Chinese Spam Filter under Adversarial Impact

Yaqing Zhao[1,a], Yan Xu[1,2,b], Xiaodan Zhao[1]

[1] Beijing Language and Culture University, Beijing 100083, China

[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 10080, China

[a] zhaoyaqing1991@163.com , [b] xuy@blcu.edu.cn

Keywords: Anti-spam, Machine Learning, Adversarial Attacks, Adversarial Impact

Abstract. Machining learning techniques have achieved great success in anti-spam area. But because of the limitations of these techniques, classifiers derived from them often get attacked by spam senders thus posing a threat to the whole Spam filtering system. This article briefly describes the type of attacks to classifiers and then simulates an attack on a public Chinese spam corpus to analyze the adversarial impact of several major classifiers.

Introduction

E-mail has been widely used since it appeared in the 1970s. According to the report of NetEase Company of China in 2014, each netizen owns 3.8 e-mail boxes on average, 87% of which use e-mails every day. However Spam Research Center (ASRC) points out that over 70 % mails are spams in Security Trends Report which also states that 3/4 of spams are from China. Spams not only consume network resources, reduce network operational efficiency, causing a great threat to network security, but also invade privacy which results in leakage of personal information.

Machine learning algorithms have been successfully applied to anti-spam filter system [1,2,3], and they have become the target of deliberate obstruction from spam senders in mainly 2 aspects, generating new spam variants and attacking classifier learning. The most common spam variant is adding spam information including texts, pdf documents, images with advertising message [4], html documents into attachments in order to escape detection. As for classifier attacking, spam senders usually try to change the classification and identification in the training process [5] by modifying the training data on purpose. Machine learning techniques assume that the training corpus can be a good representative of the real data and ignore the artificial data modification, classifier attacks can often reduce the accuracy of the classifier and undermine its credibility [6]. Thus classifiers' adversarial impact under malicious attacks is the research priority.

This article firstly introduces the related work and the types of classifier attacks especially in spam filter area, then describes several machining learning classifiers widely used in anti-spam area briefly and at last simulates an attack what we call Confusion Attack on a public Chinese spam corpus to analyze the adversarial impact of these classifiers.

Related Work

Adversarial classification is proposed for the first time by Dalvi, et al [6] in 2004 who view the classification as a game between the classifier and its adversary and formalize the problem into a frame and an algorithm which acquires a more optimal classifier. Considering that the attacker may not have perfect knowledge of the classifier, Lowd, et al [7] introduce a theoretical framework , adversarial classifier reverse engineering (ACRE) , for studying adversary and classifier which determines whether an adversary can efficiently learn enough about a classifier to minimize the cost of defeating it. Barren, et al [8] present a taxonomy of different types of attacks on machine learning techniques and a variety of defenses against those attacks. Wei Liu, et al [9] model the interaction between a data miner and an adversary as a Stackelberg game with convex loss functions ,then solve the Nash equilibrium problem. Battista Biggio, et al [10] use multiple classifier to resist adversarial attacks. Wei Deng, et al used the idea of injecting malicious information to corpus raised by Battista Biggio [11] to perform good word attacks in Chinese spam corpus and receive good effects. Xiaohui Pei, et al compare the performance of linear classifiers on Chinese spam corpus under good word attacks, and prove that SVM performs better.

Attacks to Spam Classifier

The attacks to classifiers are defined in 3 aspects: influence, specificity and security violation [8].In terms of influence, attacks can be divided into Causative Attacks and Exploratory Attacks. The difference between which is Causative Attacks have some measure of control over the training of the learner while Exploratory Attacks have not and can only use other techniques such as offline analysis to discover information. As to specificity, attacks are classified into Targeted Attacks and Indiscriminate Attacks. Targeted Attacks focus on a particular or a small set of points. However Indiscriminate Attacks have a flexible goal of involving a general class of points, for example, “any false negative”. The third is security violation which can be separated to Integrity Attacks and Availability Attacks. The main issue of Integrity Attacks is increasing false negatives, but Availability Attacks have a much more boarder influence of resulting in so many classification errors including false negatives and false positives.

