Imbalanced dataset classification algorithm based on NDSVM

Yueting LIU

School of Communication Engineering, Lanzhou University of Arts and Science, Lanzhou, GanSu, 730000, CHN

Corresponding author’s e-mail: 1000264@luas.edu.cn

Abstract: Because of uneven distribution and indistinct boundary in imbalanced dataset, imbalanced dataset classification algorithm based on neighbors density support vector machine (NDSVM) is proposed. In this algorithm, the neighbor range density of each sample in the majority class is calculated firstly. According to the density value, the data which on the majority class border or close to the border is equal to the minority samples in quantity, which are selected, then the minority class complete SVM initial classification. Then the resulting support vector machine and residual data in the majority class optimize the initial classifier. The simulation results of experiments on the manual and UCI dataset show that compared with WSVN, ALSMOTE-SVM and SVM, NDSVM has better classification performance, which effectively improve the classification performance of SVM algorithm on the uneven distribution and indistinct boundary in imbalanced dataset.

1. Introduction
Classification is defined as analyzing and learning input training samples to get decision model, and then predicting unknown samples. It has become an important research direction in the field of machine learning. At present, many classical algorithms can achieve good classification performance of balanced data, such as support vector machine, fuzzy classification algorithm, cost sensitive learning method and decision tree algorithm. However, many application areas have obvious unbalanced data, such as network intrusion, commercial fraud, text classification and other datasets[1-3], people attach great importance to the information of the minority class. The traditional classifier always tend to the majority class in the classification decision. The minority samples are wrongly assigned to the majority class, resulting in a very high rate of error, and the performance of the classifier is not ideal[4]. Therefore, how to improve the classification performance of unbalanced data has become a research hotspot[2].

At present, the research on imbalanced data classification mainly includes data processing and algorithm improvement. Data processing includes under sampling, over sampling and mixed sampling. Under sampling, the datasets is balanced by reducing the samples of the majority class, which may result in the loss of information and reduce the performance of the classifier[3]. Over sampling, the datasets is balanced by copying and interpolation to add the minority samples, but it can lead to over fitting, and increase the space and time consumption of the classifier[4]. Mixed sampling is an effective combination of under sampling and over sampling to balance the distribution of datasets.

Algorithm improvement is the operation of the classification algorithm itself, including the improvement of the traditional algorithm, the integration of many algorithms, etc. The improved algorithm is mainly to adjust the bias of the algorithm on the datasets by adjusting the classification boundary, changing the probability density, which make the decision surface tend to the minority class
in order to improve the classification performance\cite{5}. Literature\cite{6} proposed SVM pin down the number of neighbors firstly, then complete the classification by the KNN algorithm. Literature\cite{7} proposed WSVM algorithm, which selects the majority samples different effects on the classification surface by clustering weights the majority class of selected samples and the minority class complete SVM training. The characteristics of imbalanced datasets are presented in literature \cite{7-8}. Based on the idea of literature\cite{7-8}, a balance method based on neighbor density is put forward. According to the density values in the $K$ neighborhood of the majority class, the samples are selected by different effects on the classification surface. According to the function of the classification surface, the majority samples are reorganized to train with the same number of the minority samples.

### 2. SVM

SVM based on statistical learning theory and the structural risk minimization principle to find the optimal classification surface. Nonseparable problem in the original feature space can be mapped to a higher dimensional feature space by kernel function, which is transformed into a quadratic programming problem.

Given samples $(x_i, y_i), i = 1, 2, \ldots, n, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$, SVM uses the maximum interval to find the decision classification hyper plane, and the decision function is

$$f(x) = w^T x + b$$  \hspace{1cm} (1)

in the formula, $w \in \mathbb{R}^n$ are weight vectors, $b \in \mathbb{R}$ is threshold value.

Constructing SVM optimization model

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, n$$ \hspace{1cm} (2)

in the formula, $\xi_i$ is slack variable, which reflects the degree of the samples are wrongly classified. $C$ is penalty constant.

