Information Sciences 178 (2008) 3214–3227

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An information granulation based data mining approach for classifying imbalanced data

Mu-Chen Chen [a,*], Long-Sheng Chen [b], Chun-Chin Hsu [c], Wei-Rong Zeng [d]

[a] Institute of Traffic and Transportation, National Chiao Tung University, 4F, 118, Section 1, Chung-Hsiao W. Road, Taipei 10012, Taiwan
[b] Department of Information Management, Chaoyang University of Technology, 168, Jifong E. Road, Wufong Township, Taichung County 41349, Taiwan
[c] Department of Industrial Engineering and Management, Chaoyang University of Technology, 168, Jifong E. Road, Wufong Township, Taichung County 41349, Taiwan
[d] Information Management Department, Entie Commercial Bank, Taipei, Taiwan

ARTICLE INFO

Keywords:
Information granulation
Granular computing
Data mining
Latent semantic indexing
Imbalanced data
Feed-forward neural network

ABSTRACT

Recently, the class imbalance problem has attracted much attention from researchers in the field of data mining. When learning from imbalanced data in which most examples are labeled as one class and only few belong to another class, traditional data mining approaches do not have a good ability to predict the crucial minority instances. Unfortunately, many real world data sets like health examination, inspection, credit fraud detection, spam identification and text mining all are faced with this situation. In this study, we present a novel model called the "Information Granulation Based Data Mining Approach" to tackle this problem. The proposed methodology, which imitates the human ability to process information, acquires knowledge from Information Granules rather then from numerical data. This method also introduces a Latent Semantic Indexing based feature extraction tool by using Singular Value Decomposition, to dramatically reduce the data dimensions. In addition, several data sets from the UCI Machine Learning Repository are employed to demonstrate the effectiveness of our method. Experimental results show that our method can significantly increase the ability of classifying imbalanced data.

1. Introduction

In recent years, we have seen an increase in research activities in the class imbalance problem. This increased interest resulted in two workshops being held, one by AAAI (American Association for Artificial Intelligence) in 2000 and another one by International Conference on Machine Learning (ICML) in 2003. *SIGKDD Explorations* also published one special issue in 2004. The problem is caused by imbalanced data, in which one class is represented by a large number of examples while the other is represented by only a few [4]. Imbalanced data will result in a significant bottleneck in the performance attainable by standard learning methods [20,27] which assume a balanced class distribution as shown in Fig. 1. It is regarded as one of the most relevant topics of future machine learning researches.

When learning from imbalanced data, traditional data mining methods tend to produce high predictive accuracy for the majority class but poor predictive accuracy for the minority class [39,45]. That is because traditional classifiers seek accurate performance over a full range of instances. They are not suitable to deal with imbalanced learning tasks [6,12,19,23] since they tend to classify all data into the majority class, which is usually the less important class. Fig. 2 illustrates this situation. If data mining approaches cannot classify minority examples such as medical diagnoses of an illness, or the abnormal products

* Corresponding author.
E-mail addresses: ittchen@mail.nctu.edu.tw (M.-C. Chen), lschen@cyut.edu.tw (L.-S. Chen), cchsu@cyut.edu.tw (C.-C. Hsu).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.03.018

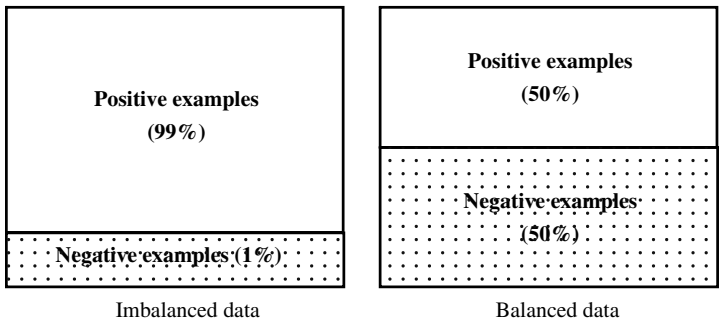

Fig. 1. Imbalanced and balanced data sets.

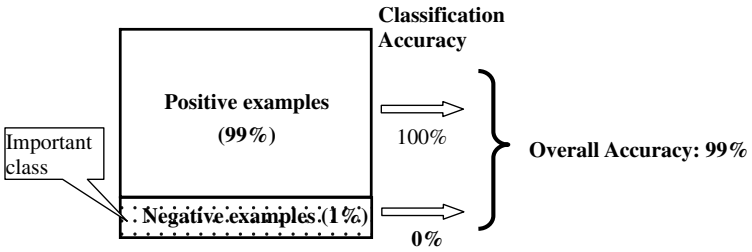

Fig. 2. The illustration of class imbalance problems.

of inspection data, the extracted knowledge becomes meaningless and useless. Recently, this problem has been recognized in a large number of real world domains, like medical diagnosis [37], inspection of finished products [38], identifying the cause of power distribution faults [39], surveillance of nosocomial infections [14], prediction of the localization sites of protein [3], speech recognition [25], credit assessment [20], and functional genomic applications [41].

To address the class imbalance problem, two major groups of techniques are proposed in the available literature. The first group involves five approaches: (1) *under-sampling*, a method in which the minority population is kept intact, while the majority population is under-sampled; (2) *over-sampling*, methods in which the minority examples are over-sampled so that the desired class distribution is obtained in the training set [6,11,19]; (3) *cluster based sampling*, methods in which the representative examples are randomly sampled from clusters [2]; (4) *moving the decision threshold*, methods in which the researcher tries to adapt the decision thresholds to impose bias on the minority class [11,21,24] and (5) *adjust costs matrices*, a method in which the prediction accuracy is improved by adjusting the cost (weight) for each class [15]. Besides, Liu et al. [27] also presented a weighted rough set method for this problem. However, all of these techniques have some disadvantages [2]. For instance, the computational load is increased and overtraining may occur due to replicated samples in the case of over-sampling. Under-sampling does not take into account all available training data which corresponds to loss of available information. Huang et al. [21] indicated that these supervised methods lack a rigorous and systematic treatment of the imbalanced data.

