtopic e.g. C11(Strategy) was taken as the assigned subtopic. Our assumption here is that the first subtopic used to tag a particular Reuters news item is the one which is most relevant to it. The text data was then processed two ways to generate data vectors in two different formats.

3.2 Semantic Significance Vector Generation

We use a vector representation which represents the significance of the data and weighs different words according to their significance for different topics. Significance Vectors [9, 10] are determined based on the frequency of a word in different semantic categories. A modification of the significance vector called the semantic vector uses normalized frequencies where each word w is represented with a vector $(c_1,c_2,..,c_n)$ where c_i represents a certain semantic category and n is the total number of categories. A value $v(w,\ c_i)$ is calculated for each element of the semantic vector as follows:

$$v(w,\ c_i) = \frac{\text{Normalised Frequency of } w \text{ in } c_i}{\sum_k \text{Normalised Frequency of } w \text{ in } c_k}$$

$$\text{where} \quad k \in \{1..n\}$$

For each document, the document semantic vector is obtained by summing the semantic vectors for each word in the document and dividing by the total number of words in the document. Henceforth it is simply referred to as the *significance vector*. The TMG Toolbox [11] was used to generate the term frequencies for each word in each news document. The word vector consisted of 54 columns (for 4 main topics and 50 subtopics) for the two Reuters Corpus datasets. While calculating the *significance vector* entries for each word, its occurrence in all subtopics of all main topics was taken into account. The vector generated this way was called the *Full Significance Vector*. We also generated the *Conditional Significance Vector* [8] where a word's occurrence in all subtopics *of only a particular main topic* is taken into account while calculating the word significance vector entries.

3.3 Data Vector Sets Generation

As will be described below, two of these different vector representations (Full Significance Vector and Conditional Significance Vector) were generated for our experiments. The Conditional Significance Vectors were processed further to generate four main category-wise data vector sets.

3.3.1 Full Significance Vector

Here, the document vectors were generate from the full significance word vectors as explained in section 3.2 For each Reuters Full Significance document vector the first four columns, representing four main topics – CCAT, ECAT, GCAT & MCAT, were ignored leaving a vector with 50 columns representing 50 subtopics. A train/test split of 90/10 was taken for both datasets

3.3.2 Category-wise Conditional Significance Vectors

Here, the conditional significance word vectors (see section 3.2) were used to generate the document vectors and a train/test split of 90/10 was taken generating 9000 training and 1000 test vectors for each Reuters dataset, The training set was then divided into four sets according to the main topic labels. For each of these sets, only the relevant subtopic vector entries (e.g. C11, C12, etc for CCAT; E11, E12, etc for ECAT) for each main topic were retained. Thus the CCAT category training dataset had 18 columns for 18 subtopics of CCAT. Similarly the ECAT training dataset had 8 columns, the GCAT training dataset had 20 columns and the MCAT training dataset had 4 columns. These four training sets were then used to train the four 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

3.3.3 Upper Limit for Hybrid Classifier Accuracy

For the Reuters headlines dataset, the accuracy of choosing the correct main topic by selecting the maximum significance level 1 entry was 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 for the Reuters Headlines dataset. Similarly, this upper limit was measured to be 82.50% for the hybrid parallel classifier for the Reuters Full Text dataset.

3.4 Classification Algorithms

Six classification algorithms were tested with our datasets, namely Random Forest, J48 (C4.5), the Multilayer Perceptron, Naïve Bayes, BayesNet and PART. Random Forests [14] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently. C4.5 [15] is an inductive tree algorithm with two pruning methods: subtree replacement and subtree raising. The Multilayer Perceptron [16] is a neural network which uses backpropagation for training. Naive Bayes [17] is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. BayesNet [18] implements Bayes Network learning using various search algorithms and quality measures. A PART [19] decision list uses C4.5 decision trees to generate rules.

IV. Results and Analysis

We tested various hybrid 2-classifier and 4-classifier combinations. For the hybrid 2-classifier combinations, a classifier of one type was combined with classifiers of other types in a large variety of combinations. The performance of each single classifier on the full data was compared with the performance of the hybrid

2-classifier combinations in which this particular classifier also participated. For the single classifier experiments, Full Significance Vector representation was used whereas for the hybrid classifier experiments, the category-wise separated Conditional Significance Vector representation was used. The standard text classification train/test split [20] of 90:10 was taken in all cases. The numerical values in Fig. 2 and Fig. 3 show the average of 10 runs for each classifier.