To solve the formula (2) quadratic programming problem, Lagrange function is

$$L_{(w, b, \xi)} = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$ \hspace{1cm} (3)

in the formula, $\alpha_i$ and $\beta_i$ are lagrange operator. Take formula (3) partial respect to $w$, $b$, $\xi_i$ respectively, then set zero to get the formula (2) dual problem.

$$\max_{\alpha} \quad - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n} \alpha_i$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$$ \hspace{1cm} (4)

According to the KTT. That is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

$$b^* = y_j - \sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x_j)$$ \hspace{1cm} (5)

The discriminant function is

$$f(x) = \text{sgn} \left( \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^* \right)$$ \hspace{1cm} (6)

in the formula, $K(x_i, x_j)$ is a kernel function. The kernel function in input space instead of the inner product of the two training samples in high dimensional feature space. That is
3. Neighbors density support vector machine

3.1. Problem analysis

In reality, a lot of imbalanced datasets have the characteristics of uneven distribution and fuzzy class boundary, or have different density in different regions of the data space. Which lead to the traditional classification algorithm is difficult to achieve the desired results. The uneven density problem is illustrated by the dataset of Figure 1. Figure 1 dataset, a, b, c, respectively, on behalf of the scope of the 3 clusters, the density relationship a>b>c. the traditional classification algorithm is difficult to find a unified neighborhood radius to find 3 categories a, b, c.

![Fig1. The uneven distribution dataset](image)

Generally speaking, it is very difficult to get the density distribution of the whole dataset. So determine the region of the sample with the $K$ neighborhood density value in this paper. The density of the sample can be calculated by the reciprocal of the average distance of all samples in the neighborhood. The average distance is smaller, the density is bigger, the samples are more dense, the sample is located in the region close to the class center. On the contrary, the average distance is larger, the density is smaller, the samples are more sparse, the sample is located in the region close to the class boundary.

3.2 Near neighbors density

Definition: near neighbors density. A dataset $Z$ that contains $q$ data samples of $m$ dimension. For any $Z( z_i \in Z ,1 \leq i \leq q$), it has $K$ near neighbors $z_i(k)$. Formula (8) is euclidean distance between $z_i$ and $z_i(k)$. Formula (9) is $K$ near neighbors density of sample $z_i$.

\[
d(z_i,z_i(k)) = \sqrt{\sum_{i=1}^{K} (z_i - z_i(k))^2}
\]

\[
\rho_i = \frac{K}{d(z_i,z_i(k))}
\]

Form formula (9), $K$ near neighbors density of sample $z_i$ is the reciprocal of the euclidean distance mean from $z_i$ to $K$ near neighbors. The larger the $\rho_i$ value, the more dense the near neighbor of $z_i$. Otherwise, the smaller the $\rho_i$ value, the more sparse the near neighbor of $z_i$.

3.3 Algorithm process

In this paper, a SVM algorithm based on near neighbors density for the classification of unbalanced datasets is proposed. The neighbor range density of each sample in the majority class is calculated firstly. According to the density value, the data which on the majority class border or close to the border is equal to the minority samples in quantity to ensure the balance of the two classes. Then the
resulting support vector machine and residual data in the majority class optimize the initial classifier. NDSVM algorithm process is

Step1: Initialize variable. The minority class set is $P$. The samples number in the minority class is $p$. The majority class set is $N$. The samples number in the majority class is $n$. The set the majority of samples in descending order according to the density value is $N_{\text{order}}$. $N_{\text{order}_\text{behind}}$ is a set of $N_{\text{order}}$ the last $p$ samples. $N_{\text{order}_\text{before}}$ is a set of $N_{\text{order}}$ the $p$ samples before last. $N_{\text{order}_\text{other}}$ is a set of $N_{\text{order}}$ the remaining samples.

Step2: The majority class is isolated from the training samples.

Step3: Select sample $i$ from set $N$. Calculate the subset radius by formula (8). Taking $n_i$ as the center, and $r$ as the radius, which constitutes a subset $SC(n_i)$. Calculate the density of each sample in the $SC(n_i)$. In turn, calculate the density of all the samples in set $N$. All samples in set $N$ are listed in descending order by density value. Then that is set $N_{\text{order}}$.

Step4: Judge the relationship between $p$ and $n$. If $p < n \leq 2p$, the training set can be considered as a balanced set. Train it with SVM. If $n > 2p$, the training set can be considered as an imbalanced set. Transfer to Step5.

