

# Unsupervised Learning for Document Classification: Feasibility, Limitation, and the Bottom Line

Ji He<sup>†</sup>, Chew-Lim Tan<sup>‡</sup>, Hwee-Boon Low<sup>#</sup>, and Dan Shen<sup>#</sup>

<sup>†‡</sup>School of Computing, National University of Singapore,  
3 Science Drive 2, Singapore 117543

<sup>†</sup>mail@jihe.net; <sup>‡</sup>tancl@comp.nus.edu.sg

<sup>#</sup>Institute for Infocomm Research,  
21 Heng Mui Keng Terrace, Singapore 119613  
{hweeboon, shendan}@i2r.a-star.edu.sg

## Abstract

While unsupervised learning methods are usually proposed to handle document clustering, in the literature, there exist practices that apply these methods to document classification as well. This paper analyzes the feasibility and the limitation of such practice, and studies its efficacy through a preliminary case study on the Reuters-21578 document collection.

## 1 Introduction

Document clustering and document classification are of great importance in the text mining study. Technologies for document clustering and document classification have been widely used in enormous number of real-life applications.

In essence, both clustering and categorization involve the regression of a mixture probability model  $M(K, w, C, Y)$ , of which the domain topology is approximated as  $K$  patterns  $C_i, i = 1, \dots, K$ . With this, the probability model can be formalized as

$$P(\mathbf{x}) = \sum_{i=1}^K w_i \cdot P(\mathbf{x}|C_i, Y_i(\mathbf{x})), \quad (1)$$

where  $w_i$  is the mixture weight and  $Y_i(\mathbf{x}) \equiv \mathbf{x} \rightarrow C_i$  is a mapping from the data point  $\mathbf{x}$  to the pattern  $C_i$ . In contrast to a classification task, in which  $K$  is fixed, instances of  $\mathbf{x}$  are given corresponding mapping labels  $y(x)$ , and the objective of the learning is to estimate  $w$  as well as the distribution of  $C$  in a supervised way so that the mismatch in predicting  $y(x)$  is minimal, a clustering task usually involves a more general unsupervised learning process in the sense that all parameters of the model, namely  $K$ ,  $w$ ,  $C$ , and  $Y$ , are not known.

Based on the common modelling, there are practices in the literature that apply unsupervised learning algorithms to the document classification task. This paper reviews the principle of this idea, analyzes its feasibility and limitation, and reports our preliminary case study on the Reuters-21578 document collection.

## 2 Unsupervised Learning for Document Classification

It is widely believed that with an appropriate feature modelling, the document domain falls into a mixture Gaussian approximation. That is, the document domain can be quantized into a number of sub-spaces, documents in each sub-space following the Gaussian approximation, i.e. hyper-spherically grouped with decreasing density in response to the distance to the centroid. Various clustering studies based on this modelling have shown satisfactory performance on the document domain. Examples include the generic EM algorithm (Dempster et al., 1977) and its specialized version K-Means (Tou and Gonzalez, 1974), ISODATA (Venkateswarlu and Raju, 1992), SOM (Kohonen, 1997), as well as ART 2A (Carpenter et al., 1991b).

Based on the mixture Gaussian approximation modelling, most document classification studies presume there is a high correlation between the document patterns (classes) and their natural grouping. That is, on a well-defined problem domain, documents in the same class tend to gather in the same Gaussian sub-space while documents in different classes tend to fall into different sub-spaces<sup>1</sup>. To formalize,

$$C_i \simeq \text{mix}\{\text{Gauss}_{i\langle j \rangle}(X)\}, \quad (2)$$

<sup>1</sup>A multi-class tie can be disengaged by converting the task into multiple binary-class sub-tasks.

where  $\text{Gauss}_{i\langle j \rangle}(X)$  is the Gaussian approximation of the sub-space  $i\langle j \rangle$ . Probably one of the best proofs on the validity of this presumption is an intuitive while effective so-called “lazy learning” method,  $k$  Nearest Neighbor (kNN) (Dasarathy, 1991), which predicts the class label of a test sample according to the class labels of the  $k$  training samples that are most similar to it.

This close correlation makes unsupervised learning, while originally proposed for clustering, suitable for document classification as well. In the machine learning literature, a large number of clustering algorithms have been extensively studied to handle the classification task, either directly or after minor modification. These include the SOM-based models such as WEBSOM (Kohonen et al., 2000) and LabelSOM (Rauber et al., 2000), the ART-based models such as ARTMAP (Carpenter et al., 1991a) and ARAM (Tan, 1995), as well as various hybrid models. Among them, some have been applied to document domain and have shown satisfactory results (He et al., 2003b; Luo and Zincir-Heywood, 2003).