The attacks to spam classifier are mainly about the attacks for spam recognition. The spam senders intend to disturb the identification and investigation of the receiving end by sending some confusion or poison information. Among all types of attacks, a kind of Exploratory Attacks named Evasion Attack [12] is the most commonly used. In Evasion Attack, spam senders disguise spam content by removing a portion of spam words or blurring these words, and adding some legitimate content, so that a spam looks more like a legitimate message and then escapes the detection of classifiers, releases successfully. Evasion Attack can degrade the accuracy of spam filters and let a camouflage spam escape filter detection. Dictionary Attack [13], Frequent Word Attack [13], Frequency Radio Attack [13], Weak Statistical [13], Sparse Data Attack, Obfuscation are frequently seen in Evasion Attack.

In Causative Attacks for spam classification, Poison Attack [14] is most frequently used by attackers. Spam senders add samples doped with misleading information to training set to mislead the learning procedure of classifiers. This will lead to a result that classifiers generate much more false positives in test set [15].

Furthermore, some other ways like adding junk words into legitimate e-mails with the aim to reduce the junk attributes of these words, or alter mails with spam titles and legitimate body to reduce the junk attributes of spam titles in classifiers are also popular.

Introduction to Main Classifiers in Spam filter

Support Vector Machine (SVM)

Support Vector Machine method is widely used in data mining, pattern recognition and some other areas since it was firstly proposed in 1995 [16]. The basic idea of SVM is that feature vectors will be mapped to a high-dimensional space, in which the data can be separated properly by a plane with a maximum interval. As illustrated in Fig.1, in a two-dimensional plane, point set $\mathrm{T} = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ where $x_i \in R^n$ and $\mathrm{y}_i \in \{+1, -1\}$ denoted in the map by ○ and× respectively. The straight line in Figure 1 represents the maximum interval plane which can be formalized as a function.

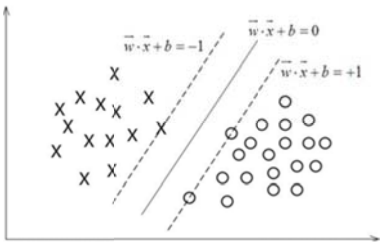
$\overline{w}\cdot\overline{x}+b=-1$
$\overline{w}\cdot\overline{x}+b=0$
$\overline{w}\cdot\overline{x}+b=+1$

Fig.1 SVM schematic

$$\{\mathrm{x}|w^T x + b = 0\}$$
$$\mathrm{s.t.}\ \ \max(w) = \sum_{i=1}^{n} y^i w^T x^i + b \tag{1}$$

The data in Fig. 1 is linear separable. When faced with the point set of the linear inseparable, SVM maps them into a high dimensional space to separate. But the consequences of such deal are likely to lead to a rapid increase in high-dimensional space dimension. In order to solve this problem, SVM will use the RBF (Radial Basis Function) whose nature is calculating in the low-dimensional space but expressing the essence of classification in high-dimensional space. The choice of RBF may influence the performance of SVM and should be based on specific issues.

Naive Bayesian Model (NBM)

Bayesian Model is firstly introduced by Sahami, Dumais Heckeman et al [17] and applied to spam filtering for the first time. The basic idea of NBM is calculating the posterior probability of an object using bayes formula when given its prior probability, and select the class with the largest posterior probability as the class of the given object. Assuming that $\mathrm{X} = (x_1, x_2, \dots, x_n)$ is the feature vector of an e-mail, x_i means the feature value in ith position, n represents the number of feature dimensions. Let $\mathrm{C} \in \{spam\ ,\ \ ham\}$ represents categories. Then NBM use the formula below to calculate the conditional probabilities of each mail $\mathrm{P}(\mathrm{C}_i|\mathrm{X})$, namely the probability of sample X belongs to category C_i.

$$\mathrm{P}(C_i|\mathrm{X}) = \frac{P(C_i)P(X|C_i)}{P(X)}$$
$$\mathrm{P}(X|C_i) = \mathrm{P}(x_1|C_i) * \mathrm{P}(x_2|C_i) * \mathrm{P}(x_3|C_i) * \dots * \mathrm{P}(x_\mathrm{n}|C_i) = \prod_{j=1}^{j} \mathrm{P}(x_\mathrm{j}|C_i) \tag{2}$$

In the formula, $P(C_i)$ is the priori probability of class C_i. $P(X)$ is the input probability, i.e. the probability of generating the feature vector X, regardless of the category.