Step5: Set $P$ and set $N_{\text{order}_\text{behind}}$ compose a balanced set $M_{PN1}$. Train $M_{PN1}$ with SVM. Get support vector machine $S_{PN1}$. The number of support vector machine of majority class is $n_{\text{neg1}}$. The number of support vector machine of minority class is $n_{\text{pos1}}$.

Step6: Set $P$ and set $N_{\text{order}_\text{behind}}$ compose a balanced set $M_{PN2}$. Train $M_{PN2}$ with SVM. Get support vector machine $S_{PN2}$. The number of support vector machine of majority class is $n_{\text{neg2}}$. The number of support vector machine of minority class is $n_{\text{pos2}}$.

Step7: It can reduce the training time by using the support vector set instead of the training set, which has no influence on classification accuracy. $M_{PN3}$ is composed of $S_{PN1}$, $S_{PN2}$, and $N_{\text{order}_\text{other}}$. Extract $(n_{\text{pos1}} + n_{\text{pos2}} + n_{\text{neg1}} + n_{\text{neg2}})$ of all support vector set from $M_{PN3}$. Extract $(n_{\text{pos1}} + n_{\text{pos2}} + n_{\text{neg1}} + n_{\text{neg2}})$ of the majority samples. SVM classifier complete the iterative training, and complete the support vector update. When $T < 0.9$ is satisfied, return to Step4. Otherwise, the iterative training stops. The classification results will output.

$$T = \frac{\text{The total number of support vector set of majority class}}{\text{The total number of support vector set of minority class}}$$  \hspace{1cm} (10)

4. Experimental analysis

4.1. Evaluation index

According to the characteristics of imbalanced datasetss, the performance indexes of traditional classifiers are severely deficient. After research, scholars put forward the following indicators: in imbalanced dataset, the minority class is $P$, and the majority class is $N$.

- TP: the number of samples the actual positive class is predicted as positive.
- FN: the number of samples the actual positive class is predicted as negative.
- FP: the number of samples the actual negative class is predicted as positive.
- TN: the number of samples the actual negative class is predicted as negative.

(1) Recall: The correct rate of the minority class samples.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
(2) Specificity: The correct rate of the majority class samples.

\[ \text{Specificity} = \frac{TN}{TN + FP} \]

(3) Precision: Precision is the ratio of positive class samples correctly classified to all samples divided into positive class.

\[ \text{Precision} = \frac{TP}{TP + FP} \]

(4) G-mean: The classification performance of minority and majority class. If the classifier is biased in one class, it will affect the correct classification rate of the other.

\[ G\text{-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \]

(5) F-measure: F-measure is an evaluation method combination of recall and precision. which responds the classifier sensitivity the minority class samples.

\[ \text{F-measure} = \frac{2 \times \text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}} \]

(6) AUC: AUC is a kind of evaluation tool for unbalanced data, which calculate area under the curve of ROC(Receiver Operating Characteristic). It can fully describe the performance of the classifier in different decision thresholds.

\[ \text{AUC} = \frac{1 + \frac{TP}{FP + FN} - \frac{FP}{TN + FP}}{2} \]

### 4.2 Comparison of experimental results for each algorithm

In order to verify the feasibility of this algorithm, the program is wrote by Matlab2014a. Dataset and UCI datasets are selected as experimental data. Compare the test results with the literature [5] ALSMOTE-SVM algorithm, the literature[8]WSVM algorithm and SVM algorithm. In Figure 2, the Dataset is a uneven artificial dataset, which contains 1018 data points. As shown in table 1, select light rate of imbalance such as Iris and Glass Identification datasets and high rate of imbalance Spectf Heart and Ecoli datasets from the UCI library. The SVM classifier parameters setting: select the Gauss function as the kernel function, \( \sigma = 1 \), \( C = 1000 \), \( \alpha = 0.2 \), iterative experiments run 20 times. The results of 4 algorithms on 5 datasets are shown in Table 2. As shown in Figure 3~Figure 5, the AUC curve results of 4 algorithms on Dataset, Glass Identification and Spectf Heart datasets.

