8—36

Chinese Character Classification Based on Rough Set and SVM Algorithm.

Fan Jinsong*
Department of Electronic Engineering
Uni. of Science & Technology of China

Fang Tingjian †
Hefei Institute of Intelligent Machines
Chinese Academy of Sciences

Abstract

In the paper, we present a integrated approach combined Rough Set theory and SVM algorithm. The approach will be divided into two steps. The first step is classified roughly with Rough Set, rule should be induced in this step by information system. The second step should be classified precisely based on SVM Algorithm, in this step we present two new fundamental principles to help us select basic attributes for SVM algorithm. In virtue of Rough Set and SVM, we can identify characters fast and well. The paper gives handwriting Chinese as an example to show that the method can be used practically.

1 Introduction

As a new tool for analyzing information systems [1], Rough Set theory has been used widely in KDD, signal processing and characters classification [2,3,4,5]. It is fast and easy to deal with information system, even the quantified information in system is uncertainty and vagueness. The weakness of rough set is appeared when the ratio of classification is highly required.

Comparing with Rough Set theory, the SVM algorithm is a new machine learning method[6], which can identify characters correctly with higher ratio of classification.

The SVM algorithm, which overcame some fatal weakness of Neural Networks, had single universally accepted theoretical framework, and well applied for predictive learning. It has already applied to human face recognition [7], KDD [8] and signal processing [9],

The SVM algorithm is resulted in the constructive learning methodology based on Quadratic Programming (QP) optimization technique [10]. The time of calculating increases rapidly with dimensions or number of samples increased, so it is emergent to be solved.

In the paper, we present an integrated approach combined Rough Set theory and SVM algorithm to solve above problems. The approach will be divided into two steps. The first step is classified roughly with Rough Set, rule should be induced in this step by information system. The second step should be classified precisely based on SVM Algorithm, in this step we present two new foundational principles to help us select basic attributes for SVM algorithm. In virtue of Rough Set and SVM, we can identify characters fast and well. The paper gives handwriting Chinese as an example to show that the method can be used practically.

2 Rough Classification based on Rough Set

Rough Set theory was introduced by Z.Pawlak as a tool to deal with uncertainty and vagueness in the analysis of information systems [1]. An information system is a formal representation of the analyzed data set and is defined as a pair $S = (U, A)$ where U is a finite set of objects and A is a finite set of attributes.

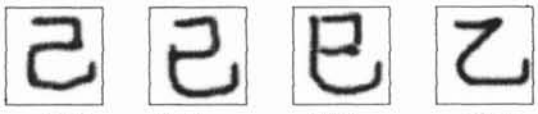
*Address:P.O.Box 1130,Hefei, Anhui, 230031, China.,
E-mail: jsfan@mail.iim.ac.cn

† Address: P.O.Box 1130,Hefei,Anhui,230031, China.
E-mail: tjfang@mail.iim.ac.cn

Set of values V_a is associated with every attribute $a \in A$,. Each attribute 'a' determines a function $f_a : U \to V_a$, In practice, we are mostly interested in discovering dependencies in a special case of information system called attribute. The decision attribute determines the partition of U into k disjoint classes $X_1, X_2, ..., X_k$ (where k is a number of different values of attribute d) called decision classes.

The rough set theory is based on an observation whose objects may be indiscernible due to limited available information. For a subset of attributes $B \subseteq A$ the indiscernibility relation is defined by $Ind(B) = \{(x,y) \in U \times U : f_a(x) = f_b(y), \forall a \in B\}$. The classes of this relation are called B-elementary sets. An elementary equivalence class (i.e. single block of the partition $U / I(B)$) containing element x is denoted by $I_B(x)$.

Figure 1 gives an sample of four similar Chinese Characters. We will explain how to classify roughly by Rough Set.



"己" "已" "巳" "乙"

Figure 1 84x84 handwriting Chinese characters

We can use image technology--SVD (Singular Value Decomposition) algorithm to obtain basic features vectors in table 1, whose dimension is 74x1. To classify those vectors and induce their rules an integrating symbolic learning and statistical analysis technique[11] is applied .

