

Unsupervised Feature Selection for Text data

Nirmalie Wiratunga, Rob Lothian, and Stewart Massie

School of Computing,
The Robert Gordon University,
Aberdeen AB25 1HG, Scotland, UK
{nw|rml|sm}@comp.rgu.ac.uk

Abstract. Feature selection for unsupervised tasks is particularly challenging, especially when dealing with text data. The increase in online documents and email communication creates a need for tools that can operate without the supervision of the user. In this paper we look at novel feature selection techniques that address this need. A distributional similarity measure from information theory is applied to measure feature utility. This utility informs the search for both representative and diverse features in two complementary ways: CLUSTER divides the entire feature space, before then selecting one feature to represent each cluster; and GREEDY increments the feature subset size by a greedily selected feature. In particular we found that GREEDY's local search is suited to learning smaller feature subset sizes while CLUSTER is able to improve the global quality of larger feature sets. Experiments with four email data sets show significant improvement in retrieval accuracy with nearest neighbour based search methods compared to an existing frequency-based method. Importantly both GREEDY and CLUSTER make significant progress towards the upper bound performance set by a standard supervised feature selection method.

1 Introduction

The volume of text content on the Internet and the widespread use of email-based communication have created a need for text classification, clustering and retrieval tools. There is also growing research interest in email applications, both within the Case-Based Reasoning (CBR) community [6, 12] and more generally in Machine Learning [15]. Fundamental to this interest is the challenge posed by unstructured content, large vocabularies and changing concepts. Understandably, much of the research effort is directed towards mapping text into structured case representations, so as to facilitate meaningful abstraction, comparison, retrieval and reuse.

Feature selection plays an important role for the indexing vocabulary acquisition task. Often this initial selection can be either directly or indirectly applied to identify

representative dimensions with which structured cases can be formed from unstructured text data. Applied directly, each selected feature corresponds to a dimension in the case representation. When applied indirectly, selected features are first combined to identify new features in a process referred to as feature extraction before they can be used as dimensions for case representation [4, 25]. Although feature extraction is undoubtedly more effective than feature selection at capturing context, our experiences with supervised tasks suggests that feature selection is an important complementary precursor to the extraction phase [24]. In this paper we are interested in feature selection applied directly to derive case representations for unsupervised tasks involving text data.

Feature selection reduces dimensionality by removing non-discriminatory and sometimes detrimental features, and has been successful in improving accuracy, efficiency and comprehension of learned models for supervised tasks in both structured [8, 10] and unstructured domains [26]. Feature selection in an unsupervised setting is far more challenging, especially when dealing with text data. Typical applications (e.g. email, helpdesk, online reports) involve clustering of text for retrieval and maintenance purposes. The exponential increase in on-line text content creates a need for tools that can operate without the supervision of the user. However, in spite of this need, current research in feature selection is mainly concerned with supervised tasks only.

The aim of this paper is to apply unsupervised feature selection to text data. We introduce feature selection methods that are applicable to free text content as in emails and to texts that are sub-parts of semi-structured problem descriptions. The latter form is typical of reports such as anomaly detection or medical reports. Analysis of similar words and their neighbourhoods provide insight into vocabulary usage in the text collection. This knowledge is then exploited in the search for representative yet diverse features. In a GREEDY search, the next best feature to select is one that is a good representative of some unselected words, but also unlike previously selected words. This procedure maintains representativeness while ensuring diversity by discouraging redundant selections. Greedy search can of course result in locally optimal, yet globally non-optimal feature subsets. Therefore, a globally informed search, CLUSTER selects representative features from word clusters.

Central to feature selection methods introduced in this papers is the notion of similarity between words. Word co-occurrence behaviour is a good indicator of word similarity, however co-occurrence data derived from textual sources is typically sparse. Hence, distance measures must assign a distance to all word pairs, whether or not they co-occur in the data. Distributional similarity measures (obtained from information theory) achieve this by comparing co-occurrence behaviour on a separate disjoint set of target events [18]. In this paper events are all other words. Intuitively, if a group of words are distributed similarly with respect to other words then selecting a single representative from a neighbourhood of words will mainly eliminate redundant information. Consequently, this selection process will not hurt case representation, but will significantly reduce dimensionality. A further advantage of exploiting co-occurrence patterns is that it provides contextual information to resolve ambiguities in text such as similar meaning words that are used interchangeably (synonyms) and the same word being used with different meaning (polysemies). In both situations similar cases can be overlooked during retrieval if these semantic relationships are ignored.

