Research Article

Application of Big Data Unbalanced Classification Algorithm in Credit Risk Analysis of Insurance Companies

Xian Wu and Huan Liu

Science and Technology Finance Key Laboratory of Hebei Province, Hebei Finance University, Baoding, Hebei 071051, China

Correspondence should be addressed to Xian Wu; wuxian@hbfu.edu.cn

Received 10 January 2022; Revised 7 February 2022; Accepted 28 February 2022; Published 25 March 2022

Academic Editor: Miaochao Chen

Copyright © 2022 Xian Wu and Huan Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The 2008 global financial crisis triggered by subprime mortgage crisis in the United States and the ongoing European debt crisis have urged governments and academics to pay high attention to financial industry risk supervision. The financial industry has actively implemented comprehensive risk management. As an important component of the financial industry, the insurance industry implements comprehensive risk management to control the risks of insurance companies. Propose an integrated learning model based on imbalanced dataset resampling and apply it to UCI dataset (University of California Irvine). First, resampling technology is used to preprocess the unbalanced dataset to obtain a relatively balanced training set. Then, use the classic backpropagation neural network, classic k-nearest neighbor, and classic Naive Bayes three algorithms as the base classifier and use the Bagging strategy to get the ensemble learning model. In order to verify its effectiveness, F-measure and G-mean methods are used to measure the performance of the classifier. The subject mainly focuses on the classification of relevance vector machine (RVM) in two types of large-scale datasets, imbalanced and balanced, and proposes solutions for these two types of data. This explains the effectiveness of the disequilibrium classification algorithm used in the risk analysis of insurance companies.

1. Introduction

Bank credit risk, control risk, and commercial risk are the main risks facing my country’s current financial market. Among them, credit risk is the main type of financial risk. It not only affects the development of microeconomic individuals but also has a huge impact on the stability of the entire macroeconomy. The manifestation of credit risk is becoming more and more complex, and the risk is gradually expanding. Insurance companies are also facing huge challenges in credit risk management [1]. With the improvement of my country’s socialist market economic system and the deepening of financial system reforms, my country’s insurance industry has entered a period of rapid development. The internal and external environments for the operation and development of insurance companies are changing, and the corresponding risk management of insurance companies has become a necessary task [2]. In some fields, the importance of the minority class samples being correctly classified is often higher than that of the majority class. However, most classical classification algorithms assume that the prior probability distribution of the sample is balanced or the cost of misclassification is equal. In the face of unbalanced distribution of data, the information of the minority samples is often covered by the information of the majority samples, resulting in the classification error rate from the minority samples being much higher than the majority samples. Therefore, the research of imbalanced data classification has attracted more and more researchers’ attention [3]. With the continuous emergence of new technologies and the in-depth research of data mining theory, classifier design has become a hot issue in the field of data processing. The classifier design belongs to the category of predictive analysis. Its basic operation is to use the training set to learn the classifier model and test the performance of this model. People’s research on classifiers generally focuses on the classification of a large number of high-dimensional data and the classification of unbalanced data. A large
number of experimental studies have shown that we want to
find a classification model that can reflect the fit of all
datasets and optimize the performance of the classifier.
Therefore, in the design of the classifier, different datasets
usually need to use different classification models to achieve
the best [4]. As shown in Figure 1, with the continuous in-
depth study of classification learning algorithms by a large
number of scholars, people have discovered that there is a
large amount of unbalanced data of class distribution in the
real world [5]. Unbalanced data refer to the fact that, in the
same type of data centralized processing, some samples are
more than other samples. We define the class with more
samples as the majority class, and vice versa; the class with
fewer samples is called the minority class. Unbalanced data
are common in all areas of daily production and life. In
general, only when the ratio of the minority class to the
majority class distribution is less than or equal to 1:2, the
data have the characteristics of unbalanced class distribution
[6–9].