Tab.1 Features of four similar handwriting characters

U	X_1	X_2	X_3	...	Y
"己"	33.58	15.52	9.32	...	a_1
"己"	34.52	15.90	8.85	...	a_2
"已"	33.31	17.12	9.86	...	b_1
"已"	33.87	16.71	9.79	...	b_2
"巳"	35.06	13.02	10.24	...	c_1
"巳"	35.75	14.62	11.66	...	c_2
"乙"	29.25	14.77	8.16	...	d_1
"乙"	30.30	14.80	10.05	...	d_2

We can transfer from table 1 to table by quantified method. Table 2 is a kind of information system based on equal distance between values. For other quantified methods one may refer to paper [4].

Tab.2 Information system of handwriting characters

U	X_1	X_2	X_3	...	Y
"己"	δ_{17}	δ_8	δ_5	...	a_1
"己"	δ_{18}	δ_8	δ_5	...	a_2
"已"	δ_{17}	δ_9	δ_5	...	b_1
"已"	δ_{17}	δ_9	δ_5	...	b_2
"巳"	δ_{18}	δ_7	δ_6	...	c_1
"巳"	δ_{18}	δ_8	δ_6	...	c_2
"乙"	δ_{15}	δ_8	δ_5	...	d_1
"乙"	δ_{16}	δ_8	δ_6	...	d_2

Now we use the indiscernibility relation to classify these characters according to different attributes:

$$U / ind(X_1) = \{\{a_1, b_1, b_2\}, \{a_2, c_1, c_2\}, \{d_1\}, \{d_2\}\} \quad (2\text{-}1)$$

Formula (2-1) classifies the system into four groups: $\{a_1, b_1, b_2\}$, $\{a_2, c_1, c_2\}$, $\{d_1\}$ and $\{d_2\}$.

$\{a_1, b_1, b_2\}$ and $\{a_2, c_1, c_2\}$ are indiscernible. If below rules are defined, we can identify character " 乙 " from others.

$$X_1 \mid \delta_{15} \vee X_1 \mid \delta_{16} \to \{"乙"\} \quad (2\text{-}2)$$

$$X_1|\delta_{17} \vee X_1|\delta_{18} \rightarrow \{"\square","\square","\square"\} \qquad (2\text{-}3)$$

If database of character very large, all these rules, which induced from several information system, will classify database into lots of small groups. Each group may contains similar handwriting characters. Then we can start following step for precise classification.

3 Precise classification based on SVM algorithm

SVM (Support Vector Machines) algorithm is a new and promising classification technique developed by Vapnik and his groups of AT&T Bell Laboratories [6,7,9,10,12]. This new leaning algorithm can be seen as an alternative training technique for Polynomial, Radial Basis Function and Multi-Layer Perceptron classifiers. The main idea behind the technique is to separate the classes with a surface that maximizes the margin between them. An interesting property of this approach is an approximate implementation of the Structural Risk Minimization (SRM) induction principle[6,7]. The derivation of Support Vector Machines-- its relation with SRM and geometrical insight, are discussed in papers [7,9,10,12]

Similar with step one, in step two rules can be induced from several important attributes to make a decision. To compare with step one it is based rather on original message inside features than quantified information. How to choose these important attributes is a key of step two.

Here we present two fundamental principles: Maximum Positive Margin Principle and Minimum Negative Margin Principle[13].

1. Maximum Positive Margin Distance Principle: The positive margin is equal to twice distance between the nearest spots and hyper-plane. The selected attributes with the bigger positive margin have better results.
2. Minimum Negative Margin Distance Principle: The negative margin is equal to sum distance between all mistaken spots and hyper-plane. The selected attributes with the smaller negative margin have better results.

According to above fundamental principles, we can select basic attributes to be classified precisely by SVM. Here we give an example of Table 1 to show how to select attributes.

First we select "linear" as a kernel of SVM algorithm. Then $\{X_1,X_2\}$ and $\{X_1,X_5\}$ are chosen as basic attributes, which can be classified by linear hyper-plane. It means that empirical risk is minimized, and a pair of attributes with bigger positive margin distance is better basic attributes we should find. Finally we use SVM algorithm to train model of selected attributes. Both the trained model of SVM and the selected attribute are saved into database for further identification.