Section 2 presents existing work in unsupervised feature selection and work related to distributional distance measures and clustering based indexing schemes. Next we establish our terminology before presenting the baseline method in Section 3. Details of distributional distance measures and the role of similarity for unsupervised feature selection is discussed in Section 4. Section 5 introduces the two similarity-based selection methods, GREEDY and CLUSTER. Experimental results are reported on four email datasets in Section 6, followed by conclusions in Section 7.

2 Related Work

Feature selection for structured data can be categorised into filter and wrapper methods. Filters are seen as data pre-processors and generally, unlike wrapper approaches, do not require feedback from the final learner. As a result they tend to be faster, scaling better to large datasets with thousands of dimensions, as typically encountered in text applications. Comparative studies in supervised feature selection for text have shown heuristics based on Information Gain (IG) and the Chi-squared statistic to consistently outperform less informative heuristics that rely only on word frequency counts [26].

Unlike with supervised methods, comparative studies into unsupervised feature selection are very rare. In fact, to our knowledge there has only been one publication explicitly dealing with unsupervised feature selection for text data [16]. Generally, existing unsupervised methods tend to rely on heuristics that are informed by word frequency counts over the text collection. Although frequency can be a fair indicator of feature utility it does not consider contextual information. Ignoring context can be detrimental for text processing tasks because ambiguities in text can often result in poor retrieval performance. A good example is when dealing with polysemous relationships such as “financial bank” and “river bank”, where the word frequency for “bank” is clearly insufficient to establish its context and hence its suitability for indexing or case comparison.

In Textual Case-Based Reasoning (TCBR) research [22] the reasoning process can be seen to generally incorporate contextual information in two ways: as part of an elaborate indexing mechanism [2]; or as part of the case representation [24]. The latter requires simpler retrieval mechanisms, hence is a good choice for generic retrieval frameworks; while the former, although better at capturing domain-specific information, is more demanding of the retrieval process. A further distinguishing characteristic of TCBR systems is the different levels of knowledge sources employed to capture context [14]. These levels vary from deep syntactic parsing tools and manually acquired generative lexicons in the FACIT framework [7]; to semi-automated acquisition of domain-specific thesauri with the SMILE system; to automated clause extraction exploiting keyword co-occurrence patterns in PSI [25]. Of particular interest to this paper is the capture of co-occurrence based, contextual information within the case representation. Current research in this area is focused on feature extraction, which unlike feature selection aims to construct new features from existing features. Interest in this area has resulted in extraction techniques for both supervised (e.g. [25, 27]) and unsupervised settings (e.g. [4, 11]).

In text classification and applied linguistic research the problem of determining context is commonly handled by employing distributional clustering approaches. Introduced in the early nineties for automated thesaurus creation [18], distributional clustering has since been widely adopted for feature extraction with supervised tasks, such as text classification [1, 20]. Word clusters are particularly useful because contextual information is made explicit by grouping together words that are suggestive of similar context. Additionally, word clusters also provide insight into vocabulary usage across the problem domain. Such information is essential if representative features are to be selected. Of particular importance for word clustering are distributional distance measures. These measures ascertain distance by comparison of word distributions conditioned over a disjoint target set. Typically, class labels are the set of targets and so cannot be applied to unsupervised tasks.

The textual case retrieval system SOPHIA introduced a novel approach to combining distributional word clustering with textual case base indexing [17]. Here feature distance is measured by comparing word distributions conditioned on other co-occurring words (instead of class labels). Indexing is enabled by identifying seed features that act as case attractors. They argue that seed features are those that have non-uniform distributions having low entropy, referred to as specific word contexts. However the entropy based measure cannot distinguish between representative and diverse features even if they have specific contexts.

In structured CBR, clustering is commonly employed as a means to identify representative and diverse cases for casebase indexing. A good example is the footprint-driven approach [21] where a footprint case is: representative of its neighbourhood because of its influence; and diverse because its area of competence cannot be matched by any other case. This notion of identifying diverse yet representative cases has also been exploited in casebase maintenance [6, 23].

In summary, the representativeness and diversity of an entity can be measured by analysing its neighbourhood. In this paper the entity is the feature and representativeness and diversity are also important for feature selection. Central to feature neighbourhood analysis is a good distance metric. When features are words, the distance metric must take context into account. Distributional distance measures do this by exploiting word co-occurrence behaviour.

