



An Efficient Algorithm to Analyse the Cost Sensitive Measures on Datasets

Vishnupreeth K.G, Talit Sara George

Pursuing M.Tech, Dept. of CSE, Caarmel Engineering College, MG University, Kerala, India

Assistant Professor, Dept of CSE, Caarmel Engineering College, MG University, Kerala, India

ABSTRACT: An efficient algorithm to analyse the cost sensitive measure is proposed. The main goal of this paper is to measure the sum of weighted sensitivity and specificity, total misclassification cost. To achieve the maximum sum value and minimum cost value, cost sensitive gradient descent algorithm and constraint based gravitational search algorithm is used. The proposed algorithm can work well for anomaly detection in bioinformatics and cancer datasets. The algorithm can work highly efficient and accurate for classify sensitive data. Finally, compare the performance of proposed algorithm with other existing algorithm.

KEYWORDS: Sensitivity, Specificity, Misclassification cost, CSOGD.

I. INTRODUCTION

Nowadays in communities of data mining and machine learning, classification [2] and online learning [1] have been widely examined. It is used to develop efficient and scalable algorithms for mining substantial massive volume of data. In general, the goal of online learning is to incrementally realize few prediction models to make accurate predictions on a stream of samples that arrive consecutively. Online learning is highly efficient and scalable for large application such as gene profiling, medical diagnosis, credit card fraud detection. Researchers developed different online learning algorithm, which support online classification that have been proposed in literature [2].

Cost sensitive classification is an important problem in data mining which has to address varied misclassification cost [10]. In many real world applications, misclassification cost can be quite large. For example, if considered cancer is regarded as the positive class and non- cancer as negative class. The patient is actually positive but it is classified as negative. So the classification is wrong thereby the patient could lose her/his life. Most of the existing algorithm might not be able to provide optimal solution with 100% prediction rate and accuracy, which is obviously cost insensitive and thus inappropriate for many real world applications in data mining.

Researchers have proposed three meaningful metric in literature, such as sensitivity [7], specificity [7], and misclassification cost [10]. Researchers have been developed different batch classification algorithm to improve the cost sensitive measures, but these algorithms are not suitable for online classification applications.

In this paper, we describe how to solve the classification problems with balance accuracy. To achieve the accuracy and reduce the misclassification error, we proposed a new framework of cost sensitive classification. The main goal of this framework is to measure the weighted sum of sensitivity and specificity, total misclassification cost by using cost sensitive online gradient descent technique to tackle the optimization task. We are also trying to develop a constraint based gravitational search algorithm which reduces error due to misclassification and also time complexity, thereby increasing the accuracy. Then apply these two algorithms to online anomaly detection task such as bioinformatics and breast cancer. Finally compare the performance of both algorithms. The proposed algorithm can work well for online anomaly detection task and highly efficient and scalable for classification problems.

The rest of the paper is organized as follows. Section 2 formulates the problem definition. Section 3 shows the system architecture and module description. Section describes the two algorithms. Section 5 describes the experimental setup on online anomaly detection task. Section 6 concludes the work.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

II. PROBLEM DEFINITION

The classifier is built from a set of training data whose class label is '+1' or '-1'. At each round, learner receives an instance and predicts its class label. After making the prediction, the learners receive the true label of the instance and suffer a loss if the prediction is incorrect. The result of each prediction can be classified into four cases [1]:

- (1) True positive (TP) if $\hat{y} = y = +1$;
- (2) False positive (FP) if $\hat{y} = +1$ and $y = -1$;
- (3) True negative (TN) if $\hat{y} = y = -1$;
- (4) False negative (FN) if $\hat{y} = -1$ and $y = 1$;

Where $\hat{y} = \text{sign}(W \cdot x)$. The value $|W \cdot x|$ known as the margin is used by the learner on prediction. The true label for instance x is denoted as $y \in [-1, +1]$. If $\hat{y} \neq y$, the learner makes mistake; otherwise it is a correct prediction. To make the prediction accuracy better, a most appropriate metric is considered to measure the weighted sum of sensitivity and specificity [1].

$$\text{Sum} = \delta p * \text{sensitivity} + \delta n * \text{specificity}$$

Where $\delta p + \delta n = 1$ and $0 \leq \delta p, \delta n \leq 1$ are two parameters to trade off between sensitivity and specificity. Notably, when $\delta p = \delta n = 0.5$, the corresponding sum is the well known balanced accuracy. In general, higher the sum value, better the classification performance. Another approach is to measure the total cost [1], which is defined as:

$$\text{Cost} = C_p * M_p + C_n * M_n$$

Where $C_p + C_n = 1$ and $0 \leq C_p, C_n \leq 1$ are the misclassification parameters for positive and negative classes respectively. The lower cost value, better the classification performance.

