Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/108360, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Using Genetic Approach for Learning from Imbalanced Text Corpora

Lincy Mathews[1*] and Hari Seetha[2]

[1]School of Information Technology and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India; lincymm99@gmail.com,
[2]SCOPE, Vellore Institute of Technology University, Vellore - 632014, Tamil Nadu, India; hariseetha@gmail.com

Abstract

Aiming at the ever-present problem of imbalanced data in text classification, the paper employs the Genetic Algorithm approach for tackling the imbalance problem in a binary classed text data. One of the inherent characteristics of imbalanced data is the highly uneven distribution of data among the classes. Consequentially, the traditional classifier algorithms such as the Nearest Neighbor have shown a decreased performance due to the under representation of the interested class. A hybrid approach has been used to incorporate the oversampling technique with the advantages of Genetic Algorithm for generation of the artificial patterns for the minority class. This approach employs avoidance of over fitting as the fitness function to decide the stopping criterion for generation of synthetic samples. Efficient evaluation measures analyze the increase in performance of the proposed hybrid-learning model.

Keywords: Genetic Algorithm, Imbalance Data, Nearest Neighbor, Oversampling, Synthetic Data, Text Data

1. Introduction

Availability of huge volume of text corpus makes learning a much-needed requirement in terms of information retrieval, machine learning and information processing. A number of machine learning algorithms have been introduced to deal with text mining. However, a severe drawback in these learning algorithms is the assumption of balanced text data. Now, attention is paid to the quality of training data on which the models are trained. One of the parameters that affect the quality of the training data is the distribution of data between the classes. Imbalance in a two class data exist when the number of instances that represent the class of interest is much less compared to the other class.

In an unbalanced data, the class that is well - represented is known as majority class. The class of interest usually under-represented is called as minority class. As indicated by[1], unbalanced data sets often appear in many practical applications like fraud detection, medical diagnosis, text categorization and biomedicine. Considering an example in fraud detection cases, the number of fraud cases will be lesser compared to the occurrence of non-fraudulent cases. The drawback of most data mining algorithms is the bias towards majority class (well-represented classes) when the classifiers are applied to imbalanced data. Implemented works towards the problem of imbalanced data are classified into mainly three categories - Data Driven techniques, algorithmic techniques and cost sensitive methods

Occurrence of imbalanced datasets can be due to various reasons. Imbalance in data can be inherent (pertaining to the domain) or due to external factor like cost constraints, restrictions or disturbances in data collection etc. In a study of building a manufacturing centered technical paper corpus[2], due to the costly efforts demanded for human labeling and diverse interests in the papers, the dataset naturally ended with a skewed collection. Certain data intrinsic characteristics of imbalanced data can affect the performance of the classifier such as identification of clusters recognized as small disjuncts[3,4], lack of infor-

*Author for correspondence

mation in the training data[20], overlapping of instances between classes[5,6], presence of noisy data[7,8], etc.

Data-Driven techniques involve oversampling, under sampling and hybrid methods. Under sampling removes instances from the majority class. This however can lead to loss of information. Over sampling is the sampling of existing instances or creation of new instances to reduce the degree of imbalance. SMOTE is one of the most known works in the field of oversampling. The synthetic data is produced by interpolating several minority class instances that lie together for oversampling the training set. Various variations to the SMOTE has been proposed such as Borderline-SMOTE[9], Adaptive Synthetic Sampling[10], Safe-Level-SMOTE[11] and SPIDER2[12] algorithms. Cost Sensitive algorithms focus on the increasing the costs of misclassification so that the models train harder on the misclassified instances. Cost-sensitive learning solutions incorporating both the data (external) and algorithmic level (internal) approaches assume higher misclassification costs for samples for the minority class and seek to minimize the high cost errors[13–15].

In[16] first proposed the Genetic Algorithm (GA). The algorithm incorporated the evolution process found in nature. Its purpose is to conduct a search for an optimal solution to a given problem by mimicking natural selection. Several studies have demonstrated the advantages of the GA in solving high dimensionality and Feature Selection problems[17,18].