3 Frequency based Unsupervised Feature Selection

We first introduce the notation used in this paper to assist presentation of the different feature selection techniques. Let \mathcal{D} be the set of documents and \mathcal{W} the set of features, which are essentially words. A document d is represented by a feature vector, $\mathbf{x} = (x_1, \dots, x_{|\mathcal{W}|})$, of frequencies in d of words from \mathcal{W} [19]. In some applications, the frequency information is suppressed, in which case the x_i are binary values indicating the presence or absence of words in d . The main aim of unsupervised feature selection is to reduce $|\mathcal{W}|$ to a smaller feature subset size m by selecting features ranked according to some utility criterion. The selected m features then form a reduced word vocabulary set \mathcal{W}' , where $\mathcal{W}' \subset \mathcal{W}$ and $|\mathcal{W}'| \ll |\mathcal{W}|$. The new representation of document d is the reduced word vector \mathbf{x}' , which has length m .

Frequency counts are often used to gauge feature utility particularly in an unsupervised setting. The Term Contribution (Tc) is one such measure, showing promising results in [16]:

$$Tc(w) = \sum_{\substack{i,j \\ i \neq j}} F(w, d_i) * F(w, d_j)$$

$$F(w, d) = f(w, d) * \log_2 \frac{|\mathcal{D}|}{n}$$

Here F computes the tf*idf score which is a measure of the discriminatory power of a word given a document. Term frequency f is the within document frequency count of a feature and n is the number of documents containing feature w . Tc 's frequency based ranking and selection of features is the base line feature selection method used in this paper and we will refer to it as BASE (Figure 1).

```

m = feature subset size
BASE
  For each  $w_i \in \mathcal{W}$ 
    calculate  $Tc$  score using  $\mathcal{D}$ 
  sort  $\mathcal{W}$  in decreasing order of  $Tc$  scores
   $\mathcal{W}' = \{w_1, \dots, w_m\}$ 
  Return  $\mathcal{W}'$ 

```

Fig. 1. Feature selection with Tc based ranking.

Tc will typically rank frequent words appearing in fewer documents above those appearing in a majority of documents. In this way the BASE method will attempt to ignore overly frequent (or rare) features. Its main drawback is its inability to address the need for both representative and diverse features. This leads to selection of non-optimal dimensions that fail to sufficiently capture the underlying document content.

4 Role of Similarity for Unsupervised Feature Selection

A representative feature subset is one that can discriminate between distinct groups of problem-solving situations. In a classification setting, these groups are identified by their class labels and are typically exploited by the feature selection process. However in the absence of class knowledge, we need to identify and incorporate other implicit sources of knowledge to guide the search for features.

Similar problem situations are typically described by a similar set of features forming an operational vocabulary subset. When these subsets are discovered the search for features can be guided by similarity in problem descriptions. In particular knowledge

about feature similarity enables the search process to address both the need for representative and diverse features. The question then is how do we define similarity between features. A good starting point is to analyse feature co-occurrence patterns because features that are used together to describe problems are more likely to suggest the same operational vocabulary subset than features that rarely co-occur. In the rest of this section we look at how feature utility can be inferred from similarity knowledge extracted from feature co-occurrence patterns.

4.1 Feature Utility Measures

For a given word $w \in \mathcal{W}$, our first metric estimates the average pair-wise distance \overline{Dist} between w and its neighbourhood of k nearest word neighbours.

$$\overline{Dist}(w, \mathcal{A}, k) = \frac{1}{k} \sum_{w_N \in N_k(w, \mathcal{A})} Dist(w, w_N)$$

where N_k returns the k nearest neighbours of w chosen from $\mathcal{A} \subseteq \mathcal{W}$, and $Dist$ is the distance of w from its neighbour w_N . Lower values for \overline{Dist} suggests representative words that are centrally placed within dense neighbourhoods.

An obvious distance measure for words is to consider the number of times they co-occur in documents [19]. However the problem with such a straight forward co-occurrence count is that similar words can be mistaken as being dissimilar because they may not necessarily co-occur in the available document set \mathcal{D} . This is typical with text due to problems with sparseness [4].

