A Danger Theory Inspired Learning Model and Its Application to Spam Detection

Yuanchun Zhu and Ying Tan

Key Laboratory of Machine Perception (MOE), Peking University
Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China
{ychzhu,ytan}@pku.edu.cn

Abstract. This paper proposes a Danger Theory (DT) based learning (DTL) model for combining classifiers. Mimicking the mechanism of DT, three main components of the DTL model, namely signal I, danger signal and danger zone, are well designed and implemented so as to define an immune based interaction between two grounding classifiers of the model. In addition, a self-trigger process is added to solve conflictions between the two grounding classifiers. The proposed DTL model is expected to present a more accuracy learning method by combining classifiers in a way inspired from DT. To illustrate the application prospects of the DTL model, we apply it to a typical learning problem — e-mail classification, and investigate its performance on four benchmark corpora using 10-fold cross validation. It is shown that the proposed DTL model can effectively promote the performance of the grounding classifiers.

Keywords: artificial immune system, danger theory, machine learning, spam detection.

1 Introduction

The development of Artificial Immune System (AIS) is usually promoted by the proposal of novel Biological Immune System (BIS) paradigms. In recent years, a novel biological paradigm — Danger theory (DT), proposed by Matzinger [1], has become popular in explaining the mechanism of BIS. According to the DT, an immune response is not triggered by the detection of 'non-self' but the discovery of 'danger', and immunity is controlled by an interaction between tissues and the cells of the immune system. Although there are still debates on the relation between the DT and classical viewpoint, the DT does contain enough inspiration for building relative AIS [2]. Based on DT, novel AIS paradigms have been proposed and applied to web mining and intrusion detection. Secker et al. [3] presented a DT based adaptive mailbox, where high number of unread messages were defined as the source of danger. Aickelin et al. [4] gave thoughts about the way of building a next generation Intrusion Detection System (IDS) based on DT. In Ref. [5], the development and application of two DT based algorithms for intrusion detection, namely the Dendritic Cell Algorithm and the Toll-like Receptor Algorithm, were presented.

Y. Tan et al. (Eds.): ICSI 2011, Part I, LNCS 6728, pp. 382–389, 2011.

In this paper, we propose a DT based learning (DTL) model for combining classifiers. Mimicking the mechanism of DT, signal I, danger signal and danger zone are designed for machine learning task, and then the framework of the model is presented. Among the three components, danger zone is the most important one leading to the success of the DTL model. The danger zone defines a specific way of interaction between two grounding classifiers. To illustrate the application prospects of the DTL model, we apply it to a typical classification task — spam detection, and investigate its performance on four benchmark corpora using 10-fold cross validation. Experiments were conducted to analyze the effect of the danger zone, and compare the DTL model with classical machine learning approaches. It is shown that the proposed model can effectively promote the performance of the grounding classifiers.

The remainder of the paper is organized as follows. In Section II, we present how to transplant the three main concepts of DT into the machine learning task, and the framework of the DTL model is given. In Section III, the DTL model is implemented for spam filtering task. Section IV discusses our experimental results. The conclusions are presented in Section V.

2 Danger Theory Based Learning Model

The immune system has the ability of detecting and responding to dangerous things, according to DT. This phenomenon implies that the immune system can discriminate between danger and non-danger. Thus, it is logical to build a DT based Learning model for two-group classification problem. In this section, we concern with how to transplant the three main concepts of DT, namely Match — Signal I, Danger Signal and Danger Zone, into the field of machine learning.

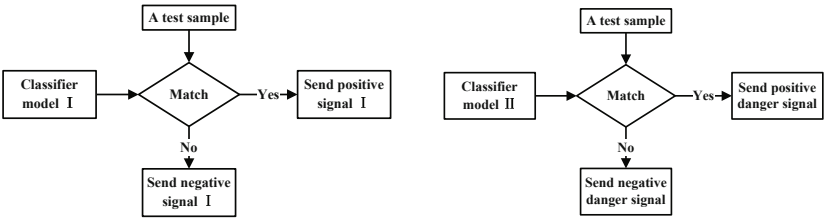
2.1 Generating Signals

The signal I is generated using the classifier I for each test sample in the DTL model. The process is depicted in Fig. 1(a). When the classifier I classify a test sample as positive class (match occurs), it will send a positive signal I to the sample. Otherwise, it will send negative one to the sample, if no match occurs.