The second group is related to Granular Computing (GrC) models. These GrC models [37,38] which copy the human instinct of information processing can increase classification performance by improving the class imbalance situation. However, these models use the concept of sub-attributes to describe Information Granules (IGs) which are collections of objects arranged together based on their similarity, functional adjacency and indistinguishability [5,9,40,42]. When handling continuous data, the drawback of sub-attributes is that computational loads will increase dramatically due to the generation of a huge number of sub-attributes. Therefore, by introducing the Latent Semantic Indexing (LSI) based feature-extraction technique, this study proposes a novel GrC model called the "Information Granulation Based Data Mining Approach" to solve the class imbalance problem. In addition, for highly skewed data, we present a new IG construction strategy which only builds IGs from majority examples and keeps minority instances intact. Finally, the experimental results show the superiority of our method for classifying imbalanced data.

2. Granular computing

Humans have a remarkable capability to perform a wide variety of physical and mental tasks without any measurements/computations, such as playing computer game, driving, and cooking. Human beings use perceptions of direction, speed, time and other attributes of physical/mental objects, instead of numerical data. Basically speaking, reflecting the limited ability of human brains, perceptions are inaccurate. In more concrete terms, perceptions are granular. It means that the boundaries of

perceived classes are not sharp; and the values of attributes are granulated [43]. Consequently, when making decisions, we tend to shy away from numbers and use aggregates to ponder the question instead [5,43]. This is especially true when a problem involves vague, uncertain, or incomplete information. It may be sometimes difficult to differentiate distinct elements, and so one is forced to consider IGs [9,40,42].

Zadeh [44] summarized four reasons/situations why we need to process perception based information: (1) the bounded ability of sensory organs, and ultimately the brain, to resolve detail and store information; (2) when numerical information may not be available; (3) when an attribute is not quantifiable; (4) when there is a tolerance for imprecision which can be exploited through granulation to achieve tractability, robustness and economy of communication.

The process of constructing IGs was first pointed out in the pioneering work of Zadeh [42] who coined the term "information granulation" and emphasized the fact that a plethora of details does not necessarily amount to knowledge. In addition, GrC is oriented towards the representation and processing of IGs. It is a new direction of Artificial Intelligence [43]. Recently, GrC is quickly becoming an emerging conceptual and computing paradigm of information processing, particularly in soft computing [5,28]. Albert Einstein (1879–1955) said "As far as the laws of mathematics refer to reality, they are not certain; as far as they are certain, they do not refer to reality". His words can be used to explain why researchers paid lots of attentions on uncertainty/vagueness in human decision making, such as fuzzy sets, rough sets, granular computing, etc. [43]. These researches are not intended to replace traditional measurement information based methods which operate numerical data. Their purposes are to let the developed computational theories refer to reality. GrC which operate perception based information is developed under this background.

Castellano and Fanelli [9] indicated that the main issues of GrC are how to construct the IGs and how to describe IGs. One particular question that arises is how to determine the level of granularity. In the issue of constructing IGs, there are many approaches, such as the Self Organizing Map (SOM) network, Fuzzy C-means (FCM), rough sets, shadowed sets [5,9] used to do this. In the issue of representing IGs and determining the level of granularity, Bargiela and Pedrycz [5] proposed the "hyperbox" and "inclusion and compatibility" to measure IGs. Su et al. [37] presented "sub-attributes" to describe IGs and "*H*-index and *U*-ratio" to determine the level of granularity. However, most of GrC related researches focused on such fundamental issues. We need an advanced/integrated mechanism to imitate human ability of processing information, such as extracting knowledge from IGs and making decision based on them. But, if we want to acquire knowledge from IGs, we must try to solve these three questions mentioned above.

This study presents a new GrC model which can discover knowledge from IGs. Our proposed methodology follows the procedure shown in Fig. 3 which involves three steps: IG construction, IG representation, and knowledge acquisition. The details of these three steps will be provided in the following subsections.

2.1. Information granules construction

This subsection introduces how to construct IGs, including how to determine the level of granularity. In available works, SOM, Fuzzy Adaptive Resonance Theory (ART) [37,38], rough sets [8,31,46,47], etc. have been proposed to construct IGs. However, for the purpose of being simple and clear, this study uses K-means [29], which is one of the simplest and most widely used unsupervised learning algorithms for clustering, to build IGs. However, before implementing K-means, one issue needs to be addressed. "What level of granularity is appropriate for building IGs?" This question is equal to "how to determine the number of IGs (clusters)?" K-means needs to be given a number of clusters before implementing this clustering algorithm.

Data exist at different levels of granularity. We usually group IGs of similar "size" (that is granularity) in a single layer. Take sale data for example, there are different levels of granularity such as daily, weekly, and monthly sales. Daily data (the lower level of granularity) can provide the most detailed information. But, some useful knowledge may be buried into unnecessary details. On the other hand, using monthly data (the higher level of granularity) might reduce some information. But, it can provide a better insight into the essence of data, rather than get buried in all the unnecessary details. By changing the granularity, we can hide or reveal more or less details [5]. Using this concept, Chen and Yao [13] proposed a multi-view approach that provides a unified framework for integrating multiple views of intelligent data analysis.

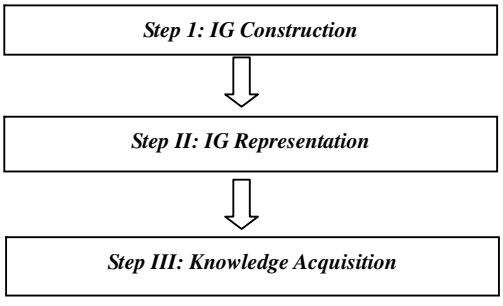

Fig. 3. Three steps of acquiring knowledge from IGs [37,38].

To sum up, if more detailed processing is required, smaller IGs are selected. Then, we are concerned with numeric processing in this low level of granularity. This is a domain completely taken over by numeric models, such as differential equations, regression models, neural networks, etc. At the intermediate level, we see larger IGs (viz. those embracing more individual elements). The top level is solely devoted to symbol based processing, and as such invokes well-known concepts of Petri nets, qualitative simulation, etc. [5]. However, what level of granularity is suitable? This work employs objective indexes, *H*-index and *U*-ratio, developed by Su et al. [37] to solve this question.

