Multiple Sets of Rules for Text Categorization

Yaxin Bi[1,2], Terry Anderson[3], and Sally McClean[3]

[1]School of Computer Science,
Queen's University of Belfast, Belfast, BT7 1NN, UK
[2]School of Biomedical Science,
University of Ulster, Coleraine, Londonderry, BT52 1SA, UK
[3]Faculty of Engineering,
University of Ulster, Newtownabbey, Co. Antrim, BT37 0QB, UK
y.bi@qub.ac.uk, {tj.anderson, si.mcclean}@ulster.ac.uk

Abstract. This paper concerns how multiple sets of rules can be generated using a rough sets-based inductive learning method and how they can be combined for text categorization by using Dempster's rule of combination. We first propose a boosting-like technique for generating multiple sets of rules based on rough set theory, and then model outcomes inferred from rules as pieces of evidence. The various experiments have been carried out on 10 out of the 20-newsgroups – a benchmark data collection – individually and in combination. Our experimental results support the claim that "k experts may be better than any one if their individual judgements are appropriately combined".

1 Introduction

Appropriately combining evidence sources to form a more effective output than any of the individual sources has been investigated in many fields. The challenges of integrating evidence have gone under pattern recognition [1], sensor fusion [2], and a variety of ensemble methods [3]. Ensemble methods first solve a classification or regression problem by creating multiple classifiers that each attempts to solve the task independently, then use the procedure specified by the particular ensemble method for selecting or combining the individual classifiers. The two most popular ensemble methods include bagging and boosting [4]. In this research, we investigate an approach for combining multiple decisions derived from multiple sets of rules based on Demspter's rule of combination. Each set of rules is generated by a single rough sets-based inductive learning method, and is referred to as a classifier as in the boosting method. The advantage of our approach is its ability to combine multiple sets of rules into a highly accurate classification rule by modelling the accumulation of evidence.

We apply these methods to 10 out of the 20-newsgroups – a benchmark data collection – individually and in combination. Our experimental results show that the performance of the best combination of the multiple sets of rules on the 10 groups of the benchmark data can achieve 80.47% classification accuracy, which is 3.24% better than that of the best single set of rules.

T. Yakhno (Ed.): ADVIS 2004, LNCS 3261, pp. 263–272, 2004.

2 Rough Sets for Generating Text Classifier

Inductive learning can be loosely defined as *learning general rules* from specific instances [5]. In other words, inductive learning can be seen as a process of synthesizing mappings from a sample space consisting of individual instances. The result often is to reduce the space containing individual instances, leading to a new smaller space containing a set of representative instances, which serves the same role as the original one. By contrast, a rough sets-based inductive learning is aimed at learning a covering set of attributes in terms of a reduct, which is a minimal sufficient subset of a set of condition attributes. It preserves the dependency degree with respect to a set of decision attributes that has the same ability to discriminate concepts as a full set of attributes.

A rough set-based approach to inductive learning consists of a two-step process. The first step is to find multiple single covering solutions for all training instances held in a decision table. Specifically, given a set of condition attributes A and a subset $B \subseteq A$, a covering attribute set is found directly by computing its dependency degree with the decision attribute. The direct solution involves adding an attribute at a time, removing the attribute covered by the attribute set, and then the process is repeated until the dependency of B is equal to that of A. At the end of the induction of conjunctive attributes, more than one covering set – reduct – will be found.

The second step is to transform rules from multiple sets of reducts and weight each rule based on counting the identical attribute values. As a result, a number of rule sets will be produced, denoted by $\mathfrak{R} = \{R_1, R_2,\ldots, R_{|\mathfrak{R}|}\}$, where $R_\mathrm{i} =\{ r_{\mathrm{i}1}, r_{\mathrm{i}2,\ldots,} r_{|R\mathrm{i}|} \}$, $1 \le i \le |\mathfrak{R}|$. Each set of rules is called a *intrinsic* rule set, referred to as a classifier. It plays an independent role in classifying unseen instances. The relation between two sets of intrinsic rules is in disjunctive normal form (DNF) as are the rules within R_i. To examine the effectiveness of using multiple classifiers to classify unseen cases, our approach does not involve any rule optimzation between multiple sets of rules. More details about these algorithms can be found in [6].

