A Novel Approach to Handle Imbalanced Data for Classification

R. J. Kuo
Department of Industrial Management
National Taiwan University of Science and Technology
No. 43, Section 4, Kee-Lung Road, Taipei, Taiwan
Tel: 886-2-27376328, E-mail: rjkuo@mail.ntust.edu.tw

L. Lin
Gemtek Technology Co., Ltd.
5F.-4, No.186, Jian First Road., Zhonghe District, New Taipei City, Taiwan
Tel: 886-2-82271792 ext. 6724 E-mail: amati223@gmail.com

F. E. Zulvia
Department of Industrial Mangement
National Taiwan University of Science and Technology
No. 43, Section 4, Kee-Lung Road, Taipei, Taiwan
Tel: 886-2-27333141 ext. 7103, E-mail: feranizulvia@gmail.com

Abstract. This paper attempts to propose a particle swarm *K*-means optimization (PSKO)-based granular computing (GrC) model to preprocess the skewed class distribution in order to enhance the classification accuracy for class imbalance problem. The GrC model acquires knowledge from information granules rather than from numerical data. It also processes multi-dimensional and sparse data by using singular value decomposition and latent semantic indexing (LSI). Four benchmark data sets are employed to demonstrate the effectiveness of the proposed model. Experimental results indicate that the proposed model has better performance to classify imbalanced data by using support vector machine.

Keywords: Granular computing, Class imbalance, Classification.

1. INTRODUCTION

In order to deal with imbalanced datasets, some appro aches such as sampling (Batista et al., 2004) the cost-matrices adjustment, and moving the decision thresholds (C hawla et al., 2002) have been proposed. However, these me thods have some weakness. They tend to produce high accu racy for the majority class but not for the minority class (Za drozny et al., 2003). However, this minority class is usually the important part of the data.

Therefore, this paper aims to develop a novel granular computing method for tackling imbalanced data. Granular computing (GrC) is an innovative computing model of information processing. Generally speaking, granular computing (GrC) is a process of complex information entities called information granules, which arise in the process of data abstraction and derivation of knowledge from information. The idea of information granularity has been explored in a number of fields such as rough sets, fuzzy sets, cluster analysis, database, machine learning, and data mining (Bargiela and Pedrycz, 2003). The proposed algorithm extracts knowledge from information granules (IG) by employing a metaheuristic-based algorithm, proses multi-dimensional and sparse data using singular value decomposition (SVD) and latent semantic indexing (LSI) and builds a prostate cancer prognosis system. In order to verify the proposed algorithm, some experiments are conducted. The results are compared with cluster-based GrC model and an improved cluster-based GrC model by support vector machine (SVM)

The reminder of this paper is organized as follows. Section two presents the proposed algorithm while the experimental results are discussed in Section three. Finally, the concluding remarks are made in Section four.

2. METHODOLOGY

The proposed algorithm comprises of several parts: (1) data preprocessing, (2) granular construction, (3) feature extraction and knowledge discovery, and (4) classification for imbalanced data. The first part preprocesses the data. The size of majority class data is reduced by using granular construction in the second part. The dimension reduction and feature extraction are also implemented in the second part. This paper employs the LSI to tackle critical attributes. Then, several classifiers are applied to acquire knowledge from these preprocessed data. A concise procedure is explained as follows:

Part I: Data preprocessing: delete missing value, and normalization.

Part II: Granular construction

Step 1: Granularity selection criteria. Determine the thresholds of *H-index* and *U-ratio*.

Step 2: Determine the level of granularity for IGs. The number of IGs is determined by *H-index* as well as *U-ratio*.

Step 3: Execute granular construction by clustering techniques.

Step 4: Compute *H-index* $\sum_i (n \,/\, m)/i$ of IGs. Where m represents the number of all objects in one granule, i is the number of all IGs and m is the amount of objects possessing the majority class. The *U-ratio* is $(u \,/\, i)$, where *u* represents the number of undistinguishable granules and *i* represents the quantity of all IGs.