A brief introduction of *H*-index and *U*-ratio has been provided as below. *H*-index is used to measure the consistency of the class of the objects in one IG. The *H*-index is defined as

$$H\text{-index} = \sum_{m} \frac{i}{n} \Big/ m \tag{1}$$

where n represents the number of all objects in one granule, m is the number of all IGs and i is the amount of objects possessing the majority class. The second index is the *U*-ratio. Before defining this index, we should clarify what an "undistinguishable granule" is. One IG may involve more than one example. Usually, the class label of majority examples is assigned to be the class label of the IG. If we have an IG which cannot be distinguished from the majority class of the IG, then we call that granule an "undistinguishable granule". We can define the *U*-ratio as

$$U\text{-ratio} = \frac{u}{m} \tag{2}$$

where u represents the number of undistinguishable granules and m represents the quantity of all IGs. This index calculates the proportion of undistinguishable granules to all IGs. If there are 10 IGs and 3 of them are undistinguishable granules, which means u is equal to 3 and m is equal to 10, then the *U*-ratio is equal to 0.3. In addition, we need a "granularity selection criteria" to determine the suitable level of granularity. This criteria can be described as "the larger the *H*-index the better it is" and "the smaller the *U*-ratio the better it is". After solving the question of determining the suitable level of granularity, K-means is employed to construct IGs. More detailed information about *H*-index and *U*-ratio can be found in Su et al. [37].

2.2. Information granules description

After building IGs, the next question is how to describe these constructed IGs. This study uses the concept of sub-attributes [37] to describe IGs so that we can discover knowledge from these defined IGs without changing the mechanism of data mining algorithms. Fig. 4 provides an illustrative example to show the 2-phase implementation procedure of sub-attributes. There are two IGs, A and B, which merely have one attribute, X_i. In phase 1, we use the lower and upper limit of the objects to represent IGs. Therefore, IGs A and B can be described as $[a_{\min}, a_{\max}]$ and $[b_{\min}, b_{\max}]$, respectively. However, data mining approaches which are designed for numerical numbers cannot discover knowledge from these constructed IGs. This is especially true when overlaps between IGs occur. We tackle this problem in phase 2.

Phase 2 divides the value interval of attribute into overlapping and non-overlapping areas. Then, these sub-intervals are named as X_{i1}, X_{i2}, and X_{i3}, which are so-called "sub-attributes". Next, we use the Boolean variable, 0 or 1, to represent if the IG contains these intervals or not. This procedure is illustrated by Tables 1 and 2. From these tables, the IG A (B) has been rewritten from $[a_{\min}, a_{\max}] + Y_1$ ($[b_{\min}, b_{\max}] + Y_2$) to 1 1 0 + Y_1 (0 1 1 + Y_2), where $Y_1(Y_2)$ are class labels. Finally, we can acquire knowledge from IGs by using this data format.

From the results of Table 2, it is easy to find that the number of input variables increases to three times (from 1 to 3). If we have to cope with continuous data, this situation will become worse. Therefore, we propose an LSI based feature extraction method using Singular Value Decomposition (SVD) to solve this problem. The illustration of SVD can be found in Section 3.1. The reasons we propose the LSI based feature extraction method are: (1) LSI can greatly reduce the number of attributes and

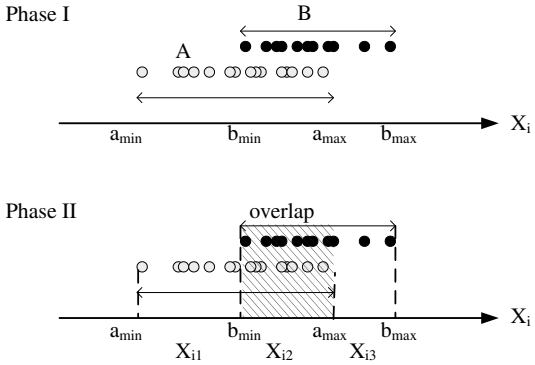

Fig. 4. IG description: the concept of "sub-attributes".

Table 1
IG A and IG B

IGs	Attribute	
	X_i	Class
IG A	$[a_{min}, a_{max}]$	Y_1
IG B	$[b_{min}, b_{max}]$	Y_2

Note: X_i is the ith attribute of objects. Y_1 and Y_2 are class labels. a_{min} and a_{max} are lower and upper limits of objects of IG A.

Table 2
The implementation of sub-attributes

IGs	Sub-attribute			Class
	X_{i1}	X_{i2}	X_{i3}	
IG A	1	1	0	Y_1
IG B	0	1	1	Y_2

Note: X_{i1}, X_{i2}, and X_{i3} denote sub-attributes of IGs. Boolean variables are employed to represent IG contains the value interval of attribute or not.

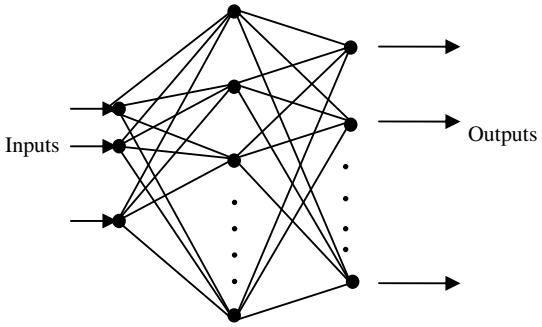

Fig. 5. The feed-forward neural network structure.

enhance the performance of the classifiers; (2) the IGs described by sub-attributes usually satisfy the behavior of sparse data. LSI is a popular and effective technique in the text mining domain which often copes with sparse data.

2.3. Knowledge acquisition

The feed-forward neural network (NN) with back-propagation learning algorithm is employed to extract knowledge. Neural nets have been used widely in pattern recognition, function approximation, optimization, and clustering. Generally speaking, neural nets can be classified into two categories, feed-forward and feedback networks. In this study the feed-forward network, shown in Fig. 5, was employed it because of its superior classification ability.

The back-propagation learning algorithm [36] is the best known training algorithm for neural networks and still one of the most useful. This iterative gradient algorithm is designed to minimize the mean square error between the actual output of a multilayer feed-forward perceptron and the desired output. According to the rule of thumb and reports of available published papers, the number of hidden layers should be one or two. The back-propagation algorithm includes a forward pass and a backward pass. The purpose of the forward pass is to obtain the activation value and the backward pass is to adjust weights and biases according to the difference between the desired and actual network outputs. These two passes will be gone through iteratively until the network converges. The detailed information about network training by back-propagation can be found in related references [18,32].

3. Proposed methodology

3.1. Latent semantic indexing

In machine learning, the number of necessary sample points grows exponentially with the dimension of the feature space. This problem has been known as the "curse of dimensionality". A large feature set often contains redundant and irrelevant information, and can actually degrade the performance of the classifier [30]. Therefore, one needs techniques to reduce the dimension of examples and should use either features extraction, features selection or a combination of the both [22].

