Feature Selection Based on Correlation between Fuzzy Features and Optimal Fuzzy-Valued Feature Subset Selection

Jirong Li
North China Electric Power Univercity, China
little_deer13@163.com

Abstract

Feature selection plays an important role in classification or recognition. The aim of feature selection is to reduce the number of features used in classification. In the whole feature space, there might be strong correlation between the features. Feature selection based on information theory is proposed for avoiding redundant features. However, such algorithm only focuses on the case that the feature values are discrete. This paper proposes a method includes correlation between features based on fuzzifying the numeric-value features. In this paper, we suggest a method of constructing compact feature space before feature selection. It aims at removing redundant features which may be correlative with some other features in the original feature space and improvements in classification performance.

1. Introduction

Methods of feature selection try to find a subset of features that should lead to the best results. A general approach to feature selection explicitly attempts to select "most relevant" subset of features. This technique aims to reduce the entire feature space to a highly predictive subset of the space. These methods treat each feature in an independent way, trying to determine how useful they may be for the classification. This reduction may improve the performance of data mining algorithms to be used, in terms of speed, accuracy, and simplicity. The previous study on relation-based feature selection is based on a similar principle: features should be highly related with the target but not with each other. Relevant to the target could be described as:

A feature x_i is relevant to a target concept c if there exists a pair of examples A and B in the instance space such that A and B differ only in their assignments to x_i and $c(A) \neq c(B)$.

However in many database used in practice, there might be strong correlation between the features that describe the instance space. That is, value of one feature may be consistent with the value of another one. If the two features have great denotation to the classification, then both of them will be in the feature subset after feature selection, this introduces redundancy. If values of one feature could predict values of the other one, then they are correlative. Correlation between features could be described as:

A feature x_i is correlative to another feature x_j if there exists a pair of examples A and B in the instance space such that A and B differ in their assignments to x_i and x_j $(A) \neq x_j$ (B).

Paper[1][2] estimate correlation between features using information method. Information theory could help to determine the correlation between the discrete-value features properly, while it is not adapted to the numeric-value features. Paper[6] adopts the method of fuzzifying the numerical numbers into linguistic terms, which could be represented by fuzzy sets. Compared with the simply discretization method, it avoids data information losing effectively.

This paper proposes a method that includes correlation between features based on fuzzifying the numeric-value features into linguistic terms. We begin with the problem of filtering features based on correlation between features. We then turn to the problem of performing feature selection on the compact representative feature subset based on optimal fuzzy-valued feature subset selection (OFFSS). We apply the method to the fuzzy decision tree induction. The experimental data shows that the method is feasible and results in the low-dimensional data with good classification accuracy.

978-0-7695-3278-3/08 $25.00 © 2008 IEEE
DOI 10.1109/IIH-MSP.2008.292

2. Correlation between Features

Analysis of the correlation between numerical data needs the help of fuzzifying method. Fuzzy clustering algorithm is one of the methods, which can fuzzify numerical data into linguistic terms based on self-organized learning and can generate some type of membership functions[3]:

Let X be the considered data set. We intend to cluster X into k linguistic terms T_i, ($i=1, 2, \cdots, k$). For simplicity, we assume the shape of membership function to be triangular. These linguistic terms T_i will have triangular membership functions as follows:

$$T_1(x)=\begin{cases}1 & x \le m_1 \\ (m_2-x)/(m_2-m_1) & m_1 < x < m_2 \\ 0 & x \ge m_2\end{cases}$$

$$T_i(x)=\begin{cases}0 & x \ge m_{i+1},\ x \le m_{i-1} \\ (m_{i+1}-x)/(m_{i+1}-m_i) & m_i < x < m_{i+1}, 1 < i < k \\ (x-m_{i-1})/(m_i-m_{i-1}) & m_{i-1} < x < m_i\end{cases}$$

$$T_k(x)=\begin{cases}1 & x \ge m_k \\ (x-m_{k-1})/(m_k-m_{k-1}) & m_{k-1} < x < m_k \\ 0 & x \le m_{k-1}\end{cases}$$

