ELSEVIER

Expert Systems with Applications 31 (2006) 531–541

Expert Systems with Applications

www.elsevier.com/locate/eswa

Knowledge acquisition through information granulation for imbalanced data

Chao-Ton Su [a,*], Long-Sheng Chen [b], Yuehwern Yih [c]

[a] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, 101, Kuang Fu Road, Sec. 2, Hsinchu 300, Taiwan, ROC

[b] Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu 300, Taiwan, ROC

[c] School of Industrial Engineering, Purdue University, IN 47907-2023, USA

Abstract

When learning from imbalanced/skewed data, which almost all the instances are labeled as one class while far few instances are labeled as the other class, traditional machine learning algorithms tend to produce high accuracy over the majority class but poor predictive accuracy over the minority class. This paper proposes a novel method called 'knowledge acquisition via information granulation' (KAIG) model which not only can remove some unnecessary details and provide a better insight into the essence of data but also effectively solve 'class imbalance' problems. In this model, the homogeneity index (*H*-index) and the undistinguishable ratio (*U*-ratio) are successfully introduced to determine a suitable level of granularity. We also developed the concept of sub-attributes to describe granules and tackle the overlapping among granules. Seven data sets from UCI data bank, including one imbalanced diagnosis data (pima-Indians-diabetes), are provided to evaluate the effectiveness of KAIG model. By using different performance indexes, overall accuracy, *G*-mean and Receiver Operation Characteristic (ROC) curve, the experimental results comparing with C4.5 and Support Vector Machine (SVM) demonstrate the superiority of our method.

Keywords: Information granulation; Fuzzy ART; Granular computing; Knowledge acquisition; Imbalanced data

1. Introduction

Learning from imbalanced/skewed data is an important topic and rises very often in practice. In such kind of data, one class might be represented by a large number of examples while the other is represented by only a few. Many real world data have these characteristics, such as fraud detection, text classification (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Chawla, Japkowicz, & Kolcz, 2004) telecommunications management, oil spill detection, risk management, medical diagnosis/monitoring, financial analysis of loan policy or bankruptcy (Batista, Prati, & Monard, 2004; Chawla et al., 2004; Grzymala-Busse, Stefanowski, & Wilk, 2004) and protein data (Provost & Fawcett, 2001). Traditional classifiers seeking an accurate performance over a full range of instances are not suitable to deal with imbalanced learning tasks (Batista et al., 2004; Chawla et al., 2004; Guo & Viktor, 2004) since they tend to classify all data into the majority class, which is usually the less important class. Therefore, these traditional algorithms often produce high accuracy over the majority class, but poor predictive accuracy over the minority class.

To cope with imbalanced data sets, there are some methods proposed in literatures, such as the methods of sampling (Batista et al., 2004; Chawla et al., 2002; Guo & Viktor, 2004), adjusting the cost-matrices (Cristianini & Shawe-Taylor, 2000), and moving the decision thresholds (Chawla et al., 2002; Huang, Yang, King, & Lyu, 2004; Jo & Japkowicz, 2004). Sampling methods reduce data imbalance—by 'down-sampling' (removing) instances from majority class or 'up-sampling' (duplicating) the training instances from the minority class or both. The second kind of methods improves the prediction accuracy by adjusting the cost (weight) for each class or changing the strength of rules (Batista et al., 2004). The third school of methods tries to adapt the decision thresholds to impose bias on the minority class. However, these three schools of methods lack a rigorous and systematic treatment on imbalanced data (Huang et al., 2004). For example, down-sampling the data will lose information, while up-sampling will introduce noise.

In this study, we introduce the concept of 'information granulation' to solve class imbalance problems. There are two

* Corresponding author. Tel.: +886 357 429 36; fax: +886 357 222 04.
E-mail addresses: ctsu@mx.nthu.edu.tw (C.-T. Su), alexchn.iem91g@nctu.edu.tw (L.-S. Chen), yih@purdue.edu (Y. Yih).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.082

reasons why we propose this concept to tackle this issue. The first one is human instinct. As human beings, we have developed a granular view of the world. When describing a problem, we tend to shy away from numbers and use aggregates to ponder the question instead. This is especially true when a problem involves incomplete, uncertain, or vague information. It may be sometimes difficult to differentiate distinct elements, and so one is forced to consider 'information granules' (IG) which are collections of entities arranged together due to their similarity, functional adjacency and indistinguishability (Bargiela & Pedrycz, 2003; Castellano & Fanelli, 2001; Yao & Yao, 2002; Zadeh, 1979). A typical example is the theory of rough sets (Walczak & Massart, 1999). The process of constructing IGs is referred to as information granulation. This was first pointed out in the pioneering work of Zadeh (1979) who coined the term 'information granulation', and emphasized the fact that a plethora of details does not necessarily amount to knowledge. Granulation serves as an abstraction mechanism for reducing an entire conceptual burden. The essential factor driving the granulation of information is the need to comprehend the problem and have a better insight into its essence, rather than get buried in all the unnecessary details. By changing the size of the IGs, we can hide or reveal more or less details (Bargiela & Pedrycz, 2003).

The second reason is about the behavior of data. In many practical datasets, such as medical/diagnosis, inspection, fault monitoring and fraud detecting data, the normal group and abnormal group are considered separate populations. Taguchi & Juoulum (2002) thought every abnormal condition (or a condition outside 'healthy' group) is considered unique, since the occurrence of such a condition is different. Tolstoy's quote in *Anna Karenina*: 'All happy families look alike. Every unhappy family is unhappy after its own fashion' is also noted to illustrate their opinions (Taguchi & Juoulum, 2002). Therefore, we can clearly understand the normal group (i.e. healthy patients, good products) looks alike while the abnormal group (i.e. sick patients, defective products) is unique. If we construct IGs by the similarity of numerical data, the amount of IGs in normal group will be remarkably smaller than the size of normal numerical data. In other words, if we consider IGs instead of numerical data, it might increase the proportion of abnormal data and improve imbalanced/skewed situation of data.