Feature selection is to select a subset of most representative features from the original feature space. Feature extraction is to transform the original feature space to a smaller feature space to reduce the dimensionality. Liu et al. [26] indicated that feature extraction can greatly reduce the dimensions of the feature space compared with feature selection.

The most representative feature extraction algorithm is LSI [16], which is an automatic method that transforms the original textual data to a smaller semantic space by taking advantage of some of the implicit higher-order associations of words with text objects [7,16]. The transformation is computed by applying truncated SVD to the term-by-document matrix. After SVD, terms which are used in similar contexts will be merged.

Fig. 6 briefly introduces the concept of SVD. Let A be an $m \times n$ matrix of rank r with rows representing documents and columns denoting terms (variables). Let the singular values of A (the eigenvalues of $A \times A^{\mathrm{T}}$) be $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r$. The *singular value decomposition* of A expresses A as the product of three matrices $A = USV^{\mathrm{T}}$, where $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ is an $r \times r$ matrix, $U = (u_1, \ldots, u_r)$ is an $m \times r$ matrix whose columns are orthonormal, and $V^{\mathrm{T}} = (v_1, \ldots, v_r)^{\mathrm{T}}$ is an $r \times n$ matrix. LSI works by omitting all but the k largest singular values in the above decomposition, for some suitable k (k is the dimension of the low-dimensional space). It should be small enough to enable fast retrieval and large enough to adequately capture the structure of the corpus. Let $S_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$, $U_k = (u_1, \ldots, u_k)$ and $V_k = (v_1, \ldots, v_k)$. Then $A_k = U_k S_k V_k^{\mathrm{T}}$ is a matrix of rank k, which is the approximation of A. The rows of $V_k S_k$ above are then used to represent the documents. In other words, the row vectors of A are projected to the k-dimensional space spanned by the row vectors of U_k; we sometimes call this space the LSI space of A.

To sum up, SVD is an optimal linear transformation for dimensionality reduction. It allows the arrangement of the space to reflect the major associative patterns in the data and ignores the smaller, less important influences. SVD transformation also has the advantage of yielding zero-mean and uncorrelated features [10]. Moreover, it has been reported that SVD can be applied to education, solving linear least-squares problems and data compression [1]. Therefore, SVD is employed as the feature extraction tool in this study.

3.2. An information granulation based data mining approach

In this subsection, we propose a new methodology which combines "Information Granulation" and LSI to solve class imbalance problems. Briefly, our proposed method involves two major parts: First, we introduce information granulation to reduce the data size. Second, the LSI technique is employed to reduce dimensions of features and then we acquire knowledge from these constructed IGs. The main advantage of our proposed method is to reduce both the size of attributes and the size of the data. Fig. 7 shows the procedure of the proposed methodology. A concise algorithm is provided as follows:

Part 1: Information granulation

Step 1: Determine the thresholds of H-index and U-ratio.
Step 2: Determine the number of IGs.
Step 3: Execute K-means.

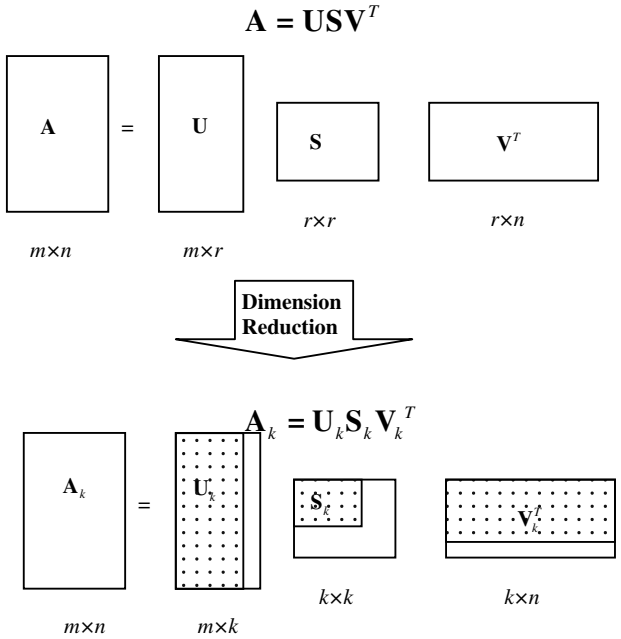

Fig. 6. The singular value decomposition [16].

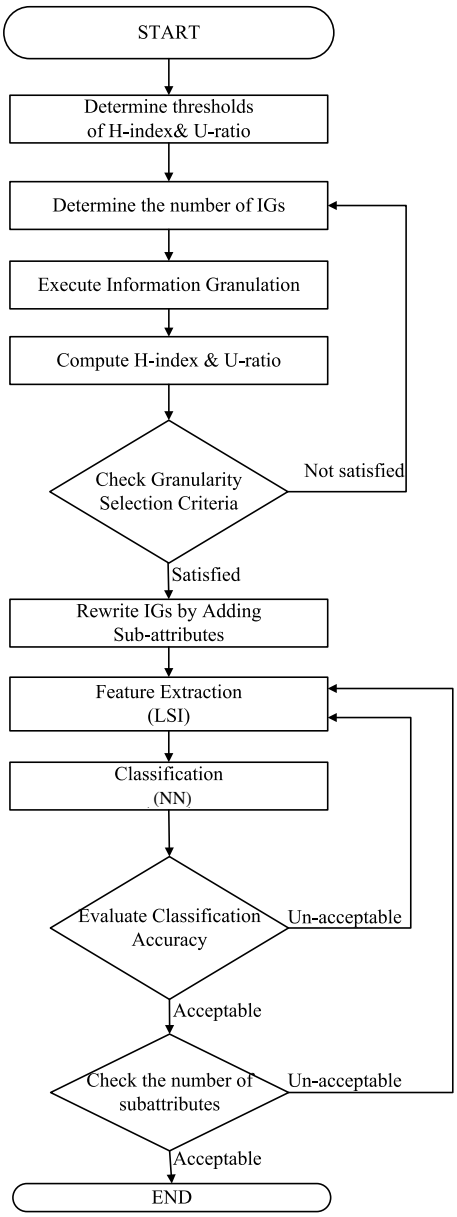

Fig. 7. The procedure of our proposed methodology.

Step 4: Compute *H*-index and *U*-ratio of IGs.
Step 5: Check two indexes, *H*-index and *U*-ratio, do they satisfy the pre-determined threshold or not?
(a) If the answer is "True", go to Step 6.
(b) If the answer is "False", repeat Steps 2–4 till *H*-index and *U*-ratio satisfy the minimum threshold requirement.
Step 6: Check if the data type is continuous or discrete.
(a) If the data is "continuous", go to Step 7.
(b) If the data is "discrete", go to Step 8.
Step 7: Data discretization.
Step 8: Divide original attributes into sub-attributes.