A general DNF model does not require mutual exclusivity of rules within a set of intrinsic rules and/or between different sets of intrinsic rules. The DNF used in this context differs from the conventional way in which only one of the rules is satisfied with a new instance. Instead, all the rules will be evaluated on a new instance. Rules for either the same classes or different classes can potentially be satisfied simultaneously. In the case of different classes, conflicting conclusions occur. One solution for this is to rank rules for each class according to a class priority as established in some way, such as *information gain*, where the latest class is taken as the final class [7, 8]. The other solution is based on the majority voting principle, in which the conflicting conclusions are resolved by identifying the most satisfied rules [9]. In contrast, our approach makes use of as much rule-based evidence as possible to cope with conflicting conclusions through Dempster's rule of combination.

3 Demspter Shafer Theory of Evidence

The Demsper-Shafer (D-S) theory of evidence allows us to combine pieces of evidence from subsets of the frame of discernment that consists of a number of exhaustive and mutually exclusive propositions h_i, i = 1, .., n. These propositions form a universal set Θ. For any subset $H_i = \{h_{i1}, \ldots, h_{ik}\} \subseteq \Theta$, h_{ir} ($0< r \leq k$) represents a proposition, called a *focal element*. When H_i is a one element subset, i.e. $H_i = \{h_i\}$, it is called a *singleton*. All the subsets of Θ constitute powerset 2^Θ, i.e. $H \subseteq \Theta$, if and only if $H \in 2^\Theta$. The D-S theory uses a numeric value in the range [0, 1] to represent the strength of some evidence supporting a subset $H \subseteq \Theta$ based on a given piece of evidence, denoted by $m(H)$, called the *mass function*, and uses a sum of the strengths for all subsets of H to indicate the strength of belief about proposition H on the basis of the same evidence, denoted by $bel(H)$, often called the *belief function*. Notice that $bel(H)$ is equal to $m(H)$ if the subset H is a singleton [10].

4.1 Derive Mass Functions

In the previous section, we have given a general form of a text classifier, R. As stated in Section 2, given multiple reducts obtained from a collection of documents, the multiple corresponding sets of intrinsic rules will be generated, denoted by $\Re = \{R_1, R_2, \ldots, R_{|\Re|}\}$, where $R_i =\{ r_{i1}, r_{i2,\ldots,} r_{|Ri|}\}$ and $1 \leq i \leq |\Re|$. We now examine how to connect each classifier to a piece of evidence in order to formulate a mass function.

Let $\Theta = \{c_1, c_2, \ldots, c_{|\Theta|}\}$ be a frame of discernment, and let $R_i =\{ r_{i1}, r_{i2,\ldots,} r_{i|Ri|}\}$be a set of intrinsic rules as above. Given a test document d, if k rules are activated, i.e. $r_{ij+1}, r_{ij+2}, \ldots, r_{ij+q}$ where $1 \leq j$, $q \leq |R_i|$, then q decisions are inferred from R_i. Formally, this inference process can be expressed by $r_{ij+1}(d) \rightarrow h_1|stg_{j+1}$, $r_{ij+2}(d) \rightarrow h_2|stg_{j+2}$, …, $r_{ij+q}(e) \rightarrow h_q|stg_{j+q}$, where $h_s \in 2^\Theta{}_,$ $s \leq q$, and stg_{j+s} are rule strengths expressing the extent to which documents belong to the respective categories in terms of degrees of confidence. At the end of the inference process, a set of decisions will be obtained, and denoted by $H' = \{h_1, \ldots, h_q\}$, where $H' \subseteq 2^\Theta$.

With respect of the number of the rules fired, there are two situations, i.e. either $q = |R_i|$ or $q < |R_i|$. When $q = |R_i|$, his means all the rules in R_i are completely satisfied with a given document. We exclude this case since it may not play any role in classifying any documents. When $q < |R_i|$, $stg_{j+1} + stg_{j+2} + \ldots + stg_{j+q} < 1$, so H'does not constitute a frame of discernment. Therefore Demspter's rule of combination can be not applied. To use Demspter's rule of combination appropriately to pool all the conclusions to draw a final decision, we need a way to normalize the outcomes obtained. For convenience later, we define a function ϖ such that $\varpi(h_j) = stg_{i+j}$.