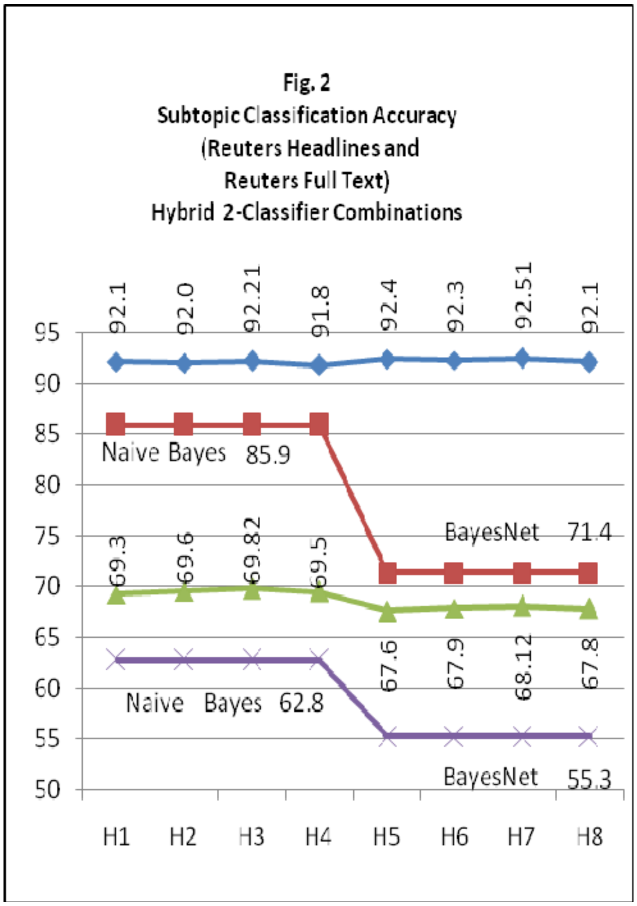


Fig. 2
Subtopic Classification Accuracy
(Reuters Headlines and
Reuters Full Text)
Hybrid 2-Classifier Combinations

<u>Fig. 2 Hybrid 2-Classifier Combinations Index</u>:

H1 - NB/J48	**H9** - J48/NB	**H17** - PART/NB
H2 - NB/RF	**H10** - J48/BN	**H18**-PART/BN
H3 - NB/MLP	**H11** - J48/MLP	**H19**-PART/J48
H4 - NB/PART	**H12** - J48/PART	**H20** -PART/RF
H5 - BN/J48	**H13** - RF/NB	**H21** - MLP/NB
H6 - BN/RF	**H14** - RF/BN	**H22** - MLP/J48
H7 - BN/MLP	**H15** - RF/MLP	**H23** - MLP/BN
H8 - BN/PART	**H16** - RF/PART	**H24 -** MLP/PART
		H25 - MLP/RF

<u>Abbreviations:</u>

NB - Naive Bayes	RF - Random Forest
BN - BayesNet	

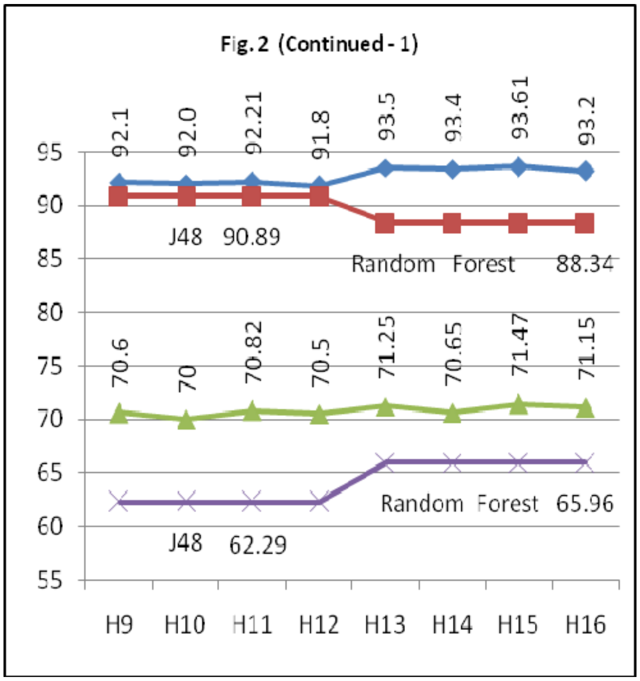


Fig. 2 (Continued - 1)