In[19] proposed a meta heuristic approach to locate an optimal learning set from imbalanced data without empirical studies that are normally required to find an optimal class distribution. The SVM classifier was employed with the under sampling technique. This method provided a capable and valuable solution for imbalanced data learning with an SVM[20] for large-scale imbalanced datasets. In[21] applied Genetic Algorithm to increase the ratio of optimistic samples and cluster the training data to remove any noisy samples. An improvement was shown in the classification measured as the area under the Receiver Operating Characteristics (ROC) curve[22].

The performance of Nearest Neighbor classifier, a non-parametric algorithm, is biased towards the imbalanced text data. This paper focuses on reducing the bias of the Nearest Neighbor model to the majority class by incorporating oversampling technique. The oversampling technique employs genetic approach to produce synthetic samples. The paper follows by a brief explanation on the proposed approach in Section 2. Section 3 describes the various datasets used to investigate the approach proposed and the experimental results. Section 4 concludes the paper.

2. Proposed Approach

The principle behind the genetic approach is the survival of the fittest. The individuals with the highest fitness function survive. These individuals are formed by the crossover of the parents. The same principle has been employed in forming the synthetic samples. The samples generated for the minority class will be of the same distribution and quality of the existing minority instances. The oversampling technique is adopted by employing cross-join of the training instances to generate the synthetic samples for the minority instances. In addition, the Genetic Algorithm (GA) approach utilizes the ratio between minority and majority samples for effective model performance. The binary class text data is represented by instances belonging to two classes, minority and majority class. The proposed approach works on the minority instances of training data.

The notations and definitions are explained below followed by the pseudo code of the proposed approach.

2.1 Notations and Definitions

The notations used in this paper are defined as follows. Consider binary Imbalanced data, TS. Given a training set, $\mathbf{TR}$, it comprises of positive and negative samples; indicated as $\mathbf{TR^+}$ and $\mathbf{TR^-}$ respectively. Therefore, $\mathbf{TR^+ \cup TR^- = TR}$. Let, $|\mathbf{TR^+}|$ be denoted as m. Let each pattern be represented as $(x_1, x_2, \ldots, x_p)$ where p indicates the number of features that exist in the dataset. i.e. $\mathbf{TS} = \{f_1, f_2, \ldots f_p\}$. The set of classes $\mathbf{C} = \{C^+, C^-\}$ indicate the class for minority and majority data respectively.

Let $\mathbf{BT}$ denote the partitioned blocks for the minority training text data. Thus, $\mathbf{BT} = \{B_1, B_2\}$, where $\mathbf{B_i}$ indicates the ith block. The feature set of block number $\mathbf{B_1}$ and $\mathbf{B_2}$ be $\mathbf{F_1}$ and $\mathbf{F_2}$ respectively where $\{F_1\} \cap \{F_2\} = \emptyset$. The sub pattern indicated by $\mathbf{X_A}$ is the projection of data instance $\mathbf{x}$ on to the subset of features of each block $\mathbf{B_i}$.

2.2 Genetic Approach for Efficient Nearest Classifier Model

The GA begins with an initial population of minority instances from the training data represented by absence or presence of word indicated by 0 and 1 respectively (Binary representation). Two approaches of crossover are applied in this paper. For the crossover process, the minority training set is partitioned vertically to two or three sub dataset based on random selection of split points among the attributes. The random selection of attributes is based on one point and two point crossover.

After generation of samples, the synthetic instances are selected based on the closest similarity to the existing minority instances. A cutoff is maintained on the distance measure. This process of genetic generation of instances is repeated until maximum efficiency can be attained or until the ratio of instances between majority and minority is maintained.

Algorithm-1: Explains the genetic process for generating synthetic samples. It is as follows:

Step 1: Accept Minority Training Instances as Initial Population, $\mathbf{TR}^{+}$ as $\mathbf{TR}^{S+}$

Step 2: Evaluate the required minority instances on comparison with number of majority instances.

Step 3: Create new populations by repeating the following operation.

A. Apply crossover operation on $\mathbf{TR}^{S+}$

B. Evaluate the synthetic instances closest to the existing instances in $\mathbf{TR}^{S+}$ based on distance criterion.

C. Update population with the selected synthetic instances.

Step 4: Repeat Step 2

2.3 Artificial Pattern Generation by Crossover

Investigation of two techniques of crossover is been briefed here. For one point crossover, a random feature is selected to split the Boolean representation of the minority sample into two. An example is shown in Figure 1. Parent 1 and Parent 2 will form two instance at cross-site 6 (representing a feature of an instance). The first part of the 1st parent minority training instance is cross joined with the second part of the parent 2 minority training instance and vice versa to create two synthetic instances as children.

