Hierarchical Text Categorisation based on Neural Networks and Dempster-Shafer Theory of Evidence

Gertrud Jeschke and Mounia Lalmas
Department of Computer Science,
Queen Mary University of London, London E1 4NS, United Kingdom
mounia@dcs.qmul.ac.uk, gertrud_jeschke@westlb-systems.co.uk

Abstract

This paper investigates an approach for text categorization that combines neural networks and the Dempster-Shafer Theory of Evidence to represent hierarchical relationships between categories.

Keywords: automatic text categorization, hierarchical relationships, Dempster-Shafer theory of evidence, neural network.

1 Introduction

The objective of text categorization is to appropriately assign predefined categories to text documents [7,12]. In this paper, we are interested in automatic categorization, where categories display hierarchical relationships: some categories, the super-categories, are broader than others (e.g., animal vs. mammal), whereas other categories, the sub-categories, are narrower than others (e.g., mammal vs. animal).

This paper investigates an approach for text categorization that provides for the explicit representation of hierarchical relationships. The approach, based on the work in [6,10], assigns categories to documents using evidence coming from categories lower in the hierarchy. This evidence is modeled using the Dempster-Shafer Theory of Evidence [8], where evidence is quantified through the so-called basic probability assignment and belief functions.

In our approach, a basic probability assignment value represents the extent to which a document belongs to a category taking into account the category alone, and the belief value represents this extent but taking into account the sub-categories using their basic probability assignment values. One important issue is the estimation of the basic probability assignment values.

In this paper, we propose to use neural networks to estimate these values. Neural networks are information-processing systems that have the ability to associate input patterns with a particular response by adapting to their environment through learning [1]. Neural networks have been used to automatically categorize documents according to predefined categories, and have been shown to provide good results. Therefore, they seem a good candidate to our estimation task.

Section 2 provides a brief background on neural networks and the Dempster-Shafer Theory of Evidence. Section 3 describes our approach. Section 4 describes small-scale experiments carried out to evaluate our approach, as well as the analysis of the results. We conclude in Section 5.

2 Background

2.1 Neural Networks

Neural networks consist of many individual processing units, the so-called neurons, connected by links, which are able to work in parallel. These links have weights associated to them, and allow neurons to activate other neurons, thus leading to the collective processing of all the units in the network. A learning algorithm is used to find suitable weight values for the links, such that a set of inputs applied to the network produces a desired set of outputs.

Training of neural networks is usually based on two distinct sets, the training set and the test set. The training set contains the patterns with which the neural network acquires its domain knowledge during the training phase. The test set consists of patterns that are used to estimate how well the neural network performs.

Back-propagation is one of the most widely used training algorithms, and is divided into two main phases. In the first phase, inputs traverse the layers of the neural network from the input layer to the output layer, modified by the weight values associated with the links between the units, also referred to as nodes, and their corresponding activation functions (e.g., the sigmoid function). The output layer produces output values for the particular inputs it was presented with. Between this and the next phase the actual output of the neural network is compared against the desired output, and the difference between the two outputs (i.e. the error) is calculated.

In the second phase, all weights in the neural network are adjusted to minimize the error using the gradient descent method. This method takes the contribution of layers/nodes to the error into account and updates all weight values proportionally to their contribution to the total output of the network. This process starts from the output layer and proceeds backward in adjusting the weights. Parameters such as a learning rate, and an error-tolerance are part of the algorithm. Other parameters include the number of iterations (epochs) and the number of hidden units.

2.2 The Dempster-Shafer Theory of Evidence

Dempster-Shafer Theory of Evidence is an extension of probability theory, which allows the representation of uncertainty and the combination of evidence. Dempster-Shafer Theory starts with the definition of all possible values, *U*, called a frame of discernment that a variable can take. An exact belief value is assigned to each subset of *U* and represents the uncertainty that the value of the variable belongs to the set. All such belief values form the so-called basic probability assignment (bpa) function and satisfy:

$$\sum_{A \subseteq U} m(A) = 1 \quad \text{and} \quad m(\emptyset) = 0$$

If $m(A)$ >0 then A is called a focal element. The focal elements and their bpa values define a body of evidence. The total belief provided by a body of evidence for any subset of *U* is given by:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

$Bel(A)$ is the uncertainty that the value of the variable belongs to A taking into account all possible evidences, i.e. the subsets of A.