Many researchers have begun to study the problem of
unbalanced data classification. The critical parameter
analysis method proposed by Semeniuk and Povodzinskiy is
used to improve the formulation of technical specifications
for fermentors and for the development and scaling of
culture processes for active pharmaceutical ingredient
production in biopharmaceutical production [10]. Onan
et al. examined the performance of La tent
Dirichlet allocation in text sentiment classification in four
dimensions through a five-class algorithm and an ensemble
approach [11]. In addition, Onan, on the basis of empirical
analysis, proposes an ensemble classification scheme that
integrates random subspace ensembles of random forests
and four types of features [12]. By exploring purposeful
sampling methods, Nugrahanti found no empirical evidence
that the impact of corporate government mechanisms
covering corporate governance structure and corporate
governance quality can reduce the increase in earnings
management in the Indonesian banking sector [13]. Alek-
seeva and Milgunova, through a comprehensive assessment
of the risk of distortions in the information environment,
proposed the method which will allow one to form ac-
counting information and manage control procedures in a
systematic and high-quality manner [14]. Different from
these papers, in this study, we draw on the superior methods
they used in their research and compare and judge the
results through simulation experiments. Find a model
(classifier) of a class of suitable complexity that avoids
overfitting and apply it to the test dataset to predict the class
labels of unknown records. Therefore, the classifier shows a
good classification accuracy in the training sample set, which
solves the problem that the classification effect of new
samples is not ideal. In the traditional insurance company
credit risk analysis samples, the prior probability distribu-
tion is balanced or the cost of misclassification is equal. In
the real world, this assumption is not necessarily true. When
faced with unbalanced data, the information of minority
samples is greatly reduced. Overwhelmed by most samples,
the former has a much higher classification error rate than
the latter, and the generalization ability of the classifier is
poor. Therefore, the problem of unbalanced data classifi-
cation is one of the main problems facing the field of data
mining and machine learning [15]. In recent years, with the
continuous in-depth research on the problem of unbalanced
data, classification algorithms based on unbalanced data
have attracted a large number of researchers. International
conferences on the theme of unbalanced data classification
have been held all over the world. The methods to improve
the overall performance of the unbalanced data classifier
mainly include resampling technology at the data recon-
struction level and the improvement of the classic classi-
fication algorithm at the algorithm level and the proposed
new algorithm [16]. Under the background of the high
frequency of credit problems of current insurance com-
panies and the extremely uneven data of various attack
types, this study focuses on how to build a risk analysis
model with excellent performance and good classification
performance for a small number of attack types. The
problem is mainly analyzed and studied from the following
aspects [17]. (1) Introduce each method in ensemble
learning and analyze the advantages and disadvantages,
and finally, choose GBDT (Gradient Boosting Decision
Tree) algorithm for data classification. In the theoretical
part, the formula derivation process to realize GBDT is
listed in detail, and the focus is on how to apply GBDT to
the classification of insurance company risk analysis pro-
cess. (2) Through analysis and research, the corresponding
GBDT classifier is constructed using the one-to-one
method, combined with the processing of risk data by
DBN, and an insurance company credit risk security risk
model under unbalanced data is proposed [18, 19].