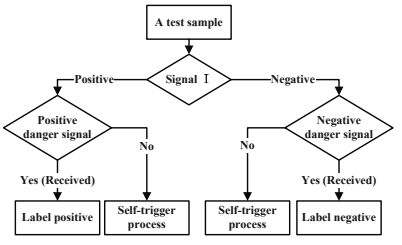
Figure 1(b) shows how a danger signal (Signal II) is triggered by the classifier II. Although the generating process of a danger signal seems to be quite similar as that of a signal I, the transmission coverage of a danger signal is quite different from that of a signal I. When a signal I is triggered, it will be sent only to the specific sample, upon which the signal is arisen. However, a triggered danger signal will be sent to all the test samples within the danger zone, besides the specific sample.

2.2 Classification Using Signals

This phase is the key procedure of the DTL model, which defines an immune based interaction between the two classifiers. As shown in Fig. 1(c), a test sample



(a) The process of generating signal I (b) The process of generating signal II



(c) Classification process using signals

Fig. 1. The process of classification using signals

is labeled only if the two signals that it received agree with each other. Otherwise, a self-trigger process is utilized to get the test sample classified.

The weighted result given by the interaction between the two classifiers is defined as Eq. 1.

$$E(x_i) = \sum_{x_j \in D} \delta(c_1(x_i), c_2(x_j)) K(d(x_i, x_j)), \tag{1}$$

where x_i and x_j are test samples, D denotes the test set, $c_1(x)$ and $c_2(x)$ are the two classifier models, $d(x_i, x_j) = \parallel x_i - x_j \parallel$ is the distance between two samples, $K(z)$ is defined in Eq. 2, and $\delta(y_1, y_2) = 1$, if $y_1 = y_2$, and 0 otherwise.

$K(z)$ defines the effect of the danger zone as follows.

$$K(z) = \begin{cases} 1 & if\ z \leqslant \theta \\ 0 & otherwise \end{cases}, \tag{2}$$

where θ is the size of the danger zone.

After obtaining the weighted result $E(x_i)$, the sample x_i can get its class label using Eq. 3.

$$L(x_i) = \begin{cases} c_1(x_i) & if\ E(x_i) \geqslant 1 \\ f(x_i) & otherwise \end{cases}, \tag{3}$$

where $f(x)$ denotes the class label given by the self-trigger process.

Algorithm 1. Framework of the DTL model

```
Select two uncorrelated classifiers: classifier I, II;
Train the two classifiers respectively on train corpus;
for each sample x_i in test corpus do
  Trigger a signal I upon x_i using classifier I and send the signal to x_i;
  Trigger a danger signal upon x_i using classifier II and send the signal to the test
  samples within the danger zone of x_i;
end for
for each sample x_i in test corpus do
  if x_i has received a positive signal I then
    if x_i has received a positive danger signal then
      Label x_i as positive class;
    else
      Call self-trigger process;
    end if
  else
    if x_i has received a negative danger signal then
      Label x_i as negative class;
    else
      Call self-trigger process;
    end if
  end if
end for
```

Self-Trigger Process: for the test samples which get conflict results from classifier I and II, a self-trigger process is designed. An intuitional thought is to get the sample activated using its nearest neighbor. Thus, the Nearest Neighbor (NN) approach is applied in this phase [6]. In future work, we intend to incorporate other approaches for self-trigger process into the DTL model and compare their performance.

2.3 The Framework of the DTL Model

Algorithm 1 summarizes the framework of the DTL model, in which two grounding classifiers interact through two signals. In the model, two grounding classifiers are first chosen and trained independently. Then the signal I and the danger signal, simulating the signals in the DT, are triggered upon each test sample utilizing the two classifiers. Finally, each test sample gets labeled by considering the interaction between the two classifiers.

2.4 Analysis of the DTL Model

For any machine leaning model, the essence of it is the conditional probability $P(y_k|x_i)$ of class y_k that it computes for each test sample x_i. In the DTL model, a test sample x_i gets a label y_k in two cases as follows.

(1)The two grounding classifiers give a consistent label y_k to the sample x_i: Suppose the two grounding classifiers are conditionally independent, given a test

sample x_i. Then the probability $P(y_k \mid x_i, case_1)$, which denotes the probability that the two grounding classifiers give consistent label y_k to the sample x_i, is computed as follows.

$$\begin{aligned} &P(y_k \mid x_i, case_1) \\ \leqslant &P(c_1(x_i) = y_k \mid x_i) \cdot \sum_{x_j \in D} P(c_2(x_j) = y_k \cap K(\parallel x_i - x_j \parallel) = 1 \mid x_i, x_j). \end{aligned} \quad (4)$$

(2)There is confliction between the two grounding classifiers, and the self-trigger process gives the label y_k to the sample x_i. The probability $P(y_k \mid x_i, case_2)$, which denotes the probability that this case may happen, is defined as follows.

$$P(y_k \mid x_i, case_2) = P(E(x_i) = 0 \cap f(x_i) = y_k \mid x_i). \quad (5)$$

Following the above analysis, the probability $P(y_k \mid x_i)$, computed by the DTL model, is presented in Eq. 6.

$$P(y_k|x_i) = P(y_k \mid x_i, case_1) + P(y_k \mid x_i, case_2), \quad (6)$$

which can be expanded using Eqs. 4 and 5.