3 The Model

Let *C* be the set of hierarchically structured categories. The set of all atomic categories (i.e. categories in *C* having no sub-categories) constitutes the frame of discernment, i.e., $U \subseteq C$. For example, let *U* = {‘Cows’, ‘Ducks’, ‘Geese’, ‘Goats’, ‘Hens’}. All other categories can be viewed as labels for sets of lower-level categories; they are hence modeled as subsets of *U*. For example, suppose that ‘Poultry’ is composed of the atomic sub-categories ‘Hens’, ‘Ducks’, and ‘Geese’. Therefore, ‘Poultry’ is modeled by the subset {‘Hens’, ‘Ducks’, ‘'Geese’}.

We define next a body of evidence to represent the categories assigned to a document. Each assigned category constitutes a focal element, and its bpa *m* captures the exact belief that the document belongs to the category. For instance, m({‘Hens’, ‘Ducks’, ‘Geese’}) = 0.2 and m({‘Hens’}) = 0.4 means that the document belongs more to ‘Hens’ than to ‘Poultry’. The belief function *Bel* expresses the overall belief that the document belongs to a category and considers the category and its sub-categories. In our example, *Bel*({‘Hens’, ‘Ducks’, ‘Geese’}) = 0.2+0.4 = 0.6, which is the total belief that the document belongs to the category ‘Poultry’.

The next step is the estimation of the bpa values. We use neural networks for this purpose. For each category, we build a classifier using neural network. Term weights (we use tf.idf) derived from document indexing are used as input values to the neural network. The bpa values are derived from the neural network outputs of each classifier. To obtain appropriate estimations, we require effective classifiers. Therefore, we build classifiers following the approach in [5], which has shown to achieve good results.

For each classifier, we determine the terms that form the input layer (i.e. input nodes) of the neural network. We then examine the terms indexing a training document. If a term corresponds to an input node, the input for this document to the input node is set to the weight associated with the term in the document; otherwise a value of 0 is used.

For each classifier, one output node is used, which indicates whether a document belongs to the corresponding category.

$$\frac{(N_{r+} \times N_{n-} - N_{r-} \times N_{n+})\sqrt{N}}{\sqrt{(N_{r+} + N_{r-}) \times (N_{n+} + N_{n-}) \times (N_{r+} + N_{n-}) \times (N_{r-} + N_{n-})}}$$

N_{r+} is the number of positive documents in which the term occurs.
N_{n+} is the number of negative documents in which the term occurs.
N_{r-} is the number of positive documents in which the term does not occur.
N_{n-} is the number of negative documents in which the term occurs.
N is the number of distinct terms.

Figure 1: Ranking formula for determining the terms to be used as inputs to the neural network.

Table 1: Performance when using neural networks only.

	MAX F1
f1_HDC	0.3867069
f1_aorco	0.2702703
f1_CVA	0.4074074
f1_DAP	0.2962963
f1_tetfal	0.6666667
f1_TranGV	0.125
f1_HSDA	0.5172414

	MAX F1
f1_HVD	0.1008403
f1_AVI	0.4086022
f1_AVS	0.5668449
f1_MVI	0.3736264
f1_MVS	0.6666667
f1_PVS	0.1904762
f1_TVI	0.2439024

	MAX F1
f1_cordis	0.5016026
f1_angpec	0.2492212
f1_corart	0.1699346
f1_corthrom	0.3272727
f1_corvas	0.4736842
f1_angun	0.4814815

The activation function is the sigmoid function and therefore output values range between 0 and 1. In the training phase, we set the output value of an input pattern of a document to 1 if the document belongs to the corresponding category. Otherwise, the output value is set to 0. This trains the network to respond with high output values to inputs for which it finds enough indications to assign the category and to respond with low output values, otherwise

Following [4,5], we build a “queryzone”, based on the work of [9], to obtain a good selection of documents to train each classifier. The selected documents consist of positive and negative documents, i.e. documents that belong and do not belong to the corresponding category, respectively. The number of negative documents will usually be much larger than that of positive documents, which may bias the classifier. Therefore, a strategy is needed to appropriately select negative documents.

For each category, we create an average vector from the representations (vectors d) of a subset of *n* of its positive documents:

$$\mu = \tfrac{1}{n}\sum_{1}^{n} d_i$$

The created average vector is used as a query to retrieve documents from the collection using the vector space model. The retrieved documents are then ranked in order of their estimated relevance to the query. We take the top ranked 10,000 documents and add to them all other selected positive documents that have not been retrieved. This set constitutes the “queryzone”, which we use as the training set.

To determine a suitable set of input terms, we use the formula in Figure 1 to rank terms based on whether they occur frequently in positive documents but infrequently in negative documents. We select the top 25 terms. We built a back-propagation network with 50 hidden units and train it with 100 epochs with a learning rate of 0.5.

