THREE-PHASE TOURNAMENT-BASED METHOD FOR BETTER EMAIL CLASSIFICATION

Sabah Sayed, Samir AbdelRahman, and Ibrahim Farag

Computer Science Department, Faculty of Computers and Information
Cairo University, Egypt
{s.sayed, s.abdelrahman, i.farag}@fci-cu.edu.eg

ABSTRACT

Email classification performance has attracted much attention in the last decades. This paper proposes a tournament-based method to evolve email classification performance utilizing World Final Cup rules as a solution heuristics. Our proposed classification method passes through three phases: 1) clustering (grouping) email folders (topics or classes) based on their token and field similarities, 2) training binary classifiers on each class pair and 3) applying 2-layer tournament method for the classifiers of the related classes in the resultant clusters. The first phase evolves K-mean algorithm to result in cluster sizes of 3, 4, or 5 email classes with the pairwise similarity function. The second phase uses two classifiers namely Maximum Entropy (MaxEnt) and Winnow. The third phase uses a 2-layer tournament method which applies round robin and elimination tournament methods sequentially to realize the winner class per cluster and the winner of all clusters respectively. The proposed method is tested for various K settings against tournament and N-way methods using 10-fold cross-validation evaluation method on Enron benchmark dataset. The experiments prove that the proposed method is generally more accurate than the others.

KEYWORDS

Email Classification, Round Robin Tournament, Elimination Tournament, Clustering Tournament and Multi-Class Binarization.

1. INTRODUCTION

The internet user requires minimizing time and effort spent to read and organize his bunch of daily emails. Email classification becomes a crucial tool to tackle such a problem. Its goal is to automatically assign new incoming emails to one of the pre-defined set of classes (folders) based on their contents and properties [7]. Unfortunately, emails are written informally and they don't usually belong to a specific domain. This results in low performance of most classifiers [1, 13].
According to email classification literature, there are main five design aspects. (1) *Classifiers*: many classifiers have been probed such as Maximum Entropy (MaxEnt) [7], Winnow, Naive Bayes, Support Vector Machine (SVM) [1, 6]. (2) *Multi-Classification Methods*: two main types of methods have been evolved to classify an email into user folder topics, namely N-way [7] and tournament methods [12]. (3) *Classifier Evaluation Methods*: various classifier evaluation methods have been used such as F-measure [12] and Accuracy [1]. (4) *Corpus*: researchers usually work on private email dataset [5, 6] or they may use some benchmark corpora [1]. (5) *Corpus Trainig/Test Criterion*: various corpora' training/testing methods have been investigated such as corpus splitting [2], co-training [5], mixing all users' folders and choosing the biggest folders [12] and incremental time-based splitting [1].

DOI : 10.5121/ijaia.2012.3606

This paper proposes a tournament-based method to evolve email classification performance utilizing World Final Cup rules as its solution heuristics. These rules divide football teams into four groups of four teams. The first round is to apply round robin rule among each team pair to find out the winner team; the winner is the team which obtains the highest accumulated score. For the higher rounds, the elimination rule is applied between team pairs to eliminate the weaker ones. By these rules, each team plays with considerable number of other teams in a fast and exciting manner. Nonetheless, the distribution of the competitive teams and the repetition of each round rule through the game rounds are still questionable and differ according to the competition to ensure fair play.

Our proposed classification method passes through three phases. In the first phase, we cluster email folders using K-means algorithm [9, 11] with predefined number of cluster classes (3, 4 or 5). We use Un-weighted Pair-Group Method with Arithmetic Mean (UPGMA) [3, 9] function as our clustering algorithm similarity function of the email classes' tokens. Our aim is to distribute the nearest classes among the clusters such that each cluster has an optimal number of classes; we test 3, 4 and 5 as cluster sizes. By using clustering in the first phase, we are able to exclude all binary classifiers of distant classes from the round robin competition in order to resolve the trade-off between the classification errors and the accuracy. In the second phase, we train binary classifiers on each class pair to reduce the multi-classification errors [12]. In our experiments, we use Maximum Entropy and Winnows classifiers since they showed evidence in [1] that they provided competitive results. In the third phase, round robin tournament is applied for the first round on the cluster classes and then elimination tournament is used for the succeeding rounds. This tournament combination imitates the World Final Cup rules to allow each classifier competes with considerable number of others in a fast manner. Based on our experiments using Enron benchmark dataset[1] [10], we prove that the combination of the proposed three phases improve the email classifier performance against current email classification methods.