4.2 Distributional Distance Measures

Often, related words do not co-occur in any document in a given collection, due to sparsity and synonymy. This limits the usefulness of similarity measures based purely on simple co-occurrence. Distributional distance measures circumvent this problem by carrying out a comparison based on co-occurrence with members of a separate disjoint target set [18]. Applied to text, the idea measures distances between word pairs by comparing their distributions conditioned over the set of other words. Since the conditioning is undertaken over a separate disjoint set, distances between non co-occurring word pairs need no longer remain unspecified.

Let us first demonstrate the intuition behind distributional distance measures by considering three words, a , b and c , and their fictitious word distribution profiles (see Figure 2). The x-axis contains a set of target events w_i , while the y-axis plots the conditional probabilities $p(w_i|w)$, for $w = a, b, c$. Comparison of the three conditional probability distributions suggests a higher similarity between a and b (compared to profiles of a and c). When target events on the x-axis are words, then a comparison between conditional probability distributions provides a similarity estimate based on word co-occurrence patterns. The next question then is how can we measure distance between feature distributions.

Let q and r be two features from \mathcal{W} whose similarity is to be determined. For notational simplicity we write $q(w_i)$ for $p(w_i|w = q)$ and $r(w_i)$ for $p(w_i|w = r)$,

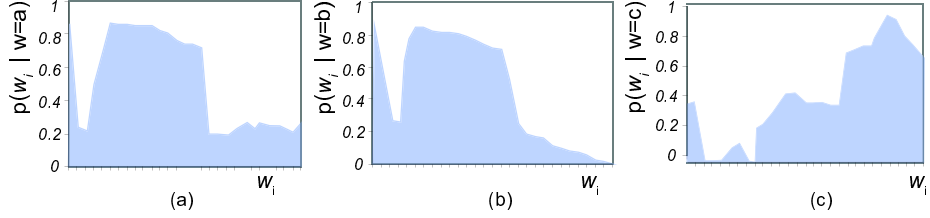


Fig. 2. Conditional probability distribution profiles.

where $w_i \in \mathcal{W} \setminus \{q, r\}$ and p denotes probabilities calculated from the training data \mathcal{D} . Research in linguistics has shown that the α -Skew metric is a useful measure of the distance between word distributions, when applied to the task of identifying similar noun pairs [13]. It is argued that the asymmetric nature of this distance measure is appropriate for word comparisons, since one word (e.g. ‘fruit’) may be a better substitute for another (e.g. ‘apple’) than vice-versa. Here we adopt this metric to compare word distributions and thereby determine the distance from word $q \in \mathcal{W}$ to word $r \in \mathcal{W}$.

$$Dist(q, r) = \sum_i r(w_i) \log \frac{r(w_i)}{q(w_i)}$$

is the Kullback-Leibler (KL) divergence, which is derived from information theory. It measures the average inefficiency in using $r(w_i)$ to code for $q(w_i)$ [3].

In our context, a large value of $Dist(q, r)$ would suggest that the word q is a poor representative of the word r , but not necessarily vice-versa. However, the $Dist$ is undefined if there are any words for which $q(w_i) = 0$, but $r(w_i) \neq 0$. The α -Skew metric avoids this problem by replacing q with $\alpha q + (1 - \alpha)r$, where the parameter α is less than one. In practice, our $Dist$ is the α -Skew metric with $\alpha = 0.99$, as suggested in [13].

5 Similarity based Unsupervised Feature Selection Methods

\overline{Dist} is the simplest measure that can be employed to rank features. However, we wish to use it so that a diverse yet representative set of features is discovered. This can be achieved in two alternatively ways: a GREEDY search that is locally informed; or a more globally informed CLUSTER-based search.

5.1 Greedy Search for Features

What we propose here is a greedy local search for the best feature subset. At each stage, the next feature is selected to be both representative of unselected features and distant from previously selected features. The feature utility score FUS_k , combines the average neighbourhood distance \overline{Dist} from both the selected and unselected feature neighbourhoods as follows:

$$FUS_k(w) = \frac{\overline{Dist}(w, \mathcal{S}, k)}{\overline{Dist}(w, \mathcal{U}, k)}$$

where $\mathcal{U} \subseteq \mathcal{W}$ contains previously unselected features, and $\mathcal{S} = \mathcal{W} \setminus \mathcal{U}$ contains previously selected features. Here the numerator penalises redundant features while the denominator rewards representative features.