4 Experiments and Analysis

4.1 Set-up

Experiments were carried out using a small subset of the MEDLINE collection, which consists of bibliographical references to medical journals, and MeSH, which is a thesaurus of controlled vocabulary terms for indexing medical abstracts. The subset of MEDLINE used is a subset of OHSUMED composed of MEDLINE abstracts from the years 1987-1991. We use the sub-tree of ‘Heart Diseases’ located in MeSH under the category ‘Diseases’, which includes a total of 119 subcategories.

The evaluation consists of finding whether our approach assigns the correct categories to the documents of the test set. We use the F1 measure for this purpose [7]:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

where precision = a/a+c and recall = a/a+b, and a = number of documents for which our approach assigns the correct category, b = number of documents for which our approach assigns the incorrect category, and c = number of documents for which our approach fails to assign the correct category.

We select three categories and their sub-categories to evaluate our approach. The three categories are ‘Heart Diseases, Congenital’ (HDC), ‘Heart Valve Diseases’ (HVD) and ‘Coronary Disease’ (cordis). The results for the three selected categories and their sub-categories are shown in Table 1. Various threshold values were used to decide whether a category is assigned to a document. We only provide the results for those thresholds that lead to the highest F1 measures.

We observe that the individual neural networks classifiers perform very differently across categories and sub-categories. The same was observed in [4,5] (variance: 0.1606 vs 0.15937; standard variation: 0.0258 and 0.0254, respectively). Our average F1 value over all categories is 0.375, which is lower compared to that obtained in [4,5] (average F1 of 0.445), but the performance of our classifiers are still acceptable. Therefore their outputs can be used as inputs to estimate our bpa values.

4.2 Estimating the bpa Values

The output values of the neural networks that were built for each category were used as inputs to estimate the bpa values. Normalization was applied on the output values to obtain a bpa function. We carried out two sets of experiments. First, we consider all sub-categories of the three selected main categories to calculate the belief values. Second, we use the direct sub-categories only. Again, various threshold values were used to decide whether a category is assigned to a document. We provide the results for those thresholds that lead to the highest F1 measures.

Table 2 shows the results for the three main categories. We also show the difference in performance between our approach (referred to as DS) and using the neural networks alone (referred to as NN). _suball and _subdir refer to the first and the second experiments, respectively.

The results indicate that our estimations based on the outcomes of the neural networks are not appropriate, apart for the category ‘Heart Valve Diseases’. However, in all cases, better results are obtained using direct sub-categories only. The latter would indicate that using direct evidence is better than using all possible evidences in calculating the belief values.

The neural network classifier for the category ‘Heart Valve Diseases’ is by comparison to the other classifiers performing badly. It shows a maximal performance of 0.10084, which is below the average performance (0.375) over all categories. This suggests that estimating the bpa values based on neural networks may be of benefit for weakly performing classifiers. To test this hypothesis a new set of experiments was carried out.

4.3 Weakly Performing Categories

The category ‘Heart Valve Diseases’ is the only weak performing main category. The other two main categories perform well, so we degraded their performances by training them with a lower number of positive documents. Table 3 shows the results for the three main categories.

Significant performance improvements with respect to using neural networks alone were achieved with our approach. This would indicate that our approach could be used in addition to neural networks to obtain more effective categorization for weakly performing classifiers.

Although, these gains in performance seem promising, they should be compared against other methods that take into account sub-categories. Therefore, we carried out a final set of experiments.

4.4 Training using Sub-categories

The aim of this set of experiments is to determine whether training the classifiers with inputs from sub-categories as well as the super-categories would have led to performance comparable to training the classifiers without sub-categories but using the belief functions to take into account the sub-categories.

We train each classifier using documents belonging to the “queryzone” determined for the category and its sub-categories.

Table 2: Performance with the bpa values estimated using neural networks.

	MAX F1 - NN	MAX F1 - DS	%
HDC_suball	0.386707	0.355438	-8.085966895
HDC_subdir		0.362177	-6.772931467
HVD_suball	0.10084	0.107801	6.903014677
HVD_subdir		0.109955	9.039071797
cordis_suball	0.5161	0.511832	-0.833867363
cordis_subdir		0.512304	-0.740966301

Table 3: Performance for weakly performing categories.

	MAX F1 - NN	MAX F1 - DS	%
HDC25_subdir	0.194631	0.199288	2.392732915
HVD_subdir	0.10084	0.109955	9.039071797
cordis25_subdir	0.318599	0.361846	13.57411668

Table 4: Performance when training neural networks with sub-categories.