The normalization process starts by finding the duplicate conclusions within H', and then the corresponding rule strengths are added up, resulting in a new set of the decisions. Formally, for any two h_j, $h_{i+s} \in H'$, if $h_j = h_s$, $j \neq s$, then $\varpi(h_j) \leftarrow \varpi(h_j) + \varpi(h_s)$ and h_s is eliminated. After this processing, a set of decisions is reconstructed, denoted by $H = \{h_1, h_2, \ldots, h_{|H|}\}$, where $H \subseteq 2^\Theta$. The definition of a mass function for H is as follows:

Definition 5. A mass function is defined as m: $H \rightarrow [0,1]$. There are four different situations based on the inclusive relations between Θ and H.

1) if Θ ∈ *H*, then we define a mass function as follows:

$$m(\{h_i\}) = \frac{\varpi(h_i)}{\sum_{j=1}^{|H|} \varpi(h_j)} \quad (1\leq \text{i} \leq |H|) \tag{1}$$

2) if Θ ∉ *H*, and |*H*| < 2, then *H* ← *H* ∪ Θ and we define a mass function as follows:

$$m(\{h_i\}) = \varpi(h_i)\ (1\leq \text{i} \leq |H|\text{-}1) \tag{2}$$

$$m(\Theta) = 1 - \sum_{i=1}^{|H|-1} \varpi(h_i) \tag{3}$$

3) if *H* = Θ, and ϖ(*h*$_i$) ≠ 0 for any element *h*$_i$ ∈ *H* (1≤ i ≤|*H*|) then we define:

$$m(\{h_i\}) = \varpi(h_i)\ (1\leq \text{i} \leq |H|) \tag{4}$$

4) if *H* = Θ, and ϖ(*h*$_i$) = 0 for any element *h*$_i$ ∈ *H* (1≤ i ≤|*H*|) then we define: *m*(*H*) =1.

We have elsewhere provided a proof that the rule strength satisfies the condition of a mass function [6]. However, as in the first case above, some conclusions cannot be inferred from a specific piece of evidence, so these conclusions remains unspecified. Thus it is necessary to redistribute mass among known conclusions. We believe such a redistribution for the unknown state of hypotheses could be valuable in the coherent modeling and basic assignment of probabilities to potential hypotheses and for making decisions over an incomplete frame of discernment.

The second case means that the added Θ represents our ignorance about the unknown state of hypotheses in inference processes. It absorbs the unassigned portion of the belief after the commitment to *H*. The addition of ignorance about the likelihood of future hypotheses provides us with all the information we need for the inference process. This also means that the system does not require complete knowledge about all potential hypotheses since we represent an implicit set of unmodeled future hypotheses by including an additional Θ.

For the third case, the conclusions obtained are exactly the same as these integral hypotheses within Θ, through we directly replace strengths with a mass function.

The fourth case means that the conclusion obtained does not have knowledge about any individual hypotheses within the frame of discernment Θ, and its complement is an empty element. In this situation, we reassign its degree of total belief as 1.0.

4.2 Decision Fusion

Having defined the mass function, now we examine the problem of combining multiple classifiers. Suppose we are given multiple classifiers $\mathfrak{R} = \{R_1, R_2, \ldots, R_{|\mathfrak{R}|}\}$ and a set of categories $\Theta = \{c_1, c_2, \ldots, c_{|\Theta|}\}$, for a new document *d*, the category predictions of multiple classifiers R_1, R_2, …, $R_{|\Re|}$ will be applied to the document, resulting in $R_i(d) = H_i$. If only one of the classifiers is activated, such as $R_1(d) = H_1$, then H_1will be ranked in decreasing order. If the top choice of H_1 is a singleton, it will be assigned to the new document, otherwise lower ranked decisions will be considered for further selection. When K classifiers are activated, the multiple sets of classification decisions H_1, $H_{2, \ldots,}$ H_K are obtained, where $H_i = \{h_{i1}, h_{i2}, \ldots, h_{|H_i|}\}$, $H_i \subseteq 2^C$, and the corresponding rule strengths are $\varpi(H_i) = \{\varpi_i(h_{i1}), \varpi_i(h_{i2}), \ldots, \varpi_i(h_{|H_i|})\}$. After normalizing $\varpi(H_i)$ by using the method introduced in Section 4.1, we can obtain K mass functions, denoted m_1, m_2, …, m_K. With all of these outcomes along with the mass functions, we can gradually combine them to decide the final decisions using Equation (4) as follows:

$$[\ldots[m_1 \oplus m_2] \oplus \ldots \oplus m_K] \tag{5}$$

The combined results will be ranked and the final decision will be made by Equation (6). Notice that we are interested in the case where h_{ij} is a singleton, i.e. a single category, given H_i, so we have $m(h_{ij}) = bel(h_{ij})$ as stated in Section 3.