The FUS_k based ranking and selection of features is the first unsupervised feature selection method introduced in this paper and we will refer to it as GREEDY (Figure 3). Unlike Tc , FUS_k 's reliance on distributional distances to capture co-occurrence behaviour undoubtedly makes it far more computationally demanding. However this cost is justified by FUS_k 's attempt to address the need for both representative and diverse features. One problem though is that GREEDY is a hill-climbing search where the decision to select the next best feature is informed by local information, hence it can select feature subsets that, although locally optimal, can nevertheless be globally non-optimal.

```

m = feature subset size
 $\mathcal{S} = \emptyset; \mathcal{U} = \mathcal{W}$ 
GREEDY
  Repeat
    Foreach  $w_i \in \mathcal{U}$ 
      calculate  $\text{FUS}_k$  score
    sort  $\mathcal{U}$  in decreasing order of  $\text{FUS}_k$  scores
     $w_j =$  top ranked feature in  $\mathcal{U}$ 
     $\mathcal{S} = \mathcal{S} \cup \{w_j\}$ 
     $\mathcal{U} = \mathcal{U} \setminus \{w_j\}$ 
  Until ( $|\mathcal{S}| = m$ )
   $\mathcal{W}' = \mathcal{S}$ 
  Return  $\mathcal{W}'$ 

```

Fig. 3. GREEDY method using FUS_k based ranking.

5.2 Clustered Search for Features

Clustering of words provides a global view of word vocabulary usage in the problem description space. Each cluster contains words that are contextually more similar to each other than to words outwith their own cluster. Partitioning the feature space in this way facilitates the discovery of representative features because each cluster can now be treated as a distinct sub-part of the problem description space.

We use a hierarchical agglomerative (bottom-up) clustering technique, where at the beginning every feature forms a cluster of its own. The algorithm then unites features with greatest similarity in small clusters and these clusters are iteratively merged until m number of clusters are formed. The decision to merge clusters is based on the furthest neighbour principle, where those two clusters with least distance between their most dissimilar cluster members are merged. Typically, this form of cluster merging leads to tightly bound and balanced word clusters.

from each cluster. The main steps appear in Figure 6. Here the number of clusters formed is equal to the desired feature subset size, m . This determines the stopping criterion for clustering. Like GREEDY, CLUSTER also addresses the need for representativeness and diversity, however, we expect CLUSTER to have an edge over GREEDY because its selection is influenced more globally.

```

 $m$  = feature subset size
 $\mathcal{W}' = \emptyset$ 
generate set of word clusters  $\{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ 
CLUSTER
  Foreach  $\mathcal{C}_i \subset \mathcal{W}$ 
     $w_j$  = feature with max  $FUS_{\mathcal{C}}$  in  $\mathcal{C}_i$ 
     $\mathcal{W}' = \mathcal{W}' \cup w_j$ 
  Return  $\mathcal{W}'$ 

```

Fig. 6. CLUSTER method using $FUS_{\mathcal{C}}$ based ranking.

6 Evaluation

We wish to determine the effectiveness of the two similarity-based searches for features, compared to the frequency-based search:

- GREEDY, introduced in this paper with ranking using FUS_k ¹ (Figure 3);
- CLUSTER, also introduced in this paper, exploits clustering and ranking using $FUS_{\mathcal{C}}$ (Figure 6); and
- BASE, the baseline with ranking on Tc (Figure 1).

The Tc -based ranking used by BASE is the only unsupervised method that has up to now been shown to perform better than the basic document frequency and the term strength methods [16]. We would hope to significantly improve upon the performance of BASE. Now the upper-bound for any unsupervised technique is its supervised counterpart, therefore, we also compare all our unsupervised methods with the standard IG-based SUPERVISED feature ranking and selection method.

It is generally harder to carry out empirical testing within a truly unsupervised setting compared to a supervised one. This is because, the absence of supervised labels calls for alternative sophisticated evaluation criteria, such as comparison of retrieval rankings or establishing measures of cluster quality. Instead, we applied our unsupervised methods on labelled data ignoring labels until the testing phase. Essentially we are exploiting class labels only as a means to evaluate retrieval performance which indirectly measures the effectiveness of the case representation. Note that we are not interested in producing a supervised classifier.

¹ In our experiments $k=15$ is used as FUS_k 's neighbourhood size.