The remainder of this paper is organized as follows: section 2 gives a brief overview of related work in email classification. In section 3, we briefly give a background of the tournament classification methods which we use to compare our proposed approach with. In section 4, we state our design decisions in some issues such as data pre-processing and features construction. Section 5 explains the proposed email classification approach. Section 6 shows the experimental analysis and results. Finally, section 7 discusses the work contributions and some important future directions.

2. RELATED WORK

In [5], an email classification system is proposed which applies the co-training algorithm aiming to use a small set of labelled data to produce a weak initial classifier. It then boosts the classifier by using the remaining unlabelled data. The authors investigate how to apply this algorithm in the email domain. Their results show that the performance of co-training process depends on the classifier at hand. Also it concludes that Support Vector Machine (SVM) significantly outperforms Naive Bayes on the email classification task.

In [6], an email classification system is proposed which uses email temporal (time-related) features. Instead of using the traditional content-based features only, the system uses the timestamp of the email as a feature in email classification task. The authors find that when they use temporal features, SVM and Naïve Bayes outperform decision trees. Also they conclude that temporal features are not enough alone to obtain the best email classification results.

[1] http://www.cs.cmu.edu/~enron.

In [1], a benchmark case study of email classification is presented on both Enron and SRI[2] email datasets. The authors use four classifiers, namely Maximum Entropy, Naive Bayes, Support Vector Machine (SVM) and Winnow. Four experiments are carried out using the four classifiers. The experiments' accuracies show that SVM has the best outcomes while Naïve Bayes has the worst results. They also propose an incremental time-based splitting method for evaluating the classifiers performance. Unfortunately, they concluded that this evaluation method resulted in low accuracy results compared with the random training/testing splits.

In [12], a probabilistic tournament-like classification method is presented where the authors show that the tournament methods, namely round robin and elimination, improve the N-way method by 11.7% precision. Also, they noticed that round robin method [4] has more execution time and complex implementation than the elimination method does; fortunately, it acquired better results. They used an email corpus having 10 users and 50 folders by mixing all the folders and choosing the biggest 15 folders. None withstanding, the selected only 15 folders are not enough to present the relationships among the whole users and the related folders.

3. TOURNAMENT METHODS

Tournament classification against the classic N-way method is proposed in [12]. The classic N-way method is a probabilistic classification method in which all classes modelled based on the multinomial distribution and then the classification is performed using the maximum likelihood estimator. So, all classes are considered in the training and classification process. Current probabilistic classifiers use N-way concept in their implementation.

In their proposed method, [12] authors presents two methods the elimination tournament (ET) and round robin tournament (RRT). In ET method, every class is required to compete against a set of other classes. After each competition, the winner of the two classes remains in the next round competition and the loser is eliminated. The winner class in the last round is the optimal class to classify the incoming email message.

In RRT Method, scores are assigned for both classes after each competition. Then, every class accumulates the total scores of its competitions. The class which has the highest score is considered the optimal class for the incoming email message. In RRT method, a tie may be occurred in the final score table [12]; the tie occurs when more than two classes have equal final score in the competition table.

4. THE PROPOSED METHOD: DESIGN ASPECTS

4.1 The Classifier Aspects

Based on the analysis presented in [1] for various email classifiers, we decide to select Maximum Entropy and Winnow as our competitive classifiers. The classifier features are the email fields, namely the subject, the sender, the recipient, and the body content. We tokenize the subject and the body texts which are presented as a vector of lower-case word counts.