Multi-Layer Perceptron neural network

Multi-Layer Perceptron neural network (MLP) is a kind of neural network composed of a group of sensing units that are connected to all the units from adjacent layers, but no junctions among units in the same layer. The network has an input layer, one or more intermediate layers (hidden layer), and an output layer [18]. Each unit has several inputs x_i with a weight w_i and one output y, namely the activation value of the neuron. Formula below quantifies the perception of unit calculation.

$$\mathrm{y} = \mathrm{f}(w^T\mathrm{x}) = \mathrm{f}\left(\sum_{i=1}^{d} w_i x_i\right) \tag{3}$$

The training procedure of MLP consists of 2 sections, forward section and backward section. In forward section a training sample X is provided to MLP, the activation value y of X is passed from the input layer to the output layer through each intermediate layer, and finally produce the input responses of the network from all sensing units in the output layer. In backward section, MLP modifies all connection weights w_i from the output layer to input layer through each intermediate layer with the aim to reducing the actual errors.

Adaptive Aadboost

The algorithm's idea of boosting originates from Probably Approx Corret (PCA) Model proposed by Valiant [19]. The nature of the boosting is to enhance weak classifiers whose recognition error rates are lower than 0.5 to a strong classifier by combination. However boosting has a defect that it requires the prior knowledge of the weak classifier. Thus FreundY et al [20] put forward an improved algorithm Aadboost which overcomes the shortcoming.

In the training process, each sample is assigned with a weight, T iterations. After each iteration, the weights of miscategorized samples will increase. After T iterations, Aadboost produces T weak classifiers with different weights, and the final prediction function H in classification problem is produced by weighted voting method using all T weak classifiers.

Experiment Introduction

Confusion attack

This article simulates a Causative Attack what we call Confusion attack on Chinese spam corpus CCERT. In Confusion Attack we modify the corpus manually by adding legitimate email content into a portion of spams, so the whole confusion corpus consists of ham, spam and confusion spam. Then we use the original corpus and modified corpus, in other words confusion corpus, to train the classifiers. By comparing the classification performance of several classifiers on balance dataset and imbalanced dataset, we give an analysis on Chinese spam filter under adversarial impact.

Corpus composition

The experiment in this article is based on a public Chinese spam corpus named CCERT corpus which contains 2 subsets, 2005-Jun data set and 2005-Jul data set. There are 25088 spams and 9272 hams in 2005-Jun data set, 20308 spams and 9042 hams in 2005-Jul data set. All hams in CCERT are collected from forum and spams are gathered by honeypot technique.

Considering the fact that the number of spam is far higher than the hams in actual situation, and the imbalance of corpus also interferes classifier [21]. So we set a Balance Dataset A, a Balance Confusion Dataset B and an Imbalanced Confusion Dataset C as illustrated in Fig. 2. Balance Dataset A consists of all 9042 hams and 9042 spams randomly chosen from 2005-Jul data set. Balance Confusion Data B has the same hams and 4542 spams with Balance Bata A, and other 4500 spams injected with hams randomly chosen from 2005-Jun data set. Imbalanced Confusion Dataset C is composed of 9042 hams, 11208 spams and 9100 spams containing hams that is also randomly chosen from 2005-Jun data set.

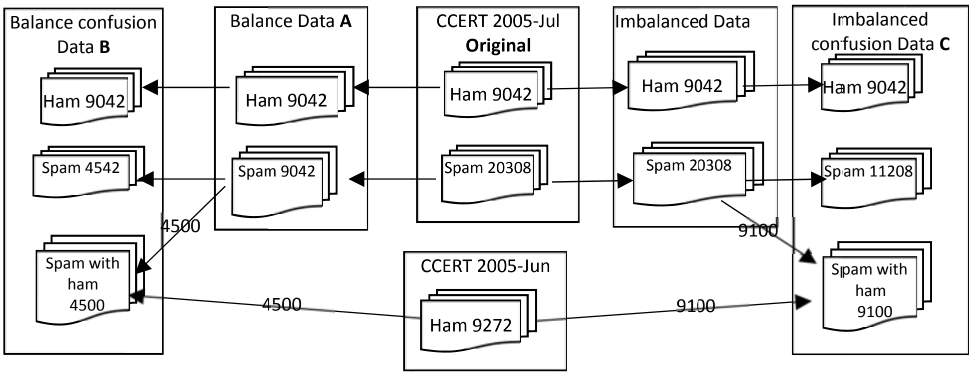

Fig. 2 experiment corpus composition

Data Processing

Data preprocessing and feature representation are key issues in applying machine learning to solve problem, and in the meanwhile determine the quality of classifiers trained. We process the experiment corpus as shows in Fig. 3.