3 Filter Spam Using the DTL Model

3.1 Feature Extraction

At the beginning, terms are selected according to their importance for classification, which can be measured by Information Gain (IG) [7].

Bag-of-Words (BoW), also referred to as vector space model, is usually utilized as the feature extraction approach for spam filtering [8]. In BoW, an email is transformed into a d-dimensional vector $\langle x_1, x_2, \ldots, x_d \rangle$ by calculating occurrence of previously selected terms. For BoW with Boolean attribute, x_i is assigned to 1 if t_i occurs in the e-mail, or it is assigned to 0 otherwise. In our experiments, 300 features were selected by using IG, and a BoW with Boolean attribute was applied to the feature extraction phase.

3.2 Selection of Classifiers

Support Vector Machine (SVM) and Naive Bayes (NB) are chosen as the two grounding classifiers of the DTL model, as they are two of the most prevalent and effective classifiers especially for spam filtering [9, 10].

3.3 Performance Measures

To validate the effectiveness of the proposed DTL model, two overall performance measures were adopted in our experiments, namely accuracy and F_β measure [8]. The two components of F_β measure are also given in the experiment results.

4 Experiments of Spam Detection

Experiments were conducted on four benchmark corpora PU1, PU2, PU3, and PUA[1] [11], using 10-fold cross validation. The corpora have been preprocessed when published, by removing attachments, HTML tags, and header fields except for the subject. The details of the corpora can be found in Ref. [11].

4.1 The Effects of the Danger Zone

The specific interaction between the two grounding classifiers is implemented by the design of the danger zone. To some extent, the success of the DTL model lies in a proper danger zone design and an optimal size of the danger zone. In this subsection, we investigate the impact of the danger zone size on the overall performance of the DTL model. Experiments of the DTL model with different danger zone size were conducted on PU1, using 10-fold cross validation. The results are depicted in Fig. 2(a), which shows the variational performance of the DTL model, as the size of the danger zone growing larger. At initial stages, the accuracy and F_1 measure increases as the size of the danger zone is enlarged. Then, the performance of the DTL model peaks at a size of 20. After that, the performance declines as the size growing even larger.

4.2 Comparison Experiments

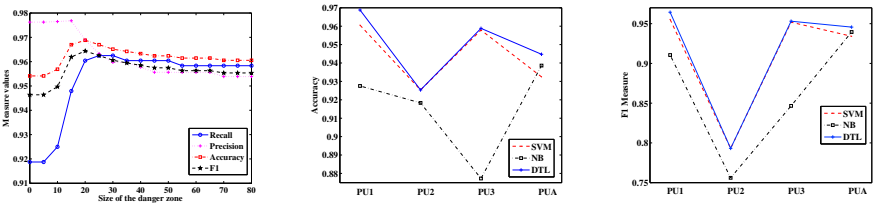
Comparison experiments were conducted on four benchmark corpora PU1, PU2, PU3, and PUA to validate the proposed DTL model, using 10-fold cross validation. As the four corpora have already been preprocessed when published, our experiments began at the phase of feature extraction. First, 300 discriminative words were selected by using the IG method. Then based on this, each e-mail was transformed to a 300-dimensional feature vector. Finally, the two grounding classifiers were built from the feature vector set.

In the experiments, two performance measures — accuray and F_1 measure were adopted as mentioned in Section 3.3. Fig. 2(b) depicts the comparison of accuracy among SVM, NB and DTL, while Fig.2(c) shows the comparison of F_1 measure among the three approaches. More details on the comparison are shown in Table 1, where the two components of F_1 measure, namely spam recall and spam precision, are also given. Besides, the performance of NN, which was utilized in self-trigger process, is also presented in the table.

On corpus PU1, PU3 and PUA, the DTL model outperforms SVM and NB in terms of both accuracy and F_1 measure. On corpus PU2, the DTL model performs equally as SVM and outperforms NB. From these results, we can draw a preliminary conclusion that the proposed DTL model can effectively improve the performance of classifiers.

