



2004-01-01

# An analysis of case-base editing in a spam filtering systems

Sarah Jane Delany

*Dublin Institute of Technology, Sarahjane.Delany@dit.ie*

Padraig Cunningham

*Trinity College Dublin*

Follow this and additional works at: <http://arrow.dit.ie/scschcomcon>

 Part of the [Computer Sciences Commons](#)

## Recommended Citation

Delany, Sarah Jane and Cunningham, Padraig: An analysis of case-base editing in a spam filtering systems. *Advances in Case-Based Reasoning (Proceedings of the 7th. European Conference on Case Based Reasoning, ECCBR-04)*, edited by P.Funk and Gonzales Calero , LNAI 3155, pp.128-141.

This Conference Paper is brought to you for free and open access by the School of Computing at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact [yvonne.desmond@dit.ie](mailto:yvonne.desmond@dit.ie), [arrow.admin@dit.ie](mailto:arrow.admin@dit.ie), [brian.widdis@dit.ie](mailto:brian.widdis@dit.ie).



# An Assessment of Case-Based Reasoning for Spam Filtering<sup>\*</sup>

Sarah Jane Delany ([sarahjane.delany@comp.dit.ie](mailto:sarahjane.delany@comp.dit.ie))

*Dublin Institute of Technology, Kevin Street, Dublin 8*

Pádraig Cunningham ([padraig.cunningham@cs.tcd.ie](mailto:padraig.cunningham@cs.tcd.ie)) and Lorcan Coyle ([lorcan.coyle@cs.tcd.ie](mailto:lorcan.coyle@cs.tcd.ie))

*University of Dublin, Trinity College, Dublin 2*

**Abstract.** Because of the changing nature of spam, a spam filtering system that uses machine learning will need to be dynamic. This suggests that a case-based (memory-based) approach may work well. Case-Based Reasoning (CBR) is a lazy approach to machine learning where induction is delayed to run time. This means that the case base can be updated continuously and new training data is immediately available to the induction process. In this paper we present a detailed description of such a system called ECUE and evaluate design decisions concerning the case representation. We compare its performance with an alternative system that uses Naïve Bayes. We find that there is little to choose between the two alternatives in cross-validation tests on data sets. However, ECUE does appear to have some advantages in tracking concept drift over time.

**Keywords:** Case Base Reasoning, Spam Filtering

## 1. Introduction

Spam classification is a challenging task for a number of reasons. Not least of these is the fact that something of an “arms race” has developed between spammers and the filtering systems developed to identify spam. The content and structure of spam messages is constantly changing as spammers attempt to bypass the techniques used by the filtering systems to catch the spam. This poses a difficult challenge as systems need to identify and learn new types of spam as this arms race continues.

Lazy learning is good for dynamically changing situations. With lazy learning the decision of how to generalise beyond the training data is deferred until each new unseen instance is considered. In comparison to this, eager learning systems determine their generalisation mechanism by building a model based on training data in advance of considering any new unseen instances. In this paper we present E-mail Classification Using Examples (ECUE), a lazy learning system using CBR that seamlessly incorporates new training data.

---

<sup>\*</sup> This research was supported by funding from Enterprise Ireland under grant no. CFTD/03/219 and funding from Science Foundation Ireland under grant no. SFI-02IN.II111

Another challenge facing effective spam filtering using machine learning is dealing with large amounts of training data. A dynamic system which integrates new training data will require some means of managing the training data. CBR research offers a number of case-base management techniques to remove noisy and redundant training data and so effectively manage the size of the training data or case base over time. ECUE incorporates an effective case-based editing technique (Delany and Cunningham, 2004) which allows the number of training cases to remain at a manageable and efficient level.

The existing research on using a memory or case-based based approach (Androutsopoulos et al., 2000b; Sakkis et al., 2003) has a number of limitations. Firstly the evaluations are based on a restrictive data set incorporating legitimate email messages sent to a linguistics mailing list and “old-fashioned” spam emails that contain few of the obfuscations common in spam email today. Secondly all evaluations are static evaluations and do not take into account the changing nature of spam. In addition to static cross-validation tests, our evaluation of the approach presented in this paper includes dynamic evaluation of two independent datasets of over 10,000 email messages each, received over the period of a year.

This paper begins with an overview of other work using machine learning techniques in spam filtering in Section 2. Section 3 presents ECUE, our case-based spam filtering approach and describes the feature selection, case retrieval and case-base management techniques we use. The evaluations of ECUE and comparisons with Naïve Bayes (NB) are presented in Sections 4 and 5. Section 6 outlines directions for future work while our conclusions are presented in Section 7.

## 2. Spam Filtering and Machine Learning

Existing research on using machine learning for spam filtering primarily uses NB as the technique of choice (Androutsopoulos et al., 2000b; Pantel and Lin, 1998; Sahami et al., 1998; Androutsopoulos et al., 2000a) with many unpublished implementations reported on the Web. In addition to NB there has been work using Support Vector Machines (Androutsopoulos et al., 2000c; Drucker et al., 1999), Latent Semantic Indexing (Gee, 2003), and work using memory based classifiers (Androutsopoulos et al., 2000b; Sakkis et al., 2003; Cunningham et al., 2003). Sakkis et al. (2003) reported that their memory based classifier compared favourably to NB for spam filtering mailing lists and newsgroups while our preliminary findings (Cunningham et al., 2003) suggested that CBR would outperform NB.