It is interesting to note that among the above methods for document classification, the nearest neighbor prediction also plays an important role. In essence, most unsupervised learning algorithms applied to classification form a hybrid model. The hybrid involves the topology-preserving proximation of the primitive training domain via unsupervised learning, followed with the nearest-neighbor-like predictions of the testing inputs. This process is summarized as follows:

1. **Proximation:** Through unsupervised learning, proximate the training samples on the problem domain  $X$  with a number of patterns on the prototype domain  $Z$ , preserve the mapping  $X \rightarrow Z$ .
2. **Association:** Based on the training instances  $x$  and their associated class labels  $y(x)$ , generate the mapping  $Z \rightarrow Y$  between the learnt patterns and the predefined patterns (classes).
3. **Prediction:** Given an unknown instance  $x'$ , find its association  $y(x')$  through the mapping  $X \rightarrow Z \rightarrow Y$  based on nearest neighbor prediction.

The captivating part lies on the association stage. Techniques applied in this stage can be heuristic,

which also can be extended to be a complicated classification algorithm, or non-heuristic, which can be as simple as majority voting. We discuss the later case which more strictly suites in the above hybrid. Equation 2 assumes the natural document sub-groups are homogenous, i.e. the mapping  $Z \rightarrow Y$  is many-to-one. This presumption does not always hold true in practice, due to the proximation error incurred in the clustering stage. To satisfy the presumption, some practices in the literature adopted the majority voting, i.e. to assign each pattern  $z$  with the class label that has the dominating training instances mapped to the pattern. This however compromises the topology preservation of the input. As such, it is necessary to introduce soft links (many to many mapping) between  $Z$  and  $Y$  that reflect their association. Alternatively, ARTMAP (Carpenter et al., 1991a) and ARAM (Tan, 1995) add constraints on the mapping  $X \rightarrow Z$  which force all training instances mapped to the same pattern to have the same class association. This in turn makes the hybrid rather inline, as the class label information is used during the clustering process.

Compared with the naive kNN method, one notable advantage of introducing unsupervised learning into the hybrid is the significantly lower computational cost during prediction, particularly due to the approximation mapping  $X \rightarrow Z$ , which is typically many-to-one. This advantage is of great importance to document analysis, as a document collection is typically sparse and large scaled. However, unsupervised learning does not optimize the model parameters in response to the prediction error. Hence it is foreseeable that the performance of the hybrid is predominated by the success of the nearest neighbor prediction, i.e. the validity of the presumption on the document model, as expressed in Equation 2. Previous studies on ARAM has reflected this (He et al., 2003b).

### 3 A Case Study with ART-C 2A

We did a case study with a relatively new unsupervised learning, feed-forward neural network, named ART-C 2A (He et al., 2003a), to the classification task on the extensively studied Reuters-21578 data collection<sup>2</sup>. We evaluated the performance of ART-C 2A with that of kNN, using a set of quantitative

<sup>2</sup><http://www.daviddlewis.com/resources/testcollections/reuters21578/>.

Table 1: The performance of ART-C 2A and kNN on Reuters-21578 top seven classes, in terms of precision ( $P$ ), recall ( $R$ ), and  $F_1$ . In the *McNemar’s Test Result* column, “Yes” and “No” indicates whether the difference between the prediction of the two classifiers reaches 0.05 level.

Class	Trn#	Tst#	ART-C 2A ( $C = 500, k = 1$ )			kNN ( $k = 19$ )			McNemar’s Test Result
			$P$	$R$	$F_1$	$P$	$R$	$F_1$	
earn	2,600	969	0.9821	0.8493	0.9109	0.9854	0.8338	0.9033	No
acq	1,616	703	0.8839	0.8336	0.8580	0.8987	0.8834	0.8910	Yes
money-fx	536	179	0.6325	0.8268	0.7167	0.6697	0.8156	0.7355	No
grain	433	149	0.6761	0.3221	0.4363	0.6406	0.2752	0.3850	Yes
crude	388	189	0.8592	0.6455	0.7372	0.8451	0.6349	0.7251	No
trade	369	117	0.8614	0.7436	0.7982	0.8598	0.7863	0.8214	No
interest	347	130	0.7692	0.3846	0.5128	0.7375	0.4538	0.5619	No
Micro-Average (overall):			0.8827	0.7652	0.8198	0.8859	0.7746	0.8265	
Macro-Average (average):			0.8092	0.6579	0.7100	0.8053	0.6690	0.7176	

measures.

### 3.1 The ART-C 2A Feed-Forward Neural Network

The ART-C 2A feed-forward neural network is a variance of the ART-C network (He et al., 2002), which in turn is a modification of the ART network (Carpenter et al., 1991b). Both ART and ART-C have shown competitive efficiency in handling large scale data (He et al., 2003a). Compared with conventional ART which generates indefinite number of recognition neurons (output clusters) on an unknown problem domain, ART-C provides a novel alternative by generating a fixed number of recognition neurons. We consider this more suitable for our study as the scale of ART-C’s output is easily controllable. ART-C 2A uses the symmetric cosine similarity to measure the pattern proximity. The cosine similarity has been widely used in document analysis studies.

