

An Application on Text Classification Based on Granular Computing *

**Xia Zhang
Yixin Yin**

School of Information Engineering
University of Science and Technology Beijing

Haiyan Yu

School of Economics and Management
University of Science and Technology Beijing

ABSTRACT

Machine learning is the key to text classification, a granular computing approach to machine learning is applied to learning classification rules by considering the two basic issues: concept formation and concept relationships identification. In this paper, we concentrate on the selection of a single granule in each step to construct a granule network. A classification rule induction method is proposed.

INTRODUCTION

Knowledge discovery and data mining are frequently applied as a process extracting interesting information or patterns from large databases. It is actually a technique or program which to do automatic inductive reasoning learning, identification and searching for knowledge, patterns, and regularities from data.

Knowledge Discovery in Text (KDT), which uses Text Mining techniques to extract and induce hidden knowledge from unstructured text data, surges in the data and natural language processing research. KDT is a multidiscipline of Artificial Intelligence, machine learning with a stressing on its IE (Information Extraction)-based induction and specific fields practices. Text classification is one of the practices of KDT.

Classification is one of the well studied problems in machine learning and data mining as it involves of discovery knowledge. Text classification is the task of deciding whether a piece of text belongs to any of a set of pre-specified categories. Automatic text classification includes the next six steps: set up data set (training data set and testing data set), text knowledge indexing, feature (key words) extracting and selecting and feature, design a classifier through machine learning, test the classifier with testing set, evaluate the method. Among these steps, designation of a classifier is the most important.

Granular computing (GrC) is an umbrella term which covers any theories, methodologies, techniques, and tools that make use of granules (i.e., subsets of a universe) in problem solving. A subset of the universe is called a granule in granular computing. Basic ingredients of granular computing are subsets, classes, and clusters of a universe. It deals with the characterization of a concept by a unit of thoughts consisting of two parts, the intension and extension of the concept.

In the sight of the fact, Y.Y. Yao presented a granular computing view to classification problems and proposed a granular computing approach to classification. He provided a modeling data mining with granular computing to resolve classification problem.

This paper put this model to an application on text classification based on this model of granular computing.

* National Natural Science Foundation of China(60374032)

GRANULAR COMPUTING BASIC FOR CONSISTENT CLASSIFICATION PROBLEMS

There are two aspects of a concept, the intension and extension of the concept. In the granular computing model for knowledge discovery, data mining, and classification, a set of objects are represented using an information table. The intension of a concept is expressed by a formula of the language, while the extension of a concept is represented as the set of objects satisfying the formula. This formulation enables us to study formal concepts in a logic setting in terms of intensions and also in a set theoretic setting in terms of extensions.

Representation of granular

In order to formalize the problem, an information table was introduced in. An information table can be formulated as a tuple:

$$s = (U, At, L, \{V_a \mid a \in At\}, \{I_a \mid a \in At\}) \quad (1)$$

Where

- U is a finite nonempty set of objects,
- At is a finite nonempty set of attributes,
- L is a language defined using attributes in At ,
- V_a is a nonempty set of values for $a \in At$,
- $I_a : U \rightarrow V_a$ is an information function.

In the language L, an atomic formula is defined as $a=v$, where $a \in At$ and $v \in V_a$. If \neg, \wedge, \vee and \models can be defined with negation, conjunction, disjunction and satisfiability, the following condition can be understood.

$$x \models a = v \text{ iff } I_a(x) = v \quad (2)$$

$$x \models \neg\phi \text{ iff } I_a x \not\models \phi \quad (2)$$

$$x \models \phi \wedge \psi \text{ iff } x \models \phi \text{ and } x \models \psi \quad (3)$$

$$x \models \phi \vee \psi \text{ iff } x \models \phi \text{ or } x \models \psi \quad (4)$$

If ϕ is a formula, the set $m_s(\phi) = \{x \in U \mid x \models \phi\}$, is called the meaning of the formula ϕ in S. ϕ is intension of concept $(\phi, m(\phi))$, and $m(\phi)$ is the extension of concept $(\phi, m(\phi))$. For an atomic formula $a=v$, we can obtain a granule $m(a=v)$, if $m(\phi)$ and $m(\psi)$ are granules corresponding to formulas ϕ and ψ , we obtain granules $m(\phi) \cap m(\psi) = m(\phi \wedge \psi)$ and $m(\phi) \cup m(\psi) = m(\phi \vee \psi)$. (5)