Algorithms incorporating the NB classifier have proven to be among the most successful learners in the categorisation of text documents (Lewis and Ringuette, 1994) and are good for high dimension data, hence their popularity in spam classification.

## 2.1. NAÏVE BAYES FOR TEXT CLASSIFICATION

NB is a probabilistic classifier that can handle a large number of features that other machine learning techniques cannot. It is naïve' in the sense that it assumes that the features are independent. Consider a group of documents that are labelled as one of a set of classifications  $c_i \in C$ . Each document is described by a set of attributes  $\{a_1, a_2, \dots, a_n\}$  where  $a_i$  indicates the presence of that attribute in the document. The classification returned from a NB classifier for a new document is given in Equation 1.

$$c_{NB} = \operatorname{argmax}_{c_i \in C} P(c_i) \prod_j P(a_j|c_i) \quad (1)$$

Due to the significance of false positives (legitimate emails identified incorrectly as spam) in spam filtering, the NB classifier is not generally used in this simple argmax form. In practice the classification threshold is set to bias the classifier away from false positives (see Section 5.2).

The conditional probabilities can be estimated by  $P(a_i|c_j) = n_{ij}/n_j$  where  $n_{ij}$  is the number of times that attributes  $a_i$  occurs in those documents with classification  $c_j$  and  $n_j$  is the number of documents with classification  $c_j$ . This provides a good estimate of the probability in many situations but in situations where  $n_{ij}$  is very small or even equal to zero this probability will dominate, resulting in an overall zero probability. A solution to this is to incorporate a small-sample correction into all probabilities called the Laplace correction (Niblett, 1987). The corrected probability estimate is given by Equation 2, where  $n_{ki}$  is the number of values for attribute  $a_i$ . Kohavi et al. (1997) suggest a value of  $f = 1/m$  where  $m$  is equal to the number of training documents.

$$P(a_i|c_j) = \frac{n_{ij} + f}{n_j + fn_{ki}} \quad (2)$$

### 3. Case Based Spam Filtering

This section describes ECUE, the case-based system we have implemented for spam filtering. The description includes details of the feature extraction and representation, feature selection, case retrieval and case-base editing techniques we have used.

#### 3.1. FEATURE EXTRACTION

In order to identify the possible lexical features from the training set of emails, each email was parsed and tokenised. No stop word removal, stemming or lemmatisation was performed on the text before tokenisation. Email attachments were removed before parsing but any HTML text present in the email was included in the tokenisation. The datasets used were personal datasets, i.e. all emails in each dataset were sent to the same individual. Hence it was felt that certain headers may contain useful information so a selection of header fields, including the Subject, To and From headers were included in the tokenisation.

Three types of features were identified:

- word features (i.e. sequences of characters separated by white space or separated by start and end HTML tag markers),
- letter or single character features,
- statistical features, e.g. the proportion of uppercase or lowercase characters.

No domain specific feature identification was performed at this stage although work by Sahami et al. (1998) has indicated that the effectiveness of filters will be enhanced by their inclusion.

#### 3.2. FEATURE REPRESENTATION

In a CBR learner, examples in the training data are represented as cases in a case base. For the spam filtering domain, each training example is a case  $e_i$  represented as a vector of feature values,  $e_i = (f_1, f_2, \dots, f_n, s)$ . In text classification the lexical features are normally represented in one of two ways: (a) binary i.e. if the feature exists in the email,  $f_i = 1$ , otherwise  $f_i = 0$ , or (b) numeric where  $f_i$  is a number indicative of the frequency of occurrence of the feature in the email. Feature  $s$  represents the classification of the email, in our situation either *spam* or *non-spam*.

For numeric features the standard way to determine the value of  $f_i$  for feature  $x_i$  is to use the normalised frequency  $f_{ij}$  of the feature, see

Equation 3, where  $freq_{ij}$  is the number of times that feature  $x_i$  occurs in email  $e_j$ .

$$f_{ij} = \frac{freq_{ij}}{\max_k freq_{kj}} \quad (3)$$

In the evaluation presented here we used this normalised frequency for word and letter features (separate normalisations for each type) and simply the proportion calculated for the statistical features (which is between zero and one by definition).

A binary representation for the different types of feature is not so straightforward. For word features we use the simple existence rule that if the word exist in the email the feature value  $f_i = 1$  otherwise  $f_i = 0$ . However for letter features, almost all letters or characters will occur within an email so using the existence rule is not useful. For letter features we use the Information Gain (Quinlan, 1997) value of the feature as calculated during the feature selection process (see Section 3.3) to determine whether to set  $f_i = 1$ . If the normalised frequency of the letter feature is greater than or equal to the normalised frequency which returns the highest information gain for that letter then the feature value is set to one in the case representation, otherwise it is zero. Given that statistical features are also values between zero and one, this rule was also applied to features of this type to determine their binary representation.

A series of experiments to evaluate whether features should be represented as binary or numeric features was performed (as discussed in Section 4). It is more normal in text classification for lexical features to carry frequency information but the results of our evaluations showed no significant improvements were demonstrated when using numeric features over binary features.

In addition using binary features allowed use of an efficient case retrieval algorithm (discussed in Section 3.4) to improve performance.