To map one multi-class problem to a set of binary-class problems, [4] presents two approaches, namely one-against-all and pair-wise approaches. In one-against-all approach, each class is assigned by one classifier such that the number of binary classifiers equals to the number of classes (C). Each classifier is trained on all the training data set where training examples of one

[2] http://www.ai.sri.com/project/CALO.

class are considered as positive examples and all the others as negative examples. In pair-wise approach, on the other hand, there is one classifier for each pair of classes (C) such that the number of classifiers equals to C(C-1)/2. A classifier of two classes is trained only on the related training data ignoring all other training records (emails). [4] proves that pair-wise binary classification approach makes each classifier use fewer training examples than those in one-against-all approach. The author in [4] concludes, then, that pair-wise approach improves the accuracy more than one-against all approach does.

In this paper, we map the multi-class problem to a set of binary-class problems using pair-wise approach. Furthermore, we use the following accuracy equation [1] for evaluating the competitive classifiers:

$$Accuracy = \frac{\#\text{ of correct classified examples}}{\#\text{ of all tested examples}} \qquad (1)$$

4.2 The Classification Method

We select tournament methods [12] to be our core research paradigm. Moreover, we compare our proposed method with 1) N-way method, implemented in Maximum Entropy and Winnow classifier paradigms, and 2) round robin tournament method (RRT), as it performs better than the elimination tournament method (ET) in [12].

In this paper, we break any RRT resultant tie by applying ET method among the tie classes to eliminate the losers and acquire the group winner. For example, if the tie has C1, C2 and C3 equal scores, we apply 1) the binary classifier of C1 and C2, then, 2) the binary classifier of the winner and C3 to find out the final winner.

4.3 The Email Dataset Evaluation

Currently there are some email corpora, such as Ling-spam, PU1, PU123A, Enron Corpus that have been released for the public [7]. We evaluate our experiments on Enron benchmark corpus for the email classification problem. The complete Enron dataset with its origin's explanation are available at Email Dataset (http://www.cs.cmu.edu/~enron) [10]

Table 1: Statistics of Enron Preprocessed User Folders

User	Number of folders	Number of message	Size of smallest folder (messages)	Size of largest folder
beck-s	16	1030	39	166
Farmer-d	16	3578	18	1192
kaminski-v	16	3871	21	547
Kitchen-l	16	3119	21	715
lokay-m	11	2489	6	1159
sanders-r	16	1077	20	420
williams-w3	12	2740	11	1398

Inspired by [1] pre-processing steps, we removed the basic folders "all documents", "calendar", "contacts", "deleted items", "discussion threads", "inbox", "notes inbox", "sent", "sent items" and "sent mail" and then flatten the folder hierarchies. Also we removed the attachments and the X-folder field from the email headers because it contains the class label. Further we decided to use

the email directories of seven Enron employees that have large number of emails, namely beck-s, farmer-d, kaminski-v, kitchen-l, lokay-m, sanders-r and williams-w3.

Additionally, from each Enron user directory, we chose the largest folders according to the number of emails. This resulted in sixteen folders for beck-s, farmer-d, kaminski-v, kitchen-l, and sanders-r. For lokay-m and williams-w3, we chose the largest folders which almost contain more than ten messages which are eleven and twelve folders in order. Our filtered dataset version contains around 17,904 emails (~89 MB) for 7 users distributed among 103 folders. Table 1 shows statistics on the seven resulting datasets. It is notable from the table that the least number of emails is larger than 1000 emails which is a suitable number for an email classifier. 10-fold cross-validation method is chosen for evaluating our experiment.

5. THE PROPOSED METHOD

5.1 The Clustering Similarity Function

The Un-weighted Pair-Group Method with Arithmetic mean (UPGMA) [3, 9] is a popular distance analysis algorithm. It uses a pair-wise distance function in which the distance between every two classes is calculated as the average distance (similarities) among all pairs of emails in these two classes excluding the internal similarities of the emails in each class.

In this research, we use UPGMA similarity function to calculate the similarities between each class pair, C1and C2 such that,

$$similarity\,(C1, C2) = \frac{\sum_{\substack{e_m \in C1 \\ e_n \in C2}} sim(e_m, e_n)}{N1 \times N2} \qquad (2)$$

where N1, N2 are number of emails in the first class C1 and the second class C2, $e_m (m = 1, 2, ..., N1)$, $e_n (n = 1, 2, ..., N2)$ are email documents in C1, C2 respectively and $sim(e_m, e_n)$ is the multiplication of the two email vectors.