Part II: Feature extraction (LSI) and knowledge acquisition (NN)

Step 9: Implement SVD.
Step 10: Determine the optimal number of features *k*.

Step 11: Reduce the number of dimensions of the sub-attributes to k.
Step 12: Implement NN and calculate the classification accuracy.
Step 13: Validate the classification performance.
(a) If the performance is acceptable, go to Step 14.
(b) If the performance is unacceptable, raise k and repeat Steps 10–12.
Step 14: Verify the dimensions of the sub-attributes.
(a) If the number of dimensions is acceptable, terminate the procedure.
(b) If the number of dimensions is unacceptable, reduce the size of the sub-attributes, k, and repeat Steps 10–13.

In Part 1, "information granulation" phase, we construct IGs by using K-means. Then we set the "granularity selection criteria" (i.e., the threshold of H-index and U-ratio) to determine the suitable level of granularity. Next, we describe the constructed IGs by sub-attributes. Before implementing sub-attributes, we need to check the data type. If the data is continuous, it will be discretized. Part 2 is the "dimension reduction" phase. We should determine the optimal number of features k and build a classifier by feed-forward neural network (NN).

4. Implementation

In this section, several data sets from data bank will be employed to demonstrate the superiority of our proposed method. These data sets involve balanced and imbalanced data. We will validate the effectiveness of our method in different skewed class distributions.

4.1. Datasets

The data sets come from the UCI Machine Learning Repository which is available at the website: http://www.ics.uci.edu/~mlearn/MLRepository.html. Detailed information including data size, the number of attributes, data characteristics and class distribution can be found in Table 3. Originally, we only use the first four data sets which contain two balanced (Contraceptive and Wine) and two imbalanced data (Credit Screening and Pima). However, the experimental results show that our proposed GrC method has good potential in handling imbalanced data. Therefore, for the purpose of validation, another imbalanced data, BSWD, is added to demonstrate the benefits of our method. In addition, the missing value examples have been removed because LSI cannot handle missing data. A 10-fold cross validation (CV) experiment is employed in this study. In other words, the data set is portioned into 10 equal sized sets and each set is then in turn used as the test set. Section 4.3 shows the evaluation results of our method and Section 4.4 provides the experiments on highly imbalanced data.

4.2. Evaluation measures

Traditionally, the performance of a classifier is evaluated by considering the Overall Accuracy against test cases. However, when dealing with imbalanced data, this index may be insufficient as can be seen from Fig. 2. We can construct a classifier with an accuracy of 98% in a domain where the majority class proportion corresponds to 98% of the instances, by simply forecasting every new example as belonging to the majority class. However, for imbalanced data, this high overall accuracy (98%) may mean nothing if the classifier is not able to identify any single minority example. Another fact is that the metric considers different classification errors to be equally important. However, we know that a highly imbalanced class problem

Table 3
Summary of data sets

Data set	Title of data	Data size	Attributes	Attributes' value	Class distribution
Credit screening	Credit approval	653	15 condition attributes and 1 class attribute	Continuous: 6 Discrete: 5 Binary: 4	Unacceptable: 45% Acceptable: 55%
Contraceptive	Contraceptive method choice	1473	9 condition attributes and 1 class attribute	Continuous: 2 Discrete: 4 Binary: 3	No-use: 43% Long-term: 23 % Short-term: 34%
Wine	Wine recognition data	178	13 condition attributes and 1 class attribute	Continuous: 13	Class 1: 33.15% Class 2: 39.89% Class 3: 26.96%
Pima	Pima-Indians-Diabetes	768	8 condition attributes and 1 class attribute	Continuous: 8	Healthy: 65% Diabetic: 35%
BSWD	Balance scale weight and distance database	625	4 condition attributes and 1 class attribute	Discrete: 4	Left: 46.08% Balance: 7.84% Right: 46.08%

has non-equal error costs that favor the minority class, which is often the class of primary interest. Therefore, based on the available papers [6,17,19,33,34], Overall Accuracy, Positive Accuracy, Negative Accuracy, and G-Mean (of Positive Accuracy and Negative Accuracy) are employed to evaluate the performance of classifiers in this work.

The easiest way to evaluate the performance of classifiers is based on the confusion matrix as shown in Table 4. From this table, let the True Positives (TP) denote the number of positive examples correctly recognized as being positive, and False Negatives (FN) represent the number of positives incorrectly recognized as being negative. Similarly, TN and FN represent the number of negative examples correctly identified as being negative, and incorrectly identified as being positive, respectively. Overall Accuracy can be easily calculated as

$$\frac{TP + TN}{FN + TP + TN + FP} \tag{3}$$

Moreover, the effectiveness of a classifier is also frequently measured by using the Specificity and Sensitivity [35] when tested with a test data set. Theses values can then be used to define the metrics described below:

– *Positive Accuracy:* The Positive Accuracy measures the proportion of positive instances being correctly recognized as being positive:

$$\text{PositiveAccuracy} = \frac{TP}{TP + FN} \tag{4}$$

– *Negative Accuracy:* The Negative Accuracy measures the proportion of negative instances being correctly recognized as being negative:

$$\text{NegativeAccuracy} = \frac{TN}{TN + FP} \tag{5}$$

The last index is the geometric mean (G-mean) of Positive Accuracy and Negative Accuracy. G-mean can be defined as

$$\text{G-mean} = \sqrt{\text{PositiveAccuracy} \times \text{NegativeAccuracy}} \tag{6}$$

This measure is to maximize the accuracy for each of the two classes while keeping these accuracies balanced. For example, a high Positive Accuracy with a low Negative Accuracy will result in a poor G-mean.

4.3. Experimental results

Without considering class distribution, this subsection provides the experimental results of implementation. NN is employed as the basic learner. As shown in Table 5, the optimal parameters settings of NN including learning rate, momentum, training iterations and network architectures are obtained by trail and error. The comparisons between our method and the other two methods (using NN to discover knowledge from IGs and using NN to extract knowledge from numerical data) are made and the results are summarized in Table 6. To clarify the results in this table, we use "GrC" to denote the proposed method and "InG" represents the method which discovers knowledge from IGs without implementing LSI. All methods use NN to discover knowledge. GrC and InG deal with IGs and NN copes with numerical data.