### 3.3. FEATURE SELECTION

Tokenising 1000 emails results in a very large number of features, (tens of thousands of features). Feature selection is necessary to reduce the dimensionality of the feature space. Yang and Petersen's (1997) evaluation of dimensionality reduction in text categorisation found that Information Gain (IG) (Quinlan, 1997) was one of the top two most effective techniques for aggressive feature removal without losing classification accuracy. We calculated the IG of each feature and the top 700 features were selected. Our cross validation experiments,

varying between 100 and 1000 features across 4 datasets, indicated best performance at 700 features.

### 3.4. CASE RETRIEVAL

A CBR learner assigns a classification to a previously unseen example or target case by identifying and analysing the training cases that are most similar to it. Most of these classifiers use the  $k$ -NN algorithm to determine the  $k$  most similar training cases and then use these to classify the target case. The standard  $k$ -NN algorithm individually calculates the similarity of each case in a case base to the target case. This approach is quite inefficient in domains where there is feature-value redundancy and/or missing features in cases. Because our spam cases have both of these characteristics, and our feature representation was binary, we use an alternative similarity retrieval algorithm based on Case Retrieval Nets (CRNs) (Lenz et al., 1998).

A CRN is a memory structure which allows an efficient yet flexible retrieval of cases. They borrow ideas from neural networks and associative memory models. They are made up of the following components:

- Case nodes represent stored cases.
- Information Entity Nodes (IEs) represent feature-value pairs within cases
- Relevance Arcs link case nodes with the IEs that represent them. They have weights that capture the importance of the IE.
- Similarity Arcs connect IEs that refer to the same features, and have weights relative to the similarity between connected IEs.

The idea behind the CRN architecture is that a target case is activated by connecting it to the net via a set of relevance arcs and this activation is then spread across the net. Each of the other case nodes accumulates an activation score appropriate to its similarity to the target case. The case nodes with the highest activation are the most similar cases to the target case.

We implemented a CRN for case retrieval that was configurable for different  $k$ -nearest neighbour classifiers. As the features in our case representation are binary (implemented as boolean values), IEs are only included for features with a true value and similarity arcs are not needed. The relevancy arcs are all weighted with a weight of one.

Figure 1 depicts an example of our CRN for spam filtering. Our implementation of the CRN is similar in some respects to a Concept

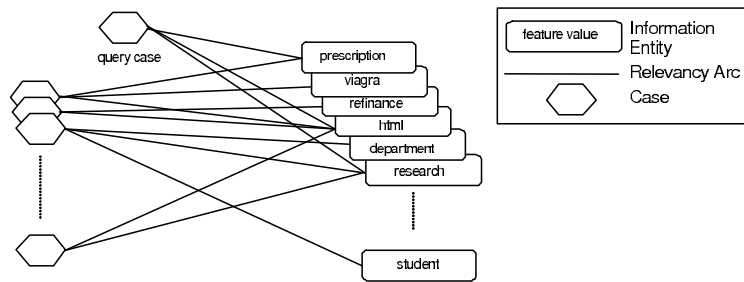


Figure 1. Case Retrieval Net

Network Graph (CNG) (Ceglowski et al., 2003) with thresholds set so that the activations are not spread beyond the first level of nodes.

### 3.5. CASE BASE MANAGEMENT

Research to date on machine learning for spam filtering has focused on static evaluations on datasets of manageable size. For instance, the LingSpam corpus (Sakkis et al., 2003; Drucker et al., 1999) contains 481 spam emails. Since a working spam filter could face this number of spam messages in a week there is a need to actively manage the training data. A key step in managing the training data is the case base editing process that deletes noisy examples and removes redundant cases from the case base.

Case base editing techniques involve reducing a case base or training set to a smaller number of cases while endeavouring to maintain or even improve the generalisation accuracy. There is significant research in this area (McKenna and Smyth, 2000; Wilson and Martinez, 1997; Brighton and Mellish, 2002). The case base editing technique that we used is Competence Based Editing (Delany and Cunningham, 2004) which uses the competence properties of the cases in the case base to identify noisy and redundant cases to remove.

The Competence Based Editing (CBE) technique initially builds a competence model of the case base identifying for each case its usefulness (represented by the cases that it contributes to classifying correctly) and also the damage that it causes (represented by the cases that it causes to be misclassified). These properties of each case are used in a two step process to identify the cases to be removed. The first stage is the competence enhancement or noise reduction stage which removes mislabelled or exceptional cases. The second stage is the competence preservation or redundancy reduction stage. Redundant cases are those that are in the centre of a cluster of cases of the same classification and



are not needed for classification. The advantage of our CBE technique applied to the spam domain is that it results in a conservative pruning of the case base which we found resulted in larger case bases but better generalisation accuracy (Delany and Cunningham, 2004).

## 4. Static Evaluation

Two types of evaluation of ECUE were performed. Firstly, evaluations on four static datasets of 1000 emails each were performed to determine which feature representation was appropriate for the cases in the case base and to evaluate how a case-based classifier would perform compared to a Naïve Bayes classifier. These evaluations are discussed in this section. The second type of evaluation, the performance of the case-based system, ECUE, in a dynamic situation is discussed in Section 5.

It is worth noting that rudimentary feature extraction techniques, described in Section 3.1, were used for all evaluations. To achieve a high performance comparable with existing commercial spam filtering systems, such as Spamassassin, “commercial grade” feature extraction techniques need to be implemented.