In this study, we propose a 'knowledge acquisition via information granulation' (KAIG) model which can improve classification performance by controlling the reduction of unnecessary details. In KAIG model, Fuzzy ART (Adaptive resonance theory) neural network is utilized to construct IGs. The two indexes, the homogeneity index (H-index) and the undistinguishable ratio (U-ratio), are developed to determine a suitable level of granularity. The concept of sub-attributes is presented to tackle the overlapping among granules. Six data sets (one for illustrative example) from data bank are employed to illustrate our method and evaluate the effectiveness of our proposed model. Besides, one imbalanced diagnosis dataset, pima-Indians-diabetes, is provided to demonstrate the superiority of our method in solving class imbalance class problem by using the indexes, overall accuracy, G-mean and receiver operation characteristic (ROC) curve.

2. Granular computing

Granular computing, which is oriented towards the representation and processing of IGs, is quickly becoming an emerging conceptual and computing paradigm of information processing (Bargiela & Pedrycz, 2003). It is a superset of the theory of fuzzy information granulation, rough set theory and interval computations, and is a subset of granular mathematics. Granular computing as opposed to numeric computing is knowledge-oriented. Numeric computing is data oriented. The main issues (Castellano & Fanelli, 2001) of granular computing are how to construct the IGs, and to describe IGs. One particular question that arises is how to determine the level of granularity. We discuss these issues in the next sections.

2.1. Construction of information granules

In the issue of constructing IGs, there are many approaches, such as the Self Organizing Map (SOM) network (Castellano & Fanelli, 2001), Fuzzy C-means (FCM), rough sets, shadowed sets (Bargiela & Pedrycz, 2003) used to do this. Because IGs exist at different levels of granularity, we usually group granules of similar 'size' (that is granularity) in a single layer. If more detailed processing is required, smaller IGs are selected. Fig. 1 illustrates this concept of granularity. At the lowest level, we are concerned with numeric processing. 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. (Bargiela & Pedrycz, 2003). In this study, the Fuzzy ART is utilized to construct IGs.

ART is a well established neural network theory developed by Carpenter, Grossberg, and Rosen (1991). The ART network is also a famous method of clustering. Instead of clustering by a given number of clusters, it assigns patterns onto the same cluster by comparing their similarity. The detailed algorithm of Fuzzy ART can be found in (Serrano-Gotarredona, Linares-Barranco, & Andreou, 1998).

The major difference between ART and other unsupervised neural networks is the so called vigilance parameter (ρ) which is viewed as a granularity and can be adjusted by the users to

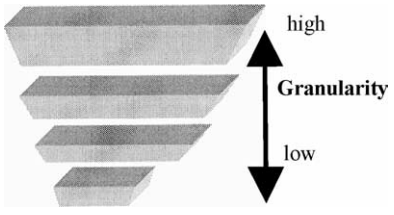

Fig. 1. An information-processing pyramid (Bargiela & Pedrycz, 2003).

control the degree of similarity of patterns placed on the same cluster. In an ART, the degree of similarity between a new pattern and a stored pattern is defined. This similarity, compared to ρ, is a measure to ensure whether the new pattern is properly classified or not. The other unsupervised learning neural networks which do not implement vigilance may cause a significantly different input pattern to be forced into an inappropriate cluster. In contrast to some other cluster methods, an ART network will not automatically force all input vectors onto a cluster if they are not sufficiently similar. This is the reason why the ART network is employed in this study to construct the IGs.

There are three similar ART architectures, namely ART 1, ART 2, and Fuzzy ART. ART 1 is designed for binary-valued input patterns, and ART 2 is for continuous-valued patterns. Fuzzy ART is the most recent adaptive resonance framework that provides a unified architecture for both binary and continuous valued inputs. There are several factors that motivated us to use Fuzzy ART, and they are as follows (Burke & Kamal, 1995):

(1) Unlike ART1, Fuzzy ART does not require a completely binary representation of the parts to be grouped. In addition, Fuzzy ART possesses the same desirable stability properties as ART1 and a simpler architecture than that of ART2.
(2) ART2 can experience difficulty in achieving good categorizations if the input patterns are not all normalized to a constant length. However, such normalization can possibly destroy valuable information. Besides, there is a serious dependency of classification results in the case of ART1 on the sequence of input presentation.

As a result, the Fuzzy ART network is employed to construct IGs in this study.

2.2. Selection of granularity

Selecting an appropriate size of IGs is a difficult task. Enough background knowledge is required to determine how similar objects should be gathered together to form one IG. An objective index is needed to select the appropriate similarity of granules. We proposed *H*-index and *U*-ratio to solve this problem.

The basic assumption of the *H*-index is that the classes of objects should be equal if their values of attributes are sufficiently similar. This implies that we always make the same decision under a similar condition. Because we form granules by the similarity of objects, the objects in the same granule should have the same class. The *H*-index is used to measure the consistency of the class of the objects in one IG. The *H*-index is defined as

$$H-\text{index} = \frac{i}{n} \tag{1}$$

where n represents the number of all objects in one granule and i is the amount of objects possessing the majority class.

Table 1
The information granule-iris example

Condition attributes				Decision attribute (classes)
A	*B*	*C*	*D*	
5.8	2.7	4.1	1	Versicolor
6.2	2.2	4.5	1.5	Versicolor
5.6	2.5	3.9	1.1	Versicolor
5.9	3.2	4.8	1.8	Versicolor
5	3.3	1.4	0.2	Setosa

For example, Table 1 shows one IG involving five objects ($n=5$). There are four condition attributes (namely A–D) in the iris data. The decision attribute (class) of the first four objects is 'versicolor', but the last one has a different decision attribute, 'setosa'. In this example, 'versicolor' is the majority class and $i=4$. The *H*-index of this IG is 4/5.

Another index for selecting similarity is the *U*-ratio. In the preceding example, 'versicolor' is the majority of the classes. So it is assigned to be the class of this IG. If there was another granule described as Table 2, and we are unable to distinguish the class of the IG, then we call that granule an 'undistinguishable granule.' The *U*-ratio is defined as

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

where u represents the number of undistinguishable granules and m represents the quantity of all granules.