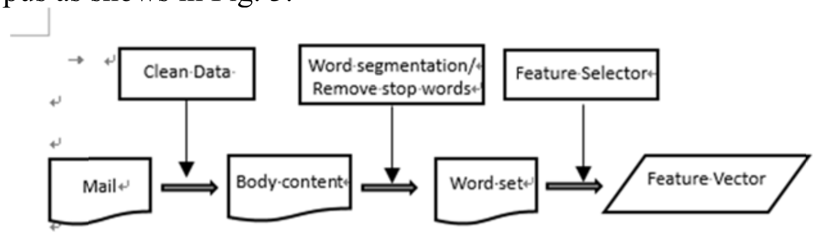

Fig.3 Data processing flow chart

In the Clean Data stage, we extract the body of a mail for the reason that it is rich in information.

According to the Chinese language features, there is no space to separate words. We perform word segmentation using ICTCLAS segmentation system developed by ICT. It is also necessary to remove stop words which occur frequently but meaningless to get corpus with analyzable value, namely Word set in Fig. 4.

Feature Extraction and Presentation

Vector Space Model (VSM) [22] is widely used in text classification and information retrieval area and performs well. In this paper we use VSM to represent each mail, a mail can be expressed as a feature vector $\mathrm{X} = (x_1, x_2, \dots, x_n)$ where x_i indicates the weight of ith word, n represents the total number of feature words, in other words, feature dimension.

Feature words are selected by feature selector as illustrated in Figure 4. Commonly used Chinese text feature selection methods are chi-square statistic (CHI) [23], information gain (IG) [23] and DF [23]. According to the result of Siyao Han et al [24] on anti-spam study, we use CHI to choose feature words and the weight of each feature word is its TF-IDF value.

We perform data preprocessing and feature representation procedure on Balance Dataset A, Balance Confusion Dataset B and Imbalanced Confusion Dataset C, then transform each mail into a feature vector.

Assessment criteria

Precise and Recall are 2 important indicators to evaluate a classifier. Precise represents the degree that a classifier classifies objects correctly, for example in this article, how many mails are really spams among the spams assigned by the classifier. Recall means the classification integrity of a classifier, for example, how many spams are assigned to spam in all spams. Sometimes there may be contradictions between Precise and Recall, so a new evaluation standard F-Measure which combines Precise and Recall together is applied. This article takes Precise, Recall and F-Measure (F1) as the assessment criteria. The calculation methods are shown in Table 1.

Table 1 Calculation methods table

	Ham	Spam
Judged as Ham	f00	f01
Judged as Spam	f10	f11
$\mathrm{Precise} = \frac{\mathrm{f11}}{\mathrm{f10+f11}}$ $\mathrm{Recall} = \frac{\mathrm{f11}}{\mathrm{f01+f11}}$ $\mathrm{F1} = \frac{2*\mathrm{Precise}*\mathrm{Recall}}{\mathrm{Precise+Recall}}$		

Experiment Tools and Parameters of Classifiers

Weka is a collection of machine learning algorithms for data mining tasks and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Our experiment is based on Weka. LibSVM, Naive Bayesian, MLP, AdaBoostM1 in Weka are the classifiers we want to analysis.

For the parameters selection, we use Grid search to find the optimal parameters and use Cross Validation to select the suitable models which avoid over-fitting. LibSVM classifier performs better with gamma 5, cost 25 and nonlinear kernel RBF. Native Bayesian classifier takes default parameters in Weka. AdaBoostM1 classifier is trained with 500 iterations by using weak classifier Decision stump. For MLP the learning rate is 0.3 and the momentum is set to 0.2.

Experiment Results

On dataset A, B and C we randomly split 80% corpus for training and the remaining 20% for testing, 20-200 features with interval of 10, namely 10 different dimensions. All training models are trained with 10-fold cross-validation to get the most suitable classifiers. The result is in the following.

The Precision Result

Fig. 4 shows the precision results on all dataset, balance and imbalance, normal and confusion.

Adboost, MLP and SVM have some a little influence under confusion attack, drop by some 3-4% in average when compared with the result on Balance Dataset which shows in the figure 5 above. Also the imbalance of corpus doesn’t influence precision as show in in the figure 5 below.