### 4.1. EXPERIMENTAL SETUP

The objectives of the static evaluations were two-fold, to determine the most appropriate case representation and to compare a case-based classifier with a Naïve Bayes classifier. Four datasets were used. The datasets were derived from two corpora of spam and legitimate email collected by two individuals over a period of approximately eighteen months up to and including December 2003 for Dataset 1 and up to and including January 2004 for Dataset 2. The legitimate emails in each corpus include a variety of personal, business and mailing list emails.

Four datasets of one thousand cases were extracted from each corpus. Each included five hundred spam emails and five hundred non-spam or legitimate emails. Datasets Feb-1 and Feb-2 consisted of 500 consecutive spam and legitimate emails received up to and including February 2003 while Datasets Nov-1 and Nov-2 consisted of 500 spam and legitimate consecutive emails received between February 2003 and November 2003. Given the evolving nature of spam it was felt that these datasets gave a representative collection of spam.

## 4.2. EVALUATION METRICS

Since False Positive (FP) classifications (legitimate emails classified incorrectly as spam) are much more serious than False Negative (FN) classifications (spam emails classified incorrectly as legitimate), accuracy (or error) as a measure of performance does not present the full picture. Two filters with similar accuracy may have very different FP and FN rates.

In previous work on spam filtering a variety of measures have been used to report performance. The most common performance metrics are precision and recall (Gee, 2003). Sakkis et al. (2003) introduce a weighted accuracy measure which incorporates a measure of how much more costly an FP is than an FN. Although these measures are useful for comparison purposes, the actual FP and FN rate are not visible so the true effectiveness of the classifier is not evident. For these reasons, where appropriate, we will use the rate of FPs, the rate of FNs, and the average within class error rate,  $AvgError = (FPRate + FNRate)/2$  as our evaluation metrics.

## 4.3. EVALUATION OF FEATURE REPRESENTATION

The objective of this evaluation was to determine whether a binary or numeric feature representation resulted in better generalisation accuracy. For each dataset we used 50 fold cross-validation, dividing the dataset into 50 stratified divisions or folds. Each fold in turn is considered as a test set with the remaining 49 folds acting as the training set.

For each test fold and training set combination we built two case bases, the first using a binary feature representation for the cases and the second using numeric features. Section 3.2 discusses how these representations were achieved. Each case base representation was then edited using CBE (see Section 3.5). We then calculated the performance measures of the test set against each of the four case base configurations; binary and numeric feature representation, edited and not edited. Confidence levels were calculated using McNemars test (Dietterich, 1998) between each combination of two case base configurations to determine whether significant differences existed. For each test example the result is recorded and, in order to compare case base configuration A with B, a table such as Table I is constructed.

The total number of test examples is  $n = n_{00} + n_{11} + n_{01} + n_{10}$ . If no difference exists between the two case base configurations then  $n_{10} = n_{01}$ . McNemars test requires the statistic in Equation 4 to be calculated. This statistic is distributed (approximately) as  $\chi^2$  with one degree of freedom.

Table I. McNemar's Results Table

$n_{00}$ = the number of examples misclassified by both case base configurations	$n_{01}$ = the number of examples misclassified by case base configuration A but classified correctly by B
$n_{10}$ = the number of examples classified correctly by case base configuration A but misclassified by B	$n_{11}$ = the number of examples classified correctly by both case base configurations

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (4)$$

The advantage that McNemars test has over the cross-validated paired t-test is a lower Type I error (the probability of incorrectly detecting a difference when no difference exists) but it also has good power (the ability to detect a difference where one exists) (Dietterich, 1998).

The results of our evaluations for each dataset and the average over all datasets are presented in Figure 2. It is worth noting that in the overall results, since we are calculating confidence levels using 4000 test examples, significance can be observed where the effect is quite marginal.

The results can be summarised as follows:

- (i) Case base editing improves performance for both case representations although the performance for numeric features is not as significant. (The difference for numeric features is not significant for any of the individual datasets and only measures as significant at 95% confidence level for the overall result whereas 3 of the 4 datasets demonstrate significant improvement for binary features at 95% level or higher with an overall difference significant at the 99.9% level).
- (ii) Case base editing also improves the performance on FPs with the binary feature representation showing higher levels of significance.
- (iii) Using numeric features on a full (non-edited) case base has significantly better performance (at 99% or higher) than binary features in 3 of the 4 datasets. However, the FP performance is not significantly different except in Dataset 1 (Feb) (at the 95% level).