This process leads to generation of m*(m-1) synthetic instances where m denotes the number of minority training instances.

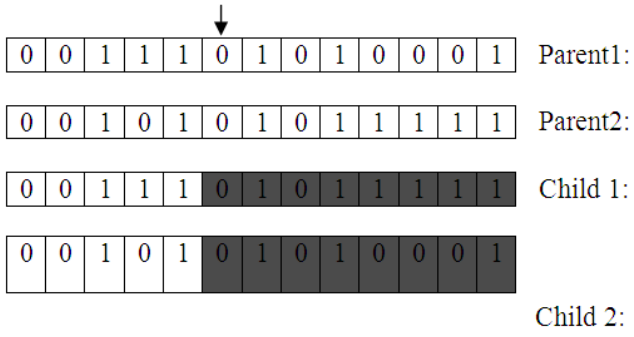

Figure 1. One split crossover.

In case of two-point crossover, the point of crossover will be between two features. Figure 2 explains this technique. If the cross-site 1 is four and cross-site 2 is ten, the strings between four and ten are exchanged between the parent 1 and parent 2.

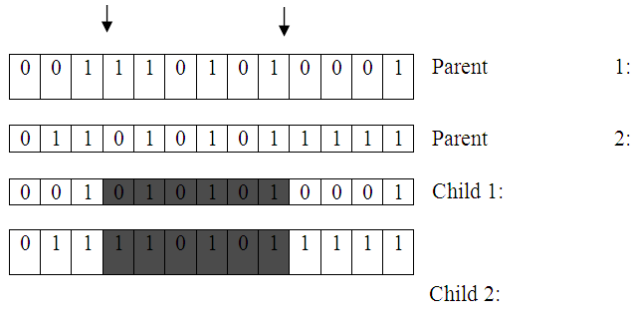

Figure 2. Two split crossover.

Description of the Nearest Neighbor (NN) classifier model and the generation of synthetic patterns by min_JOIN1 are explained here. Cartesian product of the Nearest Neighbor sub patterns produces the synthetic instances for the minority instances. The synthetic instances are produced till the balance ratio between the classes is maintained. The balance ratio acts as the fitness function. NN classifier is trained on this balanced training dataset.

2.3.1 Formal Description of the k Neighbor Classifier with Crossover Technique

- One Split Crossover:

This section first explains the implementation of one split crossover. The centroid C of the initial population of

the minority instances is computed. Assume the minority instances of the training data; TR^+ be partitioned vertically into two blocks; $BT = \{B_1, B_2\}$ at the random split feature selected. Assuming split at Feature f_n, then the Block features of B_1 are $\{f_1, f_2, \ldots f_n\}$ and B_2 are $\{f_{n+1}, f_{n+2}, \ldots f_p\}$. Cartesian product formed between the sub patterns of the blocks are indicated as TR^S . Only instances closer to centroid C (identified from TR^+), indicated by TR^{S+} are augmented to the the original majority training data, TR^- . The process of generation of synthetic samples continues until balance distribution between classes exists.

Therefore, $|TR^S| = m * (m-1)$. The classification model applied here is the k Nearest Neighbor. Algorithm 2 describes the modified Nearest Neighbor classifier, k_neighbor for the synthetic dataset generated by min_JOIN1.

Algorithm - 2: k_neighbor:

Step 1: Perform min_JOIN1 to generate synthetic minority set, TR^{S+}

Step 2: Find the k Nearest Neighbor of the Test pattern T from $TR^+ \cup TR^-$. Let it be denoted as δ.

Step 3: Classify T to the class based on the majority vote among the k global neighbors from δ

To generate synthetic_minority data TR^{S+}, algorithm 3: min_JOIN1 is implemented.

To generate the synthetic minority instances using GA, Algorithm 3 is shown below.

Algorithm - 3: min_JOIN1:

Input Data:

A Partitioned Minority Training data TR^+ , into Blocks $= \{B_1, B_2\}$.