Each pair of adjacent membership functions crosses at the membership value 0.5. Parameters needed to be determined are k centers $\{m_1, m_2, \cdots, m_k\}$. An effective method to determine these centers is Kohonen-feature-maps algorithm. At initial time, k centers are set to be distributed evenly on the range of X. Let:

$$M=\{m_1, m_2, \cdots, m_k\}$$

$$d(X,M)=\sum_{x\in X}\min|x-m_i|$$

The centers will be adjusted iteratively. Each iteration consists of three steps:

(1) randomly take a value x from X, denoted by $x[n]$;

(2) search for an integer i such that $|x[n]-m_i[n]|=\min_j|x[n]-m_j[n]|$;

(3) put $m_i[n+1]=m_i[n]+\alpha(x[n]-m_i[n])$ and keep other centers unchanged.

Where n is the iteration time and α is the learning rate. The iteration ends when $d(X, M)$ converges.

After fuzzifying the feature space, each feature value could be regarded as a fuzzy set, which has linguistic meaning.

If the possibility is nearly 1, when the value of feature A is A_i while the value of feature B is B_j, this denotes that feature A and feature B are correlative. So they are mutually redundant and one of them stays in the subset is enough. Base on the fuzzy-value data, the correlation between any two features can be measured according to the similarity degree of fuzzy sets.

Definition 1: Suppose X is a given domain of discourse, $x \in X$. A, B are two fuzzy sets, $A(x)$, $B(x)$ are their membership grades, A(x), B(x) $\in$ [0,1], the similarity measure formula is defined as:

$$SM(A,B)=\max_{x\in X}(\min(A(x),B(x)))^2$$

Based on the similarity measure we define the correlation between the features, and then divide them into groups. Correlation measure of any pair of features in the entire feature space could be defined on the measure method of the fuzzy set similarity. In the following, any feature in the feature space is denoted as *FeatureX* and its value is denoted as X_i.

Definition 2: Suppose *FeatureA* and *FeatureB*, where *FeatureA* has n kind of values, *FeatureB* has k kind of values. Values of *FeatureA* when *FeatureB* gets the value B_i ($i=1, 2, \cdots, k$) are denoted as:

$$A_m^i \qquad (m=1, 2, \cdots, n)$$

Definition 3: Suppose A, B are two fuzzy sets, the correlation vector responds to A and B is defined as:

$$CV(A, B)=r,\quad r=SM(A,B).$$

Definition 4: The correlation vector of A_m^i and B_j is defined as $CV(A_m^i, B_j)$, here:

$$CV(A_m^i,B_j)=\begin{pmatrix}CV(A_m^i,A_1^j)\\ \bullet \\ CV(A_m^i,A_n^j)\end{pmatrix}$$

The correlation vector of B_i and B_j is defined as $CV(B_i, B_j)$, here:

$$CV(B_i,B_j)=\begin{pmatrix}CV(A_1^i,B_j)\\ \bullet \\ CV(A_m^i,B_j)\end{pmatrix}$$

The correlation vector of each pair of *FeatureB* values is defined as $CV(B_1, B_2, \cdots, B_k)$, here:

$$CV(B_1, B_2, \bullet \quad B_k) = \begin{pmatrix} CV(B_1, B_2) \\ CV(B_1, B_3) \\ \bullet \\ CV(B_1, B_k) \\ \bullet \\ CV(B_{k-1}, B_k) \end{pmatrix}$$