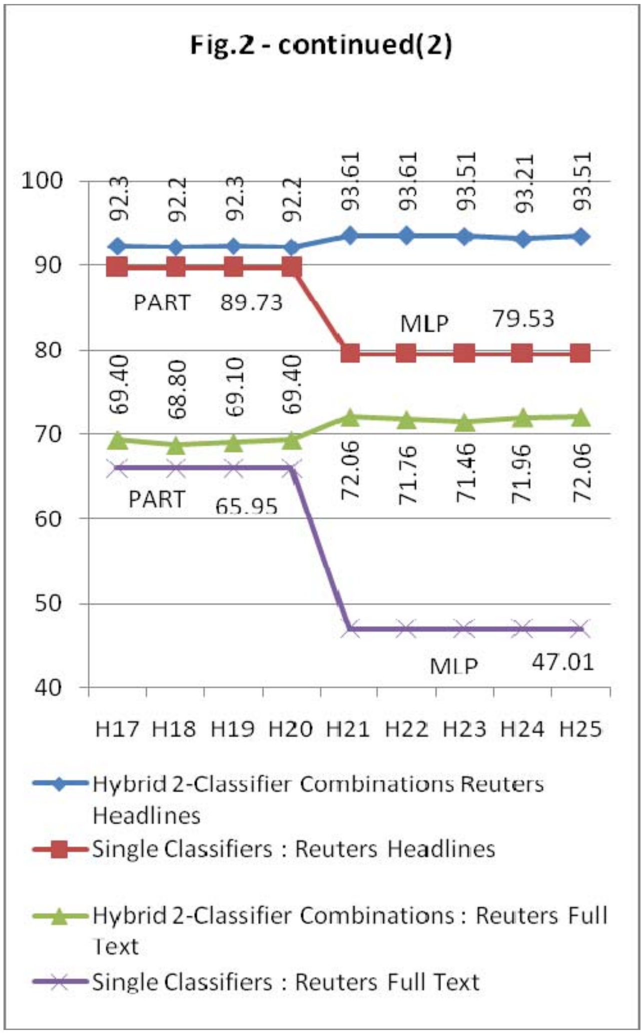


Fig.2 - continued(2)