CONSISTENT CLASSIFICATION PROBLEMS

In supervised classification, each object is associated with a unique and pre-defined class label. Suppose an information table is used to describe a set of objects. Without loss of generality, we assume that there is a unique attribute class taking class labels as its value. The set of attributes is expressed as $At = F \cup \{class\}$, where F is the set of attributes used to describe the objects. The goal is to find classification rules of the form, $\phi \Rightarrow class = c_i$, where ϕ is a formula over F and c_i is a class label.

If all objects with the same description over F have the same class label, namely, if $I_F(x) = I_F(y)$, then $I_{class}(x) = I_{class}(y)$, then this information table with attributes of $F \cup \{class\}$ can provide a consistent classification.

MEASURES ASSOCIATED WITH GRANULES

There are three types of quantitative measures associated with granules. Measures of a single granule, measures of relationships between a pair of granules, and measures of relationships between a granule and a family of granules, as well as a pair of family of granules.

The measure of a single granule $m(\phi)$ of a formula ϕ is generality $G(\phi) = |m(\phi)| / |U|$, which indicates the relative size of granule $m(\phi)$.

Given two formulas ϕ and φ , it is usually used to measure the relationship between these two granules $m(\phi)$ and $m(\varphi)$.

The confidence or absolute support of φ provided by ϕ is:

$$AS(\phi \Rightarrow \varphi) = \frac{|m(\varphi \wedge \phi)|}{|m(\phi)|} = \frac{|m(\varphi) \cap m(\phi)|}{|m(\phi)|} \quad (6)$$

The coverage φ provided by ϕ is: Φ

$$CV(\phi \Rightarrow \varphi) = \frac{|m(\varphi \wedge \phi)|}{|m(\varphi)|} = \frac{|m(\varphi) \cap m(\phi)|}{|m(\varphi)|} \quad (7)$$

A family of $\{m(\psi_1), \dots, m(\psi_n)\}$ granules which provided by a family of formulas $\Psi = \{\psi_1, \dots, \psi_n\}$, then we obtain the following probability distribution in terms of $\phi \Rightarrow \psi_i$ is:

$$P(\Psi | \phi) = (AS(\phi \Rightarrow \psi_1), \dots, AS(\phi \Rightarrow \psi_n)) \quad (8)$$

The conditional entropy:

$$H(\Psi | \phi) = - \sum_{i=1}^n AS(\phi \Rightarrow \psi_i) * \log(AS(\phi \Rightarrow \psi_i)) \quad (9)$$

Suppose another family of formulas $\Phi = \{\phi_1, \dots, \phi_m\}$ define a family of granules $\{m(\phi_1), \dots, m(\phi_m)\}$, the strength of these two family of granules can be measured by the conditional entropy:

$$H(\Psi | \Phi) = \sum_{j=1}^m \sum_{i=1}^n P(\psi_i \wedge \phi_j) * \log P(\psi_i | \phi_j) \quad (10)$$

PARTITIONS AND COVERING

Partitions and coverings are two simple and commonly used granulations of universe. A partition consists of disjoint sub-sets of the universe, and a covering consists of possibly overlap subsets. Partitions are a special type of coverings. In granular computing, we treat each element of a partition or covering as a granule. Each granule can also be further divided through partition or covering.

Classification algorithm

(1) ID3

ID3 is an attribute oriented approach. Based on a measure of connection between two partitions, one selects an attribute to divide the universe into a partition. If an equivalence class does not belong to one user defined class, it is further divided by using another attribute. The process continues until one finds a decision tree that correctly classifies all objects. Each node of the decision trees labeled by an attribute, and each branch is labeled by a value of the parent attribute.

This top-down construction of a decision tree for classification searches for resolving a partition problem. If we search a covering solution, we must modify this decision tree to represent the results.

(2) PRISM

PRISM is an attribute-value pair oriented approach. PRISM generates rules from training set directly. The learning of PRISM algorithm is followed:

For $i=1$ to n

Repeat

- (1) calculate the confidence of class i for each attribute-value pair
- (2) select the attribute-value pair with the maximum confidence, and select all the instances with this attribute-value pair. Through this step create a subset of the training set.
- (3) Repeat (1) and (2) with this subset until it contains only instances of class i . The conjunction of all attribute-value pairs selected create a induced rule of class i
- (4) Remove all instance covered by this rule from training set.