III. SYSTEM ARCHITECTURE

The system architecture includes different components such as data collector, pre - processor, learning algorithm, data model etc. The classification techniques usually require a considerable amount of training data in order to build good classification model. The data is fed to the data collector from LIBSVM [3] library. Second step is data pre-processing. Data pre- processing [9] is used to transform the raw data into a format that makes it possible to apply data mining techniques. Data extraction [9] and normalization [9] are the most important tasks in pre-processing. In the training phase, the evaluation parameters of the algorithm are analysed and specified based on the given training data within the dataset. Here more than one learning algorithm is used for building a good data model. The data model is then validated against the test data given by the dataset to evaluate classification result. The classification result are explained in terms of cost sensitive measures i.e.; weighted sum and cost. In general, higher the sum value, better the classification performance. The lower cost value, better the classification performance. The build classifier can also be used for anomaly detection.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

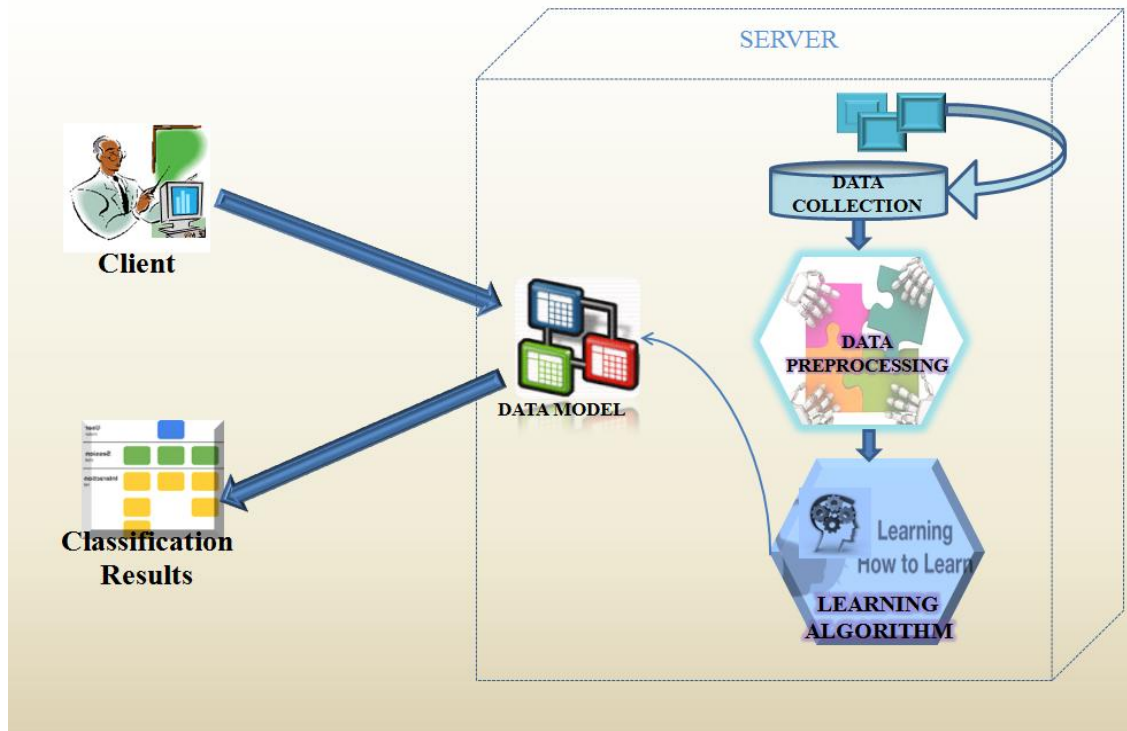


Fig 1: System Architecture

IV. ALGORITHMS

Algorithm 1: CSOGD algorithm [1]

INPUT: learning parameter λ , bias parameter $\rho = \frac{\delta p T n}{\delta n T p}$ for sum and $\rho = \frac{C_p}{C_n}$ for cost

Step 1: Initialization $w_1 = 0$

For loop $i = 1 \dots I$;

Step 2: Instance X_i arrives

Step 3: Make a prediction based on existing weights

$$y_i^{\wedge} = \text{sgn}(w_i \cdot x_i)$$

Step 4: Observe the true class label $y_i \in \{+1, -1\}$

Step 5: Suffer loss $l_i(w_i) = l * (w_i; (x_i, y_i))$

Step 6: if $(l_i(w_i) > 0)$

Update classifier $w_{i+1} = w_i - \lambda \nabla - t(w_i)$;

End if

End for

OUTPUT: Updated weight.