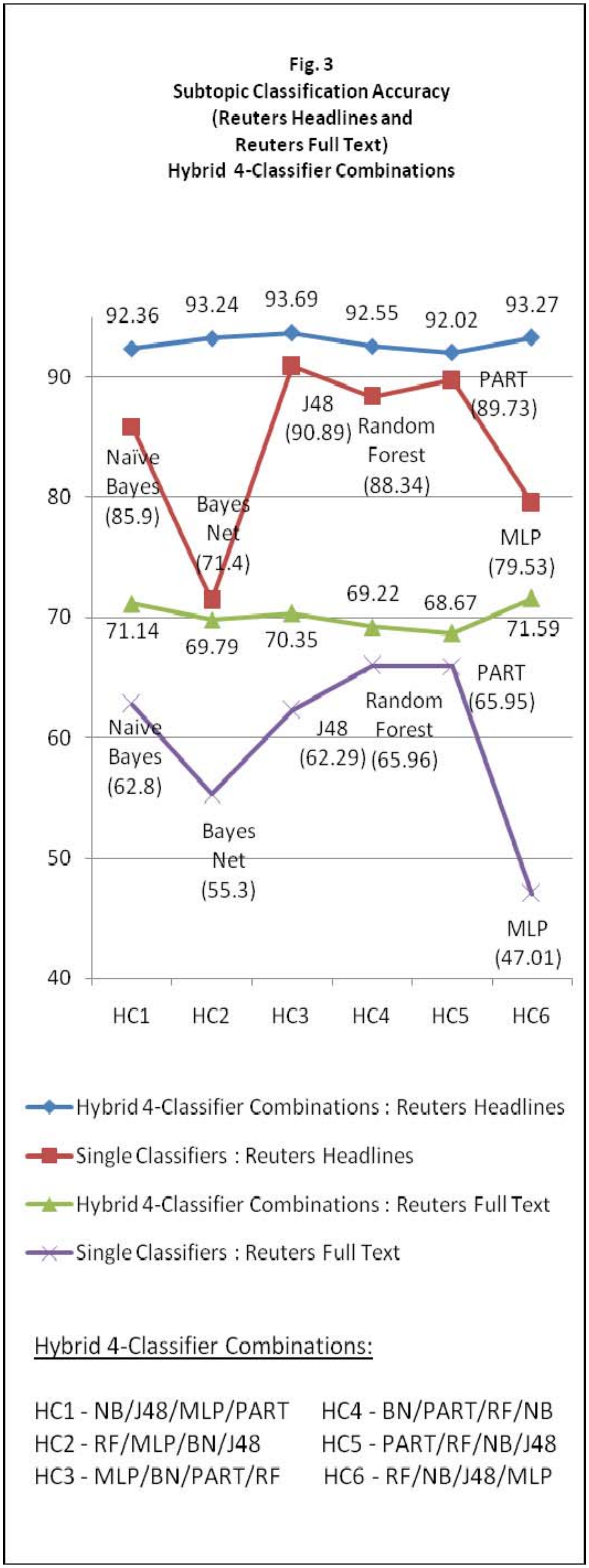


Fig. 3
Subtopic Classification Accuracy
(Reuters Headlines and
Reuters Full Text)
Hybrid 4-Classifier Combinations

In all combinations, it was observed that hybrid 2-classifier combinations performed better than the single basic classifier.

Fig. 2 shows the subtopic classification accuracy of the hybrid 2-classifier combinations along with the subtopic classification accuracy of single basic classifiers for both the Reuters Headlines as well as the Reuters Full Text datasets. Both datasets follow a similar pattern where all the hybrid classifiers perform better than any of the single classifiers. In both cases, this was statistically significant (Wilcoxon Signed Rank V = 325, p = 1.304e-05). Numerically, the classification accuracy values for the Reuters Headlines are better than those of Reuters Full Text.

The single classifier performances also show a similar pattern for both datasets. In the tree based classifiers, J48 performs better than Random Forest for Reuters Headlines and vice-versa for Reuters Full Text. In Fig. 2, the hybrid classifier data points immediately above a particular single classifier show the 2-classifier combinations which include that single classifier e.g. the hybrid classifier data points H1-H4 which are above the single classifier Naïve Bayes, show the two-classifier combinations which include Naïve Bayes. As can be seen in the figure all the hybrid 2-classifier combinations perform better than the corresponding single classifiers.

Fig. 3 shows the subtopic classification accuracy of the hybrid 4-classifier combinations along with the subtopic classification accuracy of single basic classifiers for both the Reuters Headlines as well as the Reuters Full Text datasets. Here again, both datasets follow a similar pattern where all the hybrid classifiers perform better than any of the single classifiers. Once again, this was statistically significant for both headlines and full text (Wilcoxon Signed Rank V = 21, p = 0.03125). In this case too, the numerical classification accuracy values for the Reuters headlines are better than those of Reuters Full Text.

V. Conclusion

Our experiments highlight the fact that Reuters Headlines perform better than Reuters Full Text for the purpose of news categorization. This finding is consistent across all types of experiments – single classifiers, hybrid 2-classifier combinations as well as hybrid 4-classifier combinations. This can be attributed to the fact that Reuters Full Text contains a lot of text which is introduced to make reading interesting. From a text processing point of view, this acts as noise which interferes with the relevant words (see Example 1 below). On the other hand, Reuters Headlines provide a concise summary of the news article which improves classification accuracy.

Our results also show that combining a basic classifier in parallel with classifiers of other types in a hybrid combination boosts the classification accuracy of the concerned basic classifier where the data is divided into distinct semantic categories. They also show that combining various types of classifiers in a hybrid combination results in a classification accuracy better than that of all the constituent single classifiers. The experiments confirm the fact that the maximum significance value is very effective in detecting the relevant subspace of a test document and that training different classifiers on different subsets of the original data enhances overall classification accuracy.

Example 1

Headline:

Planet Hollywood launches credit card

Body Text:

If dining at Planet Hollywood made you feel like a movie star now you can spend money like Arnold Schwarzenegger with a new credit card from the themed restaurant chain.