Until no all instances of class I have been removed

It is easy to find that PRISM is a covering based method.

construction of a granule network

With the extension of ID3 and PRISM, Y.Y. Yao proposed a method of construction of a granule network to practice machine learning problem.

In a granule network, each node is labeled by a subset of objects. The arc leads from a larger granule to a smaller granules labeled by an atomic formula. In addition, the smaller granule is obtained by selecting those objects of the larger granule that satisfy the atomic formula. The family of the smallest granules thus forms a conjunctively definable covering of the universe.

Atomic formulas define basic granules, which serve as the basis for the granule network. The pair $(a = v, m(a = v))$ is called a basic concept. Each node in the granule network is a conjunction of some basic granules, and thus a conjunctively definable granule. The granule network for a classification problem can be constructed by a top-down search of granules.

APPLICATION

The precondition of our research

As we all know, text classification is a large and complex project. Automatic text classification includes the next six steps: set up data set(training data set and testing data set),text knowledge indexing, feature(key words) extracting and selecting and feature, design a classifier through machine learning, test the classifier with testing set, evaluate the method.

Many techniques are involved in automatic classification, such as feature extracting, feature selecting and classification algorithms. For a text, feature extracting is to extract important features from the text, such as extracting key words and topic words of the text. There are many typical methods to select the feature include DF (Document Frequency), IG(Information Gain), CE (Cross Entropy),MI(Mutual Information), χ^2 Statistics(CHI), WET (the Weight of Evidence), etc. These techniques are appropriate to treat very large feature spaces and a pre-processing step to reduce the feature dimensionality sufficiently.

In this paper, we provide an idea or method under the condition of a lower feature dimensionality.

Supposed we have an example of text training set is provided in the following.

Text1={feature1,feature2,feature3,feature4, feature5, feature7, feature8,feature9}
 Text2={ feature2, feature5, feature6,feature9}
 Text3={ feature2, feature4, , feature7, feature8,feature10}
 Text4={feature1,feature2, feature6, feature7, feature8,feature9}
 Text5={ feature4, feature5, feature7, feature8,feature10}
 Text6={feature1,feature3, feature5, feature7, feature10}
 Text7={feature2, feature7, feature8,feature9}
 Text8={feature1,feature2,feature3,feature4, feature5 }

Basic concept of text classification

We define $FS=\{Feature_1,Feature_2,\dots,Feature_n\}$, which is a set concluding all text features of training data. It is easy to obtain a database S. Let S be a set of texts, where each text T is a set of items such that $T \subseteq FS$.

The set of attributes is: $At = FS \cup \{class\}$

Because all objects with the same description over FS have the same class label, namely, if $I_{FS}(x) = I_{FS}(y)$, then $I_{class}(x) = I_{class}(y)$, it can conclude that $At = FS \cup \{class\}$ provide a consistent classification.

Table 1: Text information table.

TEXT	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	class
text1	y	y	y	y	y	n	y	y	y	n	1
text2	n	y	n	n	y	y	n	n	y	n	1
text3	n	y	n	y	n	n	y	y	n	y	2
text4	y	y	n	n	n	y	y	y	y	n	2
text5	n	n	n	y	y	y	y	y	n	y	2
text6	y	n	y	n	y	n	y	n	n	y	3
text7	n	y	n	n	n	n	y	y	y	n	3
text8	y	y	y	y	y	n	n	n	n	n	3

Construct a granule network for the text training set

The algorithm for constructing a granule network is summarized as follows.

- (1) **Construct** the family of basic concept with respect to atomic formulas
 $BC(U) = \{(a = v, m(a = v)) \mid a \in FS, v \in V_a\}$
- (2) **Set** the unused basic concepts to the set of basic concepts:
 $UBC(U) = BC(U)$
- (3) **Set** the granule network to $GN=(\{U\}, \varphi)$, which is a graph consists of only one node and no arc.
- (4) **While** the set of smallest granules in GN is not a covering solution of the classification problem:
 - (4.1) **Compute** the fitness of each unused basic concept.
 - (4.2) **Select** the basic concept $C = (a = v, m(a = v))$ with maximum value of fitness.
 - (4.3) **Set** $UBC(U) = BC(U) - \{C\}$

- (4.4) **Modify** the granule network GN by adding new nodes which are the intersection of $m(a=v)$ and the original nodes of GN; connect the new nodes by arcs labeled by $a=v$.