Algorithm 2: Proposed Constraint based Gravitational Search Algorithm

The constraint based gravitational search algorithm is the latest nature inspired population based stochastic search algorithm which is widely used to solve the optimization problems. The gravitational search algorithm is based on Newton's theory. Newton's law of gravity states that every particle attracts another particle by means of some gravitational force [11]. A constraint based search is initiated within the existing gravitational search algorithm to provide better accuracy.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

- Step 1: Initialization
 For loop $i = 1 \dots N$;
 - Step 2: Evaluate the fitness for each agent
 - Step 3: Update constant G at each iteration, best and worst of the population.
 - Step 4: Calculate masses and acceleration for each agent is calculated at iteration.
 - Step 5: Update velocity and the position of the agents at next iteration.
 - Step 6: Return best solution.
- End.

V. EXPERIMENTAL RESULTS

Table 1: Sample data sets

Dataset	#Examples	Features
a9a	48842	123
German	1000	24
Spambase	4601	57

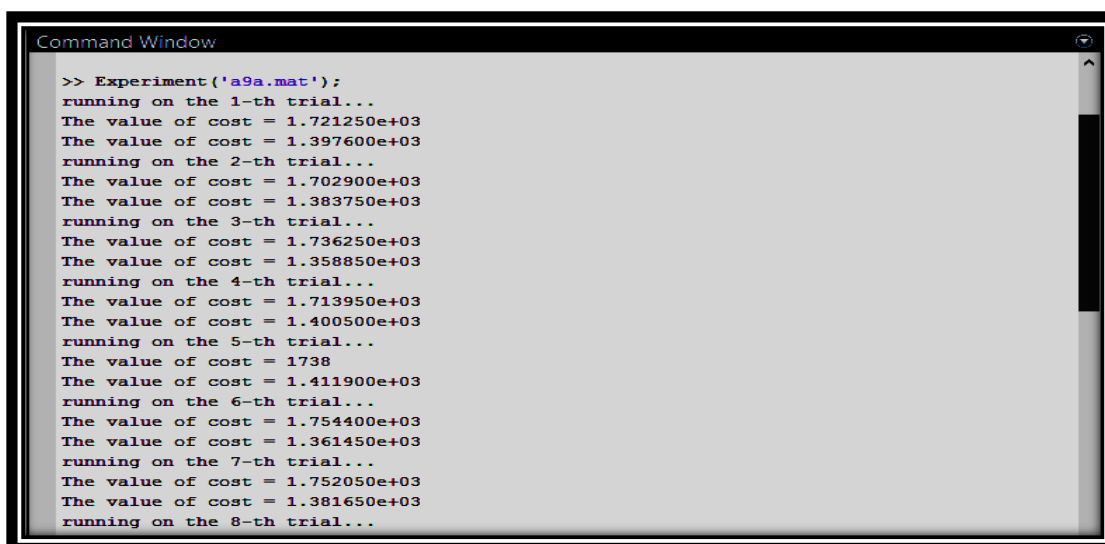
To evaluate the classification performance several metrics are considered:- weighted sum of sensitivity and specificity, and weighted cost. From the previous observations, it is found that CSOGD algorithm perform well for all sample data sets. And also found that the CSOGD algorithm shows lesser misclassification cost. For example, on a9a, the total sum made by CSOGD is 79.130 ± 0.059 and the total cost is 1385.223 ± 17.186 . In the case of german data sets, the total sum is 70.690 ± 0.846 and total cost is 77.313 ± 3.514 . For spambase datasets, the total sum is 91.473 ± 0.166 and total cost is 86.235 ± 3.652 .

```
Command Window
>> Experiment('a9a.mat');
running on the 1-th trial...
The sum of weighted sensitivity and specificity = 7.885733e-01
The sum of weighed sensitivity and specificity = 7.909004e-01
running on the 2-th trial...
The sum of weighted sensitivity and specificity = 7.902650e-01
The sum of weighted sensitivity and specificity = 7.920953e-01
running on the 3-th trial...
The sum of weighted sensitivity and specificity = 7.862016e-01
The sum of weighed sensitivity and specificity = 7.918886e-01
running on the 4-th trial...
The sum of weighted sensitivity and specificity = 7.898436e-01
The sum of weighed sensitivity and specificity = 7.911208e-01
running on the 5-th trial...
The sum of weighted sensitivity and specificity = 7.889007e-01
The sum of weighed sensitivity and specificity = 7.913485e-01
running on the 6-th trial...
The sum of weighted sensitivity and specificity = 7.910168e-01
The sum of weighed sensitivity and specificity = 7.922681e-01
running on the 7-th trial...
The sum of weighted sensitivity and specificity = 7.873052e-01
The sum of weighed sensitivity and specificity = 7.920583e-01
running on the 8-th trial...
The sum of weighted sensitivity and specificity = 7.891233e-01
The sum of weighed sensitivity and specificity = 7.894376e-01
The sum of weighed sensitivity and specificity = 7.919475e-01
```