To our surprise, the precision of Naive Bayes has a dramatic increase about 20 percent. And the shape of Naive Bayes shows a rising downward trend along with the increase of feature numbers. Feature selection method CHI that calculates the correlation between variables and classes is used to select feature words in our experiment. In confusion attack, hams are injected into spams, which reduces the relevancy of certain legal words and hams, and in the meanwhile improves the ranking of junk words. But when more features words are added, more legal words are added. This will lead to the rising downward trend of Naive Bayes.

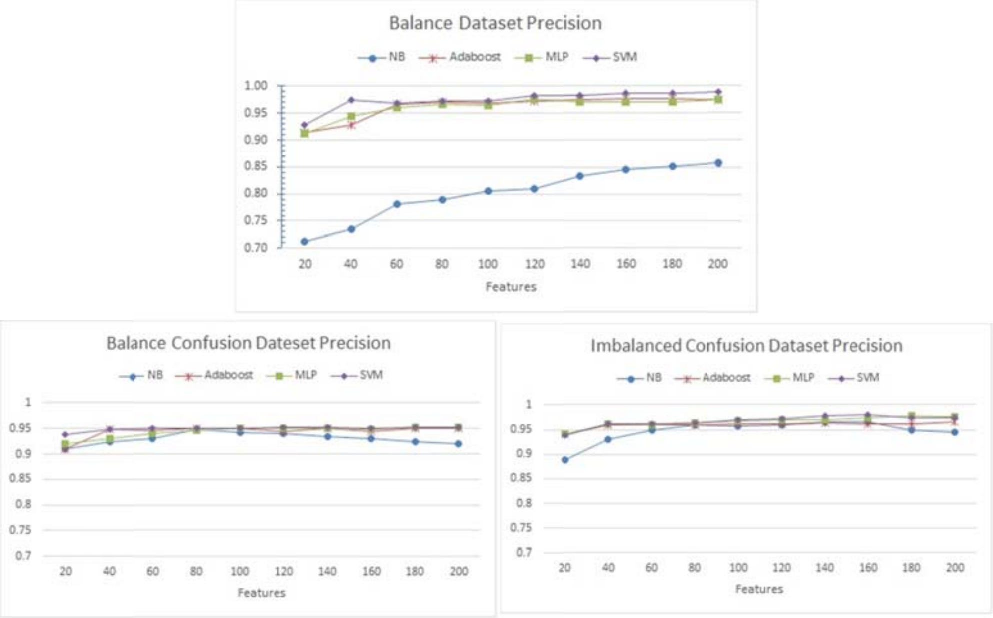

Fig. 4 The precision results on all dataset

The Recall Result

As show in Fig.5, confusion attack doesn’t make a big influence on Adboost, MLP and SVM classifiers in recall, the value is about 95%. But for Naive Bayes, things are different. The recall rate decreases by 20% and 10% respectively on Balance Confusion Dataset B and Imbalanced Confusion Dataset C. Recall is an index to measure how many spams are judged as spam in this article. The decline of recall means more spams escape the interception of classifiers. Native Bayes classifier classifies a mail by comparing the probabilities the mail belongs to spam and ham. In confusion attack, more legal words are selected as feature words, and the ratio of junk words in the mail body will decline for the reason that we add hams into part of spam corpus. All these lead to a low spam probability of a disguise spam, then it will massively more likely be judged as a ham. From Figure 6, we also draw a conclusion that the imbalance of corpus doesn't degrade the recall performance of various classifiers, however, the recall of Native Bayes get improved when adding more spam data.