When applied to classification, we first do clustering of the training instances with ART-C 2A, which produces  $C$  output clusters. We then build the soft links  $L_i = (l_{i1}, l_{i2}, \dots, l_{iM})$  between each  $i$ th output cluster of ART-C 2A and the predefined  $M$  classes, according to  $l_{ij} = \frac{N_{ij}}{\sum_{j=1}^M N_{ij}}$ , where  $N_{ij}$  is the number of training documents associated with cluster  $i$  and labelled with class  $j$ . Given an unknown document  $x$  during testing, the link weights of its  $k$  nearest neighbor clusters are averaged into the prediction vector  $\bar{L} = (\bar{l}_1, \dots, \bar{l}_M) = \frac{1}{k} \cdot \sum_{kNN} L_i$  and the  $J$ th class with the maximal weight is identified as the prediction:

$$J = \operatorname{argmax}\{\bar{l}_j : \text{for } j = 1, \dots, M\} \quad (3)$$

### 3.2 The Experiments

Our study was carried on the top seven classes of Reuters-21578, in terms of the number of positive training samples with ModLewis split, each class containing over 100 positive testing samples. Following a common practice, the documents were modelled with Unigram. The seven-class classification task was divided into seven binary-class tasks to facilitate feature selection using CHI statistics (Yang and Liu, 1999). 365 keywords were selected after text clean up and stop list removal. TF-IDF (Yang and Liu, 1999) and the second-level Euclidean normalization was used in document feature representation. The numbers of training and test samples of each class, after null vector removal, are listed in Table 1. The prediction efficacy of the classifiers was evaluated using precision ( $P$ ), recall ( $R$ ), and  $F_1$  rating (Yang and Liu, 1999). In addition, we empirically evaluated the efficiency of the two classifiers by comparing the CPU times used during training and testing. Since the training and testing portion of the corpus are fixed according to ModLewis split, we adopted the McNemar’s test (Dietterich, 1998) to identify the significance of our comparison results. Table 1 reports the comparative results on the prediction efficacy. Table 2 reports the comparative results on the training/testing efficiency.

Due to the time constraint in preparing this manuscript, the classifiers’ parameters were not optimized. Our controlled experiments however show the prediction efficacy of ART-C 2A and kNN are reasonably comparable to each other. Specifically, McNemar’s test falls below significance level 0.05 on five categories, while on the rest two categories, the comparison results are mixed. In terms of effi-

Table 2: The CPU time cost of ART-C 2A and kNN on Reuters-21578 top seven classes.

	Training	Testing
Total Document #	6,289	2,436
ART-C 2A Avrg Time Cost	97.1 (s)	26.4 (s)
kNN Avrg Time Cost	–	315.8 (s)

ciency, kNN’s time cost is dominated by the computational cost during prediction, as it adopts “lazy learning”. While there is a relatively high time cost on ART-C 2A’s learning process, ART-C 2A’s prediction is of significantly higher efficiency over kNN. The total time cost of ART-C 2A during training and testing is also significantly lower than that of kNN.

#### 4 Concluding Remarks

We reviewed the idea of applying unsupervised learning algorithms to document classification. The feasibility of this practice is analyzed and studied through our preliminary case study.

However, it’s important to remark on the limitation of such practice. Due to the nature of unsupervised learning, which does not optimize the model’s parameters in response to the prediction accuracy, the performance of the hybrid highly depends on the success of the nearest neighbor prediction. Hence, how to improve the classification efficacy of such a straight forward practice remains a challenging work.

On the other hand, our review and case study suggest the bottom line of this practice. That is, compared with the generic kNN classification, an appropriately applied unsupervised learning algorithm may still be capable of producing satisfactory performance for document classification, with notable higher efficiency over kNN.

#### References

G.A. Carpenter, S. Grossberg, and J.H. Reynolds. 1991a. ARTMAP: Supervised real-time learning and classification of nonstationary data by self-organizing neural network. *Neural Networks*, 4:565–588.

G.A. Carpenter, S. Grossberg, and D.B. Rosen. 1991b. ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4:493–504.

B.V. Dasarathy. 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Las Alamitos, California.

A. Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B(39)*:1–38.

T.G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

J. He, A.H. Tan, and C.L. Tan. 2002. ART-C: A neural architecture for self-organization under constraints. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 2550–2555.

J. He, A.H. Tan, and C.L. Tan. 2003a. Modified ART 2A growing network capable of generating a fixed number of nodes. *IEEE Transactions on Neural Networks*. In press. Available via <http://www.jihe.net/publications.htm>.

J. He, A.H. Tan, and C.L. Tan. 2003b. On machine learning methods for chinese documents classification. *Applied Intelligence*, 18:311–322.

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. 2000. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

T. Kohonen. 1997. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, second edition.

X. Luo and N. Zincir-Heywood. 2003. A comparison of som based document categorization systems. In *International Joint Conference on Neural Networks*.

A. Rauber, E. Schweighofer, and D. Merkl. 2000. Text classification and labelling of document clusters with self-organising maps. *Journal of the Austrian Society for Artificial Intelligence*, 19(3):17–23, October.

A.H. Tan. 1995. Adaptive resonance associative map. *Neural Networks*, 8(3):437–446.

J.T. Tou and R.C. Gonzalez. 1974. *Pattern Recognition Principles*. Addison-Wesley, Massachusetts.

N. Venkateswarlu and P. Raju. 1992. Fast ISO-DATA clustering algorithms. *Pattern Recognition*, 25(3):335–342.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49.