2. Comprehensive Risk Management for
Insurance Companies

Comprehensive risk management is considered to be a
natural evolution in the development of risk management.
We all know that traditional risk management focuses on
reducing or avoiding losses. The goal of minimizing losses is
to manage pure risks, and it is a silo management model in
which different types of risks are managed by different
departments. The globalization trend of the world economy
is intensifying, and financial crises frequently break out. The
risk relationships faced by companies are complex and often
exist in a complex form. Even a single form of risk is often
linked. The company assesses and comprehensively manages
all risks from a holistic perspective. Objective requirements
and practical needs are the objective conditions for the
formation of ideas [20]. The financial community has dis-
covered that losses can be caused by a combination of credit
risk and market risk and has paid more attention to the
integrated model of market risk and credit risk, as well as the
quantification of operational risk. Through more in-depth
consideration of risk prevention and management issues,
they realized the shortcomings and deficiencies of traditional
risk management, such as ignoring the correlation between
risks and the characteristics of the company’s own orga-
nizational structure, and fully realized that risk management
should be taken from the overall system level. In this context,
comprehensive risk management has gradually been paid attention to and valued, studied, and practiced by companies, institutions, and governments. So far, its theoretical research and practical application have developed greatly, and it is still a hot topic and academic frontier in the field of risk management. Its strategic decisions are perfectly applied and appear in the entire risk management process of the company. It is jointly executed by senior leaders, management, and other relevant personnel. The purpose is to identify potential problems and risks that may affect the subject and to provide reasonable assurance that they are now within the controllable range of the subject’s risk capability. It is implemented by people, and it also affects how people identify, evaluate, and respond to risks; it is used in strategy formulation to help management evaluate and select strategies and related goals; it runs through the entire company and is applied in various levels and unit activities. Implemented by personnel at all levels, we adopt a subject-level risk portfolio view [21]. The definition captures the key concepts of managing risk, provides a foundation for the application of different organizational forms, industries, and departments, and provides a basis for the validity of the definition. Therefore, this definition is generally accepted by academics and has a high degree of authority and recognition. This study is subject to this definition, and the relevant definitions of other organizations and institutions will not be repeated [22].

The comprehensive risk management system of an insurance company is a powerful system built on the basis of its connotation. The comprehensive risk management system of an insurance company includes all the content of comprehensive risk management, from the objectives, concepts, and policies of risk management, to supporting systems such as organizational structure, risk management system, and risk environment, as well as the core comprehensive risk management. In response to current policies, insurance companies’ own risk management capabilities will be included in the minimum capital under the first pillar as the capital requirements for risk control. Therefore, there is an urgent need to establish an effective and comprehensive risk management system that meets regulatory requirements [23]. In the comprehensive risk analysis of an insurance company, the company must first determine the business content according to its own situation and development direction, combined with the requirements of market supervision, accurately identify the various risks faced by the company, and then classify these risk factors, organize, including quantified objects that can be presented, and risks, such as operational risks and strategic and reputation risks, which are difficult to quantify, should be considered. For
quantifiable risks, risk measurement and evaluation must be carried out. For risks that are difficult to quantify, risk evaluation must be carried out in conjunction with risk preference [24]. Insurance companies should cultivate a mature risk management concept, with capital as the core, risk governance as the basis, risk preference as the orientation, and risk quantification tools and risk performance appraisal as the main means, to continuously improve risk management and technology and to dynamically control the company. Take the risk to achieve a balance between risk control and business development [25]. The last key step of risk management is risk monitoring and reporting. Risk reports are formed for the abovementioned risk identification, assessment, and processing and submitted to relevant departments for execution. In dynamic risk analysis, various risks and the ways in which the analysis affects the emergence of the enterprise are constantly changing. Dynamic risk refers to the risk related to social change; mainly the risk caused by the change of social economy, politics, technology, and organizational structure. The dynamic risk analysis of great importance due to negative factors, such as inflation, exchange rate risk, strikes, riots, consumer preferences change, and national policy changes, will produce dynamic risks for an insurance company [26]. With the development of the bank and the improvement of the management system, the original risk management system was retested. Risk quantification models and risk control methods are no longer applicable or need to be improved. Therefore, companies are required to dynamically monitor risk management and provide feedback on the monitoring results.

3. Unbalanced Data Classification Algorithm Based on the Gaussian Mixture Model

3.1. Synthetic Minority Oversampling Technique Algorithm. The classification of unbalanced datasets is one of the tasks of big data information processing. Generally, the application of traditional classification methods to unbalanced data sets will bring about problems such as the decrease of the classification accuracy of minority classes. The preprocessing of unbalanced datasets and the optimization of classifiers can be more effective. Improve the classification accuracy of minority classes. The default risk management of national student loans involves the classification of unbalanced datasets. Effective prevention and control of default risks have always been the focus of this field [27].