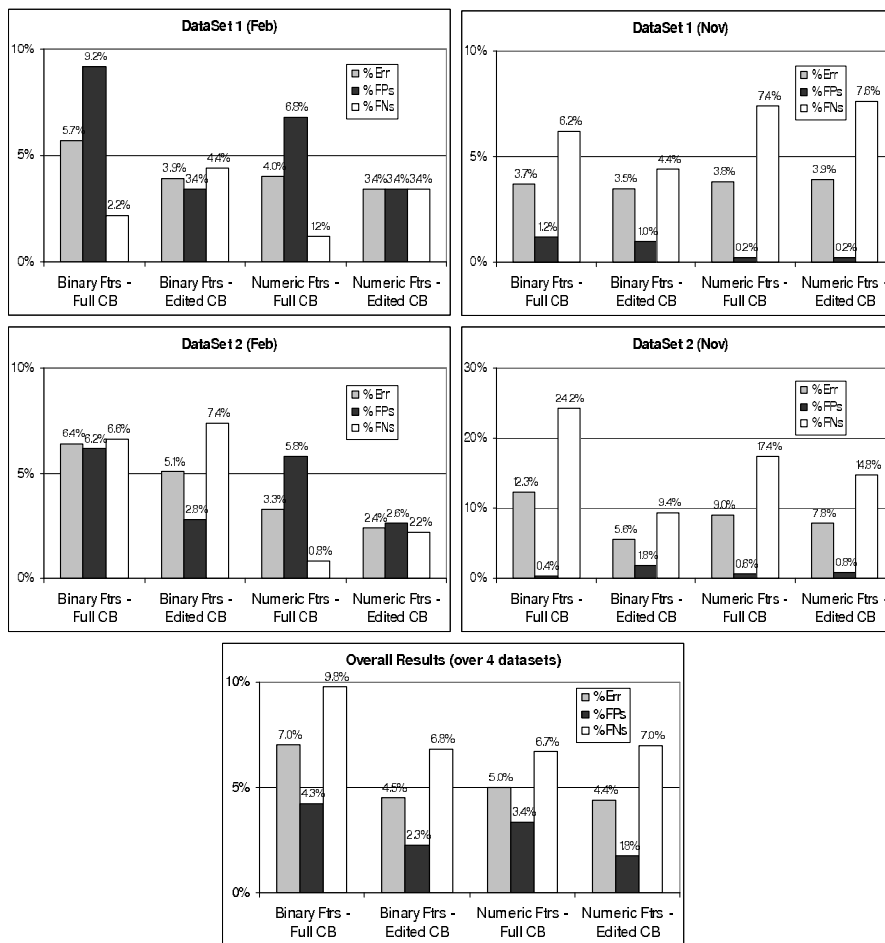


Figure 2. Results of evaluations for different feature representations

- (iv) The performance of an edited case base with binary features is not consistently significantly better than a full or edited case base with numeric features.

We then evaluated whether feature weighting improved performance or not. Each feature was weighted with a weight equal to the IG value of the feature identified during the feature selection process as suggested by Sakkis et al. (2003). We evaluated each case-base configuration with and without feature weighting. The results are shown in Figure 3.

The results can be summarised as follows:

- (i) Using feature weights has a negative effect on FP performance, with 5 of the 16 comparisons showing a significant difference (at 95% or higher) indicating that the rate of FPs is better without

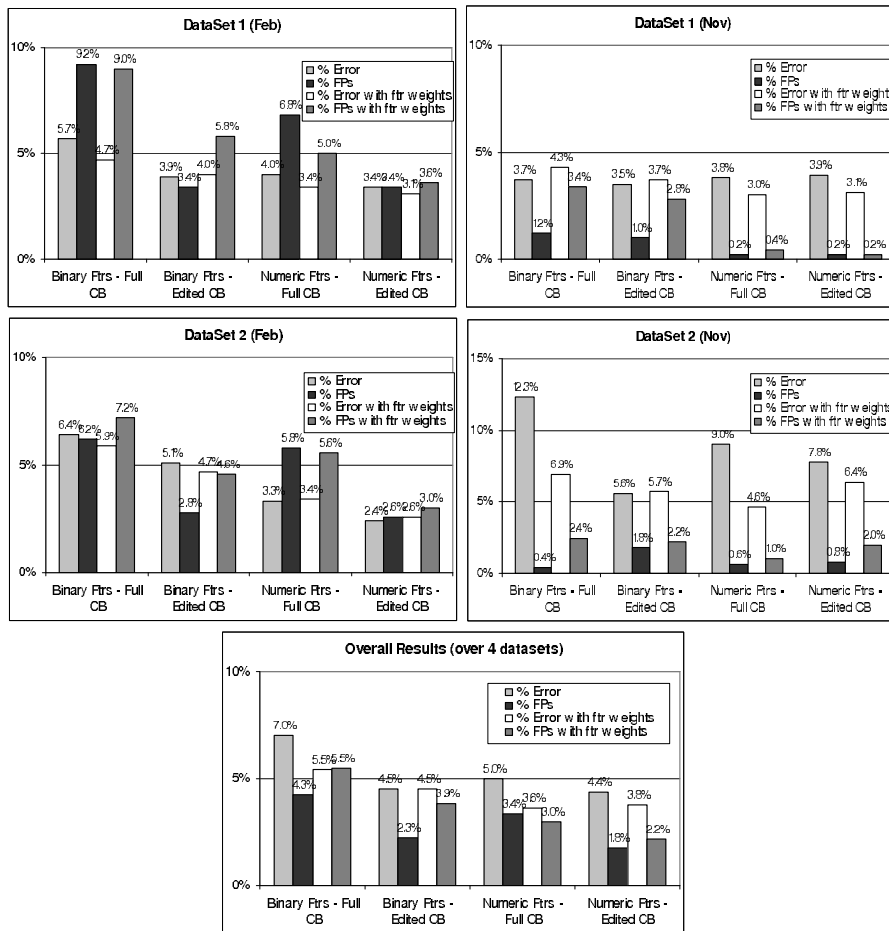


Figure 3. Results of including feature weighting

feature weighting. The remaining comparisons have no significant difference.

- (ii) Using feature weights significantly improves the accuracy only in Dataset 2 (Nov) where 2 of the 4 comparisons show a lower error rate, with feature weighting, significant at 99.9% or higher. The remaining datasets do not demonstrate any significant improvement using feature weighting.