Using Equation (2), the matrix of all classes is built in which each matrix element presents the similarity between two classes. To construct each group (cluster) of classes, we modify the number of classes in the K-means algorithm as follows,

$k = \#$ of Corpus Classes / Size , where size = 3, 4 or 5.

5.2 The Proposed Algorithm

Our core algorithmic steps can be enumerated as follows:

1. Pre-process the email corpus and tokenize each email body producing the list of tokens.
2. Construct the features' vectors from the message tokens list and message fields (section 4.1).
3. Cluster the classes of all emails into groups according to the clustering method (section 5.1); the aim is to result in the groups of similar classes.
4. Split the corpus into two separate portions; training and testing corpora according to cross-validation evaluation method (section 4.3).
5. For each training data email, output the binary classifiers (section 4.1); one classifier for each two different classes by using training data of only these two classes.

6. For the testing data emails, split them into groups of email classes generated from Step 3.
7. For the testing data email group, remove each email class label and use the related binary classifiers only.
8. For each testing email:
   a. Apply RRT method (section 3) within each group of classes (step 7) and find the winner class in that group.
   b. Apply ET method (section 3) between the winner classes from all groups.
   c. The ET method winner class is considered as a predicted label (class) of the testing email.
9. Repeat steps 6-8 for various group sizes; 3, 4 and 5 (section 5.1).

6. THE EXPERIMENTS

6.1 Experiment Settings

To evaluate the performance of our proposed email classification approach, we designed the following settings:

1. We examined various settings of grouping sizes, mainly 3, 4, and 5 classes using our clustering method (section 5.1). Additionally, we examined the grouping size equal to 4 using random class selection; we selected such a class size as it is a typical World Final Cup rule. Our aim of these settings is to verify: 1) clustering versus random grouping and 2) the adequate class group size.
2. We used the above settings to conduct two experiments validating our proposed method (section 5). Table 2 and Table 3 present the results for Maximum Entropy (MaxEnt) and Winnow classifiers respectively where the bold values highlight the best performance.
3. We used [8] as our toolkit environment for our implementation aspects.

6.2 Maximum Entropy (MaxEnt) Experiment

Table 2 shows that the accuracy orderly is 0.788, 0.789, 0.794, 0.795, 0.796 and 0.797 for N-Way, clustering with size 4, and 3, tournament, random grouping and clustering with size 5 respectively. Clustering with size 5 and random grouping methods surpass the others which highlights, in general, the superiority of the Grouping clue over the classic N-Way and Tournament methods.

Table 2: The Results Our Method with MaxEnt classifier

User Name	Random with Size=4	Clustering with Size=3	Clustering with Size=4	Clustering with Size=5	N-Way	Tournament
Beck-s	0.785	0.7824	0.7841	**0.7882**	0.7729	0.7881
Farmer-d	**0.7879**	0.7871	0.7845	0.7856	0.7799	0.7861
Kaminski-v	0.6911	0.6859	0.688	0.6823	0.6545	**0.6988**
Kitchen-l	0.7215	0.7198	0.7215	0.7206	0.7194	**0.7257**
Lokay-m	0.8088	0.8094	0.7704	**0.8102**	0.7993	0.8077
Senders-r	0.8278	0.8223	0.8207	0.8387	**0.851**	0.808
Williams-w3	0.9524	0.9532	**0.9539**	0.9544	0.9368	0.9528
Average	*0.796*	*0.794*	*0.789*	*0.797*	*0.788*	*0.795*

6.3 Winnow Classifier Experiment

Table 3 shows that the accuracy orderly is 0.178, 0.413, 0.417, 0.423, 0.425 and 0.427 for N-Way, tournament, random grouping and clustering with size 5, 3 and 4 respectively. This means that the weakest score is recorded by N-Way method and, on the extreme side, the clustering method with several settings achieve comparable best results.

7. DISCUSSION

According to Tables 2 and 3, the following conclusions could be outlined:
First, tournament classification method performs better than N-Way which is conformed with [12] results and claims.