The fast growing company whose outlets are festooned with kitsch movie memorabilia has teamed up with the William Morris talent agency and MBNA America Bank of Wilmington-Del to offer a credit card with appropriate Hollywood perks.

These include preferential seating in the restaurants, a limited edition T-shirt and discounts on food and merchandise, a statement said. Planet Hollywood joins other pop culture companies such as Rolling Stone magazine that are issuing branded credit cards that make going into debt more fun than usual.

Approved applicants don't have to pay an annual fee and there's a special introductory annual percentage rate of 5.9 percent for balance transfers and cash advance checks. Orlando, Florida based Planet Hollywood is part of Planet Hollywood International Inc.

Full Text = Headline + Body Text

References

[1] J.H. Friedman, 1997, “On Bias, Variance, 0/1—Loss, and the Curse-of- Dimensionality,” In Data Mining and Knowledge Discovery, Volume 1, Issue 1, 1997, pp 55 – 77

[2] L. Parsons, E. Haque & H. Liu, 2004, , “Subspace Clustering for High Dimensional Data : A Review,” In ACM SIGKDD Explorations Newsletter, Vol 6, Issue 1, 2004, pp 90 – 105

[3] K.R. Varshney & A.S. Willsky, 2009, “Learning dimensionality-reduced classifiers for information fusion,” In Proceedings of the 12th International Conference on Information Fusion, pages 1881–1888, Seattle, Washington, July 2009.

[4] Tin Kam Ho, 1998. "The random subspace method for constructing decision forests" IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 20, Issue 8 (Aug 1998), pp 832-844

[5] Y. Yaslan & Z. Cataltepe, 2010, "Co-training with relevant random subspaces", Neurocomputing 73 (2010) pp 1652-1661 (Elsevier)

[6] N. Garcia-Pedrajas and D. Ortiz-Boyer, 2008, "Boosting Random Subspace Method", Neural Networks 21 (2008), pp 1344-1362

[7] S.B. Kotsiantis, 2009, "Local Random Subspace Method for constructing multiple decision stumps", International Conference on Information and Financial Engineering, pp 125-129, 2009

[8] N. Tripathi, S. Wermter, C. Hung & M. Oakes, 2010, "Semantic Subspace Learning with Conditional Significance Vectors", Proceedings of the IEEE International Joint Conference on Neural Networks, pp 3670-3677, Barcelona July 2010

[9] S. Wermter, C. Panchev & G. Arevian, 1999, “Hybrid Neural Plausibility Networks for News Agents,” In Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999, pp 93-98

[10] S.Wermter, 1995, “Hybrid Connectionist Natural Language Processing”, Chapman and Hall. 1995

[11] Rose T., Stevenson M., and Whitehead M.,“The Reuters Corpus Volume 1 - from Yesterday’s News toTomorrow’s Language Resources,” In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02),2002, pp 827–833.

[12] D. Zeimpekis and E. Gallopoulos, “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

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, 2009 , “The WEKA Data Mining Software: An Update,” In ACM SIGKDD Explorations Newsletter, Volume 11, Issue 1, July 2009, pp 10-18.

[14] L. Breiman, 2001, “Random Forests,” In Machine Learning 45(1), Oct. 2001, pp 5-32

[15] J.R. Quinlan, 1993, “C4.5 : Programs for Machine Learning,” Morgan Kaufmann Publishers, San Mateo, CA. 1993

[16] B.Verma, 1997, “Fast training of multilayer perceptrons,” In IEEE Transactions on Neural Networks, Vol 8, Issue 6, Nov 1997 pp 1314-1320.

[17] L. Jiang, D. Wang, Z. Cai and X. Yan, 2007, “Survey of Improving Naïve Bayes for Classification,”Proceedings of the 3rd International Conference on Advanced Data Mining and Applications (ADMA '07), 2007, pp 134-145

[18] F. Pernkopf, 2007, “Discriminative learning of Bayesian network classifiers,” In Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications, 2007, pp 422-427

[19] E. Frank and I.H. Witten, 1998, “Generating Accurate Rule Sets Without Global Optimization,” In Shavlik, J., ed., Machine Learning: Proceedings of the Fifteenth International Conference, Morgan Kaufmann Publishers, 1998

[20] Andrew K. McCallum "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.