Why does the DTL model perform not so outstandingly on corpus PU2 as it does on the three other corpora? The preliminary investigation shows that the

[1] The four PU corpora can be downloaded from the web site: http://www.aueb.gr/users/ion/publications.html



(a) Performance of the DTL model with different danger zone size

(b) Accuracy of SVM, NB and DTL on corpus PU1, PU2, PU3 and PUA

(c) F_1 measure of SVM, NB and DTL on corpus PU1, PU2, PU3 and PUA

Fig. 2. Performance of SVM, NB and DTL on corpus PU1, PU2, PU3 and PUA

Table 1. Performance of SVM, NB, NN and DTL on four PU corpora

Corpus	Approach	**Recall**	**Precision**	**Accuracy**	F_1	**Feature dim.**
PU1	SVM	95.83%	95.39%	96.06%	95.54%	300
	NB	85.00%	98.30%	92.75%	91.06%	300
	NN	84.17%	94.43%	90.73%	88.86%	300
	DTL	96.04%	96.89%	**96.88%**	**96.44%**	300
PU2	SVM	72.86%	88.72%	**92.54%**	**79.31%**	300
	NB	65.71%	91.00%	91.83%	75.60%	300
	NN	45.71%	84.13%	87.32%	58.52%	300
	DTL	72.86%	88.72%	**92.54%**	**79.31%**	300
PU3	SVM	94.45%	96.04%	95.79%	95.19%	300
	NB	77.25%	94.03%	87.72%	84.66%	300
	NN	84.51%	95.38%	91.33%	89.57%	300
	DTL	94.73%	95.99%	**95.88%**	**95.31%**	300
PUA	SVM	94.56%	92.60%	93.25%	93.41%	300
	NB	94.39%	93.98%	93.86%	93.95%	300
	NN	90.88%	86.87%	87.98%	88.39%	300
	DTL	95.44%	93.93%	**94.47%**	**94.57%**	300

two grounding classifiers make more correlated error on corpus PU2 compared to other corpus. This reflects that the success of the DTL model lies in selection of uncorrelated grounding classifiers. Besides, the poor performance of the self-trigger process (NN) on PU2 is also a reason for the unideal performance of the DTL model.

5 Conclusions

In this paper, we have transplanted the main concepts of the DT into building an immune based learning model. In addition, the DTL model has been successfully applied to a typical machine learning problem – spam detection. The experimental results show that the proposed DTL model can promote the performance of

grounding classifiers. In the experiments, the DTL model outperformed SVM, NB and NN in terms of both accuracy and F_1 measure.

In future work, we seek to incorporate other design of danger zone and self-trigger process into the DTL model, and investigate the performance of the model under different settings. In this way, we hope to obtain a more ideal model and better performance. Finally, we intend to add other signals, which can indicate the drift of knowledge, into the DTL model. In this way, we hope it can develop into an adaptive learning model.

Acknowlegements. This work was supported by National Natural Science Foundation of China (NSFC), under Grant No. 60875080 and 60673020, and partially supported by the National High Technology Research and Development Program of China (863 Program), with Grant No. 2007AA01Z453.

References

1. Matzinger, P.: The danger model: a renewed sense of self. Science's STKE 296(5566), 301–305 (2002)
2. Aickelin, U., Cayzer, S.: The danger theory and its application to artificial immune systems. In: Proceedings of the First International Conference on Artificial Immune Systems, pp. 141–148. Citeseer (2002)
3. Secker, A., Freitas, A., Timmis, J.: A danger theory inspired approach to web mining. In: Timmis, J., Bentley, P.J., Hart, E. (eds.) ICARIS 2003. LNCS, vol. 2787, pp. 156–167. Springer, Heidelberg (2003)
4. Aickelin, U., Bentley, P., Cayzer, S., Kim, J., McLeod, J.: Danger Theory: The Link between AIS and IDS? In: Timmis, J., Bentley, P.J., Hart, E. (eds.) ICARIS 2003. LNCS, vol. 2787, pp. 147–155. Springer, Heidelberg (2003)
5. Aickelin, U., Greensmith, J.: Sensing danger: Innate immunology for intrusion detection. Information Security Technical Report 12(4), 218–227 (2007)
6. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
7. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Proceedings of International Conference on Machine Learning, pp. 412–420. Citeseer (1997)
8. Guzella, T., Caminhas, W.: A review of machine learning approaches to Spam filtering. Expert Systems with Applications 36(7), 10206–10222 (2009)
9. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Transactions on Neural networks 10(5), 1048–1054 (1999)
10. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 Workshop, vol. 62, pp. 98–105. AAAI Technical Report WS-98-05, Madison (1998)
11. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to filter unsolicited commercial e-mail. Technical report, National Centre for Scientific Research "Demokritos", Greece (2006)