The existing classification algorithms for unbalanced data mostly sample existing datasets to reduce the degree of imbalance between classes. The usual approach is to oversample the minority class, then generate a new dataset, and add it to the original minority class, while undersampling the majority class to reduce the number of samples in the majority class, so that the data becomes balanced. However, most existing algorithms (such as SMOTE) do not take into account the properties of the data itself. SMOTE is a synthetically sampling artificial data algorithm used to fix the data class imbalance. If we can accurately obtain the distribution of the minority class samples, the newly generated samples will be more representative and can better describe the minority class. Then, use the base classifier to classify the data so that the classification results will be more accurate.

The SMOTE algorithm has been widely used. The core idea of the algorithm is to use the K nearest neighbors (KNN) algorithm for the minority data sample set. The definition of the algorithm is to consider that the K points in Smin have the smallest Euclidean distance between the high-dimensional feature space and x. In order to generate a composite sample, randomly select a point x in the nearest neighbor of K, to calculate the difference between the feature vector and the data point, multiply the result by any positive number less than 1, and finally add the x calculation formula as [28]:

\[ x_{\text{new}} = x_i + (\bar{x}_i - x_i) \times \delta. \]  

Among them, \( x_i \in S_{\text{min}} \) is the minority sample, which is a center point in the nearest neighbor of K. Therefore, the synthesized sample obtained by the above method is a randomly selected point on the connecting line between the sample point \( x_i \) and the center point \( x_i \).

As can be seen from the process of the SMOTE algorithm above, the way it synthesizes samples is by selecting random sample points between two sample points. This method does not thoroughly study the attribute characteristics of the entire minority sample set, and the new samples obtained in this way are very likely to lead to the overfitting of the model. Therefore, we consider starting from studying the sample distribution of the minority class and looking for a new sample oversampling method [29].

3.2. GMM Classification Algorithm. Different from the traditional SMOTE algorithm, here we use GMM to describe the minority data. After completing the modeling and estimation of model parameters, we can use GMM to oversampling to generate new minority samples. These newly generated data can change the distribution of unbalanced data and make the majority and minority classes of the sample become balanced. Finally, the Naive Bayes classifier is used to process the balanced dataset to complete the classification. The flow of the algorithm is shown in Figure 2 [30]. For building the framework based on unbalanced classification algorithm, this study adopts the collaborative filtering algorithm to accelerate the data calculation speed for reducing the real-time response time and enhance the scalability. Thus, the accuracy, grammar, and real time of reconstruction levels are improved.

The oversampling algorithm based on GMM is mainly divided into the following steps:

Select GMM to describe the distribution of minority data.

GMM can be regarded as a linear superposition of a set of single Gaussian models, but it can better describe the distribution of data than a single Gaussian model. The probability density function of GMM can be described by the following equation:
\[ p(x_i) = \sum_{k=1}^{K} p(x_i, z_i = k) = \sum \pi_k N(x_i). \] (2)

It can be seen from the above formula that GMM is composed of \( k \) single Gaussian models. We used the Gaussian model in the branches, and these branches are superimposed to form a GMM. In order to make the description of the model and the subsequent model parameter estimation more convenient, a hidden variable \( z_i \) is introduced. When \( x_i = k \), it can indicate that \( x_i \) comes from the branch. In addition, the joint probability distribution \( P(x_i, z_i) \) can be obtained by multiplying the marginal probability distribution of \( z_i \) and the conditional probability distribution. The definition of \( \pi_k \) is determined by

\[ p(z_i = k) = \pi_k. \] (3)

Similarly, for a given \( x_i \) and \( z_i \), we can use Gaussian distribution to express the conditional probability distribution:

\[ \gamma(i, k) = \frac{p(x_i = k) p(x_i, z_i = k)}{\sum_{k=1}^{K} p(z_i = k') p(x_i, z_i = k')} = \frac{\pi_k N(x_i, \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i, \mu_k, \Sigma_k)}. \] (4)

(2) Estimate GMM parameters based on minority samples.