$$D(x) = H \text{ if } bel(H) = max_{H \in C}\, bel(H) \tag{6}$$

5 Experiment and Evaluation

There are a number of methods for evaluating the performance of learning algorithms. Among these methods, one widely used in information retrieval and text categorization is a pair of measures called precision and recall, and denoted by p and r respectively. Precision is the ratio of the true category documents to the total predicted category documents. Recall is the ratio of the predicated category documents to the true category documents. To compute overall performance on all the categories, we use the other measure the micro-averaged F_1 which is defined on the basis of the concepts of precision and recall as follows:

$$\text{micro-averged } F_1 = \frac{\sum_{i=1}^{m} F_1(c_i)}{m} \tag{7}$$

where

$$F_1(c_i) = \frac{2pr}{p + r} \tag{8}$$

The F_1 measure, initially introduced in [11], it combines Precision and Recall as a harmonic mean of the two measures. This measure will be used as an evaluation criterion in this experiment.

5.1 Newsgroup Data

For our experiments, we have chosen a benchmark dataset, often referred to as 20-newsgroup. It consists of 20 categories, and each category has 1,000 documents (Usenet articles), so the dataset contains 20,000 documents in total. Except for a small fraction of the articles (4%), each article belongs to exactly one category [12].

In this work, we have used 10 categories of documents, 10,000 documents in total, to reduce the computational requirements. The documents within each category are further randomly split into two groups, one consisting of 800 documents for training, and the other including 200 documents, but only 100 of the 200 documents are selected for testing.

5.2 The Experiment Results

For our experiments, we use information gain to select about 270 keywords after removing stopwords and applying stemming. By using the algorithms described in Section 2, ten reducts have been generated, and ten corresponding sets of intrinsic rules in turn have been constructed, denoted by R_0, R_1, …, R_9. In the following, we will not distinguish between the concepts of rules and reducts if no confusion occurs.

Prior to the experiments for evaluating the effectiveness of different combinations of reducts, we first carried out the experiments on individual reducts. Figure 1 presents the performance of each set of intrinsic rules. It can be seen that the best performing reduct is R_4.

To examine of the effectiveness of combined reducts in classification, we rank these reducts in decreasing order based on their classification accuracy, and then divide the 10 reducts into two groups to see the effect of the combinations of the reducts with high and low predictive accuracy, respectively. The first group consists of R_1, R_2, $R_{3,}$ $R_{4,}$ $R_{6,}$ and R_7, and the second group includes R_0, R_5, $R_{8,}$ and R_9. For the first group of reducts, we first take R_4 with the best performance, and then combine it with R_1, R_2, $R_{3,}$ $R_{6,}$ R_7. The combined results are denoted by R_{41}, R_{42}, $R_{43,}$ $R_{46,}$ R_{47} and they will be ranked. The best performing combination R_{46} is chosen, and in turn is combined with R_1, R_2, $R_{3,}$ R_7, resulting in ranked combinations of R_{461}, R_{462}, $R_{463,}$ R_{467}. As illustrated in Figure 2, in comparison with R_{46}, their classification accuracy performance has dropped. To examine the change in performance with the addition of more reducts, R_{461} and R_{463} are taken for further combinations, it is surprising that the performance increases. However, the performance degrades again with more reducts being combined. Therefore, it can be concluded that the combination of the best individual reduct with a reduct having a fairly modest performance is the best combination in achieving the highest predictive performance, and the performance of the best combination is 3.24% better than the best individual in the first group.

For the second group of reducts, we use the same method as the first group to examine behaviour of the combined reducts. We first take R_9 to combine with R_0, R_5, R_8. The performance of the combined reducts is shown in Figure 3. Following the same principle as above, we combine R_{59} with R_0, R_8, and combine R_{58} with R_0. As