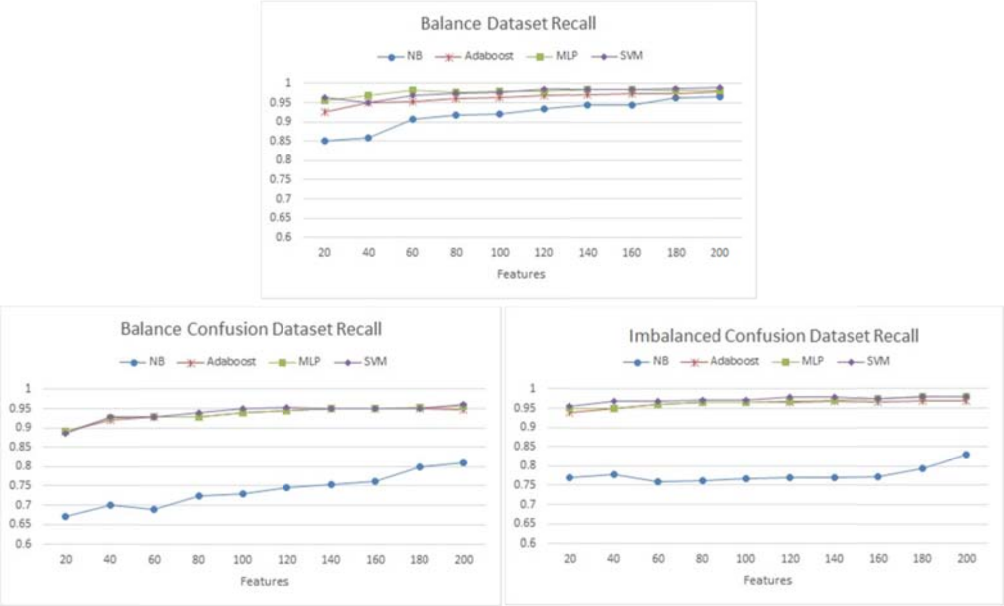

Fig. 5 The recall results on all dataset

The F1 Result

F1 value is a comprehensive criterion to evaluate a classifier. The graph above in Fig.6 is the F1 result on original balance dataset, the images below are results on confusion corpus. According to the figure, confusion attack do affect the performance of all classifiers. Adboost, MLP and SVM are less affected with only a speck of drop, you might say, they perform stably under confusion attack. Native Bayes falls about 5% in average, and the trendline on Balance Dataset A keeps rising with the increase of feature numbers, but this not happens on confusion data, they stay at about 85%. What's more, the imbalance of corpus improve the performance of all classifiers under confusion attack and we can see it from the images below.

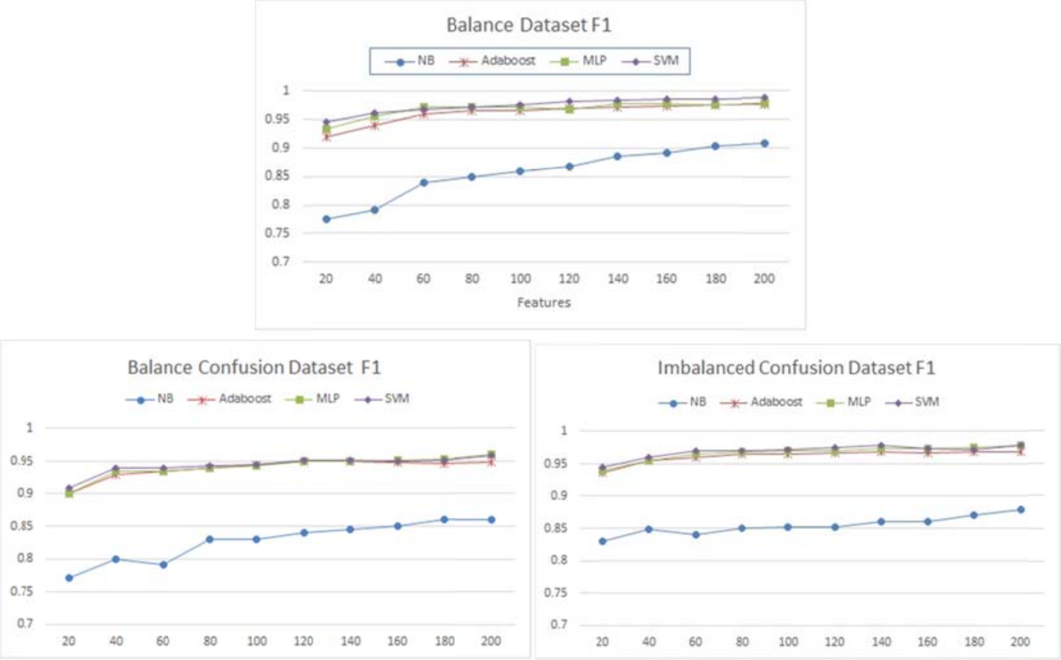

Fig. 6 The F1 result in all dataset

Summary

This article gives an overview on the machine learning classifiers and then perform a confusion attack on 4 machine learning classifiers in balance and imbalance corpus. Support Vector Machine, Adaptive Adboost and Multi-Layer Perceptron neural network perform stably under confusion attack only with a little performance loss. Native Bayes is affected seriously, and it's not smart to be used in anti-spam system. According to the experimental results, the imbalance of corpus may improve the performance of classifiers under confusion attack.