Assuming that \( x \in S_{\text{min}} \) obeys a Gaussian mixture distribution (represented by \( p(x) \)), then the mean \( \mu_k \) and variance \( \Sigma_k \) of each single Gaussian model and the mixing coefficient \( \pi_k \) of the mixture model need to be determined. In order to find some parameters to maximize the probability of generating the model, \( N \) represents classes in the second-class set \( S_{\text{min}} \). The log likelihood function of GMM is shown in the following formula:

\[ \ln P(X) = \sum_{i=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k N(x_i, \mu_k, \Sigma_k) \right). \] (5)

After completing the parameter estimation of GMM, we can use the generated model to represent the distribution of the minority sample set. Next, we can use the obtained GMM to generate minority samples.

3.3. Naive Bayes Algorithm Classification Prediction. The most important thing to implement a prediction algorithm is to build a classifier. Naive Bayes has a very complete theoretical foundation. And because of its high computational efficiency and accuracy, it has been widely used. Therefore, we will use the Naive Bayes classifier as the base classifier for the imbalanced data classification and prediction algorithm. The Naive Bayes classifier relies on an assumption: the classifier’s classification feature condition is that the known attribute values are independent. Under the condition of the attribute independence of this assumption, the Naive Bayes classifier presents a simple structure. As a large number of scholars continue to study classification learning algorithms, there are a large number of unbalanced class distribution data in the real world. The GMM-Naive Bayes algorithm can update and modify the real-time data.

As shown in Figure 3, it can be seen that the effectiveness of the model is verified. Thus, it is necessary to select different ratios and randomly select different groups of survey data for testing in an unbalanced dataset. The ratios selected here are 1:10, 1:30, and 1:60. There are 5 subsets of data for each ratio to be verified, and each subset also contains 200 samples. We use Borderline-SMOTE-Naive Bayes to compare with the GMM-Naive Bayes algorithm proposed in this section.

We can see that the GMM-Naive Bayes algorithm performs better than the Borderline-SMOTE-Naive Bayes algorithm to dealing with unbalanced data. The reason is that the Gaussian mixture model can better describe the distribution of minority sample sets. Therefore, the newly generated data are more useful for training and testing.

In short, through the results of the above experiments, we can see that the GMM-Naive Bayes algorithm can improve the precision and accuracy of minority predictions. The sampling of minority classes is very necessary. Before a user complains about an operator’s failure, it is important to predict the user’s failure behavior. Therefore, our proposed
scheme is an effective method when dealing with this kind of unbalanced dataset.

We propose an oversampling method based on this model. We use the Gaussian mixture model to replace the $K$-nearest neighbor algorithm in traditional SMOTE algorithm to describe the distribution of the sample of users reporting failures. Then, in order to obtain the parameters of the Gaussian mixture model, it is necessary to use the expectation maximization algorithm to estimate. Then, we use the generated Gaussian mixture model to generate new minority samples. Finally, the Bayesian classifier is used to complete the classification and prediction of the data set as shown in Figure 4. Experimental results show that, in order to predict user reports, this algorithm is much better than existing algorithms in terms of performance.

4. Classification Algorithm Improvement

In the processing of unbalanced datasets, the undersampling method may lose the information of many samples, while oversampling may overfit the samples. Moreover, before the traditional method of classification, the sparse area data are often deleted directly as noise data. However, due to the large difference in the number of data samples between the two types in the unbalanced dataset, if there are too many samples of the minority class in the sparse area data, directly deleting the sparse area data will make the dataset more unbalanced. Suppose a dataset is $T$, the majority class sample set is $D$, and the minority class sample set is $S$. In response to the above problems, this study proposes an analysis of the risk analysis of insurance companies based on an unbalanced dataset algorithm. First, the sample detection method based on the coefficient of variation is used to identify the sparse and dense regions of the unbalanced dataset. The basic idea is as follows:

(1) In the data set $T$, calculate the average value of the sum of the distances from the object $T$ to its $K$-nearest neighbors, where the object $D$ is the $K$-nearest neighbors of $T$ and $N\text{dist} (t)$ is the number of $D$'s $K$-nearest neighbors. The density of each point is represented by the reciprocal of the average:

$$T \cdot d(t) = \frac{1}{\sum (\text{dist}(t,d))/(N\text{dist}(t))}$$

(6)