First, we want to know the benefits of introducing LSI. From Table 6, the average overall accuracy (72.68%) and the number of sub-attributes (24.5) of GrC are better than those (70.71% and 49.8) of InG. Our method can indeed greatly reduce the

Table 4
Confusion matrix

	Predicted positive	Predicted negative
Actual positive	TP (the number of True Positives)	FN (the number of False Negatives)
Actual negative	FP (the number of False Positives)	TN (the number of True Negatives)

Table 5
The settings of BP neural network

Data set	Network structure	Learning rate	Momentum	Iterations
Credit screening	15-29-1	0.2	0.8	20,000
Contraceptive	9-37-1	0.2	0.8	10,000
Wine	13-18-1	0.2	0.8	15,000
Pima	8-27-1	0.2	0.8	20,000
BSWD	4-23-1	0.2	0.8	20,000

Table 6
Implementation results without considering class distribution

Data set	Method					
	GrC model (extract knowledge from IGs)				Numerical computing model (extract knowledge from numerical data)	
	Proposed method (GrC)		Information granulation (InG)		NN (BP)	
	No. of sub-attributes	Overall accuracy – Mean (%)	No. of sub-attributes	Overall accuracy – Mean (%)	No. of attributes	Overall accuracy – Mean (%)
Credit screening	19	92.00	63	88.00	15	84.58
Contraceptive	20	30.16	33	30.16	9	29.60
Pima	27	84.21	39	78.95	8	76.92
Wine	32	84.33	64	85.71	13	86.49
Average	24.5	72.68	49.8	70.71	11	69.40
StDev	6.14	28.57	16.07	27.30	3.65	26.85

Notes: 1. GrC denotes the proposed method and InG represents the method which discovers knowledge from IGs without implementing LSI.
2. NN is the basic learner. All methods use NN to discover knowledge. GrC and InG deal with IGs and NN copes with numerical data.

number of sub-attributes and enhance the performance. Second, the pro and cons between granular computing and numerical computing model must be validated. Therefore, we compare the results of NN and GrC. GrC (72.68%) outperforms NN (69.4%) in overall accuracy. Based on overall accuracy, the proposed method has the highest performance. Although the differences are not statistically significant, the result shows that our method can slightly increase the classification performance by 1.97% and 3.28% over InG and NN, respectively. In addition, our method can indeed solve the problem of the amount of sub-attributes, which has been mentioned in Su et al. On the average, compared with the number of sub-attributes in InG which does not introduce the LSI technique, our method decreases the dimension size by 51%. To sum up, our method cannot only drop the number of input variables, but it also raises the overall accuracy compared with InG.

Taking the execution time into consideration, Table 7 shows a comparison between NN (numerical computing) and our method GrC (granular computing). On average, our method will be shorter by more than 308 s compared with NN, if we use GrC to discover knowledge. In other words, the computational time will be reduced by a dramatic 82.42%.

If we consider class distribution, it is evident that the proposed method might be a possible solution to the class imbalance problems. From Table 8, the overall accuracies of the three methods are almost the same for balanced data sets. It shows that our method has the shortest execution time; however, it cannot improve the classification performance when dealing with balanced data sets. However, the situation is totally different for classifying imbalanced targets. The overall accuracies of our proposed method, InG and NN are 88.11%, 83.48%, and 80.75%, respectively. These results show that our method has

Table 7
The computational time

Data set	Methods	
	Granular computing model, GrC (s)	Numerical computing model, NN (s)
Credit screening	19	396
Contraceptive	211	661
Pima	28	394
WDBC	34	365
Wine	37	55
Average	65.8	374.2
StDev	81.46	215.11

Table 8
Implementation results with considering class distribution

Data type	Title	Granular computing model		Numerical computing model
		GrC (%)	InG (%)	NN (%)
Balanced data	Contraceptive	30.16	30.16	29.6
	Wine	84.33	85.71	86.49
	Average	57.25	57.94	58.04
Imbalanced data	Credit screening	92	88	84.58
	Pima	84.21	78.95	76.92
	Average	88.11	83.48	80.75

Table 9
Summary of imbalanced data sets

Data set	Name	Data size	Number of attributes	Attributes' value	Class distribution
Pima I (10%)	Pima-Indians-Diabetes	555 (499 + 56)[a]	8	All continuous	Healthy: 90% Diabetic: 10%
Pima II (5%)		527 (499 + 28)	8	All continuous	Healthy: 95% Diabetic: 5%
BSWD	Balance scale weight and distance database	625	4	All discrete	Left and Right: 92.16% Balance: 7.84%

[a] *Note:* (499 + 56) represents the data contains 499 majority examples (healthy) and 56 minority instances (diabetic).

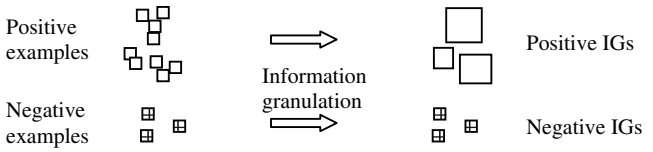

Fig. 8. The proposed IG construction strategy for imbalanced targets. (*Note:* Information granulation refers to the process of constructing IGs from objects.)

an excellent ability of classifying imbalanced targets. In order to validate the effectiveness and test the limitations of the proposed method, the next subsection will focus on highly imbalanced targets.

4.4. Implementation of the proposed method in imbalanced data sets

This subsection will evaluate the effectiveness of our proposed method for highly imbalanced data. Two skewed datasets, Pima-Indians-Diabetes (2 class labels) and BSWD (3 class labels), which are from the UCI data bank, are employed in this subsection. Table 9 provides a brief explanation of the data background. In addition, for the purpose of testing the limitations of our proposed method under highly skewed situation, we manipulate Pima-Indians-Diabetes data by reducing the proportion of diabetic patients (minority class) from 35% to 10% and 5% by randomly removing them. Then we use "Pima I" and "Pima II" to represent 10% and 5% of the skewed data, respectively. For instance, Pima II denotes that the data set contains 90% healthy examples and 10% diabetic patients.

Moreover, from Section 4.3, we also find that the few minority instances might be distributed to majority class IGs and will then be diluted by majority examples. This might result in some loss of information from the minority examples. Therefore, we propose a new IG construction strategy. We keep the relatively few minority examples intact and only construct IGs from majority examples. Fig. 8 provides the illustration of this idea.