Output Data

A Synthetic_Minority Instances, TR^{S+}

From block B_1 ε B,

Let the instances with respect to feature projection in B_1 be S_1 where $S_1 = \{X_{B_1}^{11}, X_{B_1}^{12}, \ldots, X_{B_1}^{1m}\}$

From block B_2 ε B,

Let the instances with respect to feature projection in B_2 be S_2 where

$S_2 = \{X_{B_2}^{21}, X_{B_2}^{22}, \ldots, X_{B_2}^{2m}\}$

$TR^S = (S_1 \times S_2)$.

Calculate Centroid C, using k means for TR^+

Compute **dist** (X_i, C) where $X_i \in (S_1 \times S_2)$,

Sort **dist** (x_i, C) , in ascending order

Populate $TR^{S+} = X_i, \quad where\ X_i \in (S_1 \times S_2)$ wrt dist (x_i, C) and $q < (m * m)$

synthetic_Minority = TR^{S+}

Two Split Technique

This technique is similar to the implementation of the one split technique except that the features between the two split points are merged into one block and the remaining features into the other block. The cross join procedure can be followed with respect to the two blocks. The procedure of selection is the same as mentioned in one split technique.

2.4 Fitness Function

The genetic approach was taken so that minimum synthetic instances were generated for the minority class without compromising on the classification accuracy of negative or majority instances. It is not necessary there must exist equal representation of instances between the minority and majority samples for the ideal classification accuracy. A 10-fold cross-validation procedure on the training data will be used to assess the effectiveness of the solution produced after each crossover. For every iteration, m*(m-1), minority instances are generated. The user can decide the selection of synthetic samples based on the Nearest Neighbor criterion to the existing training samples. Once instances are produced and augmented with the original training data; at some random point, Nearest Neighbor classifier is run on the extended training data. The predictive model is performed by following the 10-fold validation procedure. If the average classification performance of minority samples improves, without affecting the classification of majority samples, then the synthetic samples are generated are considered sufficient and maximum.

3. Experimental Results

3.1 Dataset Description

Text corpora from[23] were used for. The TDT2 and Reuters-21578 were used. The TDT2 corpus (Nist Topic

Table 1. Description of datasets from TDT2

Name -TDT2	Number of Classes	Examples(+,-)	Number of Attributes	Reduced Features Set	Synthetic Instances Produced both Split 1 and Split 2 approx
mat48	2	71:1821	15833	5603	213+
mat35	2	10:1844	21154	7855	152+
mat29	2	56:1828	15462	5507	112+

Table 2. Description of datasets from Reuters 21578

Name -TDT2	Number of Classes	Examples(+,-)	Number of Attributes	Reduced Features Set	Synthetic Instances Produced both Split 1 and Split 2 approx
mat39	2	298:3713	9600	3482	1998+
mat43	2	142:3713	8559	2969	224+
mat30	2	24:2055	7298	3380	127+

Detection and Tracking corpus) consists of data collected in 1998 and was taken from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). It consists of 11201 on-topic documents which are classified into 96 semantic categories[23]. Reuters-21578 corpus contains 21578 documents in 135 categories. The documents with multiple category labels are discarded. 8293 documents in 65 categories were remaining. However, for investigating imbalanced text data, the 2 classes preprocessed data has been used in both categories[23] as shown in Tables 1 and 2.

Table 3. Comparison between measures for the NN classifier for TDT2

Measusure	Classifier	mat48	mat35	mat29
AUC	**k- nearest classifier**	97.86	95.00	99.14
	k-neighbor(One Split)	99.79	99.97	99.89
	k-neighbor(Two Split)	99.82	99.98	99.95
F Measure	**k- nearest classifier**	97.81	90.00	98.33
	k-neighbor(One Split)	99.79	98.73	98.91
	k-neighbor(Two Split)	99.82	99.99	99.73
Geometric Mean	**k- nearest classifier**	97.77	90.00	99.10
	k-neighbor(One Split)	99.79	99.97	99.89
	k-neighbor(Two Split)	99.82	99.98	99.95
Sensitivity	**k- nearest classifier**	95.71	90.00	90.00
	k-neighbor(One Split)	99.58	100	100
	k-neighbor(Two Split)	99.64	100	100
tprt	**k- nearest classifier**	0.9571	0.9000	0.9833
	k-neighbor(One Split)	0.9958	1	1
	k-neighbor(Two Split)	0.9964	1	1