(2) The ratio of the standard deviation of the $k$-nearest neighbor density of the object to their average value is defined as the coefficient of variation of $T$, namely,
This method first divides the minority samples $S$ in the sparse domain into three types: safety samples, boundary samples, and noise samples. After finding the boundary samples of the minority classes, it not only finds the nearest neighbors from the minority samples but also generates new minority samples. And find the nearest neighbors among the majority samples, generate new minority samples, and finally perform appropriate undersampling processing on the sparse domain dataset after adding artificial minority minority classes, which will make the minority samples closer to their true values:

$$
\max F(w) = \sum_{i=1}^{M1} \sum_{j=1}^{M2} w_{ij} AUC_{ij}.
$$

To divide the most sample set $D$, first use the k-nearest other way to divide most sample set into the samples, for example, boundary samples and other samples, and then select a sample point $c$ in the boundary sample set, with $c$ as the center of the circle and $r$ as the radius. In the circle of $n$, represents the number of samples contained in the circle. Delete the sample point when the total number of samples in the circle is greater than $n/2$, otherwise keep the sample point when it is less than or equal to $n/2$:

$$
F(c, \varepsilon) = \frac{1}{\sum_{i=1}^{m} (y_i - f(x_i))^2 + \varepsilon}.
$$

Among them, $\varepsilon > 0$ is a relatively small number to prevent the denominator from being zero.

Compared with other algorithms, the random forest algorithm (USI) based on mixed sampling proposed in this study first identifies the sparse and dense regions of the unbalanced dataset by introducing a “variation coefficient” and then performs improved oversampling on the sparse and dense regions, respectively (Figure 5). The “variation coefficient,” such as range, standard deviation, and variance, is the absolute value reflecting the degree of data dispersion and the undersampling method. The USI algorithm proposed in this study overcomes the fact that other undersampling methods such as RU and IS algorithms may lose the information of many samples, while oversampling methods such as SMOTE, USMOTE algorithm, and mixed-sampling RU-SMOTE algorithm may make the minority class problems such as sample overfitting [29].

5. Risk Classification of Insurance Companies Based on GSSPSO-SVM Classification Algorithm

The resampling method was applied to the classification of student loan data and achieved good classification results. It is well known that one way to optimize unbalanced datasets is to optimize the sampling method. Another method is to optimize the classifier algorithm. In Section 4, the parameters’ cost and weight of the support vector machine are regarded as a particle $K(c, w)$ of the particle swarm. By using the characteristics of particle swarm optimization, we can find the best position of $K(c, w)$ to optimize support vector machines. Experiments on UCI dataset prove that the algorithm is significantly better than the classification effect before optimization. Now, this new optimization algorithm GSSPSO-SVM is applied to the risk classification of insurance companies:

$$
q = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_m}{\sum_{i=1}^{P} \lambda_i} \times 100,
$$

where $\lambda$ is the eigenvalue.

They are the fitness value curves of the standard particle swarm algorithm and improved algorithm with an unbalanced dataset as the object. The blue dotted line in the figure represents the optimal fitness value of each generation in the particle swarm optimization process and the red triangle curve, as shown in Figure 6. It represents the optimal fitness value curve of a certain particle, and the green circle represents the average fitness value curve of each generation of the particle swarm. It is the optimization curve of the standard particle swarm algorithm for the imbalanced student loan dataset. It shows that when algorithm runs to the 50th generation, the algorithm basically converges, while it shows the improved algorithm in the 20th generation of the particle swarm. It converges quickly, indicating that the improved algorithm has a better way, and accuracy of them is relatively close.