Experiments were conducted on 4 datasets; all involving email messages. Each email message belongs to one mail folder. Here folders are the class labels. As in previous experiments we used the 20Newsgroups corpus of 20 Usenet groups [9], with 1000 postings (of discussions, queries, comments etc.) per group, to create 3 sub-corpus [24]: SCIENCE (4 science related groups); REC (4 recreation related groups) and HW (2 hardware problem discussion groups, one on Mac, the other on PC). With each sub-corpus the groups were equally distributed. A further set of 1000 personal emails, used for Spam filtering research forms the final dataset, USREMAIL, of which 50% are Spam [5].

We created 15 equal-sized disjoint train-test splits. Each split contains 20% of the full dataset, selected randomly, but constrained to preserve the original class distribution. All text was pre-processed by removing stop words (common words) and punctuation and the remaining words were stemmed. In the interest of reducing time taken for repeated trials, the initial vocabulary size was cut down to a subset composed of the 500 most and 500 least discriminating words (using IG). These 1000 words then form \mathcal{W} . An effective feature selection method should eliminate the non-discriminating words and assemble a representative and non-redundant combination of the discriminating ones.

The effectiveness of feature selection is directly reflected by the usefulness of the case representation obtained. Therefore, case representations derived by GREEDY, CLUSTER, BASE and SUPERVISED are compared on test set accuracy from a retrieve-only system, where the weighted majority vote from the 3 best matching cases are used to classify the test case. For each test corpus and each method the graphs show the test set accuracy (averaged over 15 trials) computed for representations with 5, 20, 40 and 60 feature subset sizes (Figure 7).

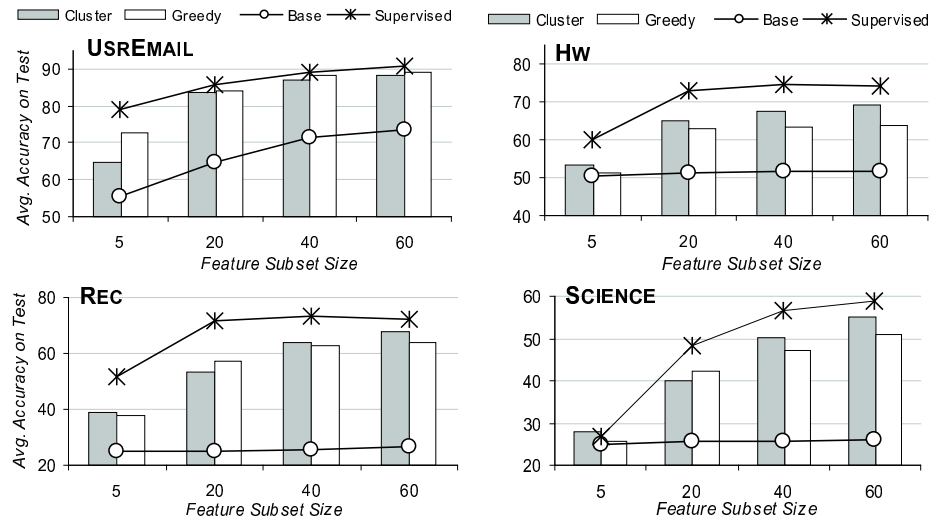


Fig. 7. kNN accuracy results for 4 datasets.

6.1 Results

Analysis of overall performance of SUPERVISED on the 4 datasets indicates that the classification of emails from USREMAIL as Spam or legitimate presents the easiest task. Here, SUPERVISED obtained 80% accuracy with just 5 features, compared with only 60% accuracy on the SCIENCE dataset. In all datasets except SCIENCE, we observe a steep rise in accuracy up to about 20 features, followed by a levelling-off as more features are added. This indicates that SCIENCE is the most difficult problem. Unlike USREMAIL, the other binary-classed HW dataset is harder, because similar terminology (e.g. monitor, hard drive) can be used in reference to both classes (i.e. PC and Apple Mac). Additionally, the same hardware problem can be relevant to both mailing lists, resulting in cross-posting of the same message.

We note that BASE performs very poorly on all datasets compared to GREEDY, CLUSTER and SUPERVISED. With the exception of the easiest problem (USREMAIL), it barely outperforms random allocation of classes and does not improve its performance as more features are added. Both GREEDY and CLUSTER clearly outperform BASE on all four datasets and improve their performance as the number of features increase. BASE's poor performance is explained by the fact that it selects features purely on the basis of term frequency information. Although frequent words will co-occur with many other words these co-occurrences will not necessarily be with similar words. Since similar words are indicative of similar areas in the problem space, BASE is not able to identify words that are representative of the problem space.