Acknowledgement

The authors give their sincere gratitude towards The National Key Technology R&D Program (Grant No.2012BAH39B02), National Natural Science Foundation of Beijing (Grant No. 4122076) for providing support.

References

[1] Androutsopoulos I , Koutsias J , Chandrinos K et al. An evaluation of Naive Bayesian anti-spam filtering. In Proceedings of the Workshop on Machine Learning in New Information Age, 11th European Conference on Machine Learning. Spain, 2000.09.17

[2] I. Androutsopoulos, G.Paliouras, E.Michel.akis, Learning to Filter Unsolicited Commercial E-Mali. Technical report2004/2. NCSR "DEmokritos", 2004.11

[3] E.Blanzieri ,A.Bryl, A survey of learning-based techniques of email spam filter. Artificial Intelligence Review. vol. 29(2008), p.63-92.

[4] FUREMA G, PILLAI I, ROLI F. Spam filtering based on the analysis of text information embedded into images. Journal of Machining Learning Research. 2006, 7:2699-2720

[5] Daniel Lowd, Christopher Meek. Adversarial Learning. In ACMPress,editor,Proceedings of the Eleventh ACM SIGKDD Internatinal Conference on Knowledge Discovery and Data Mining(KDD). Chicago, IL, 2005,p. 641-647.

[6] NileshDalvi, Pedro Domingos, Mausam, SumitSanghai, et al. Adversarial classification. In Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Seattle, 2004, p. 99-108

[7] LOWD D, MEEK C. Adversarial learning. In Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 2005, p.641-647.

[8] BARRENO M, NELSON B, SEARS R, et al. Can machine learning be secure? . In ASIACCS'06: Proceedings of the 2006 ACM Symposium on Information, computer and communications security. New York, USA, 2006. P.16–25.

[9] LIU W, CHAWLA S. Mining adversarial patterns via regularized loss minimization. Machine Learning. 2010, p. 69-83.

[10] Battisa Biggio, Giorgio F, Fabio R. Adversarial pattern classification using multiple classifiers and randomization. Proceeding of the 12th Joint IAPR International Workshop on Structural and Syntactic Pattern Recognition (SSPR 2008). Orlando, Florida: Springer, 2008, p.500-509

[11] Battista Biggio, Adversarial Pattern Classification. Sardinia: University of Cagliari, 2010

[12] Xiao, Han, Thomas Stibor, Claudia Eckert. Evasion attack of multi-class linear classifiers. Advances in Knowledge Discovery and Data Mining. Springer Berlin Herdelberg, 2012,p.207-218.

[13]D.Lowd and C.Meek. Good word attacks on statistical spam filters. In Proceedings of the 2nd conference on Email and Anti-Spam, 2005.

[14] Kloft, Marius, and Pavel Laskov. Online anomaly detection under adversarial impact. 2011

[15]Nelson Blaine A. Behavior of Machine Learning Algorithms in Adversarial Environments. No.UCB/EEECS-2010-140. California Univ Berkeley Dept of Electrical Engineering and Computer Science, 2010.

[16] V. Vapnik , S. Kotz. Estimation of dependencies based on empirical data. Springer, 1982

[17] M. Sahami, S. Dumais, D. Heckerman, et al. A Bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization. 1998.

[18] V. Vapnik and S.Kotz. Estimation of dependencies based on empirical data. Springer, 1982

[19] Valiant L. G., A Theory of the Learnable. Communications of the ACM, 1984, p. 1134-1142

[20] FreundY., SchapireR. E. A. Decision Theoretic Generalization of On Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997, p. 119-139.

[21] Yanming Sun, Andrew K.C.Wong, Mohamend S.Kamel. Classification of Imbalaced Data: A Review.International Journal of Pattern Recognition and Artificial Intelligence. Vol.23(2009),No.4, p.687–719

[22]C.D. Manning, P. Raghavan, H.Schutze . Introduction of Information Retrieval. 2008, p. 110-115.

[23] Yang Y, Pedersen J. A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th Ien nternational conference on Machine Learning. 1997, p.412-420.

[24] Siyao Han, Yan Xu. A Comparative Study on Machine Leaning Techniques in Chinese Spam. ICACI2015, 2015.03