This index is to calculate the proportion of undistinguishable granules to all granules. If there are 10 granules and two of them are undistinguishable granules, which means u is equal to 2 and m is equal to 10, then the *U*-ratio is equal to 0.2.

By using these two indexes, we can determine the similarity of the IGs. In the present study, the larger the *H*-index the better it is, because it means that more objects in one granule possess the same class. There is no need to set up the index to a fixed value. The size of the index depends on the domain knowledge or how large an error you can tolerate. On the other hand, the *U*-ratio is the opposite. As far as the *U*-ratio is concerned, the smaller the better. It's difficult to process an undistinguishable granule, so we need to view them carefully. However, we try to avoid this situation by setting the *U*-ratio as small as possible. In other words, if we select a specific similarity where the *H*-index is larger and the *U*-ratio is smaller, then this similarity is the best solution.

Table 2
The undistinguishable information granule

Condition attributes				Decision attribute
A	*B*	*C*	*D*	
5.4	2.2	3.9	1.2	Versicolor
6.8	3.4	5.6	2.4	Virginica

Table 3
Two IGs represented by hyperbox form

IGs	Attributes	
	X_1	X_2
A	(a_1^-, a_1^+)	(a_2^-, a_2^+)
B	(b_1^-, b_1^+)	(b_2^-, b_2^+)

2.3. Representing the information granules

In this section, we utilize hyperboxes to represent IGs (Pedrycz & Bargiela, 2002). A hyperbox $[b]$ defined in R^n is fully described by its lower (b^-) and upper corner (b^+), where b^- and b^+ are vectors in R^n. An important and frequently used universal set is the set of all points in the n-dimensional space. This set is denoted as R^n. Using b^- and b^+ we can express the hyperbox as $[b]=[b^-,b^+]$.Consider two IGs (hyperboxes) $A=[a]$ and $B=[b]$ defined in R^2. More explicitly, we follow a full notation $[a]=[a^-,a^+]$ and $[b]=[b^-,b^+]$. These two granules are described as Table 3.

As Fig. 2 shows, there are overlaps between two granules A and B. This makes it difficult to handle by knowledge acquisition tools. This is because most of knowledge acquisition algorithms are not designed to deal with IGs, especially when overlapping occurs between granules. Unfortunately, the overlapping situation always happens in real world. In this study, we introduce the concept of 'sub-attributes' to tackle the problem of overlaps between granules.

We can explain this idea of 'sub-attributes' by using Fig. 2. In axis X_1 (attribute 1), the overlapping part of two granules are separated into overlapping part ($[b_1^-, a_1^+]$) and non-overlapping parts ($[a_1^-, b_1^-]$ and $[a_1^+, b_1^+]$). These sub-intervals, $[a_1^-, b_1^-]$, $[b_1^-, a_1^+]$ and $[a_1^+, b_1^+]$, are named as X_{11}, X_{12}, X_{13} which are so called 'sub-attributes.' The binary variable which is employed to be the values of sub-attributes represents whether an IG contains these sub-intervals or not. The results of rewriting the IGs by using sub-attributes can be found in Table 4. We divide the original attribute X_1 into sub-attributes X_{11}, X_{12}, X_{13}; and attribute X_2 into X_{21}, X_{22}, X_{23}. Then, these two granules are rewritten by replacing the original attributes with sub-attributes. By introducing the concept of sub-attributes, we can easily extract knowledge from the granules even if the overlapping situation always exists.

Table 4
The IGs with sub-attributes

Original attributes	X_1			X_2		
Sub-attributes	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	X_{23}
IGs	$[a_1^-, b_1^-]$	$[b_1^-, a_1^+]$	$[a_1^+, b_1^+]$	$[a_2^-, b_2^-]$	$[b_2^-, a_2^+]$	$[a_2^+, b_2^+]$
A	1	1	0	1	1	0
B	0	1	1	0	1	1

This method can maintain the complete characteristics of data. The IGs with addition of sub-attributes are suitable for all knowledge acquisition algorithms. It is not required to adjust the computational architecture of these algorithms. However, too many sub-attributes may be generated in the situation of natural overlapping which the values of the condition attributes are continuous and diverse. Therefore, as we often do in data preparation phase of data mining, we suggest descretizing data before implementing KAIG model to control the number of sub-attributes.

3. Proposed methodologies

This section describes in detail the procedure of the KAIG model. First, we address how the IGs are formed from numerical data. Secondly, H-index and U-ratio are introduced to determine the level of granularity which can be used to construct IGs in Fuzzy ART. Then, we try to describe IGs by using sub-attributes and extract knowledge from them. The well-known dataset, iris, will serve as an illustrative example.

3.1. The KAIG model

Fig. 3 shows the proposed KAIG model. We explain it by the following steps:

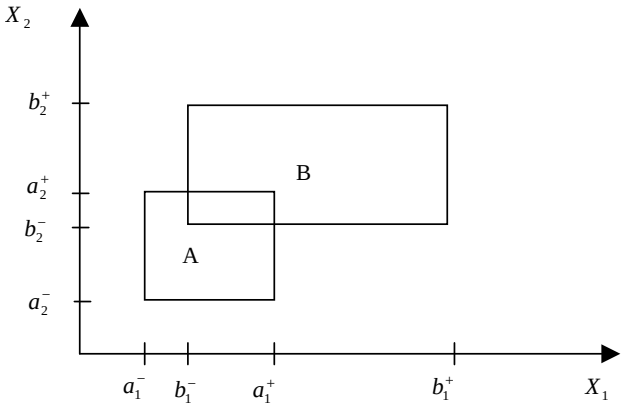

Fig. 2. The overlap between IGs.

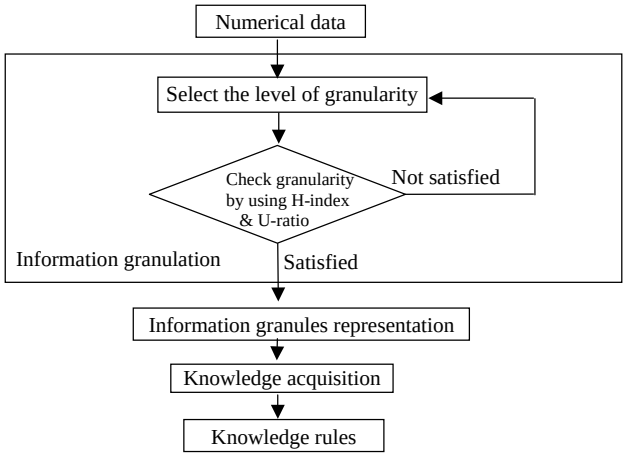

Fig. 3. knowledge acquisition via information granulation (KAIG) model.