Table 3: The Results Our Method with Winnow classifier

User Name	Random with Size=4	Clustering with Size=3	Clustering with Size=4	Clustering with Size=5	N-Way	Tournament
Beck-s	0.291	**0.3469**	0.3369	0.2958	0.1039	0.2804
Farmer-d	0.351	0.3546	0.3546	**0.3555**	0.0622	0.351
Kaminski-v	**0.1771**	0.1761	0.1757	0.1758	0.0308	0.1748
Kitchen-l	**0.2989**	0.2919	0.2919	0.2919	0.0431	0.2899
Lokay-m	0.4718	0.4718	0.4718	**0.4722**	0.0977	0.4718
Senders-r	0.4166	0.4166	0.4166	**0.4167**	0.0837	0.4166
Williams-w3	0.9104	0.9166	0.9438	**0.9537**	0.8212	0.9048
Average	*0.417*	*0.425*	*0.427*	*0.423*	*0.178*	*0.413*

Second, clustering and random grouping achieve better detailed results when they are combined with classic N-Way and tournament methods. Augmented by Winnow classifier, the grouping based methods acquire best scores of the whole seven users. However, they boost Maximum Entropy classifier for 4 out of 7 users.

Third, Table 2 shows that 2 out of 4 best grouping method scores go to clustering grouping with size equal to 5. Table 3 shows that 4 out of 7 best grouping method scores go to clustering grouping with size equal to 5.

Fourth, with careful analysis of the above observations, we realize that augmenting tournament classification with grouping via our proposed clustering method excludes all binary classifiers of distant classes from group competitions which reduces the classification errors and increases the classification accuracy.

Fifth, in spite of the fact that the best clustering-based results come from the users having large number of emails, such as Williams-w3, the random grouping method has significant effect on the users having small email numbers, such as Kaminski-v. Hence, the most adequate clustering algorithms with related similarity functions are still questionable.

Sixth, in this paper, we prove that our three-phase tournament method is more accurate than classic N-Way and tournament methods based on Enron benchmark. In the future, further dataset

benchmarks will be investigated. Furthermore, we evolved our three-phase tournament method without using any domain specific knowledge. As such we are thinking of probing ontological solutions.

REFERENCES

[1] Bekkerman, R., Mccallum, A. and Huang, G., (2004), "Automatic Categorization of Email into Folders : Benchmark Experiments on Enron and SRI Corpora", Science, 418: 1-23

[2] Diao, Y., Lu, H. and Wu, D., (2000), "A Comparative Study of Classification Based Personal E-mail Filtering", in Proceedings of PAKDD00 4th PacificAsia Conference on Knowledge Discovery and Data Mining, 1805:408-419

[3] Fayyad, P., Piatetsky-Shapiro, U. and Smyth, G., (1996), "From Data Mining to Knowledge discovery: An Overview":1-36

[4] Fürnkranz, J. (2002), "Round Robin Classification", The Journal of Machine Learning Research, 2 (4): 721-747

[5] Kiritchenko, S. and Matwin, S., (2001), "Email Classification with Co-Training", Machine Learning, CASCON

[6] Kiritchenko, S. and Matwin, S. and Abu-hakima, S., (2004), "Email Classification with Temporal Features", International Intelligent Information Systems: 523-533

[7] Li, P., Li, J. and Zhu, Q. (2005), "An Approach to Email Categorization with the ME Model", Artificial Intelligence, (2): 229-234

[8] McCallum, A. and Kachites, A., (2002), "MALLET: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu

[9] Murtagh F, (1984). "Complexities of Hierarchic Clustering Algorithms: The State of The Art". Computational Statistics Quarterly 1: 101–113.

[10] Shetty, J. and Adibi, J. (2004), "The Enron Email Dataset Database Schema and Brief Statistical Report," Distribution

[11] Slonim, N. and Tishby, N. (2001), "The Power of Word Clusters for Text Classification", Neural Computation, 1: 1-12

[12] Xia, Y., Liu, W. and Guthrie, L., (2005), "Email Categorization with Tournament Methods", in Lecture Notes in Computer Science Natural Language Processing and Information Systems, volume 351, Xia, Y., Liu, W. and Guthrie, L., (E. Springer), Berlin / Heidelberg: 150-160

[13] Youn, S. and Mcleod, D., (2006), "A Comparative Study for Email Classification", Group: 387-39