3.2 Results Obtained

Tables 1 and 2 shows the datasets that were been trained on by the classifier. The third column of Tables 1 and 2 describes the number of minority samples as against the majority instances. The features set were further reduced by following the Zip's Law. The reduced feature set is shown for all datasets in Tables 1 and 2. We remove the most frequent and least commonly occurring terms. Time and space requirement were reduced without effect in the classification performance. The Tables also indicate the number of synthetic instances produced for effective classification of minority instances without increasing the misclassifications of the majority instances. For example, in mat48, under TDT2, the number of synthetic minority instances produced was 213 instances more.

Classification Accuracy is the most common performance used to evaluate the performance of classifier models. This evaluation metric is insensitive to the data distribution. It does not indicate the degree of misclassification of minority instances. The performance measure,

therefore considered here is the Precision, Sensitivity, F-Measure and Geometric mean. All the mentioned measures use parameters that are formed by using a confusion matrix.

Table 4. Comparison between measures for the NN classifier for Reuter-21578

Measusure	Classifier	mat39	mat43	mat30
AUC	**k- nearest classifier**	97.97	87.00	94.17
	k-neighbor(One Split)	99.94	99.54	98.36
	k-neighbor(Two Split)	99.96	99.97	99.98
F Measure	**k- nearest classifier**	97.98	85.06	93.81
	k-neighbor(One Split)	99.92	99.47	97.88
	k-neighbor(Two Split)	99.96	99.90	99.71
Geometric Mean	**k- nearest classifier**	97.94	85.73	92.84
	k-neighbor(One Split)	99.97	99.54	98.32
	k-neighbor(Two Split)	99.98	99.97	99.98
Sensitivity	**k- nearest classifier**	95.97	74.00	88.33
	k-neighbor(One Split)	99.78	99.12	96.78
	k-neighbor(Two Split)	100	100	100
tprt	**k- nearest classifier**	0.9597	0.7400	88.33
	k-neighbor(One Split)	0.9770	0.9912	96.78
	k-neighbor(Two Split)	0.9992	1	1

As in Tables 3 and 4, the values were plotted against AUC, Sensitivity, Geometric Mean and True Positive rate. As explained in the paper, the Nearest Neighbor model initially is plotted against the original imbalanced dataset. The synthetic instances are produced by cross-join under two crossover techniques - k-neighbor (One Split) and k-neighbor(Two Split). Each instance produced is only selected depending on the similarity with existing minority instance. The similarity is measured through the closeness with the centroid of the existing minority instances. The advantage of this method is that the number of instances produced necessarily does not have to balance the distribution among the majority and minority class. This process of generation needs to be continued until maximum efficiency is attained among the training data. Every generation of synthetic instances, a random feature is selected for split point. This introduces less redundancy among the artificial samples generated.

The approach produced synthetic samples on a trial and run basis by performing 10 cross validation on the training data itself. This is done so that the synthetic instances do not affect the performance of the majority and minority samples. When the average performance of the training data improves, the training data is appended with the synthetic instances. The model trained on the modified training data is run on the test set and values plotted.

It is observed with all datasets that the evaluation parameters have shown an increase in performance. Both one and two-split point crossover technique has shown appreciable increase in performance. The two point crossover technique have shown a slight more better performance through F-Measure, Geometric Mean, AUC and sensitivity. This also indicates the goodness of the algorithm of the two- point crossover. G-Mean measure is to maximize the accuracy on both the classes while these accuracies are still balanced.

4. Conclusion

The paper worked on improving the classification of imbalanced text corpora. The model incorporated genetic approach with oversampling technique. The oversampling technique generated synthetic instances for the underrepresented class. The instances were formed by the one point and two point crossover technique. The genetic approach was incorporated so that instances generated produced did not affect the performances of the majority instances. The instances produced did not necessarily bring a balanced distribution among the classes. This hybrid approach was effectively used to evaluate the model on six different datasets. The evaluated results of the hybrid approach have indicated the superiority of both the method.