3.1.1. Step 1: Information granulation

In step 1, we use Fuzzy ART to construct IGs. But, first thing we need to determine is to select the suitable level of granularity (vigilance). The IGs are formed by the selected granularity. The initial value of granularity is set 1 and then decrease gradually until find one satisfying criteria of *H*-index and *U*-ratio. The found suitable granularity is employed to construct IGs.

3.1.2. Step 2: Information granules representation

IGs are represented in a suitable form that can be handled by knowledge acquisition tools. As mentioned in Section 2.3, these formed IGs are described in hyperboxes. Then, the sub-attributes are applied in these IGs to solve the problem and finally, we can extract knowledge from these IGs.

3.1.3. Step 3: Knowledge acquisition

After describing IGs appropriately and tackling the overlapping situation, we can use knowledge acquisition tools to extract knowledge rules from the granules. In this study, we will compare three famous data mining algorithms, C4.5, Rough sets and neural network (back-propagation), to evaluate their effectiveness in KAIG model.

3.2. Illustrative example

We apply the KAIG model to the well-known data set, iris. It is comprised of 150 examples. We rearrange it randomly and divide it into two subsets, training set (100 objects) and test set (50 examples). We will illustrate the process of KAIG step by step.

3.2.1. Step 1: Information granulation

We input the 100 training examples to the Fuzzy ART to form IGs. We set the parameters of Fuzzy ART $\alpha = 0.01$ and $\beta = 1$. The number of IGs varies with the different level of similarity (vigilance). In this study, similarity value varies gradually from 1 to 0. The similarity 1 represents the numerical data. Next, we need to determine which similarity is suitable by the *H*-index and the *U*-ratio. The *H*-index is 'the larger-the-better' and the *U*-ratio is 'the smaller-the-better'. In Fig. 4, we can find more than one similarity that satisfies this criterion.

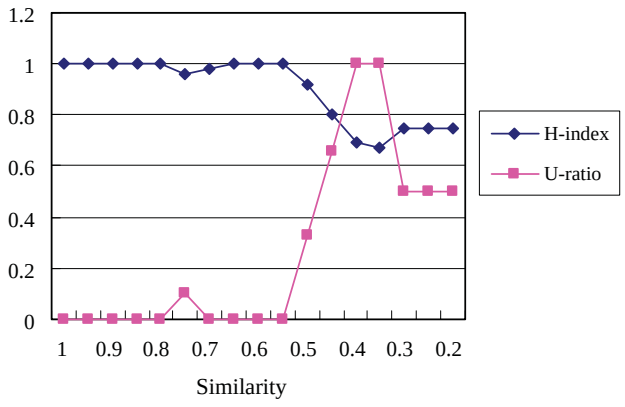

Fig. 4. The *H*-index and *U*-ratio of the iris data.

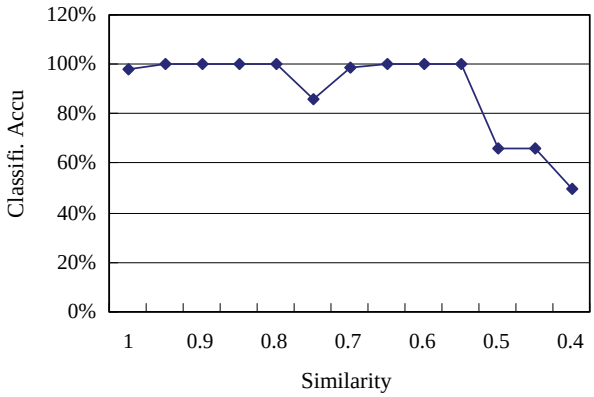

Fig. 5. The performance of classification (Iris data).

These similarities are 0.95–0.8 and 0.7–0.55, where *H*-index = 1 and *U*-ratio = 0. Their performances of classification, as described in Fig. 5, are equal to each other. All classification accuracies are equal to 100%.

When the performances are equally good, the amount of granules becomes another criterion for selecting the similarity. In this study, we use IGs instead of numerical data to acquire knowledge and make decisions. If the smaller similarity is selected, the lesser the amount of granules will be dealt with. This smaller amount of granules may save some training time during the building of the model. Therefore, we select a similarity of 0.55 and the amount of granules is 3.

3.2.2. Step 2: Representing the IGs

We describe these three granules in hyperboxes form and they are shown in Table 5. L_i represents the lower bound of attribute values, and U_i represents the upper limit of attribute values in the *i*th granule. Take granule #1 for example, it contains 33 objects. In condition attribute *A*, the minimum is 4.4 and the maximum is 5.7. We utilize the low limit and upper limit to describe all examples in the same one granule. Granule 1 possesses the same class, setosa. Granule 2 contains 33 examples which are of the same class, versicolor. Granule 3 is comprised of 34 examples which have the same class, virginica.