As expected, the SUPERVISED method achieves highest accuracy. Although both GREEDY and CLUSTER never match the performance of the supervised method, they make good progress towards the upper bound which it is expected to provide. Interestingly, CLUSTER improves relative to GREEDY as feature subset size increases and by 60 features, it is clearly better on the three more difficult datasets and only slightly worse on USREMAIL.

The fact that GREEDY is competitive with CLUSTER at lower feature subset sizes, but falls behind at higher subset sizes, suggests that GREEDY is more susceptible to overfitting. This effect can be seen in Figure 8, which plots training and test set accuracy for GREEDY and CLUSTER on the HW dataset. In these plots, data points lying significantly above the line $x = y$ are indicative of overfitting. Comparison of the scatter-plots confirms that GREEDY is more likely to overfit the selected feature subset to the training set.

6.2 Evaluation Summary

We checked the significance of observed differences between GREEDY and CLUSTER, using a 2-tailed t-test with a 95% confidence level for feature subset size, m equal to 60 (see Table 1). This test indicated that the superiority of CLUSTER over GREEDY was significant in all three datasets (bold font), but that of GREEDY on USREMAIL was not shown to be significant at this level. The superior scaling of CLUSTER can be explained by the fact that small optimal feature subsets need not be subsets of larger ones. GREEDY can be expected to suffer from overfitting at larger feature subset sizes, since the greedily chosen early features are locked in and cannot be altered to improve

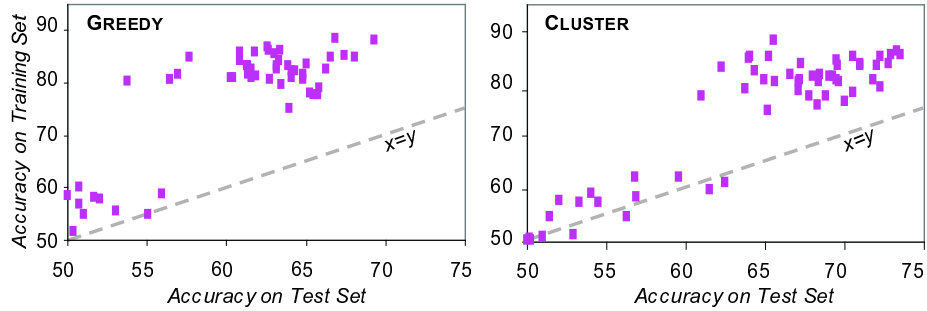


Fig. 8. Comparison of overfitting behaviour with GREEDY and CLUSTER on HW.

the global quality of a larger feature set. CLUSTER avoids this problem by dividing the entire feature set into as many clusters as required, before then selecting one keyword to represent each cluster.

Table 1. Results summary for feature subset size 60 according to significance.

60 features	USREMAIL	HW	REC	SCIENCE
GREEDY	89.3	63.7	64.0	51.0
CLUSTER	88.3	69.1	67.7	54.9
BASE	73.5	51.7	26.5	26.2
SUPERVISED	90.8	74.0	72.0	58.7

7 Conclusions

The methods introduced in this paper are particularly suited to generating case representations from free text data for unsupervised tasks. The novelty of these methods lies in their exploitation of distributional similarity knowledge to assess the utility of candidate features.

We introduce two unsupervised feature selection methods: GREEDY and CLUSTER. Key to both these methods is the selection of representative yet diverse features using similarity knowledge. Distributional distance measures are able to adequately capture feature similarity by addressing sparseness in co-occurrence data [18]. Evaluation results show significant retrieval gains with case representations derived by GREEDY and CLUSTER, over an existing proven method (BASE) from a previous comparative study [16]. It is also encouraging to report that both GREEDY and CLUSTER make good progress towards the upper bound which is provided by a standard supervised feature selection method. Generally GREEDY is able to generate good feature subsets early on in the search for features while CLUSTER’s global search approach consistently outperforms the GREEDY search with increasing feature subset sizes. This is due to the

locally informed GREEDY search identifying locally optimal, yet globally sub-optimal, subsets. Results also indicate that GREEDY is more susceptible to overfitting. We intend studying the influence of representativeness and diversity on overfitting, using a weighted form of FUS_C to control the balance between representativeness and diversity.

Previously we have shown that feature selection is a useful integral part of feature extraction when applied to text classification [24]. One difficulty that we have encountered since then, is that a majority of applications involving text are not necessarily supervised. This work is a first step towards resolving this shortcoming in existing feature discovery tools. Future work will look at combining feature selection with more powerful feature extraction methods to create comprehensive tools for text representation, indexing and retrieval for both supervised and unsupervised tasks.