Looking at the overall results for feature weighting, the best performance appears to be using numeric features on an edited case base. There is a significant difference in accuracy (at the 95% level) using feature weighting, but no significant difference in FP rate. A close second is using binary features on an edited case base with no feature weighting.

The accuracy is not as good (a difference of 0.7%) significant at the 95% level but the difference in FP rate is not significant. Therefore in terms of classification accuracy, numeric features on an edited case base using feature weights wins.

However, there is a considerable performance hit when using numeric features. The improvements in speed offered by the CRN are not realised for numeric features, only for symbolic or binary features. While numeric feature impact on response time at run-time the real performance hit comes at case base editing time which involves classifying each case in the case base multiple times. This is significant as it is clear that case based editing improves accuracy. A case base configuration that has long response time will have a large effect on the performance of a real time system. In the case of commercial applications like spam filtering, this cannot be ignored. For these reasons we chose to use a binary feature representation and an edited case base over the numeric features. We lose slightly in overall classification accuracy but the FP rate is not affected. We expect that including domain specific feature extraction methods will improve the accuracy.

#### 4.4. CBR vs NB

A key objective was to evaluate the generalisation accuracy of ECUE using a  $k$ -NN classifier with different values of  $k$ . A number of  $k$ -NN classifiers were evaluated. Once the  $k$ -NN classifier returns the cases that are determined to be closest to the query case, a voting algorithm is implemented to determine the classification of the query case.

For this evaluation we used a distance weighted voting algorithm. The vote returned for classification  $c_i$  for query case  $x_q$ , over the  $k$  nearest neighbours  $x_1, \dots, x_k$  using distance weighted voting is given in Equation 5 where  $1(a, b) = 1$  if  $a = b$ ,  $1(a, b) = 0$  if  $a \neq b$ ,  $w_j$  is given in Equation 6,  $f_i(x_j)$  is the value of feature  $i$  in case  $x_j$  and  $c_j$  is the classification of neighbour  $x_j$ . The classification with the highest vote is deemed to be the classification of the query case.

$$\text{Vote}(c_i) = \sum_{j=1}^k w_j 1(c_j, c_i) \quad (5)$$

$$w_j = \left( \sum_{m=1}^n |f_m(x_q) - f_m(x_j)| \right)^2 \quad (6)$$

The votes for spam and non spam are normalised and the spam normalised vote is compared with a set threshold. If the spam vote is greater than the threshold the query case is considered to be spam. By

varying the threshold from zero to one and plotting the resulting FP rate against one minus the FN rate an ROC curve (Bradley, 1997) can be plotted.

In order to compare ECUE with the current spam filtering technique of choice, a NB classifier was implemented using the algorithm described in Section 2.1. Normalising the probabilities returned by the NB algorithm and varying the threshold for a spam classification as described above allowed an ROC curve to be plotted for the NB classifier. The larger the area under the curve for an ROC curve, the better the classifier. The results of the best  $k$ -NN classifier, for an edited and non edited case base and the NB classifier are presented in Figure 4. To show the detail of the curve more clearly, only the top left hand corner of the graphs are presented.

The results presented below do not show that one classifier outperforms in all cases. NB seems to perform best in the February datasets while the  $k$ -NN classifier on an edited case base performs best in the November datasets.

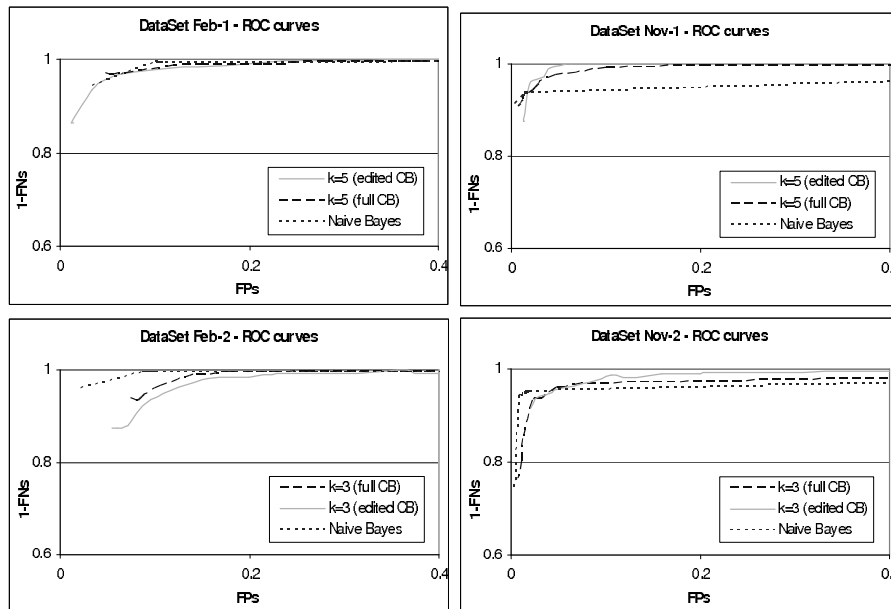


Figure 4. Results of comparing different classifiers

## 5. Dynamic Evaluation

We also evaluated how ECUE performs over a period of over a period of a year using two datasets of over 10,000 emails each, allowing the

Table II. Profile of the testing data

		Feb '03	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan '04	Total
Data	spam		629	314	216	925	917	1065	1225	1205	1830	576		<b>8902</b>
Set1	non spam		93	228	102	89	50	71	145	103	85	105		<b>1076</b>
Data	spam	142	391	405	459	406	476	582	1849	1746	1300	954	746	<b>9456</b>
Set2	non spam	151	56	144	234	128	19	30	182	123	113	99	130	<b>1409</b>

system to dynamically update its training data with examples of spam and legitimate email that were incorrectly classified.