Next, the original attributes are divided into several sub-attributes. Table 6 shows the IGs and their sub-attributes. The four original condition attributes (*A*, *B*, *C*, *D*) are divided into 17 sub-attributes ($A1,\ldots,D4$). These 17 sub-attributes are used

Table 5
The IGs with the similarity of 0.55

No. of granules		Condition attribute				Classes (No. of examples)		
		A	*B*	*C*	*D*	Setosa	Versicolor	Virginica
#1	L_1	4.4	2.3	1	0.1	33	0	0
	U_1	5.7	4.2	1.9	0.6			
#2	L_2	5	2.2	3	1	0	33	0
	U_2	6.8	3.4	5.1	1.8			
#3	L_3	5.6	2.2	4.8	1.4	0	0	34
	U_3	7.9	3.8	6.9	2.5			

Table 6
The IGs with sub-attributes

Original attributes	A					B				C				D				Classes
Sub-attributes	A1	A2	A3	A4	A5	B1	B2	B3	B4	C1	C2	C3	C4	D1	D2	D3	D4	
	4.4–5.0	5–5.6	5.6–5.7	5.7–6.8	6.8–7.9	2.2–2.3	2.3–3.4	3.4–3.8	3.8–4.2	1–1.9	3–4.8	4.8–5.1	5.1–6.9	0.1–0.6	1–1.4	1.4–1.8	1.8–2.5	
Granule No.1	1	1	1	0	0	0	1	1	1	1	0	0	0	1	0	0	0	Setosa
Granule No.2	0	1	1	1	0	1	1	0	0	0	1	1	0	0	1	1	0	Versicolor
Granule No.3	0	0	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	Virginica

as the inputs for the operation of knowledge acquisition algorithms.

3.2.3. Step 3: Knowledge acquisition

The rough sets method can be utilized to remove superfluous sub-attributes and to acquire knowledge. The theory of rough sets emerged as a major mathematical tool for discovering knowledge. A fundamental principle of a rough set-based learning system is to discover redundancies and dependencies between the given features of a problem to be classified (Mitra, Pal, & Mitra, 2002). In the rough set method, a reduct is the minimal subset of attributes that enable the same classification of objects with full attributes. All results of rough sets are operated by Rosseta software. Readers can find additional information on the theory of rough sets in the references (Hu, Cercone, Han, & Ziarko, 2002; Walczak & Massart, 1999). The knowledge rules extracted by rough set method are listed as follows:

Rule 1 IF $B4 = 1$ THEN Class = setosa;
Rule 2 IF $D2 = 1$ THEN Class = versicolor;
Rule 3 IF $B4 = 0$ AND D2 = 0 THEN Class = virginica;

These knowledge rules can be translated as follows:

Rule 1 ATTRIBUTE $B \in (3.8,4.2]$ THEN Class = setosa;
Rule 2 ATTRIBUTE $D \in (1.0,1.4]$ THEN Class = versicolor;
Rule 3 ATTRIBUTE $B \in (3.8,4.2]$ AND ATTRIBUTE $D \notin (1.0,1.4]$

THEN Class = virginica;

These knowledge rules are applied to test the remaining 50 examples. Table 7 is the minimal reduct of the testing granules. The sub-attributes of testing granules, $B4$ and $D2$, are put into these extracted knowledge rules. The predicted decisions are fully equal to the true ones. Therefore, the classification accuracy is 100%.

In this illustrative example, we reduce some unnecessary detailed information by acquiring knowledge from IGs, but the classification accuracy remains high. Also, the knowledge rules for decision-making are fewer than those extracted from numerical data, which may save the response time of a decision. Table 8 shows the comparison of classification performances.

4. Evaluation of KAIG model

To evaluate the effectiveness of the KAIG model, five data sets which come from databank of UCI machine learning group

Table 7
The minimal reduct of IGs for testing

IGs No.	B4	D2	Classes	
	3.8–4.2	1–1.4	Predicted	True
#1	1	0	Setosa	Setosa
#2	0	1	Versicolor	Versicolor
#3	0	0	Virginica	Virginica

Table 8
The comparison of processing with information granules and numerical data

Methods	Rough sets		KAIG	
Data type	Numerical data (similarity = 1.0)		Information granules (similarity = 0.55)	
Classification accuracy	100%	98%	100%	100%
No. of rules	16		3	

(http://www.ics.uci.edu/~mlearn/) are considered in this section. Table 9 provides brief explanation about the data background, including data size, number of features, data characteristics (binary/continuous), and defined classes. Before implementing, we divide all data sets into training set and testing set with the proportion of 3:1.

With the help of the *H*-index and the *U*-ratio shown in Fig. 6, we can find the suitable similarity of these data sets. According to these determined similarities, numerical data is transformed into IGs by Fuzzy ART. Then, three famous knowledge acquisition algorithms, neural network (BP), decision tree (C4.5 algorithm) and the rough set method, are utilized. Professional II PLUS is employed to build neural network in this study. The optimal neural network (BP) parameter settings, structure and learning iterations shown in Table 10 are obtained by trial and error. Decision trees built by C4.5 are models which each node is a test on an individual variable and a path from the root to a leaf is a conjunction of conditions required for a given classification. See5 (C4.5 commercial version) software was utilized to construct a decision tree in this study. In See5, there are two parameters that can be tuned, during the pruning phase: the minimal number of examples represented at any branch of any feature-value test; and the confidence level of pruning. To avoid the occurrence of over-fitting and generating a simple tree, 2 was set as the minimum number of instances at each leaf and the confidence level for pruning was set at 25%. The inputs and outputs of decision tree and the rough set method are condition attributes and defined classes, respectively.

The comparisons of implementation results are provided in Table 11. Except WDBC, KAIG model has better classification performances in the other five data sets than those of traditional methods which use numerical data. In average, the classification accuracy increases 2.33% and the number of rules is reduced by 48.67% compared with traditional methods. In KAIG model, we can use different kind of knowledge acquisition tools and the results will be different. The classification accuracy averagely increases 0.86, 2.238, 1.182% by applying Rough sets, C4.5 and BP, respectively. In addition, C4.5 has fewer number of knowledge rules (12 rules in average) than those of Rough sets (84.8 rules in average). Therefore, C4.5 is more suitable to be employed in KAIG model than the other two methods.

5. Implementation in imbalanced data

This section will apply KAIG method to overcome the class imbalance problems. C4.5 and SVM are usually utilized as benchmarks or basic learners in related works (Batista et al., 2004; Guo & Viktor, 2004; Huang et al., 2004; Jo & Japkowicz, 2004; Provost & Fawcett, 2001; Radivojac, Chawla, Dunker, & Obradovic, 2004). Therefore, the experimental results of KAIG will be compared with these two methods. A brief introduction about SVM can be found in (Cristianini & Shawe-Taylor, 2000; Wu & Chang, 2005).

5.1. Performance measures

The easiest way to evaluate the performance of classifiers is based on the confusion matrix described as Table 12. TP, FP, TN and FN are defined as bellows.