The algorithm is involved in a produce that is a refinement from the largest granule to the smallest one, which is belong to the same class.

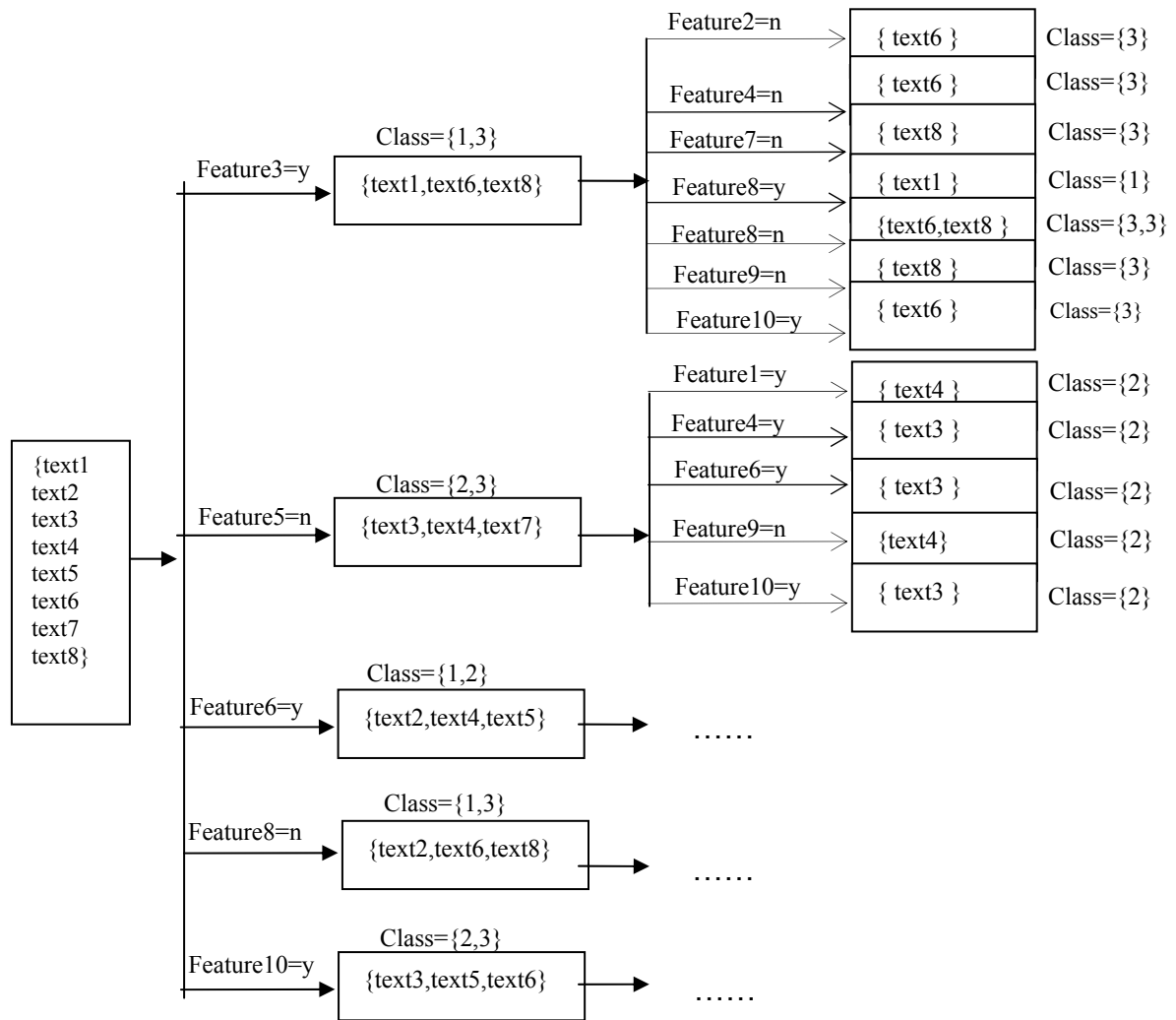
Table 2 summarizes the measures of basic concepts. First, select the minimum values of entropy. The range of entropy's values in this training set is from 0 to 2. When the value of entropy is minimum to near 0, it shows that the granule defined by this atomic formulas is the best suitable to select. There are five minimum entropies listed in table2, so they are chosen first.