Step 5: Check whether the criteria are satisfied or not.

(a) If the *H-index* is larger or equal to the threshold of H-index and *U-ratio* is smaller or equal to the threshold of *U-ratio*, the answer is ‘‘Yes.” Go to Step (b) Otherwise the answer is ‘‘No.” Repeat Steps 4–5 till criteria are satisfied.

Step 6: Rewrite attributes

Divide value interval of attribute into overlapping and non-overlapping areas into sub-attributes.

Part III: Feature extraction and knowledge acquisition

Step 7: Analysis of sub-attributes by implementing SVD to reduce the dimensionality.

Step 8: Feature extraction. Determine the optimal number of features by evaluating efficiency and accuracy.

Step 9: Sub-attributes reduction to the optimal number.

Step 10: Implement classifiers and calculate the classification accuracy.

Step 11: Validate the classification performance. If the performance is acceptable, terminate the procedure.

Part IV: Apply classification method to classify the data after granularity.

2.1 Construction of Information Granules

In order to construct the IGs, this paper utilizes Chen’s information granulation (Chen et al., 2008) and improves it using particle swarm *K*-means optimization (PSKO) algorithm (Kuo et al., 2009). K-means algorithm is very sensitive to the initial centroids. Therefore, this study applies PSKO algorithms in order to obtain a better clustering result.

2.2 Selection of Granularity

PSKO algorithm can group IGs of similar ‘‘size” (that is granularity) in a single layer according to the Euclidean distance between data and cluster center. The same degree of similarity of patterns will be placed in the same cluster. However, the measures of *H*-index and *U*-ratio are calculated according to the result of PSKO, and the levels of granularity will be adjusted by PSKO until *H*-index and *U*-ratio are satisfied.

2.3 Representation of Information Granules

This paper employs the sub-attributes concept called hyperboxes to represent IGs (Bargiela and Pedrycz, 2003). A hyperbox $[b]$ defined in R^n is fully described by lower bound $(b-)$ and upper bound $(b+)$, defined as $[b] = [b-, b+]$. Part 1 in Figure 1 shows an illustrative example to express the implementation procedure of sub-attributes. The part 2 of Figure 1 shows that there are overlaps between two granules, A and B. This makes it difficult to be handled by knowledge acquisition tools. However, data mining cannot discover knowledge from these constructed IGs because most of knowledge acquisition algorithms are designed to deal with numeric attributes. In this paper, this problem is tackled by “sub-attributes” which divide the value interval of attribute into overlapping and non-overlapping areas. Next, Boolean variable, 0 or 1, is used to represent whether the IG contains these intervals or not.

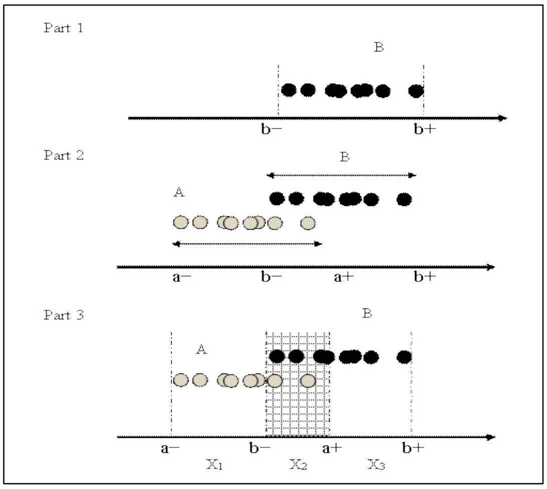

Figure 1: Representation of information granules.

2.4 Latent Semantic Indexing

Sub-attribute in this paper involves a large feature set which often contains redundant and irrelevant information. It may causes worse performance of classifiers. Therefore, this paper applies latent semantic indexing (LSI) to reduce the dimensions.