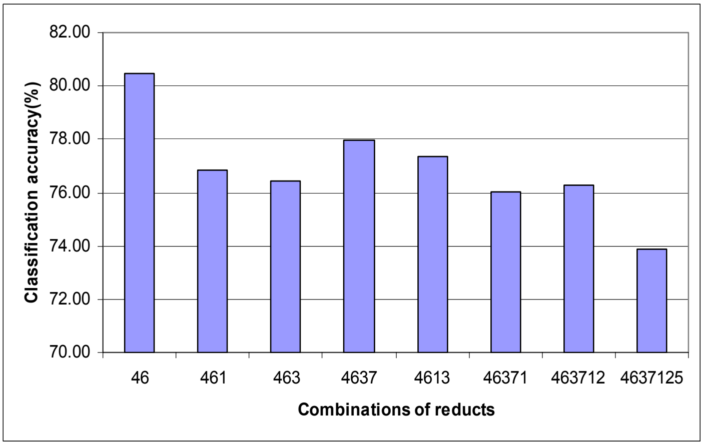

Fig. 1. The performance of different sets of intrinsic rules

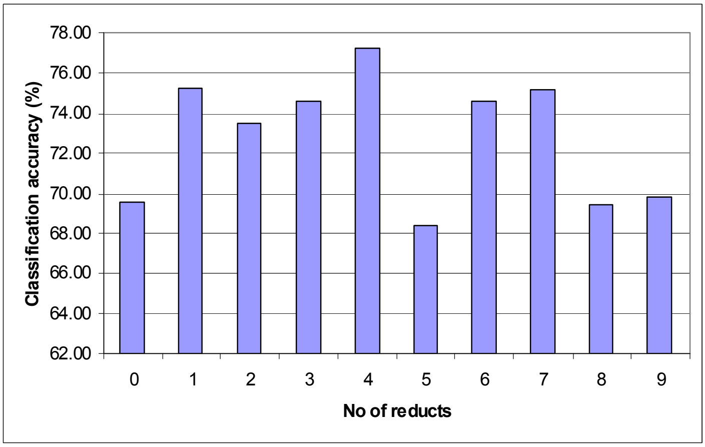

Fig. 2. The performance of the combined reducts from the first group

can be seen, the performance of these combinations drops by about 2% on average relative to R_{59} and R_{58}. However, when four reducts are combined, the performance increases again. A similar pattern to the first group of reducts is observed for this group. To analyze the effect of adding more reducts, we take the best performing R_4 and worst performing R_2 from the first group to combine with R_{5890}. The performance of this combination is not better than the previous one, this is a similar outcome to the first group of reducts.

To investigate how the performance improvement has been achieved when multiple reductes are combined, we base the outcome of the first group of reducts to

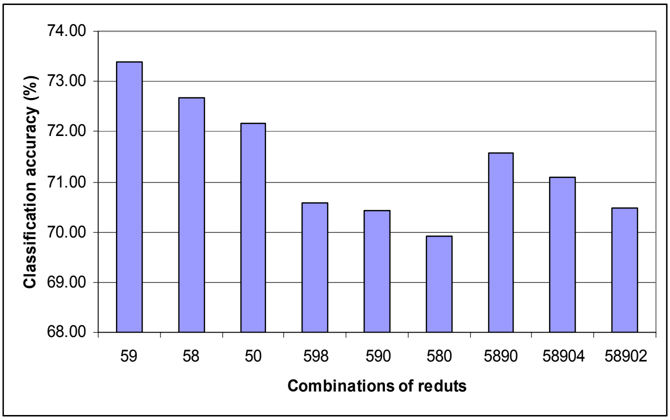

Fig. 3. The performance of the combined reducts from the second group

examine the performance variation on each category. Figure 4 presents a comparison between the performance of individual reducts and their combinations on each document category. It can be observed that with the exception of category 3, the predictive performance of the combinations is better that of individuals on all the document categories. However, it can also be conjectured that the performance of the combination of two reducts may be not better than that of two individuals, if there is a big margin between their performance on that category, e.g. category 3.

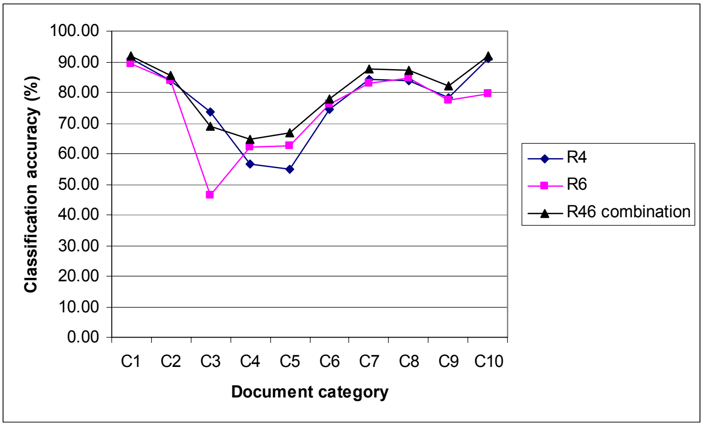

Fig. 4. The performance of reducts R4 and R6 vz the combined reducts 46

In Figure 5, we put the four combined reducts R_{46}, R_{463}, R_{4637} and R_{46371} on one graph to see the effect of the different combinations. The performance of the best com bination is mainly determined by the performance on categories C2, C3, C5, and C6.