### 5.1. EXPERIMENTAL SETUP

Two datasets were used. The datasets were derived from the same two corpora of email as described in Section 4.1. A case base of 1000 cases, 500 spam emails and 500 legitimate emails were set up in each case. This training data included the last 500 spam and non spam emails received up to and including February 2003 in the case of Dataset 1 and up to and including January 2003 in the case of Dataset 2. This left the remainder of the data for testing. Table 2 shows the number of spam and legitimate emails received each month for both datasets.

A case base was set up for each training dataset using binary word and letter features. The classifier selected was  $k$ -nearest neighbour with  $k = 3$ . Due to the fact that an FP is much more serious than an FN, the classifier used unanimous voting to determine whether the target case was spam or not. All neighbours returned had to have a classification of spam in order for the target case to be classified as spam. This corresponds to the leftmost point on the ROC curve in Fig. 4. This strongly biases the classifier away from False Positives.

Each case base was edited using the emphk-NN classifier with  $k = 3$  and the CBE editing technique. Our previous experiments with case editing using CBE and a unanimous voting classifier indicated that generalisation accuracy increased using an edited case base (Delany and Cunningham, 2004). Each email in the testing datasets, documented in Table II, was presented for classification in date order to closely simulate what would happen in a real-time situation.



## 5.2. EVALUATION METHODS

A number of experiments were performed, varying from making no updates to the original case base to updating the case base on a monthly, weekly and daily basis with those emails that were misclassified over the specified period. Our evaluation showed the best performance occurred when updating the case base on a daily basis with any emails misclassified that day. These results are presented in Figure 5.

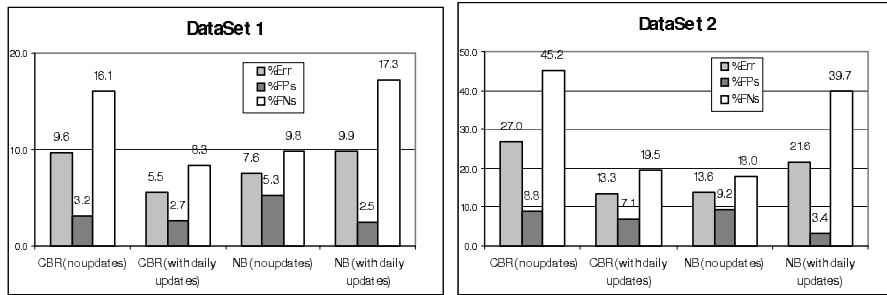


Figure 5. Results of evaluation over a period of time

The same experiments were performed using the NB classifier on unedited training data. Due to the significance of FPs, the NB classifier was configured to be biased away from false positives by setting the threshold equal to 1.0. Figure 5 also includes the results of using NB.

## 5.3. RESULTS

Although NB has a lower overall error rate over the datasets with no updating, the CBR system performs better in both datasets when dynamically updating the data to learn from incorrectly classified emails. It can be seen that daily updating of the training data with misclassified emails improves performance of the CBR system but has an overall detrimental effect on the NB classifier. NB with daily updates does improve the FP rate more than ECUE but the degradation of the FN rate has an overall negative effect on performance.

CBR only needs individual marker cases to construct its model whereas NB requires a full concept description. This may affect the performance of the NB classifier however the need to train the NB classifier on a full set of data presents its own set of data management problems.

It is worth noting that updating a system using NB with any new training data requires a separate learning process to recalculate the probabilities for all features. Updating a CBR system, such as ECUE,

with new training data simply requires new cases to be added to the case base.

## 6. Future Work

The focus of the research presented in this paper is on the case base classifier's ability to dynamically update the training data as new examples of spam and non spam are encountered. However, we envisage a hierarchy of learning within this domain where this continuous updating with misclassified emails is only the first level within three levels of learning.

As time passes and spam changes, the features selected for earlier training data may not be as predictive for new training examples. The second level of learning is to re-train the classifier by performing the feature selection process on the updated training data. This level of retraining may need to be performed infrequently, e.g. every month or every other month. The highest level of learning, performed even more infrequently than feature selection, is to allow new feature extraction techniques to be added to the system. For instance, when domain specific features are used in the system, new feature extraction techniques will allow new features to be included. The benefit of using a CRN for implementing the second and third levels of learning is that it can easily handle cases with new features. The fact that these features may be missing in old cases is not a problem.

Future work on our CRN will also incorporate CNG-type activation spreading to allow cases that do not include the actual selected features to influence the classification process.

## 7. Conclusions

The initial stage of this research focused on identifying the most appropriate case base configuration for a case-based classifier for spam filtering. Evaluations indicated that the best accuracy was provided by a numeric feature representation using feature weighting. However, this benefit compared with using binary features is marginal (0.7% in accuracy and not significant on the all important FP figures) and using numeric features has a significant impact on the speed of the system, particularly at the case editing stage. For this reason we are inclined to stay with the binary representation and seek to achieve accuracy improvements from improved feature extraction techniques. An alternative is to seek to speed up the editing of a case base that

incorporates numeric features using caching of similarity scores for case pairs.