The experiments with two Pima data sets have an optimal classification performance when the dimensions of the sub-attributes are reduced to 8 ($k = 8$). Table 10 summarizes the results of 10-fold CV experiments and then we compare these results with those of NN. In Pima I data, all evaluation metrics indicate that our method has a better performance than those of NN. On average, the overall accuracy and negative accuracy are increased by up to 7% and 16.3%, respectively. These results indicate that our method can increase not only the overall classification performance, but also will improve the ability of identifying the minority examples.

Pima II data denotes the highest skewed situation. Compared with the results of Pima I, this highly skewed situation does indeed lower the performance of the classifier. However, our method still has a better performance than that of NN. The overall accuracy and G-mean of our method are 97.33% and 94.83%, respectively, which are slightly higher than the NN results of 96.64% and 71.32%. In addition, our method can dramatically enhance the ability of detecting minority examples (an average increase of 36.67% over that of NN) without losing the ability to recognize majority examples (positive accuracy = 100%).

The implementation results of another imbalanced data set, BSWD, show a similar conclusion. Unlike Pima-Indians-Diabetes, the BSWD represents a situation of multiple classes. Table 11 provides the summary of 10-fold CV experiments on BSWD. Considering overall accuracy and G-mean, our method shows its remarkable ability to identify not only the major Left and Right class, but also the minor Balance class. Moreover, from the results of the Balance Accuracy (positive accuracy), our approach has a better (averagely increases 22.58%) and a more stable performance (standard deviation decreases from 14.18% to 0%).

5. Discussion and conclusion

Can our method always provide an optimal solution for class imbalance problem? For which situation is our method suitable to be applied? In fact, it was in order to answer these questions that we attempted to validate two ideas in Sections 4.3 and 4.4. The first idea, described in Section 4.3 was to improve an imbalanced situation by considering IGs instead of numerical

Table 10
Comparisons of results between the proposed methodology and NN (BP) on Pima I (10%) and Pima II (5%) data

Classification performance	Methodology			
	Proposed methodology ($k = 8$)		NN (BP)	
	Mean (%)	SD (%)	Mean (%)	SD (%)
Pima I (10%)				
Overall accuracy	100	0.00	93.00	0.90
Positive accuracy	100	0.00	93.38	0.85
Negative accuracy	100	0.00	83.70	4.31
G-mean	100	0.00	88.39	2.68
Pima II (5%)				
Overall accuracy	97.33	1.40	96.64	0.50
Positive accuracy	100	0.00	97.71	0.44
Negative accuracy	90.00	5.27	53.33	18.92
G-mean	94.83	2.72	71.32	11.94

Note: 1. k is the dimensions of input variables and $k = 8$ is obtained by trial and error.
2. SD denotes "standard deviation".
3. Pima I (10%) represents the data set whose proportion of minority examples is 10%. So does Pima II (5%).

Table 11
Comparisons of results between the proposed methodology and NN (BP) on BSWD data set

Classification performance	Methodology			
	Proposed methodology ($k = 8$)		NN (BP)	
	Mean (%)	SD (%)	Mean (%)	SD (%)
Overall accuracy	100	0.00	97.54	1.00
Left accuracy[a]	100	0.00	96.63	1.90
Right accuracy	100	0.00	95.83	1.30
Balance accuracy[c]	100	0.00	77.42	14.18
G-mean[b]	100	0.00	84.67	0.60

Note: SD denotes "standard deviation".
[a] "Left Accuracy" represents the proportion of examples whose class label is "Left" being correctly identified as being "Left". Similarly, "Right Accuracy" ("Balance Accuracy") measures the proportion of "Right" ("Balance") instances being correctly recognized as being "Right" ("Balance").
[b] G-mean $= \sqrt{\text{LeftAccuracy} \times \text{RightAccuracy} \times \text{BalanceAccuracy}}$.
[c] Minority class.

data. Originally, we thought that the "within-variance" of each class data might be the key factor. If the within-variance of the majority class is smaller than that of the minority class, then considering IGs (clusters) can indeed improve a skewed situation. Therefore, we consider the coefficient of variation (VC) which is a measure of dispersion of a probability distribution. It is defined as the ratio of the sample standard deviation σ to the mean μ:

$$\text{VC} = \frac{\sigma}{\mu} \tag{7}$$

If the coefficient of the minority class is larger than the majority class, it means the within-variance of the minority class is larger than that of the majority class. From Table 12, we can find that the coefficients of the minority class of Credit Screening and Pima are larger than those of the majority class. Compared with those methods which operate numerical data, if we consider IGs which are constructed by gathering similar objects together, it can indeed improve an imbalanced situation. The results in Section 4.3 proved it as well.

However, we may encounter some situations in which the coefficient of variation of the minority class is less or equal to that of the majority. Therefore, the second idea was to propose a new IG construction strategy for this situation. As we know, the process of information granulation will reduce some detailed information. Of course, the reduction comes from both majority and minority instances. In order to save the information of the minority instances and improve the class imbalanced situation, our proposed strategy described in Section 4.4 was to build IGs merely from majority examples and keep the minority examples intact. This technique does not merely save valuable information of minority instances, but it also improves the skewed class situation. The results of Pima I, Pima II, and BSWD confirmed the benefits of our proposed strategy.

To sum up, in this study a novel granular computing model called the "information granulation based data mining approach" was proposed for classifying imbalanced data. Experimental results showed that extracting knowledge from IGs has some benefits over building classifiers from numerical data. Without considering class distribution, the advantages of

Table 12
The coefficient of variation of each class of imbalanced data

Data	Coefficient of variation (VC)
Credit screening	Unacceptable:[a] 13.62 Acceptable: 6.21
Pima	Diabetic:[a] 1.90 Healthy: 1.13
BSWD	Balanced: 0.48 Left: 0.47 Right:[a] 0.47

[a] Denotes the minority class.

our method include a slight better overall accuracy and a much faster execution time than the numerical computing models. The results also show that our proposed method might be a possible solution of class imbalance problems. It has an impressive ability to improve classification performance and can dramatically increase the performances of classifying all instances, including majority and minority examples. In addition, this study indicates that introducing the LSI based feature extraction technique (SVD) into the information based data mining model will indeed reduce the amount of sub-attributes. It not only improves the classification performance, but it also saves much execution time and storage space.

Acknowledgement

The authors would like to thank the National Science Council, Taiwan, ROC for financially supporting this research under Contract No. NSC 95-2416-H-009-034-MY3.