TP the number of True Positive examples
FP the number of False Positive examples
TN the number of True Negative examples
FN the number of False Negative examples

Traditionally, the performance of a classifier is evaluated by considering the overall accuracy against test cases. However, when learning from imbalanced data sets, the measure is often not sufficient. For example, it is straightforward to create a classifier having an accuracy of 95% in a domain where the majority class proportion corresponds to 95% of the examples, by simply forecasting every new example as belonging to the majority class. Another fact is the metric considers different classification errors to be equally important. But as we know, a highly imbalanced class problem does not have equal error

Table 9
The background of five data sets

Data set	Title	No. of instances	No of attributes	Data characteristics	Class distribution
WDBC	Wisconsin diagnostic breast cancer	683 (699 minus 16 missing data)	9 (remove first attribute—'ID')	All discrete	Benign (65.5%) malignant (34.5%)
CE	Car evaluation database	1728	6	All discrete	Unacceptable (70.023%) acceptable (22.222%) good (3.993%) very good (3.762%)
TAE	Teaching assistant evaluation	151	5	1-continuous 4-discrete	Low (32.45%) medium (33.11%) high (34.44%)
BUPA	BUPA liver disorders	345	6	All continuous	Class 1 (42.03%) class 2 (57.97%)
WINE	Wine recognition data	178	12	All continuous	Class 1 (33.15%) class 2 (39.89%) class 3 (26.96%)
PIMA	Pima Indians diabetes	768	8	All continuous	Healthy (65%) diabetic (35%)

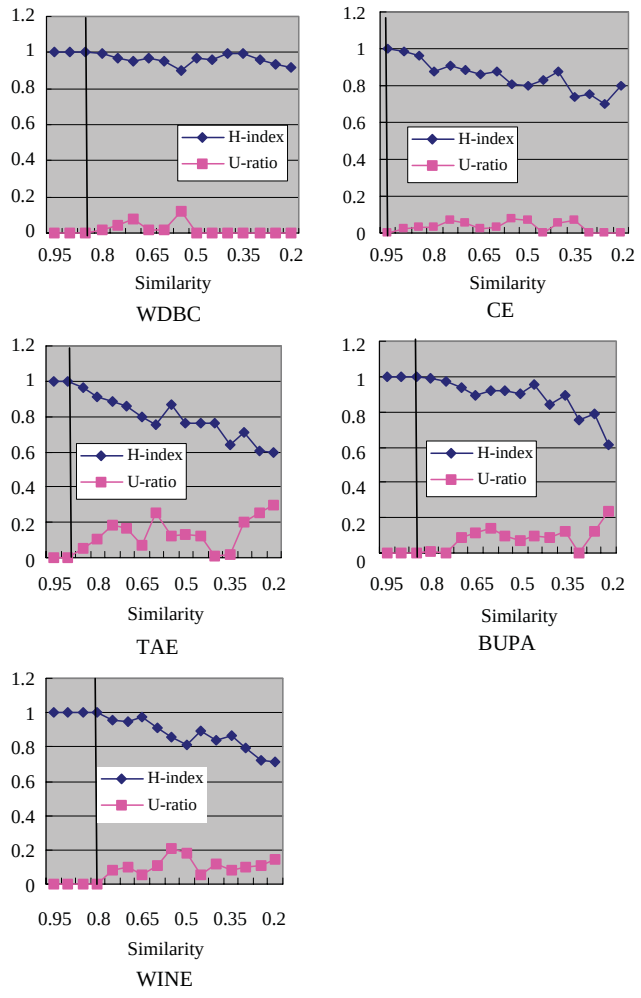

Fig. 6. The *H*-indexes and *U*-ratios of five data sets.

costs that favor the minority class, which is often the class of primary interest. Therefore, following the available studies (Batista et al., 2004; Estabrooks, Jo, & Japkowicz, 2004; Guo & Viktor, 2004; Provost & Fawcett, 2001; Radivojac, Chawla, Dunker, & Obradovic, 2004), we use overall accuracy (including positive accuracy and negative accuracy), *G*-mean and receiver operation characteristic (ROC) curve to evaluate our KAIG model. The *G*-mean is defined as

$$\sqrt{\text{Positive Accuracy} \times \text{Negative Accuracy}} \tag{3}$$

where positive accuracy and negative accuracy are calculated as TP/(FN + TP) and TN/(TN + FP). This measure is to maximize the accuracy on each of two classes while keeping these accuracies balanced. For instance, a high positive accuracy by a low negative accuracy will result in poor *G*-mean.

Another index is ROC curve, which is a technique for summarizing a classifier's performance over a range by considering the tradeoffs between TP rate and FP rate. The TP rate and FP rate are calculated as TP/(FN + TP) and FP/(FP + TN). We use the term ROC space to denote the coordinate system used for visualizing classifier's performance. In ROC space, TP rate is represented on the *Y*-axis and FP rate is represented on the *X*-axis. Each classifier is represented by the point in ROC space corresponding to its (FP rate, TP rate) pair. A ROC analysis also allows the performance of multiple classification functions to be visualized and compared simultaneously. The area under ROC curve (AUC) represents the expected performance as a single scalar. The AUC has a known statistical meaning: it equals to the Wilconxon test of ranks, and is equivalent to several other statistical measures for evaluating classification and rank models (Hand, 1997).

5.2. Diagnosis data

The imbalance class problems often happen in medical diagnosis data. Therefore, pima-Indians-diabetes whose information shows in Table 9 is employed to verify effectiveness of our model. Results for this data set, shown in Table 13, were averaged over four-fold cross validation (CV) experiments, which the data set was partitioned into four equal sized sets and each set was then in turn used as the test set. Besides, in order to test the robustness of KAIG model, we reduce the proportion of minority class from 35% to 10% and 5% by removing the number of minor examples randomly.