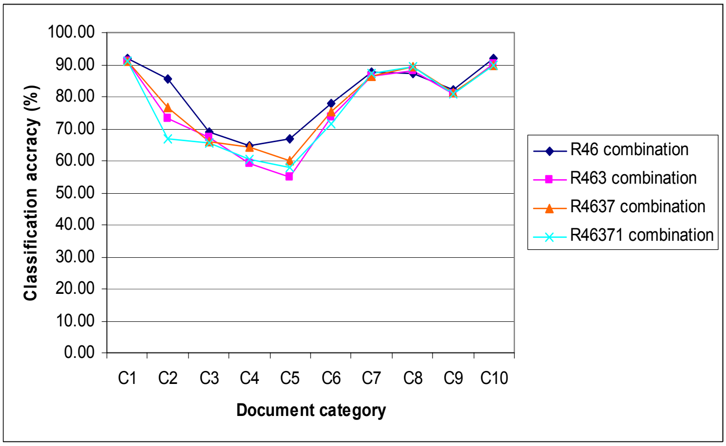

Fig. 5. The performance of the different combinations of reducts

Although R_{4637} performs better than R_{463} and R_{14637}, still worse than R_{46}, and its trend seems not provide an indication that it could be competitive to R_{46} on each category.

6 Conclusion

In this work, we have presented a boosting-like method for generating multiple sets of rules which is built on Rough Set theory, and a novel combination function for combining classification decisions derived from multiple sets of rule classifiers based on Dempster's combination rule. Preliminary experiments have been carried out on 10 of 20-newsgroups benchmark data, individually and in combination. We found that the combination which can achieve the highest predictive performance is a combination of two reduts of which one is the best, and the other should have reasonable predictive performance. The finding of which combining more 'weak learners' outperforms any individuals is consistent with the results obtained by Quinlan, and Freund and Schapire [13, 14].

To our knowledge, this work is the first attempt to use Dempster's rule of combination as a combining function for integrating multiple sets of decision rules in boosting-like methods and for text categorization. The experimental results have shown the promise of our approach. To consolidate this work, more comprehensive comparisons with the other combining functions of weighted linear and majority voting methods, and with previous results published in the literature will be carried.

Acknowledgements. This work was partly funded by the Centre for Software Process Technologies at the University of Ulster which is supported by the EU Programme for Peace and Reconciliation in Northern Ireland and The Border Region of Ireland (PEACE II).

References

1. Duda R, Hart P and Stork D (2001) Pattern classification. New York, NY: John Wiley & Sons, Inc.
2. Klein LA (1999) Sensor and data fusion concepts and applications. Society of Photo-optical Instrumentation Engineers. 2nd edition.
3. Dietterich, T. G. (2000a). Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, pp1-15, © Springer-Verlag.
4. Freund, Y and Schapire, R.(1996). Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference, pp148-156.
5. Mitchell, T. (1999). Machine learning and data mining. Communications of ACM. Vol. 42 (11).
6. Bi, Y. (2004). Combining Multiple Piece of Evidence for Text Categorization using Dempster's rule of combination. Internal report.
7. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Matero: Morgan Kaufmann.
8. Apte, C., Damerau, F., Weiss, S. (1994). Automated Learning of Decision Text Categorization. ACM Transactions on Information Systems, Vol. 12 (3), pp 233-251.
9. Weiss, S. M. and Indurkhya, N. (2000). Lightweight Rule Induction. Proceedings of the International Conference on Machine Learning (ICML).
10. Shafer, G.(1976). A Mathematical Theory of Evidence, Princeton University Press, Princeton, New Jersey.
11. van Rijsbergen, C. J. (1979). Information Retrieval (second edition). Butterworths.
12. Joachims, T. (1998). Text categorization With Support Vector Machines: Learning With Many Relevant Features. In Proceedings 10th European Conference on Machine Learning (ECML), Springer Verlag, 1998.
13. Quinlan, J, R. (1996). Bagging, boosting, and C4.5. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pages725–730, 1996.
14. Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), pp119–139.