The paper need to further incorporate the minimum error criterion of the selected synthetic instance with

respect to existing minority instances in the training data. The initial population can be selected with certain criteria that decide the best candidates for crossover. The performance of the model can further be investigated using error criterion as the fitness function.

5. References

1. Japkowicz N. Learning from imbalanced data sets: A comparison of various strategies. Proceedings of Learning from Imbalanced Data Sets, AAAI Work Shop. Technical Report; 2000.
2. Liu Y, Loh HT. Corpus building for corporate knowledge discovery and management: A case study of manufacturing. Proceedings of the 11th International Conference on Knowledge-based and Intelligent Information and Engineering Systems, KES'07, Lecture notes in artificial intelligence, LNAI. Vietri sul Mare, Italy. 2007; 4692:542–50.
3. Weiss GM. Mining with rare cases. O. Maimon, L. Rokach, editors. The Data Mining and Knowledge Discovery Handbook, Springer; 2005. p. 765–76.
4. Weiss GM, The impact of small disjuncts on classifier learning. R. Stahlbock, S. F. Crone, S. Lessmann, editors. Data Mining: Annals of Information Systems. Springer. 2010; 8:193–226.
5. Cortes C, Vapnik V. Support Vector Networks. Machine Learning. 1995; 20:273–97.
6. Denil M, Trappenberg T. Overlap versus imbalance. Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence (CCAI'10), Lecture Notes on Artificial Intelligence. 2010; 6085:220–31.
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligent Research. 2002; 16:321–57.
8. Seiffert C, Khoshgoftaar TM, Van Hulse J, Folleco A. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Information Sciences; 2013. Available from: http://dx.doi.org/10.1016/j.ins.2010.12.016
9. Han H, Wang WY, Mao BH. Borderline–SMOTE: A new over–sampling method in imbalanced data sets learning. Proceedings of the 2005 International Conference on Intelligent Computing (ICIC'05), Lecture Notes in Computer Science. 2005; 3644:878–87.
10. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN'08); 2008. p. 1322–8.
11. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe–level–SMOTE: Safe–level–Synthetic Minority Over–Sampling Technique for handling the class imbalanced problem. Proceedings of the 13th Pacific–Asia Conference on Advances in Knowledge Discovery and Data Mining PAKDD'09; 2009. p. 475–82.
12. Stefanowski J, Wilk S. Selective pre-processing of imbalanced data for improving classification performance. Proceedings of the 10th International Conference on Data Warehousing and Knowledge, Discovery (DaWaK08); 2008. p. 283–92.
13. Batuwita R, Palade V. Class imbalance learning methods for Support Vector Machines. H. He, Y. Ma, editors. Imbalanced Learning: Foundations, Algorithms and Applications; Wiley. 2013. p. 83–96.
14. Seiffert C, Khoshgoftaar TM, Van Hulse J, Folleco A. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Information Sciences. 2013. Available from: http://dx.doi.org/10.1016/j.ins.2010.12.016
15. Garcia-Pedrajas N, Perez-Rodriguez J, Garcia-Pedrajas M, Ortiz-Boyer D, Fyfe C. Class imbalance methods for translation initiation site recognition in DNA sequences. Knowledge Based Systems. 2012; 25(1):22–34.
16. Choi JM. A selective sampling method for imbalanced data learning on Support Vector Machines. Iowa State University. 2010.
17. Gunal S. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Sciences. .2012; 20(2):1296–311.
18. Uysal AK, Gunal S. Text classification using Genetic Algorithm oriented latent semantic features. Expert Systems. 2014; 41(13):5938–47.
19. Wasikowski M, Chen XW, Combating the small sample class imbalance problem using feature selection. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(10):1388–400.
20. Choi JM. A selective sampling method for imbalanced data learning on Support Vector Machines. Iowa State University. 2010.
21. Maheshwari S, Agrawal J, Sharma S. A new approach for classification of highly imbalanced datasets using evolutionary algorithms. International Journal of Scientific and Engineering Research. 2011 Jul; 2(7):1–5. ISSN 2229-5518.
22. Maheshwari S, Agrawal J, Sharma S. A new approach for classification of highly imbalanced datasets using evolutionary algorithms. International Journal of Scientific and Engineering Research. 2011 Jul; 2(7):1–5. ISSN 2229-5518.
23. Available from: http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html