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015



```
Command Window
>> Experiment('a9a.mat');
running on the 1-th trial...
The value of cost = 1.721250e+03
The value of cost = 1.397600e+03
running on the 2-th trial...
The value of cost = 1.702900e+03
The value of cost = 1.383750e+03
running on the 3-th trial...
The value of cost = 1.736250e+03
The value of cost = 1.358850e+03
running on the 4-th trial...
The value of cost = 1.713950e+03
The value of cost = 1.400500e+03
running on the 5-th trial...
The value of cost = 1738
The value of cost = 1.411900e+03
running on the 6-th trial...
The value of cost = 1.754400e+03
The value of cost = 1.361450e+03
running on the 7-th trial...
The value of cost = 1.752050e+03
The value of cost = 1.381650e+03
running on the 8-th trial...
```

Further, by examining both sensitivity and specificity metrics, we found that CSOGD often achieves the best sensitivity result, but does not always guarantee the best results for specificity and misclassification error. So we are trying to implement CBGSA algorithm that may achieve best sensitivity, specificity, and misclassification result when compare to CSOGD algorithm and also to reduce the time complexity.

VI. CONCLUSION

Different learning algorithm such as cost sensitive online gradient descent algorithm, constraint based gravitational search algorithm has been proposed in order to evaluate the accuracy of the classifier system. These algorithms can be effectively utilized for solving online anomaly detection task. By the experimental results, the proposed technique can be proved to a highly efficient and effective tool to tackle cost sensitive online classification tasks in various application domains.

ACKNOWLEDGEMENT

We would like to thank reviewers for their helpful and constructive comments. We would also like to thank the faculties, the resource providers and the system administrator for the useful feedback and constant support.

REFERENCES

1. J. Wang, P. Zhao, and S. C. H. Hoi, "Cost-sensitive online classification," in *Proc. 12th IEEE ICDM*, Brussels, Belgium, 2012.
2. S. Neelamegam, Dr. E. Ramaraj, "Classification algorithm in data mining: An overview" *IJPTT*, 2013.
3. <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>
4. X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Proc. 6th ICDM*, Washington, DC, USA, 2006, pp. 970–974.
5. H. Masnadi-Shirazi and N. Vasconcelos, "Risk minimization, probability elicitation, and cost-sensitive svms," in *Proc. 27th ICML*, Haifa, Israel, 2010, pp. 759–766.
6. Nitesh V. Chawla, "Data mining for imbalanced datasets: An overview" in *journal of artificial intelligence research* 16 (2002) 321 – 357.
7. Chih – Wei Hsu, Chih – Chung Chang and Chih – Jen Lin, "A practical guide to support vector classification", 2003.
8. Wen Zhu, Nancy Zeng, Ning Wang, "Sensitivity, Specificity, Accuracy analysis with practical implementations" 2010.
9. http://www.cs.ccsu.edu/markov/ccsu_courses/DataMining-3.html
10. R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. 15th ECML*, Pisa, Italy, 2004, pp. 39–50.
11. Yugal kumar, G. Sahoo, "A Review on Gravitational Search Algorithm and its Applications to Data Clustering & Classification" *I.J. Intelligent Systems and Applications*, 2014, 06, 79-93.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

BIOGRAPHY

Vishnupreeth K.G received his Bachelor of Engineering in Computer Science and Engineering from Anna University, Chennai, 2012. At present, he is pursuing M.Tech in Computer Science and Engineering at Caarmel Engineering College, Kerala, and Affiliated to MG University. His research interests include Data Mining, Big Data processing.

Talit Sara George is an Assistant Professor in the Computer Science and Engineering Department, Believers Church Caarmel Engineering College, Pathanamthitta, Affiliated to MG University. She received Master of Engineering degree in 2012 from Anna University, Chennai. Her research interests are Data Mining, Cloud Computing and Big Data Analytics.