When *FeatureA* gets a certain value A_m, *FeatureB* gets a certain value B_i most time. While when *FeatureA* gets other values, *FeatureB* seldom gets the same value B_i. This could be adjusted by the similarity measure of A_m^i and A_t^j ($m,t=1, 2, \cdots, n$). If the similarity measure is low，it implicates that *FeatureA* and *FeatureB* are correlative. According to definition4, the correlation vector $CV(B_1, B_2, \cdots, B_k)$ illustrates the similarity measure of A_m^i and A_t^j , so we define correlation between them base on their correlation vector. Suppose β is a given threshold value, if any value in the correlation vector $CV(B_1, B_2, \cdots, B_k)$ is less-or-equal to β, then it is named as a valid element of the vector. The correlation between *FeatureA* and *FeatureB* could be adjusted according to the number of valid elements. The main idea is:

If *FeatureA* and *FeatureB* are correlative, then there must be more valid elements in the vector. Else if there has poor correlation between *FeatureA* and *FeatureB*, then there must be less valid elements in the vector.

Definition 5: For a given example set E, $n=|E|$ is the number of all the examples. Suppose *FeatureA* and *FeatureB* are two features in the feature space, $CV(A, B)$ is their correlation vector, $NUM_{A,B}$ is the number of valid elements in the vector. Under a given threshold value $\beta \in (0.5,1]$, if $NUM_{A,B}/n \geq \beta$ then *FeatureA* and *FeatureB* are correlative, could be grouped.

According to definition 5, feature space compacting method is defined as following, which supposes the entire feature space set is *ES*, *CS* is the compact feature set, original value is null.

Step 1: For each feature *FeatureX* calculate the correlation vectors toward others and count the number of valid elements.

Step 2: Group such features with *FeatureX*, that the number of valid elements toward *FeatureX* is greater than β. Call them *GroupX*,;

Step 3: Find the group which has the most features, and use F_i to represent all the members in the group, then delete the group of features from *ES*, *ES-GroupX*, $CS = CS \cup FeatureX$;

Step 4: Repeat the process on the current *ES* until all the features are divided into a group, then *CS* is the reduced feature set.

3. Feature Selection Based on the Compact Feature Space

After reducing the redundancy of the feature space, we get the compact feature set, and then use the OFFSS method to produce the feature selection.

OFFSS method adopts the overlapping defined on the similarity measure of the fuzzy sets, to describe the partition extent of positive and negative class. The kernel idea is: consider an example set, each example consists of several features and a classification value. The feature value is defined on fuzzy subset. The classification value is determinately positive or negative. The whole example set is divided into two sets, positive set P and negative set N. Build the extension matrix $EM(P,N)$ of P with respect to N. The elements of $EM(P,N)$ is the *SM* between P and N. The element which value is less than a given threshold value T is called $under_T$ element. Select one $under_T$ element from each row we will get a path of $EM(P,N)$. Find an optimal feature subset is to search for such a path involves the minimum numbers of columns.

Algorithm process of OFFSS method based on the correlation between features is as following[6]:

Step 1: Call the algorithm of analysis the correlation between fuzzy features, for the whole feature space to analysis the feature correlation, get the compact feature space *CS*.

Step 2: Suppose S is the is the feature subset to be searched, at first it is empty, $S = \Phi$. P is the given positive class; N is the given negative class; $EM(P,N)$ is the current extension matrix of P with respect to N; and S is initially set to an empty set.

Step 3: From the current extension matrix $EM(P,N)$, find a column with the most $under_T$ element. Use F_j to denote this column, and then replace S with $S \cup \{F_j\}$.

Step 4: From $EM(P,N)$, remove the rows which include an $under_T$ element in the selected *j-th* column, and then form a new $EM(P,N)$, which is regarded as the current extension matrix.

Step 5: If $EM(P, N)$is empty, then regard S as the final result [stop]; else, go to Step 2.

4. Result and Conclusion

We applied the method into the induction of the fuzzy decision tree and chose five typical databases. Each selected dataset is first split into two parts, the training and testing sets by randomly choosing examples. For all datasets 80% of the examples are randomly selected as the training set and the remainder as the testing set. Fuzzy decision tree induction is an important way of learning from fuzzy examples. We would like to use fuzzy decision tree induction to check the performance difference between before and after feature space compacted, and to compare the performance of our proposed method.