References

1. Baker, L., McCallum, A.: Distributional clustering of words for text classification. In *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval* ACM Press (1998) 96–103
2. Bruninghaus, S., Ashley, K.: The role of information extraction for textual CBR. In *Case-Based Reasoning Research and Development: Proceedings of the 4th International Conference on CBR* Springer (2001) 74–89
3. Cover, T., Thomas, J.: *Elements of Information Theory*. John Wiley (1991)
4. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6) (1990) 391–407
5. Delany, S., Cunningham, P.: An analysis of case-base editing in a spam filtering system. In *Proceedings of the 7th European Conference on Case-Based Reasoning* Springer (2004) 128–141
6. Delany, S., Cunningham, P., Doyle, D., Zamolotskikh, A.: Generating estimates of classification confidence for a case-based spam filter. In *Case-Based Reasoning Research and Development: Proceedings of the 6th International Conference on CBR* Springer (2005) 177–189
7. Gupta, K., Aha, D.: Towards acquiring case indexing taxonomies from text. In *Proceedings of the Seventeenth International FLAIRS Conference* AAAI Press (2004) 307–315
8. Jarmulak, J., Craw, S., Rowe, R.: Genetic algorithms to optimise CBR retrieval. In Enrico Blanzieri and Luigi Portinale, editors, *Proceedings of the 5th European Workshop on CBR* Springer (2000) 137–149
9. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorisation. In *Proceedings of the Fourteenth International Conference on Machine Learning* (1997)
10. John, G., Kohavi, R., Pflieger, K.: Irrelevant features and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning* (1994) 121–129
11. Kang, N., Domeniconi, C., Barbara, D.: Categorization and Keyword identification of Unlabelled Documents. In *Proceedings of the 5th IEEE International Conference on Data Mining* (2005)
12. Lamontagne, L., Lapalme, G.: Textual reuse for email response. In *Proceedings of the 7th European Conference on Case-Based Reasoning* Springer (2004) 242–256
13. Lee, L.: On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001* (2001) 65–72
14. Lenz, M.: Defining knowledge layers for textual CBR. In *Proceedings of the 4th European Workshop on CBR* Springer (1998) 298–309

15. David D. Lewis and Kimberly A. Knowles. Threading electronic mail: A preliminary study. *Information Processing and Management* 33(2) (1997) 209–217
16. Liu, T., Liu, S., Chen, Z., Ma, W.: An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning* (2003) 488–495
17. Patterson, D., Rooney, N., Dobrynin, V., Galushka, M.: Sophia: A novel approach for textual case-based reasoning. In *Proceedings of the Nineteenth IJCAI Conference* (2005) 1146–1153
18. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* (1993) 183–190
19. Salton, G., McGill, M.: *An introduction to modern information retrieval*. McGraw-Hill (1983)
20. Slonim, N., Tishby, N.: The power of word clusters for text classification. In *Proceedings of the 23rd European Colloquium on Information Retrieval Research* (2001)
21. Smyth, B., McKenna, E.: Building compact competent case-bases. In Klaus-Dieter Althoff, Ralph Bergmann, and L. Karl Branting, editors, *Proceedings of the Second International Conference on Case-Based Reasoning* Springer (1999) 329–342
22. Weber, R., Ashley, K., Bruninghaus, S.: Textual case-based reasoning. *To appear in The Knowledge Engineering Review* (2006)
23. Wiratunga, N., Craw, S., Massie, S.: Index driven selective sampling for case-based reasoning. In *Case-Based Reasoning Research and Development: Proceedings of the 5th International Conference on CBR* Springer (2003) 637–651
24. Wiratunga, N., Koychev, I., Massie, S.: Feature selection and generalisation for textual retrieval. In *Proceedings of the 7th European Conference on Case-Based Reasoning* Springer (2004) 806–820
25. Wiratunga, N., Lothian, R., Chakraborty, S., Koychev, I.: Propositional approach to textual case indexing. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* (2005) 380–391
26. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorisation. In *Proceedings of the Fourteenth International Conference on Machine Learning* (1997) 412–420
27. Zelikovitz, S.: Mining for features to improve classification. In *Proceedings of Machine Learning, Models, Technologies and Applications* (2003)