References

[1] A.G. Akritas, G.I. Malaschonok, Applications of singular-value decomposition (SVD), Mathematics and Computers in Simulation 67 (2004) 15–31.
[2] H. Altincay, C. Ergun, Clustering based under-sampling for improving speaker verification decisions using AdaBoost, Lecture Notes in Computer Science 3138 (2004) 698–706.
[3] B.D. Anastasiadis, G.D. Magoulas, Analysing the localization sites of proteins through neural networks ensembles, Neural Computing and Application 15 (2006) 277–288.
[4] R. Barandela, J.S. Sanchez, V. Garcia, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (2003) 849–851.
[5] A. Bargiela, W. Pedrycz, Recursive information granulation: aggregation and interpretation issues, IEEE Transactions on Systems, Man, and Cybernetics 33 (2003) 96–112.
[6] G. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations 6 (2004) 20–29.
[7] M.W. Berry, S.T. Dumais, G.W. O'Brien, Using linear algebra for intelligent information retrieval, SIAM Review 37 (1995) 573–595.
[8] S. Bodjanova, Granulation of a fuzzy set: Nonspecificity, Information Sciences 177 (2007) 4430–4444.
[9] G. Castellano, A.M. Fanelli, Information granulation via neural network-based learning, in: IFSA World Congress and 20th NAFIPS International Conference, vol. 5, 2001, pp. 3059–3064.
[10] V. Castelli, A. Thomasian, C.-S. Li, CSVD: Clustering and Singular Value Decomposition for approximate similarity search in high-dimensional spaces, IEEE Transaction on Knowledge and Data Engineering 15 (2003) 671–685.
[11] N.V. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 231–357.
[12] N.V. Chawla, N. Japkowicz, A. Kolcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations 6 (2004) 1–6.
[13] Y. Chen, Y. Yao, A multiview approach for intelligent data analysis based on data operators, Information Sciences 178 (2008) 1–20.
[14] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, A. Geissbuhler, Learning from imbalanced data in surveillance of nosocomial infection, Artificial Intelligence in Medicine 37 (2006) 7–18.
[15] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.
[16] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, Journal of the Society for Information Science 41 (1990) 391–407.
[17] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling methods for learning from imbalanced data sets, Computational Intelligence 20 (2004) 18–36.
[18] A.J. Freeman, M.D. Skapura, Neural Networks: Algorithms, Applications, and Programming Techniques, Addison Wesley, Reading, MA, 1992.
[19] H. Guo, H.L. Viktor, Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, SIGKDD Explorations 6 (2004) 30–39.
[20] Y.-M. Huang, C.-M. Hung, H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Analysis: Real World Applications 7 (2006) 720–747.
[21] K. Huang, H. Yang, I. King, M. Lyu, Learning classifiers from imbalanced data based on biased minimax probability machine, in: Proceedings of the '04 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), 2004, pp. 558–563.
[22] A.K. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 4–37.
[23] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intelligent Data Analysis 6 (2002) 429–449.
[24] T. Jo, N. Japkowicz, Class imbalances versus small disjuncts, SIGKDD Explorations 6 (2004) 40–49.
[25] Y. Liu, N.V. Chawla, M.P. Harper, E. Shriberg, A. Stolcke, A study in machine learning from imbalanced data for sentence boundary detection in speech, Computer Speech and Language 20 (2006) 468–494.
[26] T. Liu, Z. Chen, B. Zhang, W. Ma, G. Wu, Improving text classification using local latent semantic indexing, in: Proceedings of the 4th IEEE International Conference on Data Mining, 2004, pp. 162–169.
[27] J. Liu, Q. Hu, D. Yu, A weighted rough set based method developed for class imbalance learning, Information Sciences 178 (2008) 1235–1256.
[28] J.-M. Ma, W.-X. Zhang, Y. Leung, X.-X. Song, Granular computing and dual Galois connection, Information Sciences 177 (2007) 5365–5377.
[29] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1967, pp. 281–297.

[30] O. Oyeleye, E.A. Lehtihet, A classification algorithm and optimal feature selection methodology for automated solder joint defect inspection, Journal of Manufacturing Systems 17 (1998) 251–262.
[31] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (2007) 28–40.
[32] D.W. Philip, Neural Computing: Theory and Practice, Van Nostrand Reinhold, New York, 1992.
[33] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (2001) 203–231.
[34] P. Radivojac, N.C. Chawla, A.K. Dunker, Z. Obradovic, Classification and knowledge discovery in protein databases, Journal of Biomedical Informatics 37 (2004) 224–239.
[35] R. Ranawana, V. Palade, Optimized precision – a new measure for classifier performance evaluation, in: 2006 IEEE Congress on Evolutionary Computation, Vancouver, Canada, 2006, pp. 2254–2261.
[36] D.E. Rumelhart, J.L. McClelland, Parallel Distributed Processing, MIT Press and the PDP Research Group, Cambridge, 1986.
[37] C.-T. Su, L.-S. Chen, Y. Yih, Knowledge acquisition through information granulation for imbalanced data, Expert System with Applications 31 (2006) 531–541.
[38] C.-T. Su, L.-S. Chen, T.-L. Chiang, A neural network based information granulation approach to shorten the cellular phone test process, Computers in Industry 57 (2006) 412–423.
[39] L. Xu, M.-Y. Chow, A classification approach for power distribution systems fault cause identification, IEEE Transactions on Power Systems 21 (2006) 53–60.
[40] Y.Y. Yao, J.T. Yao, Granular computing as a basis for consistent classification problems, in: Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining, 2002, pp. 101–106.
[41] K. Yoon, S. Kwek, An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics, in: 5th International Conference on Hybrid Intelligent Systems, 2005, pp. 303–308.
[42] L.A. Zadeh, Fuzzy sets and information granularity, in: M.M. Gupta, R.K. Ragade, R.R. Yager (Eds.), Advances in Fuzzy Set Theory and Applications, North Holland, Amsterdam, 1979, pp. 3–18.
[43] L.A. Zadeh, A new direction in AI: toward a computational theory of perceptions, AI Magazine 22 (2001) 73–84.
[44] L.A. Zadeh, Toward a generalized theory of uncertainty (GTU) – an outline, Information Sciences 172 (2005) 1–40.
[45] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 63–77.
[46] W. Zhu, Topological approaches to covering rough sets, Information Sciences 177 (2007) 1499–1508.
[47] W. Zhu, Generalized rough sets based on relations, Information Sciences 177 (2007) 4997–5011.