	MAX F1
f1_HDCsd25	0.1748252
f1_HDCsd	0.2628866
f1_HDCsd75	0.225256
f1_HDC	0.1946309
f1_DSHDC	0.1992883

	MAX F1
f1_HVDsd25	0.111588
f1_HVDsd	0.1538462
f1_HVDsd75	0.1305842
f1_HVD	0.1008403
f1_DSHVD	0.1099554

	MAX F1
f1_cordis25sd25	0.2792852
f1_cordis25sd	0.2662968
f1_cordis25sd75	0.3309625
f1_cordis25	0.3185988
f1_DScordis25	0.3618462

To consider the extent to which a sub-category contributes to the training of the network, the input values associated to the terms in the documents belonging to the sub-category were multiplied by a factor of 0.75, 0.5 (the sub-category is half as important as its super-category) and 0.25.

Table 4 contains the results for the three main categories. sd25 corresponds to a contribution of 0.25, sd of 0.5, and sd75 of 0.75 of the sub-categories,. The two remaining lines show the performance when training is only performed using documents of the main category, and when in addition to training belief functions are used, respectively.

The best performance for the category ‘Heart Disease, Congenital’ is achieved when setting the contribution of the sub-categories to 0.5 followed by using 0.75. The best performance for the category ‘Heart Valve Disease’ is the same as for ‘Heart Disease, Congenital’. The best performance for the category ‘Coronary Disease’ is achieved using our approach, followed by setting the contribution of the subcategories to 0.75. This latter result differs from the previous two, as it is the only result showing that our approach performs best.

These findings do not indicate which method is best for taking into account sub-categories (training with input from sub-categories or using after training without sub-categories belief functions), apart for a tendency of a contribution of 0.75 of the sub-categories to perform second best.

6. Conclusion

This paper proposes an approach for text categorization that takes into account hierarchical relationships between categories. The model is based on the work of [6,10] using the Dempster-Shafer Theory of Evidence. With this theory, evidence coming from lower categories is used to assign upper categories to documents. This evidence is captured through the use of a bpa function and a belief function, where the bpa values need to be estimated. The outcomes of classifiers based on back-propagation neural networks are used to estimate these values.

Experiments were carried out using a small set of data derived from the OSHUMED collection and the MeSH thesaurus. The results do not indicate that neural networks form a good basis to estimate the basic probability assignment values.

However, we could observe that using the Dempster-Shafer theory on top of neural network improves the performance of weak classifiers. Also, we saw that considering closely related sub-categories seems to perform better than considering all sub-categories. However, further experiments are needed to validate these results, as well as applying the approach on a much larger set of data (i.e. for more categories). In addition, we need to compare the proposed approach with other approaches that consider hierarchical relationships between categories (e.g. [2,3,11]).

Acknowledgements

This work was carried out as part of Gertrud Jeschke's Master in Advanced Methods in Computer Science, at Queen Mary University of London, in 2001. Gertrud Jeschke study was partly funded by an EPSRC training scholarship.

References

[1] Fausett, L. *Fundamental of Neural Network*. Prentice Hall, 1994.

[2] Gaussier, E., Goutte, C., Popat, K. and Chen, F. A hierarchical model for clustering and categorizing documents. *ECIR*, Glasgow, pp 229-247, 2002.

[3] Klas, C-P. and Fuhr, N. A new Effective Approach for Categorizing Web Documents. *Proceedings of the 22th BCS-IRSG Colloquium on IR Research*, Cambridge, 2000.

[4] Ruiz, M.E. and Srinivasen, P. Automatic Text Categorization using Neural Networks. *Advances in Classification Research vol 8., Proceedings of ASIS SIG/CR Classification Research Workshop*, pp 59-72, 1998.

[5] Ruiz, M.E. and Srinivasen, P. Combining machine learning and hierarchical indexing. *Advances in Classification Research vol 10., Proceedings of ASIS SIG/CR Classification Research Workshop*, 1999.

[6] Schocken, S. and Hummel, R.A. On the use of the Dempster-Shafer model in information indexing and retrieval applications. *Int Journal of Man-Machine Studies 39*, 1993.

[7] Sebastiani, F. Machine learning in automated text categorization. *ACM Computer Survey*, 34(1):1-47, 2002.

[8] Shafer, G. *A mathematical theory of evidence*. Princeton University Press, 1976.

[9] Singhal, A., Mitra, M. and Buckley, C. Learning routing in a query zone. *ACM SIGIR* Philadelphia, pp 65-74, 1997.

[10] Teixera de Silva, W. and Milidiiu, R.L. Belief Model for information retrieval. *JASIS* 4(1):10-18, 1993.

[11] Vinokourov, A. and Girolami, M.A. probabilistic framework for the hierarchic organization and classification of document clustering. *JIIS* 18:153-172, 2002.

[12] Yang, Y. and Liu, X. A re-examination of text categorization methods. *ACM SIGIR*, pp 42-49, 1999.