![Fig2. Artificial dataset](image)

| Table1. Experimental dataset information |
|-----------------------------------------|
| dataset | total sample | attribute | class number | sample distribution after treatment | training set(positive,negative) | proportion of training | Test set(positive,negative) |
|---------|-------------|-----------|--------------|----------------------------------|-----------------|------------------------|--------------------------|
| Iris    | 150         | 5         | 3            | before treatment 50               | 0.33            | 0.33                   | 0.33                     |
| Glass   | 214         | 9         | 2            | before treatment 50               | 0.33            | 0.33                   | 0.33                     |
| Spectf Heart | 266    | 13         | 3            | before treatment 50               | 0.33            | 0.33                   | 0.33                     |
| Ecoli   | 156         | 5         | 33           | before treatment 50               | 0.33            | 0.33                   | 0.33                     |
Table2. Comparison of experimental results

| dataset              | SVM    | ALSMOTE-SVM | WSVM    | NDSVM   |
|----------------------|--------|-------------|---------|---------|
|                      | G-mean | F-measure   | G-mean  | F-measure |
| Iris                 | 94.3   | 94.0        | 97.4    | 94.6    |
| Class Identification | 93.3   | 92.9        | 95.2    | 94.8    |
| Spectf Heart         | 75.0   | 73.8        | 89.9    | 90.6    |
| Ecoli                | 80.3   | 80.0        | 90.6    | 92.8    |

The value of AUC, namely the area between ROC curve and the horizontal axis, the value of AUC can reflect the performance of classifier. The larger the area is, the better the classifier performance is. From the Figure 3–Figure 4, For light rate of imbalance Glass Identification dataset, the AUC value got by 3 improved SVM algorithms are higher than value got by SVM. NDSVM algorithm performance is best among 3 improved SVM algorithms. NDSVM based on the local density and division of the majority class is feasible. For high rate of imbalance Spectf Heart dataset, the AUC value got by NDSVM is higher than value got by SVM is equal to 0.797. Calculate the local density of each sample in the majority class. According to the density value, the data which on the majority class border or close to the border is equal to the minority samples in quantity to ensure the balance of the two classes. NDSVM classifier performance is better than other algorithms, which verify the feasibility of the NDSVM algorithm.

5. CONCLUSION

Because of uneven distribution and indistinct boundary in imbalanced datasets, imbalanced datasets classification algorithm based on neighbors density support vector machine (NDSVM) is proposed. In this algorithm, the neighbor range density of each sample in the majority class is calculated firstly.
According to the density value, the data which on the majority class border or close to the border is equal to the minority samples in quantity, which are selected and the minority class complete SVM initial classification. Then the resulting support vector machine and residual data in the majority class optimize the initial classifier. Experimental results show that NDSVM has good classification performance on imbalanced datasets, and proves the feasibility and effectiveness. How to better coordinate the value of relevant parameters and reduce the time complexity is the goal of further research in future.

Acknowledgments
This paper was supported by scientific research project of Gansu Province in 2015(No. 2015B-132).

References
[1] Zhang J.J.,(2015)Research on Support Vector Machine Models and Algorithms for Imbalanced Data. Beijing: China Agricultural University.pp.25-34.
[2] Li Y.,Liu Z.D.,Zhang H.J.,(2014)Review on ensemble algorithms for imbalanced data classification.J.Application Research of Computers,31(5):1287-1291.
[3] Wilson D R,Martinez T R.,(2000)Reduction techniques for instance-based learning algorithms.J.Machine Learning,38:257-286.
[4] Chawla N V, Bowyer K W, Hall L O,et al.,(2002)SMOTE:synthetic minority over-sampling technique.J.Journal of Artificial Intelligence Research,16:321-357.
[5] Jun M.,(2014)Research on classification algorithm for imbalanced datasets.Jiangshu: Nanjing University of Science and Technology.pp.29-43.
[6] Wang C.H.,Lee Lam Hong,R.J. R,et al.,(2012)Hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine.J.Expert Systems with Applications,39:11880-11888.
[7] Wang C.X.,Zhang T.,Ma C.S.(2015)Resolution of classification for imbalanced dataset based on cluster-weight and grading-SVM algorithm.J. Computer Engineering and Applications,51(21):133-137.
[8] Liu W.L.,Liu S.Y.,Xue Z.X.,(2008)Balance method for imbalanced support vector machines.J.Pattern recognition and artificial intelligence,21(2):136-141.