As shown in Table1 feature subset selection on the whole feature space and the compact feature space are different. Table 2 shows the result of the compare of the two methods in the fuzzy decision tree induction.

Table 1. Feature subset selection on the entire feature space(ES) and the compact feature space(CS).

	Breast Cancer	Wine	Pima	Glass	Derm
ES	{6,2,1,8,5, 4,9}	{1,7,4,8, 13,2,11}	{2,6,1,7}	{3,9,5,4, 1,7}	{16,4,34 ,19,23}
CS	{6,3,8,7}	{8,2,11,4}	{2,8,7}	{9,4,2,7}	{16,34,23}

Table 2. Performance comparison between the ES and the CS feature selection.

		Breast Cancer	Wine	Pima	Glass	Derm
ES	Train	0.890	0.935	0.742	0.867	0.970
	Test	0.857	0.908	0.745	0.720	0.950
CS	Train	0.901	0.937	0.744	0.884	0.972
	Test	0.893	0.908	0.746	0.729	0.953

In the paper, we propose a method of analysis the correlation between features, and compare the result of using compact feature set to perform feature selection with the result of using the original feature space. The result is optimal fuzzy-valued feature subset selection based on the features correlation is helpful to filter the correlative features, which are all have great denotation to the classification in the real application.

We can see clearly from Table2, using compact feature set to perform feature subset selection instead of the original feature space, we can get higher testing accuracy, that is to say we reduce the redundant features to form a more highly predictive feature subset, then can get a better inductive rules through the decision tree to cover the positive examples and exclude the negative examples.

It is well known that feature selection is a time consuming process. However, no efforts have been made to analyze the computational complexity of the method. Our paper introduces a separate process for the purpose of finding the most compact and useful features, that occurs before the basic induction step. The preprocessing step uses general characteristics of the training set to select some features and exclude others. Because they filter out irrelevant attributes before induction occurs, thus, filter methods are independent of the induction algorithm that will use their output, and it can be combined with any such method.

5. References

[1] REN Jiang-Tao, SUN Jing-Hao, HUANG Huan-Yu, YIN J Jan, "Feature Selection Based on lnformation Gain and GA", Computer Science, vo1.33, no.10, pp.193-195.

[2] Wtodzistaw, Krzysztof Grqbczewski, Tomasz Winiarski, "Feature Selection Based on Information Theory, Consistency and Separability Indices", Proceedings of the 9th international conference on neural information processing, vol.4, pp.1951-1955.

[3] X.Z. Wang, Y.D. Wang, X.F. Xu, W.D. Ling, D.S. Yeung, "A new approach to fuzzy rule generation: fuzzy extension matrix", Fuzzy sets and systems, vol.123, pp.291-306.

[4] Wang Xi-Zhao，Lu Xiao-Ying, Zhang Feng, "Feature Selection Based On Fuzzy Extension Matrix For Multi-class Problem", Proceedings of 2004 International Conference on Machine Learning and Cybernetics, Shanghai, April 2004, pp.2032-2035.

[5] R.K.De, N.R.Pal, and S.K. Pal, "Feature analysis: Neural network and fuzzy set theoretic approaches", Pattern Reorganization, vol.30, no. 10, pp.1579-1579.

[6] E.C.C. Tsang, D.S.Yeung, X.Z.Wang, "OFFSS: Optimal Fuzzy-valued feature Subset Selection", IEEE Transactions on Fuzzy Systems, vol. 11, no. 2, pp.202-213.

[7] R.Weber, "Fuzzy-ID3: A class of methods for automatic knowledge acquisition," in Proc. 2nd Int. Conf. Fuzzy Logic Neural Networks, Iizuka, Japan, July 17-22, 1992, pp. 265-268.