3. EXPERIMENTAL RESULTS AND ANALYSIS

This paper verifies the proposed PSKO-based GrC algorithm using imbalanced datasets. For all datasets, 90% of the data ae used as the training while the 10% are testing data. The classifier applied in this paper is SVM.

The experiments using imbalanced datasets are conducted in order to evaluate the capability of the proposed algorithm in extracting data from variate skewed datasets. Table 1 lists imbalanced datasets used in this paper. Each dataset was run 30 times for each 10-fold cross validation. The appropriate parameter setting obtained from Taguchi method is as follows: the number of particles is equal to 40, learning factors, c_1, is equal to 1.47, c_2 is equal to 0.5, and inertia weight is equal to 0.5. Table 2 shows that the proposed PSOK-based GrC model with classified dataset obtain better results than other algorithms shown by its smaller mean square errors. This result proves that classifying the dataset using granularity improves the performance of the clustering algorithm. In addition, combining the *K*-means algorithm with PSO also can improves the clustering performance.

4. CONCLUSIONS

This paper has shown that the PSKO-based GrC models have capability to extract knowledge for imbalanced datasets. They have benefits over building classifiers from numerical data. PSKO-based GrC models can decrease the effect of imbalance that classifiers produce high accuracy over the majority class but poor predictive accuracy over the minority class. In addition, integrating PSKO to improve the clustering result and employing LSI to decrease the dimension size really can improve the classification accuracy as well as speed up the classification. In overall, due to the integration of PSKO and granular computing, it's really helpful to extract knowledge from a huge amount of data. Meanwhile, it can increase the classification accuracy.

REFERENCES

Bargiela, A., and Pedrycz, W. (2003). Recursive information granulation: aggregation and interpretation issues. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 33*(1), 96-112.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20-29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321-357.

Chen, M.-C., Chen, L.-S., Hsu, C.-C., and Zeng, W.-R. (2008). An information granulation based data mining approach for classifying imbalanced data. *Information Sciences, 178*(16), 3214-3227.

Kuo, R. J., Wang, M. J., and Huang, T. W. (2009). An application of particle swarm optimization algorithm to clustering analysis. *Soft Computing, 15*(3), 533-542.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Third IEEE International Conference on Data Mining*, 19-22.

Table 1: Imbalanced datasets.

Dataset	Data size	No. of attributes	Attributes' value	Class distribution
BSWD (Balance scale weight and distance)	625	4	Discrete: 4	Left: 46.08%; Balanced: 7.84 %; Right: 46.08%
PIMA (Pima-Indians-Diabetes)	768	8	Continuous: 8	Class 1: 95%; Class 2: 5%
Car Evaluation	1708	6	Discrete: 5, Binary:1	Healthy: 90%; Diabetic: 10%
Glass (Glass identification database)	214	9	Continuous: 9	Class 1: 33% ; Class 2: 36% Class 3: 8%; Class 4: 6% Class 5: 4%; Class 6: 13%

Table 2: Means square error of imbalanced dataset using SVM.

Method Results		PSKO-based GrC model+classified	PSKO-based GrC model	*K*-means-based GrC model	Numerical computing model
BSWD	Training accuracy	93.15%	53.32%	91.47%	87.82%
	Testing accuracy	74.29%	47.31%	50.61%	88.33%
	RMSE	0.3088	0.4813	0.1877	0.3454

PIMA	Training accuracy	89.06%	52.49%	99.28%	90.07%
	Testing accuracy	97.62%	64.81%	46.73%	89.81%
	RMSE	0.2872	0.7315	0.0170	0.3151
Car evaluation	Training accuracy	84.43%	55.39%	95.99%	83.84%
	Testing accuracy	91.01%	34.42%	52.98%	84.27%
	RMSE	0.3369	0.3944	0.0990	0.3362
Glass	Training accuracy	57.19%	32.89%	32.89%	61.97%
	Testing accuracy	44.44%	28.33%	29.68%	58.57%
	RMSE	0.3564	0.2851	0.2851	0.3332