In the experiments of 35%, 10% and 5%, the results indicate that KAIG model has better performance than those of SVM and C4.5 against highly imbalanced data sets, in term of the negative accuracy. In average, KAIG owns 58.08% of negative accuracy far better than 14.55% of SVM and 27.49% of C4.5. It means KAIG has excellent capability of detecting minor examples (diabetic patients). Meanwhile, KAIG does not lose overall accuracy and positive accuracy. They are even better than those of SVM and C4.5 in experiment of 35%.

Table 10
The setting of parameters in neural network (BP)

Data Set	Data type	Structure	Learning rate	Momentum	Iterations
WDBC	Numerical	9-11-1	0.2	0.9	20,000
	Granule(0.85)	90-160-1	0.2	0.9	20,000
CE	Numerical	6-9-1	0.2	0.9	10,000
	Granule(0.95)	21-35-1	0.3	0.9	20,000
TAE	Numerical	5-6-1	0.2	0.9	20,000
	Granule(0.85)	69-120-1	0.2	0.9	20,000
BUPA	Numerical	6-5-1	0.3	0.8	30,000
	Granule(0.85)	26-31-1	0.2	0.9	15,000
WINE	Numerical	13-5-1	0.3	0.7	10,000
	Granule(0.8)	35-7-1	0.2	0.8	15,000

Table 11
The comparison of classification performance

Methods	Classification accuracy	Data type					
		Numerical data (similarity = 1.0) Traditional methods			Granules (similarity = 0.85) KAIG		
		Train (%)	Test (%)	No. of rules	Train (%)	Test (%)	No. of rules
WDBC							
	Rough sets	100	92.23	212	100	89.47	58
	Decision tree (C4.5)	97.5	97.06	10	93.4	94.74	4
	Neural network (BP)	96.66	100	–	100	89.64	–
CE		Similarity = 1.0			Similarity = 0.95		
	Rough sets	100	89.58	385	100	88.96	207
	Decision tree (C4.5)	97.4	92.8	75	98.4	95.58	36
	Neural network (BP)	91.18	91.09	–	94.04	92.80	–
TAE		Similarity = 1.0			Similarity = 0.90		
	Rough sets	84.96	84.21	90	95.95	87.37	68
	Decision tree (C4.5)	60.2	47.36	13	64.9	48.39	11
	Neural Network (BP)	68.23	69.05	–	78.61	84.21	–
BUPA		Similarity = 1.0			Similarity = 0.85		
	Rough sets	100	63.95	165	100	66	80
	Decision tree (C4.5)	76.4	65.1	15	78.2	70	5
	Neural Network (BP)	69.35	64.47	–	100	66.15	–
WINE		Similarity = 1.0			Similarity = 0.8		
	Rough sets	100	93.18	31	100	95.65	11
	Decision tree (C4.5)	95.6	90.9	6	96.7	95.7	4
	Neural network (BP)	87.63	86.49	–	85.87	84.21	–

Table 12
Confusion matrix for binary class problem

	Predicted positive	Predicted negative
Actual positive	TP (the number of true positive)	FN (the number of false negative)
Actual negative	FP (the number of false positive)	TN (the number of true negative)

Both *G*-mean and ROC curves shown in Fig. 7 also demonstrate the superiority of our method. In extreme skewed data (10 and 5%), *G*-mean is more sensitive than overall accuracy. When negative accuracy decreases dramatically, *G*-mean can indicate these changes but overall accuracy cannot. ROC curves provide visual results which can easily compare these three methods and find KAIG has best performances (AUC) in different experiments.

Table 13
The results in different proportion of minor class examples

Methods	KAIG				SVM				Decision tree (C4.5)			
	Training		Test		Training		Test		Training		Test	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
(35%)												
Overall accuracy	91.97	2.5	78.78	2.5	76.82	1.4	75.52	2.8	81.50	4.28	74.22	3.1
Pos. Acc.	93.07	2.3	84.00	4.5	93.07	0.5	92.60	2.7	87.94	7.50	83.20	2.8
Neg. Acc.	85.24	2.1	70.52	8.0	46.52	4.7	43.66	3.3	71.40	8.73	57.46	8.3
G-mean	90.67	2.59	76.46	3.85	65.73	3.18	63.56	3.29	78.95	3.24	68.99	4.80
(10%)												
Overall accuracy	95.01	1.1	87.05	2.3	89.93	0	89.93	0	91.55	1.7	88.49	1.6
Pos. Acc.	99.33	0.8	92.20	1.2	100	0	100	0	98.73	1.9	96.80	3.1
Neg. Acc.	52.98	11.9	41.08	20.5	0	0	0	0	27.38	20.4	14.29	24.0
G-mean	72.16	7.91	59.73	17.0	0	0	0	0	44.43	30.6	23.63	32.1
(5%)												
Overall accuracy	97.48	0.7	94.89	1.6	94.94	0	94.70	0	96.52	0.8	93.56	1.0
Pos. Acc.	98.54	0.9	98.60	1.7	100	0	100	0	99.47	0.7	98.20	0.8
Neg.Acc.	72.50	13.2	28.57	26.1	0	0	0	0	41.25	14.9	10.72	13.7
G-mean	84	7.69	44	33.3	0	0	0	0	63	12.7	23	26.9