This section uses the Wrapper-SMOTE resampling method, and the improved particle swarm optimization improves the vector optimization method which classifies student loan data. Through experiments, it is found that both methods have effectively improved the classification accuracy of student loan data, especially the classification accuracy of the minority classes (default samples):
\[
\sum_{k=1}^{p} \rho^2(Y_k, X_i) \sigma_i = \lambda_k.
\] (11)

The problem of class imbalance in classification is also the problem of imbalance [31]. It has always been a problem in the field of data classification research. In the research of unbalanced data classification, the value of the classification accuracy of the minority class is often far greater than the accuracy of the study of the majority class. The imbalance of rate and class distribution makes traditional classification algorithms encounter many problems: the failure of traditional evaluation methods, the absolute and relative rarity of rare data, the influence of data fragments, the inductive bias, the interference of noisy data, etc.; in addition, the size of the positive and negative class ratio is also an influencing factor. Besides, the maximum, average, and minimum value of error density are obtained as shown in Figure 7. Besides, the simulating results of attributing sizes with different algorithms from other studies and ours are presented in Figure 8. It can be seen from the figure that our algorithm has certain advantages in data size.

6. RVM Classification Algorithm for Unbalanced Datasets

In these application fields, the models trained by the traditional RVM classification algorithm have obvious biases, and such models have poor recognition ability for the categories of a small number of samples, as illustrated in Figure 9. The commonly used solutions to such problems are

(1) Sampling: sampling refers to the initial sample set with unbalanced some ratios. By processing the initial sample set, the initial sample set with unbalanced positive and negative sample ratio is converted into a sample set with balanced two numbers. The model obtained from the dataset processed in this way will improve the result in most cases.

2) Data synthesis: for a small number of positive samples, data synthesis uses existing samples to generate new positive samples in a certain way. The number of positive samples has increased significantly, which also improves the balance ratio of positive and negative samples in the dataset.

3) Weighting: when the data are unbalanced, each sample in the positive sample set is very important. When each such sample is misclassified, the impact will be greater than the same negative sample classification error on the final model. To enhance the importance of data in the negative sample set, the weights to different samples are reassigned. Meanwhile, the cost of misclassification of such samples by assigning larger weights is increased. Therefore, the degree of influence of the positive sample set on the
model is approaching to that of the negative sample set:

$$\sum_{k=1}^{p} \rho^2(Y_k, X_k) = \frac{1}{\rho_{ii}} \sum_{k=1}^{p} \lambda_k \mu_k^2,$$

where $\rho$ represents the correlation coefficient.

Aiming at the problem of unbalanced data sets, this section mainly adopts the method of data sampling, combined with the idea of AdaBoost algorithm, to study the training and classification effects of the mixed-granularity RVM model when dealing with large-scale unbalanced datasets.

7. Conclusion

With the development of the computer and communications industries, the era of big data has come in full swing. The explosive growth of raw data and the rapid increase in data types have made all walks of life urgently demand data processing technology, which has brought huge opportunities to the disciplines of machine learning and data mining. Most of the data in real life is unbalanced data, such as facial age estimation, oil spill detection from satellite images, anomaly detection, identification of fraudulent credit card transactions, software defect prediction, and image annotation. Traditional algorithms generally deal with balanced datasets with equal misclassification costs. How to deal with datasets to turn unbalanced data into balanced data has become the key to data processing.

The insurance risk of insurance companies in China is mainly reflected in underwriting risk, reserve risk, and solvency risk and is mainly reflected by insurance exposure, insurance leverage, reserve equity deficit, and recognized asset-liability ratio indicators. These four indicators are all related to insurance risk. The credit risk of insurance companies in China is mainly reflected in two aspects, financial market credit risk and insurance business credit risk, and is mainly reflected by the debt asset growth rate, debt asset ratio, and surrender rate indicators. These three indicators are positive for credit risk. The market risk of our insurance companies is mainly reflected in interest rate risk and financial market investment risk and is mainly reflected by the interest rate-sensitive total asset ratio and equity asset ratio indicators, both of which are positively related to market risks. The liquidity risk of our insurance company is mainly reflected in the degree of asset solidification and the risk of repurchase business and is mainly reflected by two indicators, fixed asset ratio and financing repurchase ratio, both of which are positively related to liquidity risk.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Financial Synergy Innovation of Science and Technology Center in Hebei Province, Science and Technology Finance Key Laboratory of Hebei Province Open Fund Project, project name: Research on the Innovation of Big Data Reshaping the Insurance Value Chain, project no. STFCIC202206.

References