Figure 2 gives a result of classification with attributes $\{X_1, X_5\}$ and Figure 3 is with $\{X_1, X_2\}$. The positive margin distances of Figure 2 are [0.4396, 0.0866], and ones of Figure 3 are [0.4562, 0.4139]. Another 30 sample to be tested shows the ratio of classification with $\{X_1, X_2\}$ is 85%, and with $\{X_1, X_5\}$ is 70%. If we want to increase the ratio, we can change kernel of SVM, and integrate it with Rough Set to construct a compound classifier [14].

4 Conclusion

Combining SVM algorithm with Rough Set can be used to classify handwriting Chinese characters, and it has good results. It's not only used in characters, but also applied in other classification and regression

estimation.

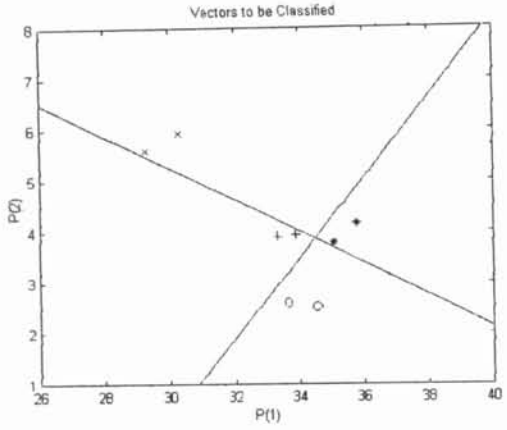


图 2 classified characters with $\{X_1, X_5\}$

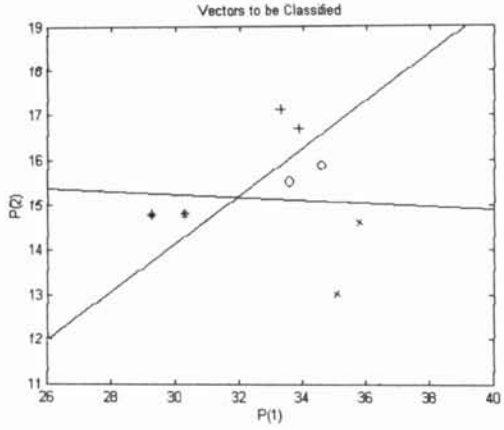


图 3 classified characters with $\{X_1, X_2\}$

REFERENCES

[1] Z.Pawlak, "Rough Sets.Theoretical Aspects of Reasoning about Data." Kluwer Academic Publishers, Dordrecht, 1991

[2] Jaroslaw Stepaniuk , "Attribute Discovery and Rough Sets," Principles of Data Mining and Knowledge Discovery , 145-155, 1997

[3] Wojciech Kowalczyk and Rrank Slisser, "Modeling Customer Retention with Rough Data Models ", Principles of Data Mining and Knowledge Discovery ,4-13, 1997。

[4] Zeng Huangling, "Rough Set Theory and Its Application", press of. Chongqing University, China, 1995

[5] Nguyen, "Classification based on Optimal Feature Extraction and the Theory of Rough Sets", Thesis of Master Degree, SDSU, 1995

[6] V.Vapnik. "The Nature of Statistical Learning Theory", Springer-Verlag. 1995.

[7] E.Eosuna, R.Freund and F.Girosi. "Support Vector Machines: Training&Applications" A.I.Memo 1602 ,MIT AI Lab, 1997

[8]L.Guyon, N.Matic, and V.Vapnik, "Discovering Information Pattern and Data Cleaning",AT&T Bell Lab,1997

[9] V.Vapnik, Golowich and Smola, "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing", AT&T Bell Laboratory ,1999

[10] C.Cortes and V.Vapnik. "Support Vector Networks." Machine Learning ,20:1-25,1995.

[11] I.F.Iman, R.S.Michalski, L.Kerschberg "Discovering Attribute Dependence In Databases by Integrating Symbolic learning and Statistical Analysis Techniques" Proc. of Knowledge Discovery in Databases Workshop, AAAI-93 (1993)264-275

[12] V.Vapnik "Estimation of Dependencds Based on Empirical Data Springer-Berlag", 1982,p364

[13] Fan Jinsong, Fang Tinjian, "Analysis and Evaluation on main factors for Feature Selection and Abstraction", Computer Engineering and Application, Beijing, China

[14] Fan Jinsong, "Research of SVM Algorithm and its Application", thesis of Ph.D. degree, USTC, Nov.2000