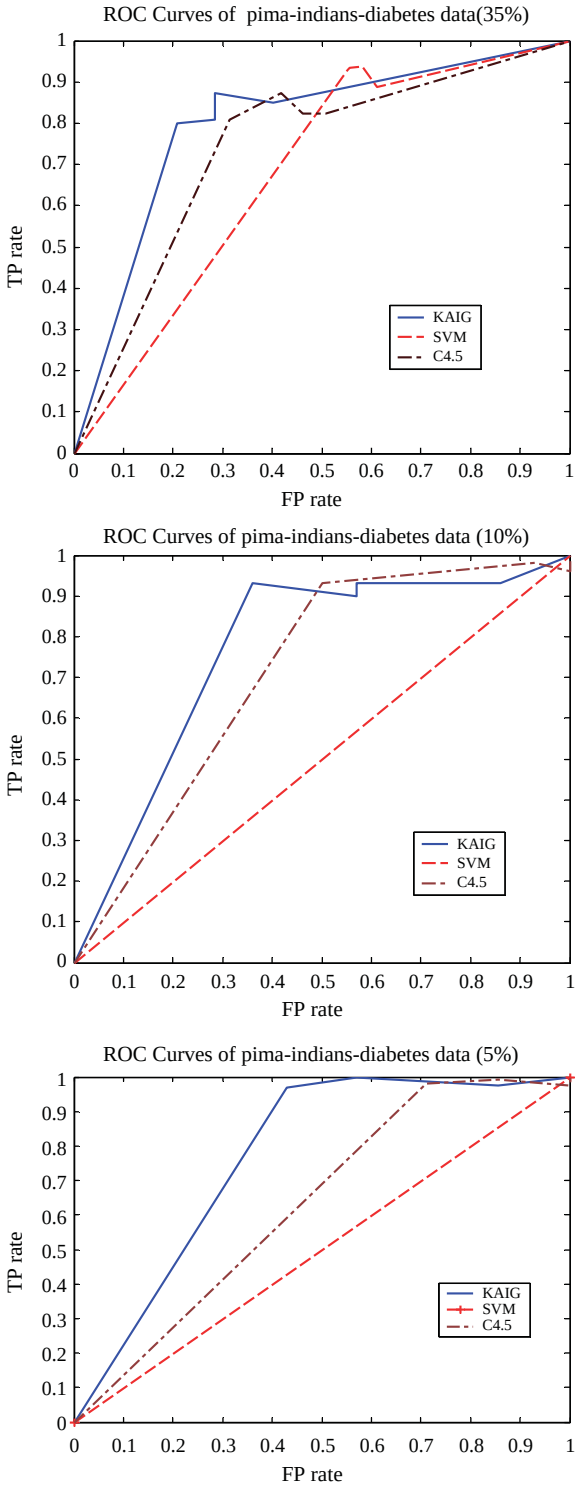

Fig. 7. ROC curves of pima-indians-diabetes data.

6. Conclusions

This study introduces the concept of information granulation to solve class imbalance problems. A novel method called KAIG model is presented. In this model, we propose two indexes to determine the level of granularity and the 'sub-attributes' concept to describe IGs. The experimental results show that the KAIG model can improve classification performance by reducing unnecessary details of information. We also demonstrate that the proposed method has excellent ability of identifying the minority examples in imbalanced learning tasks. In medical diagnosis data, our method can dramatically increase negative accuracy without losing positive accuracy and overall accuracy. ROC curves and *G*-mean also illustrate the superiority of KAIG model compared with C4.5 and SVM.

Construction of IGs is one of many interesting and important issues in granular computing. IGs are aimed at building efficient and user-centered views of the external world and supporting/facilitating our perception of the surrounding physical and virtual world. In our research, we construct IGs by objects' 'similarity', the parameter (vigilance) of Fuzzy ART. It can define the 'indistinghishable, similar, coherency and alike' relations of objects. However, other relations whose definitions are not specific/ concrete, such as 'functional adjacency', also can employ to construct IGs. But, it is hard to define these 'not specific' relations. Therefore, more efforts of studying different relations are necessary in the future researches.

Acknowledgements

This work was supported in part by National Science Council of Taiwan (Grant No. NSC 94-2213-E007-059).

References

Bargiela, A., & Pedrycz, W. (2003). *Granular computing: An introduction*. Boston: Kluwer Academic Publishers.

Batista, G., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, *6*(1), 20–29.

Burke, L., & Kamal, S. (1995). Neural networks and the part family/machine group formation problem in cellular manufacturing: A framework using fuzzy art. *Journal of Manufacturing Systems*, *14*(3), 148–159.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991). Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771.

Castellano, G., & Fanelli, A. M. (2001). Information granulation via neural network-based learning. *IFSA World Congress and 20th NAFIPS international conference*, Vol. 5 pp. 3059–3064.

Chawla, N. V., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 231–357.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations*, *6*(1), 1–6.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling methods for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36.

Grzymala-Busse, J. W., Stefanowski, J., & Wilk, S. (2004). A comparison of two approaches to data mining from imbalanced data. *Lecture Notes in Computer Science*, *3213*, 757–763.

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *SIGKDD Explorations*, *6*(1), 30–39.

Hand, D. J. (1997). *Construction and assessment of classification rules*. New York: Wiley.

Hu, X., Cercone, N., Han, J., & Ziarko, W. (2002). In T. Y. Lin, Y. Y. Yao, & L. A. Zadeh (Eds.), *GRS: A generalized rough sets model, data mining, rough sets and granular computing* (pp. 447–460). New York: Physica-Verlag.

Huang, K., Yang, H., King, I., & Lyu, M. (2004). Learning classifiers from imbalanced data based on biased minimax probability machine. *Proceedings of the 04' IEEE Computer Society conference on computer vision and pattern recognition (CVPR'04)* pp. 558–563.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *SIGKDD Explorations*, *6*(1), 40–49.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

Pedrycz, W., & Bargiela, A. (2002). Granular clustering: A granular signature of data. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, *32*(2), 212–224.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203–231.

Radivojac, P., Chawla, N. C., Dunker, A. K., & Obradovic, Z. (2004). Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics*, *37*, 224–239.

Serrano-Gotarredona, T., Linares-Barranco, B., & Andreou, A. G. (1998). *Adaptive resonance theory microchips-circuit design techniques*. Boston, MA: Kluwer Academic Publishers.

Taguchi, G., & Juoulum, R. (2002). *The Mahalanobis-Taguchi strategy—a pattern technology system*. New York: Wiley.

Walczak, B., & Massart, D. L. (1999). Tutorial: Rough sets theory. *Chemometrics and Intelligent Laboratory Systems*, *47*, 1–16.

Wu, G., & Chang, E. Y. (2005). KBA: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 786–795.

Yao, Y. Y., & Yao, J. T. (2002). Granular computing as a basis for consistent classification problems. *Proceedings of PAKDD'02 workshop on toward the foundation of data mining* pp. 101–106.

Zadeh, L. A. (1979). Fuzzy sets and information granularity. In M. M. Gupta, R. K. Ragade, & R. R. Yager (Eds.), *Advances in fuzzy set theory and applications* (pp. 3–18). Amsterdam: North Holland.