Figure2 is a portion of the granule network of this training set. The data set in rectangle is the selected granule.

Table 2: Basic granules and their measures.

formula	granule	generality	confidence			entropy
			Class 1	Class 2	Class 3	
Feature1=y	{text1,text4,text6,text8}	0.5	0.25	0.25	0.5	1.5
Feature1=n	{text2,text3,text5,text7}	0.5	0.25	0.5	0.25	1.5
Feature2=y	{text1,text2,text3,text4,text7,text8}	0.75	0.3333	0.3333	0.3333	1.585
Feature2=n	{text5,text6}	0.25	0	0.5	0.5	1
Feature3=y	{{text1,text6,text8}	0.375	0.3333	0	0.6667	0.9183
Feature3=n	{ text2, text3, text4, text5, text7}	0.625	0.2	0.6	0.2	1.371
Feature4=y	{ text1, text3, text5, text8}	0.5	0.25	0.5	0.25	1.5
Feature4=n	{ text2, text4, text6, text7}	0.5	0.25	0.25	0.5	1.5
Feature5=y	{ text1, text2, text5, text6, text8}	0.625	0.4	0.2	0.4	1.5219
Feature5=n	{ text3, text4, text7}	0.375	0	0.6667	0.3333	0.9183
Feature6=y	{ text2, text4, text5}	0.375	0.3333	0.6667	0	0.9183
Feature6=n	{text1, text3, text6, text7, text8}	0.625	0.2	0.2	0.6	1.371
Feature7=y	{text1,text3,text4,text5,text6,text7}	0.75	0.1667	0.5	0.3333	1.4591
Feature7=n	{ text2, text8}	0.25	0.5	0	0.5	1
Feature8=y	{ text1, text3, text4, text5, text7}	0.625	0.2	0.6	0.2	1.371
Feature8=n	{ text2, text6, text8}	0.375	0.3333	0	0.6667	0.9183
Feature9=y	{ text1, text2, text4, text7 }	0.5	0.5	0.25	0.25	1.5
Feature9=n	{ text3, text5, text6, text8}	0.5	0	0.5	0.5	1
Feature10=y	{ text3, text5, text6 }	0.375	0	0.6667	0.3333	0.9183
Feature10=n	{ text1, text2, text4, text7,text8 }	0.625	0.4	0.2	0.4	1.5219

Figure 2: A portion of the granule network.



CONCLUSION

The test result shows that the granular computing is a proposed approach to solve the problem of text classification. Because the value of attribute of each feature is simple, just yes or no, it will simplify the produce of computing. Although the classification rules may have overlaps with each other, they may be shorter than the rules obtained from classical decision tree methods.

REFERENCES

- JiHe, Ah-HweeTan, Chew-Lim Tan, (2000). A Comparative Study on Chinese Text Categorization Methods. PRICAI Workshop on Textand Web Mining,24-35.
- Sebastiani F., (2000). Machine learning in automated text categorization. *ACM Computing Suerveys*,1-47.
- Taorong Qiu,Xiaoqing Chen,Qing Liu,Houkuan Huang, (2006). Granular computing based text classification. 2006 IEEE International Conference , 313- 316.
- Yao.J.T, Yao.Y.Y., (2002). A granular computing approach to machine learning. Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),732-736.
- Yao.J.T, Yao,Y.Y.(2002). Induction of classification rules by granular computing. Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing, 331-338.
- Yao.Y.Y.(2001). On Modeling data mining with granular computing. Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC), 638-643.
- Yao.Y.Y, Yao.J.T.(2002). Granular computing as a basis for consistent classification problems. Proceedings of PAKDD, 101-106.
- Zhao.Y, Yao.Y.Y.(2005). Interactive user-driven classification using a granule network. Proceedings of the Fifth International Conference of Cognitive Informatics (ICCI'05), 250-259.