Using CBR for spam filtering is certainly no worse than using NB. As techniques which utilise a probabilistic classifier to detect spam e-mail are already patented (US Patent, 2000), it is necessary to find other techniques which offer at least comparable results. In fact, our research suggests that CBR demonstrates better performance for learning over time than NB. CBR as a lazy learner offers significant advantages; it provides capabilities to learn seamlessly without the need for a separate learning process and facilitates extending the learning process over different levels of learning.

## References

- Androutsopoulos, I., J.Koutsias, G. Chandrinou, G. Paliouras, and C. Spyropoulos: 2000a, 'An Evaluation of Naive Bayesian Anti-Spam Filtering'. In: G. Potamias, V. Moustakis, and M. van Someren (eds.): *Procs of Workshop on Machine Learning in the New Information Age, ECML 2000*. pp. 9–17.
- Androutsopoulos, I., J.Koutsias, G. Paliouras, V. Karkaletsis, G. Sakkis, and C. Spyropoulos: 2000b, 'Learning to Filter Spam EMail: A comparison of a naive Bayesian and a memory based approach'. In: H. Zaragoza, P. Gallinari, and M. Rajman (eds.): *Procs of Workshop on Machine Learning and Textual Information Access, PKDD 2000*. pp. 1–13.
- Androutsopoulos, I., G. Paliouras, and E. Michelakis: 2000c, 'Learning to Filter Unsolicited Commercial EMail'. *Technical Report 2004/02, NCSR "Demokritos"*.
- Bradley, A.: 1997, 'The Use of the area under the ROC curve in the evaluation of Machine Learning Algorithms'. *Pattern Recognition* **30**, 1145–1150.
- Brighton, H. and C. Mellish: 2002, 'Advances in Instance Selection for Instance-Based Learning Algorithms.'. *Data Mining and Knowledge Discovery* **6**(2), 153–172.
- Ceglowski, M., A. Coburn, and J. Cuadrado: 2003, 'Semantic Search of Unstructured Data using Contextual Network Graphs'.
- Cunningham, P., N. Nowlan, S. Delany, and M. Haahr: 2003, 'A Case-Based approach to Spam Filtering that can track Concept Drift'. In: *ICCBR 2003 Workshop on Long-Lived CBR Systems*.
- Delany, S. J. and P. Cunningham: 2004, 'An Analysis of Case-Based Editing in a Spam Filtering System'. In: P. Funk and P. González-Calero (eds.): *7th European Conference on Case-Based Reasoning (ECCBR 2004)*, Vol. 3155 of *LNAI*. pp. 128–141, Springer.
- Dietterich, D. T.: 1998, 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms'. *Neural Computing* **10**, 1895–1923.
- Drucker, H., D. Wu, and V. Vapnik: 1999, 'Support Vector Machines for Spam Categorisation'. *IEEE Transactions on Neural Networks* **10**(5), 1048–1055.
- Gee, K. R.: 2003, 'Using latent semantic indexing to filter spam'. In: *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*. pp. 460–464, ACM Press.

- Kohavi, R., B. Becker, and D. Sommerfield: 1997, 'Improving simple Bayes'. In: *Proceedings of the 9th European Conference on Machine Learning (ECML 97)*. Springer Verlag.
- Lenz, M., E. Auriol, and M. Manago: 1998, 'Diagnosis and Decision Support'. In: M. Lenz, B. Bartsch-Sp&#246;rl, H. Burkhard, and S. Wess (eds.): *Case-Based Reasoning Technology, From Foundations to Applications*. pp. 51–90, Springer-Verlag.
- Lewis, D. and M. Ringuette: 1994, 'Comparison of Two Learning Algorithms for Text Categorisation'. In: *Procs of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR 94)*. pp. 81–93.
- McKenna, E. and B. Smyth: 2000, 'Competence-Guided Editing Methods for Lazy Learning'. In: W. Horn (ed.): *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence*. pp. 60–64, IOS Press.
- Niblett, T.: 1987, 'Constructing Decision Trees in Noisy Domains'. In: I. Bratko and N. Lavrac (eds.): *Progress in Machine Learning, Procs of 2nd European Working Session on Learning (EWSL 87)*. pp. 67–78, Sigma Press.
- Pantel, P. and D. Lin: 1998, 'SpamCop: A spam classification and organisation program'. In: *Procs of Workshop for Text Categorisation, AAAI-98*. pp. 95–98.
- Quinlan, J. R.: 1997, *C4.5 Programs for Machine Learning*. San Mateo, CA.: Morgan Kaufmann Publishers Inc.
- Sahami, M., S. Dumais, D. Heckerman, and E. Horvitz: 1998, 'A Bayesian Approach to Filtering Junk Email'. In: *Procs of Workshop for Text Categorisation, AAAI-98*. pp. 55–62.
- Sakkis, G., I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos: 2003, 'A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists'. *Information Retrieval* **6**(1), 49–73.
- USPatent: 2000, 'United States Patent 6,161,130.'
- Wilson, D. and T. Martinez: 1997, 'Instance Pruning Techniques'. In: *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*. pp. 403–411, Morgan Kaufmann Publishers Inc.
- Yang, Y. and J. Pedersen: 1997, 'A Comparative Study on Feature Selection in Text Categorization'. In: *ICML '97: Proceedings of the 14th International Conference on Machine Learning*. pp. 412–420, Morgan Kaufmann Publishers Inc.

